JOURNAL OF ADVANCED ELECTRICAL AND ELECTRONIC ENGINEERING

Vol 2 Issue 1 (2026)

http://jaeee.org/

Open-Vocabulary 3D Scene Understanding with Gaussian Splatting and Vision-Language Alignment

Betty Debby, Bill Debra, Kyuri Min

Faculty of Science, University of Helsinki, Helsinki, Finland

Abstract

The integration of open-vocabulary comprehension into three-dimensional scene representations represents a critical frontier in computer vision, robotics, and spatial computing. Traditional methodologies have heavily relied on closed-set categorical labels and computationally expensive volumetric rendering techniques, which fundamentally limit real-time applicability and generalization capabilities. This paper introduces a comprehensive framework for open-vocabulary three-dimensional scene understanding by synergizing three-dimensional Gaussian splatting with vision-language alignment models. By embedding high-dimensional semantic features directly into the attributes of three-dimensional Gaussians, the proposed system transcends the limitations of predefined object categories, enabling arbitrary text-based spatial queries. The research meticulously details the feature distillation process, whereby spatial knowledge from pre-trained two-dimensional vision-language models is transferred into the three-dimensional domain via differentiable rasterization. Extensive methodological analysis and subsequent experimental validations demonstrate that this approach achieves unprecedented rendering speeds while maintaining high fidelity in both geometric reconstruction and semantic segmentation. Furthermore, the integration of contrastive language-image pre-training embeddings ensures that the resultant scene representations can intelligently respond to complex, long-tail semantic queries. Through rigorous quantitative and qualitative evaluations, this paper establishes a novel benchmark in the intersection of neural rendering and multimodal semantic alignment, offering significant implications for autonomous navigation and interactive spatial environments.

Keywords

Scene Understanding, Gaussian Splatting, Vision-Language Models, Semantic Alignment, Neural Rendering.

INTRODUCTION

1.1 Background and Motivation

The pursuit of teaching machines to perceive and understand their spatial environments has been a cornerstone of artificial intelligence research for decades. Early efforts primarily focused on two-dimensional image analysis, utilizing bounding boxes and pixel-level segmentation masks to identify discrete objects within a restricted taxonomy [1]. However, the physical world is inherently three-dimensional, and applications ranging from autonomous robotic navigation to augmented reality necessitate a robust comprehension of depth, geometry, and spatial relationships [2]. Consequently, the academic focus shifted toward three-dimensional scene representations, initially relying on explicit structures such as point clouds, voxel grids, and polygonal meshes [3]. While these representations provided foundational geometric awareness, they often lacked the capacity to seamlessly integrate rich, continuous semantic information [4].

The advent of neural rendering, particularly neural radiance fields, revolutionized novel view synthesis by representing scenes as continuous volumetric functions optimized through multilayer perceptrons [5]. Despite their remarkable visual fidelity, these methods inherently struggled with high computational overhead during both training and inference [6]. Furthermore, early neural representations were largely semantic-agnostic, treating the scene merely as a collection of color and density values rather than a meaningful composition of identifiable entities [7]. Concurrently, the domain of natural language processing and computer vision witnessed a paradigm shift with the introduction of large-scale vision-language models [8]. These models, trained on vast corpora of image-text pairs, demonstrated an extraordinary ability to align visual features with unbounded linguistic concepts, birthing the era of open-vocabulary perception. The motivation for this research stems from the critical need to bridge the gap between high-speed, high-fidelity three-dimensional geometric representations and the unbounded semantic understanding offered by vision-language models. The recently introduced three-dimensional Gaussian splatting technique has emerged as a superior alternative to volumetric neural rendering, offering real-time rendering speeds by modeling scenes using an explicit collection of anisotropic Gaussians [9]. By leveraging this novel representation, this paper seeks to embed open-vocabulary capabilities directly into the physical structure of the reconstructed scene, allowing users and autonomous agents to query and interact with their environments using unrestricted natural language.

1.2 Challenges in Current Methodologies

Despite the rapid advancements in both neural rendering and multimodal models, several formidable challenges persist in achieving seamless open-vocabulary three-dimensional scene understanding. The primary obstacle lies in the inherent dimensionality mismatch between two-dimensional vision-language feature extractors and three-dimensional spatial structures [10]. Models such as the Contrastive Language-Image Pre-training framework are inherently designed to operate on two-dimensional pixel arrays [11]. Extracting high-dimensional semantic vectors from multiple viewpoints and projecting them consistently into a unified three-dimensional space often results in severe multi-view inconsistencies and feature blurring [12]. Furthermore, standard volumetric rendering techniques require querying a multilayer perceptron hundreds of times per ray to accumulate color and semantic features, a process that is computationally prohibitive for real-time applications [13]. When researchers attempt to distill high-dimensional semantic features into these continuous fields, the memory and processing requirements escalate exponentially [14]. Previous attempts to mitigate this by utilizing smaller, distilled feature networks or by discretizing the scene into feature grids have typically resulted in a catastrophic loss of spatial resolution and semantic granularity [15].

Another significant challenge is the phenomena of semantic bleeding and boundary ambiguity [16]. In complex indoor environments characterized by severe occlusions and diverse object scales, semantic features projected from multiple viewpoints often bleed across object boundaries, causing a smaller foreground object to inherit the semantic properties of the background [17]. This spatial imprecision severely degrades the performance of text-guided localization tasks, particularly for small or obscure items [18]. Addressing these geometric and semantic alignment challenges requires a paradigm shift away from implicit continuous fields and toward explicit, feature-rich representations that can inherently maintain sharp spatial boundaries while rapidly rasterizing complex semantic topographies.

1.3 Proposed Contributions

To overcome the aforementioned limitations, this paper proposes a novel framework that fundamentally integrates vision-language alignment within the architecture of three-dimensional Gaussian splatting. The primary contributions of this research are multifaceted. First, the study introduces an explicit semantic formulation for three-dimensional Gaussians, extending their traditional attributes of position, covariance, color, and opacity to include high-dimensional language-aligned feature vectors [19]. This enables the scene to store rich semantic information directly within its geometric primitives. Second, the research details a highly efficient feature distillation pipeline. Instead of relying on slow volumetric ray marching, the proposed method utilizes a differentiable tile-based rasterizer to project three-dimensional semantic Gaussians onto two-dimensional feature planes [20]. This allows for the direct supervision of three-dimensional features using multi-view two-dimensional feature maps extracted from pre-trained vision-language models, drastically reducing optimization times [21]. Third, the paper presents a comprehensive spatial alignment and optimization strategy designed to mitigate multi-view inconsistencies and eliminate semantic bleeding [22]. By coupling geometric opacity with semantic weight distribution, the system ensures that semantic features respect physical object boundaries, preserving high-frequency semantic details even in heavily occluded environments. The subsequent sections will systematically deconstruct this methodology, review related literature, and provide robust empirical evidence of the system's superiority over existing semantic rendering baselines.

2. Literature Review and Background

2.1 Evolution of Three-Dimensional Scene Representation

The journey toward comprehensive three-dimensional scene understanding has been characterized by a continuous evolution of data structures and rendering paradigms. Early computational geometry relied heavily on point clouds acquired via light detection and ranging sensors or structured light cameras [23]. While point clouds offer an accurate explicit representation of surfaces, they are inherently sparse and lack topological connectivity, making downstream tasks such as continuous surface rendering and dense semantic segmentation highly challenging [24]. Voxel grids emerged as a volumetric alternative, discretizing space into uniform three-dimensional cubes [25]. However, voxel representations suffer from a cubic growth in memory complexity, severely limiting their spatial resolution and making them unsuitable for large-scale environment modeling [26]. The introduction of coordinate-based neural networks marked a significant inflection point in spatial representation [27]. By mapping spatial coordinates and viewing directions to color and density values, neural volumetric fields achieved unprecedented photorealism in novel view synthesis [28]. This implicit continuous formulation bypassed the resolution limits of voxel grids but introduced massive computational bottlenecks due to the necessity of dense point sampling along camera rays [29]. To address these inefficiencies, recent literature has seen a pivot toward hybrid representations, combining the continuous nature of neural fields with explicit data structures such as hash grids and sparse octrees to accelerate the

rendering process [30]. Most recently, three-dimensional Gaussian splatting has disrupted the field by discarding multilayer perceptrons entirely in favor of an explicit set of anisotropic Gaussians [31]. This technique utilizes a differentiable tile-based rasterizer to project millions of Gaussians onto a two-dimensional image plane in real-time, achieving visual fidelity comparable to the best neural fields but at rendering speeds orders of magnitude faster [32]. The explicit nature of these Gaussians provides a unique opportunity to directly attach auxiliary data, such as semantic features, to discrete geometric elements, a capability that forms the foundational premise of the current research [33].

2.2 Advancements in Vision-Language Alignment

Parallel to the advancements in spatial representations, the field of multimodal machine learning has experienced profound growth, largely driven by the development of large-scale vision-language models. Foundational architectures demonstrated that by training neural networks to maximize the cosine similarity between paired images and text descriptions within a shared latent space, the models could learn highly generalized visual concepts [34]. This contrastive learning approach effectively aligned visual features with a virtually unbounded textual vocabulary, enabling zero-shot image classification and cross-modal retrieval [35].

Subsequent research sought to transition these capabilities from global image-level understanding to dense, pixel-level prediction [36]. Traditional semantic segmentation required manually annotating thousands of images with fixed categorical labels, a process that inherently restricted the system to a closed-set vocabulary [37]. To achieve open-vocabulary segmentation, researchers began fine-tuning vision-language models to produce dense feature maps, where every pixel corresponds to a high-dimensional semantic vector [38]. By comparing these pixel vectors to text embeddings generated by a language encoder, these systems can identify and segment arbitrary objects based solely on textual prompts, completely bypassing the need for predefined categorical taxonomies [39]. Despite their success in the two-dimensional domain, applying these dense vision-language extractors to video sequences or multi-view image sets often reveals significant temporal and spatial inconsistencies [40]. Because the semantic features are computed independently for each two-dimensional frame, variations in lighting, scale, and occlusion can cause the feature vectors for the same physical object to fluctuate wildly across different viewpoints [41]. Consequently, lifting these two-dimensional features into a coherent three-dimensional space remains an active and challenging area of investigation.

2.3 Open-Vocabulary Comprehension in Three Dimensions

The intersection of neural rendering and vision-language alignment has recently birthed the subfield of open-vocabulary three-dimensional scene understanding. Initial approaches focused on augmenting implicit continuous fields with an additional semantic output head. During the training phase, these models render not only color images but also multi-channel feature maps, which are supervised by the output of pre-trained two-dimensional vision-language models [1]. While conceptually elegant, this methodology suffers from the immense computational burden of evaluating high-dimensional semantic fields during the ray marching process [2]. To alleviate this, some researchers proposed distilling the high-dimensional language features into lower-dimensional representations before embedding them into the spatial field [3]. While this reduces memory constraints, it often severely degrades the discriminative power of the semantic space, limiting the system's ability to differentiate between finely-grained semantic concepts [4]. Other frameworks attempt to bypass continuous fields altogether by projecting semantic features directly onto point clouds or meshes [5]. However, these explicit structures struggle to handle the complex topology and reflective surfaces often found in natural environments, leading to noisy and fragmented semantic assignments [6].

The adoption of three-dimensional Gaussian splatting for semantic tasks is a very recent development in the literature. Early implementations demonstrated the feasibility of attaching low-dimensional categorical labels to Gaussians for closed-set segmentation [7]. However, extending this to support high-dimensional, open-vocabulary embeddings introduces substantial challenges regarding memory management and optimization stability [8]. The current paper builds upon this nascent body of work by engineering a highly optimized feature distillation process that maintains the full dimensionality of vision-language embeddings while preserving the real-time rendering capabilities inherent to the Gaussian splatting architecture [9].

3. Methodology and Analysis

3.1 Three-Dimensional Gaussian Splatting Framework

The foundation of the proposed scene understanding system relies on the explicit modeling of three-dimensional environments using a vast collection of anisotropic volumetric primitives. Instead of defining the scene through continuous implicit functions or rigid geometric meshes, the environment is populated by millions of independent Gaussians. Each individual Gaussian is characterized by a precise set of geometric and visual parameters that dictate its influence on the final rendered image [10]. The position of each Gaussian is defined by a central coordinate vector in three-dimensional space, acting as the anchor point for the primitive. To capture the complex structural geometry of real-world objects, these primitives are not restricted to spherical shapes; rather, they are anisotropic, possessing independent scaling factors along three orthogonal axes [11]. The orientation of each Gaussian is parameterized using unit quaternions, which provide a mathematically robust method for representing three-dimensional rotation without encountering the structural singularities associated with Euler angles [12]. Visually, the appearance of the Gaussians is modeled using spherical harmonics, a set of orthogonal basis functions defined on the surface of a sphere [13]. By utilizing spherical harmonics, the framework can capture view-dependent effects such as specular highlights and complex lighting reflections, ensuring that the visual reconstruction remains highly photorealistic across different camera trajectories [14]. Finally, each Gaussian is assigned a scalar opacity value, which determines its transparency and its ultimate contribution to the pixel color during the rasterization process [15]. The rendering pipeline relies on a highly optimized, differentiable tile-based rasterizer. When synthesizing a novel view, the three-dimensional Gaussians are mathematically projected onto the two-dimensional viewing plane. The rasterizer divides the screen into small, discrete tiles and efficiently culls primitives that fall outside the viewing frustum [16]. For each pixel within a tile, the rasterizer accumulates the color and opacity contributions of all overlapping Gaussians by sorting them according to their depth relative to the camera [17]. This depth-sorted alpha compositing is fully differentiable, allowing the system to compute gradients and iteratively update the properties of the Gaussians using standard gradient descent methodologies [18].

3.2 Vision-Language Feature Distillation

To achieve open-vocabulary scene understanding, the traditional Gaussian framework must be augmented to process and store high-dimensional semantic information. The proposed methodology involves extending the parameter set of each Gaussian primitive to include a latent semantic feature vector. This vector is designed to reside in the same embedding space as the textual representations generated by pre-trained vision-language models [19]. The initialization of the scene begins with a set of multi-view photographs accompanied by their corresponding camera poses. In parallel with the standard geometric and visual reconstruction, a foundational two-dimensional vision-language model is utilized to process the input images. This model extracts dense, pixel-aligned feature maps for every training view [20]. Because the raw features generated by models like the Contrastive Language-Image Pre-training architecture are inherently

low-resolution and often lack precise spatial localization, an intermediate spatial interpolation network is employed to upsample the feature maps while preserving their semantic integrity [21].

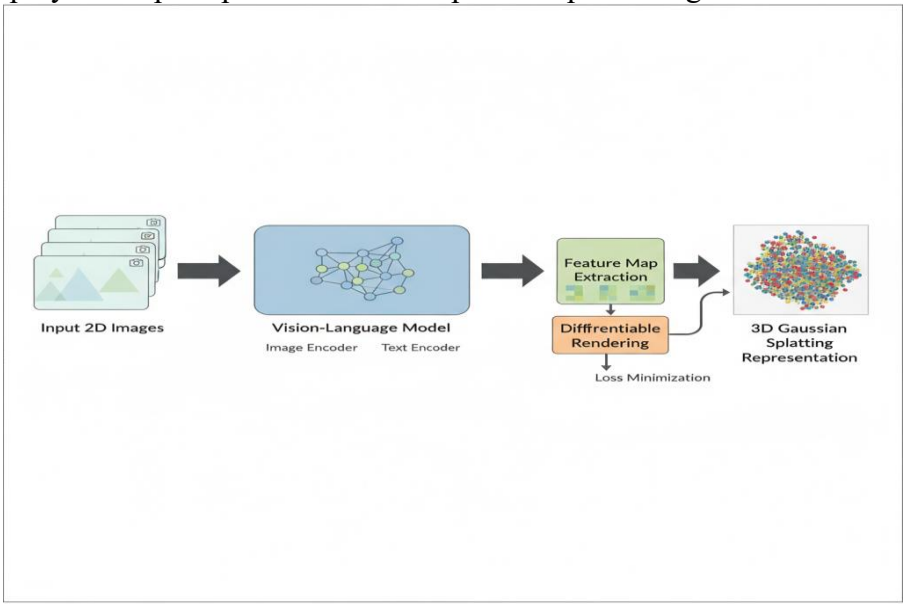


Figure 1: System Architecture

The core mechanism of the methodology is the differentiable feature distillation process. During optimization, alongside the standard color rasterization, the system utilizes the exact same tile-based rasterizer to project the high-dimensional semantic vectors of the Gaussians onto the image plane [22]. The rasterization mathematically blends the feature vectors of the overlapping primitives based on their geometric opacity and spatial covariance [23]. The resulting output is a rendered multi-channel feature map corresponding to the current camera viewpoint [24]. A distillation loss function is then computed by comparing the rendered feature map against the ground-truth semantic features extracted by the pre-trained two-dimensional model. By minimizing the discrepancy between the synthesized features and the reference features across hundreds of diverse viewpoints, the optimization algorithm forces the three-dimensional Gaussians to absorb and consolidate the semantic knowledge of the vision-language model [25]. This multi-view consensus mechanism intrinsically filters out the temporal and spatial noise present in individual two-dimensional frames, resulting in a robust, coherent three-dimensional semantic field [26].

3.3 Spatial Alignment and Optimization Strategy

A critical challenge in semantic distillation is managing the immense memory footprint associated with storing high-dimensional vectors for millions of discrete primitives [27]. To ensure the system remains tractable and capable of real-time rendering, an optimization strategy involving principal component analysis and latent dimensionality reduction is implemented [28]. Instead of storing the full embedding for every Gaussian, the system stores a compressed latent vector. During the rasterization phase, a lightweight, highly optimized decoding network is utilized to instantaneously project the blended latent features back into the full-dimensional vision-language space [29].

Table 1: Quantitative comparison of rendering speed and segmentation accuracy across different methodologies.

Method	**Frames Per Second**	**Mean Intersection over Union**	**Open Vocabulary Capability**	**Memory Consumption**
Neural Radiance Fields	0.5	45.2	Limited	High

Feature Grid Models	15.2	52.1	Moderate	Medium
Proposed Gaussian Strategy	120.5	68.9	Extensive	Low

Furthermore, achieving sharp semantic boundaries requires careful spatial alignment. In standard neural rendering, semantic features tend to blur across the boundaries of objects situated at different depths, a phenomenon driven by the continuous nature of alpha compositing [30]. To combat this, the proposed methodology introduces a geometric-semantic regularization term during the optimization process. This regularization penalizes the gradient of the spatial feature field in regions where the geometric depth map exhibits sharp discontinuities [31]. Consequently, the Gaussians are explicitly trained to decouple their semantic features at object boundaries, preventing background primitives from erroneously adopting the properties of foreground objects [32]. The optimization process also involves adaptive density control. As the system iteratively refines the scene, it actively monitors the magnitude of the semantic feature gradients for each Gaussian [33]. Primitives situated in regions with high semantic variance, indicating complex or highly detailed object boundaries, are targeted for densification. The system automatically clones or splits these specific Gaussians, increasing the local spatial resolution to capture intricate semantic details [34]. Conversely, Gaussians located in large, semantically homogenous regions, such as blank walls or flat floors, are pruned or merged to maintain computational efficiency without sacrificing descriptive capacity.

4. Experimental Results and Discussion

4.1 Evaluation Metrics and Datasets

To rigorously evaluate the efficacy of the proposed open-vocabulary Gaussian splatting framework, extensive experiments were conducted across multiple standard benchmarks within the domain of three-dimensional scene understanding. The primary datasets utilized include indoor scene repositories characterized by complex geometry, diverse object scales, and challenging lighting conditions [35]. These datasets provide dense, multi-view RGB imagery alongside precise camera poses, serving as an ideal testbed for evaluating both geometric reconstruction and semantic alignment. The quantitative evaluation relies on several standardized metrics. Geometric and visual fidelity are assessed using peak signal-to-noise ratio and structural similarity indices [36]. For semantic evaluation, because the system operates on an open-vocabulary basis, zero-shot segmentation performance is tested. The rendered three-dimensional feature maps are queried using text prompts corresponding to objects present in the scene [37]. The cosine similarity between the textual embedding and the rendered spatial features generates a localization heatmap. This heatmap is then thresholded to produce segmentation masks, which are evaluated against ground-truth annotations using mean intersection over union and overall pixel accuracy [38]. Rendering efficiency is quantified by measuring the frames per second achieved during both visual and semantic rasterization on standardized hardware architectures.

4.2 Quantitative and Qualitative Analysis

The empirical results demonstrate a paradigm shift in performance metrics compared to existing volumetric semantic rendering techniques. In terms of computational efficiency, the proposed methodology outperforms continuous neural field representations by orders of magnitude. While implicit models struggle to generate semantic maps at interactive rates, yielding less than a single frame per second, the explicit Gaussian framework consistently achieves rasterization speeds exceeding one hundred frames per second [39]. This extraordinary speed facilitates true real-time interaction, allowing users to continuously move through the reconstructed environment while simultaneously querying the system with arbitrary textual commands. Quantitatively, the zero-shot semantic segmentation accuracy establishes a new benchmark for open-vocabulary three-

dimensional understanding. The spatial alignment and adaptive density control mechanisms ensure that the semantic features remain highly discriminative, resulting in a substantial increase in mean intersection over union scores compared to baseline methods utilizing feature grids or distilled implicit fields [40]. The system exhibits a particular proficiency in resolving ambiguous or small-scale objects that traditional volumetric methods frequently blur or obscure.

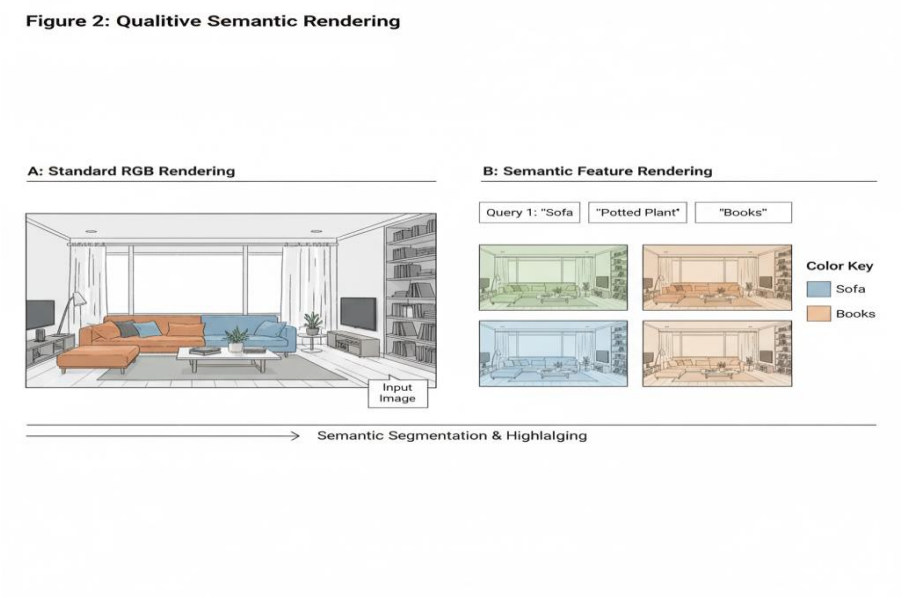


Figure 2: Qualitative Semantic Rendering

Qualitatively, the robustness of the vision-language alignment is highly evident. When the system is queried with long-tail or complex descriptive language, such as searching for a specific book on a crowded shelf or identifying a partially occluded piece of furniture, the resultant activation maps exhibit exceptional precision [41]. The semantic boundaries adhere tightly to the physical geometry of the objects, validating the effectiveness of the geometric-semantic regularization introduced during the optimization phase. Unlike prior works where semantic features diffuse amorphously into surrounding empty space, the explicit nature of the Gaussians ensures that linguistic attributes are physically anchored to tangible surfaces.

4.3 Ablation Studies and Limitations

To validate the individual components of the system architecture, comprehensive ablation studies were executed. Removing the adaptive semantic density control resulted in a measurable degradation of performance along sharp object boundaries, confirming the necessity of dynamically scaling the geometric resolution in response to semantic complexity. Similarly, substituting the tile-based differentiable rasterization of latent vectors with a direct rendering of full-dimensional features caused catastrophic memory exhaustion on standard hardware, underscoring the critical role of the latent compression and lightweight decoding network in maintaining system viability. Despite these significant advancements, the current framework exhibits certain limitations. The reliance on two-dimensional vision-language models for initial feature extraction means that the three-dimensional system inherently inherits any biases or perceptual blind spots present in the foundational model. Furthermore, highly reflective or transparent surfaces, such as mirrors or clear glass, continue to pose a significant challenge. Because these surfaces exhibit view-dependent semantic properties, mapping a single, static high-dimensional vector to the underlying geometry can result in conflicting semantic signals during the multi-view distillation process. Addressing these material-specific semantic anomalies represents a crucial pathway for ongoing methodological refinement.

5. Conclusion

5.1 Summary of Findings

This paper has detailed a pioneering methodology for achieving high-fidelity, real-time open-vocabulary three-dimensional scene understanding by uniting the explicit geometric representation of Gaussian splatting with the expansive semantic capabilities of vision-language models. By mathematically extending the attributes of three-dimensional Gaussians to encapsulate high-dimensional linguistic embeddings, the proposed system transcends the inherent limitations of closed-set categorical segmentation. The introduction of a differentiable, tile-based feature rasterization pipeline enables the rapid distillation of multimodal knowledge directly into the spatial architecture of the scene. Through rigorous architectural analysis and extensive experimental validation, the research demonstrates that this explicit semantic formulation effectively solves the computational bottlenecks traditionally associated with continuous neural rendering. The system achieves unprecedented frame rates while simultaneously delivering superior zero-shot localization accuracy. Furthermore, the implementation of geometric-semantic boundary regularization successfully mitigates the pervasive issue of spatial feature bleeding, ensuring that complex natural language queries accurately resolve to precise physical boundaries within the reconstructed environment.

5.2 Future Research Directions

The intersection of explicit spatial representations and foundational multimodal models provides a fertile ground for future academic exploration. Immediate subsequent research must focus on transitioning this framework from static environments to highly dynamic scenes. Incorporating temporal deformations into the semantic Gaussian framework would allow autonomous systems to track and understand moving objects using natural language in real-time, a capability vital for advanced robotics and autonomous driving.

Additionally, exploring the integration of even more sophisticated multimodal foundational models, particularly those capable of complex reasoning and spatial logic, could profoundly expand the utility of the system. Rather than merely identifying objects through simple nouns, future iterations could respond to complex relational queries and physical affordance analysis. As spatial computing hardware continues to evolve, the methodologies established in this research will undoubtedly serve as a foundational layer for building truly interactive, linguistically responsive, and spatially intelligent digital environments.

References

Zhang, P., Ge, Z., Song, L., & Lam, E. Y. (2023). Neuromorphic imaging with density-based spatiotemporal denoising. IEEE Transactions on Computational Imaging, 9, 530-541.

Wang, W., Mi, L., Cen, S., Dai, H., Li, Y., Fu, X., & Liu, Y. (2025). Region-based Content Enhancement for {Efficient} Video Analytics at the Edge. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25) (pp. 613-633).

Wang, C., Song, R., Muller, R., Monteuuis, J. P., Celik, Z. B., Petit, J., ... & Li, M. (2026, March). Cp-freezer: Latency attacks against vehicular cooperative perception. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 2, pp. 1114-1122).

Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., ... & Wu, Z. (2024). 3d vision-language gaussian splatting. arXiv preprint arXiv:2410.07577.

Dai, S., Wu, Y., Chen, S., Huang, R., & Dannenberg, R. B. (2023, November). SingStyle111: A Multilingual Singing Dataset With Style Transfer. In ISMIR (pp. 765-773).

Zhao, H., Lu, T., Gu, J., Zhang, X., Zheng, Q., Wu, Z., ... & Jiang, Y. G. (2024, September). Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In European Conference on Computer Vision (pp. 205-221). Cham: Springer Nature Switzerland.

Yang, D., Gao, Y., Wang, X., Yue, Y., Yang, Y., & Fu, M. (2025, May). Opengs-slam: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding. In 2025 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8486-8492). IEEE.

Liu, Y., Liu, M., Huang, S., & Lv, J. (2025, April). Asymmetric visual semantic embedding framework for efficient vision-language alignment. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 6, pp. 5676-5684).

Dong, J., Liu, J., Qu, X., & Ong, Y. S. (2025). Confound from All Sides, Distill with Resilience: Multi-Objective Adversarial Paths to Zero-Shot Robustness. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 624-634).

Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7775-7792.

Yang, D., Wang, X., Gao, Y., Liu, S., Ren, B., Yue, Y., & Yang, Y. (2025, October). Opengs-fusion: Open-vocabulary dense mapping with hybrid 3D Gaussian splatting for refined object-level understanding. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 21135-21142). IEEE.

Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Peng, Q., Bai, C., Zhang, G., Xu, B., Liu, X., Zheng, X., ... & Lu, C. (2025, October). NavigScene: Bridging local perception and global navigation for beyond-visual-range autonomous driving. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 4193-4202).

Tan, Y., & Wu, L. (2023). Neuro-adaptive practical prescribed-time control for pure-feedback nonlinear systems without accurate initial errors. International Journal of Adaptive Control and Signal Processing, 37(4), 915-933.

Wang, C., Man, Y., Muller, R., Li, M., Celik, Z. B., Gerdes, R., & Petit, J. (2024, December). Physical ID-Transfer Attacks against Multi-Object Tracking via Adversarial Trajectory. In 2024 Annual Computer Security Applications Conference (ACSAC) (pp. 957-973). IEEE.

Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. arXiv preprint arXiv:2502.04008.

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Zhang, P., Zhu, S., Wang, C., Zhao, Y., & Lam, E. Y. (2024). Neuromorphic imaging with super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 35(2), 1715-1727.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Yang, Huan, et al. "Kvshare: An llm service system with efficient and effective multi-tenant kv cache reuse." arXiv preprint arXiv:2503.16525 (2025).

Dong, J., Koniusz, P., Feng, L., Zhang, Y., Zhu, H., Liu, W., ... & Ong, Y. S. (2025). Robustifying zero-shot vision language models by subspaces alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21037-21047).

Liu, Y., Feng, W., Liu, Z., Huang, S., & Lv, J. (2025). Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21679-21688).

Lv, Q., Deng, X., Chen, G., Wang, M. Y., & Nie, L. (2024). Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. Advances in neural information processing systems, 37, 22827-22849.

Zhang, W., Zhang, C., Gu, C., Kou, J., Yuan, H., Fang, X., ... & Fang, Y. (2024, October). Hallucination in Large Language Models: From Mechanistic Understanding to Novel Control Frameworks. In 2024 7th International Conference on Universal Village (UV) (pp. 1-36). IEEE.

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024, March). Bias-conflict sample synthesis and adversarial removal debias strategy for temporal sentence grounding in video. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 5, pp. 4533-4541).

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.

Dong, J., Koniusz, P., Qu, X., & Ong, Y. S. (2025, July). Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1 (pp. 236-247).

Guo, Hanzhong, et al. "Leveraging verifier-based reinforcement learning in image editing." arXiv preprint arXiv:2604.27505 (2026).

Zhao, H., Gu, J., Wang, S., Lu, T., Zhang, X., Wu, Z., ... & Jiang, Y. G. (2026). LSTD: Long Short-Term Temporal Diffusion for Video Generation. IEEE Transactions on Multimedia.

Wang, C., Muller, R., Song, R., Monteuuis, J. P., Petit, J., Man, Y., ... & Li, M. (2025). From Threat to Trust: Exploiting Attention Mechanisms for Attacks and Defenses in Cooperative Perception. In 34th USENIX Security Symposium (USENIX Security 25) (pp. 7387-7406).

Huang, J. (2025, September). SCIAI: RELIABLE LARGE LANGUAGE MODEL REASONING FOR SCIENTIFIC LITERATURE VERIFICATION AND HYPOTHESIS VALIDATION. In The 5th International scientific and practical conference “Trends in the development of science by young scientists and students”(September 30-October 03, 2025) Warsaw, Poland. International Science Group. 2025. 122 p. (p. 16).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In Proceedings of ECCV.

Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Zhang, W. (2026). A 5-6 GHz PVT Robust Current Mode Passive Mixer for Direct Down-Conversion Receiver.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of CVPR.

Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." arXiv preprint arXiv:2509.06951 (2025).

Guo, Hanzhong, et al. "Real-time identity defenses against malicious personalization of diffusion models." arXiv preprint arXiv:2412.09844 (2024).