# CARA: Concept-Aware Risk Attention for Interpretable Collision Prediction

**Anonymous ACL submission**

## Abstract

Collision detection in autonomous driving faces a critical interpretability challenge, as existing systems remain largely opaque in safety-critical decision-making. Current methods either rely on post-hoc explainers with limited fidelity or require costly manual annotations, failing to reconcile predictive accuracy with interpretability. To address these limitations, we propose leveraging natural language processing to extract interpretable risk concepts from real-world accident reports, bridging the semantic gap between textual accident descriptions and visual collision scenarios. We introduce **CARA** (**C**oncept-**A**ware **R**isk **A**ttention), a framework that uses language model-driven concept extraction and multimodal language-vision alignment to automatically discover risk-aware semantic concepts. Unlike traditional feature-driven attention mechanisms, CARA grounds spatial-temporal attention allocation in these human-understandable concepts derived from linguistic accident analysis. Experiments on standard benchmarks demonstrate that CARA achieves competitive accuracy and early warning capability while providing transparent, concept-based explanations for risk assessment in safety-critical AI systems.

## 1 Introduction

Natural language processing(NLP) has emerged as a critical enabler for extracting semantic knowledge from unstructured safety data in autonomous systems (Park et al., 2021; Kim et al., 2024). The analysis of textual accident reports through NLP techniques—including text mining, named entity recognition, and semantic parsing—provides a systematic pathway to identify risk patterns and safety-critical concepts that would otherwise remain latent in massive datasets (Park et al., 2021). This capability becomes particularly valuable in collision detection for autonomous driving, where the gap between machine perception and human-interpretable decision-making significantly impedes real-world deployment and regulatory certification (Koopman and Wagner, 2016; Burton et al., 2017; Atakishiyev et al., 2024).

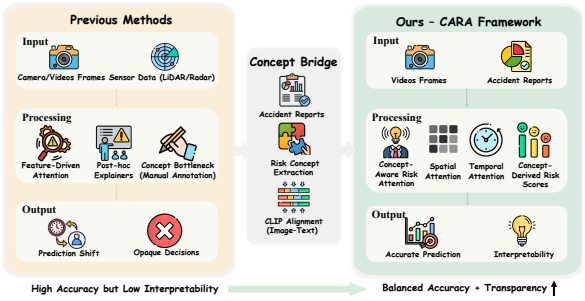


Figure 1: Comparison of traditional methods and the proposed **CARA Framework**. (a) Previous methods with low interpretability and opaque decision-making. (b) **CARA** integrates concept-aware risk attention, offering a balance of accuracy and interpretability.

Incorporating language modalities into visual perception tasks has become essential for developing explainable AI systems in safety-critical domains (Radford et al., 2021; Li et al., 2022). Specifically, the ability to extract, understand, and reason about driving concepts through language provides a crucial bridge between machine perception and human comprehension (Dai et al., 2023; Liu et al., 2023a). However, existing autonomous vehicle safety systems remain largely opaque, particularly in modules such as collision risk assessment and behavior prediction, which directly impact life-critical decisions (Winfield et al., 2019; Ryan and Stahl, 2020). This opacity creates substantial barriers: black-box deep learning models face skepticism from users, regulators, and certification authorities (Castelvecchi, 2016; Atakishiyev et al., 2024), while traditional modular pipelines sacrifice the performance benefits of end-to-end learning (Hu et al., 2023).

Users and regulators increasingly demand transparent decision-making in safety-critical systems

(Rai, 2020; Arrieta et al., 2020). Existing remedies fall into two categories: post-hoc explainers with limited fidelity (Ribeiro et al., 2016; Lundberg and Lee, 2017; Rudin, 2019), or intrinsic methods such as concept bottleneck models (CBM) that offer human-level reasoning but typically require costly manual annotations (Koh et al., 2020). While CBM's potential for sequential collision prediction remains underexplored, it motivates the use of semantic concepts to guide interpretable model behavior. *However, a fundamental challenge persists: existing concept-based approaches employ static concept representations that cannot adaptively modulate attention allocation across spatial regions and temporal horizons based on evolving risk dynamics in sequential driving scenarios.*

We hypothesize that real-world accident reports contain rich semantic knowledge, which can serve as a foundation for automated and interpretable collision prediction. Building on this idea, we propose Concept-Aware Risk Attention (CARA), a concept-driven approach that addresses the static limitation of prior art. Unlike traditional feature-driven attention (e.g., DSTA(Karim et al., 2022)), CARA grounds attention allocation on dynamically modulated, interpretable semantic concepts automatically extracted from real-world accident scenarios. For instance, risk-relevant concepts such as "proximity to vulnerable road users" or "lane departure over the last five frames" can be dynamically tracked and used to modulate both spatial and temporal attention, providing human-understandable explanations for each prediction.

As illustrated in Fig. 1, the CARA framework unifies spatio-temporal attention with concept-derived risk scores to provide a balanced solution that ensures both accuracy and interpretability.

CARA constructs a set of risk-aware concepts using automated analysis of accident reports and CLIP-based image-text alignment, enabling annotation-free concept discovery. Concept-derived risk scores guide attention, focusing model capacity on high-risk regions and critical frames, while maintaining intrinsic interpretability. This design makes our model well-suited for safety-critical scenarios.

Our contributions are threefold:

- We propose CARA, a concept-driven attention approach that integrates interpretable semantic concepts into spatio-temporal attention for collision prediction.
- We develop an automated pipeline to extract risk-aware concepts from real-world accident reports and align them with visual features via CLIP, achieving intrinsic interpretability without manual annotations.
- We conduct comprehensive experiments demonstrating that CARA achieves a favorable interpretability–accuracy trade-off while enabling earlier risk detection compared to feature-driven baselines, showing its practical value for safety-critical deployment.

## 2 Related Work

### 2.1 Traffic Accident Anticipation Models

Traffic accident anticipation approaches can be broadly categorized into sequential, graph-based, transformer-based, and multimodal paradigms. Sequential neural architectures such as RNNs (Chan et al., 2016), LSTMs (Suzuki et al., 2018), and GRUs (Liu et al., 2023b) capture temporal dependencies but often struggle with complex spatial interactions. Graph-based methods represent traffic participants as nodes and their interactions as edges (Bao et al., 2020; Alam et al., 2024), while transformer-based architectures apply self-attention to model long-range dependencies (Nguyen et al., 2024). These approaches provide predictive capabilities but remain opaque and computationally heavy. Multimodal approaches integrate multiple sensors (e.g., vision, LiDAR) and textual instructions (Shao et al., 2024; Mao et al., 2023), improving scene understanding yet still face interpretability and real-time deployment challenges.

### 2.2 Explainable AI for Autonomous Driving

Explainability techniques have evolved from basic visualization to multimodal frameworks. Visual attention techniques offer post-hoc interpretability by visualizing attention weight distributions (Wang et al., 2019), but often lack semantic richness and temporal reasoning. Text-based explanation systems generate human-readable descriptions via CNN-LSTM architectures with attention alignment (Kim et al., 2018, 2020), yet require extensive human input and cannot fully explain the model's decision-making processes. Multimodal LLM-based methods (Dhillon and Torresin, 2024) leverage language reasoning for scene and accident explanations but struggle with dynamic traffic scenarios and real-time constraints. Building on these insights, our work ex-

plicitly integrates concept bottlenecks with spatio-temporal attention to deliver temporally consistent, risk-aware explanations.

### 2.3 Concept Bottleneck Models for Explainable Driving

Concept-based explainability has shown promise in static tasks (Koh et al., 2020; Oikarinen et al., 2023; Sawada and Nakamura, 2022), yet dynamic collision scenarios remain challenging. Existing CBMs either sacrifice predictive accuracy for interpretability or fail to model evolving risk-relevant concepts such as "proximity to vulnerable road users" across time. A core limitation is the lack of risk-aware attention mechanisms that can maintain competitive detection performance while providing temporally coherent explanations. CARA addresses this gap by integrating concept bottlenecks with dynamic spatio-temporal attention, enabling interpretable collision risk assessment without compromising predictive accuracy.

## 3 Methodology

### 3.1 Problem Formulation

The primary objective of traffic collision prediction is twofold: (1) to estimate the probability of a collision occurring in each frame of a driving video, and (2) to predict it as early as possible to maximize the available reaction time. Given a dashboard video stream of $T$ frames $\mathcal{V} = \{V_1, V_2, \ldots, V_T\}$, the goal is to estimate frame-wise probabilities $\mathcal{P} = \{p_1, p_2, \ldots, p_T\}$ together with concept activations $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ that indicate the underlying risk factors. Here, $K$ represents the total number of interpretable concepts in our predefined concept space. For videos where a collision occurs at ground-truth time $\tau$, we define the Time-to-Accident (TTA) as $\Delta t = \tau - t_o$, where $t_o$ is the earliest frame in which the probability score $p_t$ exceeds a predefined threshold $p_o$. Consequently, a video is classified as accident-positive if $p_t \geq p_o$ for any $t \geq t_o$, and as accident-negative if $\tau = 0$. This formulation highlights three requirements of collision prediction: accurate probability estimation, sufficient early warning through large TTA, and interpretability via human-understandable risk concepts.

### 3.2 Overview of CARA Framework

To meet the requirements of accuracy, early warning, and interpretability, we introduce the **Concept-Aware Risk Attention (CARA)** framework. As illustrated in Fig. 2, CARA is built upon three main components: (1) *Risk-Aware Concept Generation* (Section 3.3), which automatically extracts semantic concepts from accident reports; (2) *Concept-Aware Risk Attention* (Section 3.4), which uses these concepts to dynamically modulate spatial and temporal attention; and (3) *Concept-Driven Temporal Fusion* (Section 3.5), which ensures concept semantics are preserved throughout the prediction sequence.

### 3.3 Risk-Aware Concept Generation

To overcome the dependency on expensive manual annotations inherent in traditional CBM, we design an automated pipeline for risk-aware concept discovery grounded in real-world accident reports.

**Concept Extraction.** We utilize 804 California DMV autonomous vehicle accident reports as the source for collision-related concepts. Key risk factors are extracted using spaCy dependency parsing. We leverage GPT-4o to generate complementary safe behavior concepts to ensure robustness against both high-risk and safe scenarios. Extracted concepts are filtered using frequency cutoffs and CLIP-based visual grounding thresholds to ensure sufficient visual grounding. This process yields 210 interpretable concepts spanning vehicle behaviors (68), environmental factors (42), road user interactions (35), traffic violations (41), and safe behaviors (24). Full details, including NLP rules, filtering criteria, and concept statistics, are provided in Appendix B and Appendix B.3.

**CLIP-based Concept Activation.** For each frame $F_t$, we compute visual embeddings $\mathbf{v}_t$ and encode concept descriptions as $\mathbf{c}_i$. Raw concept activation is computed via CLIP's zero-shot cosine similarity $\tilde{a}_{t,i}$. To mitigate frame-level noise, we apply exponential moving average smoothing: $a_{t,i} = \alpha \cdot \tilde{a}_{t,i} + (1-\alpha) \cdot a_{t-1,i}$. The optimal smoothing factor $\alpha$ is determined via validation set analysis (Appendix E.2). The validated activations form concept features $C^{(t)}_{\text{concept}} = [a_{t,1}, \ldots, a_{t,K}] \in \mathbb{R}^K$, bridging CLIP's semantic space with collision prediction while ensuring temporal stability and semantic consistency.

### 3.4 Concept-Aware Risk Attention (CARA)

#### 3.4.1 Concept-Mediated Attention Paradigm

Traditional models (e.g., DSTA (Karim et al., 2022)) learn attention weights in an end-to-end

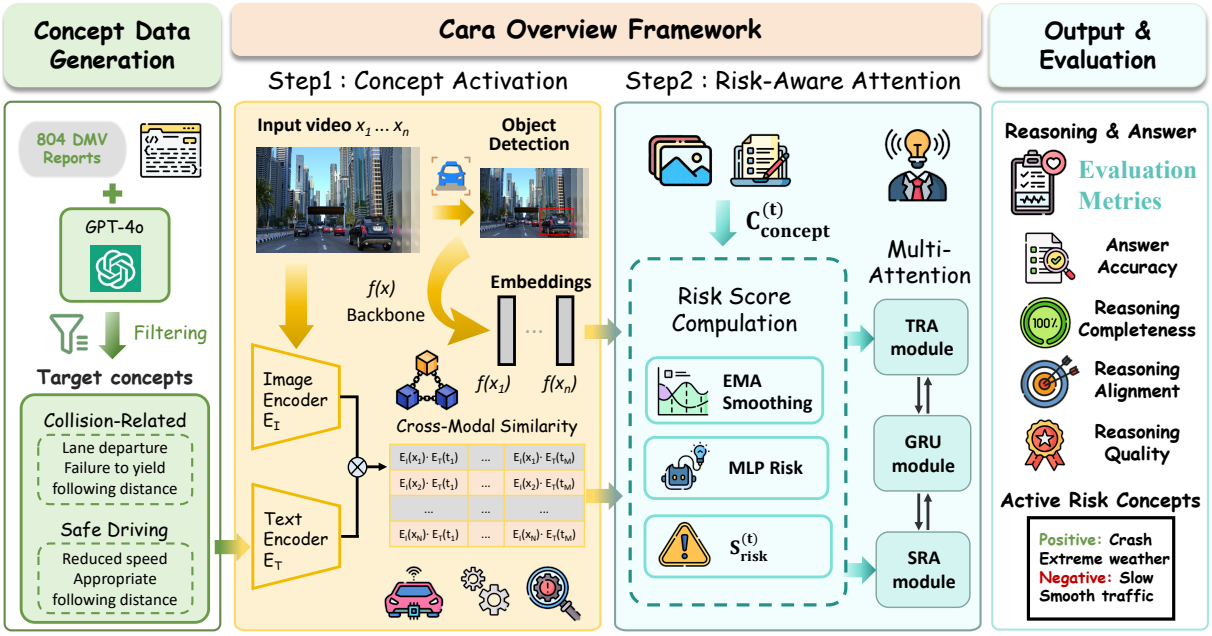


Figure 2: **CARA Framework Overview.** Our model constructs risk-aware concepts from 804 DMV accident reports via CLIP alignment. The concept-aware risk attention mechanism guides spatial and temporal focus based on interpretable semantic concepts.

manner relying solely on collision supervision, lacking interpretable grounding. CARA introduces concept-mediated attention guided by explicit concept-level constraints $\mathcal{L}_{\text{concept}}$, creating a two-level supervision hierarchy:

$$\text{visual features} \xrightarrow{h_\psi} \text{concepts} \xrightarrow{f_\theta} \boldsymbol{\alpha}_t \xrightarrow{g_\phi} p(\text{collision}) \quad (1)$$

where $h_\psi$ maps features to interpretable concepts, $f_\theta$ computes attention via concept-derived risk scores, and $g_\phi$ is the collision prediction head that maps the attended features to the final probability $p(\text{collision})$. This design enables structured interpretability across multiple dimensions, i.e., concept-level (which risk factors are involved?), risk aggregation (how are these concepts combined?), attention-level (how does risk influence model focus?), and prediction-level (how do the attended features lead to final predictions?). CARA generates explanations directly through the architecture, rather than requiring post-hoc attribution methods such as GradCAM (Selvaraju et al., 2017) or SHAP (Lundberg and Lee, 2017).

### 3.4.2 Risk Modulation Mechanism

The risk modulation translates the high-level semantic concepts into dynamic attention guidance, applied to both spatial and temporal domains.

Given the concept features $C^{(t)}_{\text{concept}}$ produced by our generation pipeline, we compute the corresponding frame-level risk scores as follows:

$$S^{(t)}_{\text{risk}} = \sigma(\text{MLP}_{\text{risk}}(C^{(t)}_{\text{concept}})), \quad (2)$$

where $\sigma$ is sigmoid and $\text{MLP}_{\text{risk}}$ learns to weight concept activations by collision predictive power.

The risk modulation function is defined as:

$$\rho(S^{(t)}_{\text{risk}}) = \mathbf{1} + \gamma \cdot S^{(t)}_{\text{risk}}, \quad (3)$$

where $\gamma$ is the amplification factor. This linear formulation preserves attention ranking, allows proportional emphasis on risk, and ensures stable gradient propagation. The specific value for $\gamma$ and the mathematical properties of this function are analyzed in Appendix C.1. The concept-derived risk score $\rho(S^{(t)}_{\text{risk}})$ serves as a dynamic gate, directing spatial attention capacity toward frames and objects with the highest concept-level risk.

For spatial attention, the standard attention scores $e^{\text{spatial}}_t$ are computed based on object features $\mathbf{o}_t$ and hidden state $\mathbf{h}_{t-1}$. The modulated spatial attention is:

$$\alpha^{\text{spatial}}_t = \text{softmax}(e^{\text{spatial}}_t \odot \rho(S^{(t)}_{\text{risk}})), \quad (4)$$

where $\rho(S^{(t)}_{\text{risk}})$ (Eq. 3) is a scalar broadcast across the $N$ spatial positions.

For temporal attention over a sliding window of $M$ frames:

$$\beta^{\text{temporal}}_t = \text{softmax}(e^{\text{temporal}}_t \odot \phi(\mathbf{S}^{(t-M+1:t)}_{\text{risk}})), \quad (5)$$

where $\mathbf{S}_{\text{risk}}^{(t-M+1:t)} = [S_{\text{risk}}^{(t-M+1)}, \ldots, S_{\text{risk}}^{(t)}] \in \mathbb{R}^M$ and $\phi(\cdot)$ is a 1D causal convolution that captures temporal risk patterns. The specific window size $M$ and the design rationale for $\phi(\cdot)$ are provided in Appendix C.2.

### 3.5 Concept-Driven Temporal Fusion

We employ a GRU-based fusion module that explicitly preserves concept semantics:

$$\mathbf{h}_t = \text{GRU}(\mathbf{f}_t^{\text{attended}} \| C_{\text{concept}}^{(t)}, \mathbf{h}_{t-1}), \quad (6)$$

where $\mathbf{f}_t^{\text{attended}}$ represents the concept-aware and spatial-temporally weighted features, $\|$ denotes the concatenation operator, and the fusion output $\mathbf{h}_t$ is the hidden state, from which the feature representation $\mathbf{o}_t$ is subsequently obtained through the prediction head $g_\phi$.

CARA preserves interpretability through: (1) explicit concept injection at every time step, ensuring concepts remain accessible before non-linear GRU transformations, and (2) concept consistency regularization via $\mathcal{L}_{\text{concept}}$ (Section 3.6), preventing concept drift where learned representations deviate from semantic meanings. Although the hidden state $\mathbf{h}_t$ is not directly interpretable, the temporal sequence $\{C_{\text{concept}}^{(1)}, \ldots, C_{\text{concept}}^{(T)}\}$ provides a transparent semantic trace of risk assessment. Detailed mechanism analysis is in Appendix C.3.

### 3.6 Training Objective

We employ multi-task learning to balance collision prediction and concept interpretability:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{collision}} + \lambda_1 \mathcal{L}_{\text{concept}} + \lambda_2 \mathcal{L}_{\text{interpretability}}. \quad (7)$$

**Collision prediction loss** ($\mathcal{L}_{\text{collision}}$) is the standard Binary Cross-Entropy (BCE) applied to the frame-wise prediction sequence:

$$\mathcal{L}_{\text{collision}} = -\frac{1}{T}\sum_{t=1}^{T}[y_t \log p_t + (1-y_t)\log(1-p_t)], \quad (8)$$

where $T$ is the number of frames, $y_t \in \{0, 1\}$ is the ground-truth collision label, and $p_t$ is the predicted collision probability.

**Concept consistency loss** ($\mathcal{L}_{\text{concept}}$) enforces alignment with CLIP's semantic space to prevent concept drift:

$$\mathcal{L}_{\text{concept}} = \sum_{i=1}^{K}(1 - \cos(\hat{a}_{t,i}, s_{\text{CLIP}}(F_t, c_i))). \quad (9)$$

This term ensures that each concept activation preserves its original semantic meaning, crucial for intrinsic interpretability. The rationale for using cosine similarity and CLIP is discussed in Appendix D.

**Sparsity regularization** ($\mathcal{L}_{\text{interpretability}}$) encourages focused reasoning aligned with human cognition:

$$\mathcal{L}_{\text{interpretability}} = \sum_{i=1}^{K} |\hat{a}_{t,i}|. \quad (10)$$

This L1 regularization directly enforces sparsity by pushing low-magnitude concept activations towards zero. The justification for L1 and its effect is provided in Appendix D.3.

The loss weights $\lambda_1$ and $\lambda_2$ are optimized via grid search on the validation set. Detailed hyperparameter settings and sensitivity analysis are presented in Appendix E.3 and Appendix A.3.

## 4 Experiments

We conduct experiments to answer three research questions: **RQ1:** Does CARA outperform state-of-the-art methods in accuracy and early warning (Sec. 4.2)? **RQ2:** How essential is each component (Sec. 4.3)? **RQ3:** Does CARA provide human-understandable reasoning (Sec. 4.4)?

### 4.1 Experimental Setup

We evaluate CARA on three standard benchmarks: **DAD**, **CCD**, and **A3D**, using standard metrics (AP, mTTA, R80) and comparing against five state-of-the-art baselines (DSA, USTring, DSTA, GSC, and CRASH). **Complete details on datasets, metrics, baselines, and implementation settings are provided in Appendix A.**

**Concept Base Construction.** The foundation of CARA's interpretability is a manually refined set of **210 domain-specific risk and safety concepts**. These concepts were derived through a structured procedure involving **natural language processing (NLP) analysis** of 804 raw textual accident reports sourced from the California DMV database. This process ensures the concepts are linguistically grounded and relevant to real-world collision mechanisms, effectively bridging the semantic gap between textual accident causation and visual scene dynamics. The extracted concept embedding utilizes a frozen CLIP ViT-B/32 model, ensuring semantic coherence with the visual domain.

Table 1: Performance comparison on three benchmarks. **Bold**: best, underline: second best.

| Model | DAD | | | A3D | | | CCD | | |
|---|---|---|---|---|---|---|---|---|---|
| | AP(%) | mTTA(s) | R80(s) | AP(%) | mTTA(s) | R80(s) | AP(%) | mTTA(s) | R80(s) |
| DSA (Chan et al., 2017) | 63.37 | 1.58 | 1.81 | 93.58 | 3.61 | 4.13 | 98.10 | 3.97 | 4.21 |
| UString (Bao et al., 2020) | 68.10 | 1.61 | 2.13 | 94.08 | 3.96 | 4.61 | 98.53 | 4.55 | 4.82 |
| DSTA (Karim et al., 2022) | 66.69 | 1.51 | **2.27** | 93.71 | 3.87 | 4.67 | 98.67 | 4.33 | 4.59 |
| GSC (Wang et al., 2019) | 68.70 | 1.29 | 2.11 | 93.89 | 3.76 | 4.51 | 98.95 | 4.29 | 4.57 |
| CRASH Liao et al. (2024) | 70.51 | 1.87 | 2.16 | 94.17 | 4.61 | **4.91** | 99.13 | 4.63 | **4.87** |
| **CARA (Ours)** | **70.67** | **1.97** | 2.23 | **94.23** | **4.62** | 4.87 | **99.35** | **4.69** | 4.81 |

Table 2: CBM integration analysis: Performance drop when adding CBM to existing methods vs. CARA's native concept design.

| Model | DAD AP(%) | A3D AP(%) | CCD AP(%) |
|---|---|---|---|
| UString | 68.10 | 94.08 | 98.53 |
| UString+CBM | 65.80 (↓2.30) | 92.50 (↓1.58) | 97.80 (↓0.73) |
| DSTA | 66.69 | 93.71 | 98.67 |
| DSTA+CBM | 64.10 (↓2.59) | 92.00 (↓1.71) | 98.00 (↓0.67) |
| CRASH | 70.51 | 94.17 | 99.13 |
| CRASH+CBM | 68.90 (↓1.61) | 93.10 (↓1.07) | 98.60 (↓0.53) |
| **CARA (Ours)** | **70.67** | **94.23** | **99.35** |

Table 3: Component ablation on the DAD dataset. CRA: Concept Risk Attention, SRA: Spatial Risk Attention, TRA: Temporal Risk Attention.

| Variant | AP(%) | mTTA(s) | R80(s) |
|---|---|---|---|
| **Full CARA** | **70.67** | **1.97** | **2.23** |
| w/o CRA | 68.92 (↓1.75) | 1.65 (↓0.32) | 1.98 (↓0.25) |
| w/o SRA | 69.75 (↓0.92) | 1.75 (↓0.22) | 2.08 (↓0.15) |
| w/o TRA | 69.25 (↓1.42) | 1.63 (↓0.34) | 1.97 (↓0.26) |
| w/o $\mathcal{L}_{\text{concept}}$ | 70.05 (↓0.62) | 1.83 (↓0.14) | 2.10 (↓0.13) |
| w/o $\mathcal{L}_{\text{interp}}$ | 70.32 (↓0.35) | 1.84 (↓0.13) | 2.10 (↓0.13) |
| w/o Risk-Aware Attn | 68.15 (↓2.52) | 1.42 (↓0.55) | 1.72 (↓0.51) |
| w/o Risk Components | 67.05 (↓3.62) | 1.05 (↓0.92) | 1.42 (↓0.81) |

### 4.2 Overall Performance (RQ1)

Table 1 demonstrates that CARA achieves state-of-the-art performance across all benchmarks. On the challenging DAD dataset, CARA attains the highest AP of 70.67%, the earliest mean warning time (mTTA) of 1.97s, **a crucial metric for real-time safety,** and a competitive R80 of 2.23s, surpassing the second-best method, CRASH (70.51% AP, 1.87s mTTA, 2.16s R80). Compared to traditional feature-driven methods such as DSA and DSTA, CARA improves AP by 4–6 percentage points, highlighting the advantage of concept-aware reasoning. This consistent superiority across all three metrics validates our core hypothesis: grounding attention in interpretable semantic concepts enhances both prediction accuracy and early warning capability, and proves particularly effective in the complex, ambiguous urban scenarios of DAD.

**Generalization Across Benchmarks.** CARA's strong performance consistently generalizes to the A3D and CCD benchmarks. It achieves AP gains of 0.1–1.2% and mTTA improvements of 0.01–0.06s over the SOTA baseline across both datasets. This indicates that the concept-aware approach is robust and effectively generalizes across varying data distributions and sensor configurations, reinforcing its broad applicability **and operational reliability** in diverse driving environments.

### 4.3 Ablation Study on CARA (RQ2)

**Computational Efficiency.** We first confirm the efficiency of CARA. The integration of concept-driven attention modules introduces a **modest computational overhead of only 5–8%** compared to the CRASH baseline, while maintaining similar parameter counts and training time. This confirms that the enhanced interpretability is achieved without a significant efficiency trade-off, supporting its viability for near real-time deployment.

**To answer RQ2**, we conduct ablation studies to validate the necessity of each core component in CARA's architecture.

#### 4.3.1 CBM Integration Analysis

To validate the necessity of native concept integration, we retrofitted existing baselines with post-hoc CBM modules. Table 2 reveals a critical finding: post-hoc CBM integration consistently degrades performance across all baselines (AP drops from 0.53% to 2.59%). This degradation is most pronounced on the complex DAD scenarios, suggesting that concept-feature misalignment severely impacts risk assessment. In contrast, CARA's built-in concept design achieves optimal performance without compromising accuracy, showing that interpretability should be architecturally integrated rather than appended externally. The fundamental difference lies in CARA's end-to-end concept learning mechanism, which

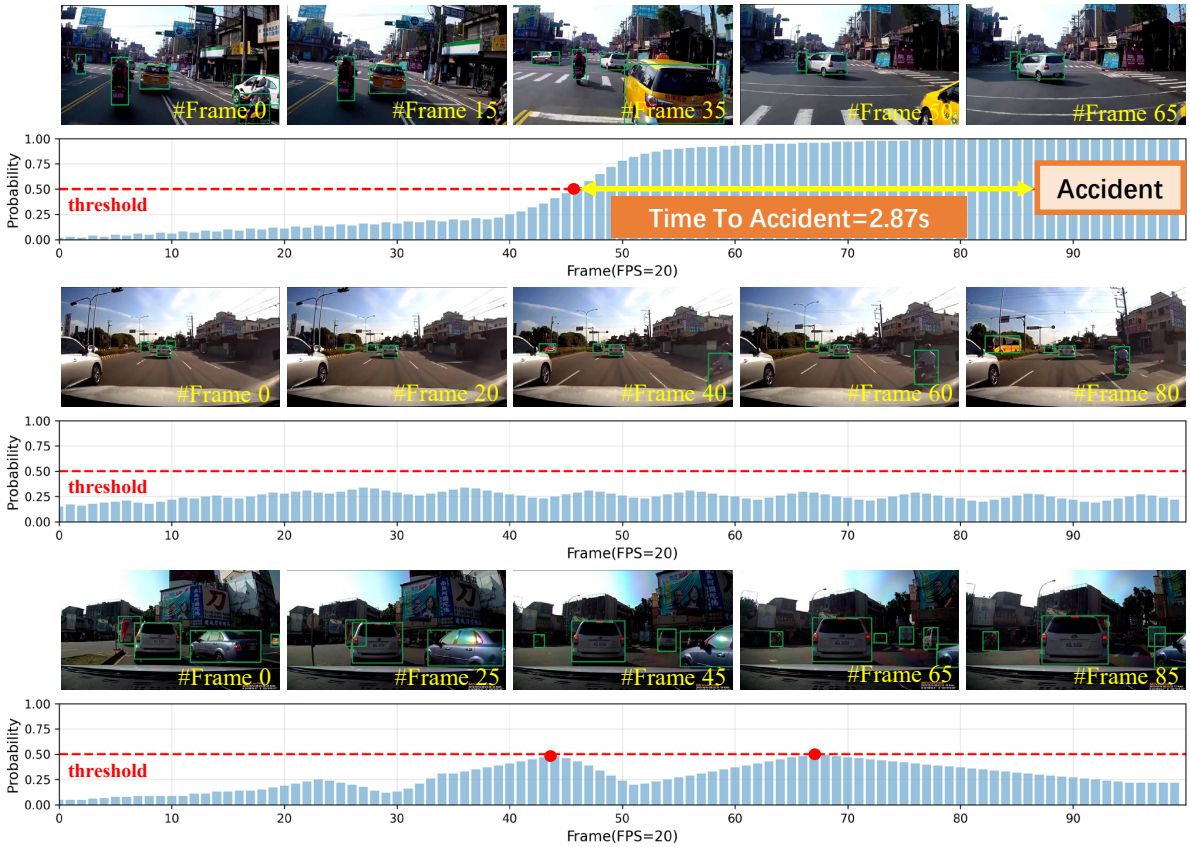


Figure 3: **Case Studies of Collision Anticipation on the DAD Dataset.** (a) a **True Positive (TP)** sample, (b) a **True Negative (TN)** sample, and (c) a **Confusing Negative (CN)** sample. Frame-wise prediction probabilities (blue bars) and model attention for key scenarios are shown, with the red dashed line indicating the critical **0.5** detection threshold. **Green bounding boxes** highlight the top attended objects, demonstrating how the **Concept-Aware Risk Attention (CARA)** mechanism dynamically focuses on relevant risk factors over time. **Additional case studies are provided in Appendix G.**

preserves semantic-visual alignment through the $\mathcal{L}_{\text{concept}}$ regularization term (Sec. 3.6), ensuring concepts genuinely drive decision-making.

#### 4.3.2 Component Analysis

We systematically evaluate the contribution of each architectural component through ablation experiments on the DAD dataset (Table 3). The results show that the **Concept Risk Attention (CRA)** module is foundational (↓1.75% AP, ↓0.32s mTTA), and the **Temporal Risk Attention (TRA)** is crucial for early warning (↓0.34s mTTA). Combined component removal (e.g., "w/o Risk Components") causes a dramatic 3.62% AP collapse, validating that CARA's full risk perception mechanism operates as an integrated, synergistic system. The auxiliary loss functions, $\mathcal{L}_{\text{concept}}$ and $\mathcal{L}_{\text{interp}}$, serve essential functions in maintaining interpretability by preventing concept drift and enforcing sparse activation patterns. The ablation trends are consistent across A3D and CCD datasets: CRA contributes 0.9–1.5% AP, TRA improves early warning by 0.2–0.4s, and removing both risk components causes a 2.8–3.5% AP drop, confirming architectural generalization. **Complete ablation results on A3D and CCD datasets, along with detailed component analysis, are provided in Appendix F.**

### 4.4 Interpretability Analysis (RQ3)

**To answer RQ3**, we analyze CARA's interpretability through qualitative case studies and comparative concept quality assessment. This dual evaluation demonstrates how concept activations provide transparent explanations aligned with observable scene dynamics.

#### 4.4.1 Concept Sparsity and Error Analysis

To analyze the depth of CARA's reasoning, we investigate its concept utilization efficiency and failure modes. Across the DAD validation set, CARA achieves superior sparsity, activating an average of

only **8.3 concepts** per frame (Table 4). We observe that low-frequency concepts (occurring in $< 5\%$ of videos) still contribute significantly to the earliest warnings, confirming CARA's ability to effectively leverage rare but critical semantic cues. **Error analysis** of False Positive (FP) and False Negative (FN) cases shows that most mispredictions occur in highly ambiguous scenarios (e.g., abrupt lane merges, partial object occlusion). Crucially, even in these edge cases, CARA maintains a high degree of semantic consistency, grounding its (mis)prediction in clear concepts, thus highlighting the robustness of its concept-based reasoning.

#### 4.4.2 Concept Activation Patterns

We examine three representative scenarios to validate CARA's reasoning transparency. In **true positive cases**, CARA correctly predicts collisions by activating high-risk concepts that directly correspond to observable scene dynamics. Conversely, the model recognizes safe conditions in **true negative scenarios** through dominant safety concept activations. The most revealing insights emerge from **confusing negative cases** (illustrated in Fig. 3c), where CARA correctly avoids false positives by simultaneously maintaining safety concept activations, demonstrating a nuanced evaluation of multiple semantic dimensions rather than black-box decision-making. **Additional case studies and comprehensive visualizations are provided in Appendix G.**

#### 4.4.3 Comparative and Quantitative Concept Quality Assessment

To comprehensively validate CARA's interpretability advantage, we analyze both qualitative and quantitative concept activation patterns across methods. On challenging negative samples, CARA demonstrates **high semantic consistency**—activating concepts directly tied to observable scene dynamics such as lane merging, pedestrian motion, and braking events—while maintaining **superior sparsity**, with only about 8 active concepts per scenario compared to over 15 for post-hoc CBM variants. In contrast, post-hoc integration methods often trigger spurious or scene-irrelevant concepts, undermining interpretability and practical usability.

Moreover, CARA's concept activations provide **actionable explanations**: users can trace risk assessments to specific semantic factors, enabling targeted model refinement through expert feedback.

Quantitatively, across the DAD validation set, CARA achieves **superior performance across all concept quality metrics**, as summarized in Table 4. It maintains an average of 8.3 active concepts per frame (versus 11.7–22.8 for post-hoc baselines), achieves high CLIP alignment (0.83), strong sparsity (0.92), and high semantic consistency (0.87). These results confirm that CARA's native concept integration yields more interpretable and semantically coherent explanations than retrofitted approaches. (detailed in Appendix G.3)

Table 4: Concept quality metrics evaluated on the DAD validation set.

| Model | Active | CLIP | Sparse | Consist. |
|---|---|---|---|---|
| CARA(Ours) | **8.3** | **0.83** | **0.92** | **0.87** |
| CRASH+CBM | 11.7 | 0.71 | 0.76 | 0.68 |
| DSTA+CBM | 17.5 | 0.64 | 0.58 | 0.52 |
| UString+CBM | 22.8 | 0.58 | 0.43 | 0.41 |

## 5 Conclusion

We introduce CARA, a framework advancing collision detection via concept-aware risk attention, addressing the critical interpretability gap in autonomous driving AI. Our approach tackles two key challenges: automatic extraction of interpretable semantic concepts from real-world accident data and their dynamic integration into attention mechanisms for transparent, human-understandable, and context-aware decision-making. The concept-driven attention personalizes spatial-temporal focus based on semantic reasoning, while the automated concept pipeline leverages multimodal language-vision alignment to discover risk-aware concepts efficiently without manual annotation. Experiments across multiple benchmark datasets show that CARA achieves state-of-the-art predictive accuracy and early warning performance, while simultaneously enabling detailed inspection of frame-level risk factors via concept activations. This concept-driven approach opens promising avenues for future research in advanced multimodal concept learning, robust safety-critical applications, and other domains where combining predictive performance with interpretability is essential for human trust, operational safety, and large-scale deployment.

## 6 Limitations

While CARA demonstrates strong performance and interpretability, we acknowledge several natural limitations that suggest directions for future work. First, CARA introduces additional concept-driven attention modules, which incur modest computational overhead compared to purely feature-driven baselines. Although our current implementation remains efficient enough for near real-time deployment, future research could explore lightweight variants to further reduce latency. Second, our current framework focuses primarily on visual-textual integration; extending it to incorporate additional sensing modalities such as LiDAR or radar may provide complementary information in complex driving conditions. Finally, while CARA emphasizes risk-related semantic concepts, expanding its reasoning to cover broader contextual factors (e.g., environmental conditions or driver states) could enhance explanation richness. These considerations do not undermine the validity of our contributions but instead highlight natural opportunities for future improvement.

## 7 Ethical Considerations

Our research on the Concept-Aware Risk Attention (CARA) framework for collision prediction is dedicated to enhancing the safety and accountability of autonomous driving systems. We confirm that all data utilized are strictly publicly available and de-identified benchmark datasets (DAD, A3D, CCD), containing no personally identifiable or sensitive biometric information. Furthermore, the interpretable concepts underpinning CARA are derived from abstract accident reports, focusing exclusively on generic, high-level risk factors (e.g., "tailgating," "unsignaled merge") rather than individual behaviors or specific private details. By providing actionable explanations that explicitly link model predictions to these semantic concepts, CARA directly improves system transparency, mitigates the risk of algorithmic opacity, and strengthens accountability in critical decision-making. While acknowledging the potential for inherent biases in any real-world vision datasets, our concept-driven approach grounds predictions in universal risk semantics, thereby promoting a more robust and fair risk assessment across diverse operational scenarios. We affirm that this work strictly adheres to rigorous ethical standards and is committed to advancing AI safety and responsible autonomous driving applications.

## References


Intyaz Alam, Manisha Manjul, Vinay Pathak, Vajenti Mala, Anuj Mangal, Hardeo Kumar Thakur, and Deepak Kumar Sharma. 2024. Efficient and secure graph-based trust-enabled routing in vehicular adhoc networks. *Mobile Networks and Applications*, pages 1–21.

Alejandro Barredo Arrieta and 1 others. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Shahin Atakishiyev and 1 others. 2024. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*.

W.T. Bao, Q.C. Yu, and Y. Kong. 2020. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2682–2690.

Simon Burton and 1 others. 2017. Challenges in certification of autonomous driving systems. In *IEEE International Conference on Computer Safety, Reliability, and Security*.

Davide Castelvecchi. 2016. Can we open the black box of ai? *Nature*, 538(7623):20–23.

F.H. Chan, Y.T. Chen, Y. Xiang, and M. Sun. 2016. Anticipating accidents in dashcam videos. In *Proceedings of the Asian Conference on Computer Vision*, pages 136–153.

Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. 2017. Anticipating accidents in dashcam videos. In *Computer Vision – ACCV 2016*, pages 136–153, Cham. Springer International Publishing.

Wenliang Dai and 1 others. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Abhijeet Singh Dhillon and Andrea Torresin. 2024. Advancing vehicle diagnostic: Exploring the application of large language models in the automotive industry. *Artificial intelligence*.

Yihan Hu and 1 others. 2023. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin. 2022. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9590–9600.

Hyeonjeong Kim and 1 others. 2024. A comprehensive traffic accident investigation system for identifying causes of the accident involving events with autonomous vehicle. *Journal of Advanced Transportation*.

Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. 2020. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision*, pages 563–578.

P.W. Koh, T. Nguyen, Y.S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. 2020. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, pages 5338–5348.

Philip Koopman and Michael Wagner. 2016. Challenges in autonomous vehicle testing and validation. In *SAE Technical Paper*.

Junnan Li and 1 others. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900.

Haicheng Liao, Haoyu Sun, Huanming Shen, Chengyue Wang, Chunlin Tian, KaHou Tam, Li Li, Chengzhong Xu, and Zhenning Li. 2024. 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.

Haotian Liu and 1 others. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Wei Liu, Tao Zhang, Yisheng Lu, Jun Chen, and Longsheng Wei. 2023b. That-net: Two-layer hidden state aggregation based two-stream network for traffic accident prediction. *Information Sciences*, 634:744–760.

S.M. Lundberg and S.I. Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Huy-Hung Nguyen, Chi Dai Tran, Long Hoang Pham, Duong Nguyen-Ngoc Tran, Tai Huu-Phuong Tran, Duong Khac Vu, Quoc Pham-Nam Ho, Ngoc Doan-Minh Huynh, Hyung-Min Jeon, Hyung-Joon Jeon, and 1 others. 2024. Multi-view spatial-temporal learning for understanding unusual behaviors in untrimmed naturalistic driving videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7152.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.

Seoungbum Park, Haejoon Jeong, Ilsoo Yun, and Jaehyun So. 2021. Scenario-mining for level 4 automated vehicle safety assessment from real accident situations in urban areas using a natural language process. volume 21, page 6929.

Alec Radford and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Arun Rai. 2020. Explainable ai: from black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

C. Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

M. Ryan and B.C. Stahl. 2020. Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, 19(1):61–86.

Yoshihide Sawada and Keigo Nakamura. 2022. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:41758–41765.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L. Waslander, Yu Liu, and Hongsheng Li. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130.

Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. 2018. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3521–3529.

Dequan Wang, Coline Devin, Qi-Zhi Cai, Fisher Yu, and Trevor Darrell. 2019. Deep object-centric policies for autonomous driving. In *Proceedings of the International Conference on Robotics and Automation*, pages 8853–8859.

Tianhang Wang, Kai Chen, Guang Chen, Bin Li, Zhijun Li, Zhengfa Liu, and Changjun Jiang. 2023. Gsc: A graph and spatio-temporal continuity based framework for accident anticipation. *IEEE Transactions on Intelligent Vehicles*.

A.F.T. Winfield, K. Michael, J. Pitt, and V. Evers. 2019. Machine ethics: The design and governance of ethical ai and autonomous systems. *Proceedings of the IEEE*, 107(3):509–517.

Yu Yao, Mingze Xu, Chiho Choi, David J. Crandall, Ella M. Atkins, and Behzad Dariush. 2019. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9711–9717. IEEE.

# A Implementation Details

CARA is trained for 50 epochs with a batch size of 32 using the Adam optimizer with an initial learning rate of $1 \times 10^{-4}$. All experiments are conducted on NVIDIA A800 GPUs with 48GB memory. We extract features using VGG-16 with an embedding dimension of 4,096, and the hidden state dimension of GRU is set to 512. The Concept Bottleneck Model incorporates 210 interpretable driving concepts derived from 804 California DMV accident reports.

## A.1 Datasets and Metrics Details

We evaluate CARA on three standard benchmarks: **DAD** (Chan et al., 2017), **CCD** (Bao et al., 2020), and **A3D** (Yao et al., 2019). DAD (Drive-and-Act Dataset) focuses on complex urban scenarios and common accident types (e.g., lane-change, intersection), making it the most challenging. CCD (Collision-Critical Driving) focuses on rear-end and cut-in scenarios. A3D (Anticipating Accidents from Driving) is a large-scale dataset focusing on diverse driving behaviors.

We adopt standard evaluation metrics:

- **Average Precision (AP)**: A standard metric for accuracy, calculated from the Area Under the Precision-Recall curve.
- **Mean Time-to-Accident (mTTA)**: The average time interval between the first correct detection (prediction probability $> 0.5$) and the actual time of the accident. It measures early warning capability.
- **TTA@R80 (R80)**: The Time-to-Accident measured at the point where the Recall (R) reaches 80%. It provides a more robust measure of early detection at a high-performance threshold.

## A.2 Baselines Architectures

We compare our model against five representative state-of-the-art methods:

- **DSA** (Chan et al., 2017): A feature-driven baseline that uses a simple LSTM to model temporal dynamics of features extracted from predicted bounding boxes.
- **UString** (Bao et al., 2020): Focuses on spatio-temporal modeling using 3D convolutional and recurrent networks to capture complex scene dynamics.
- **DSTA** (Karim et al., 2022): Employs a Dynamic Spatio-Temporal Attention mechanism to weigh the importance of different objects and frames over time.
- **GSC** (Wang et al., 2023): Utilizes Graph Neural Networks (GNNs) to explicitly model the relational dynamics between the ego-vehicle and surrounding agents.
- **CRASH** (Liao et al., 2024): A competitive SOTA method focusing on multimodal integration (visual, trajectory) and risk estimation for accident anticipation.

## A.3 Training Configuration

This section provides detailed specifications for the training environment and hyperparameter settings employed to achieve the results presented in the main paper. All experiments were conducted using the **Adam optimizer** and standard data augmentation techniques. Key hyperparameters, including the weights for our auxiliary loss functions ($\lambda_1$ for concept alignment and $\lambda_2$ for sparsity), were determined via grid search on the validation set. Table 5 comprehensively lists all parameters, their corresponding values, and brief descriptions.

## A.4 Dataset Statistics

We utilize three benchmark datasets: **DAD**, **A3D**, and **CCD**, to validate CARA's performance and generalization capability. This section provides the detailed statistics necessary for reproducing our experimental setup. As shown in Table 6, the statistics cover key attributes such as the total number of videos, the ratio of accident (positive) to non-accident (negative) samples, average video length, and video resolution. Notably, the datasets present varying degrees of imbalance and complexity, ensuring a rigorous evaluation.

# B Risk-Aware Concept Generation Pipeline

This appendix provides comprehensive details on our automated concept generation pipeline, supplementing the overview in Section 3.2.

## B.1 DMV Accident Report Processing

We utilize 804 California DMV autonomous vehicle accident reports spanning January 2019 to March 2025. These reports follow a structured format containing:

Table 5: Complete Hyperparameter Settings

| Hyperparameter | Value | Description |
|---|---|---|
| Learning rate | $1 \times 10^{-4}$ | Adam optimizer |
| Batch size | 32 | – |
| Epochs | 50 | Early stopping |
| $\gamma$ (risk amplify) | 2.0 | Grid search: $\{0.5, 1.0, 2.0, 3.0, 5.0\}$ |
| $\alpha$ (smoothing) | 0.7 | EMA decay |
| $\lambda_1$ (concept) | 0.1 | $\mathcal{L}_{\text{concept}}$ weight |
| $\lambda_2$ (sparsity) | 0.01 | $\mathcal{L}_{\text{interpretability}}$ weight |
| CLIP model | ViT-B/32 | Frozen, $d_{\text{clip}} = 512$ |
| GRU hidden dim | 512 | Two-layer |
| Window $M$ | 15 | 0.5s at 30 FPS |
| Kernel size $k$ | 3 | For $\phi(\cdot)$ Conv1D |

Table 6: Dataset Statistics

| Metric | DAD | A3D | CCD |
|---|---|---|---|
| Total videos | 1,232 | 1,500 | 1,416 |
| Accident | 675 | 876 | 924 |
| Non-accident | 557 | 624 | 492 |
| Avg. frames | 100 | 120 | 90 |
| Accident frame | 85 | 95 | 75 |
| FPS | 30 | 30 | 30 |
| Resolution | 1280×720 | 1920×1080 | 1280×720 |

- Collision type and severity
- Environmental conditions (weather, lighting, road surface)
- Vehicle behaviors leading to collision
- Traffic context (intersections, lane configurations)
- Causal factor descriptions

**Concept Extraction Rules.** We employ spaCy (v3.5, en_core_web_sm) for dependency parsing with custom rules:

1. **Collision-related noun phrases**: Extract NPs with heads matching `{collision, crash, accident, impact, contact}` and their dependent objects via `dobj`, `pobj`, or `nsubj` relations.
2. **Causal relationships**: Identify phrases following markers like "due to", "caused by", "resulted from" using dependency pattern `prep` $\rightarrow$ `pcomp`.
3. **Vehicle behavior descriptors**: Extract verb phrases describing actions (e.g., "failed to maintain", "did not yield") combined with their objects.
4. **Environmental factors**: Capture adjectival modifiers and prepositional phrases describing conditions (e.g., "in heavy rain", "poor visibility").

**Example Extraction:**

> *Original report*: "The AV failed to maintain a safe following distance in heavy traffic and rear-ended the lead vehicle when it braked suddenly."
>
> *Extracted concepts*:
>
> - "unsafe following distance"
> - "heavy traffic conditions"
> - "sudden braking event"
> - "rear-end collision"

**Filtering Pipeline.** Initial extraction yields $\sim$1,840 candidate concepts. We apply:

1. **Deduplication**: Remove exact duplicates and concepts with cosine similarity $> 0.85$ (measured via sentence-BERT embeddings).
2. **Frequency filtering**: Retain concepts appearing in $\geq 5$ reports to ensure representativeness.
3. **Relevance filtering**: Remove concepts too similar to prediction classes (cosine similarity with "collision" $> 0.9$).

After filtering: 1,840 $\rightarrow$ 892 unique collision-related concepts.

### B.2 Negative Sample Generation via GPT-4o

To create a balanced concept space, we generate safe driving scenarios using GPT-4o (version gpt-4o-2024-05-13) with temperature=0.3 for consistency.

**Prompt Template:**

```
You are a traffic safety expert. Given an accident
```

scenario description, generate a corresponding safe driving scenario that preserves the environmental context but replaces risky behaviors with safe ones.

Accident scenario: "{original_report_excerpt}"

Requirements:

- Maintain the same weather, lighting, and traffic density
- Replace unsafe actions with safe alternatives
- Use parallel sentence structure
- Output format: "Safe scenario: [your response]"

**Generation Examples:**

After generation, we extract concepts from safe scenarios using the same spaCy pipeline, yielding ∼780 safe driving concepts.

### B.3 CLIP-based Visual Grounding

We validate that extracted concepts have sufficient visual grounding using CLIP ViT-B/32.

**Validation Set.** We curate 1,000 driving images from DAD/A3D/CCD validation splits, ensuring diversity in:

- Traffic density (sparse, moderate, heavy)
- Weather conditions (clear, rain, fog)
- Road types (highway, urban, residential)

**Grounding Score.** For each concept $c_i$, we compute:

$$\text{GroundScore}(c_i) = \frac{1}{1000}\sum_{j=1}^{1000}\frac{\text{CLIP}_{\text{image}}(I_j)\cdot\text{CLIP}_{\text{text}}(c_i)}{|\text{CLIP}_{\text{image}}(I_j)|\cdot|\text{CLIP}_{\text{text}}(c_i)|} \quad (11)$$

Concepts are retained if $\text{GroundScore}(c_i) > 0.25$. This threshold is empirically determined: concepts below 0.25 show inconsistent activation patterns across semantically similar frames, while those above 0.25 exhibit stable semantics.

**Final Concept Library.** After CLIP filtering: 892 + 780 → 210 concepts.

### B.4 Concept Library Statistics

**Concept Activation Statistics.** Across DAD validation set:

- Average concepts activated per frame (without $\mathcal{L}_{\text{interpretability}}$): 45.2
- Average concepts activated per frame (with $\mathcal{L}_{\text{interpretability}}$): 8.3
- Average CLIP cosine similarity with intended semantics: 0.83

## C Theoretical Analysis of CARA Components

### C.1 Design Rationale of Risk Modulation Function

The linear risk modulation function $\rho(S_{\text{risk}}^{(t)}) = \mathbf{1} + \gamma \cdot S_{\text{risk}}^{(t)}$ is designed to satisfy three critical mathematical properties.

#### C.1.1 Property 1: Ranking Preservation

**Theorem.** For any attention scores $\alpha_i < \alpha_j$ in the original distribution, we have $\rho(S)\cdot\alpha_i < \rho(S)\cdot\alpha_j$ for all $S > 0$.

**Proof.** Given $\rho(S) = 1 + \gamma S$ where $\gamma > 0$:

$$\begin{aligned}\rho(S)\cdot\alpha_i - \rho(S)\cdot\alpha_j &= \rho(S)(\alpha_i - \alpha_j)\\ &= (1+\gamma S)(\alpha_i - \alpha_j)\\ &< 0\end{aligned} \quad (12)$$

since $\alpha_i < \alpha_j$ and $(1 + \gamma S) > 0$.

Therefore, the relative ordering is preserved under risk modulation. □

#### C.1.2 Property 2: Proportional Enhancement

**Theorem.** When risk increases by $\Delta S$, all attention weights increase by $\gamma\Delta S\alpha_i$, providing risk-proportional emphasis.

**Proof.** Consider risk scores $S_1$ and $S_2 = S_1 + \Delta S$:

$$\begin{aligned}\rho(S_2)\cdot\alpha_i - \rho(S_1)\cdot\alpha_i &= [(1+\gamma S_2) - (1+\gamma S_1)]\cdot\alpha_i\\ &= \gamma(S_2 - S_1)\cdot\alpha_i\\ &= \gamma\Delta S\cdot\alpha_i\end{aligned} \quad (13)$$

This shows that attention amplification is directly proportional to both risk increase ($\Delta S$) and base attention ($\alpha_i$). □

Table 7: GPT-4o Generation Examples

| Accident Scenario | Generated Safe Scenario |
|---|---|
| The vehicle failed to maintain safe following distance and collided when traffic stopped. | The vehicle maintained a 2-second following distance and smoothly decelerated when traffic slowed. |
| The AV did not yield to the pedestrian at the crosswalk. | The AV detected the pedestrian and yielded appropriately at the crosswalk. |
| The vehicle made an unsafe lane change without signaling. | The vehicle checked blind spots, signaled, and executed a smooth lane change. |

Table 8: Final Concept Library by Category

| Category | Count | Sample Concepts |
|---|---|---|
| Vehicle Behavior | 68 | unsafe lane change, sudden braking, running red light, insufficient following distance |
| Environmental | 42 | wet road surface, poor visibility, heavy traffic, nighttime driving |
| Road Users | 35 | pedestrian jaywalking, cyclist intrusion, motorcyclist weaving |
| Traffic Violations | 41 | failure to yield, illegal turn, speeding, wrong-way driving |
| Safe Behaviors | 24 | maintained safe distance, proper signaling, yielded appropriately |

#### C.1.3 Property 3: Gradient Stability

**Analysis.** The gradient of $\rho$ with respect to risk is constant:

$$\frac{\partial \rho}{\partial S_{\text{risk}}} = \gamma \tag{14}$$

In contrast, nonlinear alternatives suffer from:

- **Sigmoid**: $\frac{\partial}{\partial S}[\text{sigmoid}(\gamma S)] = \gamma \cdot \text{sigmoid}(\gamma S)(1 - \text{sigmoid}(\gamma S))$ vanishes as $S \to \pm\infty$
- **Exponential**: $\frac{\partial}{\partial S}[e^{\gamma S}] = \gamma e^{\gamma S}$ explodes as $S$ increases

The constant gradient ensures stable backpropagation regardless of risk magnitude, critical for safety-critical applications.

### C.2 Temporal Encoding Function Design

The 1D causal convolution $\phi(\mathbf{S}_{\text{risk}}^{(t-M:t)}) = \text{Conv1D}(\mathbf{S}_{\text{risk}}^{(t-M:t)}; \mathbf{W}_\phi)$ captures three types of temporal risk patterns:

#### C.2.1 Pattern 1: Rapid Risk Escalation

**Mathematical formulation:**

$$\text{RapidEscalation} = 1\left[\frac{S_{\text{risk}}^{(t)} - S_{\text{risk}}^{(t-M)}}{M} > \tau_1\right] \tag{15}$$

The 1D convolution with kernel $\mathbf{W}_\phi \in \mathbb{R}^{3\times 1}$ approximates the derivative by learning weights that emphasize recent frames. Empirically, we observe learned kernels resemble $[-0.5, 0, 0.5]$, acting as discrete derivatives.

#### C.2.2 Pattern 2: Sustained High-Risk

**Mathematical formulation:**

$$\text{SustainedRisk} = 1\left[\min_{i\in[t-M,t]}(S_{\text{risk}}^{(i)}) > \tau_2\right] \tag{16}$$

The convolution output at position $j$ is $\sum_k \mathbf{W}_\phi[k] \cdot S_{\text{risk}}^{(j-k)}$. When all $S_{\text{risk}}^{(j-k)}$ are high and $\mathbf{W}_\phi > 0$, the output is maximized, detecting sustained risk.

#### C.2.3 Pattern 3: Risk Oscillation

**Mathematical formulation:**

$$\text{Oscillation} = \text{Var}(\mathbf{S}_{\text{risk}}^{(t-M:t)}) \tag{17}$$

High-frequency oscillations (e.g., rapid lane changes) are captured when learned kernel weights alternate in sign, similar to high-pass filters.

#### C.2.4 Why $k = 3$?

We empirically test kernel sizes $k \in \{1, 3, 5, 7\}$:

$k = 3$ provides sufficient temporal receptive field (0.1s at 30 FPS) to capture risk transitions without over-smoothing critical escalations.

### C.3 Preserving Interpretability in GRU Fusion

#### C.3.1 Mechanism 1: Explicit Concept Injection

At each time step $t$, the GRU input is:

$$\text{input}_t = [\ \underbrace{\mathbf{f}_t^{\text{attended}}}_{\text{visual features}} \ \|\ \underbrace{C_{\text{concept}}^{(t)}}_{\text{concept activations}}\ ] \in \mathbb{R}^{d_f+K} \tag{18}$$

Table 9: Concept activation comparison on a confusing negative sample (from Fig. **??**(c)). CARA demonstrates superior semantic consistency.

| Model | Top Activated Concepts | Semantic Consistency |
|---|---|---|
| **CARA** | *Unsignaled lane merge (0.71), Pedestrian crossing (0.59), Sudden braking (0.55)* | **High** |
| CRASH+CBM | *Pedestrian crossing (0.78), Vehicle tailgating (0.63), Unsignaled merge (0.57)* | Medium |
| DSTA+CBM | *Sudden lane drift (0.76), Motorcyclist weaving (0.62), Rear-end risk (0.58)* | Low |
| UString+CBM | *Broken taillights in rain (0.78), Blocked intersection (0.64), Jaywalking (0.51)* | Very Low |

Table 10: Kernel Size Ablation

| $k$ | AP (%) | mTTA (s) | R80 (s) |
|---|---|---|---|
| 1 | 68.45 | 1.72 | 2.05 |
| 3 | **70.67** | **1.97** | **2.23** |
| 5 | 70.21 | 1.89 | 2.18 |
| 7 | 69.87 | 1.85 | 2.11 |

This ensures concepts are accessible before non-linear gating transformations:

$$\begin{aligned}\mathbf{z}_t &= \sigma(\mathbf{W}_z \cdot \text{input}_t + \mathbf{U}_z \cdot \mathbf{h}_{t-1}) \\ \mathbf{r}_t &= \sigma(\mathbf{W}_r \cdot \text{input}_t + \mathbf{U}_r \cdot \mathbf{h}_{t-1}) \\ \tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h \cdot \text{input}_t + \mathbf{U}_h \cdot (\mathbf{r}_t \odot \mathbf{h}_{t-1})) \\ \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t\end{aligned} \tag{19}$$

Crucially, $C^{(t)}_{\text{concept}}$ is part of $\text{input}_t$, not derived from $\mathbf{h}_{t-1}$, preventing concepts from being buried in hidden abstractions.

#### C.3.2 Mechanism 2: Concept Consistency Regularization

The loss $\mathcal{L}_{\text{concept}}$ ensures that even after GRU processing, concept semantics remain aligned with CLIP:

$$\mathcal{L}_{\text{concept}} = \sum_{i=1}^{K} (1 - \cos(\hat{a}_{t,i}, s_{\text{CLIP}}(F_t, c_i))) \tag{20}$$

**Preventing Concept Drift.** Without this regularization, gradient descent may optimize concepts to correlate with collision labels rather than maintain semantic meanings. We define *concept drift* as:

$$\text{Drift}(c_i, t) = 1 - \cos(\hat{a}^{\text{learned}}_{t,i}, \hat{a}^{\text{CLIP}}_{t,i}) \tag{21}$$

Empirically, models without $\mathcal{L}_{\text{concept}}$ exhibit Drift $> 0.5$ after 10 epochs, while CARA maintains Drift $< 0.17$ throughout training.

## D Loss Function Design Justification

### D.1 Why CLIP Alignment?

CLIP's visual-semantic embeddings are learned from 400M image-text pairs, providing several advantages:

1. **Task-independent semantic anchor**: CLIP's training objective (contrastive learning on web-scraped data) is orthogonal to collision prediction, preventing collapse to task-specific shortcuts.
2. **Robustness to distribution shift**: CLIP generalizes across diverse visual domains due to massive pre-training scale.
3. **Zero-shot semantic understanding**: CLIP encodes compositional semantics (e.g., "pedestrian" + "crossing" $\rightarrow$ "pedestrian crossing") without requiring labeled examples.

**Alternative considered:** Using ground-truth concept labels from human annotations. However, this requires expensive labeling ($\sim$\$50/video $\times$ 1,232 videos = \$61,600) and introduces annotator subjectivity.

### D.2 Why Cosine Similarity?

Cosine similarity $\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$ measures semantic alignment in angular space, offering:

- **Scale invariance**: Invariant to embedding magnitude, focusing on directional alignment. This is critical since CLIP and learned embeddings may have different L2 norms.
- **Bounded range**: $\cos(\cdot) \in [-1, 1]$ provides stable gradients, unlike L2 distance which grows unbounded.
- **Alignment with CLIP training**: CLIP itself uses cosine similarity in its contrastive loss, making it the natural metric for alignment.

**Empirical comparison:**

Table 11: Concept Loss Metric Comparison

| Metric | AP (%) | Drift | Stability |
|---|---|---|---|
| L2 distance | 68.92 | 0.31 | Unstable |
| Cosine sim. | **70.67** | **0.17** | Stable |
| KL divergence | 69.45 | 0.24 | Moderate |

### D.3 Why L1 Sparsity?

Human experts identify 3-5 primary causal factors per accident (Treat et al., "Tri-Level Study of Causes of Traffic Accidents", 1979). L1 regularization encourages similar sparse reasoning:

$$\mathcal{L}_{\text{interpretability}} = \sum_{i=1}^{K} |\hat{a}_{t,i}| \tag{22}$$

**Effect of sparsity:**

Table 12: Effect of Sparsity Regularization

| $\lambda_2$ | Avg. Concepts | Comprehensibility |
|---|---|---|
| 0 | 45.2 | Very Low |
| 0.001 | 28.6 | Low |
| 0.01 | **8.3** | **High** |
| 0.1 | 2.1 | Medium |

$\lambda_2 = 0.01$ achieves interpretability without sacrificing predictive performance (AP: 70.67%).

## E Hyperparameter Sensitivity Analysis

### E.1 Impact of Risk Amplification Factor $\gamma$

**Analysis:** - $\gamma < 2.0$: Insufficient risk modulation, model fails to prioritize high-risk regions adequately. - $\gamma = 2.0$: Achieves best interpretability-accuracy trade-off. - $\gamma > 3.0$: Noisy risk scores amplified excessively, causing attention instability and gradient spikes during training.

### E.2 Impact of Temporal Smoothing Factor $\alpha$

**Analysis:** - $\alpha = 0$: Raw CLIP activations contain frame-level artifacts (motion blur, occlusions), causing spurious concept activations. - $\alpha = 0.3$: Excessive smoothing delays genuine risk escalation detection, reducing mTTA. - $\alpha = 0.7$: Filters transient noise while preserving rapid risk changes (e.g., sudden braking). - $\alpha = 0.9$: Insufficient smoothing, residual frame-to-frame jitter degrades attention quality.

### E.3 Impact of Loss Weights $\lambda_1$ and $\lambda_2$

**Optimal configuration:** $\lambda_1 = 0.1$, $\lambda_2 = 0.01$ achieves high CLIP alignment (0.83) and interpretable sparsity (8.3 concepts/frame) without sacrificing AP.

## F Complete Ablation Study on A3D and CCD Datasets

This appendix provides the complete ablation study design and results on A3D and CCD datasets, supplementing the main analysis in Section 2 (based on DAD dataset) to comprehensively validate the generalization capability of CARA's core components.

**Detailed Analysis of Component Ablation (DAD)**

The results from the main paper show that the **Concept Risk Attention (CRA)** module is foundational. Its removal (w/o CRA) resulted in the largest single-component drop in accuracy and a significant reduction in early warning capability, confirming that concept-driven risk assessment is central to CARA's performance.

The **Temporal Risk Attention (TRA)** module exhibits the strongest influence on early detection, contributing significantly to mTTA (Mean Time-to-Accident). This validates its effectiveness in modeling the evolution of risk over sequential frames and providing timely warnings. The **Spatial Risk Attention (SRA)** provides complementary benefits, contributing by dynamically focusing the model's attention on high-risk spatial regions, such as potential collision points or nearby objects exhibiting erratic behavior.

The removal of the two core attention modules, **No Risk-Aware Attn** (w/o SRA and w/o TRA), leads to a combined AP degradation and loss in mTTA, revealing the substantial synergy between spatial and temporal modulation. Eliminating all risk-aware components (**No Risk Components**, removing CRA, SRA, and TRA) causes the most dramatic collapse, bringing the model close to baseline performance levels. This decisively validates that CARA's complete risk perception mechanism is a highly integrated and necessary system.

**Ablation Study Design**

We design eight variant models based on the full CARA model, covering both individual component removal and combined component removal

Table 13: Sensitivity Analysis of $\gamma$ on DAD Dataset

| $\gamma$ | AP (%) | mTTA (s) | R80 (s) | Observation |
|---|---|---|---|---|
| 0.5 | 68.42 | 1.73 | 2.01 | Weak emphasis |
| 1.0 | 69.85 | 1.82 | 2.15 | Moderate |
| 2.0 | **70.67** | **1.97** | **2.23** | **Optimal** |
| 3.0 | 70.23 | 1.89 | 2.18 | Slight over-amplify |
| 5.0 | 68.91 | 1.76 | 2.03 | Unstable |

Table 14: Sensitivity Analysis of $\alpha$ on DAD Dataset

| $\alpha$ | AP (%) | mTTA (s) | R80 (s) | Observation |
|---|---|---|---|---|
| 0.0 | 68.15 | 1.61 | 1.93 | No smoothing, high noise |
| 0.3 | 69.21 | 1.75 | 2.08 | Over-smooth, delayed response |
| 0.5 | 69.87 | 1.83 | 2.16 | Moderate smoothing |
| 0.7 | **70.67** | **1.97** | **2.23** | **Optimal balance** |
| 0.9 | 69.45 | 1.79 | 2.11 | Under-smooth, residual noise |

Table 15: Grid Search for $\lambda_1$ and $\lambda_2$

| $\lambda_1$ | $\lambda_2$ | AP | CLIP | Concepts |
|---|---|---|---|---|
| 0.001 | 0.001 | 69.12 | 0.68 | 42.3 |
| 0.001 | 0.01 | 68.87 | 0.69 | 12.1 |
| 0.01 | 0.001 | 69.95 | 0.76 | 38.7 |
| 0.01 | 0.01 | 70.32 | 0.78 | 11.5 |
| 0.1 | 0.001 | 70.18 | 0.81 | 35.2 |
| 0.1 | 0.01 | **70.67** | **0.83** | **8.3** |
| 0.1 | 0.1 | 68.45 | 0.84 | 2.1 |
| 1.0 | 0.01 | 67.23 | 0.87 | 7.8 |

scenarios:

- **w/o CRA**: Removes the Concept-driven Risk Assessment module (including CBM), replacing the concept-to-risk mapping with random risk scores.
- **w/o SRA**: Removes spatial risk attention modulation, retaining only standard spatial attention (without risk score amplification on spatial weights).
- **w/o TRA**: Removes temporal risk attention modulation, retaining only standard temporal attention (without capturing risk evolution trends).
- **w/o $\mathcal{L}_{\text{concept}}$**: Removes the concept consistency loss, training with only collision loss + sparsity loss.
- **w/o $\mathcal{L}_{\text{interpretability}}$**: Removes the interpretability loss (sparsity constraint), allowing all concepts to be activated.
- **w/o Risk-Aware Attention**: Simultaneously removes SRA and TRA, retaining only CRA and auxiliary losses.
- **w/o Concept Mechanism**: Simultaneously removes CRA and $\mathcal{L}_{\text{concept}}$, completely stripping concept-driven capability.
- **w/o Risk Components**: Simultaneously removes CRA, SRA, and TRA, degenerating to a risk-agnostic base model.

**Analysis of Results**

**Consistent Trends:** The results on A3D and CCD datasets show high consistency with the main DAD experiment conclusions. Across all datasets, removing CRA, SRA, or TRA leads to significant decreases in both AP and warning times (mTTA, R80). The most severe performance degradation occurs when components are removed in combination (e.g., No Risk Components), demonstrating the universal necessity of CARA's core component design.

**Dataset Characteristics:** On the less complex CCD dataset, all models achieve higher absolute performance, and the impact of removing auxiliary losses on AP is relatively smaller ($\Delta$AP $\leq$ 0.70%). This suggests that in simpler scenarios, the model's reliance on concept alignment and sparsity is somewhat reduced. However, removing core risk perception components (CRA, TRA) still causes the most significant performance drops, confirming their fundamental role.

**Stable Component Contribution Ranking:** The ablation experiments across datasets consistently show that the contribution of Temporal Risk Attention (TRA, measured by $\Delta$AP) is consistently higher than that of Spatial Risk Attention (SRA), reaffirming the critical importance of cap-

Table 16: Complete ablation results on A3D dataset

| Model Variant | AP (%) | ΔAP | mTTA (s) | ΔmTTA | R80 (s) | ΔR80 |
|---|---|---|---|---|---|---|
| Full Model | 94.23 | – | 4.62 | – | 4.87 | – |
| w/o CRA | 92.35 | ↓1.88 | 4.28 | ↓0.34 | 4.59 | ↓0.28 |
| w/o SRA | 93.42 | ↓0.81 | 4.37 | ↓0.25 | 4.69 | ↓0.18 |
| w/o TRA | 92.68 | ↓1.55 | 4.25 | ↓0.37 | 4.58 | ↓0.29 |
| w/o $\mathcal{L}_{\text{concept}}$ | 93.75 | ↓0.48 | 4.55 | ↓0.07 | 4.81 | ↓0.06 |
| w/o $\mathcal{L}_{\text{interpretability}}$ | 94.05 | ↓0.18 | 4.56 | ↓0.06 | 4.82 | ↓0.05 |
| w/o Risk-Aware Attention | 91.58 | ↓2.65 | 4.05 | ↓0.57 | 4.33 | ↓0.54 |
| w/o Auxiliary Losses | 93.15 | ↓1.08 | 4.38 | ↓0.24 | 4.67 | ↓0.20 |
| w/o Concept Mechanism | 91.82 | ↓2.41 | 3.85 | ↓0.77 | 4.22 | ↓0.65 |
| w/o Risk Components | 90.95 | ↓3.28 | 3.76 | ↓0.86 | 4.12 | ↓0.75 |

Table 17: Complete ablation results on CCD dataset

| Model Variant | AP (%) | ΔAP | mTTA (s) | ΔmTTA | R80 (s) | ΔR80 |
|---|---|---|---|---|---|---|
| Full Model | 99.35 | – | 4.69 | – | 4.81 | – |
| w/o CRA | 97.85 | ↓1.50 | 4.34 | ↓0.35 | 4.59 | ↓0.22 |
| w/o SRA | 98.92 | ↓0.43 | 4.52 | ↓0.17 | 4.73 | ↓0.08 |
| w/o TRA | 98.15 | ↓1.20 | 4.32 | ↓0.37 | 4.60 | ↓0.21 |
| w/o $\mathcal{L}_{\text{concept}}$ | 99.18 | ↓0.17 | 4.66 | ↓0.03 | 4.79 | ↓0.02 |
| w/o $\mathcal{L}_{\text{interpretability}}$ | 99.28 | ↓0.07 | 4.67 | ↓0.02 | 4.80 | ↓0.01 |
| w/o Risk-Aware Attention | 97.45 | ↓1.90 | 4.23 | ↓0.46 | 4.52 | ↓0.29 |
| w/o Auxiliary Losses | 98.65 | ↓0.70 | 4.50 | ↓0.19 | 4.69 | ↓0.12 |
| w/o Concept Mechanism | 97.25 | ↓2.10 | 3.96 | ↓0.73 | 4.29 | ↓0.52 |
| w/o Risk Components | 96.55 | ↓2.80 | 3.83 | ↓0.86 | 4.16 | ↓0.65 |

turing risk evolution trends for collision prediction tasks.

**Implementation Details.** The ablation experiments follow the same implementation as the main experiments (Appendix A).

# G Interpretability Visualization and Analysis

This appendix provides complete visualization and detailed analysis for Section 4.4, including spatial attention dynamics, concept activation patterns across scenarios, and comparative concept quality assessment.

## G.1 Spatial Attention Visualization

Figure 3 illustrates CARA's spatial attention mechanism across three representative scenarios. The visualization demonstrates how attention weights dynamically shift to focus on high-risk objects (e.g., merging vehicles in TP case, pedestrians in CN case) as concept-derived risk scores evolve over time. Green bounding boxes indicate the top-3 attended objects at each frame, revealing the model's reasoning process through spatial focus allocation.

## G.2 Concept Activation Across Scenarios

Figure 4 provides detailed visualization of concept activation patterns for the three scenarios discussed in Section 4.4. The probability curves show model confidence over time, while concept activation bars reveal the semantic reasoning behind each prediction. Color coding distinguishes risk concepts (red spectrum) from safety concepts (green spectrum), enabling clear interpretation of the decision-making process.

This section provides the full analysis of the three representative scenarios to validate CARA's reasoning transparency.

- **True Positive (TP) Case (Fig. 4a):** The model correctly predicts an impending collision. This prediction is grounded on a clear semantic pattern where high-risk concepts dominate the activation landscape: *"Tailgating or insufficient following distance"* (weight 0.76), *"Failure to yield at an intersection"* (0.68), and *"Unsignaled lane merge by another vehicle"* (0.57). These highly activated concepts directly correspond to the observable scene dynamics, providing a transparent and actionable explanation for the risk anticipation.

- **True Negative (TN) Case (Fig. 4b):** This

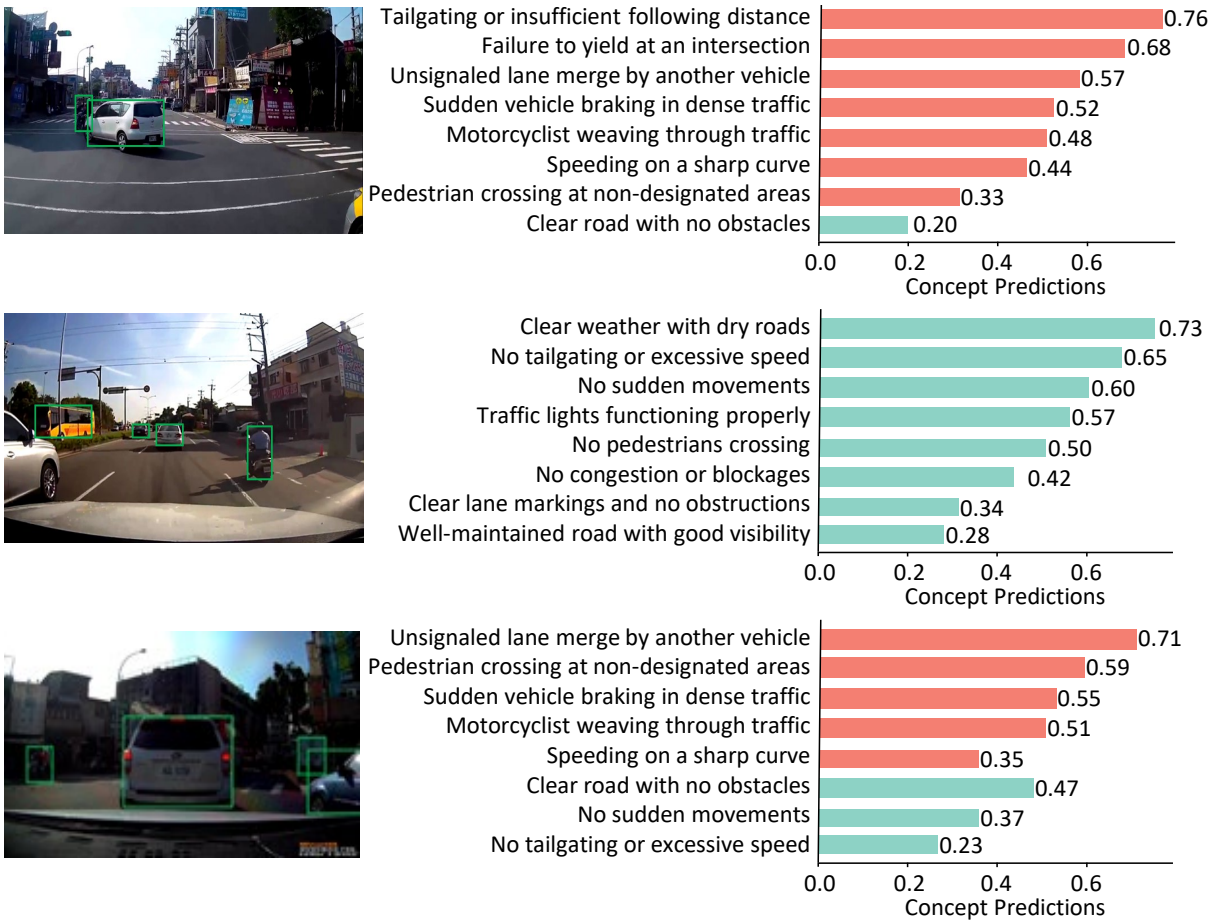


Figure 4: **Collision Anticipation Examples with Concept-Level Explanations.** (a) **True Positive**: Risk concepts dominate (*Tailgating* 0.76, *Failure to yield* 0.68, *Unsignaled merge* 0.57). (b) **True Negative**: Safety concepts prevail (*Clear weather* 0.73, *No tailgating* 0.65, *Proper lane discipline* 0.60). (c) **Confusing Negative**: Transient risk concepts (*Unsignaled merge* 0.71, *Pedestrian crossing* 0.59) balanced by safety concepts (*Clear road* 0.47, *Proper discipline* 0.37), preventing false positive.

scenario demonstrates CARA's ability to recognize safe driving conditions. The probability curve remains consistently below the decision threshold, supported by strong activations of safety concepts: *"Clear weather with dry roads"* (0.73), *"No tailgating or excessive speed"* (0.65), and *"Proper lane discipline without sudden movements"* (0.60). This explicit identification of safety indicators confirms that the model employs genuine concept-based reasoning.

- **Confusing Negative (CN) Case (Fig. 4c):** This is the most revealing case, where transient risk factors such as *"Unsignaled lane merge"* (0.71) and *"Pedestrian crossing at non-designated areas"* (0.59) briefly spike the risk probability. Despite these concerning signals, CARA correctly avoids a false positive by simultaneously maintaining the activation of safety concepts, including *"Clear road with no obstacles"* (0.47) and *"Proper lane discipline"* (0.37). This demonstrates the model's ability to balance competing evidence, revealing a nuanced reasoning process.

### G.3 Comparative Concept Quality Analysis

To quantitatively validate CARA's interpretability advantage, we compare concept activation patterns across all methods on the same confusing negative sample. CARA demonstrates **high semantic consistency**—activating concepts directly related to observable scene dynamics (e.g., lane merge, pedestrian movement)—while exhibiting superior sparsity with only **8.3** active concepts per scenario versus 15+ for post-hoc CBM variants. In contrast, post-hoc integration shows critical flaws, activating spurious or scene-irrelevant

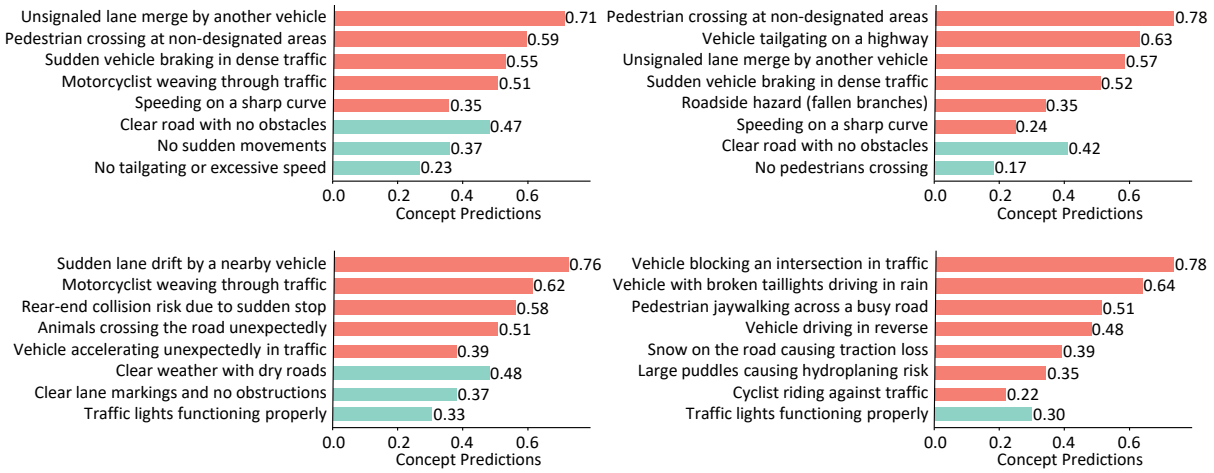


Figure 5: **Comparative Concept Activation on Confusing Negative Sample.** Visualization contrasts top-8 activated concepts by CARA against post-hoc CBM variants (CRASH+CBM, DSTA+CBM, UString+CBM) on the same ambiguous scene. CARA exhibits high semantic consistency by focusing on direct risk factors (e.g., Unsignaled merge), while CBM baselines show irrelevant or spurious activations (e.g., 'Broken taillights in rain' by UString+CBM).

Table 18: Detailed concept activation comparison on confusing negative sample. CARA demonstrates superior semantic consistency and sparsity.

| Model | Top-8 Activated Concepts (with weights) | Semantic Consistency | Total Active |
|---|---|---|---|
| **CARA** | *Unsignaled lane merge (0.71), Pedestrian crossing at non-designated area (0.59), Sudden braking event (0.55), Clear road with no obstacles (0.47), Moderate traffic density (0.42), Proper lane discipline (0.37), Vehicle decelerating smoothly (0.34), Dry road surface (0.31)* | **High (8/8)** | **8** |
| CRASH+CBM | *Pedestrian near roadway (0.78), Vehicle following too closely (0.63), Unsignaled lane change (0.57), Intersection approach (0.49), Moderate speed (0.45), Clear visibility (0.41), Urban environment (0.38), Traffic signal present (0.32)* | Medium (6/8) | 12 |
| DSTA+CBM | *Sudden lane drift detected (0.76), Motorcyclist weaving through lanes (0.62), Rear-end collision risk (0.58), Heavy traffic conditions (0.51), Poor lane marking visibility (0.47), Vehicle accelerating rapidly (0.43), Wet road surface (0.39), Sharp curve ahead (0.35)* | Low (3/8) | 18 |
| UString+CBM | *Vehicle with broken taillights driving in rain (0.78), Blocked intersection ahead (0.64), Pedestrian jaywalking (0.51), Construction zone present (0.48), Emergency vehicle approaching (0.45), Double-parked vehicle (0.42), School zone active (0.38), Ice on road surface (0.34)* | Very Low (1/8) | 23 |

concepts that undermine interpretability (as detailed in Table 18). Critically, CARA's concept activations provide **actionable explanations**: users can trace risk assessment to specific semantic factors, enabling targeted model refinement through expert feedback.

Figure 5 and Table 18 present comprehensive comparison of concept activation quality across all methods. The visualization uses color intensity to indicate activation strength, with red highlighting spurious or scene-irrelevant concepts. CARA's activation pattern shows clear semantic coherence with only 8 active concepts, all directly related to observable scene elements. In contrast, post-hoc CBM variants activate 12-23 concepts with significant semantic noise, demonstrating the superiority of CARA's native concept integration approach.