Journal of KIIT. Vol. 24, No. 6, pp. 125-138, Jun. 30, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.6.125

자율주행 위험 판단을 위한 증거 기반 설명형 추론 프레임워크(EGRR)

이세연*[1], 조신호*[2], 민경제*[3], 김주영*[4], 김건우**

EGRR: Evidence-Guided Risk Reasoning for Explainable Autonomous Driving

Se-Yeon Lee*[1], Sin-Ho Cho*[2], Gyeong-Je Min*[3], Ju-Young Kim*[4], and Gun-Woo Kim**

본 논문은 교육부와 경상남도의 재원으로 지원받은 경상남도 지역혁신중심 대학지원체계(RISE) 사업, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(RS-2026-25476256) 및 산업통상자원부 및 한국산업기술기획평가원(KEIT)의 지원을 받아 수행된 연구(RS-2025-02633048)의 결과임

요 약

본 연구는 자율주행과 같은 안전 필수 환경에서 관측 증거와 최종 결론 간의 연결성을 강화하기 위한 위험 추론 프레임워크를 제안한다. 제안한 방법은 도로 장면에서 검출된 객체 정보를 Evidence Builder를 통해 구조화된 관측 증거로 변환하고, 이를 바탕으로 CoT(Chain-of-Thought) 기반 중간 추론 단계에서 위험 판단 핵심 객체와 사실 기반 장면 설명을 순차적으로 생성한 뒤, 최종 위험도를 도출하도록 설계되었다. 이를 통해 모델은 단순히 위험 등급만 예측하는 데 그치지 않고, 어떠한 관측 근거를 바탕으로 해당 결론에 도달했는지를 함께 제시할 수 있다. 실험 결과, 객체 수준의 명시적 증거는 위험 설명 생성과 위험도 예측 성능 향상에 핵심적인 역할을 하였으며, 단계적 추론 구조는 최종 위험 판단의 일관성과 안정성을 높이는 데 기여하였다.

Abstract

This study proposes a risk reasoning framework designed to strengthen the connection between observed evidence and final conclusions in safety-critical environments such as autonomous driving. The proposed method transforms object information detected in road scenes into structured observational evidence through an Evidence Builder, and then uses this evidence within a Chain-of-Thought (CoT)-based intermediate reasoning process to sequentially generate decision-critical objects and a fact-grounded scene description before deriving the final risk level. In this way, the model does not merely predict a risk category, but also presents the observational basis on which the conclusion is reached. Experimental results show that explicit object-level evidence plays a key role in improving both risk description generation and risk level prediction, while the step-by-step reasoning structure contributes to enhancing the consistency and stability of final risk assessment.

Keywords

autonomous driving, risk reasoning, explainable AI, object-level evidence, chain-of-thought

* 경상국립대학교 컴퓨터공학부
- ORCID[1]: https://orcid.org/0009-0006-5365-4245
- ORCID[2]: https://orcid.org/0009-0003-4648-2711
- ORCID[3]: https://orcid.org/0009-0002-9184-2889
- ORCID[4]: https://orcid.org/0009-0006-9000-7258
** 경상국립대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-5643-4797

· Received: Mar. 26, 2026, Revised: Apr. 30, 2026 , Accepted: May 03, 2026
· Corresponding Author: Gun-Woo Kim
Dept. of Computer Science and Engineering, College of IT Engineering, Gyeongsang National University, Jinju, Korea
Tel.: +82-55-772-3323, Email: gunwoo.kim@gnu.ac.kr

Ⅰ. 서 론

자율주행 시스템에서 도로 장면의 위험을 정확하게 인지하고 적절히 대응하는 능력은 안전과 직결되는 핵심 요소이다. 실제 주행 환경에서는 차량, 보행자, 이륜차, 신호 체계, 교차로 구조와 같은 다양한 요소가 동시에 작용하며, 위험은 단순히 특정 객체의 존재 여부만으로 결정되지 않는다. 예를 들어 동일하게 보행자가 존재하더라도, 보행자의 위치, 이동 방향, 차량과의 상대적 거리, 교차로 여부, 야간 시인성 등의 맥락에 따라 위험 수준은 크게 달라질 수 있다. 따라서 자율주행 장면 이해는 단순 객체 인식을 넘어, 장면 내 여러 요소의 관계를 종합적으로 해석하여 현재 상황의 위험도를 판단하는 고차원적 추론 문제로 다루어져야 한다.

기존의 위험도 예측 방식은 주로 주행 이미지를 입력받아 위험 수준 또는 위험 관련 속성을 직접 분류하거나, 객체 검출, 장면 분류 등 복수의 시각 모듈을 결합하여 위험 단서를 추출하는 구조를 사용해 왔다[1]. 이러한 방식은 최종 위험 수준을 예측하는 데에는 유효하지만, 어떤 객체와 장면 요소가 위험 판단에 결정적으로 작용했는지를 자연어 수준에서 명확히 설명하지 못한다는 한계를 가지며, 이는 모델의 설명가능성을 저하시킨다. 이를 보완하기 위해 최근에는 VLM(Vision-Language Model)을 활용하여 도로 장면을 해석하고, 위험 판단의 근거를 자연어 형태로 제시하거나 단계적 추론 과정을 통해 의사결정을 설명하려는 연구가 활발히 이루어지고 있다[2]-[5]. 대표적으로 Reason2Drive와 DriveLM은 자율주행 장면에서 핵심 객체와 상황 맥락을 바탕으로 연쇄적 추론 과정을 구성함으로써 판단 근거를 보다 명시적으로 드러내고자 하였으며, DriveVLM과 X-Driver는 장면 설명, 분석, 계획을 통합한 설명 가능한 주행 프레임워크를 제안하였다.

그러나 이러한 접근은 주로 범용 장면 이해와 언어적 추론 능력에 의존하므로, 위험 판단의 핵심이 되는 객체의 종류, 위치, 상대적 거리와 같은 구체적 단서를 정밀하게 포착하는 데에는 한계가 있다. 실제로 VLM은 이미지 전반의 의미를 이해하고 이를 언어적으로 설명하는 데 강점을 보이지만, 객체의 존재 여부를 잘못 언급하거나 객체 수를 정확히 파악하지 못하는 문제가 보고되고 있으며[6][7], 이러한 특성은 안전이 중요한 자율주행 환경에서 위험 판단의 신뢰성을 저하시킬 수 있다. 따라서 자율주행 위험 추론에서는 언어적 추론 능력만으로는 충분하지 않으며, 객체 수준 정보를 구조화된 증거의 형태로 함께 활용하여 위험 판단의 근거를 명확히 보강할 필요가 있다.

본 연구는 이러한 문제의식을 바탕으로 EGRR (Evidence-Guided Risk Reasoner)를 제안한다. EGRR은 객체 탐지에 강점을 가진 YOLO 모듈을 활용하여 장면 내 핵심 객체 정보를 추출하고, 이를 요약하는 Evidence Builder를 통해 텍스트 기반의 구조화된 관측 증거로 변환한다. 이후 원본 이미지와 함께 해당 증거 텍스트를 VLM 입력에 결합함으로써, 모델이 명시적으로 정리된 객체 수준 근거를 함께 참고한 상태에서 위험을 추론하도록 구성하였다. EGRR은 최종 위험도만을 직접 예측하는 방식에 그치지 않고, 위험 판단 핵심 객체(Decision-critical objects)와 사실 기반 장면 설명(Factual scene description)을 선행적으로 생성하는 CoT(Chain-of-Thought) 기반의 중간 추론 단계를 포함한다. 이러한 설계는 설명이 단순한 사후 정당화에 머무르지 않고, 관측 근거를 단계적으로 축적하면서 결론에 이르는 구조적 위험 추론을 가능하게 한다. 특히 본 연구는 YOLO 기반 객체 검출 결과와 LiDAR 기반 거리 정보를 결합한 구조화된 객체 증거를 VLM의 위험 추론 과정에 직접 활용함으로써, 기존 VLM 기반 설명형 주행 연구에서 상대적으로 부족했던 객체 수준 근거의 명시성을 보강한다. 또한 다양한 VLM 백본 및 구성요소 제거 실험을 통해 객체 수준 증거와 CoT 기반 중간 추론이 위험 설명 생성과 위험도 예측 성능에 미치는 영향을 정량적으로 분석한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 관련 연구를 검토하고, Ⅲ장에서는 제안하는 증거 기반 설명형 추론 프레임워크를 설명한다. Ⅳ장에서는 실험 설정과 결과 분석을 제시하며, Ⅴ장에서는 결론을 정리한다.

II. 관련 연구

2.1 모듈형 주행 장면 분석 및 위험 단서 추출 연구

자율주행 장면 이해를 위한 초기 연구들은 주로 객체 검출, 장면 분류, 주행 가능 영역 분할과 같은 복수의 시각 모듈을 결합하여 주행 환경을 분석하는 모듈형 구조를 사용해 왔다. 이러한 접근은 장면 내 객체, 도로 구조, 기상 조건 등 다양한 시각 단서를 개별적으로 추출하고 이를 종합하여 주행 상황을 해석할 수 있다는 장점을 가진다. 예를 들어, Karim et al.[1]은 crash likelihood, road function, weather, time of day를 분류하는 Multi-Net과 객체 검출을 위한 YOLOv3, 주행 가능 영역 분할을 위한 DeepLab v3를 결합한 주행 장면 분석 시스템을 제안하였다. 해당 연구는 단일 이미지로부터 사고 가능성과 주행 장면 속성을 동시에 예측하고, 보행자나 근접 차량과 같은 위험 관련 객체를 탐지함으로써 충돌 위험 평가를 지원하고자 하였다. 또한 장면 분류, 객체 검출, 영역 분할 결과를 함께 활용함으로써 복잡한 도로 환경을 다각적으로 해석할 수 있음을 보였다. 이와 같은 구조는 복잡한 주행 장면에서 위험 관련 단서를 다양한 시각으로 추출할 수 있다는 점에서 의미가 있다.

그러나 이러한 모듈형 접근은 개별 시각 정보의 추출에는 효과적일 수 있으나, 최종적으로 왜 해당 장면이 위험하다고 판단되었는지에 대한 근거를 명시적인 추론 과정으로 제시하는 데에는 한계가 있다. 즉, 객체 검출 결과나 장면 분류 결과와 같은 중간 정보는 제공될 수 있으나, 이러한 정보가 어떤 방식으로 통합되어 최종 위험 판단으로 이어졌는지 설명하는 구조는 상대적으로 부족하다. 특히 기존 방식은 위험 단서를 분류 또는 검출 결과로 제시하는 데 초점이 있어, 객체 수준 증거가 최종 위험도 판단으로 연결되는 과정을 단계적으로 설명하는 데에는 제한적이다. 이러한 한계는 위험 판단의 설명 가능성을 요구하는 자율주행 환경에서 중요한 문제로 작용한다.

2.2 설명 가능한 연쇄 추론 기반 자율주행 연구

최근에는 VLM 및 MLLM(Multimodal Large Language Model)을 활용하여 자율주행 장면을 언어적으로 해석하고, 단계적 추론 과정을 통해 주행 의사결정을 지원하려는 연구가 활발히 이루어지고 있다. 이러한 흐름은 단순한 위험도 분류를 넘어, 인간 운전자의 판단 과정과 유사한 중간 추론 단계를 도입함으로써 자율주행 시스템의 해석 가능성을 높이려는 시도로 이해할 수 있다. 대표적으로 Reason2Drive[2]는 자율주행에서 해석 가능한 연쇄 추론의 필요성에 주목하여, 인지, 예측, 추론이 순차적으로 결합된 구조를 반영한 60만 개 이상의 비디오-텍스트 쌍 기반 벤치마크를 제안하였다. 이 연구는 자율주행 의사결정을 설명하는 CoT 데이터의 부족 문제를 지적하고, 다양한 기존 VLM의 추론 성능을 분석함으로써 자율주행 분야에서 해석 가능한 추론 연구의 중요성을 부각하였다.

OpenEMMA[8]는 MLLM 기반의 오픈소스 end-to-end 자율주행 프레임워크를 제안하였으며, CoT 추론을 도입하여 다양한 MLLM에서 baseline 대비 성능 향상을 보였고, 복잡한 주행 시나리오에서의 효과성과 일반화 가능성을 제시하였다. 또한 외부 visual expert를 통해 3D bounding box 정보를 결합함으로써, MLLM 단독으로는 충분히 반영하기 어려운 객체 수준의 공간 정보를 보완하고자 하였다. 그러나 이러한 연구들은 주로 일반적인 장면 이해, 주행 계획, 또는 연쇄적 의사결정 구조 자체에 초점을 두고 있어, 위험 판단이라는 특정 과제에 필요한 객체 중심 근거를 명시적으로 구조화하여 활용하는 데에는 한계를 보인다. 즉, 설명 가능한 추론 과정을 도입하더라도, 위험 판단에 직접적으로 중요한 객체 정보를 별도의 관측 증거 형태로 정리하여 활용하는 방식은 충분히 다루어지지 않았다. 따라서 자율주행 위험 판단에서는 단순한 설명 생성을 넘어, 객체 검출 기반의 구조화된 근거를 함께 활용함으로써 위험 판단의 근거를 객체 수준에서 보다 명확히 보강하고, 이를 바탕으로 위험도를 추론하는 접근이 필요하다.

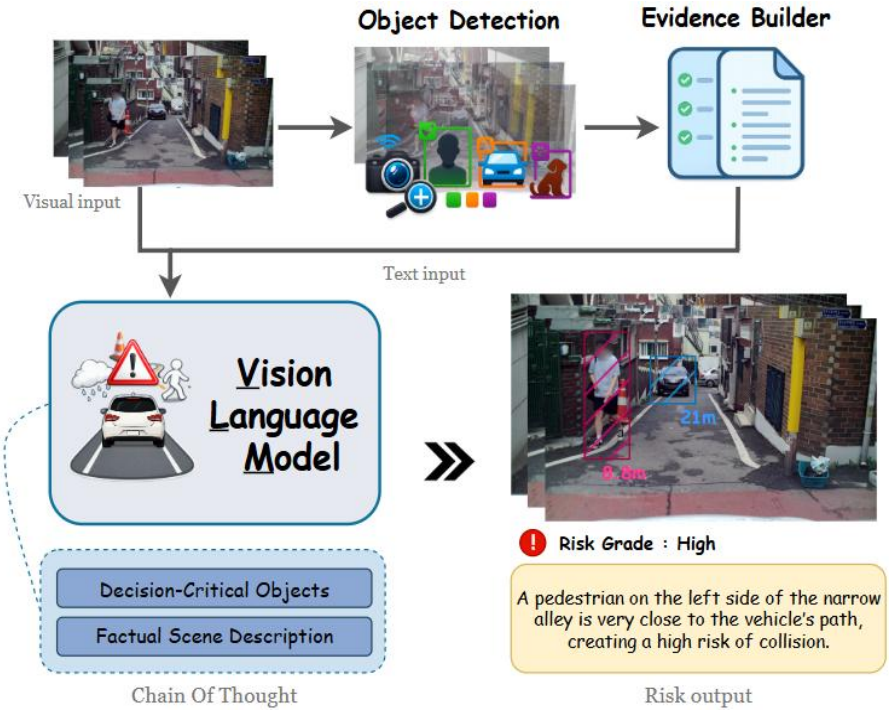

그림 1. 증거 기반 설명형 추론 프레임워크 아키텍처
Fig. 1. Architecture of the evidence-based explanatory reasoning framework

III. 자율주행 위험 판단을 위한 증거 기반 설명형 추론 프레임워크

3.1 제안 프레임워크 개요

본 연구의 전체 파이프라인은 그림 1과 같이 구성된다. 먼저 입력된 주행 장면 이미지에 대해 객체 검출 모델을 적용하여 보행자, 차량, 이륜차 등 위험 판단과 직접적으로 관련된 객체를 탐지한다. 다음으로 탐지된 객체 정보를 바탕으로 객체의 클래스, bounding box 좌표, 상대 거리, 객체 수를 정리한 구조화된 증거를 생성한다. 이렇게 생성된 증거는 원본 이미지와 함께 VLM의 입력으로 사용되며, 모델은 이를 바탕으로 위험 판단 핵심 객체와 사실 기반 장면 설명을 생성한 뒤, 최종 위험도 등급과 판단 근거를 출력한다. 3.2절에서는 Evidence Builder의 구성과 동작 방식을 설명하고, 3.3절에서는 CoT 기반 위험 추론 과정을 기술한다. 이어 3.4절에서는 최종 위험도 등급의 산정 기준과 출력 구조를 설명한다.

3.2 증거 생성기 (Evidence Builder)

본 연구는 객체 수준 정보를 보다 명시적으로 활용하기 위해, 객체 검출 결과를 구조화된 증거로 변환하여 VLM 입력에 함께 제공하는 Evidence Builder를 설계하였다. 먼저 입력 주행 장면에 YOLO 기반 객체 검출을 수행하여 보행자, 차량, 이륜차 등 도로 위험 판단과 직접적으로 관련된 객체를 탐지한다. 이후 각 검출 객체에 대해 bounding box 좌표를 추출하고, 동일 클래스 내 객체 수를 집계함으로써 장면 내 위험 요소의 위치와 분포를 구조적으로 표현한다. 또한 LiDAR 포인트 클라우드로부터 객체별 거리값을 추출한 뒤, 그 중앙값을 기준으로 자차와의 상대 거리를 산출한다. 이는 일부 이상치나 포인트 분포의 불균형으로 인한 거리 왜곡을 줄이기 위함이다. 이렇게 얻어진 객체의 종류, 개수, bounding box 좌표, 상대 거리 정보를 통합하여 텍스트 기반의 구조화된 근거를 구성하며, 특히 자차에 가까운 객체가 우선적으로 드러나도록 거리 기준으로 정렬한다. Evidence Builder의 구성 과정은 그림 2에 제시하였다. 최종적으로 구성된 Evidence

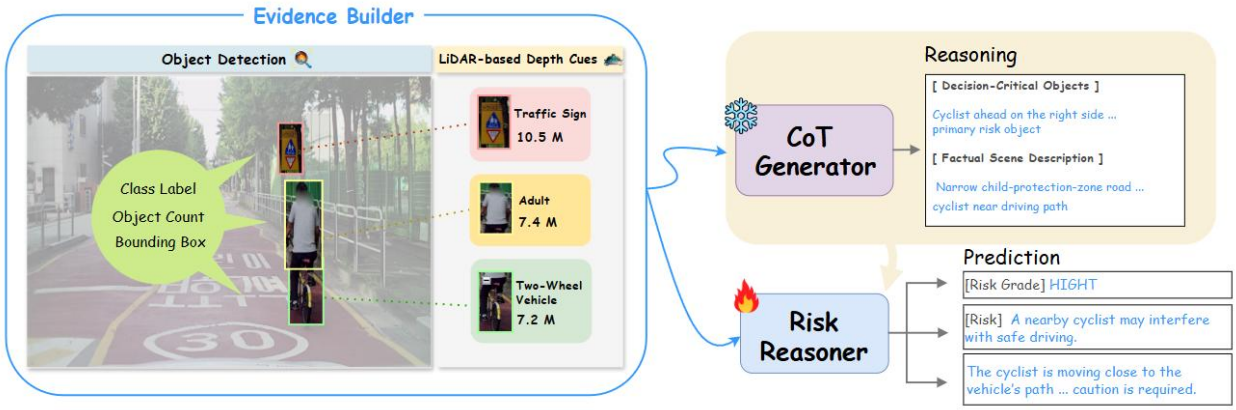

그림 2. 구조화된 객체 증거 기반 CoT 생성 및 위험 추론 학습 과정
Fig. 2. CoT generation and risk reasoning training process based on structured object evidence

Builder는 원본 이미지와 함께 VLM의 입력으로 사용되어, 모델이 장면의 전반적 시각 정보뿐 아니라 위험 판단에 직접적으로 관련된 객체 수준 근거를 함께 참고하도록 한다.

3.3 Chain-of-Thought 기반 위험 추론

기존 CoT 학습 연구에서는 사람이 직접 모든 중간 추론을 주석하는 대신, 성능이 우수한 외부 모델이 생성한 중간 추론 과정을 학습 신호로 활용하는 방식이 널리 사용된다. T-SciQ는 대형 언어모델이 생성한 CoT 추론 과정을 학습 신호로 사용하여 멀티모달 추론 모델을 학습하였고[9], 최근에는 GPT-4o가 생성한 추론 연쇄를 미세조정 데이터로 활용하여 CoT 능력을 강화한 바 있다[10]. 이에 본 연구는 사전학습된 외부 VLM을 이용해 입력 장면에 대한 Decision-Critical Objects와 Factual Scene Description을 먼저 생성하고, 이를 목표 위험 추론 모델이 학습할 CoT 정답 신호로 활용하였다. 이러한 구성은 모델이 최종 위험도를 직접 예측하도록 하는 대신, 먼저 장면 내에서 위험 판단에 직접적으로 영향을 미치는 핵심 객체를 식별하고, 이어 실제로 관측되는 도로 상황을 사실적으로 정리하도록 유도하기 위함이다. 구체적으로, Decision-Critical Objects는 자차의 주행 안전 판단에 직접적인 영향을 미치는 핵심 객체를 의미하며, 보행자, 차량, 이륜차, 신호등, 횡단보도 등 객체의 종류, 위치, 개수, 자차와의 거리 및 주행 경로와의 관계를 고려하여 정의된다. Factual Scene Description은 입력 이미지와 구조화된 객체 증거를 바탕으로 주요 객체의 존재 여부, 상대 거리, 도로 구조, 교차로 여부, 시인성, 날씨 및 시간대 등 관측 가능한 정보를 사실적으로 요약한 중간 설명이다. 이를 통해 모델은 위험 판단에 중요한 객체와 장면 맥락을 먼저 정리한 뒤, 사실 기반 이해를 바탕으로 최종 위험도를 판단하도록 학습된다. CoT 생성 및 학습 과정은 그림 2에 제시하였다.

3.4 위험도 산정 및 출력 구조

앞선 절에서 생성된 Decision-Critical Objects와 Factual Scene Description은 최종 위험도 판단을 위한 중간 근거로 활용된다. 본 연구에서는 자율주행 장면의 위험도를 LOW, MEDIUM, HIGH의 3단계로 산정하며, 최종 출력은 위험도 등급과 해당 판단 근거를 함께 포함하도록 구성하였다.

위험도 산정 기준은 도로 안전 평가 프로그램인 iRAP[12]의 위험 요인 관점을 참고하여 설계하였다. 구체적으로 취약한 도로 이용자(Vulnerable Road User, VRU)의 존재, 도로 구조, 차량 간 근접도, 시인성, 노면 상태, 규칙 이벤트 등을 주요 판단 축으로 설정하였다. 또한 객체의 단순 존재 여부만으로 위험도를 결정하지 않고, Evidence Builder에서 제공하는 객체 클래스, bounding box 위치, 객체 수, LiDAR 기반 상대 거리 정보를 함께 고려하였다. 각 판단 축과 주요 고려 요소는 그림 3에 제시하였다.

그림 3. 위험도 라벨링을 위한 판단 축과 주요 고려 요소
Fig. 3. Assessment axes and key considerations for risk labeling

이는 안전 필수 시스템에서 복수의 위험 요소를 하나의 판단으로 통합할 때 보수적인 기준을 적용하는 원칙과 맥락을 같이한다[13][14]. LOW는 핵심 위험 객체가 없거나 객체가 존재하더라도 자차로부터 충분히 이격되어 즉각적인 충돌 가능성이 낮은 장면으로 정의하였다. MEDIUM은 위험 객체가 존재하고 자차와의 상대 거리가 가까워 주의가 요구되지만, 회피 가능성 또는 통제 가능성이 남아 있는 장면으로 정의하였다. HIGH는 VRU 또는 차량이 자차에 근접해 있거나 주행 경로로 진입할 가능성이 높고, 교차로, 저시정, 복잡한 상호작용과 결합되어 즉각적인 감속이나 정지가 요구되는 장면으로 정의하였다.

Ⅳ. 실험 및 결과 분석

4.1 데이터 수집 및 구성

본 연구는 자율주행 위험 판단을 위한 데이터셋 구축을 위해 AI Hub 생활도로 데이터셋을 활용하였다[11]. 해당 데이터셋은 생활도로 환경에서 수집된 다양한 주행 장면을 포함하고 있으며, 보행자, 차량, 신호등, 횡단보도 등 위험 판단에 중요한 객체와 도로 요소를 폭넓게 포괄한다. 또한 날씨와 시간대에 대한 부가 라벨을 함께 제공하므로, 자율주행 위험 추론에서 중요한 환경적 맥락을 반영할 수 있다는 장점이 있다. 더불어 본 연구에서는 이미지 정보뿐 아니라 라이다 기반 좌표와 자차 기준 상대 거리 정보를 함께 활용하여, 주변 객체의 존재 여부뿐 아니라 자차와의 공간적 관계까지 반영할 수 있도록 하였다. 본 연구에서 사용한 데이터의 기본 통계량은 표 1과 같다.

표 1. 사용 데이터의 기본 통계량
Table 1. Basic statistics of the dataset used in this study

Category	Description
Dataset	AI Hub urban road object recognition autonomous driving dataset
Driving environment	Urban road driving scenes
Sample unit	Individual image frame
Size	3,001 images
Input	RGB images, LiDAR-based coordinates, ego-vehicle-relative distance
Object label	Pedestrians, vehicles, traffic lights, crosswalks, and other road objects

특히 이러한 데이터 구성은 객체의 종류와 위치 관계, 그리고 주변 환경을 종합적으로 반영한 위험 추론이 가능하다는 점에서 의미가 있다. 데이터 세부 클래스는 그림 4와 같이 정리하였다. 본 데이터는 Evidence Builder를 구성하기 위한 객체 검출기 학습과, 구조화된 객체 증거를 활용하는 VLM 학습에 공통으로 사용되었다.

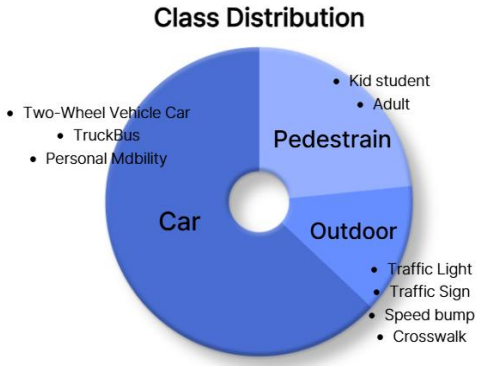

그림 4. 생활도로 데이터셋의 객체 클래스 구성
Fig. 4. Object class composition of the urban road dataset

4.2 실험 개요

Evidence Builder를 구성하기 위한 객체 검출 모듈로는 YOLOv8s와 YOLOv11s를 각각 학습하여 성능을 비교하였으며, 이를 통해 객체 수준 근거 생성에 적합한 검출기를 분석하였다. 이후 CoT 기반 중간 추론 데이터 생성을 위해 Qwen3-VL-8B-Instruct를 사용하였고, 위험 추론 모델의 비교 실험에는 Llama 3.2-11B-Vision-Instruct, LLaVA v1.6-Mistral-7B, Qwen2-VL-7B-Instruct를 활용하였다.

위험 추론 모델의 미세조정에는 4-bit 양자화 기반 QLoRA를 적용하였다[15]. LoRA 설정은 rank 16, alpha 32, dropout 0.05로 구성하였으며, 모델 입력은 원본 이미지와 Evidence Builder를 함께 사용하고 최대 64개의 객체 정보를 포함하도록 설정하였다. 학습 조건은 실험을 통해 최종 선정하였으며, 2 epoch, learning rate 2e-4, per-device batch size 1, gradient accumulation step 8을 적용하였다. 또한 최대 입력 길이는 4096으로 설정하고, 학습에는 bfloat16 정밀도를 사용하였다.

4.3 평가 지표

위험 추론 성능 평가는 위험도 등급 예측과 위험 설명 생성의 두 측면에서 수행하였다. 위험도 등급 예측 성능은 Accuracy, Macro-F1, MAE(ord)를 사용하여 평가하였다. Accuracy는 전체 샘플 중 예측 등급이 실제 등급과 일치한 비율을 의미하며, Macro-F1은 각 위험도 클래스별 F1-score를 평균한 값으로 클래스 간 성능을 균형 있게 평가하기 위해 사용하였다. MAE(ord)는 위험도 등급의 순서 정보를 반영하여 예측 등급과 실제 등급 간 차이를 평가하는 지표로, 값이 낮을수록 실제 위험도에 가까운 예측을 수행했음을 의미한다.

위험 설명 생성 성능은 BLEU, ROUGE-L, METEOR, CIDEr, SPICE, BERTScore를 사용하여 평가하였다. BLEU는 n-gram 일치도를 기반으로 생성 문장의 정확성을 평가하며[16], ROUGE-L은 최장 공통 부분수열을 활용하여 참조 문장과 생성 문장 간 유사도를 측정한다[17]. METEOR는 단어 정렬 기반 precision과 recall을 함께 고려하여 문장 간 의미적 일치를 평가한다[18]. CIDEr와 SPICE는 이미지 캡셔닝 평가에 널리 사용되는 지표로, 각각 참조 문장들과의 표현 합의 정도와 객체·속성·관계 기반의 장면 의미 구조 반영 정도를 평가한다[19][20]. BERTScore는 사전학습 언어모델의 문맥 임베딩을 활용하여 생성 문장과 참조 문장 간 의미적 유사도를 측정한다[21].

4.4 객체 검출기별 성능 비교 결과

객체 검출 성능이 전체 위험 추론 파이프라인에 미치는 영향을 분석하기 위해, YOLO 8s와 YOLO 11s를 대상으로 비교 실험을 수행하였다. 비교는 검출의 정확도를 나타내는 Precision, 실제 객체를 얼마나 놓치지 않고 검출하는지를 나타내는 Recall, IoU 0.5 기준의 전반적 검출 성능을 의미하는 mAP50, 그리고 보다 엄격한 IoU 구간에서의 평균 성능을 반영하는 mAP50-95를 기준으로 진행하였다. 표 2에 제시된 바와 같이, 두 모델의 Precision은 모두 0.77 수준으로 유사하게 나타났다. 반면 Recall은 YOLO 8s가 0.68, YOLO 11s가 0.69로, 11s가 소폭 높은 값을 보였다. mAP50 또한 YOLO 8s는 0.73, YOLO 11s는 0.74로 나타나 11s가 약간 우세하였다. mAP50-95에서도 YOLO 8s는 0.52, YOLO 11s는 0.55를 기록하여 유사한 경향 속에서 11s가 다소 높은 성능을 보였다.

종합하면, YOLO 11s는 YOLO 8s와 전반적으로 유사한 검출 성능을 보이면서도 Recall, mAP50,

mAP50-95에서 근소한 향상을 나타냈다. 그러나 두 모델 간 성능 차이는 크지 않아, 전체 파이프라인에서는 검출기의 미세한 성능 차이보다 이후 단계의 위험 추론 구조가 최종 성능에 더 중요한 영향을 미칠 가능성을 확인할 수 있었다.

표 2. YOLO 객체 검출기 성능 비교
Table 2. Performance comparison by YOLO detector

Version	Precision	Recall	mAP 50	mAP 50-95
8s	**0.77**	0.68	0.73	0.52
11s	**0.77**	**0.69**	**0.74**	**0.55**

4.5 VLM 비교 실험 결과

본 실험에서는 제안한 위험 추론 프레임워크에 여러 VLM 백본들을 적용하여 위험도 등급 예측 성능과 위험 설명 생성 성능 비교를 수행하였다. 이때 Evidence Builder를 위한 객체 검출기는 YOLO 11s 기반 모델을 공통적으로 사용하여, 백본 차이에 따른 영향을 보다 일관된 조건에서 비교하고자 하였다. 표 3에 제시된 각 모델의 base 결과를 보면, 전체적으로 Qwen2-VL-7B-Instruct와 Llama3.2-11B-Vision-Instruct가 유사하게 높은 성능을 보였으며, 두 모델 모두 LLaVA1.6-Mistral-7B 보다 전반적으로 우수한 결과를 나타냈다. 세부적으로 보면, Qwen2-VL-7B-Instruct는 BLEU 78.59, METEOR 0.827, CIDEr 7.514, SPICE 0.847로 가장 높은 성능을 기록하였고, Llama3.2-11B-Vision-Instruct는 ROUGE-L 0.900, BERT Score 0.969에서 근소하게 높은 값을 보였다. 반면 LLaVA1.6-Mistral-7B는 대부분의 설명 생성 지표에서 상대적으로 낮은 성능을 보였다. 이를 통해 Qwen2-VL과 Llama 계열은 위험 설명 생성에서 모두 강한 성능을 보였으나, LLaVA 계열은 상대적으로 설명 품질이 낮았음을 확인할 수 있다.

위험도 등급 예측 성능 또한 표 2에 함께 제시되어 있다 위험도 등급 예측에서는 Qwen2-VL-7B-Instruct가 Accuracy 0.855, Macro-F1 0.818, MAE(ord) 0.230으로 가장 우수한 성능을 기록하였다.

반면 Llama3.2-11B-Vision-Instruct는 Accuracy 0.801, Macro-F1 0.735, MAE(ord) 0.254을 기록하였고, LLaVA1.6-Mistral-7B는 Accuracy 0.786, Macro-F1 0.720, MAE(ord) 0.285로 세 모델 중 가장 낮은 성능을 보였다. 특히 Qwen2-VL은 정확도와 Macro-F1이 모두 가장 높고 MAE(ord)도 가장 낮아, 위험도 등급을 보다 안정적으로 예측하는 것으로 나타났다. Qwen2-VL-7B-Instruct는 위험 설명 생성 성능에서 최상위권을 유지하면서도, 위험도 등급 예측에서는 가장 우수한 결과를 보였다. 반면 Llama3.2-11B-Vision-Instruct는 설명 생성 품질은 우수하였으나 등급 예측 성능에서는 Qwen2-VL에 미치지 못하였고, LLaVA1.6-Mistral-7B는 두 측면 모두에서 상대적으로 낮은 성능을 보였다.

종합하면, 세 VLM 백본은 위험 설명 생성과 위험도 등급 예측에서 서로 다른 특징을 보였으며, 설명 생성 품질과 등급 예측 성능이 항상 동일한 경향을 보이지는 않았다. 특히 설명 생성 지표가 높다고 해서 위험도 등급 예측 성능까지 반드시 우수한 것은 아니었으며, Qwen2-VL-7B-Instruct가 전반적으로 가장 균형 잡힌 성능을 보였다.

그림 5는 Qwen2-VL-7B-Instruct의 테스트 결과 예시를 보여준다. 각 사례에 대해 입력 주행 장면, 정답 위험도, 예측 위험도, 그리고 모델의 CoT 추론 내용을 함께 제시하였다. 전반적으로 위험도 등급은 정답과 유사하게 예측되었으며, 장면 내 주요 객체와 위험 요인을 비교적 타당하게 설명하는 경향을 확인할 수 있었다. 특히 모델은 보행자, 전방 차량, 시인성 저하와 같은 핵심 위험 단서를 중심으로 장면을 해석하며, 이러한 단서들을 바탕으로 최종 위험도를 도출하는 모습을 보여준다.

4.6 제안 방법의 구성요소 효과 분석

본 절에서는 제안한 프레임워크의 핵심 구성요소인 Evidence Builder와 CoT 기반 위험 추론이 최종 성능에 미치는 영향을 분석하기 위해 구성요소 제거 실험을 수행하였다. 모든 실험은 동일한 학습 조건하에서 각 구성요소만 선택적으로 제거하는 방식으로 구성하여, 각 요소의 기여도를 직접적으로 비교하였다. 표 3에 제시된 바와 같이, 구성요소 제거에 따른 성능 변화는 세 가지 VLM 백본 전반에서 유사한 경향을 보였다.

표 3. 기준 모델, VLM 백본 및 구성요소 제거 설정에 따른 위험 추론 성능 비교
Table 3. Comparison of risk reasoning performance across the baseline model, VLM backbones, and ablation setting

Model	BLEU (↑)	METEOR (↑)	R-L (↑)	CIDEr (↑)	SPICE (↑)	BERT F1 (↑)	Acc (↑)	Macro F1 (↑)	MAE (↓)
Multi-Net									
Multi-Net	-	-	-	-	-	-	0.632	0.460	0.562
LLaVA1.6-Mistral-7B									
w/o Evidence	60.753	0.604	0.713	5.211	0.589	0.902	0.429	0.312	0.571
w/o CoT	**76.180**	**0.813**	**0.838**	**6.709**	**0.802**	**0.963**	0.680	0.644	0.320
EGRR	75.281	0.769	0.837	6.574	0.776	0.943	**0.786**	**0.720**	**0.285**
Llama3.2-11B-Vision-Instruct									
w/o Evidence	60.312	0.619	0.849	5.058	0.312	0.905	0.531	0.392	0.547
w/o CoT	**77.985**	**0.822**	**0.932**	7.418	**0.850**	**0.988**	0.767	0.674	0.290
EGRR	77.546	0.819	0.900	**7.429**	0.844	0.969	**0.801**	**0.735**	**0.254**
Qwen2-VL-7B-Instruct									
w/o Evidence	63.312	0.669	0.808	5.758	0.699	0.935	0.669	0.491	0.492
w/o CoT	**78.601**	**0.834**	**0.903**	**7.552**	**0.853**	**0.967**	0.752	0.668	0.247
EGRR	78.590	0.827	0.899	7.514	0.847	0.965	**0.855**	**0.818**	**0.230**

먼저, Evidence Builder를 제거한 경우에는 모든 백본에서 위험 설명 생성 성능과 위험도 등급 예측 성능이 모두 뚜렷하게 하락하였다. 예를 들어 Qwen2-VL-7B-Instruct에서는 BLEU가 78.59에서 63.31로 감소하였고, CIDEr는 7.514에서 5.758, SPICE는 0.847에서 0.699로 낮아졌다. 위험도 등급 예측에서도 Accuracy가 0.855에서 0.669로, Macro-F1이 0.818에서 0.491로 크게 하락하였으며, MAE(ord)는 0.230에서 0.492로 증가하였다. 이러한 패턴은 LLaVA-v1.6-Mistral-7B와 Llama-3.2-11B-Vision-Instruct에서도 공통적으로 나타났으며, 이는 객체의 종류, 거리, 개수와 같은 구조화된 명시적 증거가 위험 설명 생성과 위험도 판단 모두에 핵심적인 역할을 함을 보여준다.

반면, CoT를 제거한 경우에는 설명 생성 성능은 대부분의 백본에서 전체 모델과 유사하거나 일부 지표에서 소폭 향상되는 경향을 보였으나, 위험도 등급 예측 성능은 세 모델 모두에서 일관되게 하락하였다. Qwen2-VL-7B-Instruct에서는 BLEU가 78.59에서 78.60으로 거의 차이가 없었고, METEOR, ROUGE-L, CIDEr, SPICE, BERTScore에서도 미세한 향상이 나타났다. 그러나 위험도 등급 예측에서는 Accuracy가 0.855에서 0.752로, Macro-F1이 0.818에서 0.668로 감소하였고, MAE(ord) 역시 0.230에서 0.247로 증가하였다. LLaVA 및 Llama 계열에서도 동일하게 설명 생성 지표의 변화는 제한적이었지만 등급 예측 성능은 공통적으로 저하되었다. 이는 CoT 기반 중간 추론이 설명 문장의 표면적 품질 자체를 크게 높이기보다는, 최종 위험도 등급을 보다 안정적이게 판단하도록 유도하는 데 기여함을 시사한다.

종합하면, Evidence Builder는 제안 모델의 성능을 지탱하는 핵심 구성요소이며, CoT 기반 위험 추론은 특히 위험도 등급 예측의 안정성과 일관성을 향상시키는 보조적 요소로 작용하였다. 또한 세 가지 VLM 백본 전반에서 공통적으로 확인되었듯이, 본 연구의 프레임워크는 객체 수준의 구조화된 증거와 단계적 추론 구조가 결합될 때 가장 우수한 성능을 보였다. 이는 제안한 방법이 단순히 위험도를 분류하는 데 그치지 않고, 객체 기반 증거를 바탕으로 판단 근거를 보다 명확하게 제시할 수 있는 설명형 위험 추론 프레임워크로서 유효함을 보여준다.

4.7 구조화된 객체 증거 요소별 기여도 분석

본 절에서는 Evidence Builder가 생성하는 세부 증거 요소가 위험도 예측 성능에 미치는 영향을 분석한다. 앞선 4.6절에서는 Evidence Builder 전체를 제거했을 때 성능이 크게 하락함을 확인하였으며, 이에 본 절에서는 객체 수, bounding box 좌표, LiD

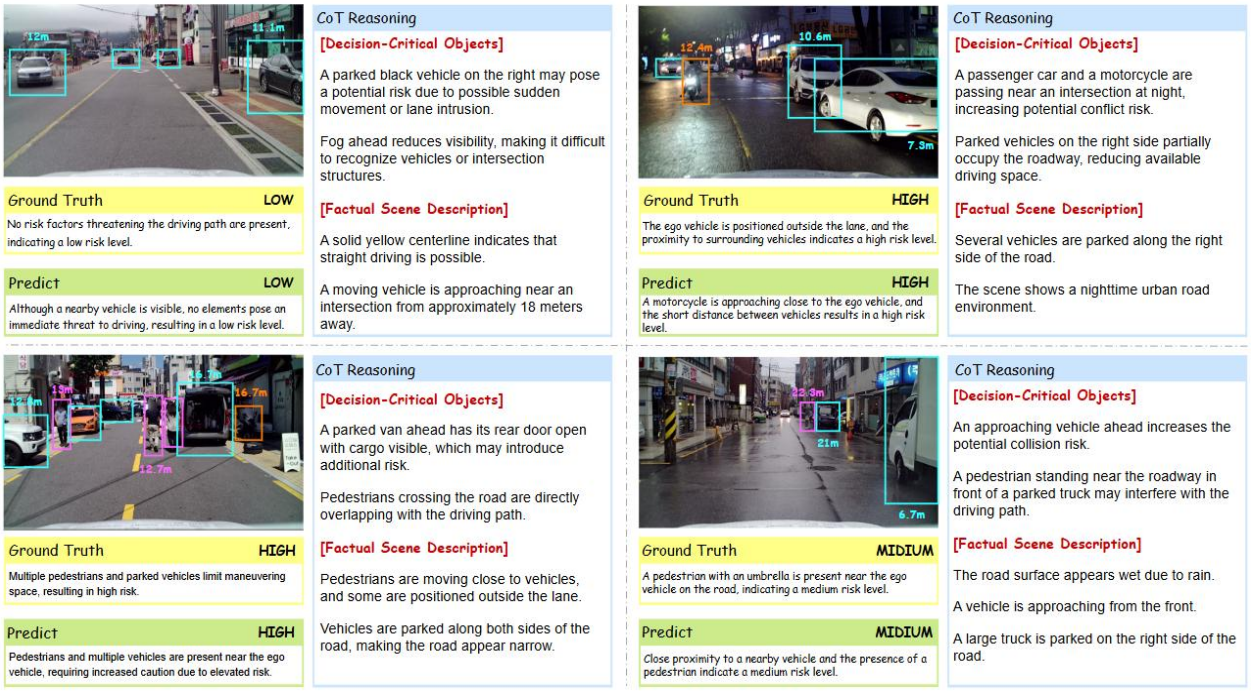

그림 5. Qwen2-VL-7B-Instruct의 테스트 결과 예시 및 CoT 기반 위험도 추론 비교
Fig. 5. Examples of test results and CoT-based risk reasoning of Qwen2-VL-7B-Instruct

AR 기반 상대 거리 정보가 각각 어떤 방식으로 위험 판단에 기여하는지 확인하기 위해 전체 evidence를 사용하는 EGRR을 기준으로 특정 증거 요소를 선택적으로 제거한 세부 제거 실험을 수행하였다. 이는 전체 evidence의 효과를 단순히 확인하는 것을 넘어, 각 증거 요소가 위험도 예측 과정에서 담당하는 역할을 구분하기 위한 것이다.

표 4와 같이 EGRR은 Accuracy 0.855, Macro-F1 0.818, MAE 0.230으로 가장 높은 성능을 보였다. Bounding box 정보를 제거한 경우 Macro-F1이 0.818에서 0.709로 가장 크게 감소하여, 객체의 공간적 위치 정보가 위험도 등급 구분에 중요함을 확인하였다. 거리 정보를 제거한 경우 Accuracy가 0.855에서 0.789로 가장 크게 감소하고 MAE가 0.230에서 0.263으로 증가하여, 객체와 자차 간 근접성이 위험도 판단의 주요 단서임을 보여준다. 객체 수 정보를 제거한 경우 MAE가 0.230에서 0.270으로 가장 크게 증가하여, 객체 밀도 정보가 위험 수준의 세밀한 조정에 기여함을 시사한다. 이러한 결과는 Evidence Builder의 세부 요소들이 서로 다른 측면에서 위험도 예측에 기여할 가능성을 보여준다.

표 4. 구조화된 객체 증거 요소별 제거 실험 결과
Table 4. Ablation results for structured object evidence components

Type	Acc (↑)	Macro F1 (↑)	MAE (↓)
w/o BBox	0.794	0.709	0.243
w/o distance	0.789	0.726	0.263
w/o count	0.802	0.754	0.270
EGRR	**0.855**	**0.818**	**0.230**

4.8 기존 방법과의 비교

본 절에서는 제안한 방법의 유효성을 검증하기 위해, 기존 논문에서 제안된 Multi-Net 기반 멀티태스크 학습 구조를 참고한 기준 모델과 비교 실험을 수행하였다[1]. 해당 기준 모델은 단일 도로 이미지를 입력으로 받아 최종 위험도뿐 아니라 보행자 위험 상황, 날씨, 시간대와 같은 장면 맥락 정보를 동시에 예측하도록 설계되었다. 이는 위험도만 단독으로 예측하는 방식보다, 보행자 존재 여부나 기상 조건과 같은 보조 정보를 함께 학습함으

로써 위험 판단에 필요한 시각적 단서를 보다 안정적으로 포착할 수 있다는 가정에 기반한다. 표 5는 기존 Multi-Net 기반 방법과 본 연구의 주요 차이를 비교한 것이다.

표 3에 제시된 바와 같이, 기준 모델인 Multi-Net은 위험도 등급 예측에서 Accuracy 0.632, Macro-F1 0.460, MAE(ord) 0.562를 기록하여, 제안한 VLM 기반 방법들에 비해 전반적으로 낮은 성능을 보였다. 한편 Multi-Net은 분류 기반 기준 모델이므로, 위험 설명 생성에 대한 텍스트 성능 평가는 비교 대상에 포함되지 않는다. 동일한 위험도 등급 예측 기준에서 Qwen2-VL-7B-Instruct는 Accuracy 0.855, Macro-F1 0.818, MAE(ord) 0.230을 기록하여 기준 모델 대비 뚜렷한 성능 향상을 보였다. 이러한 결과는 장면 맥락 정보를 함께 학습하는 멀티태스크 구조가 일정 수준의 보조적 역할을 할 수는 있으나, 실제 도로 위험 판단에 중요한 객체 수준의 세밀한 단서와 객체 간 공간적 관계를 충분히 반영하는 데에는 한계가 있음을 보여준다. 결과적으로 자율주행 위험 판단에서는 단순한 맥락 기반 멀티태스크 학습만으로는 충분하지 않으며, 객체의 종류, 위치, 거리와 같은 구조화된 근거를 명시적으로 활용하고 이를 단계적 추론과 결합하는 접근이 효과적인 것으로 나타났다.

표 5. 기존 Multi-Net 기반 방법과 제안 방법의 비교
Table 5. Comparison between the Multi-Net-based Mmethod and the proposed method

Category	Multi-Net method	Proposed method
year	2021	2026
Dataset	BDD100K	AI-Hub based custom dataset
Model	Multi-Net	Qwen3-VL
Objective	Crash likelihood and scene attribute prediction	Evidence risk prediction and explanation
Explainability	Classification-only output	Natural-language rationale provided
Metrics	Acc, F1-score, Precision, Recall	Acc, Macro-F1, MAE, text generation metrics

4.9 객체 검출기 버전에 따른 실험 결과

본 절에서는 Evidence Builder에 사용되는 객체 검출기의 성능이 최종 위험 추론 결과에 미치는 영향을 분석하기 위해, YOLO 계열의 서로 다른 버전을 적용한 비교 실험을 수행하였다. 이때 VLM 백본과 학습 조건은 동일하게 유지하고, 객체 검출기만 변경하여 그 영향을 비교하였다. 표 6은 YOLO 8s와 YOLO 11s를 적용한 경우의 위험 설명 생성 성능과 위험도 등급 예측 성능을 비교한 결과를 나타낸다.

실험 결과, YOLO 11s는 위험도 등급 예측에서 Accuracy, Macro-F1, MAE(ord) 기준으로 YOLO 8s보다 다소 우수한 성능을 보였다. 반면 위험 설명 생성 성능은 두 설정 간 차이가 크지 않았으며, 일부 지표에서는 YOLO 8s가, 다른 지표에서는 YOLO 11s가 소폭 높은 값을 보여 전반적으로 유사한 수준을 나타냈다. 이는 객체 검출기의 성능 향상이 최종 위험 추론 성능에 그대로 비례하여 반영되기보다는, VLM이 객체 검출 결과와 원본 이미지의 장면 맥락을 함께 활용하면서 검출 성능 차이의 일부를 보완하기 때문으로 해석할 수 있다. 따라서 제안한 프레임워크에서는 객체 검출기의 성능이 중요하지만, 최종 위험 추론 성능은 객체 검출 결과와 VLM의 장면 이해 및 언어적 추론이 결합되어 결정됨을 확인할 수 있다.

V. 결 론

본 연구는 자율주행 장면에서 위험도 예측의 정확성과 근거 제시의 해석 가능성을 동시에 향상시키기 위해, 구조화된 객체 수준 증거와 단계적 추론을 결합한 EGRR를 제안하였다. 실험 결과, 제안한 방법은 다양한 VLM 백본 중 Qwen2-VL-7B-Instruct에서 가장 우수하고 균형 잡힌 성능을 보였으며, 위험 설명 생성과 위험도 등급 예측 모두에서 안정적인 결과를 나타냈다. 또한 구성요소 제거 실험을 통해 Evidence Builder는 설명 생성과 위험도 예측 모두에 핵심적이며, CoT 기반 중간 추론은 최종 위험도 등급 판단의 안정성과 일관성을 높이는 데 기여

표 6. 객체 검출기 버전에 따른 위험 설명 생성 및 위험도 예측 성능 비교
Table 6. Comparison of risk description generation and risk-level prediction performance across detector versions

Model	BLEU (↑)	METEOR (↑)	R-L (↑)	CIDEr (↑)	SPICE (↑)	BERT F1 (↑)	Acc (↑)	Macro F1 (↑)	MAE (↓)
YOLO variant									
11s	78.590	**0.827**	0.899	7.514	**0.847**	**0.965**	**0.855**	**0.818**	**0.230**
8s	**78.647**	0.825	**0.901**	**7.577**	0.844	**0.965**	0.840	0.806	0.249

함을 확인하였다. 기존 Multi-Net 기반 기준 모델과의 비교에서도, 제안한 방법은 객체 수준의 세밀한 단서를 보다 효과적으로 반영하여 더 우수한 성능을 보였다.

종합하면, 본 연구는 자율주행 위험 판단에서 객체 기반의 구조화된 증거를 통해 판단 근거를 보강하고, 이를 바탕으로 위험 판단의 근거를 함께 제시할 수 있는 설명형 프레임워크의 유효성을 보여준다. 다만 CoT 기반 중간 추론은 외부 모델이 생성한 설명을 학습 신호로 활용하므로, 특정 장면 조건이나 표현 방식에 대한 편향이 일반화 성능에 영향을 줄 가능성이 있다. 따라서 향후 연구에서는 다양한 도로 환경, 날씨, 시간대, 교통 상황을 포함한 데이터로 평가 범위를 확장하고, 객체 증거와 CoT 설명 간의 일치성을 검증하는 절차를 추가함으로써 실제 자율주행 환경에서도 안정적으로 적용 가능한 위험 판단 프레임워크로 발전시키고자 한다.

References

[1] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A system of vision sensor based deep neural networks for complex driving scene analysis in support of crash risk assessment and prevention", arXiv preprint arXiv:2106.10319, pp. 1-11, Jun. 2021. https://doi.org/10.48550/arXiv.2106.10319.

[2] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2Drive: Towards Interpretable and Chain-Based Reasoning for Autonomous Driving", Computer Vision – ECCV 2024, Vol. 15084, pp. 292-308, Oct. 2024. https://doi.org/10.1007/978-3-031-73347-5_17.

[3] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "DriveLM: Driving with Graph Visual Question Answering", Computer Vision – ECCV 2024, Vol. 15110, pp. 256-274, Nov. 2024. https://doi.org/10.1007/978-3-031-72943-0_15.

[4] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models", Proc. of the 8th Conference on Robot Learning, Proceedings of Machine Learning Research, Munich, Germany, Vol. 270, pp. 4698-4726, Nov. 2025.

[5] W. Liu, J. Zhang, B. Zheng, Y. Hu, Y. Lin, and Z. Zeng, "X-Driver: Explainable Autonomous Driving with Vision-Language Models", arXiv preprint arXiv:2505.05098, pp. 1-8, Jun. 2025. https://doi.org/10.48550/arXiv.2505.05098.

[6] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 14455-14465, Jun. 2024. https://doi.org/10.1109/CVPR52733.2024.01370.

[7] Y.-H. Liao, R. Mahmood, S. Fidler, and D. Acuna, "Can Large Vision-Language Models Correct Semantic Grounding Errors By Themselves?", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 14667-14678, Jun. 2025. https://doi.org/10.1109/CVPR52734.2025.01367.

[8] S. Xing, C. Qian, Y. Wang, H. Hua, K. Tian, Y. Zhou, and Z. Tu, "OpenEMMA: Open-Source Multimodal Model for End-to-End Autonomous Driving", Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Tucson, AZ, USA, pp. 1001-1009, Feb. 2025. https://doi.org/10.1109/WACVW65960.2025.00113.

[9] L. Wang, Y. Hu, J. He, X. Xu, N. Liu, H. Liu, and H. T. Shen, "T-SciQ: Teaching Multimodal Chain-of-Thought Reasoning via Large Language Model Signals for Science Question Answering", Proc. of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, Vol. 38, No. 17, pp. 19162-19170, Feb. 2024. https://doi.org/10.1609/aaai.v38i17.29884.

[10] R. Zhang, B. Zhang, Y. Li, H. Zhang, Z. Sun, Z. Gan, Y. Yang, R. Pang, and Y. Yang, "Improve Vision Language Model Chain-of-Thought Reasoning", Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, pp. 1631-1662, Jul. 2025. https://doi.org/10.18653/v1/2025.acl-long.82.

[11] AI-Hub, "Urban Road Object Recognition Autonomous Driving Data", AI-Hub, 2023. https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71784. [accessed: Mar. 23, 2026].

[12] iRAP, "iRAP Star Rating and Investment Plan Manual", Austroads, 2024. https://austroads.gov.au/__data/assets/pdf_file/0033/836295/iRAP_Star_Rating_and_Investment_Plan_Manual_English.pdf. [accessed: Mar. 23, 2026].

[13] U.S. Department of Defense, "MIL-STD-882E: Department of Defense Standard Practice—System Safety", Sep. 2023. https://quicksearch.dla.mil/qsDocDetails.aspx?ident_number=36027. [accessed: Mar. 23, 2026].

[14] International Organization for Standardization, "ISO 26262-1:2018, Road Vehicles—Functional Safety—Part 1: Vocabulary", 2018. https://www.iso.org/standard/68383.html. [accessed: Mar. 23, 2026]

[15] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs", Advances in Neural Information Processing Systems, New Orleans, Louisiana, USA, Vol. 36, pp. 10088-10115, Dec. 2023. https://doi.org/10.52202/075280-0441.

[1] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A system of vision sensor based deep neural networks for complex driving scene analysis in support of crash risk assessment and prevention", arXiv preprint arXiv:2106.10319, pp. 1-11, Jun. 2021. https://doi.org/10.48550/arXiv.2106.10319.

[5] W. Liu, J. Zhang, B. Zheng, Y. Hu, Y. Lin, and Z. Zeng, "X-Driver: Explainable Autonomous Driving with Vision-Language Models", arXiv preprint arXiv:2505.05098, pp. 1-8, Jun. 2025. https://doi.org/10.48550/arXiv.2505.05098.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, pp. 311-318, Jul. 2002. https://doi.org/10.3115/1073083.1073135.

[17] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out, Barcelona, Spain, pp. 74-81, Jul. 2004. https://aclanthology.org/W04-1013/.

[18] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, pp. 65-72, Jun. 2005. https://aclanthology.org/W05-0909/.

[19] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-Based Image Description Evaluation", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 4566-4575, Jun. 2015.

https://doi.org/10.1109/CVPR.2015.7299087.

[20] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation", Computer Vision – ECCV 2016, Lecture Notes in Computer Science, Vol. 9909, pp. 382-398, Oct. 2016. https://doi.org/10.1007/978-3-319-46454-1_24.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT", Proc. of the International Conference on Learning Representations (ICLR), Apr. 2020. https://doi.org/10.48550/arXiv.1904.09675.

저자소개

이 세 연 (Se-Yeon Lee)

2023년 3월 ~ 현재 : 경상국립대학교 컴퓨터공학과 학사과정
관심분야 : 인공지능, 컴퓨터비전, 이미지 생성, 대규모 언어모델

조 신 호 (Sin-Ho Cho)

2024년 3월 ~ 현재 : 경상국립대학교 컴퓨터공학과 학사과정
관심분야 : 자율주행, 컴퓨터비전

민 경 제 (Gyeong-Je Min)

2023년 3월 ~ 현재 : 경상국립대학교 컴퓨터공학과 학사과정
관심분야 : 인공지능, 자율주행

김 주 영 (Ju-Young Kim)

2025년 2월 : 경상국립대학교 컴퓨터공학부(공학사), 융합전공(USG융합학사)
2025년 3월 ~ 현재 : 경상국립대학교 컴퓨터공학과 석사과정
관심분야 : 인공지능, 컴퓨터 비전, 이미지 생성, 온디바이스 AI

김 건 우 (Gun-Woo Kim)

2006년 12월 : 호주뉴캐슬대학교 컴퓨터공학과(공학사)
2007년 9월 : 호주뉴캐슬대학교 정보공학과(공학석사)
2017년 8월 : 한양대학교 컴퓨터공학과(공학박사)
2021년 9월 ~ 현재 : 경상국립대학교 컴퓨터공학과 부교수
관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝