Integrating Multimodal Large Language Models with Point Clouds for Urban Vehicle Detection

Horia Ameen[1*)], Antón Núñez-Seoane[2)], David Conde[3)], Mario Soilán[4)]

[1*)] CINTECX, Universidade de Vigo, GeoTECH Group, As Lagoas Marcosende, 36310, Vigo, Spain; horia.ameen@uvigo.gal; ORCID: 0009-0000-5928-3373
[2)] CINTECX, Universidade de Vigo, GeoTECH Group, As Lagoas Marcosende, 36310, Vigo, Spain; anton.nunez.seoane@uvigo.gal; ORCID: 0000-0002-9435-5923
[3)] CINTECX, Universidade de Vigo, GeoTECH Group, As Lagoas Marcosende, 36310, Vigo, Spain; david.conde.morales@uvigo.gal; ORCID: 0000-0002-9397-8731
[4)] CINTECX, Universidade de Vigo, GeoTECH Group, As Lagoas Marcosende, 36310, Vigo, Spain; msoilan@uvigo.gal; ORCID: 0000-0001-6545-2225

https://doi.org/10.29227/IM-2025-02-02-113

Submission date: *2025-07-24* | Review date: *2025-08-24*

Abstract

Road infrastructure assessment plays a vital role in urban planning, traffic management, and the development of autonomous systems. Recent advances in Multimodal Large Language Models (MLLMs) have shown promising capabilities for zero-shot object detection and classification in complex visual scenes. However, these models are primarily limited to 2D image analysis, lacking the spatial awareness required for precise geometric understanding in real-world environments. Meanwhile, conventional 3D object detection methods using point clouds face challenges in handling diverse, dynamic, and unstructured road environments. This work introduces a hybrid approach that combines the zero-shot visual reasoning capabilities of MLLMs with the geometric precision of 3D LiDAR point clouds collected from Mobile Mapping Systems (MMS). Specifically, the proposed pipeline uses an MLLM (Google Gemini 2.0 flash) to detect vehicles in road scene images based on natural language prompts, generating 2D bounding boxes without the need for task-specific training. These 2D detections are then projected into 3D space using calibration parameters from the KITTI-360 dataset, enabling localization and clustering of corresponding LiDAR points. Finally, 3D bounding boxes are generated, and temporal merging is performed to remove redundant detections across consecutive frames. Experiments on two sequences of the KITTI-360 dataset demonstrate that this fusion approach enables flexible and scalable vehicle detection in complex urban environments. The findings highlight the potential of integrating language-driven perception with spatially grounded 3D reasoning, offering a promising direction for automated road asset monitoring and maintenance strategies.

Keywords: *Multimodal Large Language Models, Zero-Shot Object Detection, 3D Point Clouds Processing, Road Infrastructure Assessment*

1. Introduction

Ensuring road safety is a critical priority in the era of autonomous vehicles [1] and intelligent transportation systems [2], where precise perception and understanding of road environments form the backbone of effective decision-making. According to the World Health Organization, road traffic accidents are responsible for approximately 1.19 million deaths annually, making it a major global public health concern [3]. Governments and research communities alike have been working towards reducing these fatalities through improved traffic management, urban planning, and safer road infrastructure.

A key component of road safety and mobility lies in the accurate detection, classification, and localization of dynamic and static objects particularly vehicles, pedestrians, and infrastructure elements within urban scenes as it enables timely decision-making for autonomous systems and supports collision avoidance, lane keeping, and traffic regulation enforcement. Traditionally, this task has been approached using two distinct sensing paradigms: (1) vision-based 2D object detection using RGB images [4], and (2) 3D geometric modelling through LiDAR point clouds [5]. However, each of these modalities, when used in isolation, faces significant limitations. Vision-based detection methods, often powered by Convolutional Neural Networks (CNNs), offer strong semantic understanding but lack spatial awareness, making it difficult to reason about object size, depth, and true 3D positioning. Conversely, LiDAR-based methods provide precise geometric information but typically require complex heuristic rules, expensive manual annotations, or supervised learning to derive object-level semantics. This disconnect between semantic reasoning and spatial accuracy has motivated research into multi-sensor fusion approaches, where 2D imagery and 3D LiDAR data are combined to enhance perception. Datasets such as KITTI-360 [6] have played a vital role in advancing this field by offering synchronized multi-modal data and comprehensive ground truth annotations for benchmarking detection and tracking algorithms in real-world urban environments. Despite these advancements, conventional fusion approaches such as early fusion [7] (e.g., concatenating image and LiDAR features) or late fusion [8] (e.g., decision-level combination of 2D/3D detectors) still largely rely on supervised learning, handcrafted features, or domain-specific tuning. This limits their flexibility and scalability, particularly in zero-shot scenarios or when encountering rare or underrepresented object classes.

Recent advances in Multimodal Large Language Models (MLLMs) [9], a class of deep learning models designed to jointly process and reason over multiple data types such as text and images, have revolutionized how machines interpret and generate content

across different modalities including vision and language. One of the earliest breakthroughs in vision-language understanding came with CLIP (Contrastive Language-Image Pretraining), which demonstrated how large-scale training on image-text pairs could align visual and textual representations in a shared embedding space. Building upon this foundation, newer multimodal models have evolved to support more advanced reasoning capabilities over both images and text. Recent families of such models developed by organizations like OpenAI (e.g., GPT-4 with vision capabilities), Google (Gemini models), and Anthropic (Claude), integrate vision and language more tightly, enabling tasks such as zero-shot object detection, caption generation, and visual question answering through natural language prompts. These models are trained in vast amounts of image-text pairs, enabling them to associate semantic meanings with visual regions. This allows them to generalize to new tasks in a zero-shot setting, where the model performs without having seen any task-specific examples or a few-shot setting, where only a very small number of examples are provided, thus eliminating the need for extensive task-specific training data [10]. One of the standout abilities of MLLMs is their capacity to generate structured outputs, such as object labels, bounding boxes, or even captions, purely from natural language instructions. For example, by simply prompting the model with a sentence like "Detect all cars visible in the image and return their bounding boxes", the model can return a set of bounding boxes tightly aligned with the queried object class. This bypasses the need for manually crafted models, annotated datasets, or complex training pipelines traditionally required in computer vision.

While MLLMs have shown promising capabilities in 2D vision-language tasks, their potential in 3D spatial reasoning remains largely unexplored. Most existing models lack native 3D spatial grounding, making them unsuitable for applications that require geometric precision and object localization, such as autonomous driving or infrastructure mapping. Recent research efforts, such as 3D-LLMs, have started addressing this limitation by extending LLMs to incorporate 3D point cloud inputs and spatial awareness, highlighting both the gap and the direction for future progress in 3D multimodal reasoning [11]. Specifically, the objective is to investigate whether zero-shot visual detections generated by MLLMs from 2D images can be effectively projected and grounded into 3D space, allowing for accurate object localization, shape extraction, and evaluation in complex road environments. By leveraging the strengths of both modalities, this approach aims to reduce the reliance on large, annotated 3D datasets, improve flexibility in object definition, and pave the way for more scalable solutions for road infrastructure monitoring and autonomous perception.

The format of this paper is as follows. Section 2 presents related work and defines the knowledge gap and objectives of this work. Section 3 describes the methodology and case study, while Section 4 offers concluding thoughts and suggestions for future study directions.

2. Related Work

Object detection in 2D imagery is a fundamental task in computer vision, particularly critical for road scene understanding in autonomous driving systems. Over the years, deep learning models, especially convolutional neural networks (CNNs) have significantly improved detection performance, especially in terms of robustness, speed, and scalability. Among these, YOLO (You Only Look Once) models have emerged as state-of-the-art solutions, known for their real-time processing capabilities and strong accuracy across multiple object categories [12]. More recent iterations such as YOLOv13 have advanced the detection pipeline further by introducing hypergraph-based correlation modelling, which improves feature fusion and enhances detection robustness in complex scenes like crowded or occluded road environments [13]. One implementation of YOLOv3 using the Darknet-53 CNN backbone focuses on detecting four common road object classes cars, trucks, pedestrians, and cyclists [14]. Their model was trained on the KITTI dataset, which provides rich annotated imagery captured from forward-facing vehicle cameras. Their findings emphasized that cars were detected with high precision, owing in part to their larger size and prevalence in the dataset. However, smaller objects like pedestrians and cyclists suffered from higher false negatives, a limitation tied to the grid-cell-based detection strategy of YOLO, especially when larger cell sizes are used (e.g., 32×32 pixels). The results suggest that while CNN-based models perform well on larger objects, detecting smaller or occluded objects remains a challenge, especially in complex road scenes. This motivates continued research into improving detection granularity and model generalization, especially under varied environmental and traffic conditions.

The use of 3D LiDAR data has enabled fine-grained understanding of urban scenes through geometric reconstruction and spatial reasoning. Datasets like KITTI and its extension, KITTI-360, have become standard benchmarks for 3D scene analysis. KITTI-360 specifically introduced 3D point clouds with semantic labels and panoramic imagery, providing a rich platform for infrastructure mapping and spatial analytics [6]. Moreover, the work by Behley et al. (2021) KITTI Panoptic Benchmark extends these contributions by offering a benchmark for panoptic segmentation, encouraging unified evaluation of both semantic segmentation and instance-level object understanding in 3D [15].

The integration of RGB images and 3D point cloud data has become a cornerstone technique in modern urban scene understanding [16]. This multimodal fusion enables systems to capitalize on the complementary strengths of each modality with 2D images providing high-resolution semantic and texture information, and 3D LiDAR offering precise spatial structure and depth cues. Together, they enable more robust and context-aware perception, critical for downstream tasks such as urban mapping, object detection, and multi-object tracking.

A significant milestone in the direction of complex 3D scene understanding is the development of the H3D dataset, which targets full-surround 3D multi-object detection and tracking in highly interactive and crowded urban traffic scenes. Unlike earlier benchmarks like KITTI that offer limited occlusion and interaction diversity, H3D provides over 1 million labelled instances across 27,000 frames and introduces complex environmental conditions, such as dense occlusions and diverse object types. While H3D itself is purely LiDAR-based, its introduction has motivated the development of perception algorithms that can go beyond traditional 3D reasoning, encouraging richer spatial understanding and even multimodal fusion. The dataset serves as a standardized benchmark

and highlights key challenges, like long-term tracking failures under occlusion and the need for cross-frame continuity—challenges that recent multimodal models are now beginning to address [17].

Recent developments in MLLMs have introduced the ability to perform semantic visual reasoning in a zero-shot or few-shot fashion. While early applications of MLLMs have focused primarily on 2D vision tasks, such as image captioning, segmentation, and object detection, recent research demonstrates their growing utility in specialized and challenging environments. For instance, in the Chain of Visual Perception framework [18], MLLMs are successfully applied to zero-shot camouflaged object detection, showcasing the ability to identify subtle or hidden patterns in complex backgrounds using textual cues alone. Similarly, the use of MLLMs in thermal image-based scene understanding, showing how they can extract object-level semantics from low-resolution or ambiguous inputs, a task traditionally difficult for standard detectors [19].

While MLLMs have demonstrated strong capabilities in visual understanding tasks, their application to 3D spatial reasoning is still in its early stages. Several recent efforts have attempted to bridge this gap. For example, LLMI3D leverages spatial-enhanced local feature mining and geometry-aware reasoning for 3D perception from a single image, demonstrating improved performance in 3D object detection tasks [20]. Similarly, 3D-Aware Representation Supervision (3DRS) enhances MLLMs' spatial understanding by aligning them with supervision from pretrained 3D models [21]. Uni3D-LLM takes this further by unifying perception, generation, and editing in point cloud scenes using LLMs, showcasing the versatility of natural language-driven 3D reasoning [22]. These works collectively highlight the growing interest in combining language and 3D spatial understanding, although fully generalizable and accurate 3D reasoning within MLLMs remains an open challenge.

This work addresses this gap by introducing a hybrid pipeline that uses Google Gemini 2.0 flash [23] to detect semantic objects in RGB images via natural language prompts and then projects and aligns those detections into 3D space using LiDAR-based geometry. This strategy allows us to ground high-level visual understanding into the spatial reality of road environments. The main contributions and goals of this work are:

1) Use Gemini 2.0 flash to detect vehicles directly from RGB images using intuitive natural language prompts, without any model fine-tuning.
2) Develop a hybrid pipeline that projects language-based 2D detections into 3D space using depth and LiDAR point cloud alignment. While the KITTI-360 projection and calibration tools assist in aligning modalities, the fusion strategy and filtering logic were custom-designed for this work.
3) Validate the approach on the KITTI-360 dataset, showing that meaningful 3D object detection is possible without any 3D annotated training data, while also analyzing performance and limitations.

3. Methodology

3.1 Dataset Description

The case study in this work focuses on the publicly available KITTI-360 dataset, which provides a comprehensive benchmark for urban scene understanding in both 2D and 3D domains. KITTI-360 comprises over 73,000 high-resolution stereo images and corresponding 3D LiDAR point clouds, collected along 100 km of road scenes in Karlsruhe, Germany. The KITTI-360 dataset provides a rich multimodal benchmark for urban scene understanding, offering a combination of high-resolution 2D imagery and densely sampled 3D point clouds. The image data includes front-facing stereo images as well as full-surround fisheye views, all of which are rectified to maintain stereo consistency. The 3D point clouds are captured using a roof-mounted Velodyne HDL-64E LiDAR scanner in conjunction with a SICK laser, ensuring 360-degree coverage of the scene with high geometric fidelity. Alongside the raw sensor data, the dataset includes extensive ground-truth annotations. These comprise dense semantic and instance-level labels projected consistently across the RGB images and 3D point clouds, contributing over 1 billion labelled 3D points and approximately 150,000 annotated images. Each frame is also associated with 3D bounding box annotations for dynamic objects such as cars and pedestrians, enabling rigorous benchmarking of detection and tracking algorithms. Synchronization across modalities is achieved with high temporal precision, using Velodyne timestamps as the reference point and aligning GPS/IMU data within a 5-millisecond tolerance window. This ensures consistent alignment of visual and spatial information throughout the dataset. Calibration files provide both intrinsic and extrinsic parameters for the cameras, along with rigid-body transformations between all sensors coordinate frames, including LiDAR and GPS/IMU [6]. These calibration parameters are crucial for accurate projection between 2D and 3D modalities. Overall, the dataset's precise synchronization, dense labelling, and multi-sensor calibration make it exceptionally well-suited for research involving multimodal perception pipelines, such as this approach which maps 2D visual outputs into 3D spatial domains and evaluates them against annotated LiDAR-based ground truth. For the experiments in this study, we utilized two specific sequences from the KITTI-360 dataset: sequence 0000 (frames 2 to 385) and (frames 376 to 610). These sequences were selected due to their high object density and diversity in urban traffic scenarios, providing a robust basis for evaluating 3D detection using vision-language models.

3.2 Case-Study

The case study evaluates the proposed hybrid object detection pipeline through a sequence of experiments designed to integrate semantic understanding from images with geometric localization in 3D point clouds. The process begins with synchronized RGB images and accumulated 3D LiDAR point clouds, a composite of multiple sequential LiDAR frames aligned over time to enhance object completeness and spatial density 'Fig. *1*'. A language prompt is passed to Google Gemini 2.0 flash, which performs zero-shot object detection on the images, producing 2D bounding boxes around vehicles of interest. These 2D detections are then projected into 3D space using the provided calibration parameters, aligning image-based detections with the LiDAR point cloud. The

corresponding 3D regions are processed further by applying semantic filtering, geometric constraints, and density-based clustering to accurately localize object candidates within the point cloud. To handle redundancies caused by overlapping detections across consecutive frames, a temporal merging strategy based on Euclidean clustering and Non-Maximum Suppression (NMS) is applied to eliminate duplicates. The pipeline outputs final 3D bounding boxes, which are subsequently evaluated against ground-truth annotations to assess detection accuracy and robustness.

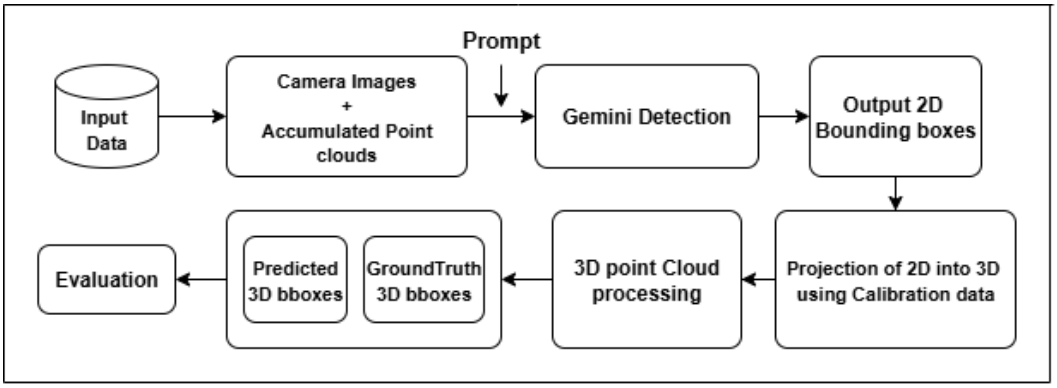


Fig. 1. Overview of the proposed hybrid 2D-3D object detection pipeline.

3.2.1 MLLM-Based Zero-Shot Object Detection. The first phase of the proposed pipeline applies Gemini 2.0 flash for zero-shot object detection on RGB images extracted from two sequences of the KITTI-360 dataset, each comprising over 300 frames. Rather than relying on supervised training or pre-labeled data, the method leverages natural language prompting to guide the model's visual reasoning. The model was prompted using explicit instructions aimed at accurate identification of vehicles, formulated through iterative refinement to achieve consistent performance. Specifically, the prompts provided were:

"Identify all clearly visible vehicles in the provided image and return their bounding boxes as JSON formatted coordinates. Ignore distant or heavily occluded vehicles."

This prompt-driven approach allows for flexible semantic filtering without the need for retraining. Google Gemini processes each image and returns structured outputs in the form of JSON format, where each detected vehicle is represented by a set of 2D bounding box coordinates, enabling easy parsing and downstream processing. An example of the response is shown below:

[{"label": "car", "bounding_box": [488, 477, 581, 515]},

{"label": "car", "bounding_box": [485, 584, 781, 722]}]

In this format, each detected object is associated with a category label and its corresponding pixel-level bounding box coordinates in the form [xmin, ymin, xmax, ymax]. These detections reflect the model's interpretation of the natural language query and the visual content of the image. To mitigate potential inaccuracies in object boundary estimation, the bounding boxes are systematically padded by a fixed pixel margin (e.g. 15 pixels). This padding ensures that the projected regions of interest in subsequent 3D mapping steps fully encompass the target vehicles, thereby reducing the risk of partial or incomplete detections. For example, 'Fig. 2' illustrates the output of Gemini's zero-shot object detection on two example frames from the KITTI-360 dataset. The red bounding boxes highlight the vehicles detected directly from the 2D images based on natural language prompts, demonstrating the model's ability to localize multiple objects of interest under urban scene conditions.

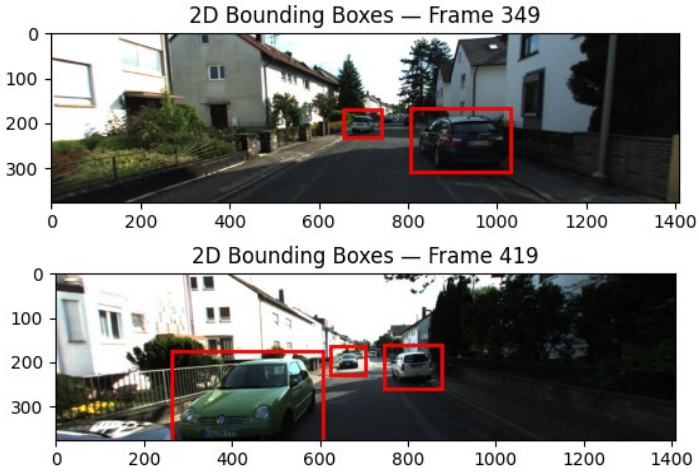


Fig. 2. 2D vehicle detections using Google Gemini on KITTI-360 frames

3.2.2 3D Projection & Point Cloud Localization. The subsequent phase focuses on transforming the 2D detections produced by Gemini into their corresponding 3D spatial representations within the LiDAR point cloud. This transformation is performed using the official KITTI-360 projection tools [24], which apply accurate calibration-based projection using the intrinsic and extrinsic parameters provided with the dataset. The intrinsic parameters define the camera's internal geometry, while the extrinsic parameters establish the rigid-body transformation between the camera and LiDAR coordinate frames. By leveraging this calibration, each 2D bounding box from the image is precisely mapped into a 3D frustum within the point cloud. To ensure the reliability of this projection, ground filtering is first applied to remove planar road points that often merge with the lower parts of vehicle detections. Without this step, such points can lead to inaccurate 3D bounding boxes by artificially inflating the vertical extent of clustered objects. The removal of ground points helps in isolating true object surfaces from the road, resulting in cleaner segmentation. Following this, semantic filtering is performed to retain only those LiDAR points associated with relevant object categories, including vehicles (e.g., cars and trucks) as well as other roadside objects present in the scene. This step eliminates non-object points, such as buildings, trees, and pedestrians, which do not match the detection target. To refine these regions, a custom geometric filtering method is applied to exclude points that exhibit abrupt spatial discontinuities or points that fall outside a ±5 meter depth window from the object's central region. This combined filtering approach refines the projected regions, ensuring that only spatially consistent clusters likely corresponding to actual vehicles are preserved. Through this integrated projection and refinement process, the 2D semantic outputs from Gemini are effectively grounded within the 3D LiDAR environment, enabling accurate spatial localization of detected objects.

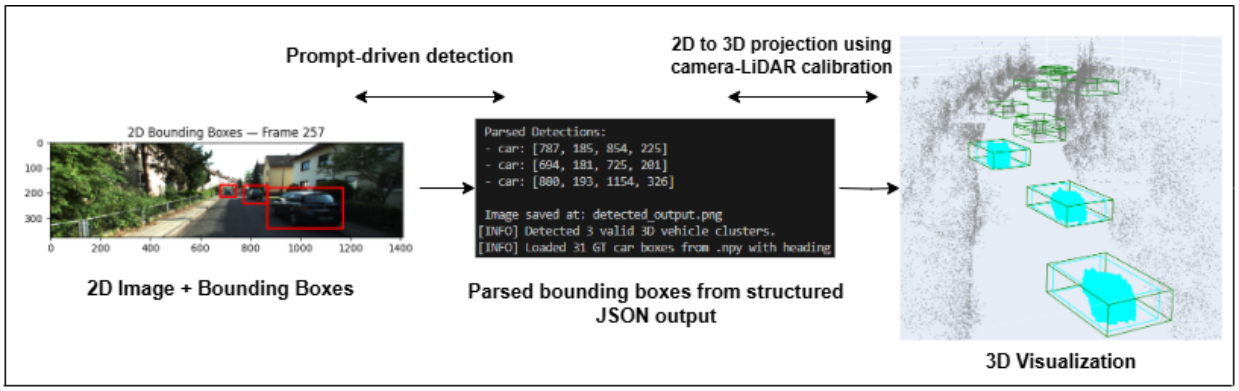


Fig. 3. End-to-end workflow from 2D detections to 3D spatial localization in point clouds

The process illustrated in 'Fig. 3' demonstrates the key steps involved in projecting the 2D detections generated by the Gemini model into the 3D spatial domain for localization within the LiDAR point cloud. Starting from the left, the system receives RGB images along with natural language prompts to detect visible cars, resulting in 2D bounding boxes overlaid on the images. These detections are then exported as structured JSON outputs, containing the class labels and corresponding pixel coordinates of each bounding box. Using the camera intrinsic and extrinsic calibration parameters provided by the KITTI-360 dataset, the 2D bounding boxes are geometrically projected into the 3D point cloud space. In this spatial domain, additional semantic and geometric filtering is applied to refine the detected regions, isolating points that correspond to vehicles while excluding ground surfaces and irrelevant background. The result, shown on the right, visualizes the localized 3D bounding boxes within the LiDAR environment, effectively bridging 2D semantic understanding with 3D spatial grounding. This hybrid approach enables accurate object detection and localization in dynamic urban scenes without requiring explicit 3D annotations during detection.

3.2.3 Filtration, Clustering, & Temporal Merging. Once the relevant 3D points are localized within the LiDAR space, the next step focuses on generating precise 3D object representations through clustering and bounding box construction. To achieve this, the method applies the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [25] algorithm to group spatially coherent point clusters within each projected region. DBSCAN is particularly well-suited for this task as it does not require prior knowledge of the number of objects and is robust to noise and varying object densities common characteristics of outdoor LiDAR scans. Each detected cluster undergoes rigorous filtering to eliminate spurious detections. Clusters are retained only if they meet predefined geometric constraints based on real-world vehicle dimensions, In particular, the axis-aligned length, width, and height of each cluster must fall within the range of 0.5 to 7.0 meters. Additionally, only clusters with at least 30 points are considered to reduce noise. To further ensure consistency, a depth continuity filter of ±5.0 meters is applied relative to the reference depth at the center of each bounding box. The DBSCAN algorithm is configured with neighborhood radius ε = 0.6 and minimum 3 points per cluster to detect spatially compact clusters. This filtering step ensures that only clusters resembling vehicle-like shapes are preserved for further processing. For every valid cluster, the algorithm computes tight axis-aligned 3D bounding boxes by determining the minimum and maximum spatial extents along the X, Y, and Z axes.

Mathematically, a candidate cluster Ck is retained if:

$$\Delta x = \max(x_i) - \min(x_i), \Delta y = \max(y_i) - \min(y_i), \Delta z = \max(z_i) - \min(z_i) \quad (1)$$

$$|C_k| \geq 30 \; and \; 0.5 < \Delta i < 7.0 \; \forall i \in \{x, y, z\}$$

where in equation (1) Δi represents the axis-aligned extent in each spatial dimension. The input to DBSCAN is filtered using a depth continuity constraint:

$$|\mathrm{d_i} - d_c| \leq 5.0 \qquad (2)$$

Where:

di = depth of each LiDAR point projected inside the bounding box

dc = reference depth of the point closest to the 2D bounding box center

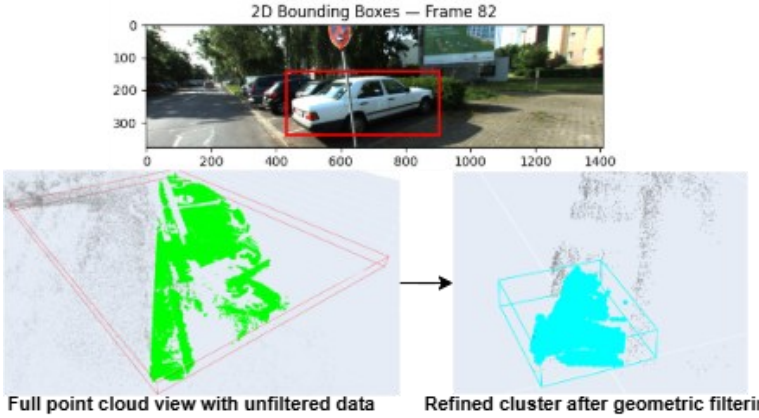


Fig. 4. Visual comparison of the 3D filtering process guided by 2D prompt-based detection

'Fig. 4' illustrates the geometric filtering and clustering process used to isolate relevant 3D object regions from raw LiDAR data. The top image shows the 2D detection generated by the Gemini 2.0 flash model. The bottom-left view presents the full unfiltered point cloud, while the bottom-right shows the refined 3D cluster after applying depth continuity constraints and DBSCAN clustering. This step is crucial to remove background and ground noise and retain only meaningful object structures aligned with 2D prompts.

However, since the KITTI-360 sequences used in this study consist of hundreds of consecutive frames captured by a moving vehicle, the same road objects such as parked or moving cars are frequently detected multiple times across adjacent frames. This leads to redundant 3D detections that can inflate false positives and degrade performance metrics if left unaddressed. To resolve this, a two-stage temporal merging process is applied to all predicted 3D bounding boxes collected across the entire sequence. In the first stage, Euclidean distance-based clustering is performed on the centers of the 3D bounding boxes using a distance threshold of 1.5 meters, which empirically balances merging temporally redundant detections of the same object while avoiding the fusion of nearby distinct objects. Within each spatial group, the second stage applies 3D Non-Maximum Suppression (NMS) [26], which uses Intersection-over-Union (IoU) computed on the 3D volumes to suppress redundant detections. Only the highest confidence bounding box in each group is retained as the final prediction. This combined strategy effectively reduces duplicate detections while maintaining spatial consistency and accuracy. The final output consists of a clean, non-redundant set of 3D bounding boxes that are temporally stable and ready for comparison against ground-truth annotations for quantitative evaluation.

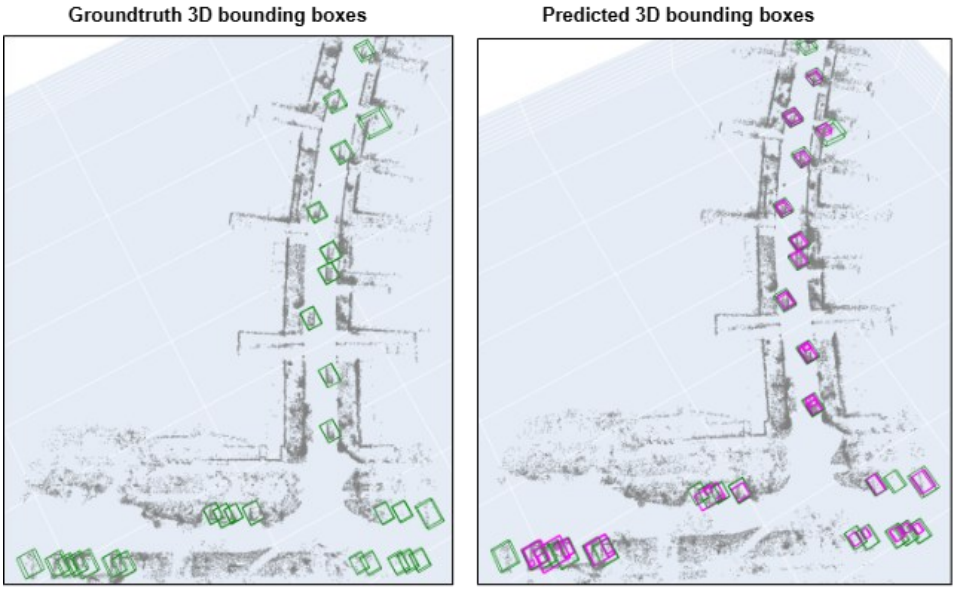
Additionally, by removing LiDAR points corresponding to the road surface, the method avoids the common issue where ground points merge with object points, leading to distorted bounding box geometries. This preprocessing step ensures that only true object surfaces contribute to the clustering process, resulting in tighter and more reliable 3D bounding boxes.

3.2.4 Evaluation Against Ground Truth. To quantitatively assess the accuracy of the proposed detection pipeline, the final set of predicted 3D bounding boxes is compared against the official ground truth annotations provided by the KITTI-360 dataset. The ground truth data consists of carefully annotated 3D bounding boxes for dynamic objects such as vehicles, each specified by the box's center coordinates, physical dimensions, and heading angle, along with a semantic class label. To ensure alignment between predictions and annotations, each ground truth box is reconstructed using the provided center, size, and orientation parameters. This reconstruction employs a standard 3D bounding box formulation with rotation applied around the vertical (Z) axis, following the KITTI-360 coordinate system. The rotated bounding box is then transformed into its corner points to obtain a precise spatial representation. Only objects labeled as cars (semantic ID 26) are retained for evaluation to ensure consistency with the detection classes targeted in the pipeline.

Similarly, each predicted bounding box generated from the clustered point clouds is saved in the same format, including 3D center, size dimensions, a fixed heading angle (as orientation is not explicitly estimated), semantic ID, and confidence score. These predictions are exported as *.npy* files in accordance with KITTI-360 evaluation standards, facilitating seamless downstream evaluation. The comparison between the predicted and ground truth bounding boxes is performed by computing standard metrics such as Average Precision (AP) and Average Recall (AR) at multiple Intersection over Union (IoU) thresholds, typically 0.25 and 0.50. This allows for robust measurement of both detection accuracy and completeness. The use of the original KITTI-360 evaluation protocol ensures that the performance of the proposed method can be benchmarked fairly and consistently against existing approaches

in literature. This ground truth comparison not only validates the effectiveness of the proposed hybrid detection system but also highlights areas for improvement, such as the need for orientation estimation or further refinement of the 3D clustering process.

'Fig. 5' provides a visual comparison between the ground truth 3D bounding boxes and the predicted 3D detections generated by the proposed hybrid pipeline. 'Fig. 5a', the official KITTI-360 annotations for vehicle instances are displayed in green, showing the precise spatial extent of each labeled object. In 'Fig.5b', the output of the detection pipeline is illustrated, with magenta boxes representing the vehicles identified using the Gemini-driven 2D detection followed by 3D projection and clustering. The side-by-side visualization highlights the effectiveness of the method in localizing most vehicles accurately while also revealing cases where false positives or minor misalignments may occur, serving as a useful tool for qualitative assessment.



(a) (b)

Fig. 5. Visualization of ground truth (a) and predicted (b) 3D bounding boxes overlaid on the accumulated point cloud

3.3 Result and Discussion

The performance of the proposed multimodal object detection pipeline is quantitatively evaluated using two widely adopted metrics in 3D detection benchmarks: Average Precision (AP) and Recall (R), computed at two different Intersection over Union (IoU) thresholds of 0.25 and 0.50. These thresholds are selected to assess both coarse and stricter alignment between the predicted 3D bounding boxes and the ground truth annotations provided by the KITTI-360 dataset. To compute precision and recall, we followed a standard IoU-based matching protocol: predicted 3D boxes were matched to ground truth boxes per frame, and labelled as True Positives (TP) if the IoU exceeded the threshold and the ground truth was unused; otherwise, they were considered False Positives (FP), while unmatched ground truths were counted as False Negatives (FN). For example, the equation (3) defines the Intersection over Union (IoU) metric, which measures the overlap between the predicted 3D bounding box P and the ground truth box G. It is computed as the ratio of the volume of their intersection to the volume of their union. A higher IoU indicates better alignment between prediction and ground truth.

$$IoU = \frac{Volume\ (P\ \cap\ G)}{Volume\ (P\ \cup\ G)} \tag{3}$$

Tab. 1 summarizes the detection results across two test sequences as well as their combined average. At IoU@0.25, the pipeline achieves a Car-AP of 65.64% and 53.21% for Sequence 1 and Sequence 2, respectively, with corresponding recall values of 70.97% and 69.23%. When increasing the matching stringency to IoU@0.50, Car-AP naturally decreases to 41.41% and 49.96%, while recall values drop to 51.61% and 61.54%, highlighting the inherent trade-off between detection precision and spatial overlap tightness. The combined average Car-AP across both sequences is 59.10% (IoU@0.25) and 37.91% (IoU@0.50), with recall reaching 70.45% and 54.55%, respectively.

Tab. 1. Detection Performance Summary across Sequences (IoU@0.25 and IoU@0.50)

Sequence	IoU Threshold	Car Average Precision (Car-AP)	Car Recall (Car-R)
Sequence 1	IoU@0.25	65.64%	70.97%
Sequence 1	IoU@0.50	41.41%	51.61%
Sequence 2	IoU@0.25	54.67%	69.23%
Sequence 2	IoU@0.50	49.96%	61.54%
Combined (Avg)	IoU@0.25	60.31%	70.45%
Combined (Avg)	IoU@0.50	40.36%	56.82%

A key factor contributing to the improvement in detection accuracy is the introduction of an enhanced temporal merging strategy. Initial attempts using naive merging resulted in redundant multiple detections for the same object across consecutive frames, inflating false positives and lowering AP. By incorporating distance-based clustering followed by 3D Non-Maximum Suppression (NMS), the pipeline effectively reduces duplicate predictions and retains only the most confident detections, resulting in measurable gains in precision. Specifically, this refinement improved precision by approximately 4–5% across the evaluated sequences. Visual inspection of both 2D and 3D results further validates the effectiveness of the proposed method. In the 2D domain, Gemini accurately identifies most vehicles, even under varying lighting and occlusion conditions. In the 3D visualization, most detected vehicles align well with ground truth bounding boxes, showcasing the success of the projection and clustering steps. Despite these strengths, some residual failure cases were observed. During the merging process, instances were observed where multiple overlapping bounding boxes were generated for the same object, resulting in a "box-within-box" or nested boxes phenomenon. This typically occurs when detections from different viewpoints or frame sequences are not properly consolidated. Additionally, frame-rate redundancy where the same object is detected multiple times across consecutive frames remains a limitation, particularly in stationary or slow-moving scenes. While the merging strategy significantly improves precision overall, these residual cases often caused by imperfect temporal association can still introduce local inconsistencies that affect average precision (AP) scores. The fusion of visual semantics from Gemini with geometric filtering in 3D space proves to be a significant strength of this approach. It allows flexible object detection without retraining, accommodates varied object appearances, and reduces reliance on large-scale annotated 3D datasets. However, limitations persist. The system exhibits sensitivity to frame-to-frame redundancy, particularly in densely packed urban scenes, where multiple detections for the same object can still emerge despite merging. NMS thresholds of 0.25, 0.30, and 0.50 are evaluated. A threshold of 0.30 offers the best balance between eliminating redundant detections and preserving true positives. Lower values (e.g., 0.25) are overly strict and discard valid boxes, while higher values (e.g., 0.50) allow excessive overlap between detections. Additionally, the strictness of the clustering and NMS thresholds influences the balance between precision and recall tighter thresholds reduce false positives but increase false negatives, and vice versa. To further improve the system, several enhancements are identified. These include dynamic adjustment of clustering parameters based on scene density, improved temporal association of object tracks to reduce redundancy, and fine-tuning of prompts or model parameters for better small object detection in the 2D stage. Incorporating temporal consistency checks or lightweight tracking mechanisms could also help improve the stability and accuracy of detections across sequences.

4. Conclusion

This work presents a hybrid framework that integrates zero-shot visual reasoning from Multimodal Large Language Models (MLLMs), specifically Google Gemini 2.0 flash, with 3D LiDAR point cloud localization to perform automated vehicle detection in urban road environments. By leveraging natural language prompts, the proposed framework eliminates the need for supervised training pipelines traditionally required for object detection, offering a flexible and scalable solution for semantic scene understanding. Through extensive experimentation on the KITTI-360 dataset, the proposed method demonstrates the potential of combining 2D vision-language outputs with 3D geometric reasoning for detecting vehicles in complex urban settings. The approach achieves a Car Average Precision of up to 65.64% at IoU@0.25 and maintains moderate recall, highlighting its ability to generalize across diverse scenarios without additional model fine-tuning. The introduction of temporal merging techniques and robust clustering significantly reduces redundant detections and false positives, which are common challenges in frame-by-frame processing.

In addition to enhancing detection performance, this pipeline contributes to the broader objective of road asset inventory automation. Traditional road asset management systems rely heavily on manual annotation and pre-trained object detectors that are difficult to adapt to new categories or dynamic environments. By contrast, the presented method allows for dynamic and prompt-based object detection, enabling streamlined inventorying of vehicles, road objects, or traffic participants without the need for retraining. This capability holds promise for scalable, cost-effective deployment in applications such as smart city infrastructure monitoring, autonomous driving, and digital twin creation for road networks. However, certain limitations remain. The system exhibits sensitivity to noisy or sparse LiDAR data, which can affect clustering reliability, and faces challenges in detecting small or partially occluded objects. Furthermore, while the use of MLLMs introduces flexibility, it also inherits limitations such as prompt sensitivity and potential hallucination in object classification. Addressing these challenges through temporal smoothing, adaptive prompt refinement, and integration of lightweight tracking modules represents a promising direction for future work.

Overall, this study demonstrates that vision-language models, when effectively combined with 3D geometric reasoning, can offer a viable path toward automating road scene understanding and asset detection, marking a step forward in bridging the gap between high-level semantic perception and spatial awareness in intelligent transportation systems.

Acknowledgments

This work has been supported by the Spanish Ministry of Science, Innovation and Universities through Grant RYC2021-033560-I funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR, and by grant ED431F 2024/02 funded by Xunta de Galicia, Spain-GAIN.

References

1. D. Parekh *et al.*, “A Review on Autonomous Vehicles: Progress, Methods and Challenges,” *Electronics*, vol. 11, no. 14, Art. no. 14, Jan. 2022, doi: 10.3390/electronics11142162.
2. M. Alam, J. Ferreira, and J. Fonseca, “Introduction to Intelligent Transportation Systems,” in *Intelligent Transportation Systems: Dependable Vehicular Communications for Improved Road Safety*, M. Alam, J. Ferreira, and J. Fonseca, Eds., Cham: Springer International Publishing, 2016, pp. 1–17. doi: 10.1007/978-3-319-28183-4_1.
3. “Road traffic injuries.” Accessed: July 10, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
4. D. Ponsa, J. Serrat, and A. M. López, “On-board image-based vehicle detection and tracking,” *Transactions of the Institute of Measurement and Control*, vol. 33, no. 7, pp. 783–805, Oct. 2011, doi: 10.1177/0142331209103039.
5. J. Wang, Q. Si, Z. Song, D. Wang, and H. Yao, “3D Geometry Modeling and Safety Compliance Assessment of In-Service Roads Using Massive LiDAR Point Clouds,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 4977–4986, June 2024, doi: 10.1109/TITS.2023.3339643.
6. Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023, doi: 10.1109/TPAMI.2022.3179507.
7. L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, “LIF-Seg: LiDAR and Camera Image Fusion for 3D LiDAR Semantic Segmentation,” Aug. 17, 2021, *arXiv*: arXiv:2108.07511. doi: 10.48550/arXiv.2108.07511.
8. S. Pang, D. Morris, and H. Radha, “CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection,” Sept. 02, 2020, *arXiv*: arXiv:2009.00784. doi: 10.48550/arXiv.2009.00784.
9. D. Caffagni *et al.*, “The Revolution of Multimodal Large Language Models: A Survey,” June 06, 2024, *arXiv*: arXiv:2402.12451. doi: 10.48550/arXiv.2402.12451.
10. H. Meshkin *et al.*, “Harnessing large language models’ zero-shot and few-shot learning capabilities for regulatory research,” *Briefings in Bioinformatics*, vol. 25, no. 5, p. bbae354, Sept. 2024, doi: 10.1093/bib/bbae354.
11. Y. Hong *et al.*, “3D-LLM: Injecting the 3D World into Large Language Models,” presented at the Thirty-seventh Conference on Neural Information Processing Systems, Nov. 2023. Accessed: July 21, 2025. [Online]. Available: https://openreview.net/forum?id=YQA28p7qNz&utm_source=chatgpt.com
12. A. Vijayakumar and S. Vairavasundaram, “YOLO-based Object Detection Models: A Review and its Applications,” *Multimed Tools Appl*, vol. 83, no. 35, pp. 83535–83574, Oct. 2024, doi: 10.1007/s11042-024-18872-y.
13. M. Lei *et al.*, “YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception,” June 21, 2025, *arXiv*: arXiv:2506.17733. doi: 10.48550/arXiv.2506.17733.
14. G. Al-refai and M. Al-refai, “Road Object Detection using Yolov3 and Kitti Dataset,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110807.
15. J. Behley, A. Milioto, and C. Stachniss, “A Benchmark for LiDAR-based Panoptic Segmentation based on KITTI,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 13596–13603. doi: 10.1109/ICRA48506.2021.9561476.
16. N. Liu, X. Zhang, L. Wang, L. Xu, and S. Zhu, “3-D Point Cloud Object Completion via RGB Images With Complex Geometric Topology in Urban Scenes,” *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 1–5, 2025, doi: 10.1109/LGRS.2024.3488724.
17. A. Patil, S. Malla, H. Gang, and Y.-T. Chen, “The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes,” in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 9552–9557. doi: 10.1109/ICRA.2019.8793925.
18. L. Tang, P.-T. Jiang, Z.-H. Shen, H. Zhang, J.-W. Chen, and B. Li, “Chain of Visual Perception: Harnessing Multimodal Large Language Models for Zero-shot Camouflaged Object Detection,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, in MM ’24. New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 8805–8814. doi: 10.1145/3664647.3680730.
19. H. I. Ashqar, T. I. Alhadidi, M. Elhenawy, and N. O. Khanfar, “Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems,” *Automation*, vol. 5, no. 4, Art. no. 4, Dec. 2024, doi: 10.3390/automation5040029.
20. F. Yang *et al.*, “LLMI3D: MLLM-based 3D Perception from a Single 2D Image,” Feb. 13, 2025, *arXiv*: arXiv:2408.07422. doi: 10.48550/arXiv.2408.07422.
21. X. Huang, J. Wu, Q. Xie, and K. Han, “MLLMs Need 3D-Aware Representation Supervision for Scene Understanding,” June 02, 2025, *arXiv*: arXiv:2506.01946. doi: 10.48550/arXiv.2506.01946.
22. D. Liu *et al.*, “Uni3D-LLM: Unifying Point Cloud Perception, Generation and Editing with Large Language Models,” Jan. 09, 2024, *arXiv*: arXiv:2402.03327. doi: 10.48550/arXiv.2402.03327.
23. “Gemini 2.0 Flash - Model Card”.
24. “autonomousvision/kitti360Scripts: This repository contains utility scripts for the KITTI-360 dataset.” Accessed: July 23, 2025. [Online]. Available: https://github.com/autonomousvision/kitti360Scripts/tree/master
25. D. Deng, “DBSCAN Clustering Algorithm Based on Density,” in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Sept. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.
26. M. Gong, D. Wang, X. Zhao, H. Guo, D. Luo, and M. Song, “A review of non-maximum suppression algorithms for deep learning target detection,” in *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, SPIE, Mar. 2021, pp. 821–828. doi: 10.1117/12.2586477.