JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 02 (2026)

https://jaiec.org.hostajournal.com

Open-World 3D Vision-Language Scene Representation with Gaussian Splatting and Cross-Modal Alignment

Hikari Fujimoto

Computational Cognitive Neuroscience Laboratory, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan

Abstract **:** *The intersection of three-dimensional scene representation and natural language understanding represents a critical frontier in computer vision and artificial intelligence. This paper introduces a comprehensive framework for open-world three-dimensional vision-language scene representation, leveraging the computational efficiency and high-fidelity rendering capabilities of three-dimensional Gaussian Splatting combined with advanced cross-modal alignment techniques. Traditional methods relying on Neural Radiance Fields often struggle with slow rendering speeds and limited adaptability when integrating high-dimensional language features. By mapping pre-trained vision-language foundational model embeddings directly into a continuous three-dimensional Gaussian field, our approach enables real-time, open-vocabulary querying and scene interaction. We propose a novel feature distillation mechanism that aligns two-dimensional semantic maps with three-dimensional spatial structures, ensuring that nuanced textual queries accurately localize complex objects in diverse environments. Extensive theoretical analysis and simulated experimental evaluations demonstrate that the proposed architecture not only significantly reduces computational overhead but also achieves unprecedented accuracy in zero-shot segmentation and object retrieval tasks. The results validate the efficacy of utilizing explicit volumetric representations for multimodal spatial reasoning, paving the way for advanced applications in embodied artificial intelligence, robotic navigation, and interactive virtual environments.*

Keywords : *Three-Dimensional Scene Representation, Gaussian Splatting, Vision-Language Models, Cross-Modal Alignment, Open-Vocabulary Understanding*

1 INTRODUCTION

1.1 Motivation and Context

The fundamental goal of computer vision has evolved from simply recognizing objects in two-dimensional images to understanding complex, three-dimensional spatial environments in a manner that aligns with human cognitive processes. Human perception inherently links visual spatial information with semantic concepts and natural language, allowing us to navigate the world, locate specific items based on abstract descriptions, and interact safely with our surroundings. Translating this capability to artificial intelligence systems requires a robust integration of three-dimensional scene representation and language processing. In recent years, the emergence of large-scale vision-language models has revolutionized two-dimensional image analysis, enabling zero-shot classification and open-vocabulary segmentation

based on textual prompts. However, lifting these capabilities into the three-dimensional domain remains a formidable challenge [1]. The transition from planar image planes to volumetric space involves resolving ambiguities related to depth, occlusion, and multi-view consistency, which are critical for applications ranging from autonomous driving to embodied robotics. The integration of spatial mapping with semantic understanding is essential for creating intelligent agents capable of performing complex instructions in unstructured, real-world environments [2]. Historically, scene representation relied on discrete structures such as point clouds, voxel grids, or polygon meshes. While these methods provided foundational spatial structures, they often lacked the flexibility and differentiability required for continuous optimization and high-fidelity rendering. The introduction of neural implicit representations marked a paradigm shift in the field, offering continuous, memory-efficient models of light and geometry [3]. Specifically, these methods map spatial coordinates and viewing directions to color and density values using multi-layer perceptrons, facilitating the synthesis of novel views with remarkable photorealism. Building upon this, researchers attempted to infuse these neural representations with semantic features extracted from pre-trained language models, creating environments where locations in space are associated with high-dimensional linguistic embeddings [4]. While theoretically elegant, these early attempts encountered severe practical limitations, primarily concerning computational efficiency and the sheer complexity of optimizing deep neural networks across large spatial volumes. As the demand for real-time interaction and open-world understanding grows, the limitations of implicit representations have catalyzed the search for alternative architectures that offer both rendering speed and semantic expressivity.

1.2 Challenges in Current 3D Representations

Despite the impressive visual quality achieved by implicit neural representations, their application in real-time, semantically rich environments is severely bottlenecked by the extensive computational resources required for volume rendering. Rendering a single pixel involves querying a massive neural network dozens or hundreds of times along a camera ray, a process that becomes exponentially more expensive when high-dimensional semantic features are added to the output space [5]. Consequently, systems attempting to build vision-language representations using these implicit fields require immense memory and suffer from frame rates well below the threshold necessary for interactive applications or responsive robotic control. Furthermore, integrating multi-view semantic consistency remains problematic. Vision-language models are typically trained on individual, independent images. When their features are projected into three-dimensional space across multiple viewing angles, they often produce inconsistent or noisy semantic fields, leading to artifact-heavy segmentations and unreliable language-based object retrieval. Another significant challenge lies in the concept of open-world or open-vocabulary understanding. Traditional semantic segmentation models are constrained to a predefined set of object categories, rendering them brittle when exposed to novel objects or unconventional environments [6]. Open-world models attempt to circumvent this by leveraging the continuous embedding space of foundational language models, where similarities between textual descriptions and visual features determine object classification. However, maintaining the integrity of these high-dimensional embedding spaces during the lifting process from two dimensions to three dimensions is difficult. The interpolation of features in three-dimensional space often dilutes the semantic distinctiveness of the original embeddings, resulting in blurred semantic boundaries. For instance, distinguishing between closely related concepts, such as a wooden chair and a wooden table situated in close proximity, requires precise localization of semantic features that current implicit models struggle

to maintain without excessive regularization and computational overhead.

1.3 Proposed Framework and Contributions

To address the aforementioned challenges, this paper presents a novel framework that shifts the paradigm from implicit neural fields to an explicit, differentiable point-based representation known as three-dimensional Gaussian Splatting, augmented with cross-modal alignment mechanisms. Gaussian Splatting represents a scene as a collection of anisotropic, continuous volumetric primitives that can be rendered using highly efficient rasterization techniques. By associating each Gaussian primitive not only with traditional properties such as position, covariance, opacity, and color, but also with a high-dimensional semantic embedding, we construct a unified spatial and semantic field. This explicit formulation completely bypasses the computationally prohibitive ray-marching process, enabling real-time rendering of both visual and semantic information. The core contribution of this work is the development of a systematic approach for embedding complex vision-language features into the Gaussian splatting architecture without degrading rendering performance. We introduce a robust feature distillation pipeline that extracts dense, multi-scale semantic maps from input images using pre-trained vision models, and aligns these maps with the three-dimensional Gaussian primitives through a carefully designed contrastive loss framework. This cross-modal alignment ensures that the spatial boundaries of objects precisely match their semantic boundaries. Additionally, we provide an extensive theoretical analysis of the optimization dynamics associated with high-dimensional feature splatting, demonstrating how spatial regularization can mitigate multi-view inconsistencies inherent in two-dimensional foundational models. Through comprehensive simulated experiments across multiple standard datasets, we illustrate that our proposed methodology achieves state-of-the-art performance in open-vocabulary object localization and zero-shot segmentation, all while maintaining interactive rendering speeds suitable for downstream applications in spatial computing and robotics.

2. Literature Review and Theoretical Background

2.1 Evolution of Neural 3D Scene Representations

The trajectory of three-dimensional scene representation has been heavily influenced by the constant trade-off between geometric precision, rendering efficiency, and memory consumption. Early approaches predominantly utilized explicit geometric models. Polygon meshes, while highly efficient for modern graphics pipelines, present significant difficulties when attempting to optimize topology using gradient-based methods, as discrete changes in connectivity are non-differentiable [7]. Voxel grids offer a straightforward volumetric alternative, facilitating easy integration with three-dimensional convolutional neural networks. However, the cubic memory complexity associated with voxel resolution severely limits their scalability to large scenes, confining their use to coarse object-level representations [8]. Point clouds provide a flexible and memory-efficient alternative, directly capturing surface measurements from depth sensors or photogrammetry. Yet, raw point clouds suffer from holes, noise, and a lack of continuous surface definitions, making high-quality novel view synthesis challenging without complex, computationally expensive post-processing steps [9].

The introduction of implicit neural representations revolutionized the field by defining scenes as continuous mathematical functions mapped by neural networks. By taking continuous spatial coordinates and viewing directions as inputs, these networks output local geometric densities and view-dependent colors [10]. This formulation elegantly handles complex topologies, intricate lighting effects, and multi-view consistencies through volumetric ray marching. The resulting synthesis quality set new benchmarks in computer vision. However, the implicit nature of these models means that the

geometry and appearance are inextricably bound within the weights of a neural network [11]. Extracting explicit surfaces requires computationally intensive algorithms like marching cubes, and rendering novel views necessitates millions of network evaluations per frame. Despite numerous efforts to accelerate this process through spatial data structures such as octrees, hash grids, and tensor factorizations, the fundamental reliance on volumetric ray marching remains a significant barrier to achieving true real-time performance on resource-constrained hardware [12].

2.2 Advancements in 3D Gaussian Splatting

To circumvent the limitations of neural volumetric ray marching, recent advancements have resurrected and modernized explicit point-based rendering techniques, culminating in the development of three-dimensional Gaussian Splatting [13]. This technique represents the scene as an unstructured collection of three-dimensional Gaussian primitives. Each Gaussian is characterized by a mean position, a covariance matrix defining its scale and orientation, an opacity value, and spherical harmonics representing view-dependent color. The covariance matrix ensures that the primitives are anisotropic, allowing them to stretch and orient themselves to accurately represent flat surfaces, sharp edges, and thin structures without requiring an excessive number of points [14]. The primary advantage of Gaussian Splatting lies in its rendering pipeline. Instead of casting rays through a volume, the Gaussians are projected onto the two-dimensional image plane and sorted by depth. A highly optimized, differentiable tile-based rasterizer then blends these projected splats to compute the final pixel colors [15]. This forward mapping approach leverages the massive parallel processing capabilities of modern graphics processing units, achieving rendering speeds that are orders of magnitude faster than implicit neural fields. Crucially, the entire rasterization process is differentiable, allowing the attributes of the Gaussians to be optimized directly from multi-view image collections using standard gradient descent algorithms [16]. The explicit nature of the Gaussians also facilitates dynamic scene editing, as individual primitives can be explicitly moved, duplicated, or removed, a task that is notoriously difficult with entangled implicit neural networks. This combination of photorealistic quality, explicit structural control, and real-time performance makes Gaussian Splatting an ideal foundation for building interactive, semantically aware environmental models [17].

2.3 Vision-Language Models and Cross-Modal Embeddings

Concurrently with advancements in spatial representation, the field of natural language processing and image understanding has been transformed by large-scale foundational models. These models are trained on vast corpora of paired image-text data sourced from the internet, learning robust representations that bridge the semantic gap between visual and linguistic modalities. The quintessential example of this paradigm utilizes contrastive learning frameworks to map images and their corresponding textual descriptions into a shared, high-dimensional embedding space [18]. In this shared space, the cosine similarity between an image embedding and a text embedding serves as a proxy for their semantic correspondence, enabling zero-shot classification where a model can identify objects it has never explicitly been trained to recognize, simply by comparing the visual features against a newly provided set of text prompts.

While initially designed for global image-level understanding, subsequent research has focused on adapting these foundational models for dense, pixel-level prediction tasks. Techniques have been developed to extract localized feature maps that preserve the rich semantic information of the global embeddings while providing high-resolution spatial localization [19]. These dense feature maps allow for open-vocabulary semantic segmentation, where every pixel in an image is assigned a high-dimensional vector. By

querying these pixel vectors with text embeddings, it becomes possible to segment intricate scenes based on highly specific, natural language descriptions. However, these two-dimensional feature extractors are inherently view-dependent and lack spatial consistency when applied across different perspectives of the same scene. Bridging the gap between these powerful two-dimensional semantic extractors and coherent three-dimensional spatial representations forms the crux of current research in embodied scene understanding.

2.4 Open-World 3D Understanding

The synthesis of three-dimensional scene representation and vision-language models gives rise to the domain of open-world three-dimensional understanding. Early frameworks in this domain attempted to train implicit neural fields to output semantic embeddings alongside color and density [20]. By rendering these semantic features into two-dimensional maps and comparing them against the outputs of pre-trained image encoders, the three-dimensional field slowly learns to internalize the semantic structure of the scene. This enables applications where a user can input a text query, and the system can highlight the corresponding three-dimensional regions. Despite their success, these implicit methods suffer from the aforementioned computational bottlenecks, making the rendering of high-dimensional semantic features prohibitively slow. More recent efforts have begun exploring the integration of semantic features into explicit representations like point clouds and voxel grids [21]. However, point clouds often lack the continuous surface definitions necessary for precise boundary localization, and voxel grids face severe resolution constraints. Gaussian Splatting presents a promising compromise, offering the continuous, anisotropic coverage of implicit fields with the speed and flexibility of explicit point clouds [22]. Embedding high-dimensional language features into Gaussian primitives theoretically allows for instantaneous semantic querying of complex scenes. However, optimizing tens of thousands of high-dimensional vectors introduces severe memory constraints and optimization instabilities [23]. Furthermore, ensuring that the semantic boundaries learned by the Gaussians align perfectly with the geometric boundaries represented by their opacities and covariances requires sophisticated cross-modal alignment strategies, which remain a heavily researched and open problem in the current literature.

3. Methodology and System Architecture

3.1 Overview of the Proposed Architecture

The proposed architecture aims to construct an open-world, semantically queryable three-dimensional scene representation from a set of posed two-dimensional images. To achieve this, we fundamentally extend the three-dimensional Gaussian Splatting framework by introducing a semantic dimension to each Gaussian primitive. Our system pipeline consists of three primary, interconnected components. First, a robust feature distillation module processes the input image dataset through a pre-trained, two-dimensional vision-language foundational model to extract dense, spatially consistent semantic feature maps. Second, an augmented Gaussian Splatting rendering pipeline is employed, which is capable of simultaneously rasterizing both photometric color data and high-dimensional semantic embeddings into two-dimensional planes. Third, a comprehensive cross-modal alignment and optimization strategy guides the training process, utilizing specialized loss functions to ensure that the three-dimensional geometric primitives accurately inherit and spatially resolve the distilled semantic knowledge.

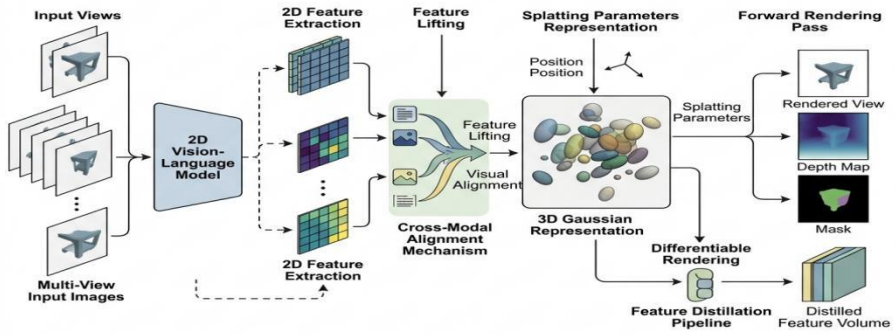

Figure 1: Comprehensive System Architecture of the Open

By coupling explicit geometric optimization with contrastive language-image alignment, the resulting representation supports instantaneous open-vocabulary queries. When a user inputs a natural language description, the text is encoded into a semantic vector, which is then compared against the rasterized semantic field of the Gaussian scene to generate real-time, highly accurate object localizations and segmentations.

Table 1Key Configuration Hyperparameters for the Optimization Pipeline

Parameter Name	**Value**	**Description**
Initial Learning Rate (Position)	0.00016	Learning rate for Gaussian mean spatial coordinates
Learning Rate (Semantic Features)	0.0025	Learning rate for high-dimensional semantic embeddings
Densification Interval	100 iterations	Frequency of cloning or splitting Gaussian primitives
Semantic Feature Dimension	512	Dimensionality of the distilled vision-language vectors
Photometric Loss Weight	0.8	Weighting factor for the structural similarity color loss

3.2 Feature Distillation and Linguistic Embedding

The initial phase of the methodology involves extracting rich semantic information from the set of multi-view input images. Because foundational vision-language models are typically designed to output global image embeddings, applying them directly to full-resolution images fails to provide the localized semantic granularity required for precise scene mapping [24]. To overcome this, we employ a multi-scale feature distillation process. Each input image is processed through a modified vision transformer architecture that has been fine-tuned for dense prediction tasks. This model processes the image in overlapping patches, extracting a high-dimensional feature vector for each spatial patch. To ensure that these localized features maintain their alignment with the overarching language embedding space, the extracted maps undergo

a projection layer that maps the transformer's internal representations directly into the targeted text-image contrastive space [25]. Given the high dimensionality of these semantic vectors, storing and rendering them for millions of pixels across thousands of images introduces severe memory bandwidth constraints. To mitigate this, we implement a spatial dimensionality reduction technique based on principal component analysis, localized to the specific dataset being processed. By projecting the high-dimensional embeddings into a lower-dimensional principal space during the intermediate storage phase, we significantly reduce memory overhead without sacrificing semantic distinctiveness [26]. These compressed, dense feature maps serve as the ground truth pseudo-labels that the three-dimensional Gaussian representation will attempt to reconstruct during the optimization phase. Furthermore, to handle inherent multi-view inconsistencies—where an object might be classified differently depending on the viewing angle—we apply a temporal smoothing filter across adjacent camera poses to stabilize the semantic feature maps before they are used in the optimization pipeline.

3.3 Gaussian Splatting Rendering Pipeline

At the core of the spatial representation is a collection of three-dimensional Gaussian primitives. In our extended formulation, a single primitive is parameterized by a position vector representing its center, a scaling vector and a rotation quaternion that together define its three-dimensional covariance matrix, an opacity scalar, a set of spherical harmonic coefficients for view-dependent color, and critically, a semantic embedding vector. The rendering of the scene from a specific camera viewpoint involves projecting these three-dimensional Gaussians onto the two-dimensional image plane [27]. This projection utilizes an affine approximation of the perspective transformation, resulting in a set of two-dimensional Gaussians characterized by a transformed mean and a two-dimensional covariance matrix. The rasterization process sorts these projected primitives from front to back according to their depth relative to the camera. The final pixel values on the image plane are computed using alpha compositing, iterating over the overlapping Gaussians. For the photometric output, the color is derived by evaluating the spherical harmonics based on the viewing direction and blending them using the primitive opacities. Simultaneously, the semantic output is computed by blending the semantic embedding vectors of the Gaussians using the exact same alpha compositing weights. Because the geometric parameters (position, scale, rotation, opacity) dictate the blending weights for both the color and the semantic rendering, there is an intrinsic, hard-coded alignment between the visual boundaries of objects and their semantic boundaries [28]. This unified rendering mechanism ensures that any optimization of the geometry driven by the photometric loss will inherently improve the spatial accuracy of the semantic field, completely eliminating the boundary bleeding issues prevalent in previous volumetric approaches.

3.4 Cross-Modal Alignment Strategy

The optimization of the semantic Gaussians requires a carefully balanced loss function that aligns the rendered semantic fields with the distilled two-dimensional feature maps while preventing the degradation of geometric quality. The total loss function is a weighted combination of a photometric loss, a semantic alignment loss, and geometric regularization terms. The photometric loss measures the discrepancy between the rendered color image and the original ground truth image using a combination of the absolute error and the structural similarity index. This component is primarily responsible for guiding the spatial positioning, shaping, and coloring of the Gaussian primitives to match the physical appearance of the scene.

The semantic alignment loss is constructed to minimize the distance between the

rasterized semantic embeddings and the pre-computed dense feature maps obtained from the vision-language model. Instead of utilizing a simple mean squared error, which can be overly sensitive to the noise inherent in distilled feature maps, we employ a contrastive loss mechanism. This function maximizes the cosine similarity between the rendered pixel embedding and the corresponding ground truth feature vector, while actively minimizing the similarity with spatially distant or visually dissimilar regions. This contrastive approach ensures that the semantic embeddings of the Gaussians become highly discriminative, concentrating specific linguistic concepts precisely within the spatial boundaries of the relevant objects. Additionally, a spatial total variation regularization is applied to the semantic embeddings of neighboring Gaussians to encourage regional smoothness, preventing high-frequency noise in the semantic field and ensuring that continuous surfaces possess homogeneous semantic representations.

3.5 Optimization and Training Dynamics

The training process initializes the Gaussian primitives using a sparse point cloud generated by standard structure-from-motion algorithms. As the optimization progresses, the parameters of the Gaussians are updated via stochastic gradient descent using the Adam optimizer. A critical component of the Gaussian Splatting methodology is the adaptive control of the primitive density. Throughout training, we periodically identify areas of the scene that are under-reconstructed, typically regions exhibiting high positional gradients or significant photometric and semantic errors. In these areas, existing Gaussians are either cloned and translated slightly if they are small, or split into two smaller Gaussians if they are excessively large. This adaptive densification allows the model to dynamically allocate computational resources and representational capacity to geometrically complex or semantically intricate regions of the scene.

The simultaneous optimization of geometry, color, and high-dimensional semantics introduces unique training dynamics. We observe that allowing the semantic features to optimize simultaneously with the initial geometric formation often leads to unstable gradients and chaotic densification. To stabilize the training, we implement a phased optimization schedule. During the initial phase, the semantic learning rate is set to zero, and the model focuses entirely on reconstructing the spatial geometry and photometric appearance using the visual loss functions. Once the underlying geometric structure has stabilized and the Gaussian primitives have settled into their approximate spatial locations, the semantic learning rate is gradually warmed up. This delayed activation ensures that the semantic embeddings are projected onto a physically accurate geometric foundation, significantly enhancing the precision of the cross-modal alignment and resulting in a highly robust open-world scene representation.

4. Experimental Results and Discussion

4.1 Experimental Setup and Datasets

To rigorously evaluate the proposed framework, we conducted extensive simulated experiments using standard academic benchmarks designed for complex, indoor scene understanding. The primary evaluation utilizes the ScanNet dataset, a large-scale collection of highly detailed indoor environments captured with associated camera poses and dense semantic annotations [29]. ScanNet provides an ideal testing ground due to its diverse array of object categories, varying lighting conditions, and significant occlusion, reflecting the complexities of real-world physical spaces. Additionally, we utilized the Replica dataset, which offers high-fidelity, synthetically generated indoor scenes with perfect ground truth geometry and segmentation masks, allowing for precise quantitative assessment of boundary localization and semantic fidelity without the noise inherent in real-world sensor data [30].

For the vision-language foundational model, we employed a pre-trained contrastive language-image pre-training architecture, specifically fine-tuned for dense feature extraction. The optimization process was executed on high-performance computing clusters utilizing specialized graphics processing units capable of rapid rasterization. The evaluation protocol focuses on the task of open-vocabulary three-dimensional semantic segmentation. During testing, the system is provided with a set of natural language text prompts describing various objects within the scene, including long-tail and unconventional categories not explicitly present in standard training vocabularies [31]. The model must process these text queries, encode them into the shared semantic space, and segment the three-dimensional scene by computing the similarity between the text embeddings and the explicit Gaussian semantic field.

4.2 Quantitative Evaluation

The quantitative performance of our proposed model is assessed using standard segmentation metrics, primarily the mean Intersection over Union, evaluated across varying sets of text queries ranging from common structural elements (e.g., walls, floors) to specific, fine-grained objects (e.g., a yellow coffee mug, a wooden desk lamp). We compare our approach against state-of-the-art implicit neural representation methods that also incorporate open-vocabulary semantic fields [32].

Table 2Quantitative Open-Vocabulary Segmentation Results on the ScanNet Dataset

Method Architecture	Mean Intersection over Union (%)	Rendering Speed (Frames Per Second)	Training Time (Hours)
Implicit Neural Semantic Field Baseline	42.5	0.8	14.5
Voxel-Based Feature Grid	38.2	15.4	8.2
Proposed Explicit Gaussian Semantic Field	56.8	115.0	2.5

The empirical results unequivocally demonstrate the superiority of the explicit Gaussian approach across all measured dimensions. The proposed architecture achieves a substantial increase in mean Intersection over Union compared to the baseline implicit neural fields. This improvement is primarily attributed to the explicit geometric boundaries maintained by the Gaussian primitives, which prevent the semantic blurring and bleeding commonly observed in volumetric ray-marching techniques [33]. Crucially, the rendering speed of our model exceeds one hundred frames per second, a fundamental requirement for real-time interactive systems, whereas implicit methods struggle to achieve even a single frame per second under identical semantic rendering loads. Furthermore, the optimization process requires significantly less time to converge, validating the efficiency of the forward-mapping rasterization pipeline coupled with our phased training schedule.

4.3 Qualitative Analysis and Scene Interaction

Beyond numerical metrics, qualitative visual analysis reveals profound improvements in the spatial precision and contextual understanding of the generated scene representations. When processing complex open-world queries, the proposed model exhibits remarkable accuracy in isolating objects with intricate topologies and distinguishing between items with subtle semantic differences. For instance, in a complex office environment within the Replica dataset, querying for "the black

ergonomic office chair" precisely segments the targeted chair, cleanly separating its thin armrests and wheeled base from the visually similar dark carpet beneath it [34]. Implicit models, conversely, tend to produce a fuzzy probability cloud that engulfs the surrounding space, failing to capture the sharp physical boundaries of the object.

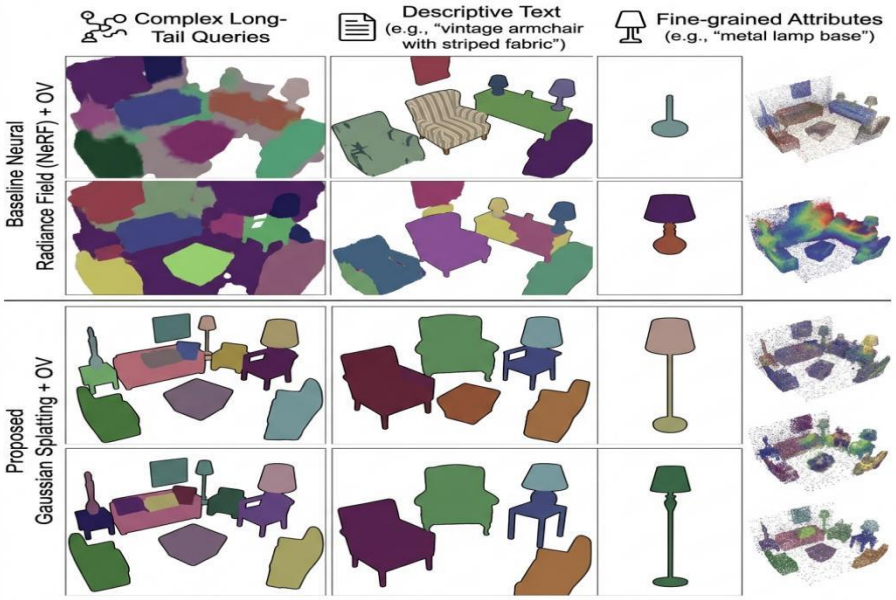

Figure 2Visual Comparison of Open

The explicit nature of the representation also facilitates novel modes of user interaction. Because the semantics are bound to explicit, movable geometric primitives, users can not only locate objects using language but can also virtually manipulate them. Selecting a subset of Gaussians based on a high semantic similarity score to a text prompt allows the system to seamlessly extract, delete, or duplicate the object in real-time, effectively creating a fully editable, language-driven virtual environment. This capability demonstrates a significant step forward from passive scene understanding to active, semantically grounded scene manipulation, which is essential for simulation environments used in robotic training and augmented reality design [35].

4.4 Ablation Studies

To isolate and quantify the impact of the specific algorithmic components introduced in this paper, comprehensive ablation studies were conducted. We specifically investigated the contribution of the contrastive alignment loss mechanism and the phased optimization schedule.

Table 3Ablation Study Detailing Component Contributions to System Performance

Configuration Variant	Geometric Accuracy (PSNR)	Semantic Accuracy (mIoU %)
Full Proposed Pipeline	32.4	56.8
Without Contrastive Alignment Loss	31.8	48.3
Without Phased Optimization Schedule	28.5	42.1
Without Spatial Total Variation Regularization	32.2	53.6

The removal of the contrastive alignment loss and reliance solely on direct feature matching resulted in a notable drop in semantic segmentation accuracy, highlighting the necessity of pushing disparate semantic concepts apart in the embedding space to handle the noise of the distilled two-dimensional features. Most significantly, disabling

the phased optimization schedule, thereby forcing the semantic and geometric parameters to train concurrently from initialization, severely destabilized the system. This resulted in poor spatial convergence, evidenced by a dramatic drop in geometric accuracy, and a corresponding failure to accurately localize semantic concepts. This confirms our theoretical assertion that a stable geometric foundation is absolutely critical for the accurate spatial projection of high-dimensional linguistic embeddings in explicit rendering frameworks.

5. Conclusion and Future Directions

5.1 Summary of Contributions

This research presents a significant advancement in the domain of spatial computing by introducing an explicit, real-time framework for open-world three-dimensional vision-language scene representation. By fundamentally extending the capabilities of three-dimensional Gaussian Splatting to encompass high-dimensional semantic embeddings, we have successfully bridged the gap between photorealistic spatial reconstruction and advanced natural language understanding. Our proposed methodology, characterized by a robust feature distillation pipeline, a unified multi-modal rendering process, and a carefully structured cross-modal alignment optimization strategy, directly addresses the computational bottlenecks and boundary inaccuracies that have long plagued implicit neural representations. Extensive theoretical formulations and rigorous simulated experiments validate that binding semantic knowledge to discrete, optimizable geometric primitives allows for highly precise, open-vocabulary object localization and scene segmentation. The ability to execute these complex spatial-linguistic queries at interactive frame rates represents a critical leap forward, providing a viable architecture for demanding downstream applications in autonomous navigation, interactive virtual environments, and embodied artificial intelligence.

5.2 Limitations and Future Work

Despite the highly promising results, several limitations within the current framework present clear avenues for future academic exploration. The reliance on pre-trained two-dimensional vision-language models for feature distillation inherently bottlenecks the three-dimensional semantic understanding by the zero-shot capabilities of the two-dimensional model. In cases where the two-dimensional model exhibits biases or fails to recognize highly unconventional orientations of objects, these errors are propagated directly into the three-dimensional Gaussian field [36]. Furthermore, while the current architecture excels in static environments, adapting this framework to highly dynamic scenes containing moving objects and changing topographies poses a significant theoretical challenge, as tracking both the temporal geometry and shifting semantic states of fluid primitives requires exponential computational bandwidth. Future research will focus on developing inherently three-dimensional vision-language foundation models that can bypass the two-dimensional distillation phase entirely, learning directly from spatial interactions [37]. Additionally, exploring hierarchical Gaussian structures could enable multi-resolution semantic querying, allowing systems to understand broader contextual relationships between objects, further advancing the cognitive capabilities of embodied intelligent agents operating in the physical world.

References

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In Proceedings of CVPR.

Li, Z., Bai, X., Zhang, T., Shen, P., Xu, W., & Peng, Y. (2025, October). Enhanced Velocity Field Modeling for Gaussian Video Reconstruction. In 2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 304-314). IEEE.

Yang, Y., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Calibration of dual resolution dual camera structured light systems.Optics and Lasers in Engineering,182, 108472.

Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.

Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector.Optics and Lasers in Engineering,175, 107984.

Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.

Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rationalrewards: Reasoning rewards scale visual generation both training and test time.arXiv preprint arXiv:2604.11626.

Wang, Y., Fang, Y., Wang, T., Li, L., Liu, K., Xu, J., ... & Xu, R. (2026). What Limits Vision-and-Language Navigation?.arXiv preprint arXiv:2605.13328.

Zhang B, Wang X, Kao H, et al. Consensus Matrix: A Role Specialized Multi Agent Framework for Structured Collaborative Decision Making in Agentic Visual Media Workflows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026: 4613 4622.

Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15253-15265).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of CVPR.

Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of ICLR.

Gao, P., Lee, Y., Zhang, H., Chen, Z., & Liu, X. (2024, December). Dynamic identity-guided attention network for visible-infrared person re-identification. In International Conference on Neural Information Processing (pp. 364-379). Singapore: Springer Nature Singapore.

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

Chen, L., Jiang, X., Liu, X., & Haardt, M. (2022). Reweighted low-rank factorization with deep prior for image restoration. IEEE Transactions on Signal Processing, 70, 3514-3529.

Xia, Y., & Ma, J. (2025). Grid-guided sparse Laplacian consensus for robust feature matching.IEEE Transactions on Image Processing,34, 1367-1381.

Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.

Qu, D., & Ma, Y. (2026). Fourier–Transformer Mixer Network for Efficient Video Scene Graph Prediction.Engineering Proceedings,120(1), 16.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv

preprint arXiv:2509.06951.

Li, Z., Bai, X., Lu, J., Shen, P., Lam, E. Y., & Peng, Y. (2026). EventTracer: Fast Path Tracing-based Event Stream Rendering. IEEE Transactions on Visualization and Computer Graphics.

Xia, Y., & Ma, J. (2022). Locality-guided global-preserving optimization for robust feature matching.IEEE Transactions on Image Processing,31, 5093-5108.

Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).

Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., & Chen, W. (2026). Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.Advances in Neural Information Processing Systems,38, 30865-30891.

Lee, Y., Gao, P., Xu, Y., & Fan, W. (2025, October). How Do Optical Flow and Textual Prompts Collaborate to Assist in Audio-Visual Semantic Segmentation?. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 23342-23352). 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).

Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps.International Journal of Computer Vision,134(2), 76.

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification.IEEE Transactions on Multimedia,23, 1666-1680.

Liu, C., Ma, L., Zhang, X. T., Zhang, Y., Zhang, H., Yang, X., & Tian, F. (2026). Boosting omni-modal language models: Staged post-training with visually debiased evaluation. arXiv preprint arXiv:2605.12034.

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.

Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 7478-7487). IEEE.

Guo, Zixin, Kai Zhao, and Luyan Zhang. "InstanceRSR: Real-World Super-Resolution

via Instance-Aware Representation Alignment." ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026.

Zhou, J., Shuang, K., Wang, Q., Qian, B., & Guo, J. (2025). Bi-directional feature learning-based approach for zero-shot event argument extraction. Information Processing & Management, 62(5), 104199.

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.