CALIBER: Benign-Calibrated Dual-Score Rejection for Label-Scarce Open-World Sequence Classification

Jiayu Li
East China Normal University
Shanghai, China
51275902029@stu.ecnu.edu.cn

Xiangxue Li*
East China Normal University
Shanghai, China
xxli@cs.ecnu.edu.cn

***Abstract*—Open-world sequence classification under label scarcity requires not only accurate recognition of known categories but also reliable rejection of previously unseen patterns under an explicit false-alarm budget. This challenge arises prominently in in-vehicle CAN message streams, where attack labels are limited and new attack behaviors may emerge after deployment, causing closed-set detectors to misclassify unknowns as known types. We propose CALIBER, a label-efficient framework that couples lightweight temporal representation learning with calibrated unknown rejection. CALIBER converts raw bus frames into fixed-length token sequences by fusing payload content, identifier cues, and short-horizon timing and rate signals, enabling a compact sequence encoder to learn discriminative representations from limited supervision. At inference time, CALIBER performs calibrated rejection via a dual-score scheme that combines an energy score from classifier logits with a representation-space Mahalanobis distance. Rejection thresholds are determined using benign-only quantile calibration, with optional mode-aware indexing to improve operating-point stability. Experiments on two public in-vehicle datasets with leave-one-attack-out zero-day evaluation show near-saturated recognition on known traffic while consistently rejecting unseen attacks at low benign false-alarm rates.**

***Index Terms*—open-set recognition, self-supervised learning, temporal representation learning, in-vehicle networks**

I. Introduction

Modern vehicles are networked cyber-physical systems where many electronic control units exchange time-critical messages over embedded in-vehicle networks. The Controller Area Network (CAN) remains a dominant in-vehicle bus, and once an adversary gains access via physical interfaces or compromised gateways, message injection and manipulation on the bus become feasible [1], [2]. CAN traffic is organized as broadcast frames. As shown in Fig. 1, a frame includes an arbitration identifier, a length field, and a payload, followed by check and delimiter fields. This structure induces identifier-conditioned regularities in content and timing, shaping how abnormal behaviors appear and how models characterize normal traffic.

Many learning-based CAN IDSs are trained and evaluated in closed-set settings where test-time attacks are covered by the training taxonomy, yet real deployments face scarce labels, shifting conditions, and continual emergence of unseen attacks [1], [2]. In such settings, classifiers may over-confidently map unseen attacks into known categories. Deployment also requires controlling benign false alarms under a specified budget, where quantile-based calibration provides a natural mechanism for threshold selection [3].

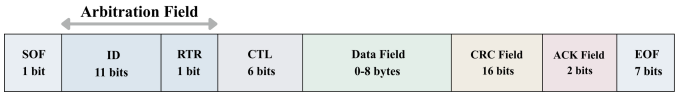

Fig. 1. Simplified field format of CAN data frames.

Motivated by these constraints, we propose CALIBER, a label-efficient open-world IDS for CAN traffic that integrates lightweight temporal tokenization with calibrated rejection. CALIBER converts raw frames into fixed-length token sequences by fusing payload content, identifier and structural cues, and short-horizon timing and rate signals, enabling a compact sequence encoder to learn discriminative representations from limited supervision. At inference time, CALIBER combines an energy score from logits with a representation-space Mahalanobis distance and selects thresholds using benign-only quantile calibration with optional mode-aware indexing for operating-point stability.

The main contributions are as follows:

- We propose CALIBER, a label-efficient open-world IDS for CAN traffic that converts raw frames into temporal tokens and learns compact sequence representations with a lightweight encoder.
- We introduce a dual-score rejection mechanism using energy and Mahalanobis distance, with benign-only quantile calibration to control false alarms without attack labels.
- We evaluate CALIBER under open-world settings including leave-one-attack-out testing, operating-point sweeps, label-budget scaling, and stability analysis, demonstrating consistent rejection of unknown attacks under a fixed false-alarm budget.

The implementation of CALIBER is publicly available at

*Corresponding author.

https://github.com/Kerwin828/CALIBER.

II. Related Work

Learning-based intrusion detection for in-vehicle CAN traffic has evolved from classical machine-learning pipelines to deep sequence models that exploit identifier-conditioned regularities in payload and timing [1], [2], [4]. Recent surveys and systematizations further summarize the design space of feature construction, temporal modeling, and deployment-oriented constraints for CAN-based IDSs [1], [2], [4], [5]. A common theme across these works is the use of multi-view cues derived from frame content and temporal dynamics to improve robustness under diverse driving conditions.

To reduce reliance on labeled attacks, label-efficient training has been widely explored through semi-supervised, self-supervised, and unsupervised objectives that learn from predominantly unlabeled traffic, using limited labels only for light supervision or downstream decision rules [6]–[9]. These methods often emphasize representation learning on benign data and limit supervised components to lightweight classifiers or thresholding rules. While they improve practicality under sparse annotations, open-world deployments still require mechanisms that handle attacks outside the training taxonomy.

Accordingly, recent work increasingly frames CAN intrusion detection as open-set recognition and unknown-attack awareness [10], [11]. Related scoring principles for unknown rejection include confidence-based baselines, representation-space deviation, and energy-based uncertainty, which provide complementary signals for distinguishing known from unknown inputs [12]–[14]. Beyond separation quality, operational IDS deployment also requires explicit control of benign false alarms. Conformal-style quantile calibration provides distribution-free thresholding, and analyses under non-exchangeable or shifted settings clarify why achieved false-alarm rates can deviate from nominal targets [3], [15], [16].

III. Threat Model

We assume an attacker can access the in-vehicle CAN bus via physical interfaces or a compromised component, enabling passive observation and active injection of crafted CAN frames. The bus provides no per-frame authentication, so injected frames contend in arbitration and may be received by ECUs as ordinary messages. Evaluation follows a label-scarce open-world setting where attacks encountered after deployment may fall outside the training taxonomy and decisions are constrained by a benign false-alarm budget.

Rate-based injection: High-frequency injections increase bus utilization and reduce transmission opportunities for legitimate messages, degrading time-critical control communication.

Fuzzing: Frames with randomized identifiers and payload bytes disrupt identifier-conditioned regularities, yielding atypical value distributions and short-range temporal inconsistencies.

Spoofing: Legitimate identifiers are impersonated while payload semantics are altered. Frames remain structurally valid but encode incorrect states, inducing systematic deviations in content and timing correlations.

Replay: Previously observed frames are re-transmitted in mismatched contexts. Individual frames may appear benign, but the resulting sequences violate expected temporal alignment and state evolution.

These attack families provide concrete known conditions for evaluation but do not cover the full space of manipulations. Injection intensity can vary and previously unseen identifiers may be targeted, producing samples outside the known taxonomy. This motivates unknown rejection under a controlled false-alarm budget.

IV. Methodology

CALIBER adopts an offline–online design for unknown-attack-aware intrusion detection under limited supervision. The method has three phases. Phase 0 constructs a shared tokenized representation from raw bus traffic. Phase 1 performs offline training and benign-only calibration to produce model and threshold artifacts. Phase 2 applies these artifacts for online scoring and rejection.

A. Phase 0: Shared Input Construction & Tokenization

As shown in Fig. 2 Phase 0, a time-ordered CAN stream is converted into a fixed-shape token sequence that is reused in both offline training and online inference. Let the i-th frame be $x_i = \langle \mathrm{id}_i, \mathrm{pay}_i, \ell_i, t_i \rangle$, where id_i is the identifier, pay_i is the payload byte sequence, ℓ_i is the payload length, and t_i is the timestamp. At step t, we form a fixed-length window of the most recent L frames

$$W_t = \{x_{t-L+1}, x_{t-L+2}, \ldots, x_t\}. \tag{1}$$

If fewer than L frames are available, we pad missing positions and record a binary mask $\mathrm{mask}_t \in \{0,1\}^L$ so the encoder can ignore padded tokens.

For each element x_{t-L+j} in W_t with $j \in \{1, \ldots, L\}$, we build a token by concatenating seven complementary views:

- **Payload content** e_j^{pay} as compact byte-derived channels
- **Identifier cue** e_j^{id} as a normalized scalar from id
- **Length cue** e_j^{len} as a normalized scalar from ℓ
- **Within-window position** e_j^{pos} as a normalized scalar from j
- **Bus-level timing** δ_j^{bus} as a normalized inter-arrival gap
- **ID-specific recency** δ_j^{id} as a normalized time since the last occurrence of the same identifier
- **ID-window frequency** ρ_j^{id} as a normalized count of the identifier within the window

Phase 0 uses deterministic and parameter-free encodings to ensure strict offline–online consistency. Payload content is mapped by a lightweight fixed encoder

$$e_j^{\mathrm{pay}} = \mathrm{Enc}_{\mathrm{pay}}\left[\mathrm{pay}_{t-L+j}\right], \tag{2}$$

while NormId, $\mathrm{NormLen}$, and $\mathrm{NormPos}$ apply fixed scaling to map identifier, length, and position to comparable numeric ranges.

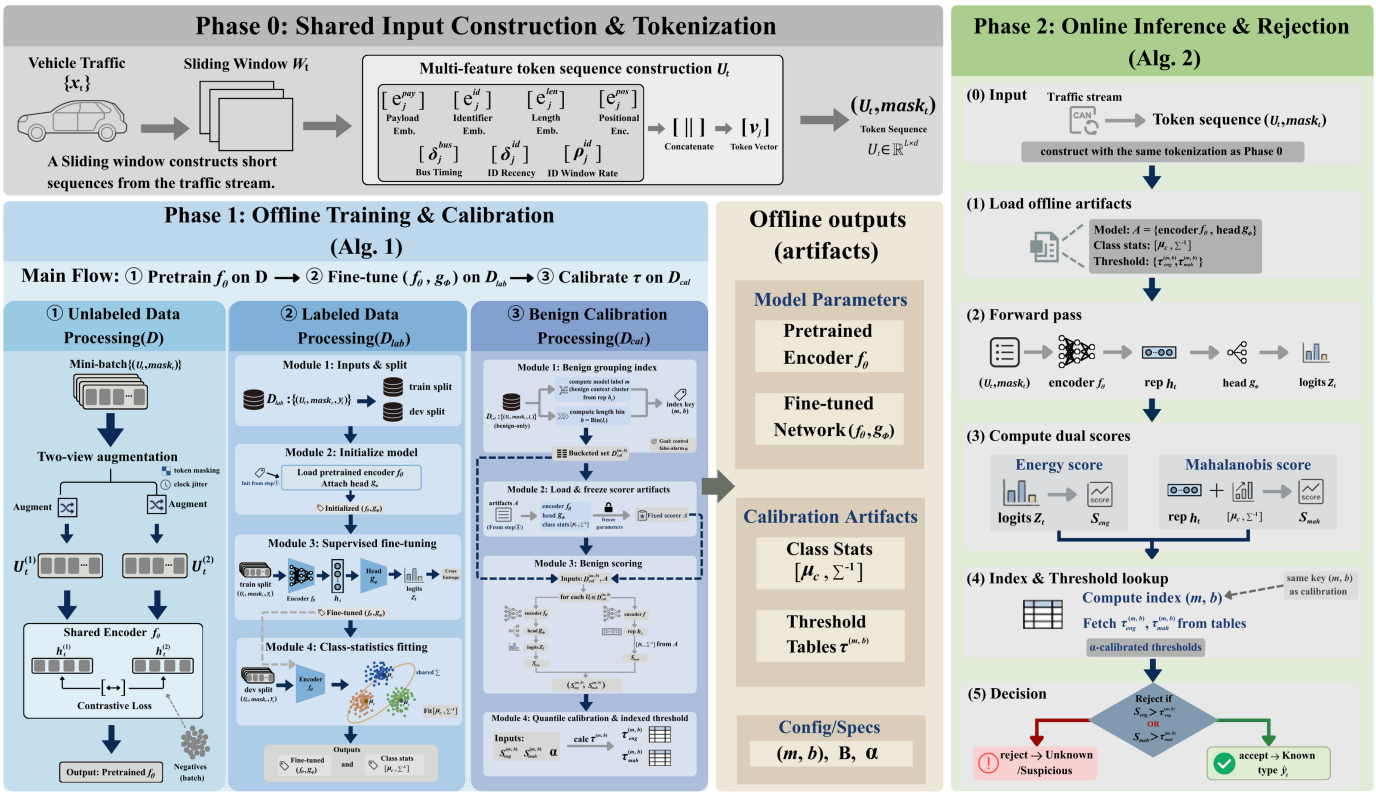

Fig. 2. CALIBER overview: tokenization, offline training with benign calibration, and online dual-score rejection with indexed thresholds.

For timing, the bus-level inter-arrival gap is

$$\Delta t^{\text{bus}}_{t-L+j} = t_{t-L+j} - t_{t-L+j-1}, \tag{3}$$

and the corresponding channel uses log scaling and normalization

$$\delta^{\text{bus}}_j = \text{NormTime}\big[\log\big[1 + \Delta t^{\text{bus}}_{t-L+j}\big]\big]. \tag{4}$$

For ID-specific recency, let $t^{\text{prev}}[\text{id}_{t-L+j}]$ denote the timestamp of the most recent earlier frame with the same identifier. If the identifier has not appeared before, we use the elapsed time since stream start. The recency gap and channel are

$$\Delta t^{\text{id}}_{t-L+j} = t_{t-L+j} - t^{\text{prev}}[\text{id}_{t-L+j}], \tag{5}$$

$$\delta^{\text{id}}_j = \text{NormTime}\big[\log\big[1 + \Delta t^{\text{id}}_{t-L+j}\big]\big]. \tag{6}$$

To capture short-horizon burstiness, we count occurrences of the current identifier within the window and apply log scaling

$$\rho^{\text{id}}_j = \log\Big[1 + \text{Count}\{\text{id}_{t-L+j} \in W_t\}\Big], \tag{7}$$

then normalize it by a fixed constant to obtain a bounded numeric channel.

Finally, the token vector $v_j \in \mathbb{R}^d$ is formed by concatenation

$$v_j = \text{Concat}\big[e^{\text{pay}}_j,\ e^{\text{id}}_j,\ e^{\text{len}}_j,\ e^{\text{pos}}_j,\ \delta^{\text{bus}}_j,\ \delta^{\text{id}}_j,\ \rho^{\text{id}}_j\big], \tag{8}$$

and stacking the L tokens in temporal order yields the shared input

$$U_t = \text{Stack}\big[\{v_j\}_{j=1}^{L}\big] \in \mathbb{R}^{L\times d}. \tag{9}$$

The model input is the pair $\langle U_t, \text{mask}_t\rangle$.

B. Phase 1: Offline Training and Benign Calibration

As shown in Fig. 2, Phase 1 learns a closed-set scorer under scarce labels and calibrates benign-only rejection thresholds. Three Phase 0 token datasets are used: an unlabeled set $D = \{\langle U_t, \text{mask}_t\rangle\}$ for representation pretraining, a labeled set $D_{\text{lab}} = \{\langle U_t, \text{mask}_t, y_t\rangle\}$ for supervised adaptation and class-statistics fitting, and a benign calibration set $D_{\text{cal}} = \{\langle U_t, \text{mask}_t, \ell_t\rangle\}$ for threshold calibration to a target false-alarm rate α. Here mask_t marks padded positions, and ℓ_t is a per-window payload-length scalar for length-binned calibration. If unavailable, we set $\ell_t = 0$ and use a single bin. Algorithm 1 summarizes Phase 1 and matches the three steps in Fig. 2. The quantities T_{ssl} and T_{sup} denote self-supervised and supervised update steps, and η is the learning rate.

Step 1: Unlabeled data processing on D. The encoder f_θ is pretrained to capture stable structure in CAN token sequences without attack labels. Lines 3–8 implement an in-batch contrastive objective. For each mini-batch, two augmented views are generated per sequence and optimized to pull together views from the same sequence while pushing apart others. The augmentation operator $\text{Aug}[\cdot]$ applies two protocol-agnostic perturbations aligned with Phase 0 token construction. Payload-channel masking randomly drops a small subset of payload-derived channels in U_t to discourage memorizing raw byte patterns and to promote reliance on structural cues. Timing jitter perturbs timing-related channels

to improve robustness to clock noise and scheduling variance. The padding mask mask_t is preserved so padded positions remain ignored.

Step 2: Labeled data processing on D_{lab}. The pretrained encoder is adapted to the known classes and class statistics are estimated for distance-based scoring. Line 9 splits D_{lab} into D_{tr} and D_{dev}. Lines 10–14 fine-tune the encoder and classifier with cross-entropy on D_{tr}, producing logits over C known classes. Mahalanobis statistics are then fit on D_{dev} to improve stability under scarce labels. Line 15 extracts representations $h_t \in \mathbb{R}^{d_h}$ for each $\langle U_t, y_t \rangle$ in D_{dev}. Lines 16–18 estimate per-class means $\{\mu_c\}$ and a shared covariance Σ with a small ridge term $\lambda > 0$ to avoid ill-conditioning,

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} h_i, \tag{10}$$

$$\Sigma = \frac{1}{N} \sum_i [h_i - \mu_{y_i}] [h_i - \mu_{y_i}]^\top . \tag{11}$$

A shared covariance keeps distances comparable across classes and avoids unstable per-class estimates. Line 19 packages the scorer artifacts A by saving the fine-tuned model and fitted statistics for reuse in Phase 2.

Step 3: Benign calibration on D_{cal}. The fixed scorer is converted into lookup-indexed thresholds that control the benign false-alarm rate and account for benign heterogeneity. Line 20 clusters benign representations into M benign modes, producing centers $\{c_m\}$. Lines 21–27 score each benign sequence and assign it to an index key. For each $\langle U_t, \mathrm{mask}_t, \ell_t \rangle$ in D_{cal}, we compute a representation h_t and logits z_t, then compute an energy score S_{eng} and a Mahalanobis score S_{mah} conditioned on the predicted class $\hat{y}_t$. A mode key $m_t \in \{1, \ldots, M\}$ is assigned by nearest-center lookup over $\{c_m\}$, and a length bin b_t is assigned by discretizing ℓ_t using fixed bin edges B. Benign score pairs are bucketed by $\langle m_t, b_t \rangle$, and per-bucket thresholds are set as empirical $1 - \alpha$ quantiles. We use a convention where larger scores indicate stronger evidence of being unknown, and rejection triggers when either score exceeds its indexed threshold. When a bucket has insufficient samples, a global fallback reservoir is used. Together with A, $\{c_m\}$, B, and α, these threshold tables form the offline artifacts queried in Phase 2.

C. Phase 2: Online Inference and Rejection

Phase 2 performs online detection for each incoming token sequence using the frozen artifacts produced in Phase 1. Algorithm 2 summarizes the per-sample inference and rejection procedure. Phase 0 tokenization is reused so online inputs remain aligned with the offline-fitted statistics and calibrated thresholds.

For each $\langle U_t, \mathrm{mask}_t \rangle$, the scorer produces a pooled representation h_t, logits z_t, and a closed-set prediction $\hat{y}_t$. Then, two complementary rejection scores are computed: an energy score S_{eng} from logits and a Mahalanobis score S_{mah} measuring the representation distance to the predicted-class center under the shared precision matrix. To keep the operating point stable, thresholds are retrieved by indexed lookup. A mode key m_t is obtained by assigning h_t to its closest benign prototype, and a length key b_t is obtained by discretizing the payload length using bin edges B. The index $\langle m_t, b_t \rangle$ selects the corresponding thresholds. A sample is rejected as unknown if either score exceeds its indexed threshold, otherwise it is

Algorithm 1 Offline Training And Benign Calibration

1: **function** PHASEONE($D, D_{\mathrm{lab}}, D_{\mathrm{cal}}, \alpha, M, \mathrm{B}$)
2: Initialize θ, ϕ
3: **for** $i = 1$ to T_{ssl} **do**
4: Sample mini-batch $\{(U_t, \mathrm{mask}_t)\} \subset D$
5: $(U_t^{(1)}, U_t^{(2)}) \leftarrow \mathrm{Aug}(U_t)$
6: $h_t^{(1)} \leftarrow f_\theta(U_t^{(1)}, \mathrm{mask}_t)$; $h_t^{(2)} \leftarrow f_\theta(U_t^{(2)}, \mathrm{mask}_t)$
7: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{con}}(h_t^{(1)}, h_t^{(2)})$
8: **end for**
9: Split D_{lab} into D_{tr} and D_{dev}
10: **for** $i = 1$ to T_{sup} **do**
11: Sample $(U_t, \mathrm{mask}_t, y_t) \subset D_{\mathrm{tr}}$
12: $z_t \leftarrow g_\phi(f_\theta(U_t, \mathrm{mask}_t))$
13: $(\theta, \phi) \leftarrow (\theta, \phi) - \eta \nabla \mathcal{L}_{\mathrm{ce}}(z_t, y_t)$
14: **end for**
15: $h_t \leftarrow f_\theta(U_t, \mathrm{mask}_t)$ for all $(U_t, \mathrm{mask}_t, y_t) \in D_{\mathrm{dev}}$
16: $\mu_c \leftarrow \mathrm{Mean}\{h_t \mid y_t = c\}$ for each class c
17: $\Sigma \leftarrow \mathrm{Cov}_{\mathrm{shared}}(\{(h_t, y_t)\})$
18: $\Sigma^{-1} \leftarrow \mathrm{Inv}(\Sigma)$
19: $A \leftarrow \{f_\theta, g_\phi, \{\mu_c\}, \Sigma^{-1}\}$
20: Fit $\{c_m\}_{m=1}^M$ by k-means on $\{f_\theta(U_t, \mathrm{mask}_t) \mid (U_t, \mathrm{mask}_t, \ell_t) \in D_{\mathrm{cal}}\}$
21: **for all** $(U_t, \mathrm{mask}_t, \ell_t) \in D_{\mathrm{cal}}$ **do**
22: $h_t \leftarrow f_\theta(U_t, \mathrm{mask}_t)$; $z_t \leftarrow g_\phi(h_t)$; $\hat{y}_t \leftarrow \arg\max z_t$
23: $S_{\mathrm{eng}} \leftarrow -\log \sum_{c=1}^C \exp(z_{t,c})$
24: $S_{\mathrm{mah}} \leftarrow (h_t - \mu_{\hat{y}_t})^\top \Sigma^{-1} (h_t - \mu_{\hat{y}_t})$
25: $m_t \leftarrow \arg\min_m \|h_t - c_m\|$; $b_t \leftarrow \mathrm{Bin}(\ell_t; \mathrm{B})$
26: Append $(S_{\mathrm{eng}}, S_{\mathrm{mah}})$ to bucket $\langle m_t, b_t \rangle$
27: **end for**
28: **for all** buckets $\langle m, b \rangle$ **do**
29: $\tau_{\mathrm{eng}}^{\langle m,b \rangle} \leftarrow \mathrm{Quantile}_{1-\alpha}(\{S_{\mathrm{eng}}\}^{\langle m,b \rangle})$
30: $\tau_{\mathrm{mah}}^{\langle m,b \rangle} \leftarrow \mathrm{Quantile}_{1-\alpha}(\{S_{\mathrm{mah}}\}^{\langle m,b \rangle})$
31: **end for**
32: **return** $A, \{c_m\}_{m=1}^M, \{\tau_{\mathrm{eng}}^{\langle m,b \rangle}\}, \{\tau_{\mathrm{mah}}^{\langle m,b \rangle}\}, M, \mathrm{B}, \alpha$

Algorithm 2 Online Inference And Unknown-attack Rejection

function DETECT($U_t, \mathrm{mask}_t, f_\theta, g_\phi, \{\mu_c\}, \Sigma^{-1}, \tau, \mathcal{C}, \mathrm{B}$)
$h_t \leftarrow f_\theta(U_t, \mathrm{mask}_t)$, $z_t \leftarrow g_\phi(h_t)$
$\hat{y}_t \leftarrow \arg\max_{c \in \{1,\ldots,C\}} z_{t,c}$
$S_{\mathrm{eng}} \leftarrow -\log \sum_{c=1}^C \exp(z_{t,c})$
$S_{\mathrm{mah}} \leftarrow (h_t - \mu_{\hat{y}_t})^\top \Sigma^{-1} (h_t - \mu_{\hat{y}_t})$
$m_t \leftarrow \arg\min_{c_m \in \mathcal{C}} \|h_t - c_m\|_2$ // $\mathcal{C} = \{c_m\}_{m=1}^M$
$\ell_t \leftarrow \mathrm{Len}(U_t)$, $b_t \leftarrow \mathrm{Bin}(\ell_t; \mathrm{B})$
$\tau_e \leftarrow \tau_{\mathrm{eng}}^{\langle m_t, b_t \rangle}$, $\tau_m \leftarrow \tau_{\mathrm{mah}}^{\langle m_t, b_t \rangle}$
if $S_{\mathrm{eng}} > \tau_e$ **or** $S_{\mathrm{mah}} > \tau_m$ **then return** REJECT
else return $\hat{y}_t$

TABLE I
CAR-HACKING DATASET

Class	Normal messages	Injected messages
Normal	988872	0
DoS	3,078,250	587,521
Fuzzy	3,347,013	491,847
Spoofing(Gear)	3,845,890	597,252
Spoofing(RPM)	3,966,805	654,897

TABLE II
CAR HACKING ATTACK AND DEFENSE CHALLENGE SUB-DATASET

Attack type	Stationary messages	Driving messages
Normal	1,467,427	1,545,284
Flooding	171,927	173,932
Fuzzy	96,328	90,244
Spoofing	23,971	26,368
Replay	48,830	61,644

accepted with label $\hat{y}_t$.

V. EXPERIMENTS

This section introduces the datasets and the evaluation metrics used in our label-scarce, unknown-attack-aware setting. Then we report the main comparative results, followed by label-budget studies and ablations of key design choices to evaluate our design.

A. Datasets

We evaluate CALIBER on two public in-vehicle intrusion datasets released by the Hacking and Countermeasure Research Lab: the 2018 *Car-Hacking Dataset* [1] and the 2021 *Car Hacking Attack and Defense Challenge dataset* [17]. Both provide time-ordered CAN traffic with timestamps, arbitration IDs, payload bytes, and intrusion labels. The *Car-Hacking Dataset* contains one normal trace and four attack traces, covering DoS, fuzzy injection, and two targeted spoofing attacks on Gear- and RPM-related messages, enabling leave-one-attack-type-out evaluation. The *Attack & Defense dataset* includes multiple attack types and two driving conditions, stationary and driving. Due to class imbalance, we use a subset of this dataset for a more balanced evaluation set.

For each dataset, we use time-ordered train, dev, and test splits with a 6:2:2 ratio to avoid temporal leakage. For attack traces, splits are anchored around the end of the injection period with a short post-attack context, so each split contains benign background traffic and injected messages. We construct an unlabeled set D by dropping labels for representation learning, a labeled set D_{lab} under a label budget for supervised adaptation, and a benign-only calibration set D_{cal} for threshold calibration at a target false-alarm rate α. Under label scarcity, D_{lab} uniformly subsamples labeled training samples, using 5% in our setup. Unknown-attack evaluation follows leave-one-attack-type-out, where the held-out attack is treated as unknown.

TABLE III
CLOSED-SET MULTICLASS RESULTS ON THE KNOWN TEST SET.

Dataset	Attack Type	Precision	Recall	Macro-F1	Latency (μs)
Car-Hacking [1]	Normal	0.9999	0.9999	0.9999	77.20
	DoS	0.9999	0.9999	0.9999	77.20
	Fuzzy	0.9992	0.9979	0.9985	77.18
	Spoofing(Gear)	0.9954	0.9983	0.9968	77.27
	Spoofing(RPM)	0.9967	0.9973	0.9971	77.29
Attack & Defense [17]	Normal	0.9999	0.9998	0.9998	78.55
	Flooding	0.9997	0.9996	0.9996	78.54
	Fuzzy	0.9994	0.9982	0.9988	78.66
	Spoofing	0.9985	0.9991	0.9988	78.71
	Replay	0.9995	0.9996	0.9995	78.59

TABLE IV
LOAO RESULTS FOR UNKNOWN-ATTACK DETECTION ON EACH DATASET.

Attack Type	AUROC	AUPR	Unknown Rej.	Benign FAR	Known Rej.	Latency (μs)
Car-Hacking [1]						
DoS	0.9748	0.9706	0.9764	0.018	0.022	78.24
Fuzzy	0.9477	0.9494	0.9473	0.018	0.015	80.36
Spoofing(Gear)	0.9703	0.9655	0.9724	0.018	0.020	79.02
Spoofing(RPM)	0.9675	0.9781	0.9773	0.018	0.024	78.88
Attack & Defense [17]						
Flooding	0.9764	0.9722	0.9773	0.020	0.022	80.02
Fuzzy	0.9575	0.9555	0.9503	0.020	0.022	79.90
Spoofing	0.9604	0.9619	0.9634	0.020	0.021	78.24
Replay	0.9773	0.9502	0.9691	0.020	0.023	81.68

B. Evaluation Metrics

CALIBER outputs a closed-set prediction over known classes and a reject decision. For known-class evaluation, we report Accuracy and macro F1-score, together with per-class precision, recall, and F1-score. We report closed-set performance without rejection and performance after applying the accept or reject rule to reflect deployed behavior.

For unknown-attack awareness, the held-out attack is treated as the positive class and all known-class test traffic as the negative class, where known-class traffic includes benign samples and known attack types. We compute AUROC and AUPR using a continuous rejection score defined as the maximum of the normalized energy and Mahalanobis scores. We also report the achieved benign false-alarm rate, the reject rate on known-class test traffic under the same operating point, and the reject rate on unknown attacks only. Finally, we report per-sample inference latency measured at batch size 1 after warm-up, excluding Phase 0 tokenization.

C. Closed-set Recognition and Unknown-attack Rejection

We evaluate CALIBER under two complementary settings. For closed-set recognition, we perform multiclass classification on the known test set, where benign traffic is treated as a normal class together with all known attack types. We compute per-class precision, recall, and F1-score and summarize results as class-wise averages across splits under a fixed taxonomy. Table III reports the closed-set results and inference latency.

Closed-set recognition is strong and balanced across benign traffic and known attacks, with near-saturated class-wise F1-scores under limited supervision and stable latency under the same runtime setting. For unknown-attack awareness, we adopt a leave-one-attack-out protocol, holding out one

TABLE V
COMPARISON WITH REPRESENTATIVE LITERATURE MODELS UNDER THE SAME EVALUATION PROTOCOL.

Method	Known Macro F1	AUROC	Unknown Rej.	Latency(μs)	Params(M)
LW-CNN [5]	0.9983	0.9527	0.9542	54.88	0.014
CNN-LSTM [18]	0.9987	0.9567	0.9500	1168	0.122
Transformer [19]	0.9989	0.9689	0.9733	2246	0.259
Ours	0.9984	0.9651	0.9684	77.23	0.071

attack type as unknown while training on benign traffic and the remaining attacks. Unknown detection treats unknown attacks as positives and all-known test traffic as negatives, where all-known includes benign traffic and known attacks. We report AUROC and AUPR from a continuous rejection score consistent with the final decision rule, along with the unknown reject rate at the calibrated operating point. Table IV summarizes results across folds. Under the calibrated operating point, benign false-alarm rates remain controlled and unknown reject rates stay high, while known reject rates remain limited, indicating that rejection preserves closed-set recognition while improving robustness to unseen attacks.

D. Comparison with Representative Literature Models

In addition to the main results, we provide a compact comparison against representative literature baselines to assess the accuracy–efficiency trade-off under the same open-world protocol. Table V reports dataset-level metrics averaged over all leave-one-attack-out folds.

All methods use the same data splits and evaluation protocol, while each baseline follows standard settings from prior work on the same dev split. The baselines cover common sequence encoders, including lightweight convolutional models, hybrid convolution–recurrent architectures, and attention-based encoders for longer-range dependencies. Closed-set recognition on known traffic is near-saturated across methods, with Known Macro F1 values concentrated in a narrow range. Unknown-attack awareness varies modestly, where higher-capacity encoders can yield slightly stronger AUROC and unknown reject rates under the same false-alarm budget. These gains come with substantially higher per-sample latency due to increased computational complexity. Overall, the results illustrate the accuracy–efficiency trade-off, and our design achieves competitive unknown-attack rejection with markedly lower inference time.

E. Operating-point Calibration Sweep

To characterize the operating characteristic of calibrated rejection, we sweep the benign calibration target α and evaluate behavior on the test distribution. We report achieved benign false-alarm rate because it reflects the deployment budget, and mismatch between α and the achieved rate can arise from finite-sample quantile estimation, benign distribution shift, and mode-dependent heterogeneity in benign dynamics. At each operating point, the unknown reject rate is computed on unknown attacks only under the same decision rule as in the main evaluation, capturing the trade-off between sensitivity to unseen attacks and tolerance to false alarms.

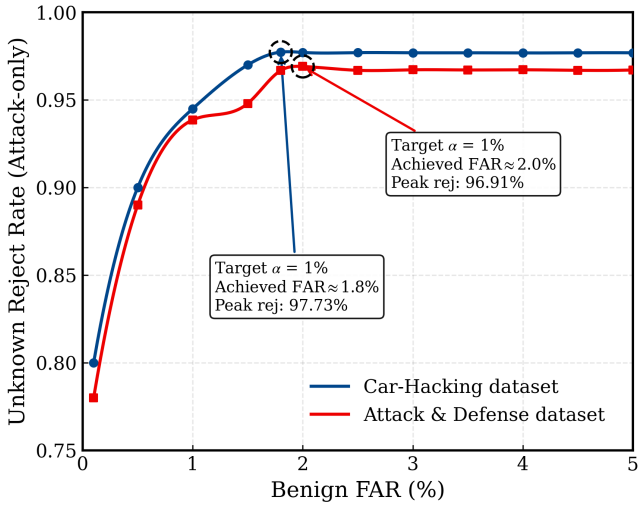

Fig. 3. Unknown reject rate versus achieved benign FAR obtained by sweeping the calibration target α, with annotations highlighting $\alpha = 1\%$.

We fix one leave-one-attack-out setting per dataset to obtain a concrete curve, holding out Spoofing RPM and Replay, respectively. Fig. 3 shows a clear knee region: unknown rejection increases rapidly in the low-FAR regime and then saturates as the budget loosens. At $\alpha = 1\%$, the achieved benign false-alarm rate is 1.8% and 2.0% on the two datasets, while the corresponding unknown reject rates are 97.73% and 96.91%. Further relaxing the budget yields diminishing returns, indicating that most unseen attacks are already filtered and residual errors are dominated by harder, confidence-preserving unknown patterns. This sweep motivates selecting α by targeting the achieved FAR on the deployment distribution rather than treating α as the achieved operating point.

F. Stability Across Random Seeds

We assess robustness to stochasticity by repeating the full Phase 1–Phase 2 pipeline under multiple random seeds while keeping data splits, hyperparameters, and the operating point fixed. To decouple seed variance from fold variance, we use a fixed unknown-attack setting per dataset.

Table VI shows that closed-set recognition is essentially invariant to the seed choice. With fixed partitions and supervision, the representation geometry and decision boundaries over the known taxonomy are strongly constrained, so training converges to similar optima across runs. Rejection exhibits slightly larger yet well-bounded variation because it depends on benign-tail quantile thresholding and indexed calibration artifacts such as benign-mode prototypes and per-bucket score summaries. These intermediate estimates are more exposed to finite-sample noise and stochastic optimization and mini-batch ordering, so small shifts in score distributions near the operating point can translate into modest deviations in the achieved benign false-alarm rate and the resulting rejection behavior. Overall, the pipeline remains stable under stochastic training, and randomness primarily affects second-order calibration

TABLE VI
PERFORMANCE VARIATION ACROSS RANDOM SEEDS UNDER FIXED SPLITS AND A FIXED UNKNOWN-ATTACK SETTING FOR EACH DATASET.

Seed	Acc	Macro-F1	AUROC	AUPR	Unknown Rej.	Benign FAR	Known Rej.
Car-Hacking [1]							
10	0.9986	0.9970	0.9580	0.9690	0.9650	0.0205	0.0280
20	0.9988	0.9973	0.9625	0.9740	0.9720	0.0190	0.0265
30	0.9987	0.9971	0.9608	0.9725	0.9705	0.0198	0.0272
40	0.9989	0.9975	0.9675	0.9781	0.9773	0.0180	0.0240
50	0.9988	0.9972	0.9640	0.9755	0.9735	0.0188	0.0258
Mean±Std	**0.99876±0.00011**	**0.99722±0.00019**	**0.96256±0.00355**	**0.97382±0.00340**	**0.97166±0.00450**	**0.01922±0.00096**	**0.02630±0.00152**
Attack & Defense [17]							
10	0.9988	0.9979	0.9685	0.9390	0.9570	0.0220	0.0260
20	0.9990	0.9981	0.9720	0.9445	0.9625	0.0210	0.0250
30	0.9989	0.9980	0.9708	0.9420	0.9600	0.0215	0.0258
40	0.9991	0.9983	0.9773	0.9502	0.9691	0.0200	0.0230
50	0.9990	0.9981	0.9735	0.9460	0.9640	0.0208	0.0242
Mean±Std	**0.99896±0.00011**	**0.99808±0.00015**	**0.97242±0.00328**	**0.94434±0.00422**	**0.96252±0.00454**	**0.02106±0.00075**	**0.02480±0.00123**

TABLE VII
LABEL-BUDGET SCALING RESULTS ON CAR-HACKING AND ATTACK & DEFENSE DATASETS.

Dataset	Label (%)	Macro-F1	AUROC	Unknown Rej.	Benign FAR
Car-Hacking [1]	1	0.8350	0.6200	0.5500	0.0300
	3	0.9300	0.8200	0.8500	0.0230
	5	0.9975	0.9675	0.9773	0.0180
	7	0.9975	0.9670	0.9769	0.0178
	10	0.9968	0.9666	0.9766	0.0179
	15	0.9967	0.9663	0.9765	0.0181
	20	0.9966	0.9661	0.9764	0.0180
Attack & Defense [17]	1	0.8600	0.6500	0.5000	0.0280
	3	0.9450	0.8450	0.8300	0.0220
	5	0.9983	0.9773	0.9691	0.0150
	7	0.9981	0.9769	0.9688	0.0148
	10	0.9984	0.9766	0.9686	0.0149
	15	0.9975	0.9763	0.9685	0.0151
	20	0.9974	0.9761	0.9684	0.0150

granularity rather than altering the underlying separability between all-known traffic and the held-out unknown attack.

G. Label-budget Scaling

We vary the labeled fraction in Phase 1 while keeping splits, architecture, and benign calibration fixed. Table VII summarizes performance on known traffic and unknown rejection under the LOAO setting.

As the label budget increases, the model transitions from limited class supervision to a well-structured representation space. With very few labeled samples, class prototypes and decision boundaries are weakly constrained, which reduces separability between unseen attacks and known behaviors and makes rejection scores less reliable. Adding labels quickly sharpens class geometry and improves alignment between closed-set training and benign-only calibration, making both the scorer and calibrated thresholds more consistent across folds and attack holds. Gains diminish once the representation becomes sufficiently anchored by labeled data, and performance stabilizes around the 5% budget. This supports using 5% labels as the default operating point, which achieves near-saturated behavior while maintaining label efficiency under realistic annotation constraints.

H. Ablation Experiments

In this section, we conduct ablation studies to quantify the contributions of key components in CALIBER. We first ablate groups of token views, and then ablate the rejection scoring components used for unknown-attack detection.

1) Group-wise Ablation of Token Views: We perform a group-wise ablation to quantify the contribution of each token-view group to closed-set recognition and unknown rejection. Each variant removes one view group from Phase 0 while keeping Phase 1 training, benign calibration, and Phase 2 rejection unchanged. Under LOAO, we report fold-average and worst-fold results, where the worst fold corresponds to the most challenging held-out attack.

Table VIII indicates that known-class recognition is supported by multiple redundant cues, whereas unknown rejection depends more strongly on whether the token retains the signals that define normality in CAN traffic. Removing payload content weakens instance-level evidence, making it harder to detect semantic perturbations that still look structurally plausible. Removing identity and structure cues breaks ID-conditioned regularities and short-horizon ordering, which reduces separability among known behaviors and increases confusion between valid and abnormal sequences. Removing temporal and rate cues mostly preserves static class separation, but discards rhythmic and burstiness patterns that distinguish unseen manipulations from benign dynamics. Overall, the results indicate that content, ID-structured context, and timing cues play complementary roles in supporting recognition and rejection.

2) Ablation on Rejection Scores for Unknown Detection: We ablate the rejection scoring components while keeping tokenization, Phase 1 training, and benign calibration unchanged. We compare energy-only, Mahalanobis-only, and the dual-score rule. All variants are calibrated on benign traffic under

TABLE VIII
GROUP-WISE ABLATION OF TOKEN VIEWS UNDER LOAO.

Missing Features	Macro-F1	AUROC	AUPR	Unknown Rej.	Benign FAR	Known Rej.
Car-Hacking [1]						
Seven Features	0.9981 / 0.9968	0.9650 / 0.9477	0.9659 / 0.9494	0.9684 / 0.9473	0.0180	0.020
w/o Content (e^{pay})	0.9858 / 0.9806	0.9492 / 0.9208	0.9445 / 0.9051	0.9550 / 0.9023	0.0187	0.026
w/o Identity&Structure (e^{id}, e^{len}, e^{pos})	0.9721 / 0.9634	0.8920 / 0.8127	0.8618 / 0.7425	0.9036 / 0.7812	0.0195	0.041
w/o Temporal&Rate ($\delta^{bus}, \delta^{id}, \rho^{id}$)	0.9946 / 0.9928	0.8425 / 0.7741	0.8012 / 0.6904	0.8610 / 0.7485	0.0183	0.024
Attack & Defense [17]						
Seven Features	0.9992 / 0.9985	0.9679 / 0.9575	0.9599 / 0.9502	0.9650 / 0.9503	0.0200	0.022
w/o Content (e^{pay})	0.9896 / 0.9848	0.9625 / 0.9340	0.9190 / 0.8120	0.9518 / 0.8955	0.0206	0.030
w/o Identity&Structure (e^{id}, e^{len}, e^{pos})	0.9834 / 0.9752	0.9128 / 0.8460	0.8355 / 0.6982	0.9026 / 0.8034	0.0214	0.045
w/o Temporal&Rate ($\delta^{bus}, \delta^{id}, \rho^{id}$)	0.9969 / 0.9956	0.8712 / 0.8015	0.7830 / 0.6406	0.8895 / 0.7768	0.0248	0.028

TABLE IX
REJECTION-SCORE ABLATION UNDER LOAO.

Rejection Score	AUROC	AUPR	Unknown Rej.	Benign FAR	Known Rej.
Car-Hacking [1]					
Energy only (S_{eng})	0.9182 / 0.8625	0.9050 / 0.8424	0.8915 / 0.8102	0.0180	0.014
Mahalanobis only (S_{mah})	0.9340 / 0.8850	0.9220 / 0.8580	0.9100 / 0.8350	0.0180	0.048
Dual-score (Ours)	0.9650 / 0.9477	0.9659 / 0.9494	0.9684 / 0.9473	0.0180	0.020
Attack & Defense [17]					
Energy only (S_{eng})	0.9280 / 0.8750	0.8935 / 0.8210	0.9014 / 0.8350	0.0200	0.016
Mahalanobis only (S_{mah})	0.9400 / 0.8920	0.9050 / 0.8380	0.9200 / 0.8480	0.0200	0.052
Dual-score (Ours)	0.9679 / 0.9575	0.9599 / 0.9502	0.9650 / 0.9503	0.0200	0.022

the same target α, and LOAO results are summarized by fold-average and worst-fold performance.

Table IX supports that the two scores address different mismatch patterns between known and unknown traffic. Energy-based rejection relies on the classifier confidence surface, which can remain sharp even for out-of-taxonomy inputs, so confidently misclassified unknown samples are less penalized. Mahalanobis-based rejection measures deviation in representation space and can better capture distributional shift, but it may over-trigger when class clusters overlap or when benign contexts differ from those represented in labeled data, increasing rejection of valid known samples. Combining the two cues provides a more balanced decision boundary, improving sensitivity to confident-looking unknowns while reducing unnecessary rejection on known traffic under the same operating point.

VI. CONCLUSION

We present CALIBER, a label-scarce open-world sequence learning framework for CAN traffic that couples closed-set classification with calibrated unknown rejection under an explicit false-alarm budget. Across two public datasets, CALIBER achieves near-saturated recognition on known traffic while retaining strong sensitivity to held-out unknown attacks in leave-one-attack-out evaluation. The proposed dual-score rejection combines logit-based energy and representation-space Mahalanobis distance, providing complementary uncertainty and deviation cues that improve unknown detection at a fixed operating point.

Despite these promising results, performance remains influenced by unknown distribution shift and the choice of the calibration target α, which can move the rejection boundary under changing benign contexts. Moreover, limited labeled coverage can bias representation geometry toward seen classes, reducing separation for certain unknown patterns. Looking ahead, we will study more robust calibration under non-stationary conditions, improved mode discovery for benign heterogeneity, and adaptive score fusion with lightweight online updates. Finally, we plan to explore alternative self-supervised objectives that better preserve anomaly-sensitive features while maintaining efficiency.

REFERENCES

[1] E. Seo, H. M. Song, and H. K. Kim, "Gids: Gan based intrusion detection system for in-vehicle network," in *2018 16th annual conference on privacy, security and trust (PST)*. IEEE, 2018, pp. 1–6.

[2] S. C. Kalkan and O. K. Sahingoz, "In-vehicle intrusion detection system on controller area network with machine learning models," in *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020, pp. 1–6.

[3] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," *Advances in neural information processing systems*, vol. 32, 2019.

[4] M. Almehdhar, A. Albaseer, M. A. Khan, M. Abdallah, H. Menouar, S. Al-Kuwari, and A. Al-Fuqaha, "Deep learning in the fast lane: A survey on advanced intrusion detection systems for intelligent vehicle networks," *IEEE Open Journal of Vehicular Technology*, vol. 5, pp. 869–906, 2024.

[5] S. Huan, X. Zhang, W. Shang, H. Cao, H. Li, Y. Yang, and W. Liu, "T-shaped can feature integration with lightweight deep learning model for in-vehicle network intrusion detection," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[6] T.-P. Nguyen, J. Cho, and D. Kim, "Semi-supervised intrusion detection system for in-vehicle networks based on variational autoencoder and adversarial reinforcement learning," *Knowledge-Based Systems*, vol. 304, p. 112563, 2024.

[7] X. Zhang, R. Zhao, Z. Jiang, Z. Sun, Y. Ding, E. C. Ngai, and S.-H. Yang, "Aoc-ids: Autonomous online framework with contrastive learning for intrusion detection," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 581–590.

[8] A. Binbusayyis and T. Vaiyapuri, "Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class svm," *Applied Intelligence*, vol. 51, no. 10, pp. 7094–7108, 2021.

[9] M. Nakıp and E. Gelenbe, "Online self-supervised deep learning for intrusion detection systems," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 5668–5683, 2024.

[10] L. Du, Z. Gu, Y. Wang, and C. Gao, "Open world intrusion detection: An open set recognition method for can bus in intelligent connected vehicles," *Ieee Network*, vol. 38, no. 3, pp. 76–82, 2024.

[11] L. Du, Y. Chai, Y. Jia, B. Fang, H. Li, and Z. Gu, "Toward open-world network intrusion detection via open recognition and inspection," *IEEE Transactions on Information Forensics and Security*, 2025.

[12] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[13] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[14] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21 464–21 475, 2020.

[15] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[16] R. I. Oliveira, P. Orenstein, T. Ramos, and J. V. Romano, "Split conformal prediction and non-exchangeable data," *Journal of Machine Learning Research*, vol. 25, no. 225, pp. 1–38, 2024.

[17] H. Kang, B. I. Kwak, Y. H. Lee, H. Lee, H. Lee, and H. K. Kim, "Car hacking and defense competition on in-vehicle network," in *Workshop on automotive and autonomous vehicle security (AutoSec)*, vol. 2021. NDSS San Diego, CA, 2021, p. 25.

[18] W. Lo, H. Alqahtani, K. Thakur, A. Almadhor, S. Chander, and G. Kumar, "A hybrid deep learning based intrusion detection system using spatial-temporal representation of in-vehicle network traffic," *Vehicular Communications*, vol. 35, p. 100471, 2022.

[19] T.-D. Le, H. B. H. Truong, D. Kim *et al.*, "Multi-classification in-vehicle intrusion detection system using packet-and sequence-level characteristics from time-embedded transformer with autoencoder," *Knowledge-Based Systems*, vol. 299, p. 112091, 2024.