Multimodal Large Language Models to road inventory with non-photorealistic Point Cloud visualization

Horia Ameen, Mario Soilán, Henrique Lorenzo, Jesús Balado

CINTECX, Universidade de Vigo, GeoTECH, 36310 Vigo, Spain - (horia.ameen,msoilan,hlorenzo,jbalado)@uvigo.gal

Keywords: LiDAR**,** Mobile Laser Scanning, Mobile Mapping Systems, Visual Language Models, Deep Learning.

Abstract

Accurate road inventories are crucial for maintenance, safety, and resource allocation, with automation improving efficiency but often lacking user-friendly human-machine interaction. This paper evaluates how non-photorealistic rendering of 3D point clouds impacts Multimodal Large Language Models (MLLMs) interpretation for road inventory, testing three methods on real road data in Santarém (Portugal). From 3D point clouds coloured with RGB information, non-photorealistic techniques are implemented and compared: Ambient Occlusion (AO), Eye-Dome Lighting (EDL) and Multi Feature-Rich Synthetic Color (MFRSC). Several state-of-the-art MLLMs are also tested: GPT5, Gemini2.5-Pro, Gemini2.5-Flash, CogVLM2, MiniCPM-V, Llama4-scout-17b, Mistral-Small3.2, Qwen 2.5vl and Gemma3. The results indicate that non-photorealistic techniques do not hinder the identification of road elements by MLLMs, indicating their potential for 3D point cloud classification tasks even when true RGB colour is not available. Furthermore, the overall performance metrics, with F-scores over 80% for proprietary, state-of-the-art models (GPT5, Sonnet 4.5 and Gemini) show that 2D captures of 3D point clouds can be a suitable data source for zero-shot object classification. Rather than proposing new algorithms, this work contributes an empirical evaluation of how non-photorealistic point-cloud visualizations affect VLM-based road inventory interpretation.

1. Introduction

Accurate and up-to-date inventories are essential for ensuring the functionality and safety of roadways, identifying areas requiring maintenance, and optimizing resource allocation for infrastructure projects. With the increasing availability of geospatial data, the automation of these inventories has begun to save hours of manual inspection (Barros-Sobrín et al., 2024; Rúa et al., 2023; Tardy et al., 2023). However, human-machine interaction with automated methods, such as Deep Learning (DL), is unfriendly.

Visual Language Models (VLM) improve the interpretation of visual data using natural language. This allows for a more nuanced understanding of the environment, helping to identify road features, conditions and anomalies in an intuitive way (Gavrikov et al., 2024; Umeike et al., 2025). However, 3D point clouds have limitations in their visualization, such as seeing through surfaces, large point density variations and errors in colouring (González et al., 2022; Remondino, 2003).

This paper evaluates the influence of point cloud non-photorealistic methods on VLM interpretation for road inventory. The novelty of this work lies in this evaluation and the resulting insights, not in introducing a new VLM or rendering algorithm. Non-photorealistic methods are techniques used in computer graphics to create visual representations that do not aim to mimic real-life appearances as possible, but focus on stylized, artistic, or abstract visualizations. The implemented methods are Ambient Occlusion, Eye-Dome Lighting and Multi Feature-Rich Synthetic Colour, and they are tested on 1km of real roads.

The structure of this paper is as follows. Section 2 presents a compilation of works related to VLM applied to point clouds. The proposed method is detailed in Section 3. Section 4 focuses on presenting, analysing and discussing the results. Section 6 concludes the paper.

2. Related Work

Multimodal Large Language Models (MLLMs) integrate text with other modalities such as images or audio, enabling advanced reasoning and generative capabilities over different modes of information. Early models focused on tasks such image captioning and Visual Question Answering (VQA) based on Convolutional Neural Networks (CNNs) (Vinyals et al., 2015). Then, transformer-based approaches such as ViLBERT (Lu et al., 2019) enabled improved multimodal reasoning enhancing VQA performance. The introduction of CLIP by OpenAI (Radford et al., 2021), which employed a contrastive learning approach to align images and text in a shared embedding space, was trained on over 400 million image-text pairs, showing zero-shot learning capabilities, that is, being able to generalize across a wide range of concepts without specific fine-tuning.

Since the release of ChatGPT in November 2023, access and usage of LLMs have been democratized and widely adopted at many societal and technological scales: By the start of 2025, ChatGPT has more than 400 million weekly active users, according to OpenAI. Their general-purpose model is GPT4o, a multimodal model that enhances real-time interaction and reasoning across multiple data modalities (text, images and audio) (OpenAI et al., 2024). In August 2025, OpenAI introduced GPT-5, their most advanced general-purpose model, bringing notable gains in reasoning, speed, and agentic/coding workflows. GPT-5 also powers a unified system that routes between a fast default model and a deeper reasoning variant (GPT-5 Thinking), and it’s available in both ChatGPT and the OpenAI API.

Obviously, there exist other actors in this technology market that provide proprietary multimodal models with comparable performance, such as Gemini from Google, or Claude from Anthropic, whose multimodal models are Gemini 2.0 (Google Deepmind, 2024) and Claude 3.7 (Anthropic, 2025).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-793-2026

Opposing to proprietary models, open-source initiatives have provided competitive model alternatives. Relevant examples are Llama 3.2 (Meta AI, 2024), CogVLM2 (Hong et al., 2024), or MiniCMP-V (Yao et al., 2024). Open-source models allow open usage and modification, further democratizing access to multimodal intelligence.

A current frontier in this multimodal research involves extending MLLMs to 3D data representations, particularly in the context of point cloud processing. Recent work has explored two approaches in this regard: Transferring knowledge from 2D vision-language models to 3D understanding and developing 3D-native multimodal models.

A notable example of the former is PointCLIP, which in its version V2 performs zero-shot 3D classification, segmentation and detection by projecting 3D objects into multiple, dense 2D depth maps (Zhu et al., 2023). Differently, PointLLM directly understands object point clouds and generate appropriate answers to human instructions (Xu et al., 2024), combining a point cloud encoder with an LLM. This model is on par to human annotators in 3D object captioning tasks. Similarly, Point-Bind extends an image multimodal learning framework to align point clouds, images, text and audio within a unified representation space (Guo et al., 2023).

Leveraging 2D visualizations of point clouds faces several challenges when compared to 3D-native multimodality, such as the loss of geometric fidelity or difficulties for LLMs to process multiple views simultaneously. However, 2D visualizations allow pretrained, general purpose multimodal models to interpret 3D data.

Thus, this study hypothesizes that general purpose MLLMs can classify point clouds under photorealistic visualizations, such as Ambient Occlusion, Eye-Dome Lightning, or Multi-Feature Synthetic Colour. This will provide insight to researchers into how these models process 3D point cloud data, with no specific training or adaptation to this mode of data. While mentioned related work relies on depth maps or multi-view images, in this work a case study presents a single-image classification task from an oblique aerial perspective of objects in road infrastructure environments. Under this hypothesis, the contributions of this work are:

1) A quantitative performance analysis of different state of the art MLLMs, both proprietary and open-source for a multi-class classification task on 2D projections of 3D point clouds of road infrastructure.

2) A comparative analysis of different photorealistic visualization techniques for 3D point clouds that allows to prove the feasibility of zero-shot classification with general purpose MLLMs under different visualization modes.

3. Methods

The proposed methodology follows a structured workflow (Figure 1) to assess the capability of Multimodal Large Language Models (MLLMs) in classifying road infrastructure elements from 2D visualizations of 3D point clouds. First, point cloud data is acquired and then processed into four different visualization styles—RGB, Ambient Occlusion (AO), Eye-Dome Lighting (EDL), and Multi-Feature-Rich Synthetic Colour (MFRSC)—to evaluate how non-photorealistic rendering affects model interpretation. The resulting 2D projections are subsequently fed into ten state-of-the-art MLLMs (GPT5, Gemini2.5-Pro, Gemini2.5-Flash, CogVLM2, MiniCPM-V, Llama4-scout-17b, Mistral-Small3.2, Qwen 2.5vl and Gemma3) for classification of key road elements. The classification outputs were compared against a manually labelled ground truth to compute accuracy, precision, recall, and F1-score, enabling a quantitative assessment of the models' performance across different visualization techniques.

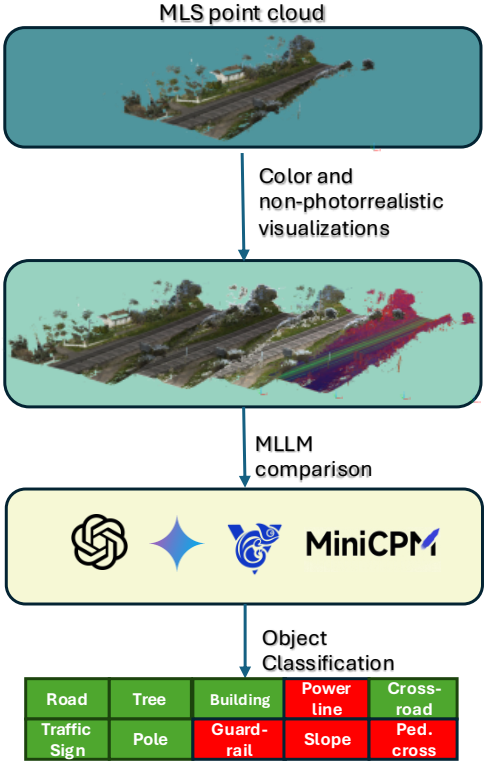

Figure 1. Methodology workflow

3.1 Ambient Occlusion

Ambient occlusion (AO) is a shading technique that simulates how ambient light is blocked by the geometry of a 3D scene. Unlike direct illumination, which comes from specific light sources, ambient light is diffuse and comes from all directions. AO calculates how much ambient light reaches each point on a surface, obscuring areas that are surrounded by other geometry. This creates a subtle but realistic shading effect that brings out the details and depth of the scene.

The AO calculation is based on integrating the visibility over a hemisphere around each point on the surface. Rays are launched from each point in random directions within the hemisphere. The number of rays hitting the nearby geometry determines the occlusion factor (1).

$$AO(i) = \frac{1}{R}\sum\big(1 - v(i,\omega)\big) \tag{1}$$

with R the number of rays and $v(i,\omega)$ the visibility of the ray in ω direction from i point

3.2 Eye-Dome Lighting

Eye Dome Lighting (EDL) is a non-photorealistic shading technique designed to improve depth perception in 3D visualizations, especially in point clouds and dense 3D models. Instead of simulating physical illumination, EDL focuses on enhancing surface details by calculating a local occlusion factor. This factor is based on the density of points near each surface

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-793-2026 | © Author(s) 2026. CC BY 4.0 License.

point, creating a shading effect that highlights variations in geometry. First, a depth buffer of the scene is generated. Then, for each pixel, an occlusion factor is calculated by comparing its depth with that of neighbouring pixels within a defined radius. This factor determines value of hidden of the pixel by the surrounding geometry. Finally, this factor is used to darken the occluded pixels, creating a shading effect that highlights surface details.

3.3 Multi-Feature Reach Synthetic Colour

Multi Feature-Rich Synthetic Colour (MFRSC) (Balado et al., 2023) is based on visualization through a color scheme generated from the features related to human perceptual descriptors. The selected features are reflectance, return number, inclination (2), depth (3), height, point density (4), linearity (5), planarity (6), and scattering (7). These features are then normalized between 0 and 1 and combined through a reduction of RGB to greyscale color conversion channels (Figure *2*).

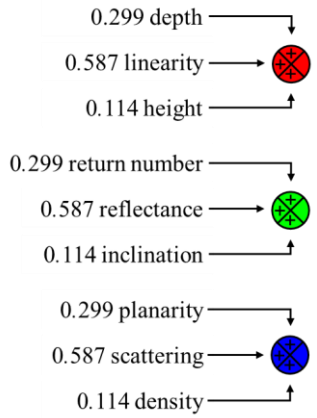

Figure 2. Schematic diagram of the feature distribution according the RGB colour channels.

$$inclination = \left| atan\left(\frac{\sqrt{N_x^2 + N_y^2}}{N_z} \right) \right| \quad (2)$$

with N the normal surface to the nearest 25 points

$$depth = \sqrt{P_x^2 + P_y^2} \quad (3)$$

with P each point of the cloud

$$point\_density = \frac{d_1}{d_4} \quad (4)$$

with d the distance to the first/fourth neighbour

$$linearity = \frac{\lambda_1 - \lambda_2}{\lambda_1} \quad (5)$$

with λ the eigenvalue

$$planarity = \frac{\lambda_2 - \lambda_3}{e_1} \quad (6)$$

$$scattering = \frac{\lambda_3}{\lambda_1} \quad (7)$$

3.4 Multimodal Language Models

For each of the point clouds processed with the visualisation methods described in the previous subsections, screenshots are taken (samples in Section 4.1). These screenshots are processed with different MLLMs to perform a multi-class classification problem, so their capacity to interpret images of 3D point clouds can be evaluated. All screenshots were generated from a consistent aerial viewpoint chosen to provide a broad overview of the road scene, which is the standard perspective used during manual inspection of road point clouds. To reduce variability, we kept the camera viewpoint strategy consistent across all scenes and visualization modes (RGB, AO, EDL, MFRSC), ensuring the same general road coverage and context in each rendered image. Viewpoint changes (e.g., different angles, zoom levels, or occlusions) may influence visibility of smaller objects such as signs, poles, and power lines.

3.4.1 Model selection and description

To compare the performance of different models, ten MLLMs were selected: GPT5, Sonnet 4.5, Gemini2.5-Pro, Gemini2.5-Flash, CogVLM2, MiniCPM-V, Llama4-scout-17b, Mistral-Small3.2, Qwen 2.5vl and Gemma3. The motivation of this selection of models followed two criteria. First, to use recently published, state-of-the-art models. Second, to cover different strategies for using a language model, from web interfaces of proprietary models to local use of open-source models:

- **Web-based model (GPT5)**: ChatGPT web interface from OpenAI allows to create custom models based in GPT5, whose behaviour and functionality can be modified by the user to align the model with the problem to solve. It requires the definition of a system prompt that guides such behaviour, and textual background can be attached to the model as its knowledge base. The system prompt is defined in section 3.4.2.
- **API-based models:** The Google family of LLMs can be accessed through its API, which offers a comprehensive platform for integrating the AI capabilities of their models into automated pipelines. This API includes a free tier whose current effective tokens-per-minute (TPM) limit is 125,000 per project. In this work we used Gemini 2.5-Flash and Gemini 2.5-Pro under that free tier, which was sufficient for the scale of our case study.
- **Local-based and open-source models**: Although proprietary models such as GPT5 and Gemini do not require of local hardware for model inference and show state-of-the-art results in different evaluation benchmarks, users have little control over the models and their usage limitations. Furthermore, there is always a data privacy concern as data is processed in external cloud services. Therefore, in this work two different open-source models are evaluated and compared. First, CogVLM2, through HuggingFace’s Transformers library. Second, MiniCPM-V 2.6, Llama4-scout-17b, Mistral-Small3.2, Qwen 2.5vl and Gemma3 through Ollama library. The weights of both models are locally downloaded so data remains private and there are no limitations on model inference as long as the hardware can allocate the models.

3.4.2 Model prompts

One of the key elements for the correct performance of a MLLM is the correct definition of its instructions through a text prompt. The prompts designed for this work are summarized in Table 1. As can be seen, the same prompt has been used for all models but CogVLM2. This is motivated by the fact that this model was found to answer inconsistently when asked to express their output in a structured format. Furthermore, given the fact that this model was locally available and does not present any usage limitations, it was found relevant to perform a multiple VQA experiment on the dataset to compare with a single prompt experiment, where

model inference is done for each object separately, subsequently parsing the answers into the same format than for the other models. The remaining four models are prompted to answer with a Python list of Booleans corresponding to the presence or absence of the objects to classify in the point cloud image. Additionally, the custom GPT is prompted to output the name of the image to build a dictionary. This step is coded in a parsing step for the other models.

Model	Prompt	Answer example
GPT5	You will only answer to messages that have images as input. If you don't have images as input, ask for images and do not interact further until the user attaches images to their input. For each image that you have as input, you will follow these steps: 1) Understand the context of the image. It is a screenshot from a 3D point cloud, so assume a sparsity in the visual content. However, it will represent an outdoor reality. The image will be taken from an aerial point of view. It may have real RGB colors, or a false color that enhances features. Try to understand the image independently of this. 2) Read this list of objects/elements: [Road, Tree, Building, Power lines, Crossroad, Traffic Sign, Pole, Guardrail, Slope, Pedestrian Crossing]. 3) Take each object of the list. Is it present in the image? If you see it explain where it is within the context of the image. 4) Then, summarize the elements you found and answer with a list of booleans (true or false) in Python format with the same structure of the object list (e.g. if you se a road, then element 0 of your answer will be a true value) *For GPT5 and Sonnet 4.5 only: [5) Answer in JSON format, with the name of the image as a key and labels list as a value.]	{"image_name": "CS01_RGB.png", "labels": [True, True, True, False, True, True, True, False, False, False]}
Sonnet 4.5		[True, True, True, False, True, True, True, False, False, False]
Gemini 2.5-Pro		[True, True, True, False, True, True, True, True, True, False]
Gemini 2.5-Flash		[True, True, True, True, True, True, True, False, False, False]
MiniCMP-V		[True, False, True, False, False, False, True, False, False, False]
CogVLM2	Is there a [object name] in the image?	Yes, there is a [object name] in the image / No, there is not a [object name] in the image

Table 1. Model prompts and answer examples

4. Results

4.1 Case Study

The proposed method was tested on 1km of N3 and N118 roads in Santarém (Portugal). The point clouds were generated with a Lynx Mobile Mapper, with a Ladybug5 360º camera and a GPS-IMU Applanix POS LV 520 (Balado et al., 2020). The point density of the clouds was 100 points per square meter on road surface. The images were generated in Cloud Compare, with a point size of 4. In total, 48 images corresponding to 12 different areas were generated with 4 visualizations (RGB without non-photogrammetric modification, AO, MFRSC, EDL) from the same perspective. Samples from the generated images can be seen in Figure *3*.

To validate the performance of the model, a manual ground truth dataset was created. This ground truth included labelled instances of 10 different infrastructure elements: Road, Tree, Building, Power Lines, Crossroads, Traffic Signs, Poles, Guardrail, Slope and Pedestrian Crossing.

4.2 MLLM evaluation

To evaluate the performance of the different MLLMs considered for this work, the answers given by the models are directly compared with the manual ground truth elaborated for the case study data. The classification metrics employed for this evaluation were Accuracy, Precision, Recall, and F1-score. Results are compiled in Table 2, Figure 4 and Figure 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

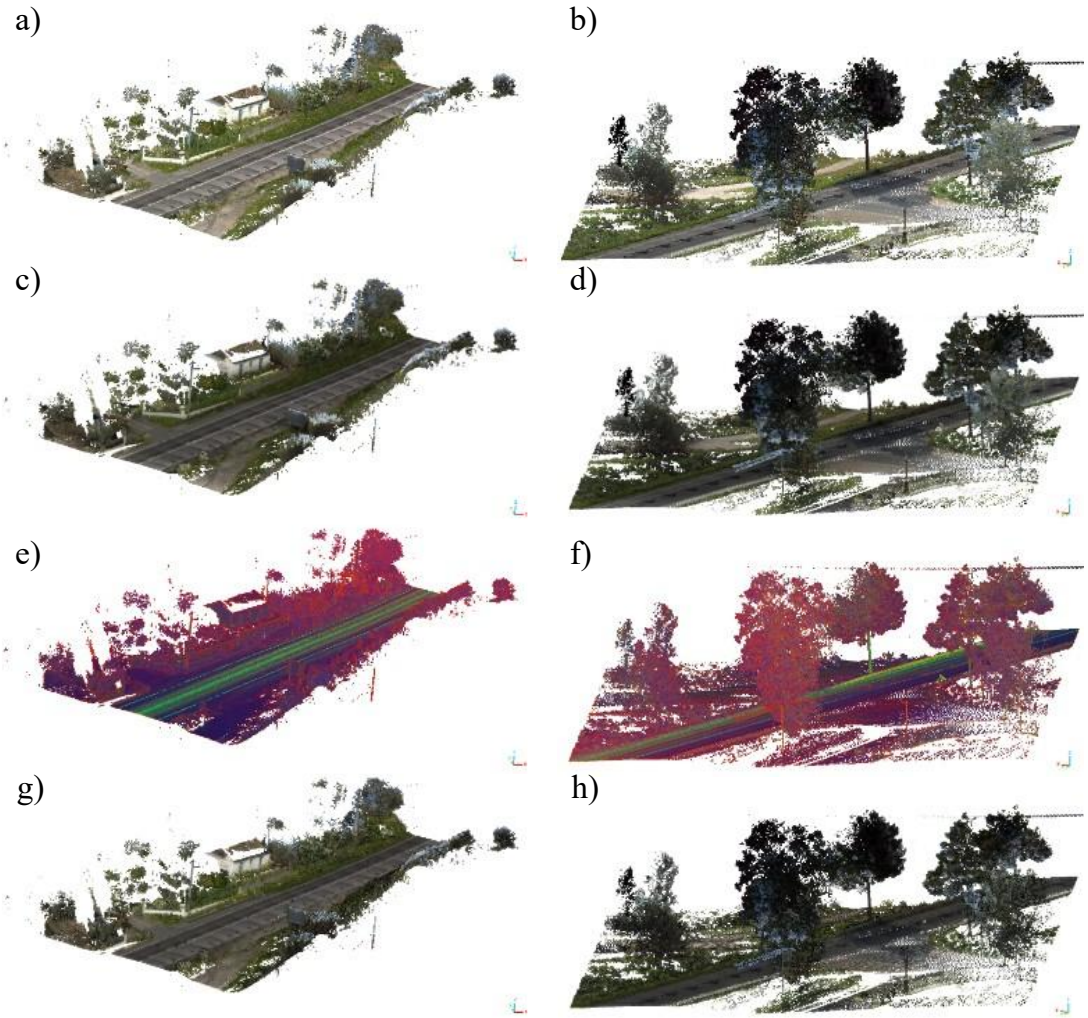

Figure 3. Samples generated from 3D point cloud road N3 (left) and N118 (right) with RGB camera (a-b), AO with artificial ambient light (c-d), MFRSC based on point cloud features (e-f) and EDL with improved depth perception (g-h).

Model	Accuracy	Precision	Recall	F1-score
GPT5	0.970	0.942	0.98	0.970
Sonnet 4.5	0.858	0.896	0.825	0.859
GPT4o	0.826	0.779	0.935	0.850
Gemini 2.5-Flash	0.792	0.773	0.853	0.811
Gemini 2.0 Flash	0.783	0.742	0.901	0.814
Gemini-2.5-Pro	0.756	0.734	0.841	0.784
Gemini 1.5 Pro	0.765	0.817	0.710	0.760
Llama4-scout-17b	0.704	0.784	0.603	0.682
Mistral-Small3.2	0.690	0.759	0.599	0.670
CogVLM2	0.660	0.611	0.972	0.750
MiniCPM-V 2.6	0.629	0.668	0.610	0.638
Qwen 2.5vl	0.625	0.607	0.810	0.694
Gemma3	0.610	0.609	0.718	0.659

Table 2. Global metrics across all visualizations.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-793-2026 | © Author(s) 2026. CC BY 4.0 License.

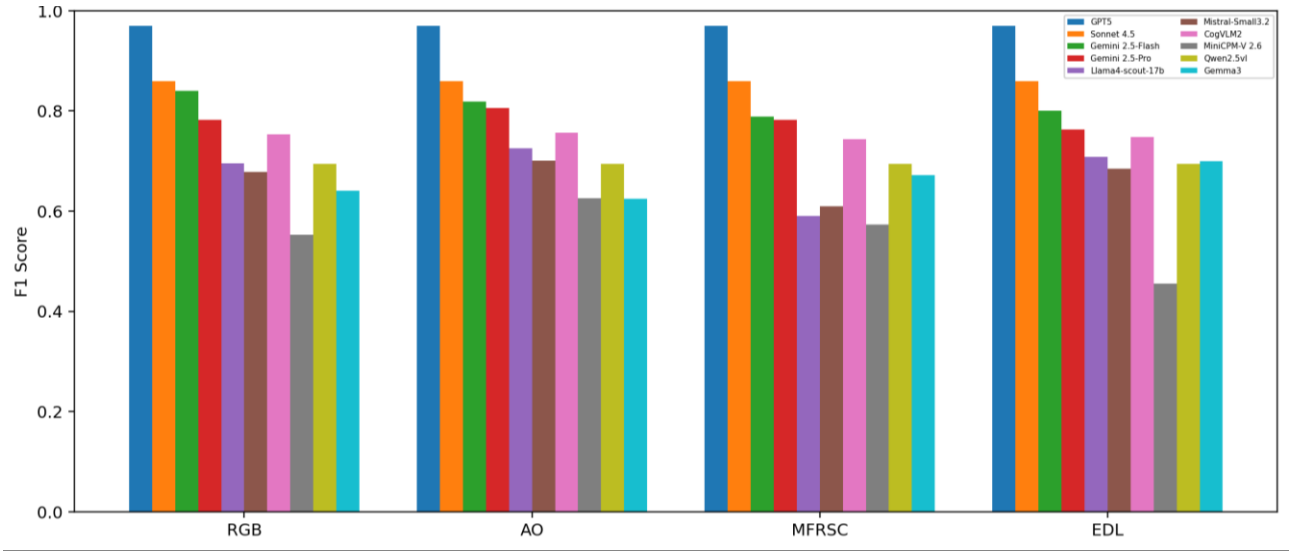

Figure 4. F1 scores per visualization method.

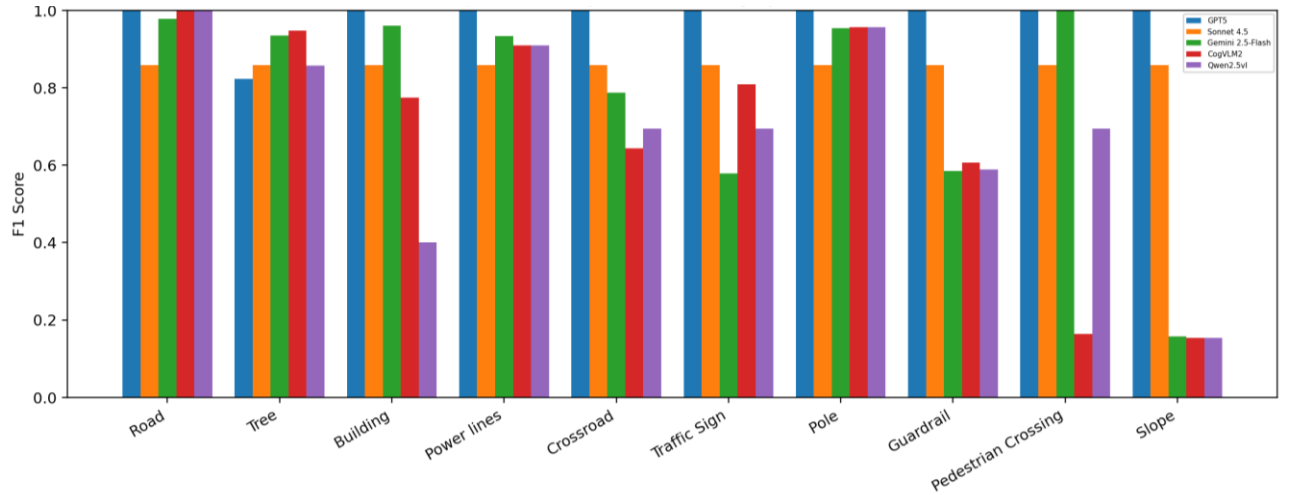

Figure 5. F1 scores per class.

5. Discussion

The results show a clear hierarchy among models and some consistent failure modes across categories and visualization types. The error distribution across classes reveals a consistent divide between structurally large, continuous objects and small, thin, or semantically defined objects. Classes such as Road, Pole, Building, and Power lines exhibit the highest stability in Figure 5. Their geometry extends over many pixels and remains recognizable under the different renderings. In contrast, Traffic Sign, Guardrail, Slope, and Crossroad concentrate most failures among Ollama models. Misclassifications in these “fragile geometry” classes are dominated by false negatives, objects go undetected when contrast is low, segments are partially occluded, or the 2D projection breaks geometric continuity.

GPT-5 is consistently at or near ceiling across categories and visualization types, with residual errors occurring almost exclusively in very thin or distant structures (guardrail spans behind vegetation, small signs without readable faces). These errors are predominantly missing rather than confusions, indicating limits set by signal quality and geometric continuity rather than semantic knowledge. The jump from GPT-4 to GPT-5 is marked by higher recall on small objects and greater invariance to rendering style (RGB vs. AO/MFRSC/EDL). Among the tested visualization techniques, RGB typically yields the highest classification performance, confirming that standard color-based representations of 3D point clouds are effective for visual interpretation by MLLMs. However, alternative techniques such as Ambient Occlusion (AO), Eye-Dome Lighting (EDL), and Multi-Feature-Rich Synthetic Color (MFRSC) show only minor variations in performance across models.

GPT-5 “thinking” runs a structured, multi-step reasoning pass before answering, planning, checking intermediate hypotheses, and self-verifying results. In practice, that yields more stable decisions under ambiguous evidence (e.g., thin/occluded objects) and better consistency across visualization styles. Qwen 2.5vl follows a similar trend at a lower operating point. Lightweight models are most sensitive to render changes and class ambiguity, which manifests as missed detections on Traffic Sign and Slope and occasional confusion between Pole and Guardrail segments. Finally, all results are zero-shot on 2D captures of point clouds; while this setting is operationally attractive, it constrains the models’ ability to exploit depth or multi-view consistency. Future

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-793-2026 | © Author(s) 2026. CC BY 4.0 License.

work should evaluate prompt-stable, geometry-aware pipelines that fuse non-photorealistic renders with minimal 3D priors to specifically target Traffic Sign, Guardrail, Slope, and Crossroad, which remain the principal bottlenecks even for state-of-the-art models.

6. Conclusions

This work explored the impact of non-photorealistic rendering techniques on the ability of Multimodal Large Language Models (MLLMs) to classify road infrastructure elements from 2D representations of 3D point clouds. The evaluation of state-of-the-art MLLMs including proprietary and open-source models showed that the classification of road elements remains feasible across different visualization styles. The results demonstrate that even in the absence of true RGB, general-purpose MLLMs can accurately interpret structural road features, achieving strong performance in the best cases.

These findings have significant implications for road inventory automation using Mobile Laser Scanning (MLS) data. The results suggest that non-photorealistic visualization techniques, in conjunction with advanced MLLMs, can support zero-shot classification tasks without requiring model fine-tuning for point-cloud data. MFRSC can serve as a viable substitute when true color information is unavailable. This could facilitate the development of more accessible and scalable road-infrastructure monitoring systems, reducing reliance on manual annotation and photorealistic imaging.

Despite the promising results, this work presents some limitations. While this study focuses on a 1 km MLS road segment to isolate the impact of visualization strategies under controlled conditions, we plan to extend the evaluation to larger and more diverse MLS areas (e.g., different vegetation densities, lighting conditions, and road topologies) and to systematically assess multiple viewpoints (e.g., alternative aerial/oblique perspectives) to better quantify generalization across environments.

This study evaluates a single viewpoint strategy representative of typical road-inventory inspection. Future work will extend the evaluation to multi-viewpoint sampling (e.g., multiple azimuth/elevation angles and zoom levels) and investigate view aggregation strategies to further improve robustness and to explore multi-view integration strategies, enabling MLLMs to process multiple 2D projections of the same 3D scene for more comprehensive object recognition. Additionally, smaller or less visually distinct elements such as power lines and guardrails exhibited higher classification variability, highlighting the need for enhanced visual prompts or multi-perspective data inputs. Moreover, investigating task-specific fine-tuning of open-source models could provide an alternative to proprietary models, ensuring greater transparency and adaptability for real-world deployment. Finally, applying this workflow to object detection in 2D within a 2D/3D data-fusion pipeline may allow for enhanced inventories of road-infrastructure assets with Mobile Laser Scanning.

Acknowledgements

Jesús Balado would like to thank the funding from Government of Spain through RYC2022-038100-I by MCIN/AEI/10.13039/501100011033 and FSE+; and from Xunta de Galicia - GAIN [EDC431C 2024/30, ED431F 2024/06]. Mario Soilán would like to thank the funding 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

Anthropic, 2025. Claude 3.7 Sonnet and Claude Code [WWW Document]. URL https://www.anthropic.com/news/claude-3-7-sonnet (accessed 3.3.25).

Balado, J., González, E., Arias, P., Castro, D., 2020. Novel Approach to Automatic Traffic Sign Inventory Based on Mobile Mapping System Data and Deep Learning. Remote Sens. 12. https://doi.org/10.3390/rs12030442

Balado, J., González, E., Rodríguez-Somoza, J.L., Arias, P., 2023. Multi feature-rich synthetic colour to improve human visual perception of point clouds. ISPRS J. Photogramm. Remote Sens. 196, 514–527. https://doi.org/10.1016/j.isprsjprs.2023.01.019

Barros-Sobrín, Á., Balado, J., Soilán, M., Mingueza-Bauzá, E., 2024. Gamification for road asset inspection from Mobile Mapping System data. J. Spat. Sci. 69, 443–466. https://doi.org/10.1080/14498596.2023.2236996

Gavrikov, P., Lukasik, J., Jung, S., Geirhos, R., Lamm, B., Mirza, M.J., Keuper, M., Keuper, J., 2024. Are Vision Language Models Texture or Shape Biased and Can We Steer Them? https://doi.org/10.48550/arXiv.2403.09193

González, E., Balado, J., Arias, P., Lorenzo, H., 2022. Realistic correction of sky-coloured points in Mobile Laser Scanning point clouds. Opt. Laser Technol. 149, 107807. https://doi.org/10.1016/j.optlastec.2021.107807

Google Deepmind, 2024. Google introduces Gemini 2.0: A new AI model for the agentic era [WWW Document]. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/ (accessed 3.3.25).

Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., Heng, P.-A., 2023. Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following. https://doi.org/10.48550/arXiv.2309.00615

Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., Zhao, L., Yang, Z., Gu, X., Zhang, X., Feng, G., Yin, D., Wang, Z., Qi, J., Song, X., Zhang, P., Liu, D., Xu, B., Li, J., Dong, Y., Tang, J., 2024. CogVLM2: Visual Language Models for Image and Video Understanding. https://doi.org/10.48550/arXiv.2408.16500

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Meta AI, 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models [WWW Document]. Meta AI. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (accessed 12.18.24).

OpenAI, Hurst, A., Lerer, A., Goucher, A.P., et al., 2024. GPT-4o System Card. https://doi.org/10.48550/arXiv.2410.21276

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020

Remondino, F., 2003. From point cloud to surface: the modeling and visualization problem. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 34.

Rúa, E., Núñez-Seoane, A., Arias, P., Martínez-Sánchez, J., 2023. Automatic detection to inventory road slopes using open LiDAR point clouds. Int. J. Appl. Earth Obs. Geoinformation 118, 103225. https://doi.org/10.1016/j.jag.2023.103225

Tardy, H., Soilán, M., Martín-Jiménez, J.A., González-Aguilera, D., 2023. Automatic Road Inventory Using a Low-Cost Mobile Mapping System and Based on a Semantic Segmentation Deep Learning Model. Remote Sens. 15, 1351. https://doi.org/10.3390/rs15051351

Umeike, R., Getty, N., Xia, F., Stevens, R., 2025. Scaling Large Vision-Language Models for Enhanced Multimodal Comprehension In Biomedical Image Analysis. https://doi.org/10.48550/arXiv.2501.15370

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D., 2024. PointLLM: Empowering Large Language Models to Understand Point Clouds. https://doi.org/10.48550/arXiv.2308.16911

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M., 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. https://doi.org/10.48550/arXiv.2408.01800

Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P., 2023. PointCLIP V2: Prompting CLIP and GPT for Powerful 3D Open-world Learning, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2639–2650. https://doi.org/10.1109/ICCV51070.2023.00249