BEYOND COMMON SCENES: LONG-TAILED ACCIDENT ANTICIPATION WITH SCENE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic accident anticipation aims to issue early warnings of impending collisions from dashcam video, providing time for preventive action. Yet balanced accident and non-accident samples can conceal substantial imbalance among driving situations, leaving models with limited examples of the early warning cues in rare scenes. We formulate this scene-level long-tail problem and characterize it through complementary analyses of scene frequency, reference-model error, and expert curation, distinguishing statistical rarity, anticipation difficulty, and safety relevance. To support underrepresented scenes with related experience, we propose TailBank, a persistent scene-memory framework. TailBank retains central and peripheral representatives of training scene groups across batches and retrieves relevant knowledge to enhance current scene features. A distance-based fitness mechanism assesses how well the bank covers the current scene, modulates the fused representation, and separately flags insufficient support. This keeps scene familiarity distinct from accident likelihood. Our theoretical analysis establishes sufficient conditions for enhancement to reduce anticipation loss and for known and open scenes to be separated. Experiments on DAD, CCD, and A3D show state-of-the-art accuracy on DAD and A3D and competitive warning timeliness. Further evaluations support TailBank's effectiveness in rare and challenging scenes, generalization to unseen scenes, and identification of difficult cases.

1 INTRODUCTION

Traffic accident anticipation aims to issue early warnings of impending collisions from dashcam video, providing additional time for drivers or automated driving systems to respond. To provide this lead time, a model must assess accident risk from the video observed so far, when the evidence of an emerging hazard may still be incomplete or ambiguous. Effective anticipation therefore requires recognizing warning cues early while limiting false alarms in non-accident situations. Recent advances in deep learning have strengthened this capability through recurrent networks, graph-based models, and attention mechanisms that capture traffic participants and their temporal interactions (Chan et al., 2017; Bao et al., 2020; Karim et al., 2022b). More recent studies explore multimodal large language models for scene-aware accident anticipation (Liao et al., 2025; 2024b), while diffusion-based accident video synthesis has been used to address data bias (Fang et al., 2025).

Despite these advances, strong overall performance can conceal missed or delayed warnings in rarely observed driving situations. Real-world traffic data contain frequently observed configurations of roads, participants, and interactions, alongside a long tail of sparsely represented combinations of these factors. *Balancing accident and non-accident videos does not balance the situations in which an early warning is needed.* A model may encounter many accident videos yet have little experience with the particular interaction unfolding in the current scene. Addressing this scene-level imbalance requires understanding which situations are underrepresented, how knowledge from other scenes can help, and where that knowledge ceases to provide adequate support. We explore these challenges through three interrelated questions (**Q1-Q3**) and attempt to answer them for the first time (**A1-A3**).

Q1: How should we characterize the long tail in accident anticipation? The binary accident label describes an outcome; it does not reveal how frequently the preceding scene configuration occurs. Moreover, scene rarity, anticipation difficulty, and safety relevance capture different properties: a rare scene need not be difficult, and a large anticipation error does not establish rarity. A useful

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.

characterization must therefore go beyond outcome counts or a single model's errors. Inspired by multi-criteria analyses of long-tailed driving behavior (Zhang et al., 2025; Rao et al., 2025), we use complementary views: visual micro-cluster frequency to examine scene rarity, reference-model error to expose anticipation difficulties, and expert curation to identify safety-relevant configurations. Together, these views reveal scene-specific weaknesses that an aggregate score can obscure.

Q2: How can we anticipate accidents in scenes with few training examples? An unusual interaction may offer only subtle warning cues before a collision (Fang et al., 2024). When a driving situation appears only a few times during training, the model has limited opportunities to learn the warning signs that precede a collision. However, different situations may share useful cues, such as a vehicle moving into the ego vehicle's path. A model could therefore draw on related scenes to help interpret a rarely observed situation. This motivates retaining useful scene information during training and retrieving relevant information for the current video. The goal is to support an early warning when the current observations alone do not yet provide clear evidence of danger.

Q3: How can we distinguish rare known scenes from unseen scenes? A scene may be poorly represented because it appeared infrequently during training or no similar scene was observed. Both can appear unfamiliar to the model. Treating every poorly represented scene as unknown would reject known tail cases, whereas indiscriminate knowledge sharing could impose familiar patterns on unseen situations. This tension is central to open long-tailed recognition, but accident anticipation adds another requirement: scene familiarity must remain distinct from accident risk. A familiar scene can be dangerous, and an unfamiliar scene need not lead to a collision. Knowledge reuse should therefore account for how well stored experience covers the current traffic scene.

To address these questions, we introduce **TailBank**, a scene-knowledge framework for long-tailed accident anticipation. TailBank maintains a persistent bank of central and peripheral representatives for each training scene group, allowing context and object features to retrieve knowledge relevant to the current observation. A distance-based fitness coefficient modulates the fused representation supplied to a recurrent anticipation model and separately flags scenes with insufficient bank coverage as potentially unfamiliar. Our theoretical analysis relates scene coverage to the conditions under which the model can separate known and open scenes. It further establishes sufficient conditions for knowledge enhancement to reduce a fixed anticipation model's loss, accounting for both retrieved information and attenuation of the original feature. Overall, our contributions are threefold:

- We introduce the scene-level long-tail problem in accident anticipation and develop a multi-criteria framework to characterize it. Complementary analyses of scene frequency, reference-model error, and expert-curated scenes distinguish statistical rarity, anticipation difficulty, and safety relevance, enabling targeted evaluation beyond binary accident labels.
- We propose TailBank, which combines persistent scene-knowledge retrieval with a shared fitness mechanism for feature modulation and candidate open-scene detection, supported by analyses of scene coverage and conditional enhancement benefit.
- Evaluation results on DAD, CCD, and A3D demonstrate TailBank's state-of-the-art anticipation accuracy and warning timeliness. Further evaluations on underrepresented and challenging scenes highlight its effectiveness in long-tailed accident anticipation.

Relation to Prior Work. We leave a detailed review of the research directions and a discussion of the technical similarities and differences between TailBank and existing methods to **Appendix B**.

2 PRELIMINARIES

Accident Anticipation. The objective of accident anticipation is to issue early warnings of impending collisions from dashcam video, allowing time for drivers or automated driving systems to respond. Formally, given a video sequence $V = (x_1, \ldots, x_T)$, an anticipation model produces a frame-level probability sequence $\mathbf{p} = (p_1, \ldots, p_T)$, where $p_t = \Pr(Y = 1 \mid x_{1:t}) \in [0, 1]$ and $Y \in \{0, 1\}$ is the video-level accident label. A warning is issued when $p_t \geq \theta_{\text{acc}}$, with θ_{acc} denoting the warning threshold. For a positive video with annotated accident onset t^*, successful anticipation requires the first warning time $t_0 < t^*$, yielding a physical lead time of $(t^* - t_0)/f_{\text{fps}}$ at frame rate f_{fps}. The model should provide timely warnings while limiting false alarms on negative videos.

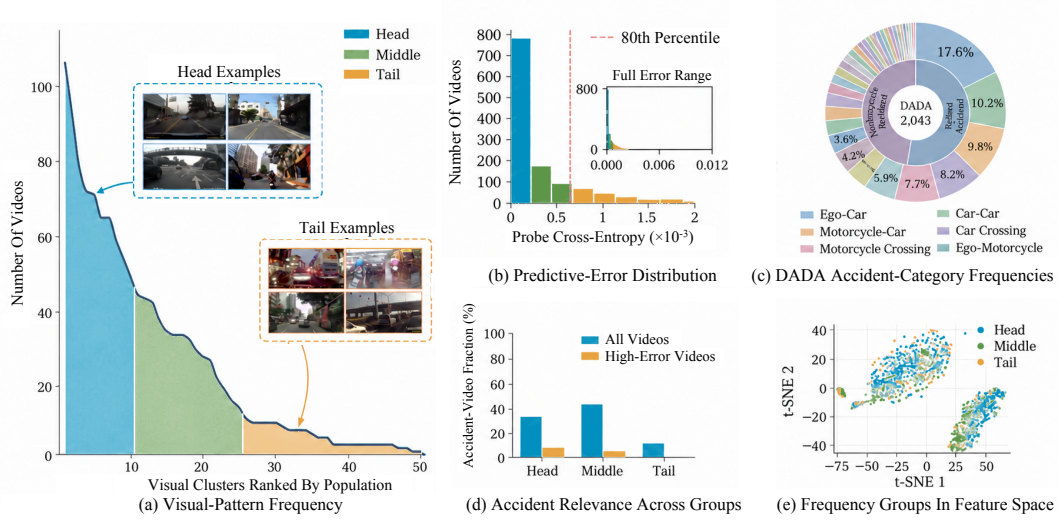


Figure 1: Multi-source characterization of long-tailed traffic accident scenarios. (a) *Micro-Cluster Tail*: visual clusters ranked by size and grouped into Head, Middle, and Tail, with representative head and tail scenes. (b) *Prediction-Error Tail*: cross-entropy distribution of the reference classifier, with the dashed line marking the 80th-percentile threshold; the inset shows the full error range. (c) *Expert-Curated Tail*: DADA accident-category frequencies provide context for selecting safety-critical scenarios. (d) Accident-video fractions across frequency groups, shown for all videos and high-error videos. (e) t-SNE projection of DAD scene features colored by frequency group.

Long-Tailed Accident Scenarios. The scene-level long tail concerns the frequencies of driving configurations rather than binary accident labels: a few configurations are common, while many others have few observations. To characterize scene rarity empirically, consider a reference video set $\mathcal{S}_{\text{ref}} = \{V_i\}_{i=1}^{n_{\text{ref}}}$, partitioned into J nonempty visual clusters $\{\mathcal{K}_j\}_{j=1}^{J}$. Their empirical frequencies, indexed in increasing order, can be formally defined as $\widehat{\pi}_j = \frac{|\mathcal{K}_j|}{n_{\text{ref}}}$, where $0 < \widehat{\pi}_1 \leq \cdots \leq \widehat{\pi}_J$. For a target video coverage $q \in (0, 1]$, we accumulate the smallest clusters until their combined frequency reaches q. The cutoff index and the corresponding frequency-based tail subset are as follows:

$$j_q = \min\left\{ j \in \{1, \ldots, J\} : \sum_{a=1}^{j} \widehat{\pi}_a \geq q \right\}, \quad \mathcal{S}_{\text{tail}}(q) = \bigcup_{j=1}^{j_q} \mathcal{K}_j. \tag{1}$$

Here, whole clusters are retained, so the selected videos cover at least a fraction q of the reference set. Rarity is relative to this set and its visual representation. The long-tailed setting seeks accurate and timely anticipation on $\mathcal{S}_{\text{tail}}(q)$ while maintaining full-set performance.

Notably, scene rarity need not coincide with anticipation difficulty (Figure 1). We therefore complement frequency-based subsets with high-error subsets selected by a fixed reference model and expert-curated safety-relevant subsets. All three evaluation families use the DAD test set (Chan et al., 2017), with identical selected videos across methods and full-set results reported alongside them. These evaluation clusters are distinct from the training scene pseudo-categories used by the knowledge bank. Together, frequency, reference-model error, and expert attributes provide an operational characterization of the scene-level long tail that distinguishes rarity from difficulty and safety relevance (**A1**). Low frequency in the evaluation set does not by itself establish limited training support or an unseen scene. Detailed selection and annotation protocols are provided in Appendix E.2.

3 Methodology

3.1 Framework Overview

As shown in Figure 2, the Context and Object Encoders extract features from each observed frame x_t. The Context-Aware and Object-Aware Modules refine these features, which are fused into the

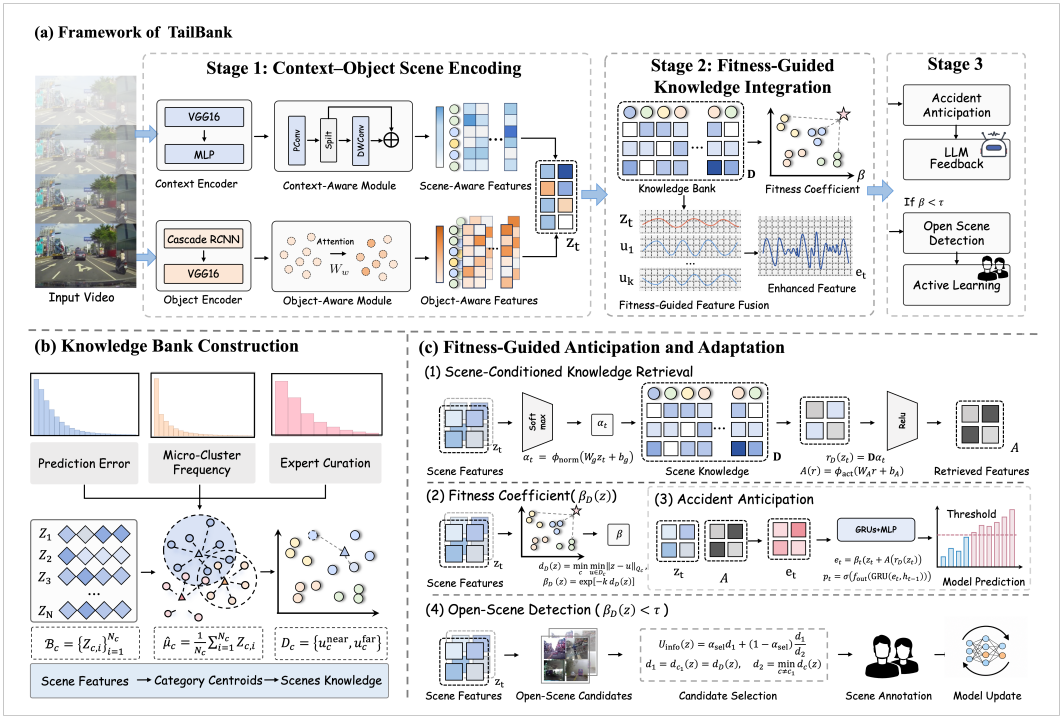


Figure 2: Overview of TailBank. (a) Context-object scene encoding and fitness-guided knowledge integration support accident anticipation, open-scene screening, and optional language feedback. (b) Knowledge-bank construction selects central and peripheral representatives of training scene groups. (c) Scene-conditioned retrieval and fitness-guided fusion produce features for recurrent accident anticipation, while candidate open scenes are selected for annotation and model updating.

scene embedding z_t. Then, the Head-to-Tail Knowledge Enhancement module retrieves $r_D(z_t)$ from the knowledge bank and combines its transformed output with z_t. Next, the fitness coefficient β_t scales the fused feature to produce e_t and also supports candidate open-scene detection. A GRU processes e_t directly, and a sigmoid readout produces the frame-level accident probability p_t.

3.2 Scene Embedding

For each frame x_t, we extract VGG-16 features from the full image and traffic participants detected by Cascade R-CNN (Cai & Vasconcelos, 2018). The Context-Aware Module (CAM) and Object-Aware Module (OAM) produce z_t^{ctx} and z_t^{obj}, respectively. Their outputs are concatenated and mapped by a learned projection to form the knowledge-bank query, which can be denoted as $z_t = \phi_{\text{fuse}}\left([z_t^{\text{ctx}}; z_t^{\text{obj}}]\right) \in \mathbb{R}^d$. We leave more details of this module in Appendix C.

3.3 Scene Knowledge Integration

Knowledge Bank Construction. To retain scene information across batches, we maintain a persistent bank of representative features. Training embeddings are grouped into M nonempty scene pseudo-categories. For category c, let $\mathcal{B}_c = \{Z_{c,i}\}_{i=1}^{N_c}$ be its construction features. We compute its centroid and select the nearest and farthest features under the chosen centroid-distance ordering:

$$\widehat{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} Z_{c,i}, \qquad D_c = \{u_c^{\text{near}}, u_c^{\text{far}}\}, \qquad D = \bigcup_{c=1}^{M} D_c. \tag{2}$$

where u_c^{near} and u_c^{far} are the features in $\mathcal{B}_c$ nearest to and farthest from $\widehat{\mu}_c$, respectively, retaining central and peripheral scene patterns. Construction and update details are provided in Appendix C.2.

Scene-Conditioned Knowledge Retrieval. Given the bank matrix $\mathbf{D} = [u_1, \ldots, u_K]$, where u_k is the k-th stored scene representative, the current embedding z_t queries the bank. Formally,

$$\alpha_t = \phi_{\text{norm}}(W_g z_t + b_g), \qquad r_D(z_t) = \mathbf{D}\alpha_t = \sum_{k=1}^{K} \alpha_{t,k} u_k, \tag{3}$$

where ϕ_{norm} denotes softmax normalization and α_t assigns relevance weights to the stored representatives. Thus, $r_D(z_t)$ provides a scene-adaptive summary of the bank.

The retrieved knowledge is then mapped to the feature space of z_t by following process:

$$A(r) = \phi_{\text{act}}(W_A r + b_A), \tag{4}$$

where ϕ_{act} denotes ReLU. This adapter A(r) produces a compatible residual feature, emphasizing scene-relevant knowledge while suppressing unrelated information.

Fitness-Guided Feature Fusion. We measure scene coverage using the distance to the nearest bank representative. Let $Q_c = \Sigma_c^{-1}$, with $\Sigma_c \succeq sI$, $s > 0$, and $\|u\|_{Q_c}^2 = u^\top Q_c u$. Formally,

$$d_D(z) = \min_c \min_{u \in D_c} \|z - u\|_{Q_c}, \qquad \beta_D(z) = \exp[-k\, d_D(z)], \quad k > 0. \tag{5}$$

With the fitness coefficient $\beta_t = \beta_D(z_t)$, we combine the current and retrieved features and scale the result, which can be formally represented as follows:

$$a_D(z_t) = z_t + A(r_D(z_t)), \qquad e_t = E_D(z_t) = \beta_t a_D(z_t). \tag{6}$$

where $a_D(z_t)$ is the unscaled fused feature and e_t is the enhanced feature passed to the GRU. A larger bank distance produces a smaller β_t, thereby attenuating features from poorly covered scenes. The same coefficient is also used for open-scene detection in Section 3.4. Persistent storage makes scene representatives available beyond the current batch, while scene-conditioned retrieval supplies a complementary feature when the current observation has limited support (**A2**). To model the temporal evolution of accident cues, we feed the enhanced feature directly into a gated recurrent unit (GRU) and compute the accident probability, which can be defined as follows:

$$h_t = H(e_t, h_{t-1}) = \phi_{\textbf{GRU}}(e_t, h_{t-1}), \qquad p_t = \sigma(f_{\text{out}}(h_t))\,. \tag{7}$$

Here, h_t is the recurrent hidden state that integrates the enhanced feature e_t with the preceding temporal context. The learnable readout f_{out} maps h_t to a scalar logit, and σ is the sigmoid function, yielding the frame-level accident probability p_t based on observations up to time t. Appendix D.3 derives a bound on the resulting probability changes under perturbations of the scene features and knowledge bank, with the recurrent predictor fixed.

Effect of Scene Knowledge Integration. Relevant retrieval does not necessarily improve anticipation, because β_t controls both knowledge injection and preservation of the original feature. We therefore analyze the complete representation change $\delta_t = e_t - z_t = \beta_t A(r_D(z_t)) - (1 - \beta_t) z_t$.

For a fixed labeled video V, bank, and downstream model, let $\mathcal{J}_V : \mathbb{R}^{Td} \to \mathbb{R}$ be the weighted anticipation loss defined in Appendix D.1. Denote the stacked features and displacements by $\mathbf{z} = (z_1, \ldots, z_T)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_T)$, and let $\xi_t = [\nabla \mathcal{J}_V(\mathbf{z})]_t$ be the gradient with respect to z_t.

Proposition 1 (Sufficient condition for loss reduction)**.** *Assume that $\mathcal{J}_V$ has an L_J-Lipschitz gradient near $[\mathbf{z}, \mathbf{z} + \boldsymbol{\delta}]$, with $L_J > 0$. Define the first-order gain as follows:*

$$\mathcal{G}_V = -\sum_{t=1}^{T} \langle \xi_t, \delta_t \rangle, \quad \textit{such that} \quad \mathcal{J}_V(\mathbf{z} + \boldsymbol{\delta}) - \mathcal{J}_V(\mathbf{z}) \le -\mathcal{G}_V + \frac{L_J}{2} \|\boldsymbol{\delta}\|_2^2, \tag{8}$$

where $\|\boldsymbol{\delta}\|_2^2 = \sum_{t=1}^{T} \|\delta_t\|_2^2$. Thus, a strict loss reduction is guaranteed when $\mathcal{G}_V > \frac{L_J}{2} \|\boldsymbol{\delta}\|_2^2$.

The proof is provided in Appendix D.1. This result clarifies when scene knowledge integration can improve anticipation. Specifically, retrieving relevant knowledge alone is insufficient; the complete feature displacement must be aligned with a loss-reducing direction, and its magnitude must remain controlled. β_t balances knowledge injection against preservation of the original scene evidence.

3.4 OPEN-SCENE DETECTION

Detection Criterion. A scene far from stored representatives may not be covered by the knowledge bank. We flag a scene as potentially unknown when its bank fitness falls below a threshold:

$$\beta_D(z) < \tau \quad \Longleftrightarrow \quad d_D(z) > r_\tau, \qquad r_\tau = \frac{\log(1/\tau)}{k}, \quad 0 < \tau < 1. \tag{9}$$

The threshold τ governs scene acceptance. Low-fitness observations are retained as candidates for the acquisition and update stage below.

Coverage and Separation. We analyze when the fitness threshold can retain known scenes while rejecting unseen ones, with the encoder, category rule, metrics, and bank fixed. For a known category c, let $Z \mid C = c \sim P_c$ have mean m_c and dispersion $v_c = \mathbb{E}_{P_c}[\|Z - m_c\|_{Q_c}^2] < \infty$. Define

$$a_c = \min_{u \in D_c} \|u - m_c\|_{Q_c}, \quad R_D = \max_c \max_{u \in D_c} \|u - m_c\|_{Q_c}, \quad d_{\text{mean}}(z) = \min_c \|z - m_c\|_{Q_c}. \tag{10}$$

Here, a_c measures coverage around the category mean, R_D bounds the extent of bank representatives around their respective means, and d_{mean} measures separation from known categories.

Proposition 2 (Coverage-based separation)**.** *For $r_\tau > a_c$, the probability of incorrectly rejecting a known-category sample satisfies*

$$\Pr(\beta_D(Z) < \tau \mid C = c) \le \min\left\{1, \frac{v_c}{(r_\tau - a_c)^2}\right\}. \tag{11}$$

Moreover, any feature satisfying $d_{\text{mean}}(z) > r_\tau + R_D$ is rejected. Let $\mathcal{O}$ be a set of unseen-scene features satisfying $\inf_{z \in \mathcal{O}} d_{\text{mean}}(z) \ge \Delta_0$. A threshold that retains each known category with probability at least $1 - \eta$ and rejects all of $\mathcal{O}$ exists whenever

$$\max_c \left(a_c + \sqrt{\frac{v_c}{\eta}}\right) < \Delta_0 - R_D, \qquad 0 < \eta < 1. \tag{12}$$

The proof of Proposition 2 is provided in Appendix D.2. This condition compares the coverage required to retain known scenes with the separation available for rejecting unseen scenes. It gives a conditional answer to distinguishing rare known scenes from unseen ones: a known scene can be retained through adequate bank coverage, whereas an unseen scene is rejected when it is sufficiently separated from that coverage (**A3**). Thus, low training frequency alone is not a rejection criterion, and the open-scene flag remains distinct from the accident probability p_t.

Scene Acquisition and Model Updating. We rank candidate scenes with $\beta_D(z) < \tau$ by bank distance and the distance ratio between the two nearest scene categories, prioritizing limited coverage and category ambiguity. Selected videos are annotated with accident labels, onset times for positive videos, and scene-group assignments. During adaptation, we freeze the feature extractor, refine existing scene representatives, add entries for new groups, and update the retrieval parameters and anticipation head. Training combines temporally weighted frame-level anticipation, auxiliary video-level classification, and prototype-margin regularization. In addition, acquisition and adaptation procedures are detailed in Appendix C.3, and the training objective in Appendix C.4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate TailBank on three traffic accident anticipation benchmarks: DAD (Chan et al., 2017), CCD (Bao et al., 2020), and A3D (Yao et al., 2019). Additional scene-level evaluations use the frequency-based, error-ranked, and expert-curated DAD subsets introduced in Section 2.

Metrics. We report Average Precision (AP) to assess anticipation accuracy, together with mean Time-to-Accident (mTTA) and TTA at 80% recall (TTA@R80) to assess timeliness.

Implementation Details. TailBank is trained for 30 epochs on an NVIDIA RTX 4090 GPU. We use Adam with a batch size of 10 and an initial learning rate of 10^{-4}, adjusted using a ReduceLROnPlateau scheduler. More details for the experimental setup are in Appendix E.2

Table 1: Evaluation on three datasets: AP (%) / mTTA (s). Bold/underline mark the best/second-best values. Purple highlights TailBank; "–" denotes unavailable results.

Method	DAD	CCD	A3D
DSA (Chan et al., 2017)	48.1 / 1.34	98.7 / 3.08	92.3 / 2.95
ACRA (Zeng et al., 2017)	51.4 / 3.01	98.9 / 3.32	–
AdaLEA (Suzuki et al., 2018)	52.3 / 3.43	99.2 / 3.45	92.9 / 3.16
UString (Bao et al., 2020)	53.7 / 3.53	99.5 / 3.74	93.2 / 3.24
DSTA (Karim et al., 2022b)	59.2 / 2.90	99.4 / 3.87	93.5 / 2.87
GSC (Wang et al., 2024)	60.4 / 2.55	99.4 / 3.68	94.9 / 2.62
DAA-GNN (Song et al., 2024)	75.2 / 1.47	–	–
XAI-Accident (Karim et al., 2022a)	52.0 / 3.79	94.0 / **4.57**	92.8 / 3.35
AccNet (Liao et al., 2024a)	60.8 / 3.58	99.5 / 3.78	95.1 / 3.26
DDM (Guan et al., 2025)	87.7 / **4.47**	99.7 / 3.70	**95.4** / 3.18
ROAR (Liu et al., 2026)	76.0 / 3.92	**99.8** / 3.24	95.3 / 3.73
TailBank (Ours)	**96.7** / 4.36	99.5 / 4.51	**95.4** / **4.53**

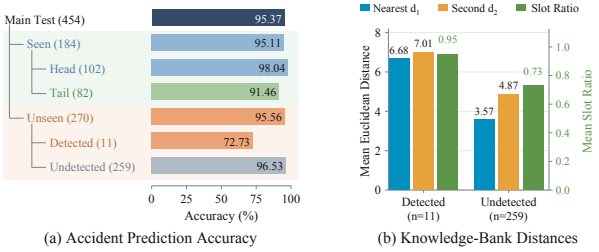
Table 2: Robustness evaluation on the DAD dataset under reduced training data and missing frames: AP (%) / mTTA (s). Bold/underline mark the best/second-best values among the displayed methods. Train: retained training videos; Drop: randomly removed frames; 1 in 5: periodic frame removal.

Setting	UString	DSTA	**TailBank**
Train 75%	51.15 / 2.67	52.79 / 2.71	**95.01 / 4.33**
Train 50%	52.68 / 2.38	52.78 / 2.52	**94.57 / 4.31**
Drop 10%	53.51 / 2.50	56.78 / 2.48	**97.10 / 4.03**
Drop 20%	52.27 / 2.47	55.89 / 2.48	**91.70 / 3.51**
Drop 50%	52.37 / 1.62	54.84 / **2.11**	**98.30** / 2.04
1 in 5	52.62 / 2.14	55.46 / 2.59	**92.80 / 3.50**

4.2 Comparison with State-of-the-Art Methods

Overall Performance. Table 1 compares TailBank with state-of-the-art (SOTA) accident anticipation methods on three benchmarks. On DAD dataset, TailBank achieves the highest AP of 96.7%, exceeding DDM by 9.0 percentage points, with a slightly lower mTTA (4.36 versus 4.47 s). On A3D dataset, it matches DDM's best AP of 95.4% and achieves the highest mTTA of 4.53 s. On CCD dataset, where several methods exceed 99% AP, TailBank obtains 99.5% AP and 4.51 s mTTA.

Data Efficiency and Robustness. We evaluate data efficiency using 75% and 50% of the training videos, and robustness under random removal of 10%, 20%, and 50% of frames, together with periodic removal of every fifth frame. On DAD, TailBank outperforms UString and DSTA in AP across all six settings and achieves the highest mTTA in five, with 50% frame removal being the exception (Table 2). Evaluation results across all three datasets are provided in Table 5.



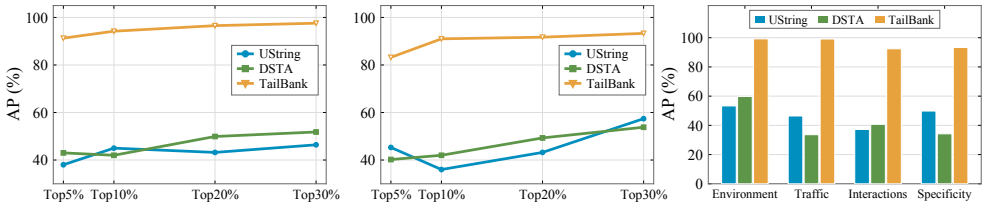
(a) Accident Prediction Accuracy (b) Knowledge-Bank Distances

Figure 3: Anticipation and screening on held-out scenarios. (a) Accident classification accuracy, with group sizes in parentheses. (b) Mean nearest and second-nearest bank-slot distances (left axis) and mean nearest-to-second-nearest slot ratio (right axis), shown separately for the 11 flagged and 259 unflagged videos from held-out scenario categories.

Long-Tail Performance. We evaluate TailBank on DAD defined by reference-model error, micro-cluster frequency, and expert curation (Appendix E.2). TailBank outperforms UString and DSTA in AP at all cutoffs in the error-ranked and micro-cluster evaluations (Figure 4(a,b)). On the 5% micro-cluster subset, it exceeds DSTA by 33.0 percentage points in AP and 0.98 s in reported mTTA, demonstrating gains on statistically rare scenes. Across four expert-curated groups, it achieves 92.4-99.2% AP (Figure 4(c)) and 4.08-4.36 s mTTA, supporting its performance in safety-relevant scenes.

4.3 Unseen-Scene Generalization and Screening

Figure 3(a) evaluates TailBank on 270 videos from held-out scene categories, yielding 95.56% accident anticipation accuracy. Fitness-based screening (Section 3.4) flags 11 videos (4.07%). At the same validation-calibrated warning threshold, accuracy is 72.73% for flagged videos and 96.53% for unflagged videos. Although the flagged subset is small, it contains 3 of the 12 observed errors (25%), indicating that screening concentrates a disproportionate share of errors within a small review set. Figure 3(b) further characterizes this subset through its distances to the knowledge bank. Flagged videos have larger nearest-bank distances and nearest-to-second-nearest bank-slot distance ratios closer to one, indicating more distant and less distinctive matches to stored representatives. These statistics describe the support available from the bank; the accompanying accuracy gap provides



(a) Prediction Error (b) Micro-Cluster Frequency (c) Expert Curation

Figure 4: Evaluation results on DAD scene subsets. AP (%) of UString, DSTA, and TailBank on subsets defined by (a) reference-model error, (b) micro-cluster frequency, and (c) expert curation. Panels (a,b) use subset cutoffs of 5%, 10%, 20%, and 30%, while (c) compares environmental conditions, traffic complexity, dynamic interactions, and scenario specificity. These views assess difficulty, rarity, and safety relevance. Higher AP is better; subset definitions are in Appendix E.2.

evidence that the selected cases are also more difficult for the anticipation model. This supports prioritizing them for human review and further annotation.

These findings clarify the distinction raised in **Q3**: an unseen scene category can still admit a correct accident decision, while limited bank support can motivate review independently of the accident score. TailBank therefore assesses scene familiarity separately from accident likelihood. The coverage analysis in Section 3.4 specifies conditions under which sparse known scenes remain accepted. Empirically, the present results support targeted review of challenging held-out scenes, although the low flagging rate indicates limited recall of unseen categories at the tested threshold (**A3**).

4.4 Ablation Analysis

Effect of Individual Components. Table 3(a) evaluates each component using retrained ablated variants. Removing Scene Knowledge Integration causes the largest performance decline: AP drops from 97.37% to 79.54%, with both timing metrics decreasing. Disabling fitness modulation or the Scene-Object-Aware Module (SOAM) reduces AP by 4.05 and 5.44 percentage points, respectively, supporting their contributions to accident discrimination. These gains do not extend uniformly to warning timeliness: removing fitness modulation slightly increases both mTTA and TTA@R80, while removing SOAM increases TTA@R80. The full model achieves the highest AP, with timing trade-offs.

Table 3: DAD ablations: (a) retrained module variants and (b) fusion interventions with the full model frozen. Here, Full denotes the checkpoint with the highest AP, whereas Table 1 reports a checkpoint selected to balance AP and mTTA. Best values within each panel are bold. ✓/×: enabled/disabled; —: not applied.

(a) Retrained module ablations

Variant	HKE	Fitness	OAM	AP (%)↑	mTTA(s)↑	TTA@R80↑
w/o HKE	×	—	✓	79.54	3.3092	3.9237
w/o fitness	✓	×	✓	93.32	**3.8463**	4.2944
w/o SOAM	✓	✓	×	91.93	3.7280	**4.3581**
Full	✓	✓	✓	**97.37**	3.7779	4.2761

(b) Frozen-Model fusion interventions

Variant	Feature		Gate		AP (%)	mTTA (s)	TTA@R80
	V	K	V	K	↑	↑	↑
Visual only	✓	×	×	×	72.03	3.2518	3.8379
Additive	✓	✓	×	×	95.42	3.7578	**4.3325**
K-gated	✓	✓	×	✓	94.47	**3.7869**	4.3222
V-gated only	✓	×	✓	×	70.86	3.3548	3.8002
Full	✓	✓	✓	✓	**97.37**	3.7779	4.2761

Effect of Fitness-Guided Fusion. Table 3(b) examines the fusion rule with all parameters of the trained model held fixed. Let $v = z$ denote the scene feature and $u = A(r_D(z))$ the transformed knowledge retrieved from the bank. We compare five decoder inputs: v, $v + u$, $v + \beta u$, βv, and $\beta(v + u)$. Adding retrieved knowledge raises AP from 72.03% to 95.42% and improves both timing metrics. Full fusion further increases AP to 97.37%, exceeding additive fusion and knowledge-only gating by 1.95 and 2.90 percentage points, respectively. However, scaling the visual feature alone yields 70.86% AP, below the visual-only input. Thus, within the trained model, the AP benefit of fitness modulation depends on its combination with retrieved knowledge; attenuation of visual evidence alone does not reproduce the benefit of the complete fusion rule.

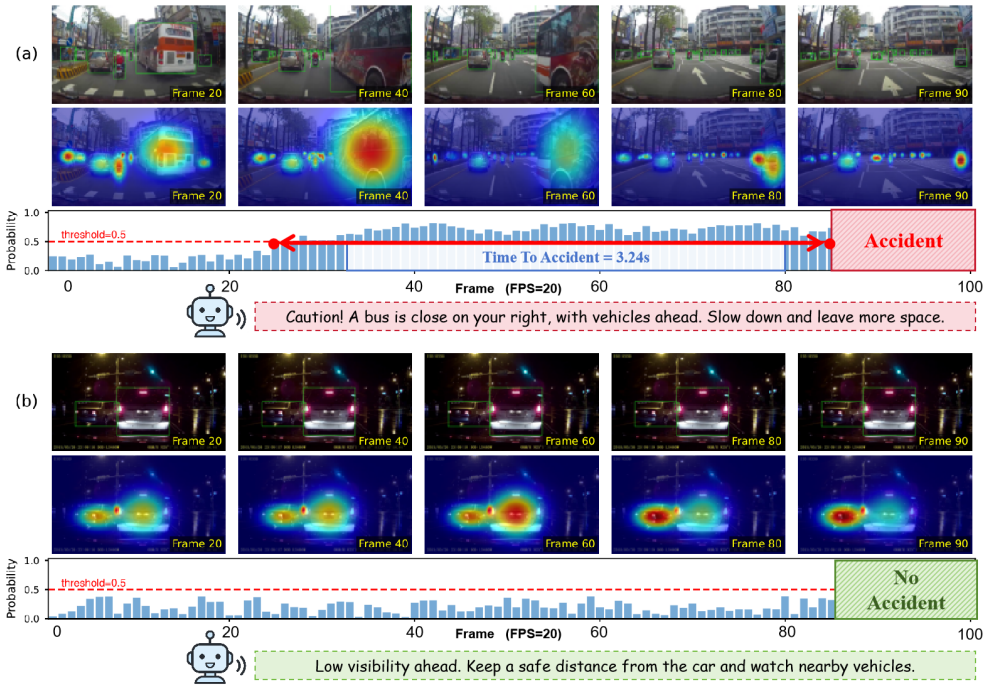


Figure 5: Qualitative results for (a) an accident sequence and (b) a non-accident sequence. Each example shows object detections, attention overlays, frame-level accident probabilities, and illustrative language feedback. Red dashed lines mark the warning threshold of 0.5 used for both sequences.

4.5 Visualization and Interpretability Analysis

Figure 5 shows how spatial attention and accident scores evolve, alongside illustrative language feedback. In (a), attention concentrates on the bus at frames 20-40 and shifts toward the right boundary as its relative position changes. Later maps show more localized responses around the remaining traffic regions. The accident score crosses the warning threshold before the marked accident interval, illustrating a response to the scene before the collision occurs. In (b), foreground vehicles also attract attention under low illumination, yet the score remains below the threshold. The contrast shows that strong attention to a vehicle need not imply high accident risk: its significance must be interpreted within the surrounding scene and temporal context. The accompanying feedback makes these contexts explicit, highlighting the nearby bus and traffic ahead in (a), and recommending visibility and following-distance precautions in (b). These examples complement the quantitative findings for **A2**, illustrating both early warning and avoidance of a false alarm in a challenging non-accident scene. Further examples under varied traffic and illumination conditions are provided in Appendix E.6.

5 Conclusion

We introduced a scene-level formulation of long-tailed accident anticipation and TailBank, which combines persistent scene knowledge with coverage-guided feature integration. Our complementary evaluations distinguish statistical rarity, anticipation difficulty, and safety relevance. A central insight is that balanced accident labels do not ensure adequate scene coverage: knowledge reuse requires assessing how well stored experience supports the current scene, while keeping scene familiarity distinct from accident likelihood. Our analysis characterizes conditions for beneficial enhancement and known/open-scene separation. Extensive experiments support this approach, with substantial AP gains on corner-case scenes and trade-offs in warning timeliness. Nevertheless, scene characterization depends on the chosen representation and grouping, and open-scene screening has limited recall at the evaluated threshold. Future work will explore more transferable scene representations and jointly assess known-tail retention and unseen-scene detection under distribution shift.

6 LLM Assistance Statement

In this work, we used generative AI tools solely to assist with writing and polishing, including improving language and readability, refining manuscript organization, and standardizing notation and equation formatting. We did not use generative AI tools for literature retrieval or discovery, research ideation or execution, drafting manuscript sections, generating synthetic datasets, or proving mathematical claims. We reviewed all AI-assisted edits to ensure that they preserved the intended technical meaning and did not introduce unsupported claims. We take responsibility for the final content of this work, including all AI-assisted text.

References

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. URL https://doi.org/10.1145/3394171.3413827.

Abhijit Bendale and Terrance E. Boult. Towards Open Set Deep Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572, 2016. doi: 10.1109/cvpr.2016.173. URL https://doi.org/10.1109/cvpr.2016.173.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018. doi: 10.1109/cvpr.2018.00644. URL https://doi.org/10.1109/cvpr.2018.00644.

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, 2017. doi: 10.1007/978-3-319-54190-7_9. URL https://doi.org/10.1007/978-3-319-54190-7_9.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953. URL https://doi.org/10.1613/jair.953.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019. doi: 10.1109/cvpr.2019.00949. URL https://doi.org/10.1109/cvpr.2019.00949.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):5890–5904, 2024. doi: 10.1109/tpami.2024.3369102. URL https://doi.org/10.1109/tpami.2024.3369102.

Jianwu Fang, Jiahuan Qiao, Jianru Xue, and Zhengguo Li. Vision-based traffic accident detection and anticipation: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 34 (4):1983–1999, 2024.

Jianwu Fang, Lei-Lei Li, Zhedong Zheng, Hongkai Yu, Jianru Xue, Zhengguo Li, and Tat-Seng Chua. Eq-taa: Equivariant traffic accident anticipation via diffusion-based accident video synthesis. *IEEE Transactions on Multimedia*, 2025.

Yanchen Guan, Haicheng Liao, Chengyue Wang, Bonan Wang, Jiaxun Zhang, Jia Hu, and Zhenning Li. Domain-enhanced dual-branch model for efficient and interpretable accident anticipation. *Communications in Transportation Research*, 5:100214, 2025. doi: 10.1016/j.commtr.2025.100214. URL https://doi.org/10.1016/j.commtr.2025.100214.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Muhammad Monjurul Karim, Yu Li, and Ruwen Qin. Toward Explainable Artificial Intelligence for Early Anticipation of Traffic Accidents. *Transportation Research Record: Journal of the Transportation Research Board*, 2676(6):743–755, 2022a. doi: 10.1177/03611981221076121. URL https://doi.org/10.1177/03611981221076121.

Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin. A Dynamic Spatial-Temporal Attention Network for Early Anticipation of Traffic Accidents. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9590–9600, 2022b. doi: 10.1109/tits.2022.3155613. URL https://doi.org/10.1109/tits.2022.3155613.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1706.02690.

Haicheng Liao, Yongkang Li, Zhenning Li, Zilin Bian, Jaeyoung Lee, Zhiyong Cui, Guohui Zhang, and Chengzhong Xu. Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3D modeling. *Accident Analysis & Prevention*, 207:107760, 2024a. doi: 10.1016/j.aap.2024.107760. URL https://doi.org/10.1016/j.aap.2024.107760.

Haicheng Liao, Yongkang Li, Chengyue Wang, Yanchen Guan, Kahou Tam, Chunlin Tian, Li Li, Chengzhong Xu, and Zhenning Li. When, where, and what? a benchmark for accident anticipation and localization with large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8–17, 2024b.

Haicheng Liao, Huanming Shen, Zhenning Li, Chengyue Wang, Guofa Li, Yiming Bie, and Chengzhong Xu. GPT-4 Enhanced Multimodal Grounding for Autonomous Driving: Leveraging Cross-Modal Attention with Large Language Models. *Communications in Transportation Research*, 4:100116, 2024c. doi: 10.1016/j.commtr.2023.100116. URL https://doi.org/10.1016/j.commtr.2023.100116.

Haicheng Liao, Haoyu Sun, Huanming Shen, Chengyue Wang, Chunlin Tian, KaHou Tam, Li Li, Chengzhong Xu, and Zhenning Li. CRASH: Crash Recognition and Anticipation System Harnessing with Context-Aware and Temporal Focus Attentions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11041–11050, 2024d. doi: 10.1145/3664647.3680672. URL https://doi.org/10.1145/3664647.3680672.

Haicheng Liao, Bin Rao, Haoyu Sun, Chengyue Wang, Qing Chang, Shengbo Eben Li, Chengzhong Xu, and Zhenning Li. Chain-of-thought guided multimodal large language models for scene-aware accident anticipation in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017. doi: 10.1109/iccv.2017.324. URL https://doi.org/10.1109/iccv.2017.324.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-Based Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Xingcheng Liu, Yanchen Guan, Haicheng Liao, Zhengbing He, and Zhenning Li. ROAR: Robust accident recognition and anticipation for autonomous driving. *Accident Analysis & Prevention*, 228:108414, 2026. doi: 10.1016/j.aap.2026.108414. URL https://doi.org/10.1016/j.aap.2026.108414.

Bin Rao, Haicheng Liao, Yanchen Guan, Chengyue Wang, Bonan Wang, Jiaxun Zhang, and Zhenning Li. AMD: Adaptive momentum and decoupled contrastive learning framework for robust long-tail trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28849–28858, October 2025.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004–4012, 2016. doi: 10.1109/cvpr.2016.434. URL https://doi.org/10.1109/cvpr.2016.434.

Wenfeng Song, Shuai Li, Tao Chang, Ke Xie, Aimin Hao, and Hong Qin. Dynamic attention augmented graph network for video accident anticipation. *Pattern Recognition*, 147: 110071, 2024. doi: 10.1016/j.patcog.2023.110071. URL https://doi.org/10.1016/j.patcog.2023.110071.

Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating Traffic Accidents with Adaptive Loss and Large-Scale Incident DB. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3521–3529, 2018. doi: 10.1109/cvpr.2018.00371. URL https://doi.org/10.1109/cvpr.2018.00371.

Tianhang Wang, Kai Chen, Guang Chen, Bin Li, Zhijun Li, Zhengfa Liu, and Changjun Jiang. GSC: A Graph and Spatio-Temporal Continuity Based Framework for Accident Anticipation. *IEEE Transactions on Intelligent Vehicles*, 9(1):2249–2261, 2024. doi: 10.1109/tiv.2023.3257169. URL https://doi.org/10.1109/tiv.2023.3257169.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to Model the Tail. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/147ebe637038ca50a1265abac8dea181-Abstract.html.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024. doi: 10.1109/lra.2024.3440097. URL https://doi.org/10.1109/lra.2024.3440097.

Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, and Ella M. Atkins. Unsupervised Traffic Accident Detection in First-Person Videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 273–280, 2019. doi: 10.1109/iros40897.2019.8967556. URL https://doi.org/10.1109/iros40897.2019.8967556.

Kuo-Hao Zeng, Shih-Han Chou, Fu-Hsiang Chan, Juan Carlos Niebles, and Min Sun. Agent-Centric Risk Assessment: Accident Anticipation and Risky Region Localization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2222–2230, 2017. doi: 10.1109/cvpr.2017.146. URL https://doi.org/10.1109/cvpr.2017.146.

Jiaxun Zhang, Yanchen Guan, Chengyue Wang, Haicheng Liao, Guohui Zhang, and Zhenning Li. Latte: A real-time lightweight attention-based traffic accident anticipation engine. *Information Fusion*, 122:103173, 2025.

Siyao Zhang, Daocheng Fu, Wenzhe Liang, Zhao Zhang, Bin Yu, Pinlong Cai, and Baozhen Yao. TrafficGPT: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, 150:95–105, 2024. doi: 10.1016/j.tranpol.2024.03.006. URL https://doi.org/10.1016/j.tranpol.2024.03.006.

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range Loss for Deep Face Recognition with Long-Tailed Training Data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5409–5418, 2017. doi: 10.1109/iccv.2017.578. URL https://doi.org/10.1109/iccv.2017.578.

A Notation and Organization

The appendix reviews related work (Appendix B), describes the implementation and training components (Appendix C), provides the theoretical derivations (Appendix D), and reports supplementary evaluation protocols and results (Appendix E).

Table 4: Principal notation. Training scene categories c and evaluation clusters j are distinct.

Symbol	Meaning
V, x_t, T, Y	Video, frame, sequence length, and binary accident label.
C, c, M	Training scene category, its index, and the number of populated categories.
$\mathcal{K}_j, J, q$	Evaluation cluster, number of clusters, and target tail-subset fraction.
$z_t, r_D(z_t), e_t$	Scene embedding, retrieved feature, and enhanced GRU input.
$D_c, D, \mathbf{D}, K$	Category bank, full bank, ordered bank matrix, and number of stored representatives.
$\widehat{\mu}_c, \nu_c, m_c$	Empirical centroid, trainable prototype, and population feature mean.
N_c, P_c, π_c, v_c	Construction sample count, conditional feature distribution, category frequency, and metric second moment.
Q_c, Σ_c, s	Category metric, regularized covariance with $Q_c = \Sigma_c^{-1}$, and eigenvalue lower bound.
$d_D, \beta_D, k, \tau, r_\tau$	Bank distance, fitness, exponential scale, fitness threshold, and corresponding distance threshold.
A, a_D, E_D	Knowledge adapter, unscaled fusion, and fitness-scaled enhancement.
$h_t, H, f_{\text{out}}, p_t$	GRU state, recurrent map, logit readout, and accident probability.
$\theta_{\text{acc}}, t^*, t_0, f_{\text{fps}}$	Accident-warning threshold, onset frame, first warning frame, and frame rate.
$\delta_t, \mathcal{J}_V, \mathcal{G}_V, L_J$	Feature displacement, fixed-predictor loss, first-order gain, and gradient Lipschitz constant.
a_c	Central coverage error relative to the category mean.
$R_D, d_{\text{mean}}, \Delta_0$	Representative radius, distance to known means, and open-scene separation level.
η	Target upper bound on the known-scene rejection probability.
$\epsilon_D, \epsilon_r, \rho_t$	Bank discrepancy, retrieval discrepancy, and scene-feature perturbation.
L_S, L_V, L_M, w_a, w_b	Frame loss, auxiliary video loss, prototype-margin loss, and auxiliary-loss weights.

B Related Work

B.1 Traffic Accident Anticipation

Video-based accident anticipation estimates the risk of a future collision from observations available before its onset. Existing methods primarily address the spatial and temporal modeling of emerging hazards. Chan et al. (Chan et al., 2017) introduced dynamic spatial attention to identify relevant objects in dashcam videos. AdaLEA (Suzuki et al., 2018) developed an adaptive loss to encourage earlier predictions, while Bao et al. (Bao et al., 2020) modeled interactions among traffic participants through an uncertainty-aware graph-based recurrent architecture. Subsequent methods further exploit attention and temporal structure: DSTA (Karim et al., 2022b) models dynamic spatiotemporal dependencies, GSC (Wang et al., 2024) incorporates graph representations and spatiotemporal continuity, and CRASH (Liao et al., 2024d) combines context-aware and temporal-focus attention.

These approaches provide mechanisms for extracting and integrating accident cues within a video. TailBank addresses the complementary question of how a temporal predictor can access scene knowledge when the current driving configuration has limited training support. It retains scene representatives across batches, retrieves a scene-conditioned feature, and modulates the fused representation before recurrent prediction. This connects temporal hazard modeling with the availability of knowledge about infrequently observed scenes.

B.2 Learning from Imbalanced and Long-Tailed Data

Imbalanced learning seeks to reduce the dominance of frequent categories during training. Data-level methods such as SMOTE (Chawla et al., 2002) increase the representation of minority classes, while focal loss (Lin et al., 2017) emphasizes difficult examples and class-balanced loss (Cui et al., 2019) adjusts training contributions according to class frequency. Representation-level approaches include lifted structured embedding (Song et al., 2016) and range loss (Zhang et al., 2017), which organize within-class and between-class feature relationships. Knowledge transfer from head to tail categories also motivates meta-model learning (Wang et al., 2017). More recently, ProCo (Du et al., 2024) estimates class-conditional feature distributions across batches and derives a probabilistic contrastive objective.

For accident anticipation, binary outcome imbalance and scene-frequency imbalance describe different properties of the data. Balancing accident and non-accident videos does not ensure adequate representation of rare road configurations, participant interactions, or environmental conditions. TailBank therefore organizes its memory by scene pseudo-categories while retaining the binary accident-prediction objective. Its bank stores finite scene representatives for retrieval and coverage estimation. The accompanying analysis relates bank coverage to known-scene retention and open-scene rejection, characterizes when feature enhancement reduces prediction loss, and bounds the sensitivity of recurrent predictions to feature and bank perturbations.

B.3 Open-Scene Detection

Open-set recognition and out-of-distribution detection assess whether an observation is supported by the training distribution. Representative approaches use maximum softmax confidence (Hendrycks & Gimpel, 2017), extreme-value modeling of activation patterns in OpenMax (Bendale & Boult, 2016), temperature scaling and input perturbations in ODIN (Liang et al., 2018), energy scores (Liu et al., 2020), or Mahalanobis distances in feature space (Lee et al., 2018). In long-tailed settings, the detection criterion must also accommodate known categories with sparse training support.

TailBank measures coverage through the distance to the nearest stored scene representative. The resulting fitness coefficient both scales the fused feature and identifies candidate open scenes. Its coverage analysis states conditions under which known categories can be retained while sufficiently separated unseen scenes are rejected. Low-fitness observations can additionally be selected for annotation and subsequent bank updates, extending the scene knowledge available to the predictor.

B.4 Vision-Language Feedback for Driving

Language and vision-language models provide interfaces for interpreting traffic observations and communicating driving-related information. TrafficGPT (Zhang et al., 2024), DriveGPT4 (Xu et al., 2024), and multimodal grounding approaches such as CAVG (Liao et al., 2024c) explore language-based interaction, driving interpretation, and the association between visual evidence and textual descriptions. These studies motivate presenting anticipation outputs in a form that relates a numerical warning to the observed scene.TailBank includes an optional interface that combines the observed video, accident-probability trace, and identified participants to produce an event description, a risk explanation, and a suggested response.

C Architecture, Training, and Adaptation

C.1 Context and Object Encoding

Let f_t^{ctx} denote the VGG-16 feature of frame x_t, and let $f_{t,i}^{\mathrm{obj}}$ denote the VGG-16 feature of its i-th detected participant. The context branch applies partial convolution, a feature split, depth-wise convolution, and a gating contraction, which can be formally represented as follows:

$$[u_{t,1}; u_{t,2}] = \mathrm{GELU}(W_1\,\mathrm{PConv}(f_t^{\mathrm{ctx}})), \tag{13}$$

$$z_t^{\mathrm{ctx}} = \mathrm{GELU}(W_2\mathcal{T}(u_{t,1}, \mathrm{DWConv}(u_{t,2}))\,. \tag{14}$$

Here, W_1 and W_2 are learned projections, PConv and DWConv denote partial and depth-wise convolution, and $\mathcal{T}$ denotes the context branch's gating contraction.

For a frame with n_t detected participants, the attention score for participant i is defined as $s_{t,i} = w_{\text{att}}^{\top} \tanh(W_{\text{att}} f_{t,i}^{\text{obj}} + b_{\text{att}})$. The normalized attention weights and the object-aware feature can then be computed as follows:

$$\omega_{t,i} = \frac{\exp(s_{t,i})}{\sum_{j=1}^{n_t} \exp(s_{t,j})}, \qquad z_t^{\text{obj}} = \sum_{i=1}^{n_t} \omega_{t,i} f_{t,i}^{\text{obj}}. \tag{15}$$

The softmax is taken over detected participants. The context and object outputs are concatenated and projected into $z_t \in \mathbb{R}^d$ as in Section 3.2.

C.2 Knowledge-Bank Construction and Updates

Scene categories and representatives. The training assignment rule partitions construction features into M populated scene groups $\mathcal{B}_c$. These groups index the bank and are separate from the evaluation clusters in Section 2. Let dist_{sel} denote the centroid-distance ordering used for representative selection. Given the empirical centroid in Equation 2, the nearest and farthest representatives are selected as follows:

$$u_c^{\text{near}} \in \arg\min_{u \in \mathcal{B}_c} \text{dist}_{\text{sel}}(u, \widehat{\mu}_c), \qquad u_c^{\text{far}} \in \arg\max_{u \in \mathcal{B}_c} \text{dist}_{\text{sel}}(u, \widehat{\mu}_c). \tag{16}$$

The matrix $\mathbf{D} \in \mathbb{R}^{d \times K}$ stores the representatives in category/slot order; $W_g \in \mathbb{R}^{K \times d}$ and $b_g \in \mathbb{R}^K$ generate the retrieval weights in Equation 3. Covariance regularization enforces $\Sigma_c \succeq sI$, so the metric in Equation 5 remains positive definite.

Persistence across batches. For an independently sampled batch of size B, the probability that a category with frequency π_c is absent can be expressed as follows:

$$\Pr(N_c^{\text{batch}} = 0) = (1 - \pi_c)^B. \tag{17}$$

Keeping D_c beyond the current batch makes its representatives available for retrieval even when the category is absent.

Prototype updates. Network optimization alternates with prototype updates. For category c in batch b, let $N_c^{(b)}$ be its feature count. The prototype gradient and the corresponding update can be defined as follows:

$$g_c^{(b)} = \frac{\sum_{i=1}^{N_c^{(b)}} (\nu_c^{(b)} - Z_{c,i}^{(b)})}{1 + N_c^{(b)}}, \qquad \nu_c^{(b+1)} = \nu_c^{(b)} - \eta_\nu g_c^{(b)}, \qquad 0 < \eta_\nu \leq 1. \tag{18}$$

An absent category has a zero update. The prototype ν_c participates in training, whereas $\widehat{\mu}_c$ is the empirical centroid used to select stored representatives.

C.3 Open-Scene Acquisition and Model Updating

Candidate scoring. For category c, define its nearest-representative distance $d_c(z) = \min_{u \in D_c} \|z - u\|_{Q_c}$. Let c_1 attain the minimum, with ties resolved consistently. The distances to the nearest and competing categories are defined as follows:

$$d_1 = d_{c_1}(z) = d_D(z), \qquad d_2 = \min_{c \neq c_1} d_c(z). \tag{19}$$

For a candidate with $\beta_D(z) < \tau$ and at least two populated categories, the acquisition score can be defined as follows:

$$U_{\text{info}}(z) = \alpha_{\text{sel}} d_1 + (1 - \alpha_{\text{sel}}) \frac{d_1}{d_2}, \qquad 0 < \alpha_{\text{sel}} < 1. \tag{20}$$

Since $d_2 \geq d_1 > r_\tau > 0$, the ratio is well defined. The first term prioritizes poor bank coverage; the second emphasizes similar distances to competing scene categories. The distance scale and α_{sel} remain fixed within each acquisition round. Candidate videos are ranked by their acquisition scores and selected within the annotation budget. Annotators review each selected video to determine

whether an accident occurs and, for positive videos, mark its onset frame. They also assess the scene configuration according to a consistent annotation protocol, assigning the video to an existing scene group or identifying a new group when the current categories do not adequately describe it. For a unique nearest category with $d_2 > d_1$, its identity is unchanged under a feature perturbation satisfying the following bound:

$$\|\Delta z\|_2 < \frac{\sqrt{s}}{2}(d_2 - d_1) = \frac{\sqrt{s}\, d_2}{2}\left(1 - \frac{d_1}{d_2}\right). \tag{21}$$

Each category distance changes by at most $\|\Delta z\|_2/\sqrt{s}$, so this condition preserves the ordering of the nearest and competing categories. At fixed d_2, a ratio closer to one gives a smaller certified radius, motivating the relative-ambiguity term in Equation 20.

Updating the model and bank. Let $\mathcal{T}_r$ be the labeled training videos before round r and $\mathcal{A}_r$ the newly annotated videos. Form $\mathcal{T}_{r+1} = \mathcal{T}_r \cup \mathcal{A}_r$. The detector and VGG backbones remain fixed, while the scene modules, retrieval parameters, adapter, and recurrent prediction components are updated with Equation 22. Prototypes follow Equation 18. New categories are initialized from their annotated training features.

After an adaptation epoch, the updated embeddings determine new empirical centroids and near/far representatives. Category covariances are re-estimated with the eigenvalue floor s. Bank slots retain a consistent category/representative ordering during optimization. If the refreshed bank has K' entries, the retrieval layer is resized to output K' logits, preserving parameters for retained slots and learning parameters for new slots. The accident readout remains binary.

C.4 Training Objective

We formulate a joint objective for TailBank by combining three loss components as follows:

$$L = L_S + w_a L_V + w_b L_M, \tag{22}$$

where L_S is the frame-level anticipation loss, L_V provides auxiliary video-level classification supervision, and L_M is the prototype-margin regularizer. The weights w_a and w_b balance the auxiliary terms against the frame-level objective. The component losses are defined below.

Frame-level anticipation. For N_{vid} training videos, let Y_v and T_v denote the label and length of video v. With nonnegative temporal weights $\omega^+_{t,v}$ and $\omega^-_{t,v}$, the frame-level loss can be defined as follows:

$$L_S = -\sum_{v=1}^{N_{\text{vid}}} \sum_{t=1}^{T_v} \left[\omega^+_{t,v} Y_v \log p_{t,v} + \omega^-_{t,v}(1 - Y_v)\log(1 - p_{t,v})\right]. \tag{23}$$

The temporal weights remain inside the frame sum and control the contribution of each observation.

Auxiliary video classification. Write $a_v^{\text{vid}} \in (0, 1)$ for the auxiliary video-level prediction, distinct from the fused feature $a_D(z_t)$. Its binary classification loss can be expressed as follows:

$$L_V = -\sum_{v=1}^{N_{\text{vid}}} \left[Y_v \log a_v^{\text{vid}} + (1 - Y_v)\log(1 - a_v^{\text{vid}})\right]. \tag{24}$$

The auxiliary score supplies training supervision; the online accident score is $p_{t,v}$.

Prototype-margin regularization. For a batch of enhanced features $\{e_i\}_{i\in\mathcal{I}_b}$, let $S_c \subseteq \mathcal{I}_b$ contain the features assigned to category c. The category-wise margin term can be defined as follows:

$$L_{M,c} = \left[\sum_{i\in S_c} \|e_i - \nu_c\|_2 - \lambda_{\text{rep}} \sum_{i\in\mathcal{I}_b\setminus S_c} \|e_i - \nu_c\|_2 + \gamma_m\right]_+, \qquad \lambda_{\text{rep}} = 0.01, \quad \gamma_m = 5.0, \tag{25}$$

where $[x]_+ = \max\{x, 0\}$. The margin γ_m sets the required gap between the weighted repulsive distance sum and the attractive distance sum, while λ_{rep} controls the contribution of the repulsive term without scaling the attractive term. The category terms are combined into L_M using the training reduction. Together, these terms form $L = L_S + w_a L_V + w_b L_M$ in Equation 22.

D THEORETICAL ANALYSIS

We prove the loss-reduction and coverage-separation results stated in the main text, and then analyze how feature and knowledge-bank perturbations propagate through the recurrent predictor. Each result specifies the quantities held fixed in its comparison.

D.1 PROOF OF PROPOSITION 1: LOSS REDUCTION

For Proposition 1, the weighted loss for a fixed predictor can be written as follows:

$$\mathcal{J}_V(\mathbf{w}) = \sum_{t=1}^{T} \lambda_{t,V} \left[\log(1 + e^{o_t(\mathbf{w})}) - Y o_t(\mathbf{w}) \right], \quad \lambda_{t,V} \geq 0, \quad \sum_t \lambda_{t,V} = 1. \tag{26}$$

The logit $o_t(\mathbf{w})$ is the output of the full fixed downstream predictor receiving sequence $\mathbf{w}$. The same labels, temporal weights, and initial hidden state are used for both feature sequences; the weights do not depend on predictions or the bank.

Complete Enhancement Displacement. For each frame, the complete enhancement displacement can be expressed as follows:

$$E_D(z_t) - z_t = \beta_t\big(z_t + A(r_D(z_t))\big) - z_t = \beta_t A(r_D(z_t)) - (1 - \beta_t) z_t. \tag{27}$$

This identity includes both knowledge injection and attenuation of the original feature. With $\xi_t = \nabla_{w_t} \mathcal{J}_V(\mathbf{z})$, the alignment in Equation 8 can be expanded as follows:

$$\mathcal{G}_V = -\sum_t \beta_t \langle \xi_t, A(r_D(z_t)) \rangle + \sum_t (1 - \beta_t) \langle \xi_t, z_t \rangle. \tag{28}$$

Although the loss sums frame contributions, $o_t(\mathbf{w})$ is computed by the full recurrent predictor, so ξ_t accounts for the influence of the input at time t on all affected outputs. No independence of frames is needed for this deterministic comparison.

Proof of Proposition 1. Fix the bank, the downstream predictor, and its initial hidden state. For the fixed input sequence $\mathbf{z}$ and enhancement displacement $\boldsymbol{\delta}$, define $\varphi(u) = \mathcal{J}_V(\mathbf{z} + u\boldsymbol{\delta})$ for $u \in [0, 1]$. Applying the fundamental theorem of calculus gives the following bound:

$$\begin{aligned}
&\mathcal{J}_V(\mathbf{z} + \boldsymbol{\delta}) - \mathcal{J}_V(\mathbf{z}) - \langle \nabla \mathcal{J}_V(\mathbf{z}), \boldsymbol{\delta} \rangle \\
&= \int_0^1 \langle \nabla \mathcal{J}_V(\mathbf{z} + u\boldsymbol{\delta}) - \nabla \mathcal{J}_V(\mathbf{z}), \boldsymbol{\delta} \rangle \, du \\
&\leq \int_0^1 \|\nabla \mathcal{J}_V(\mathbf{z} + u\boldsymbol{\delta}) - \nabla \mathcal{J}_V(\mathbf{z})\|_2 \|\boldsymbol{\delta}\|_2 \, du \\
&\leq \int_0^1 L_J u \|\boldsymbol{\delta}\|_2^2 \, du = \frac{L_J}{2} \|\boldsymbol{\delta}\|_2^2.
\end{aligned} \tag{29}$$

The two inequalities use Cauchy–Schwarz and the L_J-Lipschitz gradient assumption, respectively. Since $\langle \nabla \mathcal{J}_V(\mathbf{z}), \boldsymbol{\delta} \rangle = -\mathcal{G}_V$, Equation 8 follows. Its right-hand side is strictly negative when $\mathcal{G}_V > (L_J/2)\|\boldsymbol{\delta}\|_2^2$, proving the strict loss reduction.

D.2 PROOF OF PROPOSITION 2: COVERAGE-BASED SEPARATION

Fix the encoder, category rule, positive-definite metrics, and finite nonempty category banks. Recall that $\beta_D(z) < \tau$ is equivalent to $d_D(z) > r_\tau$. The known-scene mixture distribution and its category-wise moments can be written as follows:

$$P_Z = \sum_{c=1}^{M} \pi_c P_c, \qquad m_c = \mathbb{E}[Z \mid C = c], \qquad v_c = \mathbb{E}[\|Z - m_c\|_{Q_c}^2 \mid C = c]. \tag{30}$$

Proof of Proposition 2. Choose $u_c \in D_c$ achieving a_c. For a known scene in category c, the triangle inequality gives the following bound:

$$d_D(Z) \le \|Z - u_c\|_{Q_c} \le \|Z - m_c\|_{Q_c} + a_c. \tag{31}$$

Since $r_\tau > a_c$, Markov's inequality bounds the rejection probability as follows:

$$\begin{aligned}\Pr(\beta_D(Z) < \tau \mid C = c) &\le \Pr\big(\|Z - m_c\|_{Q_c}^2 > (r_\tau - a_c)^2 \mid C = c\big) \\ &\le \frac{v_c}{(r_\tau - a_c)^2}.\end{aligned}$$

Capping the bound at one proves Equation 11.

For every representative $u \in D_c$, the reverse triangle inequality gives the following bound:

$$\|z - u\|_{Q_c} \ge \|z - m_c\|_{Q_c} - \|u - m_c\|_{Q_c} \ge d_{\text{mean}}(z) - R_D. \tag{32}$$

Taking the minimum over all representatives proves $d_D(z) \ge d_{\text{mean}}(z) - R_D$. Hence $d_{\text{mean}}(z) > r_\tau + R_D$ implies rejection. Under Equation 12, a valid distance threshold and the corresponding fitness threshold can be chosen as follows:

$$\max_c(a_c + \sqrt{v_c/\eta}) < r_\tau < \Delta_0 - R_D, \qquad \tau = e^{-kr_\tau}. \tag{33}$$

The left inequality makes the rejection probability at most η in every known category. For every $z \in \mathcal{O}$, the right inequality yields the following distance bound:

$$d_D(z) \ge d_{\text{mean}}(z) - R_D \ge \Delta_0 - R_D > r_\tau,$$

so all of $\mathcal{O}$ is rejected. Since $r_\tau > 0$, the corresponding $\tau = e^{-kr_\tau}$ belongs to $(0, 1)$. This also covers categories with $v_c = 0$.

From representative selection to coverage. Let $\widehat{\mu}_c$ be the empirical centroid and u_c^{near} the selected near-centroid representative. The triangle inequality relates representative selection to central coverage as follows:

$$a_c \le \|u_c^{\text{near}} - m_c\|_{Q_c} \le \|u_c^{\text{near}} - \widehat{\mu}_c\|_{Q_c} + \|\widehat{\mu}_c - m_c\|_{Q_c}. \tag{34}$$

Thus, central coverage depends on both representative placement and centroid estimation accuracy.

D.3 Enhancement and Temporal Stability

We compare banks D and $\widetilde{D}$ with the same scene categories, fixed metrics Q_c, adapter A, GRU, readout, and initial hidden state. On a common feature domain, assume A is L_A-Lipschitz, both retrieval maps are L_r-Lipschitz, and both unscaled fused features have norm at most B_a. The geometric and retrieval discrepancies between the banks can be defined as follows:

$$\epsilon_D = \max_c d_H(D_c, \widetilde{D}_c), \qquad \epsilon_r = \sup_z \|r_D(z) - r_{\widetilde{D}}(z)\|_2, \tag{35}$$

where d_H is the Euclidean Hausdorff distance, defined as follows:

$$d_H(S, T) = \max\left\{\max_{u\in S}\min_{v\in T}\|u - v\|_2, \max_{v\in T}\min_{u\in S}\|v - u\|_2\right\}.$$

The geometric discrepancy ϵ_D controls bank distances, while ϵ_r accounts for changes in the retrieved feature, including changes in retrieval weights. The constants used in the stability bound are defined as follows:

$$L_E = 1 + L_A L_r + \frac{kB_a}{\sqrt{s}}, \qquad b_D = L_A\epsilon_r + \frac{kB_a\epsilon_D}{\sqrt{s}}. \tag{36}$$

Assume the probability readout $\sigma \circ f_{\text{out}}$ is L_p-Lipschitz and the GRU satisfies the following Lipschitz condition:

$$\|H(e, h) - H(e', h')\|_2 \le L_x\|e - e'\|_2 + L_h\|h - h'\|_2. \tag{37}$$

Proposition 3 (Enhancement and temporal stability)**.** *Under these assumptions, for scene-feature discrepancies $\rho_j = \|z_j - \widetilde{z}_j\|_2$ and identical initial hidden states, the probability discrepancy is bounded as follows:*

$$|p_t - \widetilde{p}_t| \le L_p L_x \sum_{j=1}^{t} L_h^{t-j}(L_E\rho_j + b_D). \tag{38}$$

Proof. Distance and fitness. The reverse triangle inequality and $\Sigma_c \succeq sI$ give $|\|z-u\|_{Q_c} - \|z'-u\|_{Q_c}| \le \|z-z'\|_2/\sqrt{s}$. Taking minima preserves this Lipschitz bound. Since $x \mapsto e^{-kx}$ is k-Lipschitz on $[0,\infty)$, the distance and fitness bounds can be written as follows:

$$|d_D(z) - d_D(z')| \le \frac{\|z-z'\|_2}{\sqrt{s}}, \qquad |\beta_D(z) - \beta_D(z')| \le \frac{k\|z-z'\|_2}{\sqrt{s}}. \tag{39}$$

The same argument, matching representatives within each category in both directions, yields the following bounds for bank perturbations:

$$|d_D(z) - d_{\widetilde{D}}(z)| \le \frac{\epsilon_D}{\sqrt{s}}, \qquad |\beta_D(z) - \beta_{\widetilde{D}}(z)| \le \frac{k\epsilon_D}{\sqrt{s}}. \tag{40}$$

Enhanced features. The map a_D is $(1 + L_A L_r)$-Lipschitz. Decomposing the difference of the products $E_D = \beta_D a_D$, and using $0 < \beta_D \le 1$ and the amplitude bound, gives the following feature stability bounds:

$$\|E_D(z) - E_D(z')\|_2 \le L_E\|z-z'\|_2, \tag{41}$$

$$\|E_D(z) - E_{\widetilde{D}}(z)\|_2 \le b_D. \tag{42}$$

A triangle inequality therefore gives $\|E_D(z_j) - E_{\widetilde{D}}(\widetilde{z}_j)\|_2 \le L_E\rho_j + b_D$.

Recurrent propagation. For $\epsilon_t^h = \|h_t - \widetilde{h}_t\|_2$, Equation 37 yields the following recurrence:

$$\epsilon_t^h \le L_x(L_E\rho_t + b_D) + L_h\epsilon_{t-1}^h, \qquad \epsilon_0^h = 0. \tag{43}$$

Unrolling the recurrence and applying the L_p-Lipschitz probability readout proves Equation 38.

Constants for the specified modules. For the linear–softmax retrieval and ReLU adapter, the Lipschitz constants can be bounded as follows:

$$L_r \le \tfrac{1}{2}\|\mathbf{D}\|_2\|W_g\|_2, \qquad L_A \le \|W_A\|_2, \tag{44}$$

with a common retrieval bound chosen for both banks. The softmax Jacobian has spectral norm at most $1/2$, since its absolute row sums are $2\alpha_i(1-\alpha_i) \le 1/2$ and it is symmetric. If $\|z\|_2 \le B_z$ and representative norms in both banks are bounded by B_u, convex averaging gives $B_a = B_z + L_A B_u + \|A(0)\|_2$ as a sufficient amplitude bound. If f_{out} is L_o-Lipschitz, then $L_p \le L_o/4$.

The finite-horizon bound requires no contraction assumption. If additionally $L_h < 1$ and $\rho_j \le \rho$, it simplifies to $L_p L_x(L_E\rho + b_D)/(1-L_h)$. The perturbations are measured in scene-feature space after detection and encoding, with the recurrent predictor and category metrics held fixed.

E Supplementary Evaluation Protocols and Results

E.1 Experimental Settings

TailBank is evaluated on DAD, CCD, and A3D. The prediction-error, micro-cluster, and expert-curated scene-subset analyses are conducted on DAD. Each video contains 100 frames at 20 fps; each frame has one global feature and 19 object slots.

We train TailBank for 30 epochs on an NVIDIA RTX 4090 GPU using Adam, a batch size of 10, and an initial learning rate of 10^{-4}. ReduceLROnPlateau monitors the training objective with a reduction factor of 0.5 and patience of five epochs. The auxiliary classification and centroid losses are weighted by 15 and 10^{-4}, respectively. The centroid loss uses a margin of 5.0 and a repulsion weight of 0.01, applied only to the repulsive term; the attractive term has unit weight.Gradients are clipped to a maximum norm of 10.

E.2 Scene-Subset Construction

Each subset contains complete videos and preserves the dataset's accident labels and onset annotations. The same membership is used for all methods compared within a subset. The evaluation subsets characterize scene frequency, reference-model difficulty, and safety-relevant attributes independently of the scene groups used to construct the training bank.

Prediction-error subsets. A fixed reference predictor consisting of a two-layer GCN, mean pooling, and a linear classifier provides a prediction-error score s_i for each evaluation video V_i. Sorting s_i in descending order yields the highest-error 5%, 10%, 20%, and 30% subsets. These nested subsets measure difficulty relative to the reference predictor.

Micro-cluster subsets. Video features are standardized, reduced using PCA, and grouped by K-means into nonempty clusters $\mathcal{K}_1, \ldots, \mathcal{K}_J$. Their frequencies are $\widehat{\pi}_j = |\mathcal{K}_j|/n_{\text{ref}}$, ordered from smallest to largest. For $q \in \{0.05, 0.10, 0.20, 0.30\}$, Equation 1 selects whole clusters until the cumulative video fraction reaches q. Retaining the boundary cluster can produce a subset larger than the nominal fraction.

Expert-curated subsets. The expert-curated collection comprises 82 DAD videos. Review considers unusual participant interactions, combinations of environmental conditions, and potentially severe event sequences. The evaluation groups are Environmental Conditions, Traffic Complexity, Dynamic Object Interactions, and Scenario Specificity. Candidate cases are reviewed independently and disagreements are resolved by consensus. Videos can receive multiple attributes, so these groups may overlap.

Frequency, reference-model error, and expert attributes describe different properties of a scene. The first characterizes representation-level rarity, the second prediction difficulty, and the third interpretable safety-relevant conditions.

E.3 Evaluation Metrics

Average Precision (AP). At an accident-score threshold θ_{acc}, a positive video is correctly anticipated if its predicted probability crosses the threshold before the annotated accident onset. A threshold crossing in a negative video constitutes a false positive. Precision and recall are computed as

$$P(\theta_{\text{acc}}) = \frac{\text{TP}(\theta_{\text{acc}})}{\text{TP}(\theta_{\text{acc}}) + \text{FP}(\theta_{\text{acc}})}, \qquad R(\theta_{\text{acc}}) = \frac{\text{TP}(\theta_{\text{acc}})}{\text{TP}(\theta_{\text{acc}}) + \text{FN}(\theta_{\text{acc}})}. \tag{45}$$

Sweeping the threshold produces a precision–recall curve. For increasing recall levels $0 = R_0 < R_1 < \cdots < R_K$, AP is computed as

$$\text{AP} = \sum_{k=1}^{K} (R_k - R_{k-1}) P_k, \tag{46}$$

where P_k is the precision associated with recall level R_k.

Mean Time-to-Accident (mTTA). For a positive video v with accident-onset frame t_v^*, the first pre-accident warning frame at threshold θ_{acc} is

$$t_{0,v}(\theta_{\text{acc}}) = \min\{t < t_v^* : p_{t,v} \geq \theta_{\text{acc}}\}, \tag{47}$$

when this set is nonempty. Let $\mathcal{I}(\theta_{\text{acc}})$ denote the set of successfully anticipated positive videos. Their mean warning lead time is

$$\overline{\text{TTA}}(\theta_{\text{acc}}) = \frac{1}{|\mathcal{I}(\theta_{\text{acc}})|} \sum_{v \in \mathcal{I}(\theta_{\text{acc}})} \frac{t_v^* - t_{0,v}(\theta_{\text{acc}})}{f_{\text{fps}}}. \tag{48}$$

Let $\mathcal{R}_{\text{eval}}$ be the set of evaluated positive recall levels and $\overline{\text{TTA}}(r)$ the corresponding mean warning lead time. We compute

$$\text{mTTA} = \frac{1}{|\mathcal{R}_{\text{eval}}|} \sum_{r \in \mathcal{R}_{\text{eval}}} \overline{\text{TTA}}(r). \tag{49}$$

A larger mTTA indicates earlier anticipation across operating points.

TTA at 80% Recall (TTA@R80). TTA@R80 measures warning timeliness at a common detection coverage. When 80% recall is available, it is defined as

$$\text{TTA@R80} = \overline{\text{TTA}}(0.8). \tag{50}$$

For discrete operating points, the nearest evaluated recall level can be used:

$$r_{80} \in \arg\min_{r \in \mathcal{R}_{\text{eval}}} |r - 0.8|, \qquad \text{TTA@R80} = \overline{\text{TTA}}(r_{80}). \tag{51}$$

E.4 Limited Training and Missing Observations

We evaluate TailBank under reduced training data and incomplete observations to examine its performance when fewer labeled videos or fewer observed frames are available. Limited-training settings retain 75% or 50% of the available training videos. Table 5 reports results on DAD, CCD, and A3D.

Reduced training data. TailBank maintains strong performance on DAD when the training set is reduced. With 75% and 50% of the training videos, it achieves AP values of 95.01% and 94.57%, respectively, exceeding the strongest baseline, CRASH, by 30.91 and 33.16 percentage points. Its corresponding mTTA values of 4.33 and 4.31 s also exceed CRASH by 1.57 and 1.69 s. Reducing the training fraction from 75% to 50% decreases TailBank's AP by only 0.44 percentage points and mTTA by 0.02 s in these runs.

On CCD, TailBank maintains 99.11% AP at both training fractions. At 50% training data, it achieves the highest AP, exceeding CRASH by 0.92 percentage points, although CRASH provides a longer mTTA of 4.71 s compared with 4.56 s. On A3D, TailBank achieves the highest AP and mTTA with 75% training data, reaching 96.69% and 4.78 s. With 50% training data, it retains the highest mTTA of 4.92 s, while its AP decreases to 93.08%, below CRASH's 94.98%. These results highlight TailBank's ability to maintain strong anticipation performance with fewer training videos, achieving the highest AP on DAD and CCD and the longest mTTA on DAD and A3D in the 50% training setting.

Random frame removal. On DAD, TailBank achieves the highest AP at all three removal rates, obtaining 97.10%, 91.70%, and 98.30% AP with 10%, 20%, and 50% frame removal, respectively. These results exceed CRASH by 31.86, 27.06, and 34.96 percentage points. TailBank also provides the longest mTTA at 10% and 20% removal, reaching 4.03 and 3.51 s. At 50% removal, however, its mTTA falls to 2.04 s, below DSTA's 2.11 s and CRASH's 2.37 s.

A similar distinction between accuracy and timeliness appears on CCD and A3D. On CCD, TailBank matches the best AP at 10% removal and achieves the highest AP at 20% and 50% removal, with scores of 99.40% and 99.60%. Its mTTA nevertheless decreases from 4.34 s at 10% removal to 2.31 s at 50% removal. On A3D, TailBank leads in AP at 20% and 50% removal, reaching 97.80% and 95.30%, but its mTTA remains below the compared baselines at all three removal rates. These results show that strong accident discrimination can coexist with shorter warning lead times when observations are heavily reduced.

Periodic frame removal. When one frame in every five is removed, TailBank achieves the highest AP and mTTA on DAD, with 92.80% and 3.50 s. Its AP exceeds CRASH by 28.42 percentage points, while its mTTA exceeds the strongest timing baseline, DSTA, by 0.91 s. On CCD, TailBank obtains the highest AP of 99.50%, but its mTTA of 3.62 s is lower than those of the compared baselines. On A3D, it obtains 93.00% AP and 3.91 s mTTA, below CRASH's 94.88% and 4.74 s.

Overall, TailBank maintains strong anticipation performance under reduced training data and missing observations. It achieves the highest AP across all evaluated settings on DAD, leads or ties for the highest AP under every missing-observation setting on CCD, and achieves the highest AP at 20% and 50% random frame removal on A3D. These results demonstrate its effectiveness across a range of data-limited and incomplete-observation settings.

E.5 Module Ablations and Fusion Interventions

Table 3(a) retrains variants without HKE, fitness modulation, or OAM. Panel (b) fixes the full model parameters and changes only the feature supplied to the recurrent predictor. With $v_t = z_t$ and $u_t = A(r_D(z_t))$, the inputs are

$$e_t^{\text{variant}} = \begin{cases} v_t, & \text{Visual only,} \\ v_t + u_t, & \text{Additive,} \\ v_t + \beta_t u_t, & \text{K-gated,} \\ \beta_t v_t, & \text{V-gated only,} \\ \beta_t (v_t + u_t), & \text{Full.} \end{cases} \tag{52}$$

Table 5: Robustness under reduced training data and missing observations on DAD, CCD, and A3D. Entries report AP (%) / mTTA; higher is better for both metrics. Train $x\%$ uses $x\%$ of the available training videos. Drop $x\%$ randomly removes $x\%$ of frames, whereas 1 in 5 removes one frame in every five. Bold indicates the best value for each metric within each dataset and setting, including ties. Underlining marks the second-best distinct value. The purple-shaded row highlights TailBank.

Dataset	Model	Reduced training data		Missing observations			
		Train 75%	Train 50%	Drop 10%	Drop 20%	Drop 50%	1 in 5
DAD	UString	51.15 / 2.67	52.68 / 2.38	53.51 / 2.50	52.27 / 2.47	52.37 / 1.62	52.62 / 2.14
	DSTA	52.79 / 2.71	52.78 / 2.52	56.78 / 2.48	55.89 / 2.48	54.84 / 2.11	55.46 / 2.59
	CRASH	64.10 / 2.76	61.41 / 2.62	65.24 / 2.84	64.64 / 2.76	63.34 / **2.37**	64.38 / 2.51
	TailBank (Ours)	**95.01 / 4.33**	**94.57 / 4.31**	**97.10 / 4.03**	**91.70 / 3.51**	**98.30** / 2.04	**92.80 / 3.50**
CCD	UString	98.55 / 4.77	97.05 / 4.27	98.71 / 4.73	96.44 / 4.36	94.52 / 4.39	96.79 / 4.44
	DSTA	98.69 / 4.58	97.21 / 4.49	98.80 / 4.79	97.18 / 4.51	94.73 / 4.01	97.95 / 4.53
	CRASH	**99.17 / 4.87**	98.19 / **4.71**	**99.30 / 4.89**	98.93 / **4.69**	98.46 / **4.53**	98.91 / **4.76**
	TailBank (Ours)	99.11 / 4.73	**99.11** / 4.56	**99.30** / 4.34	**99.40** / 3.54	**99.60** / 2.31	**99.50** / 3.62
A3D	UString	94.28 / 4.54	93.30 / 4.58	94.01 / 4.74	93.11 / 4.59	91.26 / 3.66	93.48 / 4.34
	DSTA	92.86 / 4.58	91.75 / 4.11	93.77 / 4.82	92.31 / **4.82**	91.80 / 3.75	93.54 / 4.57
	CRASH	94.92 / 4.73	**94.98** / 4.65	**95.96 / 4.88**	94.83 / 4.77	94.54 / **4.24**	**94.88 / 4.74**
	TailBank (Ours)	**96.69 / 4.78**	93.08 / **4.92**	93.50 / 4.41	**97.80** / 4.31	**95.30** / 2.28	93.00 / 3.91

All five variants use the same GRU and readout parameters. These interventions assess how retrieval and attenuation affect a fixed model, matching the comparison underlying Proposition 1.

E.6 Additional Qualitative Examples

Figure 6 provides additional examples for examining the model's temporal predictions under varied traffic and visibility conditions. All four sequences use the same displayed decision threshold of 0.5, enabling a consistent comparison of the predicted accident scores.

In Figure 6(a), the scene contains multiple road users and reflections from a wet road surface. The scores remain low throughout the sequence, with only small fluctuations below the threshold. In Figure 6(b), a vehicle occupies a substantial portion of the foreground against a brightly lit underpass exit. Despite the nearby vehicle and strong illumination contrast, the model again maintains below-threshold predictions.

Figure 6(c) exhibits a different temporal pattern: the score rises within the early portion of the sequence and then remains near or above the threshold for much of the remaining interval. In Figure 6(d), the score initially fluctuates before showing a more sustained response near or above the threshold during nighttime driving.

Taken together, these observations highlight the model's ability to maintain low scores in visually complex scenes while producing sustained elevated responses in other traffic situations. This differentiated behavior is consistent with selective risk assessment and temporal persistence, providing qualitative support for the model's capacity to produce context-sensitive predictions across varied driving conditions.

E.7 Illustrative Vision-Language Feedback

The optional VLM interface supplements TailBank's temporal predictions with natural-language feedback. Its inputs comprise the observed video prefix, the corresponding frame-level accident probabilities, identified traffic participants, and a structured prompt requesting a scene description, a risk explanation, and a suggested response. Figure 5 shows concise illustrative advisories for an accident sequence and a non-accident sequence.

Accident example. Figure 5(a) shows a daytime urban scene with a bus close to the ego vehicle's right side and other vehicles ahead. The predicted accident probability crosses the displayed warning threshold before the marked accident interval. The accompanying advisory summarizes the nearby traffic and suggests a precautionary response:



Figure 6: Additional qualitative results under varied traffic and visibility conditions: (a) a wet urban intersection; (b) a shaded underpass with a vehicle ahead and a brightly lit exit; (c) a daytime road scene with multiple vehicles; and (d) a nighttime urban road with vehicles and scooters. Each row aligns selected frames with object detections and frame-wise accident scores over a 100-frame sequence at 20 FPS. Blue bars denote predicted accident probabilities, and the red dashed line marks a common decision threshold of 0.5. Scores remain below the threshold in (a–b), whereas (c–d) exhibit sustained elevated responses over substantial portions of the sequences.

> Caution! A bus is close on your right, with vehicles ahead. Slow down and leave more space.

This feedback relates the visible traffic configuration to a practical precaution without attributing the accident to a specific participant.

Non-accident example. Figure 5(b) shows a nighttime scene with limited visibility and vehicles ahead. The predicted accident probability remains below the displayed threshold throughout the sequence. The accompanying advisory nevertheless communicates a scene-specific precaution:

> Low visibility ahead. Keep a safe distance from the car and watch nearby vehicles.

This example distinguishes general precautionary guidance from a threshold-triggered accident warning.

These examples demonstrate how LLM feedback complements accident predictions with scene-specific descriptions and precautionary suggestions. In the accident example, the feedback highlights nearby traffic and recommends slowing down and leaving more space. In the non-accident example, it emphasizes limited visibility and a safe following distance, providing useful context even when the predicted score remains below the warning threshold.