딜레마 상황 자율주행을 위한 지식그래프 기반 의미 장면 이해

조웅제1)·오현서1)·이준석2) ·김시호*3)
연세대학교 기계공학부1)·연세대학교 건설환경공학부2)·연세대학교 첨단융합공학부3)

Knowledge Graph-Based Semantic Scene Understanding for Autonomous Driving in Dilemma Situations

Woongje Cho1)·Hyeonseo Oh1)·Junseok Lee2)·Shiho Kim*3)

1) School of Mechanical Engineering, Yonsei University, Seoul 03722, Republic of Korea

2) School of Civil and Environmental Engineering, Yonsei University, Seoul 03722, Republic of Korea

3) School of Integrated Technology, Yonsei University, Incheon 21983, Republic of Korea

Abstract : For autonomous vehicles to operate safely in real-world environments, they must go beyond object detection and understand semantic relations and situational context within a scene. In particular, dilemma situations involving conflicting constraints, such as accident avoidance, pedestrian priority, and traffic rule compliance, are difficult to resolve using conventional 3D Scene Graphs (3DSGs), which mainly represent spatial structure. To address this limitation, this paper proposes a Knowledge Graph (KG)-enhanced semantic scene understanding framework tailored to autonomous driving dilemma scenarios. The proposed KG represents not only objects, attributes, and relations, but also traffic rules, class hierarchies, commonsense context, and scenario-specific semantic descriptions in a structured form. We evaluate the framework using 30 scene-understanding queries across six cognitive categories under a controlled LLM-as-judge setting (GPT-4o-mini for answering, GPT-4o for judging; N=300). Results show that the KG-based method significantly outperforms the 3DSG baseline in reasoning quality (mean 4.41 vs. 3.56, Wilcoxon $p<0.001$, Cohen's $d=0.84$), with the largest gains in dilemma reasoning (+1.71) while spatial queries confirm design fairness (−0.07). A five-condition ablation study—LLM-only (1.25), KG structure-only (3.29), 3DSG (3.56), 3DSG+NL (4.14), and KG full (4.41)—reveals that natural-language semantic descriptions (rdfs:comment) are the dominant contributor to performance; the KG framework's value lies in providing systematic infrastructure for attaching and managing domain-specific annotations rather than ontological structure alone. These findings suggest that integrating structured domain knowledge is essential for reliable semantic scene understanding in autonomous-driving dilemma situations.

Key words : Knowledge Graph(지식그래프), Autonomous Driving(자율주행), Dilemma Situation(딜레마 상황), Semantic Scene Understanding(의미 기반 장면 이해), OWL Ontology(온톨로지), LLM(대규모 언어모델), 3D Scene Graph(3차원 장면 그래프)

*교신저자, E-mail: shiho@yonsei.ac.kr

Nomenclature

KG : Knowledge Graph
3DSG : 3D Scene Graph
LLM : Large Language Model
OWL : Web Ontology Language
RDF : Resource Description Framework
SPARQL : SPARQL Protocol and RDF Query Language
RAG : Retrieval-Augmented Generation
QA : Question Answering

1. 서 론

자율주행 시스템의 안전한 운행을 위해서는 주변 환경에 대한 정확한 장면 이해(scene understanding)가 필수적이다. 기존 자율주행 시스템은 주로 HD Map이나 3D Scene Graph(3DSG)를 활용하여 "무엇이 어디에 있는가"라는 공간적 정보를 처리하는 데 초점을 맞추고 있다 [1, 2]. 그러나 실제 주행 환경에서는 황색 신호 딜레마, 불법 주차 차량 추월, 교통 흐름과 제한속도 간 갈등 등 단순 공간 인식만으로는 해결할 수 없는 딜레마 상황이 빈번하게 발생한다 [3].

이러한 딜레마 상황에서는 교통법규, 물리 법칙, 방어운전 상식 등 "왜, 어떻게 행동해야 하는가"에 대한 의미적 추론(semantic reasoning)이 요구된다. 최근 Knowledge Graph(KG)와 대규모 언어모델(LLM)을 결합한 연구가 활발히 진행되고 있으며, Hybrid-Driving [4]은 시나리오 진화 지식그래프(SEKG)를 통해 LLM 단독 대비 37.5%의 성능 향상을 달성하였고, KnowVal [5]은 교통법규와 방어운전 지식을 KG로 구조화하여 nuScenes 벤치마크에서 최저 충돌률을 기록하였다.

그러나 기존 연구들은 KG의 효과를 LLM 단독과 비교하거나, 정성적 분석에 그치는 경우가 대부분이며, 기존 3DSG 기반 접근 대비 KG의 부가 가치를 정량적으로 검증한 연구는 부재하다.

본 연구에서는 자율주행 딜레마 상황에 특화된 KG 기반 시맨틱 장면 이해 프레임워크를 제안한다. 교통법규, 물리 법칙, 방어운전 상식을 OWL 온톨로지로 구조화하고, rdfs:comment를 통해 LLM이 활용 가능한 시맨틱 설명을 제공한다. 제안 기법의 효과를 검증하기 위해 기존 3DSG 기반 접근을 baseline으로 설정하고, 동일한 지각 데이터 위에서 LLM-as-judge 프레임워크를 통해 정량적으로 평가한다. 나아가 ablation 실험을 통해 KG 효과의 원천—온톨로지 구조 vs rdfs:comment 시맨틱 설명—을 분리 분석한다.

2. 관련 연구

2.1 자율주행에서의 3D Scene Graph

3D Scene Graph는 환경을 노드(객체)와 엣지(관계)로 표현하는 구조화된 공간 표현 방법이다. Armeni et al. [18]는 의미론, 3D 공간, 카메라를 통합하는 3D Scene Graph 구조를 최초로 제안하였으며, Rosinol et al. [19]은 이를 동적 환경으로 확장하여 장소, 객체, 사람의 시공간적 상호작용을 모델링하였다.

자율주행 분야에서 CURB-SG [1]는 다중 에이전트 LiDAR 데이터로부터 협력적 3D 장면 그래프를 생성하며, GraphAD [2]는 자아-에이전트-맵 상호작용을 통합 그래프로 모델링하여 종단 간 자율주행에 활용하였다. T2SG [6]는 차선-신호-도로 위상을 명시적 장면 그래프로 구성하여 OpenLane-V2 벤치마크에서 최고 성능을 달성하였다. 그러나 이들 연구는 공간적 관계와 위상 구조에 집중하며, "왜 위험한가" 또는 "어떻게 대응해야 하는가"와 같은 의미적 추론 능력은 제공하지 못하는 한계가 있다.

2.2 자율주행 의사결정을 위한 Knowledge Graph

Knowledge Graph는 개체 간의 의미적 관계를 온톨로지 기반으로 표현하며, 도메인 지식을 구조화하여 저장할 수 있다. Wickramarachchi et al. [7]은 OWL 기반 Driving Scene Ontology(DSO)를 제안하고 7가지 뉴로심볼릭 벤치마크를 수행하였으며, nuScenes KG [8]는 4,300만 개 이상의 RDF 트리플로 구성된 대규모 주행 지식그래프를 구축하여 궤적 예측에 활용하였다. Hussien et al. [9]은 KG와 LLM을 RAG 방식으로 결합하여 보행자 횡단 및 차선 변경 예측에 활용하였으며, 설명 가능한 예측을 생성하는 데 KG가 효과적임을 보였다.

이러한 연구들은 KG가 자율주행에서 의미적 추론에 효과적임을 보여주지만, KG 기반 접근의 효과를 기존 3DSG 기반 접근과 직접적으로 정량 비교한 연구는 수행되지 않았다. 본 연구는 이러한 공백을 메우고자 한다.

2.3 LLM 기반 주행 에이전트

최근 LLM을 자율주행 의사결정에 활용하는 연구가 활발하다. DiLu [10]는 추론-반성(reasoning-

reflection) 모듈을 통한 경험 기반 학습 프레임워크를 제안하여 LLM이 주행 경험으로부터 점진적으로 학습하도록 하였고, LanguageMPC [11]는 LLM의 상식 추론을 MPC 제어와 결합하여 자연어 기반 주행 의사결정을 가능하게 하였다.

Mao et al. [14]은 도구 라이브러리와 메모리를 갖춘 LLM 인지 에이전트를 제안하였으며, Cai et al. [15]은 교통 규칙의 RAG 기반 검색과 LLM 추론을 결합하였다. 그러나 이들은 LLM의 내재된 지식에만 의존하며, 구조화된 외부 지식의 체계적 활용은 미흡하다. 본 연구는 동일한 LLM(GPT-4o-mini)에 3DSG 또는 KG 컨텍스트를 제공하여 외부 지식 구조의 효과를 분리 측정한다는 점에서 차별화된다.

3. KG 기반 시맨틱 장면 이해 기법

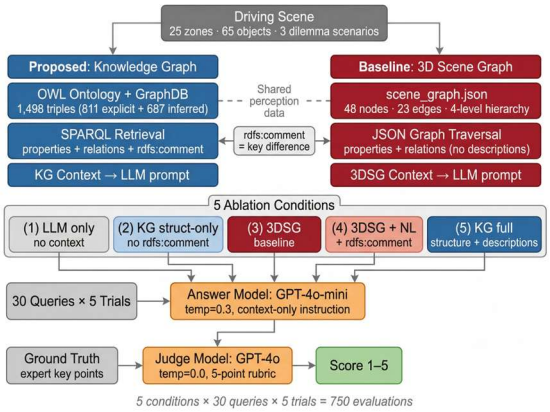

Fig. 1 Overall architecture of the proposed KG-based semantic scene understanding framework and evaluation pipeline.

Fig. 1은 제안 기법의 전체 구조를 보여준다. 주행 환경을 OWL 온톨로지로 모델링하여 GraphDB에 저장하고, 질문에 대해 SPARQL 쿼리로 관련 엔티티의 속성, 관계, rdfs:comment를 검색한다. 검색된 컨텍스트는 답변 모델(GPT-4o-mini)에 제공되며, 생성된 답변은 채점 모델(GPT-4o)이 전문가 정답 기반 5점 루브릭으로 평가한다. 3DSG baseline은 동일한 지각 데이터를 JSON 그래프 탐색으로 제공하여 시맨틱 지식의 효과만을 분리 측정한다.

3.1 주행 환경 및 온톨로지 설계

본 연구에서는 도시, 주거, 고속도로 3개 구역으로 구성된 가상 주행 환경을 설계하였다. 환경은 25개 구역(교차로 4, 차선 구간 11, 고속도로 3, 합류 구간 2, 스쿨존 1, 주차장 2, 기타 2)과 23개 연결로 구성되며, 총 65개의 도로 객체(차량 16, 보행자 5, 교통 제어 장치 18, 인프라 9 등)를 포함한다. 3가지 딜레마 시나리오—황색 신호 딜레마, 불법 주차 추월, 교통 흐름 대비 제한 속도—를 환경에 내재하였다.

온톨로지는 A.U.T.O.(Automotive Urban Traffic Ontology) [12]와 Bagschik의 5+1 계층 모델 [13]을 참조하여 OWL로 설계하였다. 클래스 계층은 TopologicalObject(RoadNetwork, District, Zone, Connection)와 PhysicalObject(Vehicle, Pedestrian, TrafficControl, Infrastructure, Obstacle)의 두 상위 클래스로 구성되며, Vehicle은 Sedan, Truck, Bus, Motorcycle 등으로, Zone은 Intersection, LaneSegment, Highway, SchoolZone 등으로 세분된다.

객체 속성(Object Property)으로 inZone, isZoneOf, isDistrictOf, connectsTo, isConnectionOf를, 데이터 속성(Data Property)으로 speed, state, speedLimit, x, y, isObstacle을 정의하였다. GraphDB에 rdfsplus-optimized 규칙셋을 적용하여 811개 명시 트리플에서 1,498개 추론 트리플을 생성하였다(1.85배 증가). 추론된 트리플에는 역속성(inverse property) 관계와 클래스 계층에 의한 상위 클래스 추론이 포함된다.

KG의 핵심 차별점은 rdfs:comment를 통한 시맨틱 설명이다. 각 객체에 물리적 속성, 상태, 맥락/어포던스, 상식적 지식을 자연어로 기술하여 LLM이 추론에 활용할 수 있도록 하였다. 예를 들어, 황색 신호 등 객체에는 자아 차량의 속도·거리 기반 정지거리 분석, 후방 차량과의 추돌 위험, 딜레마 존(dilemma zone) 개념 [3]이 기술된다.

Table 1 Examples of rdfs:comment semantic descriptions in the proposed KG

Entity	rdfs:comment (excerpt)
traffic_light_1 (Yellow signal)	At ego's distance of ~30m and speed of 55 km/h, stopping distance ≈ 35-40m. This creates a 'dilemma zone' with sedan_3 only 15m behind at 58 km/h. Proceeding is safer than emergency braking.
truck_4 (Illegally parked)	A large truck illegally parked in the right lane. Its 3.5m height blocks visibility of oncoming traffic. Passing requires temporary lane departure—technically violates markings but is standard safe response.
emergency_vehicle_12	Ambulance with sirens active, approaching from behind at ~90

	km/h. All vehicles must yield. If at an intersection, clear it before stopping—a legal obligation with priority over normal traffic rules.

3.2 3DSG baseline과의 공정한 비교 설계

Table 2 Information separation between 3DSG and KG

Information Type	3DSG	KG
Object class, coordinates	O	O
Observable state (speed, signal)	O	O
Spatial relations (edges)	O	O
Semantic description (rdfs:comment)	X	O
Class hierarchy (ontology)	X	O
Domain rules (traffic law)	X	O
Common-sense context	X	O

Table 2와 같이, 양측 모두 동일한 지각(perception) 데이터—객체 클래스, 위치 좌표, 관찰 가능한 상태, 공간 관계—를 공유하되, KG에만 시맨틱 설명, 클래스 계층, 도메인 규칙, 상식 추론 맥락이 추가된다. 이를 통해 "눈으로 볼 수 있는 정보"는 동일하게 유지하고, "알고 있어야 하는 지식"의 효과만을 분리 측정하고자 하였다.

3DSG baseline은 Armeni et al. [18]와 Rosinol et al. [19]의 3D Scene Graph 표현 형식을 참조하여, 4단계 계층 구조(Network→District→Zone→Object), 노드 속성(속도, 크기, 색상, 방향), 23개의 공간 관계 엣지(behind, adjacent, approaching, in_same_lane 등)를 포함하도록 강화하였다. 선행 연구에서 보고된 불공정 요소—빈약한 flat JSON, LLM 미사용, 순환 채점—를 모두 해소하여 3DSG가 충분한 표현력을 갖춘 공정한 baseline이 되도록 설계하였다.

KG 모드에서는 SPARQL 쿼리를 통해 관련 엔티티의 속성, 공간 관계, rdfs:comment를 검색한다. 예를 들어, "intersection_2의 교통 상황"에 대한 질문은 다음과 같은 SPARQL 패턴으로 검색된다:

```
SELECT ?obj ?prop ?val ?comment WHERE {
  ?obj :inZone :intersection_2 .
  ?obj ?prop ?val .
  OPTIONAL { ?obj rdfs:comment ?comment }
}
```

이 쿼리는 intersection_2에 위치한 모든 객체의 속성과 시맨틱 설명을 한 번에 검색하며, 추론 엔진에 의해 역속성(inZone의 역인 isZoneOf) 관계도 자동으로 포함된다. 3DSG 모드에서는 동일한 정보를 JSON 그래프의 connected_objects 배열 탐색으로 추출하되, 시맨틱 설명(rdfs:comment)은 포함되지 않는다. 양측 검색 결과는 동일한 형식의 텍스트 컨텍스트로 변환되어 LLM에 제공된다.

3.3 평가 프레임워크

제안 기법의 효과를 검증하기 위해 LLM-as-judge 프레임워크를 설계하였다. Zheng et al. [16]은 MT-Bench에서 GPT-4 judge가 인간 평가자와 80% 이상의 일치율을 보임을 입증하였으며, 단일 답변 채점(single-answer grading) 방식이 쌍대 비교(pairwise comparison)보다 일관성이 높음을 보고하였다. 본 연구는 이를 자율주행 도메인에 적용하여 reference-guided single-answer grading 방식을 채택하였다.

답변 모델의 시스템 프롬프트는 "제공된 장면 컨텍스트만을 기반으로 답변하라"는 엄격한 제약을 부여하여 LLM의 사전 학습 지식 사용을 억제하였다. 이는 KG/3DSG 컨텍스트의 효과를 정확히 측정하기 위한 핵심 설계 원칙이다. 채점 모델에게는 전문가 작성 정답의 핵심 포인트를 제공하고, 5점 루브릭(5=모든 핵심 포인트 정확, 4=경미한 누락, 3=핵심은 맞으나 공백 존재, 2=주요 추론 오류, 1=부정확 또는 위험한 권고)을 적용하였다. 답변 모델(GPT-4o-mini)과 채점 모델(GPT-4o)을 서로 다른 모델로 분리하여 평가의 독립성을 확보하고, temperature=0.0으로 채점의 결정론적 일관성을 보장하였다.

Table 3 Experiment configuration

Parameter	Value
Answer model	GPT-4o-mini (temp=0.3)
Judge model	GPT-4o (temp=0.0)
Queries	30 (6 categories)
Trials per query	5
Total evaluations	300 (30×2×5)
KG storage	GraphDB (1,498 triples)
3DSG data	scene_graph.json (48 nodes, 23 edges)
Statistical test	Wilcoxon signed-rank (one-sided)

30개 질문은 자율주행 장면 이해의 인지적 복잡도를 기준으로 6개 카테고리로 구성하였다: Spatial(공간 인식, 3개), Identification(객체 식별, 3개), Semantic(의미 이해, 7개), Hierarchy(계층 추론, 3개), Safety(안전 추론, 7개), Dilemma(딜레마 의사결정, 7

개). Spatial 카테고리는 양측이 동등한 성능을 보일 것으로 예상되어 비교의 공정성을 검증하는 내부 기준선 역할을 한다. Semantic, Safety, Dilemma(각 7개)는 KG 차별화가 기대되는 핵심 영역으로, 각 카테고리 내에서 법규, 물리, 상식 등 다양한 하위 주제를 포함하도록 설계하였다.

Dilemma 카테고리에는 황색 신호 딜레마 [3], 안전을 위한 규칙 위반의 윤리적 정당성 [17], 응급 차량 양보 프로토콜, 교통 흐름과 제한속도 간 갈등 등 복합적 추론이 요구되는 질문을 포함하였다. 각 질문에 대해 전문가가 핵심 포인트(3~6개)와 채점 시 유의사항을 포함한 정답(ground truth)을 작성하였다.

전문가 정답은 제1저자가 온톨로지 데이터와 3DSG 데이터를 모두 참조하여 작성하였다. 핵심 포인트는 양측 데이터에서 관찰 가능한 사실 기반으로 구성하여 특정 표현 방식에 유리하지 않도록 하였으나, 연구자가 KG 설계자이자 정답 작성자라는 이중 역할로 인한 잠재적 편향 가능성이 존재하며, 이에 대해서는 5장에서 논의한다.

4. 실험 결과

4.1 전체 결과

실험 결과, 제안 기법(KG)은 전체 평균 4.41점으로 baseline(3DSG)의 3.56점을 0.85점 차이로 유의미하게 상회하였다. 통계 검정은 각 쿼리의 5회 시행 평균을 하나의 관측치로 사용하여 30개 대응표본(paired observations)에 대해 Wilcoxon signed-rank test(단측, H1: KG > 3DSG)를 수행하였다. 그 결과 $p=0.000766$($p<0.001$)으로 통계적으로 유의하였으며, 효과 크기는 rank-biserial correlation $r=0.70$(large effect), Cohen's $d=0.84$(large effect)에 해당한다.

Table 4 Category-level comparison results

Category	N	KG	3DSG	Δ	p-value
Spatial	3	3.67	3.73	-0.07	n<5
Identification	3	5.00	3.73	+1.27	n<5
Semantic	7	4.26	3.49	+0.77	0.016*
Hierarchy	3	4.53	4.60	-0.07	n<5
Safety	7	4.40	3.71	+0.69	0.031*
Dilemma	7	4.60	2.89	+1.71	0.016*
Overall	30	4.41	3.56	+0.85	<0.001***

4.2 카테고리별 분석

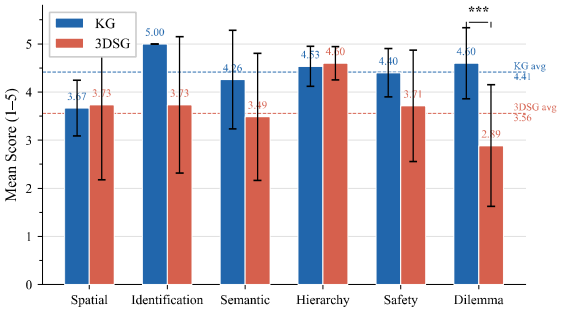

Fig. 2 Mean scores by category with error bars. Dashed lines indicate overall means. *** denotes $p<0.001$.

Fig. 2와 Table 4에서 Spatial 카테고리의 KG(3.67)와 3DSG(3.73) 차이는 -0.07로 사실상 동등하며, 이는 양측의 지각 데이터가 공정하게 설계되었음을 내부적으로 검증한다.

주목할 만한 패턴은 지식 요구도가 높은 카테고리일수록 KG의 우위가 뚜렷해진다는 점이다: Spatial(±0) → Semantic(+0.77) → Safety(+0.69) → Dilemma(+1.71). 이는 KG의 rdfs:comment가 제공하는 도메인 지식이 단순 공간 인식이 아닌, 의미적 추론이 요구되는 상황에서 결정적 역할을 함을 보여준다.

Hierarchy 카테고리에서 예상과 달리 양측이 동등(-0.07)한 것은 주목할 만한 발견이다. GPT-4o-mini가 "traffic_light", "speed_limit_sign" 등의 클래스명만으로도 카테고리 분류를 수행할 수 있어, LLM의 사전 학습 지식이 온톨로지 계층을 효과적으로 대체한 것으로 해석된다. 이는 KG 설계 시 LLM이 이미 보유한 일반 지식보다 상황 특화된 도메인 맥락에 리소스를 집중해야 함을 시사한다.

Dilemma 카테고리에서 KG는 4.60점으로 3DSG의 2.89점을 1.71점 차이로 압도적으로 상회하였다($p=0.016$). 7개 Dilemma 질문 중 6개에서 KG가 1.0점 이상의 우위를 보였으며, 이는 딜레마 의사결정에서 교통법규, 물리 법칙, 방어운전 상식이 반드시 필요함을 확인한다.

4.3 질문별 상세 분석

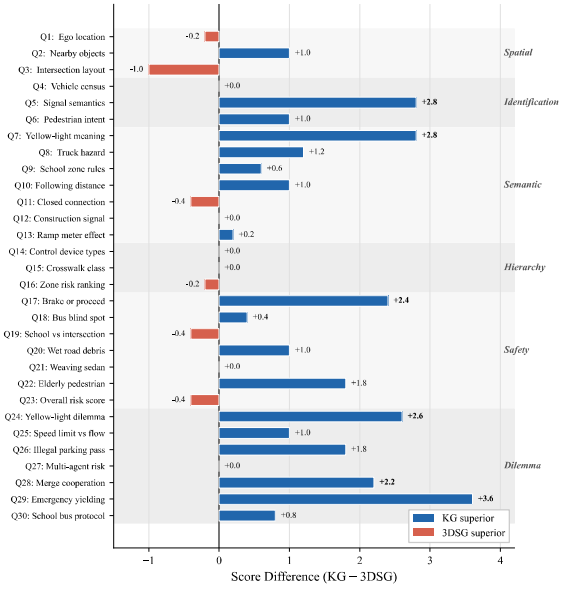

Fig. 3 Per-query score differences (KG − 3DSG). Blue: KG superior, Red: 3DSG superior.

Fig. 3은 30개 질문별 점수 차이를 보여준다. KG가 우위인 질문은 22개, 3DSG가 우위인 질문은 5개, 동률은 3개이다. KG 우위 상위 5개 질문의 구체적인 분석은 다음과 같다.

Q29(응급차량 양보 시점, Δ=+3.6)에서 KG는 5.0점, 3DSG는 1.4점을 기록하였다. KG는 emergency_vehicle_12의 rdfs:comment에 기술된 "교차로 내에서는 정차하지 말고 통과 후 양보"라는 법적 의무와 교차로 내 정차의 위험성을 종합하여 올바른 판단을 일관되게 도출하였다. 반면 3DSG는 sirens=active라는 속성만으로 즉시 정차를 권고하여 교차로 한가운데에서의 위험한 정차를 유발하였다.

Q7(황색 신호 의미, Δ=+2.8)에서도 KG의 딜레마 존 분석이 결정적이었다. KG는 정지 거리와 추돌 위험을 명시하여 5.0점을 달성한 반면, 3DSG는 signal=yellow라는 원시 데이터만 제공하여 2.2점에 그쳤다.

3DSG가 유일하게 유의미한 우위를 보인 Q3(intersection_2 공간 레이아웃, Δ=-1.0)에서는 KG의 SPARQL 쿼리가 bus_lane_1 연결을 누락한 반면, 3DSG는 connected_zones 배열에 명시적으로 포함하고 있었다. 이는 KG 자체의 한계가 아닌 SPARQL 검색 로직의 한계로, 쿼리 최적화 또는 다중 홉(multi-hop) 탐색을 통해 개선 가능한 문제이다.

Q5(차량 식별, Δ=+2.8)에서 KG는 5회 모두 만점(5.0)을 기록한 반면, 3DSG는 평균 2.2점([2,2,2,3,2])에 그쳤다. KG의 rdfs:comment에는 sedan_3의 "후방 15m에서 58km/h로 접근 중이며 급제동 시 추돌 위험"이라는 맥락이 포함되어 정확한 위험 요소 식별이 가능하였으나, 3DSG는 class=sedan, speed=58 등 원시 속성만으로 차량 간 상호작용의 위험성을 추론하지 못하였다.

Q22(노인 보행자 위험 평가, Δ=+1.8)에서 KG는 3.8점, 3DSG는 2.0점을 기록하였다. KG는 elderly_pedestrian_1의 rdfs:comment에 기술된 "보행 속도가 느려 신호 내 횡단 완료가 어려울 수 있음", "운전자의 시야 확보 주의" 등의 도메인 지식을 활용하여 다층적 위험 분석을 수행한 반면, 3DSG는 type=elderly_pedestrian, location 정보만으로 단순 위치 기반 판단에 그쳤다.

Fig. 4 Score heatmap across 30 queries (rows) and two conditions (columns). Green: high scores, Red: low scores. Category boundaries shown.

Fig. 4의 히트맵은 30개 질문 전체의 점수 패턴을 보여준다. KG 열(왼쪽)은 전반적으로 녹색(고득점)이 지배적인 반면, 3DSG 열(오른쪽)은 Dilemma와 Safety 영역에서 적색(저득점)이 빈번하다. 특히 Dilemma 카테고리의 Q24~Q30 구간에서 양측의 색상 대비가 가장 극명하게 나타난다.

4.4 추론 안정성 분석

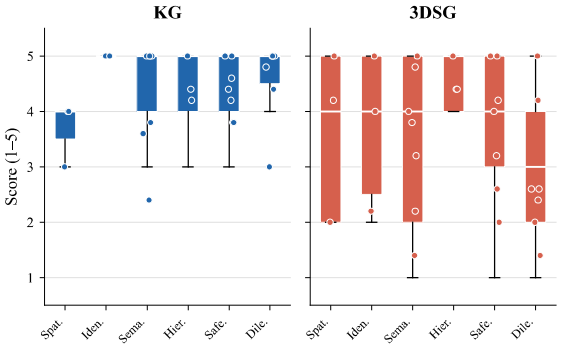

Fig. 5 Score distribution by category. Left: KG, Right: 3DSG. Dots represent per-query means.

Fig. 5는 카테고리별 점수 분포를 보여준다. KG는 모든 카테고리에서 상단(4~5점)에 밀집된 안정적 분포를 보이는 반면, 3DSG는 특히 Dilemma와 Safety에서 1~5점까지 넓게 분산되어 추론의 불안정성이 두드러진다. KG의 사분위 범위(IQR)는 Dilemma에서 4.2~5.0(0.8)인 반면, 3DSG는 1.4~4.4(3.0)로 약 3.75배 넓다.

Table 5 Standard deviation of trial scores by category (lower = more stable)

Category	KG std	3DSG std	Ratio
Spatial	0.617	1.335	2.16×
Identification	0.000	1.280	∞
Semantic	1.067	1.380	1.29×
Hierarchy	0.640	0.507	0.79×
Safety	0.736	1.274	1.73×
Dilemma	0.775	1.301	1.68×
Overall	0.837	1.328	1.59×

Table 5는 카테고리별 시행 점수의 표준편차를 비교한다. KG는 6개 카테고리 중 5개에서 3DSG보다 낮은 표준편차를 보이며, 특히 Identification에서는 표준편차 0.000(모든 질문에서 5회 반복 모두 동일 점수)으로 완벽한 일관성을 달성하였다. 유일한 예외인 Hierarchy(KG 0.640 vs 3DSG 0.507)에서는 양측 모두 높은 점수(4.5+)를 기록하여 편차의 절대적 크기가 작다.

특히 Q17(황색 신호 브레이크/진행)에서 3DSG 모드의 5회 반복 점수는 [1, 5, 1, 3, 3](표준편차 1.67)으로 극도로 불안정한 반면, KG 모드는 [5, 5, 5, 5, 5](표준편차 0.0)로 완전히 일관되었다. 동일한 3DSG 데이터를 제공했음에도 LLM이 때때로 신호등 거리를 잘못 해석하여 1점을 받은 반면, KG의 rdfs:comment는 정지거리와 추돌 위험을 명시하여 일관된 5점을 달성하였다.

유사한 불안정성은 Q11(의미, 3DSG [5,2,4,4,5] std=1.22)과 Q25(딜레마, 3DSG [2,3,3,1,1] std=1.00)에서도 관찰된다. 이는 시맨틱 설명이 부재할 때 LLM이 동일한 원시 데이터를 시행마다 다르게 해석하는 확률적 불안정성을 보여주며, 안전 관련 의사결정에서 KG의 시맨틱 가이드가 LLM의 해석 편차를 안정적으로 억제함을 시사한다.

전체 점수 분포에서 KG는 150회 시행 중 90회(60%)가 만점(5점)이고 1점은 0회인 반면, 3DSG는 만점 51회(34%), 1점이 10회(7%) 발생하였다. KG의 천장 효과(ceiling effect)는 더 도전적인 질문 설계의 필요성을 시사하지만, 동일한 질문에서 3DSG가 낮은 점수를 기록한다는 점에서 문제의 난이도가 아닌 지식의 유무가 성능 차이의 원인임을 확인할 수 있다.

4.5 Ablation: 시맨틱 설명의 기여 분석

이상의 분석에서 KG의 추론 품질과 안정성 우위가 확인되었으나, 이 우위가 온톨로지의 구조적 요소(트리플, 계층, 관계)에 기인하는지, 아니면 rdfs:comment에 담긴 자연어 시맨틱 설명에 기인하는지는 분리 검증이 필요하다.

KG의 핵심 기여가 온톨로지 구조인지, 자연어 설명(rdfs:comment)인지를 분리 검증하기 위해 5가지 조건의 ablation 실험을 수행하였다: (1) LLM only—장면 컨텍스트 없이 질문만 제공, (2) KG structure-only—rdfs:comment를 제거하여 (평균 71%의 컨텍스트 감소) 트리플과 계층만 제공, (3) 3DSG—기존 baseline, (4) 3DSG+NL—3DSG 컨텍스트에 동일한 rdfs:comment를 추가, (5) KG full—전체 KG 컨텍스트. 조건 (4)는 "KG 프레임워크 없이도 자연어 설명만 추가하면 동등한 효과를 얻을 수 있는가?"라는 질문에 대한 직접적 검증이다.

Table 6 Five-condition ablation study results

Category	LLM only	KG struct	3DSG	3DSG+NL	KG full
Spatial	1.00	3.33	3.73	4.20	3.67
Identification	1.00	3.60	3.73	4.07	5.00
Semantic	1.40	3.09	3.49	3.89	4.26
Hierarchy	1.00	4.53	4.60	4.53	4.53
Safety	1.20	3.31	3.71	4.37	4.40
Dilemma	1.49	2.77	2.89	4.00	4.60
Overall	1.25	3.29	3.56	4.14	4.41

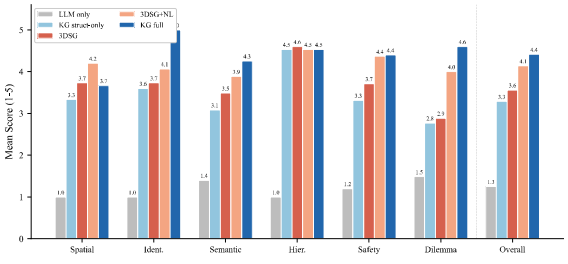

Fig. 6 Five-condition ablation results by category.

Table 6과 Fig. 6의 결과, 전체 평균 기준으로 LLM only(1.25) < KG structure-only(3.29) < 3DSG(3.56) < 3DSG+NL(4.14) < KG full(4.41)의 순서가 확인되었다. 이 순서는 세 가지 핵심 발견을 제공한다.

첫째, LLM only(1.25)는 모든 조건 중 최저 성능으로, LLM의 사전 학습 지식만으로는 특정 장면의 구체적 질문에 효과적으로 답변할 수 없음을 확인하였다. 이는 장면 컨텍스트 제공의 필요성을 정량적으로 증명한다.

둘째, 3DSG+NL(4.14)은 3DSG(3.56) 대비 +0.58점 향상되어, 자연어 시맨틱 설명의 추가가 표현 형식(KG vs 3DSG)에 관계없이 성능을 크게 향상시킴을 보여준다. 동시에 KG full(4.41)과의 차이(+0.27)는 KG 프레임워크가 제공하는 추가적 이점—SPARQL 기반 정밀 검색, 추론 엔진의 암묵적 지식 도출, 엔티티 단위 체계적 주석 관리—이 존재함을 시사한다.

셋째, KG structure-only(3.29)가 3DSG(3.56)보다 오히려 -0.27점 낮은 것은, rdfs:comment 제거 시 KG의 정보량이 3DSG의 풍부한 속성 정보보다 부족해지기 때문이다. 이는 KG의 성능 향상이 사실상 rdfs:comment 시맨틱 설명에 기인함을 명확히 보여준다.

카테고리별로 분석하면, Dilemma에서 3DSG+NL과 KG full의 차이가 가장 두드러지며, 이는 딜레마 추론에서 KG의 구조적 검색이 자연어 설명의 정확한 전달에 기여함을 시사한다. 반면 Hierarchy에서는 5가지 조건 모두 유사한 성능을 보여, LLM의 사전 학습 지식이 온톨로지 계층을 효과적으로 대체함을 재확인하였다.

이 결과는 본 연구의 핵심 기여를 보다 정확하게 정의한다. KG 프레임워크의 가치는 온톨로지 구조 자체보다, 도메인 전문 지식을 엔티티 단위로 체계적으로 부착·관리할 수 있는 인프라로서의 역할에 있다. 3DSG+NL 조건이 상당한 성능 향상을 보인 것은 자연어 설명의 효과를 입증하지만, KG가 이를 SPARQL 검색, 추론 엔진, 일관된 스키마를 통해 체계적으로 관리한다는 점에서 실용적 이점을 제공한다.

5. 논 의

5.1 핵심 시사점

본 연구의 결과는 세 가지 핵심 시사점을 제공한다. 첫째, 제안 기법은 지식 요구도가 높은 상황일수록 baseline 대비 추론 품질 우위가 뚜렷해지는 패턴을 보인다: Spatial(±0) → Semantic(+0.77) → Safety(+0.69) → Dilemma(+1.71). 이는 KG를 통한 도메인 지식 강화가 단순 지각을 넘어선 추론 과제에서 필수적임을 정량적으로 증명한다. Hybrid-Driving [4]의 37.5% 향상과 KnowVal [5]의 최저 충돌률이라는 정성적 주장을 본 연구의 통제된 실험을 통해 뒷받침한다.

둘째, Hierarchy 동등 결과와 ablation 실험은 KG 설계에 실용적 가이드라인을 제시한다. LLM이 사전 학습으로 이미 보유한 일반 지식(예: 신호등→TrafficControl 분류)에 대해서는 온톨로지 계층이 추가적 이점을 제공하지 않는다. 따라서 KG 설계 시 LLM이 모르는 상황별 맥락—딜레마 존 물리학, 방어운전 규칙, 법적 예외 상황—에 리소스를 집중하는 것이 효율적이다. 이러한 "LLM 보완 원칙(LLM-complementary principle)"은 제한된 KG 구축 비용으로 최대 효과를 달성하기 위한 전략적 지침이다.

셋째, Q3의 3DSG 우위는 공간적 열거 작업에서 3DSG의 강점을 보여주며, 실용적 시스템에서는 KG와 3DSG의 하이브리드 접근—공간 질의에는 3DSG, 의미·안전·딜레마 질의에는 KG—이 최적의 전략임을 시사한다. 이러한 하이브리드 전략은 KG 구축 비용을 절감하면서도 전체 질의 유형에 걸쳐 높은

추론 품질을 유지할 수 있다.

5.2 한계 및 향후 연구

본 연구는 LLM 기반 장면 질의응답(scene QA)과 추론 품질을 평가한 것으로, 폐쇄 루프 주행(closed-loop driving), planner/controller 연동, 또는 인지 불확실성(perception uncertainty) 반영을 포함하지 않는다. 따라서 본 연구가 입증한 것은 엄밀히 말해 "KG가 딜레마 상황의 의미 기반 질의응답과 추론 품질을 향상시킨다"는 수준이며, "실제 주행 의사결정 성능의 향상"으로 확대 해석하는 것은 적절하지 않다. 향후 CARLA 시뮬레이터 등을 활용한 폐쇄 루프 검증이 필요하다.

LLM-as-judge 방식은 재현 가능성과 확장성이 높으나 [16], 동일 모델 계열(OpenAI GPT 시리즈) 간 잠재적 선호 편향과 루브릭 해석의 불확실성이 존재한다. 본 연구에서는 답변 생성(GPT-4o-mini)과 평가(GPT-4o)에 서로 다른 모델을 사용하고 temperature=0.0으로 채점의 결정론적 일관성을 확보하여 자기 편향을 완화하였으나, 동일 OpenAI 계열이라는 근본적 한계가 있다. 향후에는 복수 평가자(인간 전문가, 이종 LLM judge) 및 평가자 간 일치도(inter-rater agreement) 분석을 병행할 필요가 있다.

전문가 정답(ground truth)은 제1저자가 작성하였으며, KG 설계자와 정답 작성자의 이중 역할로 인한 무의식적 편향 가능성을 배제할 수 없다. Spatial 카테고리의 동등 결과(-0.07)는 데이터 공정성의 내부 검증으로 기능하나, 향후 독립적 도메인 전문가(교통공학, 자율주행 엔지니어)에 의한 정답 검증이 필요하다.

본 연구는 단일 수작업 설계 시뮬레이션 환경에서 평가하였으며, 특정 환경에 대한 과적합(overfitting) 가능성을 배제할 수 없다. 25개 구역, 65개 객체, 3가지 딜레마 시나리오를 포함하는 비교적 복잡한 환경을 구성하여 단일 질문에 대한 과적합을 완화하고자 하였으나, 향후 CARLA, nuScenes 등 다양한 시나리오에서의 교차 검증이 필요하다.

추가적인 한계로는 (1) 정적 장면(snapshot)만을 다루어 시간적 변화(temporal dynamics)—신호 변경, 차량 가감속, 보행자 이동—를 반영하지 못하는 점, (2) KG 천장 효과(60%가 만점)로 인해 제안 기법의 상한 성능을 정확히 측정하지 못한 점, (3) 실시간 KG 갱신 메커니즘의 부재를 들 수 있다. 향후 연구에서는 동적 장면 기반 폐쇄 루프 검증, 다중 시나리오 확장, 인간 평가자 기반 교차 검증, 그리고 더 도전적인 시간적·다중 에이전트 추론 질문의 추가가 필요하다.

6. 결 론

본 연구에서는 자율주행 딜레마 상황에 특화된 KG 기반 시맨틱 장면 이해 프레임워크를 제안하고, 기존 3DSG 기반 접근을 baseline으로 그 효과를 정량적으로 검증하였다. 주요 결론은 다음과 같다.

1) 제안 기법(KG)은 전체 평균 4.41점으로 baseline(3DSG)의 3.56점을 통계적으로 유의미하게 상회하였다(Wilcoxon signed-rank, n=30, $p<0.001$, $r=0.70$, Cohen's $d=0.84$).

2) 딜레마 추론에서 제안 기법의 효과가 가장 두드러지며(+1.71, $p=0.016$), Spatial 카테고리의 동등 결과(-0.07)는 실험 설계의 공정성을 검증한다.

3) KG의 rdfs:comment를 통한 시맨틱 설명은 LLM의 추론 안정성을 크게 향상시키며(Q17: KG std=0.0 vs 3DSG std=1.67), 안전 관련 추론에서 일관된 판단을 가능하게 한다.

4) 5가지 조건의 ablation 실험 결과, LLM only(1.25) < KG struct-only(3.29) < 3DSG(3.56) < 3DSG+NL(4.14) < KG full(4.41)의 순서가 확인되었다. 자연어 시맨틱 설명이 성능 향상의 핵심 요인이며, KG 프레임워크는 이를 체계적으로 관리하는 인프라로서 가치를 갖는다.

5) Hierarchy 동등 결과에 기반한 "LLM 보완 원칙"은 KG 구축 시 LLM이 이미 보유한 일반 지식이 아닌, 도메인 특화 맥락(딜레마 물리학, 방어운전 규칙, 법적 예외)에 리소스를 집중해야 한다는 실용적 설계 원칙을 제시한다.

이러한 결과는 자율주행 시스템의 장면 이해에서 구조화된 도메인 지식이 필수 요소임을 실험적으로 증명한다. 다만, 본 연구는 장면 질의응답 및 추론 품질 수준의 검증이며, 향후 폐쇄 루프 주행 검증 및 다중 시나리오 확장이 필요하다.

향후 연구 방향으로는 첫째, 공간 질의에는 3DSG, 의미·안전·딜레마 질의에는 KG를 활용하는 하이브리드 접근의 구현, 둘째, CARLA 시뮬레이터와의 통합을 통한 폐쇄 루프 주행 검증, 셋째, 실시

간 지각 데이터로부터 KG를 자동 갱신하는 동적 KG(Dynamic KG) 메커니즘 개발, 넷째, 인간 전문가 평가와 이종 LLM judge를 통한 다중 평가자 교차 검증을 계획하고 있다.

Acknowledgment

본 연구는 연세대학교 교내 연구비와 모빌리티시스템융합공학과(현대자동차학과)의 지원을 받아 수행되었음.

References

[1] Greve, E., et al., "CURB-SG: Collaborative Dynamic 3D Scene Graphs for Automated Driving," Proc. IEEE ICRA, pp.11118-11124, 2024.

[2] Zhang, Y., et al., "GraphAD: Interaction Scene Graph for End-to-end Autonomous Driving," Proc. IJCAI, pp.2422-2430, 2025.

[3] Gazis, D. C., Herman, R. and Maradudin, A., "The Problem of the Amber Signal Light in Traffic Flow," Operations Research, Vol.8, No.1, pp.112-132, 1960.

[4] Wang, J., et al., "Hybrid-Driving: An Autonomous Driving Decision Framework Integrating LLMs, Knowledge Graphs and Driving Rules," Proc. AAAI, Vol.39, No.1, pp.826-833, 2025.

[5] Xia, Z., et al., "KnowVal: A Knowledge-Augmented and Value-Guided Autonomous Driving System," arXiv:2512.20299, 2025.

[6] Lv, C., et al., "T2SG: Traffic Topology Scene Graph for Topology Reasoning in Autonomous Driving," Proc. IEEE/CVF CVPR, pp.17197-17206, 2025.

[7] Wickramarachchi, R., Henson, C. and Sheth, A., "Knowledge Graphs of Driving Scenes to Empower the Emerging Capabilities of Neurosymbolic AI," Proc. ISWC, 2024.

[8] Mlodzian, R., et al., "nuScenes Knowledge Graph — A Comprehensive Semantic Representation of Traffic Scenes for Trajectory Prediction," Proc. IEEE/CVF ICCV Workshops, pp.42-52, 2023.

[9] Hussien, M., et al., "RAG-based Explainable Prediction of Road Users Behaviors for Automated Driving using Knowledge Graphs and LLMs," Expert Systems with Applications, Vol.265, Article 125914, 2025.

[10] Wen, L., et al., "DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models," Proc. ICLR, 2024.

[11] Sha, H., et al., "LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving," arXiv:2310.03026, 2023.

[12] Westhofen, L., Neurohr, C., Butz, M., Scholtes, M. and Schuldes, M., "Using Ontologies for the Formalization and Recognition of Criticality for Automated Driving," IEEE Open Journal of Intelligent Transportation Systems, Vol.3, pp.519-538, 2022.

[13] Bagschik, G., Menzel, T. and Maurer, M., "Ontology based Scene Creation for the Development of Automated Vehicles," Proc. IEEE IV, pp.1813-1820, 2018.

[14] Mao, J., et al., "A Language Agent for Autonomous Driving," Proc. COLM, 2024.

[15] Cai, T., et al., "Driving with Regulation: Interpretable Decision-Making for Autonomous Vehicles with Retrieval-Augmented Reasoning via LLM," arXiv:2410.04759, 2024.

[16] Zheng, L., et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," Advances in NeurIPS, Vol.36, 2023.

[17] Reed, N., et al., "Ethics of Automated Vehicles: Breaking Traffic Rules for Road Safety," Ethics and Information Technology, Vol.23, pp.777-789, 2021.

[18] Armeni, I., He, Z.-Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J. and Savarese, S., "3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera," Proc. IEEE/CVF ICCV, pp.5664-5673, 2019.

[19] Rosinol, A., Gupta, A., Abate, M., Shi, J. and Carlone, L., "3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans," Proc. RSS, 2020.