Unified Point Cloud Corruption-Aware Reasoning Engine

Grafika Jati[1], Martin Molan[1], Francesco Barchi[1], and Andrea Acquaviva[1]

***Abstract*— LiDAR sensors are susceptible to surface contamination such as mud, which compromises reliability, yet existing deep learning models fail to provide interpretable diagnostics for such failures. Furthermore, standard Vision–Language Models (VLMs) lack the geometric priors required to reason over LiDAR-specific sparsity and occlusion artifacts. To address this, we present a framework for natural-language explanations of LiDAR corruption through graph-based reasoning and lightweight visual semantics. The proposed method utilizes a Graph Attention Network (GAT) to detect degradation and estimate severity using node-level attention as a geometric saliency signal. Crucially, the system employs a decoupled RGB stream solely for asynchronous semantic verification, discriminating between physical obstructions and sensor-level contamination without heavy multimodal fusion. These GAT-derived signals are mapped to structured prompts and processed by a ViT–GPT2 backbone to synthesize concise, human-readable explanations. Evaluations demonstrate that the proposed method achieves promising qualitative results in contamination detection while providing transparent reliability assessments without the need for end-to-end multimodal retraining.**

I. INTRODUCTION

LiDAR contamination—ranging from dust and mud to lubricant deposits—presents a critical failure mode for autonomous perception. Unlike atmospheric degradation, physical deposits directly distort raw point clouds, creating high-confidence false negatives and "ghost" objects that clean-data detectors fail to anticipate [1]. Recent benchmarks show that even mild corruption can reduce 3D detection mAP by nearly 20% [2]. Despite advancements in real-world contaminated datasets and anomaly detection [3], current systems operate primarily as black-box classifiers; they can detect failure but cannot articulate its cause or precise spatial location. Consequently, LiDAR contamination remains detectable but fundamentally unexplainable.

Modern Vision–Language Models (VLMs) like LLaVA [4] and InternVL [5] excel at linguistic reasoning but are inherently image-centric, lacking the structural priors to interpret LiDAR sparsity and occlusion artifacts. Conversely, graph-based anomaly detectors [3] achieve high classification accuracy but cannot communicate the internal cues driving their decisions. This asymmetry creates a capability gap: LiDAR models detect but cannot articulate, while VLMs articulate but cannot "see" 3D geometric failures. This raises a pivotal question: can a LiDAR sensor be made to explain its own failure in natural language without requiring end-to-end multimodal retraining?

To bridge this gap, we introduce a framework that translates graph-driven geometric reasoning into structured natural-language explanations. The framework utilizes a Graph Attention Network (GAT) to analyze voxelized point clouds, leveraging node-level attention to generate fine-grained saliency maps. These maps are aggregated into interpretable spatial sectors (e.g., front-right, near-side) to localize degradation. Finally, a lightweight ViT-GPT2 module interprets these graph-derived cues, enabling the system to diagnose localized corruption patterns—such as foam-induced occlusion—and distinguish them from genuine empty-scene sparsity. The proposed method operates under a decoupled paradigm where RGB images are utilized for cross-sensor semantic consistency checks rather than early-stage feature fusion. This allows the system to generate high-level reasoning such as: *LiDAR sparse but camera empty—scene truly empty,"* or *Camera shows dense objects but LiDAR missing returns—likely contamination."* By distinguishing environmental emptiness from physical sensor failure, It provides a layer of semantic reliability that exceeds conventional anomaly detection.

The framework is inherently modular; the GAT backbone can be substituted with alternative graph neural networks or transformers, and the explanation head is compatible with various VLM architectures (e.g., BLIP, LLaVA, InternVL). By separating geometric reasoning from linguistic synthesis, The proposed method enables future extensions in Vision–LiDAR fusion without the computational overhead of end-to-end retraining. To our knowledge, this work represents the first attempt to enable a LiDAR sensor to explain its own physical failures through graph-driven natural-language reasoning.

Our contributions are summarized as follows:

- Making VLMs Understand Point Clouds. The first framework proposing GAT-derived saliency into structured linguistic cues, effectively enabling Vision–Language Models to analyze and describe point cloud corruption
- Spatially Grounded Corruption Reasoning. The proposed method computes node-level anomaly scores via GAT attention and aggregates them into interpretable spatial regions (e.g., front-left, near-range), allowing fine-grained explanation of where degradation occurs.
- Cross-Sensor Consistency and Real-World Bench-

*This work is funded by the European Union - Next Generation EU, Mission 4, Component 2, Investment 3.3 (Ministerial Decree 352/2022) CUP J33C22001350009. This publication is also written within the Shift2SDV project (Grant Agreement No. 101194245), which is supported by Chips Joint Undertaking and its members, including top-up funding by the national authorities of Austria, Denmark, Germany, Greece, Finland, Italy, Netherlands, Poland, Portugal, Spain, Turkey. This work is also partly support by the EdgeAI KDT-JU project (101097300). We thank FEV Italia s.r.l. for their collaboration.

[1]DEI Department, University of Bologna, Bologna, Italy grafika.jati2@unibo.it

marking. Without fusing features, the proposed method uses RGB images to assess semantic consistency—distinguishing true empty scenes from sensor failure.

II. RELATED WORKS

A. 3D-Language Alignment and Point-LLMs

Recent progress in 3D-language alignment has produced frameworks like LiDARCLIP [6] and GPT4Point [7], which map point clouds to shared text embeddings for object-level captioning and retrieval. Subsequent models, such as PointLLM-V2 [8], integrate dedicated 3D encoders to reason about object attributes. However, these models assume noise-free, object-centric data and focus on semantic recognition rather than sensor integrity. Crucially, existing 3D-LLMs lack the structural priors to interpret LiDAR-specific failures or distinguish geometric sparsity from sensor corruption.

B. LiDAR Contamination and Anomaly Detection

LiDAR sensors are susceptible to physical contaminants (e.g., dust, water, foam) that distort spatial distributions and intensity returns, leading to safety-critical "ghost" objects or false negatives [1]. While prior research has benchmarked the impact of synthetic corruptions on detection accuracy [2], [9], recent work has shifted toward real-world anomaly detection using graph-based or statistical methods [10]. Although these detectors achieve high classification accuracy, they remain "black boxes" that cannot articulate the location or cause of failure. The proposed method addresses this by converting opaque graph-level reasoning into spatially grounded, human-readable diagnostics.

Graph Neural Networks (GNNs), particularly Graph Attention Networks (GATs), excel at capturing local irregularities in sparse 3D data [3]. However, GNN explainability is typically limited to raw attention weights or saliency maps, which lack semantic meaning for end-users.

C. Vision–Language Models (VLMs) in Autonomous Driving

Modern VLMs like LLaVA [4] and InternVL [5] provide sophisticated visual reasoning but remain inherently limited to RGB data. Recent efforts, such as the RoboSense Challenge [11], have attempted to use massive models (e.g., Qwen2.5-VL) for task-specific prompting in autonomous scenarios. However, these approaches typically rely on image-level reasoning or projected representations and do not directly process raw LiDAR point clouds, thus lacking access to the underlying 3D geometric artifacts caused by sensor corruption. Furthermore, these models often utilize billion-parameter architectures that are computationally prohibitive for real-time robotics. Similarly, MAPLM [12] uses Vision-Language benchmarks for scene understanding through Question Answering (QA), but primarily focuses on semantic mapping rather than sensor integrity.

In contrast, the proposed method utilizes a lightweight ViT-GPT2 backbone and a Graph Attention Network to bridge the gap between 3D geometry and language. Unlike traditional cross-modal research that focuses on feature-level fusion for detection robustness [13], the proposed method employs the VLM for *semantic consistency verification*. By cross-referencing camera-detected scene content with LiDAR structural sparsity, our framework distinguishes between true environmental emptiness and localized sensor failure—enabling advanced cross-modal reasoning without the need for extensive multimodal retraining or massive LLM backbones.

III. PROPOSED METHOD

A. Problem Definition

Although VLMs offer sophisticated reasoning capabilities, they intrinsically fail on LiDAR for several reasons: (i) lack of LiDAR exposure during pretraining, (ii) absence of geometric inductive biases for sparse data, (iii) unmodeled corruption phenomena such as beam dropout, and (iv) loss of spatial cues during 2D projection.This motivates a hybrid approach where graph-based LiDAR reasoning provides structured signals and VLMs provide linguistic interpretability. The proposed method is the first system to integrate these components for explainable LiDAR corruption diagnosis.

We define a reasoning function f_θ that maps synchronized LiDAR and camera observations to a human-readable diagnostic $\mathcal{E}$:

$$f_\theta(\mathcal{P}, \mathbf{I}) \rightarrow \mathcal{E} \tag{1}$$

. **Input Representation**: The system processes two parallel data streams to build a unified scene representation:LiDAR Geometric State ($\mathbf{S}_{corr}$): Raw points $\mathcal{P}$ are transformed into a voxelized graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. A Graph Attention Network (GAT) extracts the latent corruption state $\mathbf{S}_{corr} = \langle \sigma, \mathcal{M}_{sal}, \mathcal{R} \rangle$, capturing failure severity (σ), geometric saliency ($\mathcal{M}_{sal}$), and the affected spatial region ($\mathcal{R}$). Visual Semantic Context ($\mathbf{F}_{rgb}$): The RGB image $\mathbf{I}$ is encoded into a high-level feature vector $\mathbf{F}_{rgb}$ to provide a "semantic witness" of the environment. **Output Explanation:** The final output $\mathcal{E}$ is a natural-language sequence generated by conditioning the decoder on the fused geometric and visual features:

$$\mathcal{E} = \text{Decoder}(\text{Prompt}(\mathbf{S}corr) \oplus \mathbf{F}rgb) \tag{2}$$

This ensures the explanation is physically grounded in the detected LiDAR anomalies and semantically verified by the camera context (e.g., distinguishing between a dirty sensor and a truly empty road).

B. Geometric Feature Extraction (LiDAR Branch)

Given a raw LiDAR point cloud $\mathcal{P} \in \mathbb{R}^{N \times 4}$, we first apply voxelization to generate a set of voxels $\mathcal{V} = \{v_i\}_{i=1}^{M}$, where each voxel stores the mean coordinate and average intensity $\bar{I}$.

To model the spatial relationships efficiently, we construct a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ using *Axis-Sorted Adjacency*. We sort nodes along each Cartesian axis $d \in \{x, y, z\}$ and connect neighboring nodes in the sorted list. This approach ensures $O(N \log N)$ complexity, making it suitable for real-time robotics. A Graph Attention Network (GAT) then computes

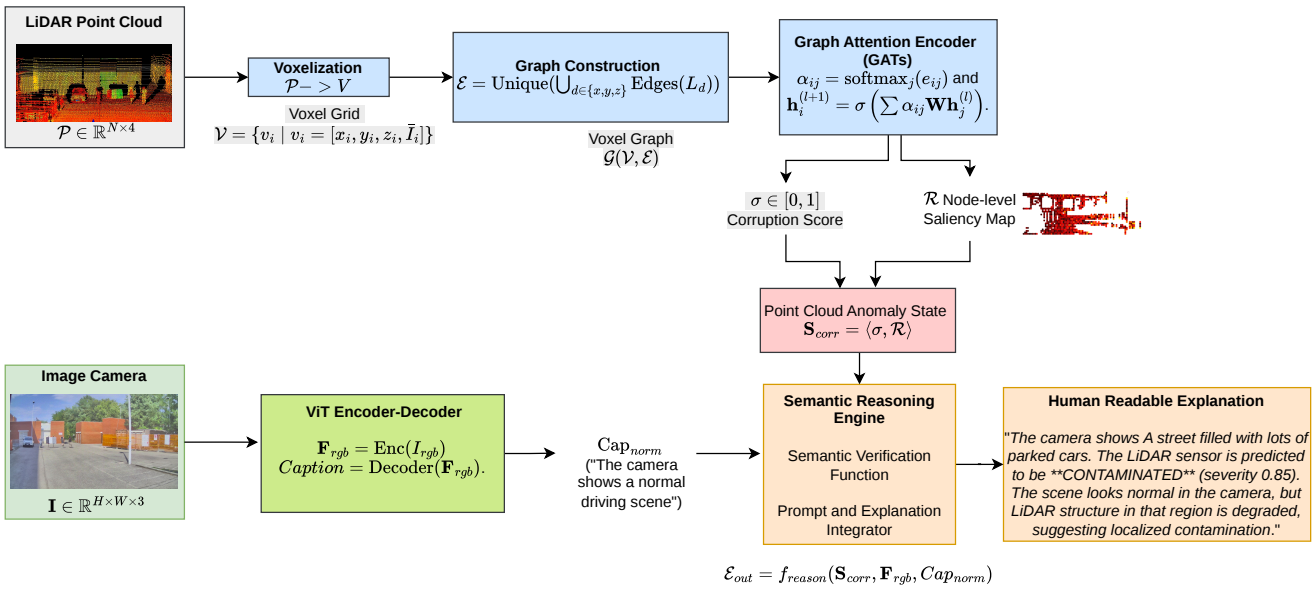

Fig. 1: Proposed Method

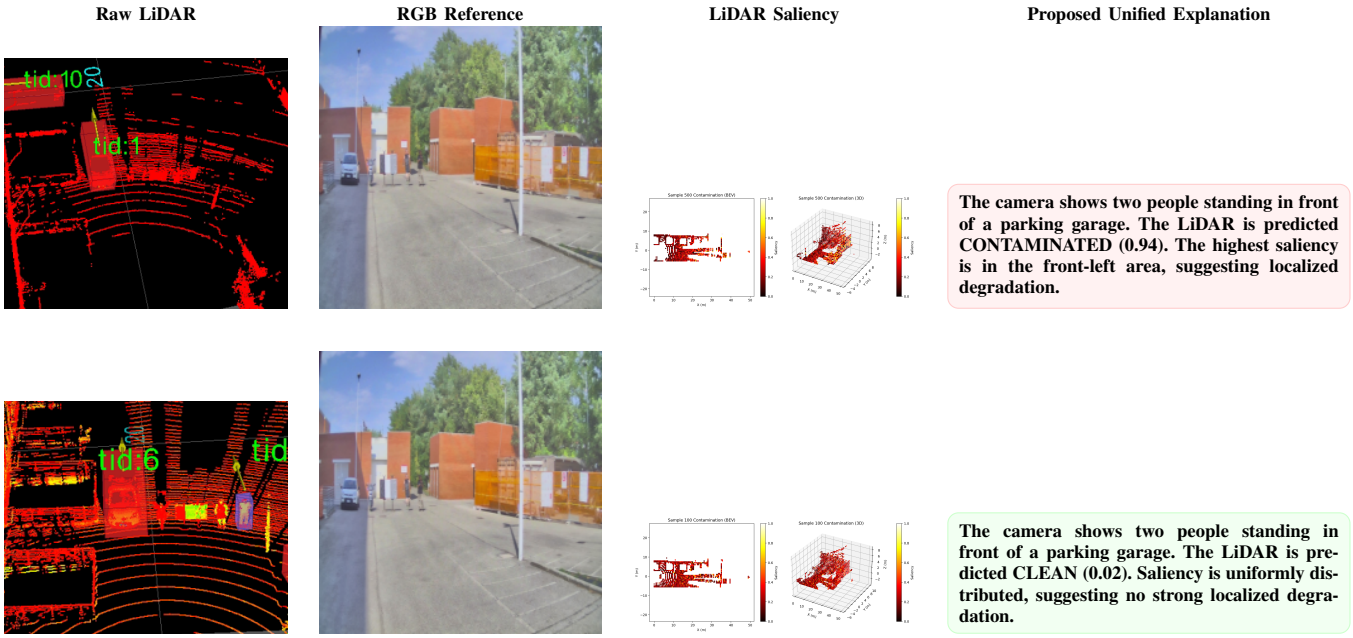

Fig. 2: Qualitative results showcasing that the proposed unified reasoning engine. The framework compares raw LiDAR geometry and visual context to identify anomalies, localized through GAT saliency, and generates grounded natural-language diagnostics.

attention weights α_{ij} to derive a node-level saliency map. The output is a corruption state vector:

$$\mathbf{S}_{corr} = \langle \sigma, \mathcal{M}_{sal}, \mathcal{R} \rangle \tag{3}$$

where $\sigma \in [0, 1]$ denotes the predicted corruption severity and $\mathcal{R}$ represents the localized region of the anomaly.

C. Semantic Contextualization (Camera Branch)

Simultaneously, the RGB camera image $\mathbf{I}$ is processed through a Vision Transformer (ViT) to extract visual embeddings $\mathbf{F}_{rgb}$. These embeddings are passed to a GPT-2 decoder to generate a normalized caption Cap_{norm}, which provides the environmental context necessary to verify the LiDAR signals.

D. Unified Reasoning Engine

The core contribution of this paper is the *Asynchronous Semantic Verification*. The engine integrates $\mathbf{S}_{corr}$ and Cap_{norm} through a rule-based prompt integrator. By comparing the structural anomalies in $\mathcal{P}$ against the visual

ground truth in $\mathbf{I}$, the system differentiates between physical obstructions and sensor-internal contamination, producing a natural-language explanation for the end-user.

IV. RESULT AND DISCUSSION

The qualitative results in figure 2 demonstrate the proposed reasoning pipeline by aligning four critical modalities: raw LiDAR geometry, RGB contextual references, GAT-based saliency maps, and synthesized natural-language diagnostics. In the contaminated scenario, the engine identifies structural anomalies—such as noise clusters or intensity drops—and localizes them to the front-left region with a high anomaly score (0.94), while the RGB reference confirms the presence of a clear scene, thereby validating that the degradation is sensor-internal. Conversely, in the clean scenario, the system maintains semantic consistency across sensors with a minimal anomaly score (0.02), proving that the linguistic output is physically grounded in the geometric attention weights of the GAT rather than being hallucinated by the vision-language backbone.

V. CONCLUSION

In conclusion, the proposed method successfully bridges the gap between low-level geometric failure detection and high-level human interpretability. By decoupling geometric reasoning from linguistic synthesis, the framework provides a scalable and modular solution for explainable sensor reliability. This approach ensures that autonomous systems can not only detect when their perception is compromised but also articulate the "where" and "why" of the failure, significantly enhancing the transparency and safety of robotic navigation in unpredictable real-world environments.

REFERENCES

[1] G. Jati, M. Molan, F. Barchi, A. Bartolini, G. Mercurio, and A. Acquaviva, "Lidaroc: Realistic lidar cover contamination dataset for enhancing autonomous vehicle perception reliability," *IEEE Sensors Letters*, vol. 8, no. 9, pp. 1–4, 2024.

[2] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3d object detection to common corruptions," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1022–1032.

[3] G. Jati, M. Molan, F. Barchi, A. Bartolini, G. Mercurio, and A. Acquaviva, "Anzil: Attention-based network for zero-risk inspection of lidar point cloud in self-driving cars," *Expert Systems with Applications*, vol. 292, p. 128412, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425020317

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[5] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

[6] G. Hess, A. Tonderski, C. Petersson, K. Åström, and L. Svensson, "Lidarclip or: How i learned to talk to point clouds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 7438–7447.

[7] Z. Qi, Y. Fang, Z. Sun, X. Wu, T. Wu, J. Wang, D. Lin, and H. Zhao, "Gpt4point: A unified framework for point-language understanding and generation," in *CVPR*, 2024.

[8] R. Xu, S. Yang, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Pointllm-v2: Empowering large language models to better understand point clouds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2025.

[9] M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool, "Fog simulation on real lidar point clouds for 3d object detection in adverse weather," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 283–15 292.

[10] D. Bogdoll *et al.*, "Anomaly detection in autonomous driving: A survey," in *the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4488–4499.

[11] A. Wu and X. Luo, "Enhancing vision-language models for autonomous driving through task-specific prompting and spatial reasoning," *arXiv preprint arXiv:2510.24152*, 2025.

[12] X. Cao, T. Zhou, Y. Ma, W. Ye, C. Cui, K. Tang, Z. Cao, K. Liang, Z. Wang, J. M. Rehg, and C. Zheng, "Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 21 819–21 830.

[13] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 682–11 692.