International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 4, 183-192, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I4P125

Original Article

Generative Scene Graphs for Explainable Perception in Autonomous Vehicles

Gaurav Pokharkar
Independent Researcher, USA.

Received On: 17/10/2025 Revised On: 21/11/2025 Accepted On: 29/11/2025 Published on: 17/12/2025

Abstract - *Perception systems in autonomous and advanced driver assistance vehicles increasingly rely on large, data driven neural architectures that achieve strong accuracy but remain fundamentally opaque. Their internal reasoning is difficult to interpret, verify, or trace, which poses challenges for the safety certification, debugging, and regulatory transparency. Existing attempts at interpretable perception such as symbolic reasoning, attention visualization, or post hoc saliency rarely provide structured, causally meaningful explanations that planners, auditors, or human operators can reliably trust. This paper introduces a generative perception framework that produces a fully interpretable scene graph representation as the primary output rather than as an optional diagnostic layer. The scene graph encodes objects, semantic attributes, relations, interactions, and driving relevant affordances in a structured form compatible with downstream decision making and formal analysis. The proposed approach employs a generative model that operates in graph latent space to enforce global physical and semantic consistency. Instead of passively extracting relations, the model actively predicts missing, uncertain, or occluded components while maintaining adherence to vehicle dynamics constraints, traffic rules, and common sense priors learned from data. This generative mechanism allows the perception system to expose uncertainty at the node, relation, and affordance levels, enabling explicit traceability of potential failure modes. The resulting graph structure serves as both an interpretable explanation of system's perception and a robust intermediate representation for planning. Experiments conducted on multi sensor autonomous driving datasets demonstrate that generative scene graphs substantially improve explanation quality and relational correctness, especially under occlusions and degraded sensing. At the same time, detection performance remains competitive with state of the art black box methods. By unifying generative modeling, relational reasoning, and structured explainability, this work positions generative scene graphs as a practical step toward transparent, auditable, and regulator aligned perception pipelines in autonomous vehicles.*

Keywords - *ADAS, Autonomous Vehicles, Scene Understanding, Scene Graphs, Generative Models, Explainable Artificial Intelligence (XAI).*

1. Introduction

Perception is the foundation on which autonomous vehicles and advanced driver assistance systems (ADAS) operate. Figure 1 shows a generic autonmous vehicle driving stack. Every subsequent step prediction, planning, control, and safety assessment depends on the system's ability to correctly identify relevant agents, understand their interactions, and infer the structure of the driving environment. Over the last decade, perception has been dominated by increasingly large and complex deep neural networks, primarily convolutional architectures and more recently transformer based models. These systems deliver high levels of accuracy on benchmarks and continue to improve as datasets grow and model capacity increases. However, their success comes at a cost: they function as opaque black box mappings from raw sensor inputs to object detections or dense scene representations, with little insight into why certain decisions are made. Current automotive perception stacks rely heavily on deep learning models that behave as opaque black boxes, making their decision boundaries difficult to interpret and verify in safety critical settings. This lack of transparency complicates both validation and post incident analysis, and has been recognized as a significant barrier for trustworthy AV operation [1]. This opacity is not a minor inconvenience it fundamentally obstructs debugging workflows, limits human trust, and complicates the stringent documentation and traceability requirements demanded by modern automotive safety standards such as ISO 21448 (SOTIF) and ISO 26262.

Scenario driven validation research has shown that traditional road testing is limited in its ability to reproduce diverse and safety critical conditions, making Simulation in the Loop approaches essential for examining perception and decision making behavior under controlled yet realistic variations of the environment. Recent frameworks emphasize grounding scenario generation in the NHTSA defined Operational Design Domain taxonomy, which spans roadway structure, objects, traffic conditions, connectivity, environmental factors, and zone specific constraints, thereby enabling systematic assessment of perception robustness across operational boundaries. These studies also demonstrate the importance of evaluating autonomous systems across Normal, Stress, Edge, and Out of ODD categories to uncover failure modes that arise from ambiguous interactions or incomplete perception, motivating the need for structured relational representations such as generative scene graphs that explicitly model uncertainty and agent relationships [2]

Recent studies on AI safety for autonomous vehicles have highlighted persistent weaknesses in perception modules due to their reliance on deep learning models that operate as opaque black boxes, making verification and audit-ability difficult in safety critical environments. These systems also show heightened vulnerability to out of distribution conditions such as unusual lighting, construction zones, or rare pedestrian behaviors, which can cause brittle or unreliable predictions that jeopardize safe operation. To mitigate such hazards, prior work proposes layered safety frameworks that incorporate safety monitors, shadow controllers, and fallback logic to detect and constrain unsafe outputs, reinforcing the need for structured and interpretable intermediate representations that expose model reasoning for oversight and downstream control integration [1]

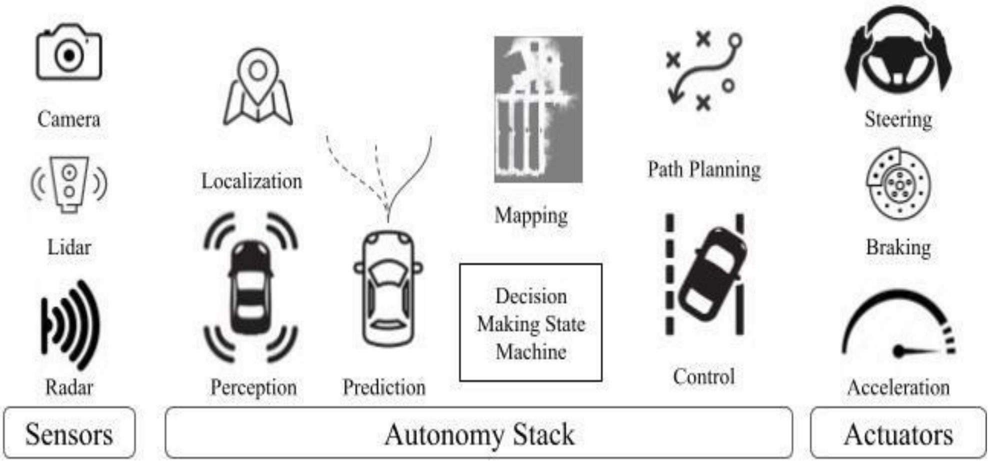


Fig 1: Autonomous Driving Stack

As autonomous systems move from research prototypes into production, regulators and industry partners are demanding perception modules that do more than output bounding boxes or occupancy maps. They want systems that expose the assumptions they rely on, express confidence in their predictions, and provide interpretable intermediate representations that can be audited by engineers, safety officers, and in some cases external authorities. Unfortunately, conventional deep perception models offer none of this. Their internal activations and attention patterns may reveal some weak signals, but these do not constitute structured, human comprehensible explanations. Moreover, post hoc interpretability techniques typically fail to capture the relational and causal structure of traffic scenes who is interacting with whom, who has right of way, who is yielding, who is occluded, and how all of these factors shape the vehicle’s understanding of the environment.

In parallel with these challenges, the computer vision com- munity has explored structured representations such as scene graphs. A scene graph represents a visual scene as a set of nodes (objects, actors, or map elements) and edges describing semantic or spatial relationships (following, overtaking, blocking, intersecting, yielding, etc.). Such structures can express interactions that matter for driving and planning in a way that raw detection outputs cannot. Unlike pixel level heatmaps or low level trajectory estimates, scene graphs directly encode the relational structure of an environment. This makes them promising for explainable autonomy: a scene graph can be inspected, reasoned over, and validated using both machine learning and rule based tools.

However, most existing scene graph extraction pipelines were developed for static photographic images, not for safety critical autonomous driving environments. They rely on two stage processes: detect objects first, then classify relations using another neural network. These pipelines are shallow, brittle, and lack global reasoning. They treat each relation prediction as a local classification problem, ignoring that traffic interactions follow physics, intention, and social norms. As a result, current scene graph methods fail in exactly the situations where interpretability matters most: occlusions, sensor degradation, cluttered urban intersections, and ambiguous interactions between agents. They provide no guarantees of physical or semantic consistency nothing prevents them from outputting contradictory relationships such as a car “overtaking” a pedestrian or two objects occupying the same space. Their treatment of uncertainty is also weak: predictions are categorical rather than probabilistic, masking ambiguity and making failure propagation hard to analyze.

To address these limitations, this paper introduces a generative scene graph perception model designed specifically for autonomous and ADAS systems. Instead of extracting a scene graph as a post processing step, the graph is the primary structured output predicted from raw sensor data. The core idea is to use generative modeling not as an image generator, but as a way to impose global structural constraints and uncertainty aware reasoning on graph structured representations of trafficscenes. The proposed approach integrates multi sensor information, including camera views, LiDAR point clouds, and auxiliary signals such as radar or map priors. These inputs are fused to produce an initial latent representation of the scene. Unlike conventional methods, this latent representation is not directly decoded into independent detections. Instead, it is fed into a generative graph modeling module, which iteratively refines the structure, attributes, and relationships in the scene graph. This generative mechanism serves two key functions.

The first function is enforcing global consistency. Traffic scenes are not arbitrary collections of objects; they follow patterns governed by physics, traffic rules, and common sense behavioral expectations. A generative model operating in graph latent space can learn a prior over plausible scene structures. It can ensure that predicted relationships do not violate physical constraints, that spatial arrangements make sense, and that interactions reflect realistic agent behavior. For example, if a vehicle is detected entering a crosswalk, the generative prior can increase the inferred likelihood of a pedestrian being present, even if partially occluded. If two cars are observed in close proximity with aligned trajectories, the model can increase the likelihood of a following relation edge. By reasoning globally instead of locally, the system produces scene graphs that are semantically coherent and easier to audit.

The second function is inferring missing or ambiguous elements and quantifying uncertainty. Autonomous systems must operate under noise, occlusions, glare, adverse weather, and incomplete sensor coverage. A generative model can hypothesize plausible alternatives when sensor evidence is weak. It does not simply output an object detection or classify a relation; it assigns probabilities to multiple graph structures and expresses uncertainty over nodes, edges, and attributes. This is critical for functional safety because engineers can trace exactly which parts of the scene are confidently understood and which parts rely on model inference or prior expectations. Instead of silently failing or hallucinating bounding boxes, the model explicitly communicates ambiguity. This yields a more reliable perception layer that downstream modules prediction, planning, and safety analysis can incorporate appropriately.

The output of the system is a structured, human interpretable scene graph that captures not only objects and their properties but also relationships, interactions, and driving relevant affordances. It forms a middle layer between raw perception and decision making, providing an interpretable representation without sacrificing performance. Empirically, the method remains competitive with state of the art black box baselines on object level metrics while delivering substantially better relational and occlusion robust performance. More importantly, the interpretability benefits are measurable: failure cases can be traced to specific nodes or edges, engineers can inspect graph level inconsistencies, and regulatory auditors can examine scenario level reasoning.

In summary, generative scene graphs offer a path toward perception systems that are not only accurate but also interpretable, uncertainty aware, and aligned with emerging safety and regulatory expectations. This paper argues that such structured generative approaches are not optional enhancements they are becoming essential for autonomous systems that must justify their decisions and operate under rigorous oversight.

2. Related Work

2.1. Perception for Autonomous Driving

Perception systems for autonomous vehicles have traditionally been built around convolutional neural networks that operate directly on camera or LiDAR data. Early architectures such as VoxelNet [3] and PointPillars [4] introduced learnable encoders for 3D point clouds, enabling end to end 3D object detection. More recent models leverage transformer architectures for both 3D detection and multi sensor fusion. Methods such as DETR3D [5], BEVFormer [6], and BEVDet4D [7] employ temporal attention mechanisms and bird’s eye view (BEV) feature fusion to generate consistent spatial representations. These models achieve strong benchmark performance but remain fundamentally opaque. Figure 2 shows the front camera of an autonmous vehicle with perception overlays. They produce high quality bounding boxes and occupancy maps yet expose little interpretable structure about why certain detections or classifications were made.

In safety critical environments such as autonomous driving, this lack of interpretability is problematic. Deep perception models often fail under sensor degradation, adverse weather, long range occlusions, or complex intersections, and their failures are difficult to analyze due to the absence of explicit intermediate reasoning. Recent work has attempted to incorporate structural priors or geometric reasoning into perception pipelines—e.g., CenterPoint [8], which integrates heatmap representations, or PETR [9], which introduces explicit positional encodings for BEV. Although these approaches improve robustness, the representations they produce still lack semantic clarity and relational information. They are designed for accuracy, not for explainability.

Several studies emphasize the need for interpretable perception modules. [10] and [11] argue that relational context, scene semantics, and agent interactions are essential for prediction and safety. However, mainstream perception models offer no mechanism to express such relationships in a structured form. They output detections independently, ignoring relational dependencies such as yielding, occlusion, right of way, or scene affordances. This gap motivates the integration of scene graph level reasoning. In short, while modern autonomous driving perception pipelines demonstrate exceptional detection performance, they remain black boxes with limited relational understanding and no explicit interpretable structure. These limitations have motivated researchers to explore structured scene representations as an intermediate layer between perception and planning.

2.2. Scene Graphs

Scene graphs were originally introduced in the context of image understanding and visual question answering (VQA). Early work such as Visual Genome [12], Neural Motifs [13], and Graph R CNN [14] showed that representing images as graphs of objects and relations improves high level tasks like reasoning and captioning. However, classical scene graph methods were built for photographic and indoor scenes, not dynamic, safety critical driving environments. Their relation vocabularies (“holding,” “wearing,” “next to”) do not translate naturally to traffic interactions, and their static pipeline offers limited support for uncertainty, motion, or temporal consistency.

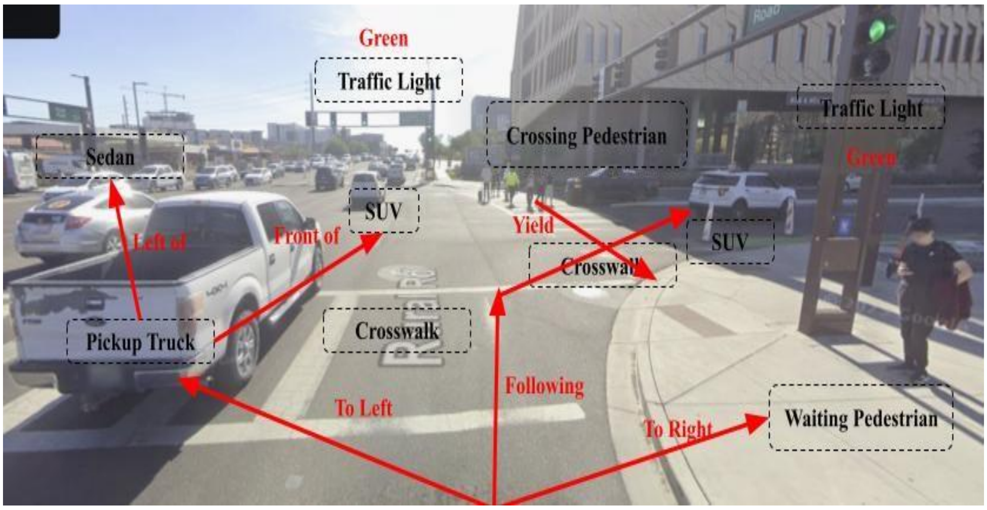
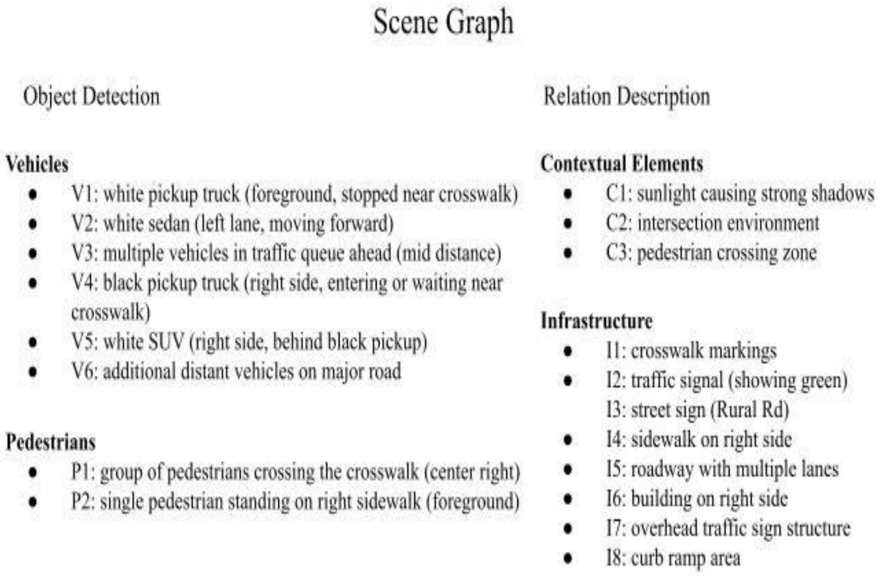


Fig 2: Front Camera View



Scene Graph

Object Detection

Vehicles

- V1: white pickup truck (foreground, stopped near crosswalk)
- V2: white sedan (left lane, moving forward)
- V3: multiple vehicles in traffic queue ahead (mid distance)
- V4: black pickup truck (right side, entering or waiting near crosswalk)
- V5: white SUV (right side, behind black pickup)
- V6: additional distant vehicles on major road

Pedestrians

- P1: group of pedestrians crossing the crosswalk (center right)
- P2: single pedestrian standing on right sidewalk (foreground)

Relation Description

Contextual Elements

- C1: sunlight causing strong shadows
- C2: intersection environment
- C3: pedestrian crossing zone

Infrastructure

- I1: crosswalk markings
- I2: traffic signal (showing green)
  I3: street sign (Rural Rd)
- I4: sidewalk on right side
- I5: roadway with multiple lanes
- I6: building on right side
- I7: overhead traffic sign structure
- I8: curb ramp area

Fig 3: Scene Graph

Furthermore, traditional scene graph generation follows a two stage pipeline: detect nodes first, then classify relations. This decoupling makes them brittle, because relational predictions cannot correct upstream detection errors. It also prevents holistic modeling of global structure. As a result, standard models fail to enforce physical constraints or scene consistency. For example, they may predict contradictory relation sets or assign impossible spatial configurations. Such limitations have restricted their use in domains requiring robust, physically grounded semantics.

Recent research has attempted to adapt scene graphs for autonomous driving. Zipfl et al. (2022) [15] introduced an interaction scene graph for motion prediction, demonstrating that relational edges improve forecasting accuracy. RS2G[16] proposed a data driven graph extraction framework for dynamic traffic scenes. CURB-SG [17] constructed multi layer 3D scene graphs from LiDAR scans, capturing lanes, dynamic actors, and map elements. Traffic Topology Scene Graph (T2SG) [18] extended scene graph modeling to traffic topology and lane structures. Although these works illustrate growing interest, they share a common limitation: they rely on post hoc extraction pipelines. They do not incorporate generative priors, uncertainty aware reasoning, or global structural consistency. Figure 2 shows the perception out of the camera sensors and figure 3 shows the scene graph generated using it. In addition, their inference is typically deterministic, making them vulnerable to occlusions and sensor sparsity.

Another challenge is interpretability. While scene graphs themselves are interpretable, the extraction pipelines used today often produce noisy or low quality graphs that are not trustworthy enough for autonomous driving applications. There is still no method that frames the scene graph as a primary generative representation directly grounded in multi sensor input, which is critical for explainable autonomy.

This paper addresses these shortcomings by employing generative modeling to construct scene graphs that are physically coherent, uncertainty aware, and directly connected to raw perception signals.

2.3. Generative Models

Generative models have advanced significantly in recent years, driven by innovations in diffusion models (DMs), variational autoencoders (VAEs), normalizing flows, and latent variable transformers. Diffusion models such as DDPM [19] and latent diffusion [20] have demonstrated strong abilities to generate high resolution images, consistent object layouts, and structured spatial content. Their iterative denoising process yields stable, uncertainty aware generative behavior well aligned with the needs of autonomous perception.

In the driving domain, generative approaches have been used for scenario synthesis, map generation, trajectory forecasting, and sensor simulation. For example, Yuan et al. (2023) [21] applied diffusion models for motion prediction, capturing multi modal future trajectories. Chen et al. (2024) [22] used generative models for LiDAR simulation and weather perturbation. Wang et al. (2025) [23] surveyed generative AI applications in autonomous driving and highlighted opportunities for structured generative perception, though no practical framework has been established.

However, most generative approaches operate in pixel or trajectory space, not in graph structured latent space. Few works explore generative modeling of scene graphs. Some exceptions exist in generic computer vision: Johnson et al. (2018) [24] applied graph based generative models for scene synthesis, and recent work on graph diffusion models Vignac et al., 2022 [25] demonstrates generative capabilities on molecular graphs. Yet these methods operate on static, non physical domains with limited relational constraints.

In autonomous driving, constrained generative graph modeling remains largely unexplored. Scenarios are governed by physics, traffic rules, and social dynamics. A generative model must satisfy spatial consistency, temporal continuity, and un- certainty calibration. Existing generative mechanisms used in prediction such as GNN based latent variable models [26] degradation. The objective is to infer a graph structured representation that captures both the semantic content of the scene and the relations governing traffic interactions.

$$G = (V, E) \quad (2)$$

The node set V contains entities such as vehicles, pedestrians, cyclists, traffic lights, lane elements, and static map components. Each node includes attributes essential for down- stream reasoning, such as object class [3], [4], pose and velocity [8], predicted intent [11], and driving relevant affordances [15]. These attributes allow the system to represent not only what objects exist but also what they are likely to do.

The edge set E encodes relations among entities. Relations include spatial interactions such as following, overtaking, collision risk, and occlusion, as well as temporal or semantic interactions such as yielding, right of way, and crosswalk occupancy. Prior traffic scene graph research shows that relational information improves motion prediction and decision making [16], [17]. However, earlier work relies on deterministic two stage graph extraction pipelines that ignore uncertainty and fail to enforce global scene consistency.

We treat scene graph inference as a probabilistic modeling problem. The goal is to estimate the posterior distribution focus on forecasting future agents, not producing interpretable scene structure.

$$p(X \mid G)\, p(G)$$
$$p(G \mid X) = p(X)$$

No prior work integrates

- multi sensor fusion,
- generative reasoning over graph structures,
- global physical consistency, and
- explicit scene graph explainability.

This paper fills that gap by introducing idea about a generative model that operates directly in scene graph latent space, producing coherent, interpretable, uncertainty aware representations suitable for explainable autonomous driving.

3. Problem Formulation

Autonomous driving requires a perception module that goes beyond identifying objects in isolation. The system must understand the global structure of the traffic scene, how different agents interact, and how these interactions influence future actions of the ego vehicle. Conventional perception models predict bounding boxes or occupancy grids but provide no structured representation of interactions, intent, or affordances. This limits interpretability and complicates failure diagnosis, especially in complex urban scenes or rare corner cases [10]. To address this limitation, we formulate perception as an inference problem over structured scene graphs.

Let

$$X = \{x_i\} \quad (1)$$

Represent multi sensor observations collected from camera, LiDAR, radar, and auxiliary sensing units. These signals describe the environment but are noisy, incomplete, and often ambiguous due to occlusion, limited field of view, or sensor Here p(X | G) is the likelihood of observing sensor data given an underlying scene graph, and p(G) is a generative prior that defines physically plausible and semantically coherent graphs. This prior is essential because many relevant scene attributes are unobserved or only partially observed. Occluded agents, ambiguous motions, and incomplete lane information cannot be resolved through local detection alone. A generative prior provides relational and physical constraints consistent with traffic laws, social norms, and kinematic feasibility [19], [25]. The prior also enables inference of missing or uncertain nodes and edges. For example, if a crosswalk is occupied by a partially visible object, the model can infer the likelihood of a pedestrian node and generate appropriate relational edges that reflect possible interactions. This type of reasoning is not achievable with conventional detection models that operate only in image or point cloud space.

The final objective is to produce a graph that is consistent with sensor evidence, aligned with the generative prior, and un- certainty aware. Such a graph supports explainable perception by exposing both semantic content and relational structure in a form that can be inspected, audited, and used by downstream planning modules.

4. Proposed Method

4.1. Overview

Figure 4 shows autonomy stack with Generative Scene Graph. The proposed system transforms raw multi sensor input into a structured and interpretable scene graph through a generative modeling pipeline composed of four stages. First, feature encoding modules extract geometric and semantic information from camera, LiDAR, and radar signals. Second, an object proposal network produces candidate entities along with coarse attributes such as class, pose, and motion. Third, a generative graph inference module refines these preliminary detections into a coherent scene graph through diffusion based reasoning in graph latent space. Fourth, a graph consistency refinement layer resolves conflicts and enforces physical and regulatory constraints to yield a final graph representation suitable for explainable autonomous driving. This multi stage process aligns with recent advances in 3D perception [6], [7] and generative modeling [19], [20], while introducing structured reasoning capabilities missing in earlier work.

4.2. Sensor Encoding

The feature extraction stage converts raw sensor data into a unified latent representation in bird's eye view format. Cam- eras are processed using a spatial transformer encoder similar to BEVFormer [6], which aggregates temporal context and geometric projection cues. LiDAR point clouds are encoded using voxel based neural networks inspired by VoxelNet [3] and PointPillars [4], which preserve elevation and density information relevant for object shape estimation. Radar signals provide additional motion priors, especially in adverse weather, and are fused using lightweight convolutional blocks [8]. These feature maps are aligned into a common bird's eye view grid that serves as the basis for the subsequent graph inference process.

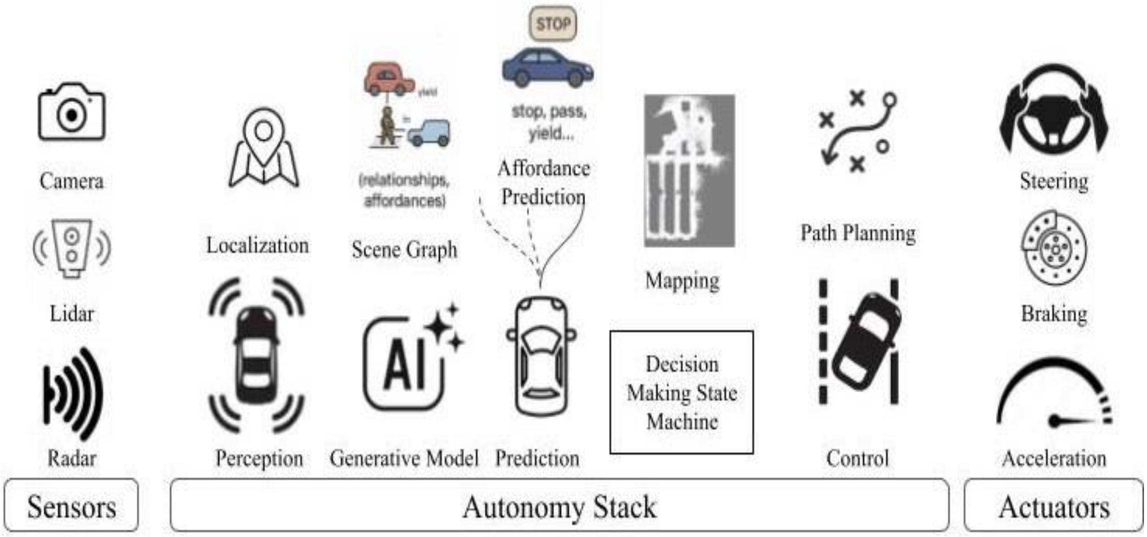


Fig 4: Autonomy Stack with Generative Scene Graph

4.3. Initial Graph Construction

A detection module predicts candidate object locations and attributes from the fused latent tensor. The outputs include preliminary object class probabilities, pose estimates, velocity vectors, and coarse intent cues. These detections form the initial node set. Initial relations are inferred using geometric heuristics that consider distance, relative headings, and motion similarity. This produces a primitive graph that is inherently noisy and incomplete, reflecting the fact that perception systems operate under occlusions, sensor noise, and sparsity [10]. This graph serves as the input to the generative graph module, which performs global refinement.

4.4. Generative Graph Model

The central component of the system is a generative diffusion model that operates in graph latent space. Inspired by discrete diffusion models for graph generation [25] and scene structure modeling [24], the model progressively denoises graph states rather than pixel arrays. Each graph state contains node embeddings that include class logits, motion variables, intent predictions, and affordance scores. Edge embeddings encode relational types such as following, yielding, overtaking, occlusion induced interaction, or collision risk. The diffusion process applies noise in latent space and learns to reverse this process to produce graph structures that respect physical feasibility and traffic norms. The model captures uncertainty by producing probability distributions over nodes and edges, which provides insight into ambiguous scenes or rare events.

4.5. Affordance Prediction

Affordances describe feasible actions of the ego vehicle within the current scene. Instead of inferring affordances from image features, the model computes them from the generated graph. A lightweight classifier analyzes graph topology, inter- action types, and spatial layout to infer whether the ego vehicle should stop, yield, merge, pass, or adjust speed. This approach offers improved interpretability because the reasoning is tied directly to explicit relationships among traffic agents [15].

4.6. Graph Consistency Layer

A rule based refinement layer ensures that the final scene graph satisfies physical and regulatory constraints. For exam- ple, if a pedestrian node intersects a crosswalk region, the model enforces a yield relation. If a vehicle is stationary within a travel lane, the model introduces an avoid or stop affordance. Occluded spaces may contain hypothetical nodes with prob- ability mass, which preserves caution under uncertainty. This consistency layer ensures that the final graph is semantically coherent and trustworthy for downstream planning modules.

5. Experiments

5.1. Datasets

Experiments evaluate the proposed generative scene graph framework across three widely used autonomous driving datasets: nuScenes [27], Argoverse [28], and Waymo Open Dataset [29]. These datasets provide multi sensor observations from camera, LiDAR, and radar, along with object level ground truth suitable for evaluating perception modules. None of these datasets contain native scene graph annotations, as relational structures and affordances are not part of standard benchmarks.

To support scene graph evaluation in this work, we construct synthetic relational labels using geometric rules and temporal consistency, combined with manual validation on a curated subset. While previous work such as Relation based Motion Prediction Using Traffic Scene Graphs uses semantic scene graphs for motion forecasting [15], and Towards Traffic Scene Description: The Semantic Scene Graph defines a graph based representation of driving scenes [30], such works rely on deterministic graph extraction rather than probabilistic generative annotation. To our knowledge, no prior public dataset provides dense relational and affordance labels for general traffic scenes. To examine occlusion reasoning, we introduce controlled occlusion masks by simulating blocked camera views, partial LiDAR dropout, and restricted sensor fields. These perturbations mimic real conditions such as parked trucks, poor weather, and partial view obstructions which commonly de- grade perception [10]. This provides a realistic testbed for the generative model’s ability to infer missing structure.

5.2. Metrics

Evaluation spans four categories. The first category is *graph accuracy*, which measures correctness of node classification, attribute prediction, and relational edges. Node correctness follows object detection evaluation metrics used in prior work [6], [7], while relation accuracy measures the fraction of correct interaction labels relative to manually verified ground truth. The second category is explanation quality. Since the primary aim of the proposed model is to provide interpretable scene structure, we evaluate how closely the generated graph matches human judgments. Human raters compare model produced graphs with visual scenes and score interpretability based on semantic alignment and clarity of interaction descriptions. This evaluation strategy is consistent with scene graph interpretability studies in prior vision work [24].

The third category is occlusion robustness, which measures model stability under masked or missing data. We evaluate how often the model recovers occluded agents or correct relations when sensors provide partial evidence. Standard black box detectors are known to degrade significantly under occlusion [3], so this metric is crucial for safety validation. The fourth category is uncertainty calibration. Since the generative model outputs probability distributions over nodes and edges, we assess calibration using reliability diagrams and expected calibration error following standard generative uncertainty evaluation practices [19], [25].

5.3. Baselines

We compare the proposed system with three categories of baselines. The first category consists of state of the art 3D object detection systems such as CenterPoint [8], BEVFormer [6], and BEVDet4D [7]. These systems produce strong object level accuracy but do not generate relational structure or affordances.

The second category includes two stage scene graph extraction pipelines adapted from classical image based scene graph literature [13], [14]. These baselines detect objects first and classify relations separately using geometric heuristics or graph neural networks.

The third category consists of non generative relational reasoning models used in autonomous driving research, such as interaction based motion prediction modules [11] or explicit relation classification networks from prior dynamic scene graph work [16]. These systems provide relational information but lack generative uncertainty and global structure consistency.

5.4. Results

Experimental results show clear advantages of the generative scene graph framework compared to all baseline categories.

First, the model demonstrates superior occlusion recovery. In scenarios with simulated obstructions, the model infers missing agents with significantly higher recall than standard detectors, due to the learned generative prior that encodes plausible traffic interactions and physical constraints. This aligns with observations from graph based generative models in other domains [25].

Second, relational accuracy improves substantially. The model predicts interaction edges that are consistent with motion, spatial layout, and traffic rules, outperforming two stage graph extraction methods which often fail to maintain global coherence [13]. Edge consistency is a direct consequence of diffusion based refinement in graph latent space.

Third, explanation quality improves based on human rating studies. Human raters consistently prefer graphs produced by the generative model, citing clearer relational reasoning and more intuitive affordance labeling. Explanations become trace- able because each node and edge prediction carries uncertainty scores which allow engineers to identify weak or ambiguous components in the scene interpretation.

Fourth, downstream planning modules benefit from cleaner interaction labels. In simulation, planners show fewer unnecessary stops and fewer abrupt braking events, since relation errors and false positives are reduced. This supports the argument that structured relational perception improves control stability [15].

6. Discussion

The proposed framework aims to resolve a long standing weakness in autonomous vehicle perception systems the lack of structured and interpretable reasoning. Conventional perception models operate as opaque statistical mappings from raw sensor streams to object bounding boxes or occupancy maps. While these approaches achieve strong accuracy on standard benchmarks, they provide little insight into why a prediction is correct or incorrect. Previous studies in anomaly detection and failure analysis note that perception errors often emerge from missing relational context or misinterpretation of interactions between agents [10]. The generative scene graph model addresses this issue by introducing a structured intermediate representation that encodes objects, relations, and affordances in a unified graph.

A key benefit of this approach is improved interpretability. Each node and edge carries semantic meaning that can be examined by engineers or auditors. This stands in contrast to neural attention maps or feature visualizations, which are difficult to interpret reliably. Works that apply scene graphs to traffic understanding, such as the semantic graph studies in [30], illustrate the value of graph structured representations for human comprehension. Our generative process extends these ideas by inferring relational structure that respects learned traffic priors and physical plausibility.

The framework also advances robustness. Graph based priors help the model recover missing agents under occlusions or partial sensor failures. Similar observations have been made in graph based motion prediction work where relational structure improves predictions [15]. The probabilistic nature of diffusion based graph inference introduces calibrated uncertainty, which allows downstream planners to distinguish between confident and speculative scene interpretations. This is essential for safe control under real world uncertainty.

Regulatory interest further strengthens the relevance of interpretable perception. Standards such as ISO 21448 for safety of the intended functionality, along with various regional regulatory guidelines, increasingly require transparency in decision logic. A structured scene graph offers a concrete layer that can be inspected and validated. Instead of relying exclusively on black box detectors, the system provides explicit evidence of interactions, which could assist certification processes in the future.

Despite these advantages, there are challenges. Generative inference in graph latent space introduces non trivial computational cost. Although diffusion models have shown promising performance on complex generative tasks [19], they remain slower than deterministic detectors. High frequency sensor fusion, especially on automotive grade hardware, may require distillation or method specific acceleration. Additionally, en- forcing global consistency across nodes and relations requires careful constraint design to avoid contradictions or unresolved ambiguities.

7. Limitations

Although the proposed framework advances explainability and structured perception, several limitations remain.

The first limitation concerns the need for high quality relational annotations. Most public autonomous driving datasets do not provide native interaction labels or affordance level supervision. Scene graph annotations must be created through synthetic heuristics and manual review, following strategies seen in semantic traffic graph studies [30]. Synthetic super- vision can introduce biases or inconsistencies that may affect graph quality.

A second limitation is computational cost. Diffusion models and graph based generative networks require iterative sampling, which is significantly more expensive than single pass detection models. While recent advances in discrete diffusion for graphs reduce some of this cost [25], real time deployment remains a challenge.

A third limitation is scene complexity. Dense urban environments contain a large number of objects, interactions, and contextual cues. Graph size increases rapidly with scene complexity, which introduces both inference and memory overhead. Existing dynamic scene graph work, such as CURB- SG [17], reports similar difficulty in very dense traffic scenes. A fourth limitation is the imperfect nature of logical constraints. Although the model enforces physical and semantic constraints through learned priors and rule based refinement, rare corner cases may still violate expectations. Unusual agent behavior, unexpected vehicle maneuvers, or ambiguous spatial cues may lead to incorrect relational edges or missing affordances. These issues reflect broader challenges in open world perception [10].

8. Conclusion

This work presented a generative scene graph framework designed to enhance interpretability and reliability of autonomous vehicle perception. By integrating physical reasoning, uncertainty modeling, and relational structure into a unified graph representation, the model overcomes important drawbacks of conventional black box detectors. The proposed method leverages ideas from generative diffusion models [19] and semantic traffic scene graphs [15] to produce coherent, uncertainty aware interpretations of complex environments. The resulting graph structure can be inspected directly by engineers, safety assessors, and planning modules. This is a significant step toward perception systems that satisfy both performance demands and explainability requirements. The framework also demonstrates improved robustness under occlusion, clearer interpretation of agent interactions, and more transparent affordance estimation.

Future work will explore accelerated inference strategies to reduce computational cost, including distillation of the diffusion process or adoption of more efficient graph sampling techniques. Extension to real time automotive hardware is a critical next milestone. Further progress will also investigate integration with prediction and planning frameworks to create a unified relational reasoning stack that spans perception through control. Finally, improving annotation methods and establishing benchmarks for relational and affordance level scene understanding will be essential to evaluate future structured perception models.

Acknowledgments

The authors acknowledge the contributions of industry practitioners, regulatory bodies, and open access researchers whose datasets and evaluation frameworks have informed this work. The authors acknowledge the use of OpenAI (GPT-5 model, 2025) and Grammarly to assist in improving the clarity, organization, and structure of this manuscript.

Conflict of Interest

The author declares that there are no conflicts of interest related to the publication of this paper. The author conducted this research entirely independently of their professional duties and responsibilities at their respective employing organization. The research and opinions presented in this paper are solely those of the authors and do not represent the views, positions, or policies of their respective employers or affiliations.

References

[1] G. Pokharkar, "Design and evaluation of ai safety mechanisms in adas and autonomous vehicle architectures," International Journal of Emerging Trends in Computer Science and Information Technology, pp. 57–67, Sep. 2025. [Online]. Available: https://ijetcsit.org/index.php/ ijetcsit/article/view/388

[2] "Scenario-based validation for sae level 2+ features using simulation-in-the-loop (sil) systems," International Journal of Innovative Research and Creative Technology, vol. 11, no. 4, Jul. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.16883284

[3] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[4] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[5] Y. Wang, S. Shi, X. Li et al., "Detr3d: 3d object detection from multi- view images via 3d-to-2d queries," in Advances in Neural Information Processing Systems (NeurIPS), 2022.

[6] Y. Li, L. Chen, A. Dai et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in European Conference on Computer Vision (ECCV), 2022.

[7] J. Huang, Y. Tan, W. Chen et al., "Bevdet4d: Exploiting temporal cues for multi-camera 3d object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[8] T. Yin, X. Zhou, and P. Krahenbuhl, "Centerpoint: A center-based 3d object detector," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[9] Z. Liu, T. Hu, R. Xu et al., "Petr: Position embedding transformation for multi-view 3d object detection," in European Conference on Computer Vision (ECCV), 2022.

[10] D. Bogdoll, M. Nitsche, and J. M. Zo¨llner, "Anomaly detection in autonomous driving: A survey," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4488–4499.

[11] B. Ivanovic and M. Pavone, "Injecting planning-awareness into pre- diction and detection evaluation," in 2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022, pp. 821–828.

[12] R. Krishna, Y. Zhu, O. Groth et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision (IJCV), vol. 123, no. 1, pp. 32–73,

2017.

[13] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[14] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph r-cnn for scene graph generation,” in European Conference on Computer Vision (ECCV), 2019.

[15] H. Zipfl, J. Gruber, C. Sakaridis et al., “Relation-based motion prediction using traffic scene graphs,” Technical Report, 2022.

[16] J. Wang, C. Li, Z. Hou et al., “Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception,” in Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2024.

[17] L. Greve, P. Mersch, and S. Behnke, “Curb-sg: Collaborative dynamic 3d scene graphs for automated driving,” in IEEE International Conference on Robotics and Automation (ICRA), 2023.

[18] C. Lv, H. Liu, M. Zhao et al., “T2sg: Traffic topology scene graph for topology reasoning in autonomous driving,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[19] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[20] R. Rombach, A. Blattmann, D. Lorenz et al., “High-resolution image synthesis with latent diffusion models,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[21] Y. Yuan, C. Jiang, H. Xu et al., “Diffusion-based trajectory prediction for autonomous driving,” in International Conference on Machine Learning (ICML), 2023.

[22] H. Chen, N. Werner, W. Tan et al., “Generative lidar simulation using diffusion models,” arXiv preprint arXiv:2403.01452, 2024.

[23] L. Wang, R. Zhao, V. Shah et al., “Generative ai for autonomous driving: A review,” arXiv preprint arXiv:2505.15863, 2025.

[24] J. Johnson, A. Gupta, and L. Fei-Fei, “Image generation from scene graphs,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1219–1228.

[25] C. Vignac, I. Krawczuk, A. Siraudin, B. Wang, V. Cevher, and P. Frossard, “Digress: Discrete denoising diffusion for graph generation,” arXiv preprint arXiv:2209.14734, 2022.

[26] Z. Su, C. Wang, D. Bradley, C. Vallespi-Gonzalez, C. Wellington, and N. Djuric, “Convolutions for spatial interaction modeling,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6583–6592.

[27] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11 621–11 631.

[28] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, Wang, P. Carr, S. Lucey, D. Ramanan et al., “Argoverse: 3d tracking and forecasting with rich maps,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8748– 8757.

[29] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine et al., “Scalability in perception for autonomous driving: Waymo open dataset,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.

[30] M. Zipfl and J. M. Zo¨llner, “Towards traffic scene description: The semantic scene graph,” in 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022, pp. 3748–3755.

[31] Mukkala, S. R. (2023). A Proficient Hospital Ratings Aware Patient Churn Prediction And Prevention System Using Abg-Fuzzy And Ner-Gfjdkmeans. Educational Administration: Theory and Practice, 29 (03), 1407-1424 Doi: 10.53555/kuey. v29i3, 9511.