2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

3D SEGMENTATION AGENT INTEGRATING 3D GAUSSIAN SPLATTING AND VISION-LANGUAGE MODELS FOR AS-BUILT ANALYSIS

Boyu Wang[1], Jun Ma[1], and Borja García de Soto [2]
[1]The University of Hong Kong, Hong Kong, China
[2]New York University Abu Dhabi (NYUAD), Abu Dhabi, United Arab Emirates

Abstract

The transition from raw as-built data to semantic object representations is a bottleneck in construction digitalization, currently hampered by manual dependencies and the "black-box" nature of existing point cloud segmentation networks. This paper introduces an autonomous self-correcting 3D segmentation agent tailored for construction environments, utilizing 3D Gaussian Splatting (3DGS) as the underlying representation and Vision Language Models (VLMs) as the reasoning core. Unlike static segmentation methods, our agent adopts a dynamic approach by performing active scene roaming. Through a closed-loop feedback mechanism, the VLM analyzes visual data, detects objects, and rigorously assesses the plausibility of segmentation results. The agent iteratively corrects its own errors until a high-fidelity, logically consistent segmentation is achieved. This method demonstrates superior interpretability and accuracy, enabling downstream applications such as as-built verification, progress tracking, and quantity takeoff. By bridging generative rendering with semantic reasoning, our approach provides a robust framework for automated 3D scene understanding.

Introduction

Construction automation is progressively evolving through four stages: data acquisition, semantic understanding, human-robot interaction, and fully autonomous construction. While the latter stages remain nascent, the industry has made significant strides in the first stage. Data acquisition has matured from manual logging to high-precision 3D laser scanning and, more recently, robust mobile LiDAR-visual Simultaneous Localization and Mapping (SLAM) systems. These technologies accurately capture the 3D geometry of construction sites. However, the resulting data remains unstructured. To enable automated tasks like as-built modeling and quality assessment, this raw geometric data must be transformed into structured, semantic representations.

Semantic 3D segmentation is the critical bridge for this transformation, yet it faces two significant barriers in construction contexts: 1) Generalization gaps**:** Conventional models trained on structured datasets (e.g., offices) often fail in dynamic, cluttered construction environments due to domain shifts. 2) Limited robustness and interpretability: "Black-box" deep learning models lack physical reasoning capabilities, often yielding results that are geometrically or topologically implausible.

To address these system-level challenges, we propose a 3D segmentation agent integrating 3D Gaussian Splatting (3DGS) and vision-language models (VLMs) inspired by the cognitive workflow of human engineers. Unlike end-to-end models, human experts adopt an iterative strategy: they first perform a rough identification, evaluate the result's plausibility (e.g., checking for completeness and topological connectivity), and actively manipulate their view including zooming in, rotating, or translating to resolve ambiguities before finalizing the segmentation.

Mimicking this paradigm, our framework, illustrated in Figure 1, integrates a 3DGS Rendering Engine with two core modules. Unlike 'black-box' neural networks that rely on implicit latent spaces, our approach achieves high interpretability by decomposing the segmentation task into an explicit, human-readable cognitive workflow. The Analysis Module utilizes VLM-based detection and topological reasoning to assess segmentation quality ("Plausibility Analysis"). If the result is unsatisfying or ambiguous, the system triggers the Operation Module. This module executes active visual manipulation such as multi-scale zooming, 3D rotation, and cropping to acquire clearer details, thereby achieving a robust, self-correcting understanding of the 3D scene.

Literature Review

Deep Learning-Based 3D Segmentation

Point cloud segmentation has evolved from manual processing to automated deep learning. Fully supervised methods, employing MLP, convolutional, or attention-based architectures, have been extensively applied across AEC scenarios ranging from Mechanical, Electrical, and Plumbing (MEP) systems to bridges (Wang et al., 2025). However, their reliance on massive, labor-intensive annotations severely limits scalability.

To mitigate data costs, weakly supervised and few-shot paradigms were developed, utilizing sparse labels or prototype alignment. Yet, these approaches often compromise accuracy and remain constrained to specific target domains.

Ultimately, these traditional methods share a critical limitation: they are "closed-set" systems. They lack the open-vocabulary reasoning required to interpret the diverse, unstructured elements found in dynamic construction sites without task-specific retraining. This bottleneck necessitates a paradigm shift toward large-scale foundation models, which liberate segmentation from fixed label sets and enable generalized, zero-shot 3D understanding.

Vision-Language Models (VLMs)

VLMs have revolutionized perception, shifting from closed-set tasks to open-vocabulary learning. While early models like CLIP (Radford et al., 2021) focused on global alignment, recent advancements like Grounding DINO (Liu et al., 2024) and SAM 3 (Carion et al., 2025) integrate semantic reasoning with precise localization. To extend this to 3D, researchers employ 2D-to-3D projection, mapping either segmentation masks or feature embeddings from multi-view images onto point clouds (Wang et al., 2024). These methods have shown promise in AEC applications ranging from indoor scanning to infrastructure inspection.

However, a significant bottleneck persists: VLMs remain inherently 2D-centric. They struggle with complex 3D construction environment, where single-view occlusion and background clutter hinder accurate detection. This limitation suggests a synergistic opportunity: leveraging 3DGS to render simplified, optimal viewpoints, thereby enabling VLMs to progressively analyze the environment and derive accurate segmentation results.

3D Gaussian Splatting (3DGS)

3D scene reconstruction has transitioned from discrete point clouds, which suffer from sparsity and artifacts, to continuous Neural Radiance Fields (NeRF) (Mildenhall et al., 2021). While NeRF achieves photorealism, its implicit nature incurs high computational costs. 3DGS bridges this gap by representing scenes with explicit, anisotropic 3D Gaussians (Kerbl et al., 2023).

Technically, 3DGS optimizes Gaussian parameters (position, opacity, covariance) via differentiable rasterization, projecting 3D ellipsoids into 2D image space for real-time rendering. Its adaptive density control dynamically densifies or prunes primitives to capture fine details efficiently. Subsequent advancements like Scaffold-GS (Lu et al., 2024) have further enhanced geometric consistency and surface alignment.

Despite these strides in visual fidelity, a critical gap remains: existing research prioritizes visual reproduction over structural interpretation. There is a distinct lack of methodologies leveraging 3DGS to semantically decompose and interpret complex civil engineering components, leaving this frontier largely unexplored.

Methodology

To address the challenges of segmenting complex 3D structures in cluttered environments, we propose a novel 3D segmentation agent. As illustrated in Figure 1, our framework establishes a dynamic synergy between the semantic reasoning capabilities of VLMs and the flexible rendering power of 3DGS. Unlike traditional static pipelines that process a fixed set of views, our agent mimics human visual inspection, treating 3D segmentation as an iterative, active perception process.

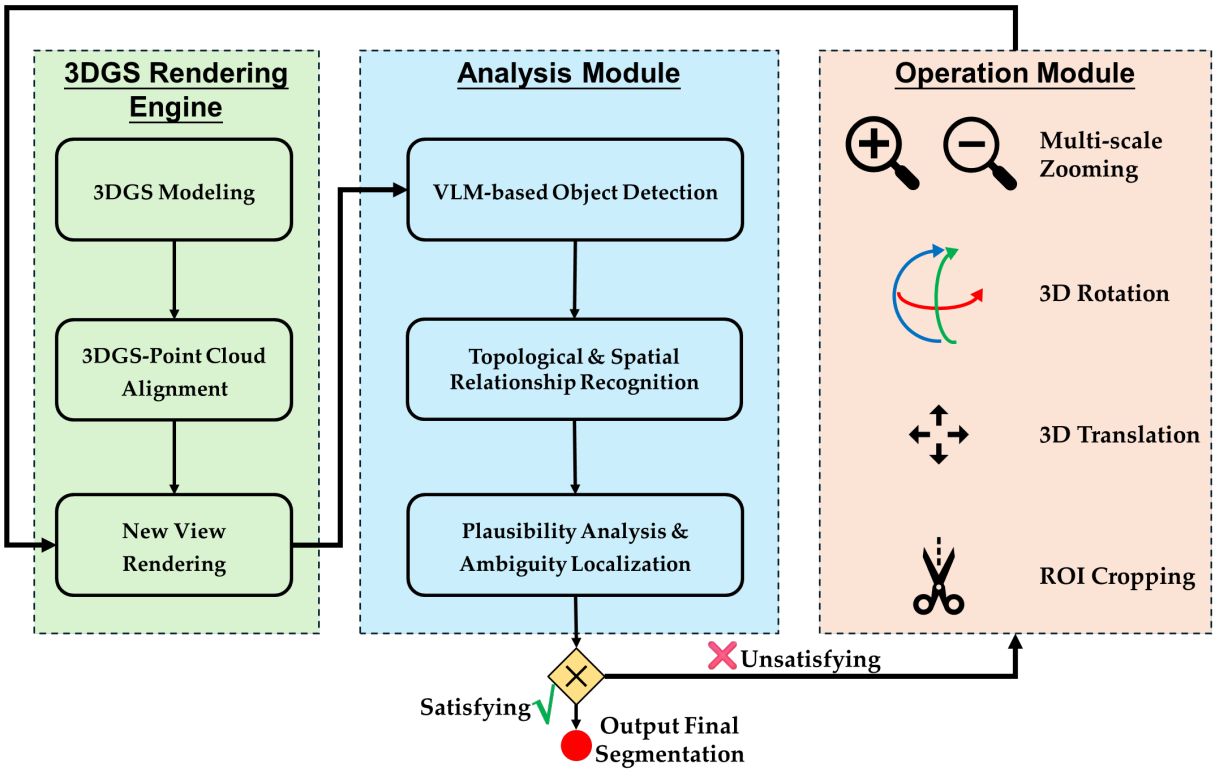

Figure 1: Overall framework of the proposed 3D segmentation agent system

The framework is structured around three interconnected components: the 3DGS Rendering Engine, the Analysis Module, and the Operation Module, forming a closed-loop active perception system:

- 3DGS Rendering Engine (Environment): This module constructs the scene via 3DGS Modeling and Point Cloud Alignment. It generates high-fidelity New View Renderings, providing the visual input necessary for semantic analysis.
- Analysis Module (Cognitive Core): Acting as the decision-maker, this module employs VLMs for object detection and topological relationship recognition. Crucially, it performs plausibility analysis to evaluate results. If the output is ambiguous, the system triggers the Operation Module.
- Operation Module (Active Control): To resolve identified ambiguities, this module executes targeted adjustments such as multi-scale zooming, 3D rotation, translation, or region of interest (ROI) cropping.

This cycle enables the agent to progressively refine its understanding of the scene until a high-confidence, topologically consistent segmentation is achieved. The following sections detail the implementation of these core components.

3DGS Rendering Engine

The 3DGS Rendering Engine acts as the bridge between the physical world and the agent's cognitive processes. Unlike traditional pipelines that rely solely on 2D images or sparse point clouds, our engine constructs a dual-representation environment that synchronizes photorealistic rendering with metric-accurate 3D geometry.

Data Acquisition and 3DGS Modeling: To capture complex as-built conditions, we utilize a handheld mobile scanning system integrating LiDAR and high-resolution RGB cameras. The system leverages Fast-LIVO2 (Zheng et al., 2024) for real-time SLAM, generating a dense point cloud with accurate global pose estimation. Simultaneously, we employ 3DGS to model the scene as a continuous radiance field. Initialized via sparse Structure-from-Motion (SfM) points, the scene is parameterized by volumetric 3D Gaussians, optimized iteratively to minimize photometric loss. This results in a representation that offers real-time, photorealistic rendering from arbitrary viewpoints, essential for the agent's active exploration.

Multi-modal Alignment and 2D-3D Projection: A critical feature of our engine is the rigorous alignment between the photorealistic 3DGS model and the metric-accurate SLAM point cloud. Since raw 3DGS reconstructions lack absolute scale, we register the two representations using a coarse-to-fine approach (anchor point estimation followed by Iterative Closest Point refinement). An example of 3DGS-Point Cloud alignment is shown in Figure 3.

This alignment establishes a deterministic mapping between the 2D rendered view and 3D space. By recording the extrinsic and intrinsic parameters of the 3DGS virtual camera for every rendered frame, we enable a precise pixel-to-point projection mechanism: Once the Analysis Module generates a 2D semantic mask, the engine could use the stored camera parameters to back-project mask pixels onto the aligned 3D point cloud. This mechanism allows the system to extract the exact spatial coordinates of semantic entities, effectively lifting 2D perception into 3D metric space for topological reasoning.

Analysis Module

The Analysis Module serves as the cognitive core of our agent, responsible for transforming raw visual data into a structured semantic understanding of the 3D environment. This process is divided into three progressive stages: VLM-based object detection, topological relationship recognition, plausibility analysis and ambiguous region identification.

VLM-based Object Detection and 3D Projection: To achieve robust segmentation without relying on pre-trained, domain-specific datasets, we adopt a "segment-then-recognize" strategy that integrates foundational 2D models with the reasoning power of Large VLMs.

The workflow begins with the 3DGS engine rendering an initial "initial look" image from a viewpoint as shown in the upper left corner of Figure 2. We utilize the Segment Anything Model (SAM) (Kirillov et al., 2023) to generate a set of high-quality, class-agnostic binary masks, where each mask corresponds to a distinct visual entity (e.g., a pipe segment, a wall patch) as shown in the upper right corner of Figure 2.

These binary masks, overlaid on the original RGB image, are fed into the VLM GPT-5.2. The VLM acts as a semantic parser, analyzing the visual context to assign specific semantic labels (e.g., "Water Pipe," "Concrete Wall," "Valve") to each mask ID as shown in the lower left corner of Figure 2. This step converts purely geometric segments into semantically meaningful objects.

To bridge the gap between 2D image space and 3D volumetric space, we perform a reverse projection. Using the known camera intrinsics and extrinsics from the 3DGS rendering process, each pixel within a 2D mask is back-projected to identify the corresponding 3D point clouds. This effectively "lifts" the 2D semantic labels onto the 3D point cloud, resulting in a semantically annotated 3D scene representation.

Topological and Spatial Relationship Recognition: Mere object detection is insufficient for complex engineering scenes; the agent must understand how objects interact. Once the 3D entities are labeled, the module extracts geometric primitives and infers spatial relationships.

The VLM analyzes these 2D visual features and 3D geometric features to construct a topological scene graph as shown in the lower right corner of Figure 2. This graph encodes connectivity and spatial adjacency, such as:

- Connectivity: Pipe 1 connects to Pipe 2 at Joint J1

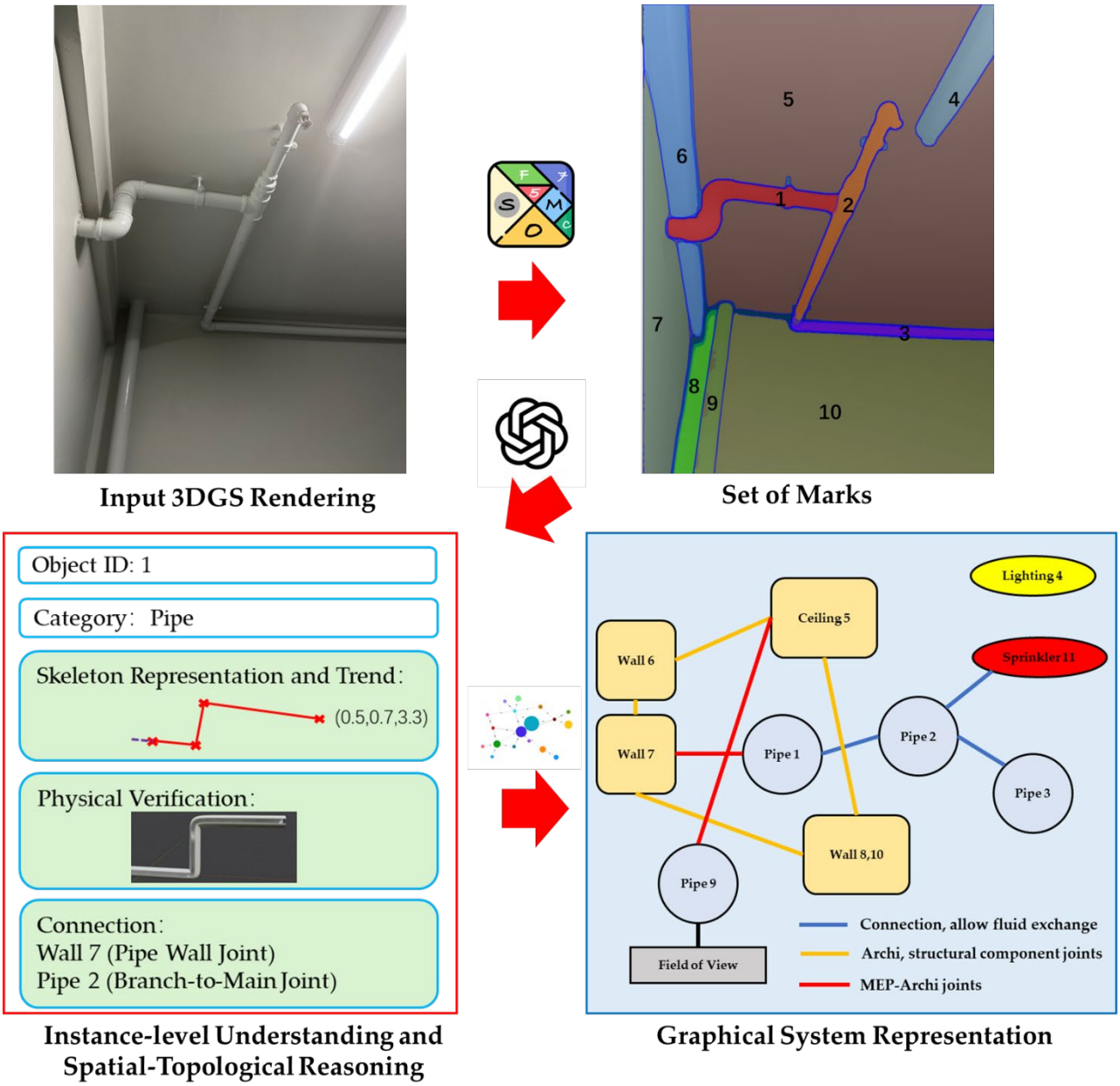

Figure 2: Analysis Module: VLM-based object detection and topological scene graph construction

- Attachment: Pipe 1 is mounted on Wall 7.
- Termination: Pipe 1 ends at coordinate (x, y, z)

Plausibility Analysis and Ambiguity Localization: With a structured scene graph established, the Analysis Module performs a "sanity check" based on physical priors. This is where the agent identifies ambiguities that require active investigation.

As illustrated in upper right corner of Figure 2, the initial segmentation successfully identifies the pipe network ("Pipe 1", "Pipe 2" and "Pipe 3") and the surrounding walls. Following this initial detection, the system evaluates the scene's topology using a hybrid verification mechanism that integrates geometric priors, structured prompting, and VLM-based visual judgment. Since the semantic labels and spatial coordinates (both 2D and 3D) of all detected components are known, the system first applies geometric priors (e.g., spatial proximity) to filter potential relationship candidates.

Applying this mechanism to the network, the topological analysis reveals a logical inconsistency regarding "Pipe 2". The system observes the skeleton of Pipe 2 and detects that it terminates abruptly in mid-air. To investigate this, a targeted logical verification chain is executed. For instance, during the Continuity Check, the system geometrically determines that "Pipe 1" is a viable candidate for connection. A targeted prompt then instructs the VLM to visually verify this specific connection within the 2D rendered image. Thus, while the verification logic is structurally constrained by geometric priors, the final determination relies on the VLM's visual reasoning. For the unattached end of Pipe 2, the VLM yields the following results:

- Continuity Check: Does the pipe connect to another segment? (Result: Negative)

- Boundary Check: Does the pipe extend outside the camera view? (Result: *Negative*)
- Structural Check: Does the pipe terminate into a wall surface? (Result: *Negative*)

According to established engineering priors, a water pipe cannot simply end in free space without a cap or a functional component. Consequently, the Analysis Module flags this region as having "Low Plausibility," hypothesizing that a small terminal device, likely a sprinkler head exists at this location but was missed due to the low resolution of the initial global view. This deduction seamlessly triggers the Operation Module, issuing a precise command: “Zoom in on the endpoint of Pipe 2 to verify the potential missing terminal component,” thereby initiating the active perception loop.

Crucially, this explicit, step-by-step verification chain provides a level of explainability absent in traditional end-to-end models. By structurally constraining the verification logic with geometric priors and outputting human-readable hypotheses (e.g., flagging a specific pipe termination as violating engineering norms), every decision made by the agent, from identifying an anomaly to actively moving the virtual camera, is fully traceable and transparent.

Operation Module

While the Analysis Module identifies what is missing or ambiguous, the Operation Module determines how to look closer. Acting as the agent's navigational controller, this module dynamically manipulates the virtual camera within the 3DGS environment to acquire supplementary views. The process is driven by a feedback loop that continues until the scene representation achieves both geometric high-fidelity and topological consistency. The Operation Module is triggered when the Analysis Module flags a region as "Ambiguous." To ensure rigorous execution, ambiguity is formally defined by two criteria: 1) Quantitative visual uncertainty, occurring when the SAM-predicted mask confidence score falls below 0.7 or the VLM semantic classification probability falls below 0.9; and 2) Qualitative topological inconsistency, occurring when a component fails the logical verification chain (e.g., a pipe exhibiting a "dangling edge" in the scene graph without connecting to a verified boundary or terminal). Upon detecting either condition, the Operation Module triggers a multi-scale zooming and rotation strategy.

- Focusing: For small or indistinct features (such as the potential sprinkler head identified in the last section), the module calculates a new camera pose that minimizes the distance to the target coordinate while maintaining the optical axis centered on the frame. This "zoom-in" operation increases the pixel density of the target, allowing the VLM to discern fine-grained details previously lost in the initial view.
- De-occlusion: If an object is suspected to be occluded, the module executes a 3D rotation around the target's centroid. By generating views from oblique angles, the system bypasses obstacles to reveal the hidden geometry.

Spatial Continuity and Boundary Extension: A common challenge in large-scale engineering scenes is that linear structures (like pipes or ducts) often extend beyond the current field of view (FoV).

- Truncation Detection: If a segmented mask touches the boundary of the rendered image frame, the system recognizes this as an incomplete observation.
- Trajectory Extension: The Operation Module computes the directional vector of the truncated object in 3D space. It then performs a 3D Translation (panning) along this vector, shifting the virtual camera to capture the subsequent segment of the structure. This ensures that long-span objects are tracked continuously across multiple frames rather than being fragmented into disjointed parts.

Active Perception Loop and Termination Criteria: The core mechanism of this module is an iterative Active Perception Loop. The system does not settle for a single pass; it continuously refines its understanding based on specific termination criteria. The loop cycles through Rendering Analysis Operation until the following three conditions are simultaneously met:

- Geometric Completeness: All segmented objects within the target block possess high-confidence mask scores (≥ 0.7), indicating clear boundaries and minimal visual noise.
- Topological Consistency: The scene graph is logically sound. There are strictly zero unexplained "dangling" edges; every segment must either connect to another component, terminate at a verified surface, or end with a recognized terminal device.
- Global Coverage: The union of all recognized entities fully occupies the spatial volume of the target block, ensuring the remaining uninspected void volume falls below a minimal threshold.

The loop iterates until all three conditions are simultaneously met, signifying convergence to a high-fidelity semantic model. To prevent infinite loops in cases of severe physical occlusion or missing scan data, the system enforces a maximum iteration limit (e.g., $N_{max} = 5$) per ambiguous target, after which the unresolved region is explicitly flagged for human review.

Experiment

To evaluate the effectiveness of the proposed pipeline, experiments were conducted on a real-world public toilet captured via mobile laser scanning. The evaluation focused on the object detection accuracy of architectural and functional components, assessed by comparing the predicted labels to manual annotations. The tested scene serves as a representative real-world case intended to validate the feasibility of the proposed pipeline, rather than to provide a comprehensive performance benchmark. Broader validation across multiple sites and building configurations is left for future work.

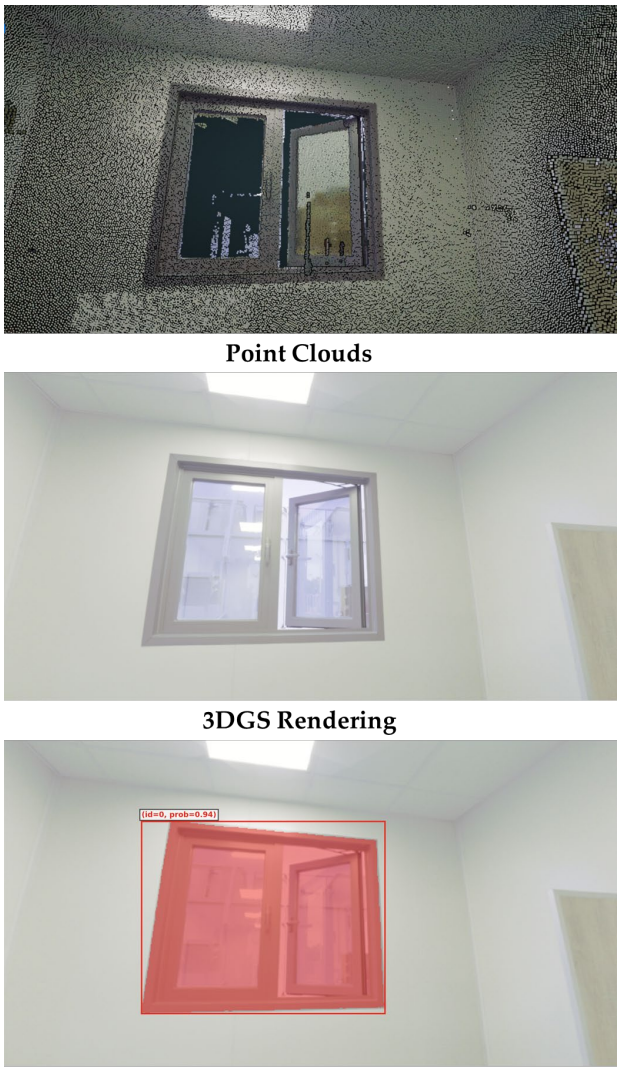

Figure 3: Alignment between 3DGS and point clouds from Rendering Engine and VLM-based object detection results in the experiment

Figure 3 shows a step-by-step breakdown of a specific detection instance (a window). The process relies on the precise alignment between the raw point cloud and the 3DGS model. The top row displays the raw point cloud view. It could be observed the laser scan data is sparse and noisy, lacking sufficient geometric definition for traditional shape-based detectors to reliably identify the object. The middle row shows the corresponding view rendered by the 3DGS. By leveraging the continuous volumetric representation, 3DGS effectively fills in the missing visual information, recovering the window frame structure and the reflective properties of the glass with photorealistic quality. The bottom row demonstrates the final detection result. The VLM performs inference on the rendered image. Because the visual features are clearly restored, the model successfully detects the target "window" with a high confidence score of 0.94, generating an accurate bounding box and segmentation mask. This example confirms that the proposed 3DGS rendering engine allows the system to overcome the sparsity inherent in point clouds, enabling accurate semantic recognition in complex scenarios.

Table 1: Results of public toilet component detection

Component Type	True Positive	False Positive	False Negative
Door	5	0	1
Window	4	0	0
Vent	5	0	0
Lighting	16	0	0
Urinal	2	0	0
Plate	11	0	0
Basin	6	0	0

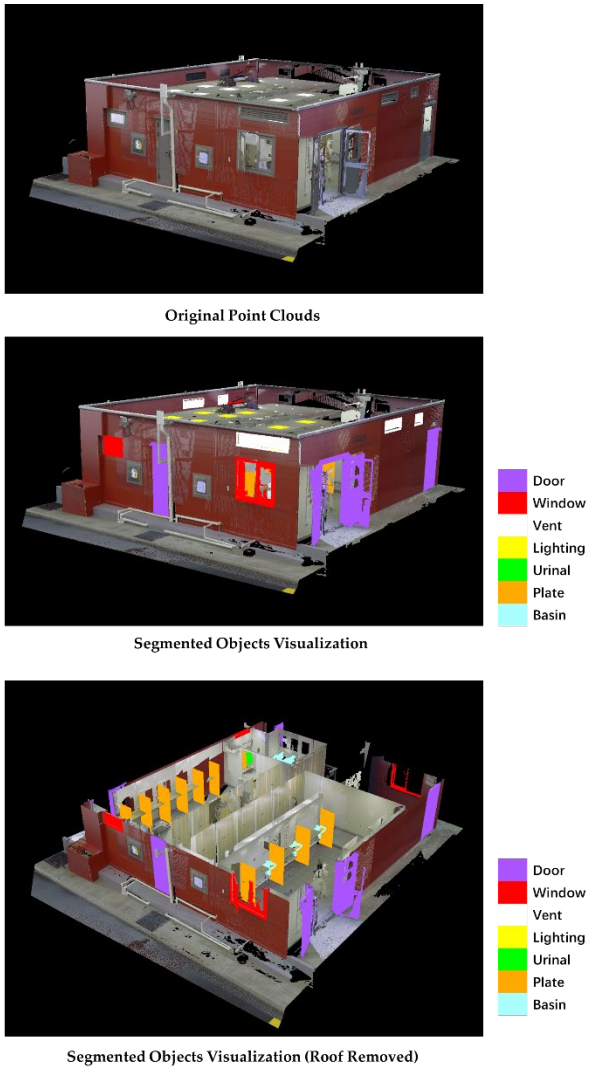

Figure 4: 3D segmentation results performed by the proposed agent

As shown in Table 1, A total of 50 components were annotated and used as ground truth. The results show that the proposed method achieves near-perfect precision

across almost all categories. Notably, repetitive functional elements such as Lights (16 instances), Partition Plates (11 instances), and Basins (6 instances) were detected with 100% precision and recall. The only detection error occurred in the Door category, where one instance was missing (5 out of 6 detected). This specific false negative was attributed to severe data incompleteness; the missing door was in a transition zone that was not fully covered by the scan path, resulting in a low-quality 3DGS for the VLM to form a valid semantic mask, even after the active operation module attempted to refine the view. However, the successful detection of most of the components confirms that the integration of 3DGS modeling and VLM-based segmentation enables high-quality detection with strong generalization across diverse indoor assets. The 3D segmentation results is shown in Figure 4.

Conclusions

This paper presents a novel 3D segmentation agent for as-built analysis by integrating 3DGS and VLMs. Through the synergy of a high-fidelity 3DGS rendering engine and an interactive operation module, the system dynamically adjusts observation perspectives via multi-scale zooming, 3D rotation, and translation. This active perception capability enables the analysis module to effectively perform VLM-based object detection, recognize complex topological and spatial relationships, and conduct plausibility analysis to localize and resolve ambiguities in the scene. Consequently, the system provides a highly interpretable and traceable decision-making process, successfully overcoming the 'black-box' limitations of conventional deep learning approaches.

Experimental results demonstrate the agent's robustness in interpreting complex as-built environments, achieving high accuracy in 3D segmentation and semantic reasoning. The approach successfully bridges the gap between geometric reconstruction and semantic understanding. However, some limitations remain, such as the automation level of the operation module is limited and the testing scenario is not complex enough. The efficiency of the proposed agent in large-scale scenes needs further validation. Furthermore, while our 3DGS-based re-rendering approach theoretically overcomes the limitations of standard 2D-to-3D projection methods, which rely on static raw images and struggle with physical occlusions, by enabling dynamic novel view synthesis, quantitative comparisons are currently absent. As part of our future work to expand this research into a comprehensive journal-length manuscript, we plan to conduct extensive quantitative ablation studies. These studies will systematically compare our active roaming framework against standard 2D-to-3D projection baselines to rigorously demonstrate the performance gains in segmentation accuracy and topological consistency. Overall, the proposed approach lays the foundation for intelligent as-built data analysis, providing a reliable and automated solution for as-built inspection and semantic enrichment.

Acknowledgments

This research was partially supported by the Center for Sand Hazards and Opportunities for Resilience, Energy, and Sustainability (SHORES), funded by Tamkeen under the NYUAD Research Institute Award CG013.

References

Carion, N., Gustafson, L., Hu, Y. T., Debnath, S., Hu, R., Suris, D., ... & Feichtenhofer, C. (2025). Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719.

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. ACM Trans. Graph., 42(4), 139-1.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4015-4026).

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2024, September). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European conference on computer vision (pp. 38-55). Cham: Springer Nature Switzerland.

Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., & Dai, B. (2024). Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20654-20664).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

Wang, B., Lin, F., Li, M., Liang, Z., Chen, Z., Wang, M., & Cheng, J. C. (2025). Informative As-Built Modeling as a Foundation for Digital Twins Based on Fine-Grained Object Recognition and Object-Aware Scan-vs-BIM for MEP Scenes. Advanced Engineering Informatics, 65, 103382.

Wang, B., Chen, Z., Li, M., Wang, Q., Yin, C., & Cheng, J. C. (2024). Omni-Scan2BIM: A ready-to-use Scan2BIM approach based on vision foundation models for MEP scenes. Automation in Construction, 162, 105384.

Zheng, C., Xu, W., Zou, Z., Hua, T., Yuan, C., He, D., ... & Zhang, F. (2024). Fast-livo2: Fast, direct lidar-inertial-visual odometry. IEEE Transactions on Robotics.