

Predict and Resist: Long-Term Accident Anticipation Under Sensor Noise

Xingcheng Liu^{*1}, Bin Rao^{*1}, Yanchen Guan¹, Chengyue Wang¹, Haicheng Liao¹,
Jiaxun Zhang¹, Chengyu Lin², Meixin Zhu³, Zhenning Li^{1†}

¹University of Macau

²Zhejiang University

³Southeast University

{yc57431,yc57416,yc37976}@um.edu.mo, chengyue.wang@connect.um.edu.mo, {yc27979,yc47415}@um.edu.mo,
linchengyu@zju.edu.cn, meixin@seu.edu.cn, zhenningli@um.edu.mo

Abstract

Accident anticipation is essential for proactive and safe autonomous driving, where even a brief advance warning can enable critical evasive actions. However, two key challenges hinder real-world deployment: (1) noisy or degraded sensory inputs from weather, motion blur, or hardware limitations, and (2) the need to issue timely yet reliable predictions that balance early alerts with false-alarm suppression. We propose a unified framework that integrates diffusion-based denoising with a time-aware actor-critic model to address these challenges. The diffusion module reconstructs noise-resilient image and object features through iterative refinement, preserving critical motion and interaction cues under sensor degradation. In parallel, the actor-critic architecture leverages long-horizon temporal reasoning and time-weighted rewards to determine the optimal moment to raise an alert, aligning early detection with reliability. Experiments on three benchmark datasets (DAD, CCD, A3D) demonstrate state-of-the-art accuracy and significant gains in mean time-to-accident, while maintaining robust performance under Gaussian and impulse noise. Qualitative analyses further show that our model produces earlier, more stable, and human-aligned predictions in both routine and highly complex traffic scenarios, highlighting its potential for real-world, safety-critical deployment.

Introduction

Traffic accident anticipation—the ability to predict collisions before they occur—represents a critical capability for autonomous driving (Zhang et al. 2025). Unlike traditional perception systems that merely detect accidents after they happen, anticipation enables proactive safety interventions, such as timely braking or evasive maneuvers, potentially preventing collisions entirely (Abdel-Aty and Ding 2024). A vehicle that can foresee danger seconds in advance transforms safety from reactive to preventive, which is the ultimate goal of intelligent transportation systems (Fang et al. 2024; Ahmed et al. 2020).

However, achieving reliable accident anticipation in real-world driving is profoundly challenging (Ali, Hussain, and

^{*}These authors contributed equally.

[†]Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

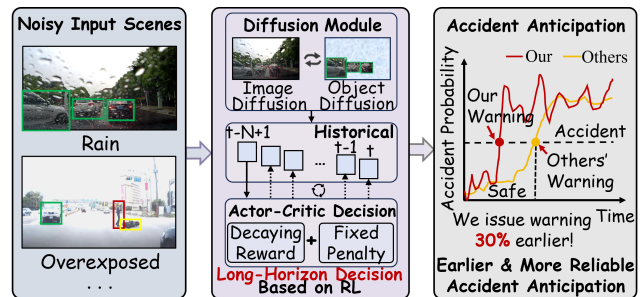


Figure 1: Overview illustration of our framework. The figure highlights the integration of the diffusion module and reinforcement learning for processing noisy input scenes, leading to earlier and more reliable accident anticipation.

Haque 2024; Liao et al. 2024b) due to two interdependent obstacles:

1) **Robustness under imperfect perception.** Autonomous vehicles operate with imperfect sensing: rain, glare, dirt, lens damage, and motion blur can obscure critical cues (visual examples are shown in **Appendix B**). In such noisy conditions, short-term predictions based on single frames become unreliable. Ironically, these are precisely the situations that demand longer temporal reasoning: by accumulating weak signals across time, a model can extract meaningful patterns even when individual frames are corrupted (Croitoru et al. 2023).

2) **The problem of “when to warn”.** Most existing approaches focus on frame-level classification or short-horizon prediction (Zeng et al. 2017): they can indicate if an accident might occur but rarely optimize when to issue an alert (Meulemans et al. 2023; Pignatelli et al. 2024). In safety-critical scenarios, timing is as important as correctness—alerts issued too late are useless, while those issued too early or too often erode trust and can even induce unsafe reactions. This is fundamentally a long-horizon credit assignment problem: the model must identify subtle early cues, maintain temporal reasoning, and determine the optimal moment to act.

Crucially, these challenges amplify each other. As shown in Fig.1, sensor degradation increases uncertainty in imme-

diate observations, which in turn magnifies the need for long-horizon temporal reasoning to stabilize predictions. Conversely, a model without effective temporal credit assignment cannot leverage redundancy across frames to overcome noisy perception. A truly deployable anticipation system must therefore treat timing and robustness as a single coupled problem.

To address these challenges, we re-frame accident anticipation as a sequential decision-making problem under uncertainty. Instead of simply classifying frames, our model learns when to warn for maximum safety utility, leveraging an actor-critic reinforcement learning framework for long-horizon credit assignment (Sutton and Barto 2018). To ensure robustness under realistic conditions, we introduce a dual-level diffusion-based denoising module that reconstructs noise-resilient features at both the image and object levels, allowing the system to preserve essential temporal cues even in degraded visual conditions (Ho, Jain, and Abbeel 2020; Song et al. 2021). Our dual-level diffusion module acts as a probabilistic feature stabilizer, conceptually similar to Bayesian evidence accumulation. By iteratively refining noisy inputs into structurally faithful and temporally coherent representations, it reduces jitter and spurious activation. This allows the actor-critic module to observe a smoother evolution of risk cues, preventing credit dilution from noisy frames and enabling more effective long-horizon reward assignment. This unified design enables early, reliable, and noise-resilient accident anticipation, bridging the gap between algorithmic capability and real-world deployment.

In summary, the contributions of this paper are as follows:

- We formulate accident anticipation as a long-horizon credit assignment problem, optimizing not only prediction correctness but also the optimal timing of alerts through an actor-critic framework.
- We design image-level and object-level diffusion modules to reconstruct robust features under sensor noise, enabling the model to retain critical temporal cues and sustain performance in real-world noisy conditions.
- Across three benchmark datasets and their noise-augmented variants, our framework achieves state-of-the-art performance in both Average Precision (AP) and mean Time-to-Accident (mTTA), showing that joint long-horizon reasoning and noise-aware feature enhancement yields earlier and more stable warnings than conventional frame-level approaches.

Related Work

Accident anticipation has become a core research problem in autonomous driving because foresight is essential for preventing collisions rather than merely reacting to them. Early studies relied on rule-based heuristics and statistical models, which captured simple patterns but failed to generalize to the complex, multi-agent interactions of real-world traffic (Grant et al. 2018).

The introduction of deep learning has driven a transition to vision-based anticipation, leveraging dashcam or onboard camera video as a rich source of spatial and temporal risk

cues (Li et al. 2023). The evolution of methods reflects a progression in how the community has tried to model risk:

- **From frame-level perception to temporal reasoning.** Initial deep models relied on CNNs to extract scene appearance and detect static risk cues (Chan et al. 2017; Fang et al. 2022). While they improved over hand-crafted features, these approaches were myopic, often missing early signals of events that develop gradually. To address this, sequential architectures such as RNNs, LSTMs, and GRUs were introduced (Fatima, Karim Khan, and Kyung 2021; Takimoto et al. 2019; Xue, Chen, and Fang 2020), enabling the capture of risk evolution over time.
- **From isolated objects to interaction-aware modeling.** Anticipating accidents requires understanding how vehicles, pedestrians, and cyclists interact. Graph Neural Networks (GNNs) explicitly encode multi-agent relationships (Karim, Yin, and Qin 2024; Liu et al. 2020; Thakur, Gouripeddi, and Li 2024; Yao et al. 2019), while transformer-based models exploit global attention to capture long-range dependencies and subtle interaction cues (Feng, Hong, and Zheng 2021; Wu et al. 2021). These approaches move beyond simply seeing the scene to reasoning about its dynamics.
- **Addressing rarity and interoperability.** True accident events are rare, creating long-tail data challenges. Generative models such as GANs and VAEs synthesize plausible traffic sequences to augment scarce critical scenarios (Bao, Yu, and Kong 2021; Yao et al. 2023). In parallel, attention mechanisms (Karim et al. 2022; Karim, Yin, and Qin 2024; Song et al. 2024) and semantic parsing (Liao et al. 2024c) improve interpretability by focusing on salient agents and regions, allowing models to highlight the cues most indicative of future risk.

Despite this progress, existing methods remain limited in two critical aspects for real-world deployment. First, they assume clean visual input, yet real-world sensors frequently face rain, glare, blur, or missing pixels, which can obscure subtle pre-accident cues and destabilize short-term predictions. Second, most models optimize for accident classification rather than the timing of warnings, leaving systems prone to delayed alerts or excessive false positives, both problematic in safety-critical applications.

These gaps highlight the need for methods that combine noise-resilient perception with long-horizon temporal reasoning, enabling accident anticipation that is both early and reliable under real-world conditions.

Methodology

Problem Formulation

We frame *traffic accident anticipation* as a sequential risk forecasting problem, where the goal is to estimate accident probabilities over time and issue an *early warning* for timely intervention.

Let a video $V = \{V_t\}_{t=1}^T$ consist of frames V_t at time step t . A learnable function f_θ predicts frame-wise accident probabilities $\mathbf{P} = \{p_t\}_{t=1}^T$ as:

$$p_t = f_\theta(V_{1:t}), \quad \text{for } t = 1, \dots, T \quad (1)$$

where $V_{1:t}$ is the frame sequence up to time t and f_θ is parameterized by θ .

To evaluate temporal performance, we define *Time-to-Accident (TTA)* as the interval between the model’s first confident prediction and the ground-truth accident frame:

$$\Delta t = \tau - t_o \quad \text{where} \quad t_o = \min\{t \in \{1, \dots, T\} \mid p_t \geq p_{th}\} \quad (2)$$

where τ is the accident frame index (or 0 for negative sequences) and p_{th} is a decision threshold. A sequence is classified as *accident-positive* if $p_t \geq p_{th}$ and $\tau > 0$; otherwise, it is *accident-negative*. The model is trained to *optimize* both accuracy and anticipation: discriminating accident-positive/negative sequences while maximizing Δt for early warnings with minimal false alarms.

Model Framework

The overall processing pipeline of our framework is illustrated in Fig. 2. The model consists of five core components: an object detector, a feature extractor, a self-adaptive object-aware module, dual diffusion modules for image- and object-level denoising, and an actor-critic decision module for long-horizon anticipation.

Given an input video sequence, the object detector and feature extractor generate global image features \mathbf{F}_{img} and object-level vectors, refined by the self-adaptive object-aware module to capture dynamic interactions and produce enhanced spatio-temporal representations $\bar{\mathbf{F}}_{obj}$. Both feature sets are denoised via diffusion modules, yielding robust, noise-resilient representations.

The denoised features are fused and processed through a GRU to capture temporal dependencies, generating accident probabilities p_t . A time-weighted layer prioritizes critical time segments, while an actor-critic module aggregates long-horizon dependencies, issuing reliable early warnings with minimal false positives. By combining noise-resistant feature extraction with reinforcement-driven temporal reasoning, the framework enables robust accident anticipation under noisy conditions. The algorithm is in the **Appendix A**. Detailed module descriptions are in subsequent subsections.

Object Detector Each frame is processed by Cascade R-CNN (Cai and Vasconcelos 2018), and top- K dynamic agents are encoded into vectors \mathbf{F}_{obj} via VGG-16 (Simonyan and Zisserman 2015), preserving appearance and spatial cues.

Feature Extractor Global features \mathbf{F}_{img} are extracted using VGG-16 and an MLP to encode scene context and enhance feature compactness for temporal reasoning.

Self-Adaptive Object-Aware Module The object-aware module refines object representations by dynamically attending to the most informative traffic participants based on temporal context and inter-object interactions.

Given the object features \mathbf{F}_{obj} and the previous hidden state \mathbf{h}_{t-1} , attention energies are first computed as:

$$\mathbf{e}_t = \tanh(\mathbf{W}_{wa}\mathbf{h}_{t-1} + \mathbf{W}_{ua}\mathbf{F}_{obj} + \mathbf{b}_a), \quad (3)$$

where \mathbf{W}_{ua} and \mathbf{W}_{wa} are learnable projection matrices, and \mathbf{b}_a is a bias vector. These intermediate energies are further

transformed as $\mathbf{e}'_t = \mathbf{W}_w \mathbf{e}_t$ and normalized using the softmax function to yield attention weights: $\alpha_t = \text{softmax}(\mathbf{e}'_t)$.

The refined object-aware features $\bar{\mathbf{F}}_{obj}$ are obtained by applying the attention weights to the original object features via element-wise multiplication:

$$\bar{\mathbf{F}}_{obj} = \alpha_t \odot \mathbf{F}_{obj}. \quad (4)$$

This mechanism adaptively prioritizes high-risk objects while adjusting to evolving scenes and encoding critical interactions temporally to produce robust spatio-temporal representations for reliable accident anticipation.

Diffusion-Based Hierarchical Feature Enhancement

To improve robustness under noise, we use a diffusion-based module that refines image- and object-level features through noise injection and learned denoising, enabling recovery of meaningful representations from degraded inputs.

1) **Adaptive Timestep Sampling.** During training, features undergo noise perturbation at a randomly sampled diffusion step $t \sim \mathcal{U}\{0, T-1\}$, where T represents the total steps, improving model generalization and gradient diversity.

2) **Variance-Preserving Diffusion.** The forward diffusion process perturbs features via a variance-stable Markov chain, ensuring consistent noise scaling across steps.

$$\mathbf{F}_{img}^{noisy} = \sqrt{\bar{\alpha}_t} \mathbf{F}_{img} + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (5)$$

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (6)$$

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s \quad (7)$$

$$\alpha_t = 1 - \beta_t \quad (8)$$

$$\beta_t = \beta_{start} + \frac{t}{T}(\beta_{end} - \beta_{start}) \quad (9)$$

with a linear schedule from $\beta_{start} = 0.001$ to $\beta_{end} = 0.02$, gradually adding noise while preserving feature variance for smooth transitions between clean and noisy states.

3) **Denoising Network Architecture.** The denoising network p_θ refines the noisy image features \mathbf{F}_{img}^{noisy} at each diffusion step t via a lightweight feedforward transformation:

$$p_\theta(\mathbf{F}_{img}^{noisy}, t) = W_2 \left(\text{ReLU}(W_1 \mathbf{F}_{img}^{noisy} + b_1) \right) + b_2 \quad (10)$$

where $W_1, W_2 \in \mathbb{R}^{d \times d}$ and $b_1, b_2 \in \mathbb{R}^d$ are learnable parameters. This two-layer structure uses ReLU for non-linearity and preserves input dimensionality, enabling stable feature recovery. Optional step embeddings can integrate timestep information, improving feature reconstruction across varying noise levels.

4) **Feature Fusion Strategy.** To ensure semantic fidelity under noise, we fuse original and denoised features via residual fusion. The enhanced image feature is computed as

$$\mathbf{F}_{img}^{enhanced} = \mathbf{F}_{img} + \lambda \cdot p_\theta(\mathbf{F}_{img}^{noisy}, t), \quad \lambda = 0.15 \quad (11)$$

where \mathbf{F}_{img} is the raw image feature, $p_\theta(\cdot)$ is the denoiser and λ adjusts residual correction. This design preserves semantics, prevents over-amplification of unstable updates, and ensures stable gradient flow via the identity path.

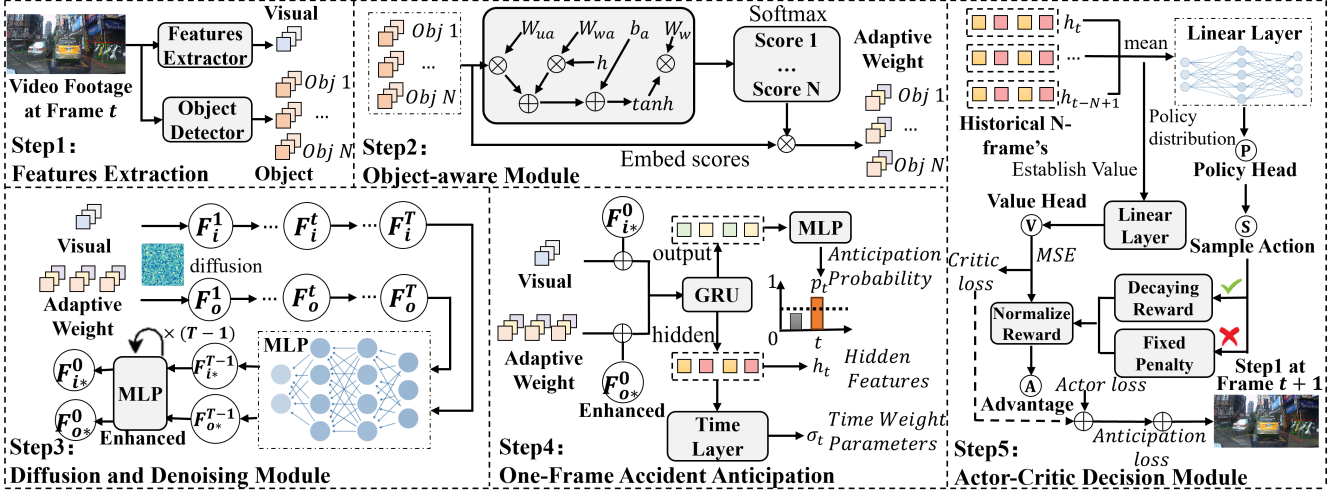


Figure 2: Overview of the proposed framework. Input frames are encoded into image and object features, refined by object-aware and diffusion modules. Fused features are processed by a GRU with time-weighted layers to predict p_t , while an actor-critic module optimizes long-horizon early warnings.

The same residual enhancement is applied to the refined object-aware features $\bar{\mathbf{F}}_{obj}$:

$$\mathbf{F}_{obj}^{enhanced} = \bar{\mathbf{F}}_{obj} + \lambda \cdot p_{\theta}(\bar{\mathbf{F}}_{obj}^{noisy}, t) \quad (12)$$

ensuring consistent robustness for both global and local modalities under gradual or abrupt input degradations.

The enhanced image and object features are concatenated and passed through a GRU to capture sequential dependencies, producing the fused representation \mathbf{X}_t and hidden state \mathbf{h}_t for frame t :

$$\mathbf{X}_t, \mathbf{h}_t = \text{GRU}(\text{concat}(\mathbf{F}_{img}^{enhanced}, \mathbf{F}_{obj}^{enhanced})) \quad (13)$$

An MLP then predicts the frame-wise accident probability $p_t = \text{MLP}(\mathbf{X}_t)$, and a time-weight layer computes the temporal weight loss $w_t = \text{fc}(\mathbf{h}_t)$ for the anticipation loss. This unified pipeline produces robust, temporally-aware features for accurate and timely accident anticipation.

State History Processing To capture short-term temporal dependencies, a rolling buffer stores the latest W hidden states. Let $\mathbf{h}_i \in \mathbb{R}^d$ be the hidden state at step i ; then at time t , the history \mathbf{H}_t is formed by concatenating the past W states (or all if $t < W$):

$$\mathbf{H}_t = \begin{cases} \text{concat}(\{\mathbf{h}_i\}_{i=1}^t) & \text{if } t < W \\ \text{concat}(\{\mathbf{h}_i\}_{i=t-W+1}^t) & \text{otherwise} \end{cases} \quad (14)$$

Summary vector obtained by averaging the buffered states: $\bar{\mathbf{h}}_t = \text{mean}(\mathbf{H}_t)$, which smooths fluctuations while retaining key context. This compact representation feeds into decision module, balancing temporal context and efficiency.

Policy and Value Estimation We adopt an actor-critic framework to model sequential decision-making, where both the policy (actor) and value (critic) functions are conditioned on the aggregated historical state $\bar{\mathbf{h}}_t$. The actor maps

$\bar{\mathbf{h}}_t$ to a discrete action distribution via a linear projection followed by softmax:

$$\pi_t = \text{softmax}(\mathbf{W}_p \bar{\mathbf{h}}_t + \mathbf{b}_p) \quad (15)$$

where $\mathbf{W}_p \in \mathbb{R}^{A \times d}$, $\mathbf{b}_p \in \mathbb{R}^A$, and A is the action space size. The action a_t is sampled as $a_t \sim \pi_t$, and its log-probability $\log \pi_t(a_t)$ is retained for policy gradient updates.

The critic predicts the expected cumulative reward from the same state using a linear value head:

$$V_t = \mathbf{w}_v^\top \bar{\mathbf{h}}_t + b_v \quad (16)$$

where $\mathbf{w}_v \in \mathbb{R}^d$ and $b_v \in \mathbb{R}$. Decoupling actor and critic stabilizes learning by guiding policy improvement with value-based estimation.

Reward Computation The reward function balances prediction correctness and temporal urgency, encouraging early and accurate decisions. The agent receives a positive reward for correct predictions, discounted exponentially to favor earlier actions:

$$r_t = \mathbb{I}(a_t = y_t) \cdot e^{-t/\tau} + \mathbb{I}(a_t \neq y_t) \cdot \gamma \quad (17)$$

where y_t is the ground-truth label, τ controls temporal decay, and γ is a fixed negative penalty for incorrect actions (set to 5 and -0.5 in our experiments).

To stabilize training and reduce reward variance across batches, rewards are normalized as:

$$\tilde{r}_t = \frac{r_t - \mu_r}{\sigma_r + \epsilon} \quad (18)$$

where μ_r and σ_r are the batch mean and standard deviation, and ϵ ensures numerical stability. This reward design promotes early, correct predictions while discouraging late or incorrect actions, aligning with the long-horizon, safety-critical nature of accident anticipation.

Model	Venue	DAD		CCD		A3D	
		AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow
DSA (Chan et al. 2017)	ACCV	48.1	1.34	98.7	3.08	92.3	2.95
ACRA (Zeng et al. 2017)	CVPR	51.4	3.01	98.9	3.32	-	-
AdaLEA (Suzuki et al. 2018)	CVPR	52.3	3.44	99.2	3.45	92.9	3.16
UString (Karim et al. 2022)	TIV	53.7	3.53	<u>99.5</u>	3.74	93.2	3.24
DSTA (Bao, Yu, and Kong 2020)	ACM MM	52.9	3.21	99.1	3.54	93.5	2.87
GSC (Wang et al. 2024)	TIV	58.2	2.76	99.3	3.58	94.9	2.62
AccNet (Liao et al. 2024a)	AAP	60.8	3.58	<u>99.5</u>	3.78	<u>95.1</u>	3.26
LATTE (Zhang et al. 2025)	IF	<u>89.7</u>	<u>4.49</u>	98.8	4.53	92.5	<u>4.52</u>
Ours	-	91.2	4.59	99.8	<u>4.29</u>	95.7	4.60

Table 1: Comparison of model performance in balancing mTTA and AP across three datasets. Best and second-best values are marked in **bold** and underline, respectively. “-” denotes missing data. \uparrow indicates that higher values are better.

Training Loss

The model is trained with a joint objective that integrates *anticipation loss* for supervised prediction and *actor-critic losses* for sequential decision-making, balancing accuracy and early warning capability.

Anticipation Loss The supervised anticipation loss \mathcal{L}_{an} combines positive and negative sample terms. For positives, a temporal penalty p encourages earlier predictions and is defined as $p = -\max(0, (t_{\text{accident}} - t_{\text{current}} - 1)/\text{fps})$

and the time weight ω_t is computed from the GRU hidden state h_t as $\omega_t = 1 + \sigma(h_t)$, where $\sigma(\cdot)$ is the sigmoid function. The positive loss is:

$$\mathcal{L}_{\text{pos}} = -(\omega_t \cdot \exp(p) \cdot \mathcal{L}_{\text{ce}}) \quad (19)$$

while the negative loss uses standard cross-entropy scaled by a constant c : $\mathcal{L}_{\text{neg}} = c\mathcal{L}_{\text{ce}}$.

The overall supervised anticipation loss \mathcal{L}_{an} averages over time and samples:

$$\mathcal{L}_{an} = \mathbb{E} \left[\sum_{i=1}^T (\alpha_i \cdot \mathcal{L}_{\text{pos}}(i) + (1 - \alpha_i) \cdot \mathcal{L}_{\text{neg}}(i)) \right] \quad (20)$$

Actor-Critic Losses To encourage early and reliable anticipation, we adopt an actor-critic formulation that combines policy and value learning. The *policy loss* guides the actor to favor actions with higher advantages while promoting exploration via entropy regularization:

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}[\log \pi_t(a_t) \cdot A_t] - \lambda_e \mathcal{H}(\pi_t) \quad (21)$$

where $A_t = \tilde{r}_t - V_t$ is the advantage and $\lambda_e = 0.1$ controls the entropy weight.

The *value loss* encourages the critic to accurately estimate the expected return:

$$\mathcal{L}_{\text{critic}} = \frac{1}{2}(\tilde{r}_t - V_t)^2 \quad (22)$$

with \tilde{r}_t denoting the normalized reward.

Finally, the complete training objective combines the supervised anticipation loss with the actor-critic components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{an} + \alpha(\mathcal{L}_{\text{actor}} + \beta\mathcal{L}_{\text{critic}}) \quad (23)$$

where $\alpha = \beta = 0.5$ in our implementation. This unified objective balances supervised accuracy with reinforcement-guided timing, driving the model to deliver both *early* and *reliable* accident anticipation.

Experiment

Experiment Setup

We evaluate our model on three benchmark datasets covering diverse real-world traffic accidents:

- **Dashcam Accident Dataset (DAD)** (Chan et al. 2017): 620 accident and 1,130 normal clips (5s@20fps, 100 frames), capturing various urban collisions.
- **Car Crash Dataset (CCD)** (Yao et al. 2019): 1,500 accident and 3,000 normal clips (5s@10fps, 50 frames), with detailed metadata and rich accident diversity.
- **AnAn Accident Detection (A3D) Dataset** (Bao, Yu, and Kong 2020): 1,087 accident and 114 normal clips (5s@20fps, 100 frames), complementing DAD with different urban contexts.

To evaluate robustness, we introduce Gaussian and impulse noise at varying levels to simulate sensor degradation, assessing the model’s ability to anticipate accidents accurately under realistic noise conditions.

Evaluation Metrics

We evaluate the model on two aspects: *accuracy* and *timeliness*, reflecting reliability and early-warning capability.

Accuracy. We report precision (P), recall (R), and Average Precision (AP), $AP = \int P(R) dR$, which measures detection performance across thresholds, with higher AP indicating more reliable recognition.

Timeliness. We measure early-warning performance with Time-to-Accident (TTA), interval between the first confident prediction and the actual accident. Mean TTA (mTTA) is averaged across thresholds as $mTTA = \int_0^1 TTA_a da$ where higher values indicate stronger early-warning capability.

Implementation Details

The framework is implemented in PyTorch 2.0 and trained for 30 epochs on an NVIDIA RTX 3050 (batch size 10) using Adam (initial LR 3×10^{-4}) with a ReduceLROnPlateau

σ	CCD		A3D	
	AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow
Original	99.8	4.29	95.7	4.60
0.5	99.6	4.00	94.3	4.10
1.0	99.6	4.04	95.3	4.23
5.0	99.6	4.35	95.2	3.98
10.0	98.0	3.43	92.9	3.97
20.0	91.6	3.05	92.9	3.96

Table 2: Comparison of Gaussian noise levels and their impact on mTTA and AP for CCD and A3D. ‘‘Original’’ denotes baseline performance without added noise.

Percents	CCD		A3D	
	AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow
Original	99.8	4.29	95.7	4.60
10%	99.5	4.54	95.3	4.06
20%	99.6	4.33	95.7	4.37
30%	99.2	4.22	93.1	3.60
50%	98.0	3.38	91.6	3.79

Table 3: Comparison of pulse noise levels and their effect on mTTA and AP for CCD and A3D. ‘‘Original’’ indicates baseline performance without added noise.

scheduler. Each frame includes up to 19 objects with 4096-D features from a VGG-16 backbone. A 256-unit GRU models temporal dynamics for efficient accident anticipation.

Evaluation Results

Compare with SOTA Baselines. Table 1 shows that our model outperforms all baselines in both accuracy (AP) and timeliness (mTTA). On DAD, it achieves 91.2% AP and 4.59s mTTA, reflecting both higher accuracy and earlier anticipation. On CCD and A3D, it delivers consistent AP gains (+0.3% and 0.6% than best SOTA) and larger mTTA balance improvements, confirming the advantage of long-horizon temporal reasoning for early warnings.

Robustness to Sensor Noise. Tables 2 and 3 evaluate the model under Gaussian and impulse noise, simulating real-world sensor degradations. Our approach maintains high AP and mTTA even under moderate corruption (e.g., $\sigma = 5.0$ for Gaussian noise yields 99.6% AP on CCD and 95.2% on A3D). Performance degrades gradually at extreme noise levels but remains meaningful, demonstrating the effectiveness of the diffusion-based denoising module.

Impulse noise shows a similar trend: up to 20% pixel corruption, the model preserves near-baseline performance, and even at 50%, it produces usable outputs. This resilience confirms that our framework can sustain reliable accident anticipation under adverse sensing conditions, a critical requirement for real-world deployment. Experiments on robustness of long-horizon credit assignment are in **Appendix D**.

Ablation Studies

Ablation Study for Core Components. Table 4 and 5 presents an ablation study on the CCD dataset to quantify the

Experiment	AP (%) \uparrow	mTTA (s) \uparrow
Our Full Model	99.8	4.29
w/o Object Aware Module	99.3	4.61
w/o Time Weight Layer	99.5	4.47
w/o Anticipation Loss	33.3	5.00
w/o Policy Gradient Loss	99.6	4.47
w/o Value Loss	92.8	3.03

Table 4: Ablation studies of different modules on CCD dataset. ‘‘w/o’’ denotes removal of a module.

contribution of each core component, including the image and object diffusion modules, the self-adaptive object-aware module, the time-weight layer, and the actor-critic components (anticipation, policy, and value losses).

Removing the anticipation loss causes AP to drop sharply to 33.3%, confirming its key role in aligning predictions with pre-accident cues. This also highlights that a higher mTTA alone is not necessarily better. Value loss removal significantly affects AP and mTTA, highlighting its role in stabilizing long-horizon decisions. The self-adaptive object-aware module and time-weight layer consistently enhance early-warning performance.

Impact of diffusion modules. We analyze the role of dual diffusion modules on the CCD dataset under Gaussian noise ($\sigma \in \{0.5, 1.0, 5.0, 10.0, 20.0\}$) by comparing the full model with variants removing each module or both.

Table 5 shows that under clean and mild noise ($\sigma \leq 1$), all models perform well, with the full model reaching 99.8% AP. At moderate noise ($\sigma = 5$), dual diffusion best preserves performance (99.6% AP), while module removal causes slight drops. Under severe noise ($\sigma \geq 10$), all degrade, and omitting image diffusion sometimes improves AP, suggesting over-denoising may harm heavily corrupted inputs.

Reward–Penalty Trade-offs in AP and mTTA

We analyze how the reward coefficient (τ) for early correct predictions and the penalty factor (γ) for mispredictions affect AP and mTTA on A3D, as summarized in Table 6.

Increasing the reward weight (from $\times 1$ to $\times 50$) steadily reduces AP (95.7% \rightarrow 92.7%) while providing slight mTTA gains, indicating that overemphasizing early predictions encourages premature actions and more false positives. Conversely, reducing the reward ($\times 0.1$, $\times 0.02$) yields the highest AP (up to 96.2%) but lowers mTTA, reflecting a more conservative strategy that prioritizes accuracy over anticipation. Penalty adjustments show a complementary trend. Strong penalties ($\times 10$) make the model overly cautious, achieving the highest mTTA (4.92s) but the lowest AP (91.2%), while low penalties ($\times 0.1$) promote earlier yet less reliable predictions (AP 92.1%).

Visual Analysis: Long-Horizon vs. Frame-Level Training

We compare our long-horizon model (history window=10) with the frame-level baseline (window=0) on three DAD scenarios (Figure 3). Additional visualizations, window-size

σ	Original		w/o Image Diffusion		w/o Object Diffusion		w/o All Diffusion	
	AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow	AP (%) \uparrow	mTTA (s) \uparrow
Original	99.8	4.29	99.6	4.50	99.6	4.65	99.6	4.54
0.5	99.6	4.00	99.4	4.45	99.5	4.35	99.4	4.36
1.0	99.6	4.04	99.5	4.38	99.5	4.64	99.4	4.11
5.0	99.6	4.35	99.5	4.02	99.3	4.07	99.0	4.05
10.0	98.0	3.43	98.6	3.75	98.8	3.48	98.2	3.89
20.0	91.6	3.05	92.8	3.03	91.4	3.06	91.0	3.19

Table 5: Impact of varying Gaussian noise levels on mTTA and AP under core module ablation on the CCD dataset. "Original" (column) denotes baseline without noise; "Original" (row) indicates the full model. Best AP per noise level are in bold.

Reward τ	Penalty γ	AP (%) \uparrow	mTTA (s) \uparrow
$\times 1$ (5.0)	$\times 1$ (-0.5)	95.7	4.60
$\times 10$	$\times 1$	93.6	4.77
$\times 50$	$\times 1$	92.7	4.70
$\times 0.1$	$\times 1$	96.2	4.47
$\times 0.02$	$\times 1$	95.8	4.46
$\times 1$	$\times 10$	91.2	4.92
$\times 1$	$\times 0.1$	92.1	4.71

Table 6: Effect of reward and penalty scaling on AP and mTTA on A3D. Reward emphasizes early correct alerts; penalty scales false/missed costs.

comparisons, and inference time are provided in **Appendices C-G**.

(a) **Complex multi-agent scenario (False Positive)**. In a rainy intersection with multiple interacting vehicles, both models produce false positives. However, the long-horizon model generates shorter and less frequent alarms by integrating temporal evidence from preceding frames, effectively suppressing spurious alerts in noisy environments.

(b) **Typical collision (True Positive)**. For a straightforward accident, both models succeed, but the long-horizon model predicts nearly one second earlier by aggregating past frames and capturing subtle cues—such as the motorcycle’s lateral drift without deceleration—enabling timely anticipation before entering the danger zone.

(c) **Sudden complex crash (Late Prediction)**. In sudden collisions, although neither model predicts far in advance, the long-horizon model still issues alerts slightly earlier than the frame-level baseline. Even a small lead time can be crucial for timely evasive maneuvers in real driving scenarios.

Overall, these examples show that long-horizon training enables earlier and more reliable warnings with fewer false positives, a critical advantage for safety-critical driving where even brief lead times can greatly reduce risk.

Conclusion

We proposed a unified framework for traffic accident anticipation that combines diffusion-based denoising with a time-aware actor-critic architecture. This design enhances robustness under noisy sensing and improves the timing of early warnings by leveraging long-horizon temporal reasoning.

Experiments show that our method sustains high accu-

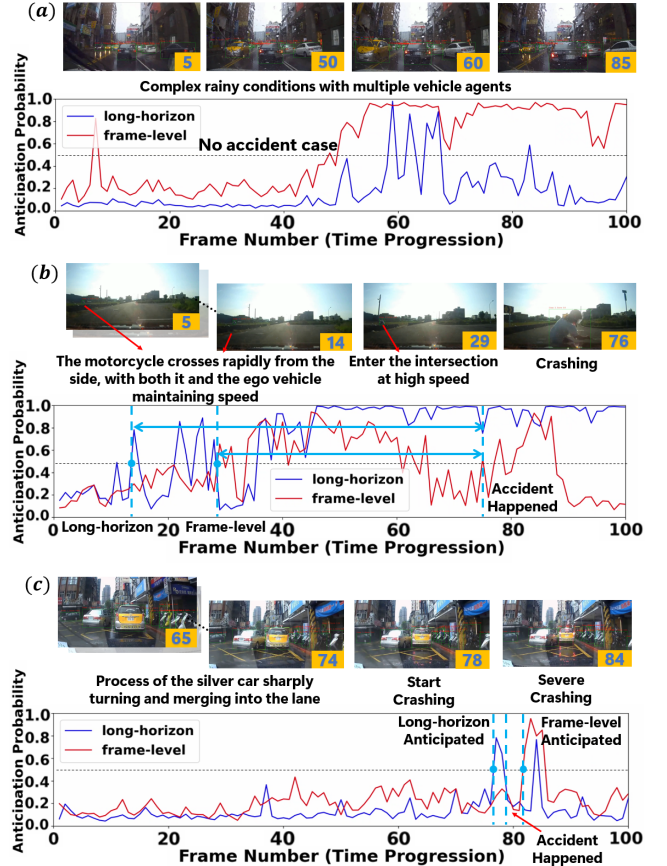


Figure 3: Visualization of long-horizon vs. frame-level anticipation on DAD (threshold 0.5). Scenarios: (a) ambiguous multi-agent rain, (b) predictable collision, (c) sudden complex crash. The long-horizon model offers earlier, more reliable predictions with fewer false positives.

racy and timeliness on both clean and degraded inputs, while qualitative analysis highlights reduced false positives and earlier predictions compared to frame-level baselines. These results demonstrate that robust, long-horizon anticipation enables safer and more proactive autonomous driving.

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau [0122/2024/RIB2, 0215/2024/AGJ, 001/2024/SKL], the Research Services and Knowledge Transfer Office, University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], the Science and Technology Planning Project of Guangdong [2025A0505010016], National Natural Science Foundation of China [52572354], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055].

References

- Abdel-Aty, M.; and Ding, S. 2024. A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nature Communications*, 15(1): 2024–2033.
- Ahmed, S. S.; Pantangi, S. S.; Eker, U.; Fountas, G.; Still, S. E.; and Anastasopoulos, P. C. 2020. Analysis of safety benefits and security concerns from the use of autonomous vehicles: A grouped random parameters bivariate probit approach with heterogeneity in means. *Analytic Methods in Accident Research*, 28: 100134.
- Ali, Y.; Hussain, F.; and Haque, M. M. 2024. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*, 194: 107378.
- Bao, W.; Yu, Q.; and Kong, Y. 2020. Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*.
- Bao, W.; Yu, Q.; and Kong, Y. 2021. DRIVE: Deep Reinforced Accident Anticipation with Visual Explanation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7599–7608.
- Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6154–6162.
- Chan, F.-H.; Chen, Y.-T.; Xiang, Y.; and Sun, M. 2017. Anticipating Accidents in Dashcam Videos. In Lai, S.-H.; Lepetit, V.; Nishino, K.; and Sato, Y., eds., *Computer Vision – ACCV 2016*, 136–153. Cham: Springer International Publishing. ISBN 978-3-319-54190-7.
- Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(09): 10850–10869.
- Fang, J.; Qiao, J.; Bai, J.; Yu, H.; and Xue, J. 2022. Traffic Accident Detection via Self-Supervised Consistency Learning in Driving Scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(7): 9601–9614.
- Fang, J.; Qiao, J.; Xue, J.; and Li, Z. 2024. Vision-Based Traffic Accident Detection and Anticipation: A Survey. *IEEE Trans. Cir. and Sys. for Video Technol.*, 34(4): 1983–1999.
- Fatima, M.; Karim Khan, M. U.; and Kyung, C.-M. 2021. Global Feature Aggregation for Accident Anticipation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2809–2816.
- Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14004–14013.
- Grant, E.; Salmon, P. M.; Stevens, N. J.; Goode, N.; and Read, G. J. 2018. Back to the future: What do accident causation models tell us about accident prediction? *Safety Science*, 104: 99–109.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.
- Karim, M. M.; Li, Y.; Qin, R.; and Yin, Z. 2022. A Dynamic Spatial-Temporal Attention Network for Early Anticipation of Traffic Accidents. *Trans. Intell. Transport. Sys.*, 23(7): 9590–9600.
- Karim, M. M.; Yin, Z.; and Qin, R. 2024. An Attention-Guided Multistream Feature Fusion Network for Early Localization of Risky Traffic Agents in Driving Videos. *IEEE Transactions on Intelligent Vehicles*, 9(1): 1792–1803.
- Li, Z.; Liao, H.; Tang, R.; Li, G.; Li, Y.; and Xu, C. 2023. Mitigating the impact of outliers in traffic crash analysis: A robust Bayesian regression approach with application to tunnel crash data. *Accident Analysis & Prevention*, 185: 107019.
- Liao, H.; Li, Y.; Li, Z.; Bian, Z.; Lee, J.; Cui, Z.; Zhang, G.; and Xu, C. 2024a. Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3D modeling. *Accident Analysis & Prevention*, 207: 107760.
- Liao, H.; Li, Z.; Wang, C.; Shen, H.; Liao, D.; Wang, B.; Li, G.; and Xu, C. 2024b. MFTraj: Map-Free, Behavior-Driven Trajectory Prediction for Autonomous Driving. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 5945–5953. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Liao, H.; Sun, H.; Shen, H.; Wang, C.; Tian, C.; Tam, K.; Li, L.; Xu, C.; and Li, Z. 2024c. CRASH: Crash Recognition and Anticipation System Harnessing with Context-Aware and Temporal Focus Attentions. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*, 11041–11050. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.
- Liu, K.; Zhu, M.; Fu, H.; Ma, H.; and Chua, T.-S. 2020. Enhancing Anomaly Detection in Surveillance Videos with Transfer Learning from Action Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia, MM '20*, 4664–4668. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

- Meulemans, A.; Schug, S.; Kobayashi, S.; Daw, N. D.; and Wayne, G. 2023. Would I have gotten that reward? long-term credit assignment by counterfactual contribution analysis. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*. Red Hook, NY, USA: Curran Associates Inc.
- Pignatelli, E.; Ferret, J.; Geist, M.; Mesnard, T.; van Hasselt, H.; and Toni, L. 2024. A Survey of Temporal Credit Assignment in Deep Reinforcement Learning. *Transactions on Machine Learning Research*. Survey Certification.
- Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, 1–14. Computational and Biological Learning Society.
- Song, W.; Li, S.; Chang, T.; Xie, K.; Hao, A.; and Qin, H. 2024. Dynamic attention augmented graph network for video accident anticipation. *Pattern Recognition*, 147: 110071.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
- Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.
- Suzuki, T.; Kataoka, H.; Aoki, Y.; and Satoh, Y. 2018. Anticipating Traffic Accidents With Adaptive Loss and Large-Scale Incident DB. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Takimoto, Y.; Tanaka, Y.; Kurashima, T.; Yamamoto, S.; Okawa, M.; and Toda, H. 2019. Predicting Traffic Accidents with Event Recorder Data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility, PredictGIS'19*, 11–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369640.
- Thakur, N.; Gouripeddi, P.; and Li, B. 2024. Graph(Graph): A Nested Graph-Based Framework for Early Accident Anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7533–7541.
- Wang, T.; Chen, K.; Chen, G.; Li, B.; Li, Z.; Liu, Z.; and Jiang, C. 2024. GSC: A Graph and Spatio-Temporal Continuity Based Framework for Accident Anticipation. *IEEE Transactions on Intelligent Vehicles*, 9(1): 2249–2261.
- Wu, J.; Zhang, W.; Li, G.; Wu, W.; Tan, X.; Li, Y.; Ding, E.; and Lin, L. 2021. Weakly-Supervised Spatio-Temporal Anomaly Detection in Surveillance Video. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 1172–1178. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Xue, R.; Chen, J.; and Fang, Y. 2020. Real-Time Anomaly Detection and Feature Analysis Based on Time Series for Surveillance Video. In *2020 5th International Conference on Universal Village (UV)*, 1–7.
- Yao, Y.; Wang, X.; Xu, M.; Pu, Z.; Wang, Y.; Atkins, E.; and Crandall, D. J. 2023. DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 444–459.
- Yao, Y.; Xu, M.; Choi, C.; Crandall, D. J.; Atkins, E. M.; and Dariush, B. 2019. Egocentric Vision-based Future Vehicle Localization for Intelligent Driving Assistance Systems. In *2019 International Conference on Robotics and Automation (ICRA)*, 9711–9717. IEEE Press.
- Zeng, K.-H.; Chou, S.-H.; Chan, F.-H.; Niebles, J. C.; and Sun, M. 2017. Agent-Centric Risk Assessment: Accident Anticipation and Risky Region Localization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1330–1338.
- Zhang, J.; Guan, Y.; Wang, C.; Liao, H.; Zhang, G.; and Li, Z. 2025. LATTE: A Real-time Lightweight Attention-based Traffic Accident Anticipation Engine. *Information Fusion*, 122: 103173.