RiskProp: Collision-Anchored Self-Supervised Risk Propagation for Early Accident Anticipation

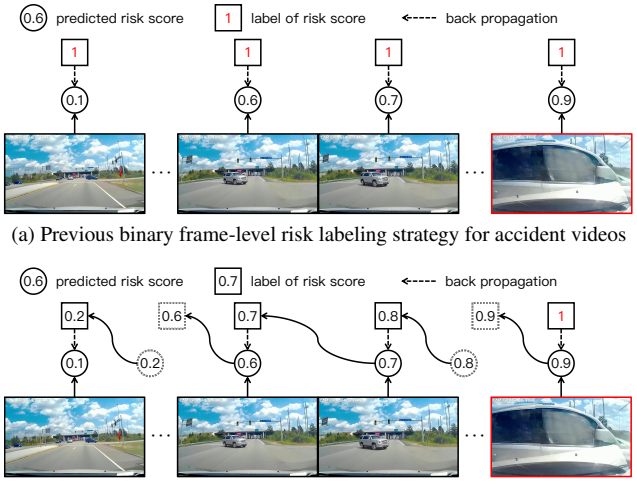
Yiyang Zou[1*] Tianhao Zhao[1,2*] Peilun Xiao[3] Hongyu Jin[3] Longyu Qi[3] Yuxuan Li[3]
Liyin Liang[3] Yifeng Qian[3] Chunbo Lai[3] Yutian Lin[1†] Zhihui Li[4] Yu Wu[1†]
[1]School of Computer Science, Wuhan University [2]Zhongguancun Academy, Beijing, China
[3]Didi Chuxing [4]University of Science and Technology of China
{yiyangzou, happytianhao, yutian.lin, wuyucs}@whu.edu.cn

Abstract

*Accident anticipation aims to predict impending collisions from dashcam videos and trigger early alerts. Existing methods rely on binary supervision with manually annotated "anomaly onset" frames, which are subjective and inconsistent, leading to inaccurate risk estimation. In contrast, we propose **RiskProp**, a novel collision-anchored self-supervised risk propagation paradigm for early accident anticipation, which removes the need for anomaly onset annotations and leverages only the reliably annotated collision frame. RiskProp models temporal risk evolution through two observation-driven losses: first, since future frames contain more definitive evidence of an impending accident, we introduce a future-frame regularization loss that uses the model's next-frame prediction as a soft target to supervise the current frame, enabling backward propagation of risk signals; second, inspired by the empirical trend of rising risk before accidents, we design an adaptive monotonic constraint to encourage a non-decreasing progression over time. Experiments on CAP and Nexar demonstrate that RiskProp achieves state-of-the-art performance and produces smoother, more discriminative risk curves, improving both early anticipation and interpretability. Code: https://github.com/xingyueye5/RiskProp/.*

1. Introduction

As autonomous vehicles and advanced driver assistance systems continue to evolve, accident anticipation has become increasingly important for improving road safety. Its goal is to predict whether an accident is likely to occur in the near future based on the current driving scene captured by a dashcam. To achieve this, the system continuously estimates a risk score in real time, and if the score exceeds a predefined threshold, an alert is triggered to enable early intervention. This allows both autonomous systems and human drivers to take preventive actions before an accident happens, thereby reducing the risk of collision and enhancing overall driving safety.



(a) Previous binary frame-level risk labeling strategy for accident videos

(b) Self-supervised frame-level risk propagation for accident videos

Figure 1. Form of supervision of previous works and ours. (a) Most previous methods treat all frames in accident-free videos as negative samples with label 0, and treat frames in accident videos—from the annotated anomaly onset frame to the collision frame—as positive samples with labeled 1. (b) Our method treats the model's prediction for the next frame as a supervision signal for the current frame, which enables the risk values to propagate gradually backward from the collision frame.

Most existing works [5, 18, 21, 38, 43] for accident anticipation train models to predict a risk score for each frame by treating the task as supervised learning with binary labels: all frames in accident-free videos are labeled 0, while all frames in accident videos—from the beginning up to collision—are labeled 1, as shown in Fig. 1 (a). To compensate for the fact that risk in pre-collision frames is typically inter-

*Equal contribution.
†Corresponding author.

mediate and evolves gradually, these methods apply exponentially decaying weights to the binary cross-entropy loss, assigning lower penalties to earlier frames. To further improve early anticipation, some works [32, 33] adopt adaptive loss weighting schemes, while others [10, 26] manually annotate the "anomaly onset" frame for accident videos and adjust loss weights relative to it. However, this binary labeling paradigm is fundamentally flawed: it forces the model to treat all pre-collision frames as equally risky, ignoring the intermediate and scenario-dependent nature of risk evolution—such as slow increases when a driver is distracted *versus* rapid spikes when a pedestrian suddenly appears. Moreover, the manual annotation of anomaly onset is subjective and inconsistent across annotators, leading to noisy and unreliable supervision. As a result, the model is misled to optimize risk predictions toward inaccurate binary targets rather than reflecting the true, dynamic progression of risk.

Given the limitations of handcrafted labels and unreliable manual annotations, we turn to data-driven observations about the nature of risk in driving scenarios. First, we observe that accident anticipation fundamentally aims to predict whether a collision will occur in the future based on current observations. Since future frames contain more definitive evidence about the occurrence of an accident, the model's risk predictions for these frames are typically more accurate and better aligned with ground truth than those for earlier frames. This suggests that future predictions can serve as reliable pseudo-supervision signals to guide the learning of current risk estimates. Second, we find that in accident videos, the underlying risk tends to follow a non-decreasing trend over time within a long temporal window before collision—reflecting the progressive deterioration of driving safety as hazardous conditions unfold.

Building on these observations, we propose RiskProp, a novel collision-anchored self-supervised risk propagation paradigm for early accident anticipation. It generates intermediate risk scores for pre-collision frames by propagating supervision backward from the collision frame without relying on binary labeling schemes or manual anomaly onset annotations. Specifically, as shown in Fig. 1 (b), for accident videos, we assign a hard risk label of 1 only to the collision frame, whose timing is reliably annotated. For all other pre-collision frames in accident videos, we generate soft supervision signals by propagating risk information backward in time instead of fixed binary labels, where the model's prediction on the next frame (detached from gradient computation) serves as the target for the current frame. This self-supervised temporal regularization enables early frames to receive intermediate and discriminative risk supervision, reflecting the gradual buildup of danger rather than being forced into arbitrary hard labels. For accident-free videos, all frames are assigned a risk label of 0. By design, RiskProp avoids the need for subjective anomaly onset annotation and learns a more nuanced risk progression grounded in temporal consistency. Furthermore, to encourage a plausible long-term trend in risk evolution, we introduce an adaptive monotonic constraint loss that penalizes violations where a later frame is assigned a lower risk than an earlier one within the same sequence. This loss operates over randomly sampled frame pairs and allows short-term fluctuations while promoting an overall non-decreasing pattern consistent with real-world dynamics.

Our contributions can be summarized as follows:

- We reveal the limitations of binary labeling and manual anomaly onset annotations in accident anticipation, and instead base our approach on two empirical observations: future frames provide more reliable risk cues, and pre-collision risk tends to grow non-decreasingly over time.
- We propose RiskProp, a novel collision-anchored self-supervised paradigm that generates intermediate risk scores by propagating supervision backward from the collision frame. It introduces a future-frame regularization loss and an adaptive monotonic constraint loss, eliminating the need for subjective annotations and encouraging early anticipation.
- Extensive experiments on two datasets demonstrate that our method achieves state-of-the-art performance and produces risk curves with better temporal consistency and earlier warning capability.

2. Related Work

2.1. Traffic Accident Anticipation

Existing accident anticipation methods predominantly adopt a frame-level binary supervision paradigm that predicts a risk score for each frame. In terms of label supervision, some works [3, 5, 33] define positive samples using a fixed temporal window before collision, while others [8, 10] rely on manually annotated anomaly onset label. Several approaches further introduce collision-time-anchored exponential or adaptive loss weighting schemes [5, 17], which emphasize frames closer to the accident in training. However, both rigid temporal windows and subjective onset annotations introduce strong assumptions about risk evolution, often resulting in ambiguous supervision and temporally volatile or physically implausible risk predictions.

Beyond supervision design, many works leverage auxiliary cues such as driver gaze [9] or causal reasoning [11] to improve interpretability, while others employ object detectors or attention maps for better spatial grounding [10, 18, 20, 32, 44]. Transformer-based architectures have also been widely adopted to capture long-range temporal dependencies in accident anticipation [1, 2, 14, 15, 19, 30, 37, 41]. Specifically, Wang et al. [35] jointly models long-term memory and short-term anticipation to achieve fast anticipation, while several works [23, 27, 36] model egocentric

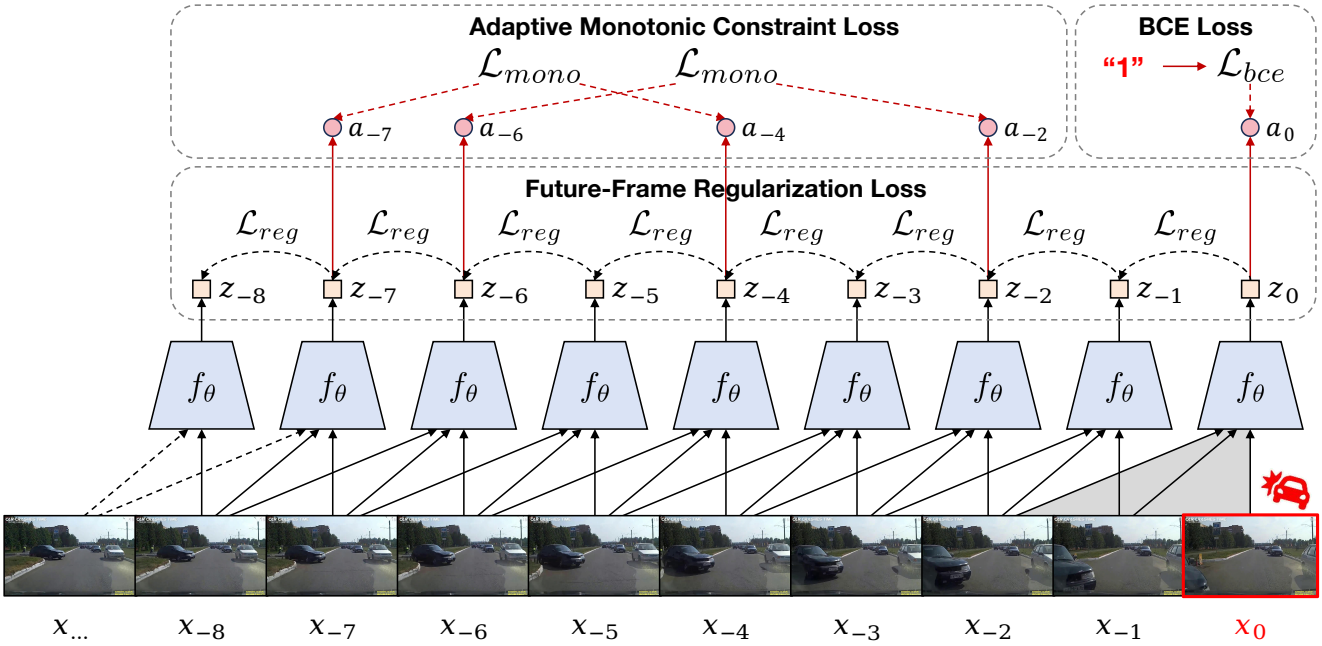


Figure 2. **Overview of our RiskProp framework.** The encoder-only model takes a snippet of consecutive frames and predicts the current frame's risk score. To train such a model without "anomaly onset" labels and with only the objective collision frame label, two losses are proposed: The **Future-Frame Regularization Loss** uses the next-frame's detached prediction as a self-supervised target for high-risk signals backward propagation, while the **Adaptive Monotonic Constraint Loss** imposes a monotonicity constraint to ensure a non-decreasing overall risk trend. And we adopt the Binary Cross-Entropy loss to provide explicit supervision only at the collision frame. This enables stable, physically plausible risk curves without manual onset annotations.

action prediction using graph structures. RAFTformer [13] further introduces a future-aware teacher–student design for real-time action forecasting. Despite these architectural advances, most existing methods remain fundamentally constrained by rigid or subjective binary supervision schemes.

2.2. Self-Supervised Temporal Modeling in Video

Self-supervised learning has emerged as a powerful paradigm for leveraging temporal structure in videos. Early methods exploit temporal order [25] or cycle consistency [7] as pretext tasks to learn video representations. Contrastive learning frameworks [16, 31] pull nearby frames closer in embedding space, enforcing local smoothness. Some works [4, 42] use label smoothing for early event prediction. A line of work uses future-aware "teacher" models to guide real-time "student" models [13, 34], enabling anticipation through knowledge distillation. Ristea et al. [29] propose a self-distilled masked auto-encoder framework for effective video anomaly detection. Our approach is conceptually related but significantly simpler: we use the next frame's prediction from the same model (with gradient detachment) as a self-supervised target. This design avoids architectural complexity while effectively propagating high-risk signals backward in time.

2.3. Monotonicity in Sequential Prediction

Modeling irreversible processes often requires monotonic behavior [6]. In accident anticipation, several works enforce increasing risk trends via weighted losses [5] or adaptive penalties [33]. Pairwise ranking losses [3] further encourage later frames to have higher scores than earlier ones, promoting local monotonicity. Some incorporate uncertainty modeling to stabilize predictions [24], but still operate within the supervised annotation paradigm and rely on manually defined labels. By enforcing a non-decreasing danger assumption, [28] imposes temporal monotonicity but fails to capture false-alarm scenarios in which perceived risk rises despite no actual accident. Our method enforces a non-decreasing long-term risk trend while allowing short-term fluctuations, yielding smooth and physically plausible risk curves aligned with real driving dynamics.

3. Method

3.1. Framework Overview

Given a dashcam video sequence, the goal of traffic accident anticipation is to estimate, at each time step, the probability that a collision will occur in the near future. The model aims to encode the past and current frames and outputs the risk score for the current frame. An accident alert is triggered if

this score exceeds a predefined threshold.

In this paper, we propose a novel RiskProp framework, which models the temporal evolution of risk by leveraging both objective supervision from the collision frame and self-supervised temporal consistency. As illustrated in Fig. 2, for the timestamp t, the input frames are denoted as $\mathbf{x}_t = \{x_{t-O+1}, \ldots, x_t\}$, where O is the number of observed frames ($O = 3$ in the figure). The model encodes this snippet and outputs a risk representation $z_t = f_\theta(\mathbf{x}_t)$, which is then transformed into a risk score through a sigmoid activation, $a_t = \sigma(z_t)$ ($a_t \in (0, 1)$).

Our training objective is designed to guide these risk predictions towards a realistic and temporally consistent evolution pattern. As shown in Fig. 2, the BCE loss and two auxiliary losses are adopted: (1) the future-frame regulation loss propagates risk signals backward from the collision frame; (2) the adaptive monotonic constraint loss encourages roughly non-decreasing risk curves. Together, these components enable RiskProp to learn risk curves that closely approximate real-world risk evolution.

3.2. Future-Frame Regularization Loss

A key characteristic of driving-risk evolution is its temporal continuity: hazardous cues emerge gradually, and future frames usually provide stronger and more decisive evidence of an upcoming accident than earlier ones. This suggests a natural self-supervision signal: the predicted risk score of the next frame can serve as a soft label for the current frame, enabling backward propagation of high-risk signals from the collision frame to earlier timesteps.

To leverage this characteristic, we introduce the **Future-Frame Regularization Loss**. Specifically, let $\texttt{detach}(\cdot)$ denote the stop-gradient operation, which prevents gradients from flowing back through $f_\theta(\mathbf{x}_{t+1})$ during backpropagation. We define our future-frame regularization loss as:

$$\mathcal{L}_{\text{reg}} = \sum_{t=1}^{T-1} \left\| \texttt{detach}(z_{t+1}) - z_t \right\|^2, \tag{1}$$

where the collision occurs at time T. The operation treats the detached variable $\texttt{detach}(z_t)$ as a *frozen target* rather than a trainable parameter. Moreover, this loss establishes a backward credit assignment pathway. Since the collision frame T has the only ground-truth label $y_T = 1$, its high predicted score propagates backward through the chain of $\mathcal{L}_{\text{reg}}$ terms:

$$\texttt{detach}(z_{T+1}) \to z_T, \quad \texttt{detach}(z_T) \to z_{T-1}, \quad \ldots$$

Thus, even without explicit onset labels, early frames receive indirect supervision from the definitive hazard signal at the collision point. Through this backward propagation of risk information, the model learns to recognize early-stage cues that precede accidents, such as subtle lane drifts, delayed reactions, or unsafe following distances.

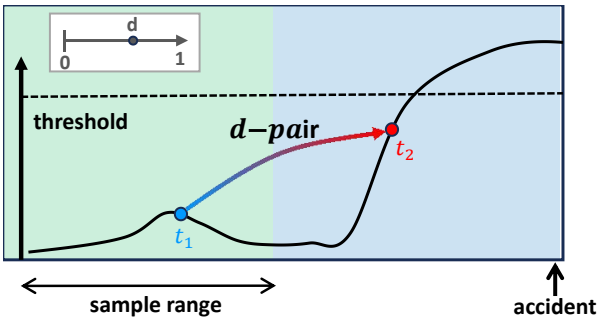


Figure 3. Illustration of sampling strategy for the adaptive monotonic constraint loss. For a randomly selected $d \in [d_{min}, d_{max}]$, a starting frame t_1 is sampled from $[0, T(1-d)]$, and $t_2 = t_1 + dT$. This encourages learning across various temporal distances.

3.3. Adaptive Monotonic Constraint Loss

In real-world accident scenarios, the level of danger generally increases as collision approaches. To capture this inherent progression, we introduce an **Adaptive Monotonic Constraint Loss** that encourages the model's predicted risk scores to exhibit an overall non-decreasing trend along the temporal axis. This enforces a physically plausible evolution of risk while maintaining flexibility for minor fluctuations.

Formally, for sampled frame pairs (x_i, x_j) where $j > i$, we expect:

$$a_j \geq a_i. \tag{2}$$

To leverage this hypothesis, we define a differentiable monotonic constraint loss to penalize cases where a later frame is predicted to be less risky than an earlier frame:

$$\mathcal{L}_{\text{mono}} = \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} \max\left(0,\, a_i - a_j + \delta(\Delta t, \bar{c}_{i:j})\right), \tag{3}$$

where $\mathcal{D}$ denotes the set of randomly sampled frame pairs, and $\delta(\Delta t, \bar{c}_{i:j})$ is an adaptive tolerance margin controlling how strictly monotonicity is enforced.

The margin δ adapts from two perspectives: (1) The temporal distance Δt. The margin increases with the temporal distance $\Delta t = t_j - t_i$, reflecting the expectation that risk may rise over longer intervals. (2) The confidence score $\bar{c}_{i:j}$. The enforcement strength depends on how confident the model is in its risk estimates. When the predicted risk scores are close to 0 or 1 (high confidence), the constraint should be stricter; when they are near the mean (low confidence), the constraint should relax. Formally, the adaptive tolerance margin is defined as:

$$\delta(\Delta t, \bar{c}_{i:j}) = \delta_0 \cdot \Delta t \cdot \bar{c}_{i:j}, \tag{4}$$

where δ_0 is a scaling coefficient, Δt measures the relative distance between two frames, and $\bar{c}_{i:j}$ denotes their average

prediction confidence. For convenience, $\bar{c}_{i:j}$ is computed as:

$$\bar{c}_{i:j} = \frac{c_i + c_j}{2}, \quad c_i = 2|a_i - \bar{a}|, c_j = 2|a_j - \bar{a}|, \tag{5}$$

where $\bar{a}$ denotes the batch-wise mean risk score computed over positive videos in the batch, reflecting the typical risk levels in driving scenarios. A prediction farther from $\bar{a}$ is treated as more confident because it lies closer to either low-risk or high-risk extremes, while predictions near $\bar{a}$ indicate ambiguous intermediate states. This formulation ensures that higher confidence leads to stronger monotonic constraints.

This adaptive formulation balances flexibility and stability by adjusting the monotonic constraint to temporal distance and prediction confidence. It relaxes the constraint for short intervals or uncertain predictions, preventing over-regularization, while tightening it when frames are far apart or confidence is high, ensuring a globally consistent rising trend of risk over time.

Sample strategy. As illustrated in Fig. 3, to ensure temporal diversity, we randomly sample frame pairs using offsets $d \in [d_{\min}, d_{\max}]$. Given d, we then sample a starting frame i uniformly at random from the interval $[0, T(1-d)]$, where T is the total length of the video clip, and set $j = i + dT$.

This sampling strategy offers two benefits. By randomly varying the temporal offset d, the model learns from both short- and long-term dynamics, improving robustness across different accident stages. Moreover, it aligns with the physical continuity of driving: while risk may rise gradually at first, it should not decrease once danger unfolds. The loss thus regularizes the risk curve to remain non-decreasing in expectation, yielding smooth and realistic progression.

3.4. Labeling Strategy and Training Objective

A major challenge in accident anticipation lies in the absence of reliable supervision. Manual annotations of "anomaly onset" are often subjective and inconsistent, whereas the collision timestamp provides an unambiguous and physically grounded reference. Accordingly, we adopt a simple labeling strategy: the actual collision frame is labeled as positive (label = 1), while only the initial start frame that is far from the accident is labeled as negative (label = 0). All intermediate frames are left unlabeled, and their risk values are inferred through self-supervised risk propagation.

We adopt a weighted Binary Cross-Entropy (BCE) loss. For accident videos, the loss is applied only to the starting frame and the collision frame:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} w_i \left[y_i \log a_i + (1 - y_i) \log(1 - a_i) \right], \tag{6}$$

where $\mathcal{S}$ denotes the set of two frames that are either the first frame of a training snippet or the accident frame, a_i is the model's predicted danger score for frame i, $y_i \in \{0, 1\}$ is the label, and w_i is the frame weight. To alleviate the severe supervision imbalance between sparse positive anchors and abundant negative frames, we assign a higher weight to the collision frames than to negative frames. For non-accident videos, since no collision anchor is available and the monotonicity assumption does not hold, both the future-frame regularization and the adaptive monotonic constraint are disabled. Therefore, we supervise these sequences by assigning all frames a negative label ($y_t = 0$ for all t) and optimizing the BCE loss over the entire video. The full training objective is computed as:

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \lambda_1 \cdot \mathcal{L}_{\text{reg}} + \lambda_2 \cdot \mathcal{L}_{\text{mono}}, \tag{7}$$

where λ_1 and λ_2 are hyperparameters balancing the contributions of the regularization terms. This formulation ensures that the model learns to anticipate danger not by memorizing subjective labels, but by discovering the underlying temporal structure of risk evolution, with the collision frame acting as the sole anchor point for supervision.

4. Experiments

4.1. Experimental Setup

Datasets. We validated the effectiveness of our method using experiments on two real-world traffic accident datasets: **CAP** [10] comprises 11,727 ego-view accident videos, totaling 2,195,613 frames, and covers 58 distinct accident categories. These videos were sourced from various existing public accident datasets, such as CCD [3], A3D [39], DoTA [40], and DADA-2000 [8], in addition to streaming platforms like YouTube, Bilibili, and Tencent Video. It provides temporal labels for keyframes, including "anomaly onset", "accident occur", and "accident end". **Nexar** [26] consists of 1,500 real-world dashcam video clips, each approximately 40 seconds long. Videos are recorded by Nexar dashcams with a resolution of 1280x720 at 30 frames per second. It was originally compiled for the Kaggle competition, Nexar Dashcam Crash Prediction Challenge. The test set is composed of 1344 videos with a duration of approximately 10 seconds, and we follow the preprocessing protocol strictly as [26].

Evaluation metrics. We follow the evaluation protocol of Nexar [26] and compute all metrics under a strict false alarm rate (FAR) constraint of $\lambda = 0.1$, as suggested by [45]. This setting ensures a realistic and fair comparison by maintaining a balanced trade-off between FAR and mean time-to-accident (mTTA). According to Nexar [26], we sample positive clips from four intervals before the accident: $[0.0, 0.5)$, $[0.5, 1.0)$, $[1.0, 1.5)$, and $[1.5, 2.0)$ seconds,

Table 1. Quantitative results comparison of different methods on the CAP [10] dataset and the Nexar [26] dataset.

Method	CAP [10]				Nexar [26]			
	$\text{mAUC}^{0.1}$	mAUC	mAP	$\text{mTTA}^{0.1}$ (s)	$\text{mAUC}^{0.1}$	mAUC	mAP	$\text{mTTA}^{0.1}$ (s)
AdaLEA [33]	0.379	0.807	0.857	1.115	0.378	0.828	0.832	0.858
XAI [17]	0.352	0.821	0.864	1.082	0.346	0.824	0.802	0.839
DSTA [18]	0.361	**0.895**	0.882	0.894	0.242	0.783	0.764	0.493
GSC [36]	0.378	0.881	**0.890**	0.911	0.322	0.802	0.811	0.815
CAP [10]	0.332	0.811	0.852	0.933	0.315	0.817	0.793	0.801
CRASH [22]	0.401	0.842	0.887	1.085	0.393	0.832	0.846	0.857
Ours	**0.483**	0.853	**0.890**	**1.207**	**0.472**	**0.869**	**0.870**	**0.958**

with an equal number of negative clips from safe driving segments. For each interval starting at τ, we compute:

- **Constrained AUC**$_\tau^\lambda$ and **AP**$_\tau$: According to [45], AUC and AP are calculated only when FAR ≤ 0.1. $\text{AUC}_{0.0\text{s}}^\lambda$ reflects accident detection capability, while later intervals measure anticipation performance.
- **mTTA**$^\lambda$: Mean TTA of true positive alarms, computed only when FAR ≤ 0.1. Note that mTTA requires the annotation of anomaly onset and is therefore used only as a compatibility metric for fair comparison with existing baselines, not for training. To avoid inflated values, we consider only alarms triggered *after* anomaly onset.

We report the primary metrics as:

$$\text{mAUC}^\lambda = \frac{1}{3} \sum_{\tau \in \{0.5, 1.0, 1.5\}} \text{AUC}_\tau^\lambda, \tag{8}$$

$$\text{mAP} = \frac{1}{3} \sum_{\tau \in \{0.5, 1.0, 1.5\}} \text{AP}_\tau, \tag{9}$$

$$\text{mTTA}^\lambda = \mathbb{E}_{\theta:\, \text{FAR}(\theta) \leq \lambda} \left[\text{TTA}^\theta \right]. \tag{10}$$

Implementation details. In our experimental setup, training snippets are sampled only from the pre-collision period. During testing we apply a causal sliding window to the pre-collision portion of each accident video to obtain frame-level predictions. We employ a 3D CNN as the snippet encoder to jointly extract spatial and temporal features. After encoding, we apply only adaptive spatial average pooling, preserving the temporal dimension of the feature sequence. In the transform pipeline, we resize each video frame to 224×224 and resample frames to a frame rate of 10 frames per second, so that the model can process the frames within a 0.1s time interval. We set the number of observed frames (the number of frames in a snippet) to 5. We use SlowOnly [12] as our snippet encoder and train our model based on their pre-trained weights. We set the adaptive monotonic constraint sampling hyperparameters $d_{min} = 0.1$, $d_{max} = 0.9$, $\delta_0 = 0.01$. We set the regularization terms balancing hyperparameters $\lambda_1 = 1.5$, $\lambda_2 = 1.1$, and optimize the model with the SGD optimizer. Our model was trained for 50 epochs on 8 NVIDIA A800 GPUs with a batch size of 64, using an initial learning rate of 0.002 that was decayed to 10% of its previous value every 20 epochs.

4.2. Comparison with State-of-the-Art Methods

We compare the quantitative results of our method with previous methods AdaLEA [33], DSTA [18], GSC [36], CAP [10], XAI [17] and CRASH [22] on both the CAP dataset and the Nexar dataset, to demonstrate its effectiveness. All baselines are re-trained following the evaluation protocol in Nexar [26], ensuring a fair comparison under the constrained false-alarm rate setting ($\lambda = 0.1$). The results are shown in Table 1.

On the CAP dataset, our method significantly outperforms previous approaches, achieving $\text{mAUC}^{0.1}$ score of 0.483. While achieving a top-tier mAP of 0.890, our model also delivers the earliest warnings, leading all compared methods with the best mean Time-to-Accident ($\text{mTTA}^{0.1}$) of 1.207 seconds. Notably, DSTA achieves the highest mAUC (0.895), but our method maintains competitive performance (0.853) while excelling in other critical metrics.

The superiority of our approach is further emphasized on the more challenging Nexar dataset. Our method establishes state-of-the-art results across all metrics: $\text{mAUC}^{0.1}$ (0.472), mAUC (0.869), and mAP (0.870), which surpass the second best method CRASH by 0.079, 0.037 and 0.024, respectively. It also achieves the highest $\text{mTTA}^{0.1}$ of 0.958 seconds, underscoring its robust and effective capability for early accident anticipation. The above results demonstrate that our model not only predicts accidents with better accuracy but also provides more timely alerts, validating its effectiveness for real-world safety systems.

4.3. Ablation Study

We conduct an extensive ablation study on the CAP and Nexar datasets to evaluate the effectiveness of our proposed components: the future-frame regularization loss (FFR) and

Table 2. Ablation study about different annotation strategies on the CAP [10] dataset, where FFR denotes Future-Frame Regularization, and AMC denotes Adaptive Monotonic Constraint.

Exp.	FFR	AMC	Anomaly Onset Label				Fixed Interval Label				Only Collision Label			
			mAUC$^{0.1}$	mAUC	mAP	mTTA$^{0.1}$(s)	mAUC$^{0.1}$	mAUC	mAP	mTTA$^{0.1}$(s)	mAUC$^{0.1}$	mAUC	mAP	mTTA$^{0.1}$(s)
I			0.441	0.829	0.871	1.162	0.412	0.816	0.872	1.173	0.358	0.783	0.832	1.009
II		✓	0.436	0.828	0.866	1.137	0.436	0.828	0.872	1.173	0.383	0.790	0.832	1.055
III	✓		0.444	0.830	0.872	1.174	0.455	0.847	0.886	**1.212**	0.474	0.850	0.850	1.202
IV	✓	✓	**0.484**	**0.856**	**0.887**	**1.198**	**0.480**	**0.851**	**0.890**	1.203	**0.483**	**0.853**	**0.890**	**1.207**

Table 3. Ablation study about different annotation strategies on the Nexar [26] dataset, where FFR denotes Future-Frame Regularization, and AMC denotes Adaptive Monotonic Constraint.

Exp.	FFR	AMC	Anomaly Onset Label				Fixed Interval Label				Only Collision Label			
			mAUC$^{0.1}$	mAUC	mAP	mTTA$^{0.1}$(s)	mAUC$^{0.1}$	mAUC	mAP	mTTA$^{0.1}$(s)	mAUC$^{0.1}$	mAUC	mAP	mTTA$^{0.1}$(s)
I			0.388	0.826	0.822	0.773	0.420	0.837	0.827	0.741	0.298	0.789	0.781	0.610
II		✓	0.408	0.844	0.842	0.902	0.445	0.815	0.830	0.815	0.302	0.797	0.785	0.654
III	✓		0.434	0.846	0.850	0.897	0.434	0.865	0.856	0.920	0.453	0.847	0.854	0.836
IV	✓	✓	**0.479**	**0.872**	**0.876**	**0.951**	**0.454**	**0.875**	**0.874**	**0.931**	**0.472**	**0.869**	**0.870**	**0.958**

the adaptive monotonic constraint loss (AMC). Tab. 2 and 3 report results on the two datasets under three different annotation paradigms: (1) Anomaly Onset, where frames from the manually annotated anomaly start to the collision are labeled as positive; (2) Fixed Interval, where a fixed 2-second window before the collision is labeled as positive; and (3) Only Collision, where only the collision frame is labeled as positive sample, while other frames are left unlabeled.

Effectiveness of the future-frame regularization loss. As shown across both tables, our FFR loss consistently improves performance across all labeling strategies and datasets. Specifically, comparing with Exp. I and Exp. III, we observe that applying only the FFR constraint provides the most significant performance gain. On CAP, under the three annotation schemes, it increases the mAUC$^{0.1}$ by 0.003, 0.043 and 0.116, respectively. On Nexar, under the three annotation schemes, it improves the mAUC$^{0.1}$ by 0.091, 0.034 and 0.174, respectively. This confirms that enforcing temporal correlation through future frame regularization effectively propagates the high-risk signal from the collision frame backward in time.

We also observe that the baseline (Exp. I) of our Only Collision setting yields the worst performance among the three settings, where we obtain mAUC$^{0.1}$ of 0.358 on CAP and 0.298 on Nexar. However, by incorporating FFR (Exp. III), the performance is dramatically boosted to a mAUC$^{0.1}$ of 0.474 on CAP and 0.453 on Nexar—a remarkable improvement of 0.116 and 0.155, respectively.

Effectiveness of adaptive monotonic constraint loss. As shown across both tables (Exp., I vs. Exp. IV), our AMC loss improves performance significantly across all labeling strategies and datasets when combined with FFR. Specifically, on Nexar, under the three annotation schemes, it increases the mAUC$^{0.1}$ by 0.02, 0.025 and 0.004, respectively. The consistent improvements demonstrate that adaptive monotonic constraint loss enforces the risk score increase over time. This enhances the reliability of the predictive curve.

The combination of both constraints yields the best overall results. Adding AMC on top of FFR (Exp. IV) further refines the predictions, leading to the highest scores in nearly all metrics on both datasets, such as improving the mAP from 0.850 to 0.890 on CAP. This indicates that the monotonic risk trend enforced by AMC loss acts as a powerful regularizer for the temporally smooth predictions generated by the future-frame regularization loss.

4.4. Analysis

Effectiveness of only-collision supervision with our self-supervised constraints. A key finding from our study is that our model, empowered by the FFR and AMC constraints, reduces the dependence on dense temporal annotations, though collision labels are still required. While the Only Collision strategy performs the worst without our constraints (Exp. I) due to a lack of full supervision, it achieves state-of-the-art results when they are applied (Exp. IV).

Remarkably, on both datasets, the full model trained with the Only Collision strategy (Exp. IV) performs on par with, and in some cases better than, the same model trained with more complex and costly labels. For instance, on the CAP dataset, its mAUC$^{0.1}$ of 0.483 is competitive with the 0.484 from the full Annotation setting. On the Nexar dataset, its mAP of 0.870 and mTTA$^{0.1}$ of 0.958s are the highest across all experiments. This is a significant result: it

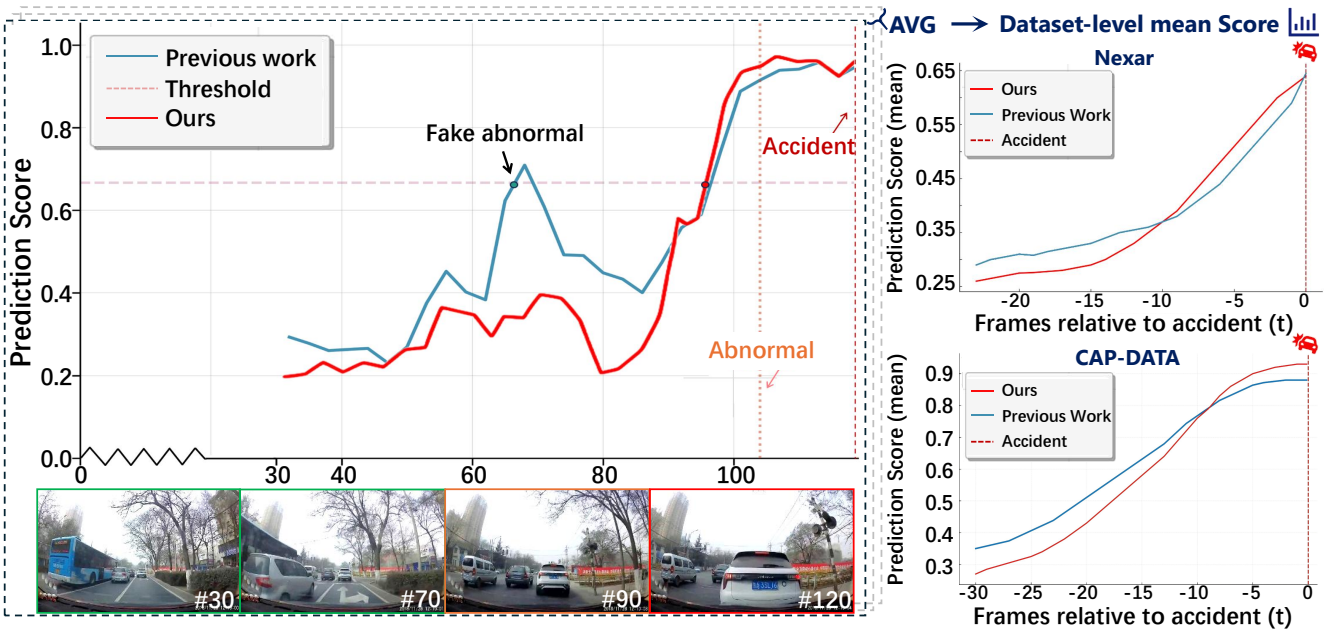


Figure 4. **Frame- and Dataset-Level Risk Prediction.** Left: On a representative accident video, supervised baselines produce early false peaks before clear risk emerges. Our RiskProp model prediction remains low during safe periods and rises sharply only when discriminative cues appear, yielding a temporally coherent risk curve. Right: Dataset-wide average risk curves on CAP and Nexar, aligned to accident timestamps. RiskProp suppresses early false positives and delivers a steeper, better-calibrated rise near the event, enabling earlier, more reliable warnings.

proves that our method can reach performance levels close to those achieved by dense annotation, without requiring humans to subjectively annotate the "start" of an accident. The proposed self-supervised constraints effectively compensate for the lack of dense supervision, enabling a more practical and scalable approach to accident anticipation.

In summary, our ablation studies validate two core conclusions: 1) FFR and AMC are essential components that significantly improve prediction accuracy. 2) When combined with these constraints, our Only Collision supervision matches the performance of densely labeled approaches, offering a powerful yet annotation-efficient solution.

Analysis of the predicted risk curve. As shown in Fig. 4, our method produces more realistic and reliable risk curves than prior work. When a vehicle appears on the side at frame 70, the baseline method falsely raises its risk score above the alert threshold, triggering a premature false alarm. In contrast, our method remains a low risk estimate during this safe period, only exhibiting a sharp, sustained rise when the leading vehicle becomes critically close around frame 90, precisely when real danger appears. This shows that our self-supervised framework, guided by future-frame regularization and adaptive monotonicity constraints, effectively suppresses false positives induced by ambiguous cues, ensuring alerts are triggered only upon clear evidence of a risky accident. Consequently, the resulting risk evolution is smooth, temporally coherent, and aligns with the physical progression of real-world driving accidents.

5. Conclusion

In this work, we propose RiskProp, a novel collision-anchored self-supervised risk propagation paradigm for modeling risk evolution in early accident anticipation. Instead of learning from binary labels that indicate whether a frame is risky, our method uses the model's own prediction at the next frame as a soft supervision signal for the current frame. Therefore, the risk information anchored at the collision point can be gradually propagated backward through the sequence. This shifts the training paradigm from static risk classification to dynamic risk evolution modeling, enabling a more temporally aware and physically plausible anticipation process. To further shape this evolution, we incorporate an adaptive monotonic constraint loss, which encourages a generally increasing long-term risk trend as the scene progresses toward an accident, while still allowing natural short-term fluctuations. Together, these losses guide the model to learn smooth, irreversible risk evolution without requiring subjective anomaly-onset labels. Extensive experiments on CAP and Nexar show that RiskProp achieves state-of-the-art performance with significantly more stable and interpretable risk curves.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62471344), the Zhongguancun Academy (Project No. 20240304), and the CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

References

[1] Mansoor G Al-Thani, Ziyu Sheng, Yuting Cao, and Yin Yang. Traffic transformer: Transformer-based framework for temporal traffic accident prediction. 2024.

[2] BM Tazbiul Hassan Anik, Zubayer Islam, and Mohamed Abdel-Aty. A time-embedded attention-based transformer for crash likelihood prediction at intersections using connected vehicle data. *Transportation Research Part C: Emerging Technologies*, 169:104831, 2024.

[3] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM MM*, pages 2682–2690, 2020.

[4] Guglielmo Camporese, Pasquale Coscia, Antonino Furnari, Giovanni Maria Farinella, and Lamberto Ballan. Knowledge distillation for action anticipation via label smoothing. In *2020 25th international conference on pattern recognition (ICPR)*, pages 3312–3319. IEEE, 2021.

[5] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *ACCV*, pages 136–153. Springer, 2017.

[6] Dangxing Chen and Weicheng Ye. How to address monotonicity for model risk management? In *International Conference on Machine Learning*, pages 5282–5295. PMLR, 2023.

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, pages 1801–1810, 2019.

[8] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention analyzed by a benchmark. In *ITSC*, pages 4303–4309. IEEE, 2019.

[9] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4959–4971, 2021.

[10] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. Cognitive accident prediction in driving scenes: A multimodality benchmark. *IEEE Intelligent Transportation Systems Magazine*, 2024.

[11] Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *CVPR*, pages 22030–22040, 2024.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, pages 6202–6211, 2019.

[13] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *CVPR*, pages 18759–18769, 2023.

[14] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, pages 13505–13515, 2021.

[15] Weili Guan, Xuemeng Song, Kejie Wang, Haokun Wen, Hongda Ni, Yaowei Wang, and Xiaojun Chang. Egocentric early action prediction via multimodal transformer-based dual action prediction. *IEEE TCSVT*, 33(9):4472–4483, 2023.

[16] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshops*, pages 0–0, 2019.

[17] Muhammad Monjurul Karim, Yu Li, and Ruwen Qin. Toward explainable artificial intelligence for early anticipation of traffic accidents. *Transportation research record*, 2676(6): 743–755, 2022.

[18] 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, 2022.

[19] Yuto Kumamoto, Kento Ohtani, Daiki Suzuki, Minori Yamataka, and Kazuya Takeda. Aat-da: Accident anticipation transformer with driver attention. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1142–1151, 2025.

[20] Lei-Lei Li and Jianwu Fang. Traffic accident anticipation via driver attention auxiliary. In *International Conference on Autonomous Unmanned Systems*, pages 348–360. Springer, 2023.

[21] 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, 2024.

[22] 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*, pages 11041–11050, 2024.

[23] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *CVPR*, pages 13904–13913, 2022.

[24] Zakaria Mhammedi. Risk monotonicity in statistical learning. *NeurIPS*, 34:10732–10744, 2021.

[25] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016.

[26] Daniel C. Moura, Shizhan Zhu, and Orly Zvitia. Nexar dashcam collision prediction dataset and challenge, 2025.

[27] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, pages 163–172, 2020.

[28] Aurel Pjetri, Davide Abbondandolo, Douglas Coimbra de Andrade, Stefano Caprasecca, Francesco Sambo, and Andrew David Bagdanov. Self-supervised road accident anticipation with non-decreasing danger. In *ECCV*, pages 65–79. Springer, 2024.

[29] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *CVPR*, pages 15984–15995, 2024.

[30] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6740–6750, 2024.

[31] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[32] Inpyo Song and Jangwon Lee. Real-time traffic accident anticipation with feature reuse. *2025 IEEE International Conference on Image Processing*, 2025.

[33] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *CVPR*, pages 3521–3529, 2018.

[34] Vinh Tran, Yang Wang, Zekun Zhang, and Minh Hoai. Knowledge distillation for human action anticipation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2518–2522. IEEE, 2021.

[35] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *ICCV*, pages 13824–13835, 2023.

[36] 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, 2023.

[37] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *CVPR*, pages 13587–13597, 2022.

[38] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020.

[39] Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. Unsupervised traffic accident detection in first-person videos. In *IROS*, pages 273–280. IEEE, 2019.

[40] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE TPAMI*, 45(1):444–459, 2022.

[41] Xi Ye and Guillaume-Alexandre Bilodeau. Vptr: Efficient transformers for video prediction. In *2022 26th International conference on pattern recognition (ICPR)*, pages 3492–3499. IEEE, 2022.

[42] Hugo Yèche, Aliz'ee Pace, Gunnar Ratsch, and Rita Kuznetsova. Temporal label smoothing for early event prediction. In *International Conference on Machine Learning*, pages 39913–39938. PMLR, 2023.

[43] 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.

[44] Tianhao Zhao, Yongcan Chen, Yu Wu, Tianyang Liu, Bo Du, Peilun Xiao, Shi Qiu, Hongda Yang, Guozhen Li, Yi Yang, et al. Improving bird's eye view semantic segmentation by task decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15512–15521, 2024.

[45] Tianhao Zhao, Yiyang Zou, Zihao Mao, Peilun Xiao, Yulin Huang, Hongda Yang, Yuxuan Li, Qun Li, Guobin Wu, and Yutian Lin. Accident anticipation via temporal occurrence prediction. *NeurIPS*, 2025.