2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Octree Diffusion for Semantic Scene Generation and Completion

Xujia Zhang, Brendan Crowe, and Christoffer Heckman

Abstract— **The completion, extension, and generation of 3D semantic scenes are an interrelated set of capabilities that are useful for robotic navigation and exploration. Existing approaches seek to decouple these problems and solve them one-off. Additionally, these approaches are often domain-specific, requiring separate models for different data distributions, e.g. indoor vs. outdoor scenes.**

To unify these techniques and provide cross-domain compatibility, we develop a single framework that can perform scene completion, extension, and generation in both indoor and outdoor scenes, which we term Octree Latent Semantic Diffusion. Our approach operates directly on an efficient dual octree graph latent representation: a hierarchical, sparse, and memory-efficient occupancy structure. This technique disentangles synthesis into two stages: (i) structure diffusion, which predicts binary split signals to construct a coarse occupancy octree, and (ii) latent semantic diffusion, which generates semantic embeddings decoded by a graph VAE into voxel-level semantic labels. To perform semantic scene completion or extension, our model leverages inference-time latent inpainting, or outpainting respectively. These inference-time methods use partial LiDAR scans or maps to condition generation, without the need for retraining or finetuning. We demonstrate high-quality structure, coherent semantics, and robust completion from single LiDAR scans, as well as zero-shot generalization to out-of-distribution LiDAR data. These results indicate that completion-through-generation in a dual octree graph latent space is a practical and scalable alternative to regression-based pipelines for real-world robotic perception tasks.

Code is publicly available on GitHub.

I. Introduction

Completion and generation of semantically rich 3D scenes is frequently required in robotics[17], autonomous driving[15], and AR/VR applications[9]. Embodied agents must reason over observed and occluded geomtries, in order to facilitate safe navigation and long-term planning. To this end, scene generation and complete are keys tools for enhancing world understanding. Each domain poses a different data distribution challenge. In outdoor environments, LiDAR sensors provide accurate but sparse geometry; in indoor settings, RGB-D sensors and LiDAR offer partial yet incomplete coverage. In addition, outdoor models must generalize across varying scales, from compact rooms to large urban scenes, while remaining efficient enough for deployment in real-world systems.

While semantic scene completion (SSC) [5, 14, 26, 22] attempts to solve this, such approaches often operate on only one category of scene or modality of sensor. Additionally, these methods are often completion-only, meaning scene extension would require a second model. Common approaches often rely on deterministic regression models, making them sensitive and incapable of scene generation.

Generative models, particularly diffusion models[10, 20] have recently gained popularity, especially for the task of scene extension and generation[12, 27, 18]. Unfortunately, such models are also often constrained to a single domain: indoor scenes, outdoor scenes, 3D objects, etc., by virtue of their training data and architecture. Furthermore, many generate occupancy only, leaving semantics to be predicted by semantic segmentation models. Most importantly, the 3D representation used in many of these models is either dense or reductive. Dense tensor representations and 3D convolutions lead to models with scalability issues. Meanwhile, methods that reduce 3D representations to 2D lose spatial relationships, leading to losses in performance and interpretability.

For these reasons we seek a generative framework while using an underlying sparse 3D representation. Our approach uses a latent diffusion model over a sparse 3D tree structure that preserves spatial relationships while greatly reducing the memory required to represent 3D space. This ubiquitous representation also enables our model to generalize to different 3D data modalities with minimal changes to the architecture. Importantly, the generative nature of diffusion models affords us the capability not only to unconditionally generate 3D geometries and semantics, but also to perform semantics scene completion and extension via a completion-through-generation framework.

Our contributions can be summarized as follows:

1) **Two-stage generative framework.** We disentangle scene synthesis into coarse structure diffusion and latent semantic diffusion, ensuring that geometry is first established before fine-grained semantics are generated.
2) **Dual octree graph patch latent representation.** We introduce a compact yet spatially local latent space based on a dual octree graph and patch-based VAE, enabling scalable generation of large indoor and outdoor scenes with voxel-level semantics.
3) **Unified generation and completion in multiple domains.** To the best of our knowledge, this is the first diffusion-based framework that unifies 3D generation, extension, and semantic completion in both indoor and outdoor scenes.

II. Related Work

A. 3D Scene Generation and Completion

Recent work in semantic scene completion (SSC) can be broadly divided into two families. End-to-end regression

Authors are with the Autonomous Robotics and Perception Group at the University of Colorado Boulder, Boulder, CO 80309, USA.

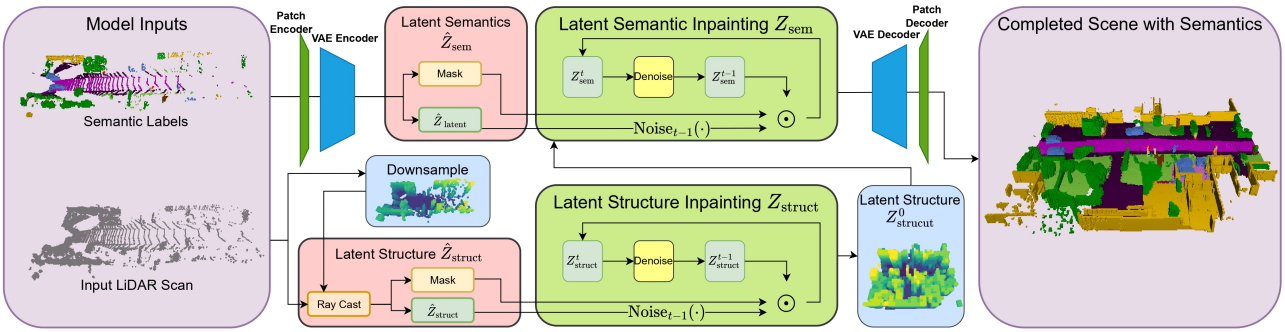

Fig. 1: Our method unifies unconditional generation and LiDAR-conditioned completion within the same framework. When available, a LiDAR scan can be voxelized and used to initialize the latent structure; partial semantic voxels, if provided, are encoded and used to anchor node latents. Both structure and semantics are then introduced during postconditioned diffusion sampling, which preserves observed regions while freely synthesizing unobserved areas. When no LiDAR or semantic input is provided, the masks default to all zeros, causing the model to perform unconditional generation and sample entire scenes from pure noise.

models, and conditioned generative models.

The regression model approach directly predicts voxel occupancies or semantics from RGB images or multi-view inputs [5, 14], often achieving strong accuracy with efficient inference. Their reliance on dense imagery and calibrated poses makes them less suitable for robotics in low-light or unstructured environments. Furthermore, the tendency of such discriminative methods to overfit makes them sensitive to viewpoint, sensor specifications, or other perturbations that shift the data distribution.

In contrast, generative methods treat completion as conditional generation, learning a distribution of plausible 3D scenes, which can then be sampled from to fill unobserved regions and outpaint beyond the field of view. Such models provide richer priors for completion and also serve as standalone generators when no observations are available.

Our work follows this latter paradigm and, importantly, accepts arbitrary point clouds or occupancy grids as inputs, decoupling performance from a specific sensor.

B. 3D Generation via Diffusion Models

Diffusion models have emerged as powerful generative frameworks across modalities. The denoising diffusion probabilistic model (DDPM) [10] and its latent-space variant (LDM) [20] laid the foundations for scalable image synthesis. Extensions into 3D include voxel- and point-based diffusion for shape generation, as well as implicit radiance-field or Signed Distance Fields (SDF)-based [7] pipelines.

Volumetric CNNs generalize 2D CNNs to 3D voxel grids and achieved promising results in 3D generation [4]. However, their cubic computational and memory complexity severely limits scalability to high-resolution or large-scale 3D scenes. Therefore, a central trend has been to pair diffusion with efficient 3D representations to overcome the prohibitive cost of dense volumetric modeling. For example, triplane features reduce volumetric data into three orthogonal 2D planes, enabling compact yet expressive volumetric generation. In parallel, octree-based approaches[27] preserve 3D spatial locality by operating directly on sparse, hierarchical structures, allowing diffusion to scale to high-resolution multiscale shapes. Together, these advances illustrate the shift towards combining diffusion priors with efficient 3D structures to enable practical scene-level generation.

C. Pre- vs. Post-Conditioned Diffusion

While unconditional diffusion enables generation, completion requires injecting partial observations (e.g., sparse LiDAR scans) into the generative process. Two main paradigms exist [8]:

a) **Preconditioned inpainting** trains a model explicitly on masked inputs, learning $p(x|y)$ where y denotes the observed region. This strategy provides high fidelity but requires retraining for each domain or mask distribution, which is costly given the variability of LiDAR viewpoints and sparsity patterns [6]. Furthermore, they demand extensive domain-specific training data covering all possible partial inputs, and requires a sophisticated condition injection mechanism designed for a particular domain, making adaptation across varying conditions (e.g., LiDAR scans from different perspectives) cumbersome and computationally expensive.

b) **Postconditioned inpainting** instead reuses an unconditional diffusion model at inference. Observed regions are preserved by blending them into the denoising trajectory at each step [2, 16]. This approach is training-free and flexible, but its effectiveness depends critically on the latent representation: it must preserve spatial locality to align masks precisely [2].

In principle, postconditioning is attractive for robotics since it avoids retraining under every sensor and data domain configuration. However, dense 3D latents are prohibitively large, and triplane features, while compact, collapse the volumetric structure onto 2D planes, making 3D masks ambiguous. This conflict motivates octree-based hierarchical representations, which remain compact yet preserve relative

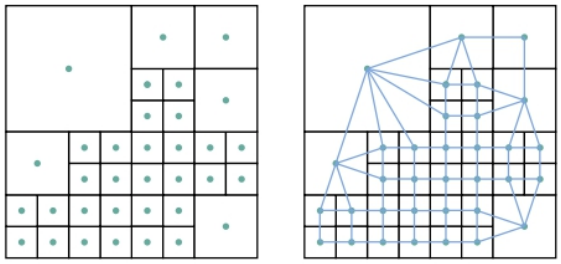

Fig. 2: **Left:** A 2D rendering of an octree. **Right:** Corresponding dual octree graph.

spatial locality, enabling accurate mask conditioning and efficient postconditioned completion at scale.

D. Octree- and Graph-Based 3D Representations

Hierarchical sparse representations like octrees provide efficient and scalable 3D modeling by focusing computation and memory on occupied regions. Octomap [11] pioneered octree-based scene mapping in robotics. OctNet [19] demonstrated that octrees can support deep convolutional architectures, enabling high-resolution volumetric learning with manageable cost. O-CNN [25] further optimized GPU-resident octree structures and showed their effectiveness for 3D classification and segmentation.

Later, a "dual" representation of octrees [13] was proposed to create a semi-regular graph over arbitrary octree inputs. Building on this idea, Dual Octree Graph Networks [23] reorganize features into a graph of face-adjacent octree nodes (dual octree graph), enabling structured graph convolutions across scales. This design improves efficiency and expressiveness by capturing both local adjacency and hierarchical depth relationships.

Most recently, OctFusion [27] integrated octree-based latent representations with multiscale diffusion, showing that a unified U-Net backbone can generate high-resolution, compact 3D shapes within seconds. This work highlighted the power of combining diffusion with hierarchical sparse structures for efficient and high-fidelity 3D generation.

Extending such paradigms from object-level shape generation to scene-level semantic generation and completion introduces new challenges. First, occupancy distributions differ: objects are compact and centered, whereas scenes are sparse and horizontally extended. Second, real-world robotics tasks require semantic prediction in addition to geometry. Third, completion with partial observations, which is essential for LiDAR perception, was not considered in prior octree diffusion work. These gaps motivate our approach, which unifies coarse occupancy diffusion, dual octree graph latent diffusion, and VAE semantic decoding, together with postconditioned blending to support LiDAR-driven inpainting and outpainting.

E. Summary and Motivation

To summarize, regression-based pipelines offer speed but lack generative flexibility, while diffusion-based models provide strong priors but demand efficient 3D representations. Recent trends show the effectiveness of triplane for scaling diffusion in 3D, and of postconditioned inpainting for training-free completion [12]. Yet no prior work has unified these advances into a framework that is (i) **two-stage**, disentangling coarse structural generation from semantic latent synthesis, (ii) **octree-based**, preserving locality with sparse efficiency, and (iii) **training-free-conditioning**, enabling LiDAR-driven completion under varied conditions, including darkness. Our approach fills this gap by combining coarse occupancy diffusion with dual octree graph latent diffusion and VAE decoding, while leveraging postconditioned blending to integrate partial observations seamlessly.

III. Methods

We design a diffusion-based framework that unifies structure prediction, semantic synthesis, and training-free conditioning. Our approach is guided by three principles: (i) **sparse 3D representation**, achieved through a dual octree graph that concentrates computation on occupied regions while preserving spatial relationships; (ii) **semantic compactness**, enabled by a patch-based VAE that prevents semantic latents from being diluted by unoccupied space; and (iii) **flexible conditioning**, realized via postconditioned blended latent diffusion that integrates partial point cloud observations without retraining. The overall pipeline proceeds in two stages: first, a diffusion model predicts the coarse occupancy structure of the scene; second, latent diffusion refines per-node semantics, which are decoded into voxel labels through the VAE decoder. This design disentangles structure from semantics, supports scalable scene generation, and enables training-free completion in diverse sensing conditions. An overview of the pipeline is shown in Figure 3.

A. Dual Octree Graph Representation

Octrees, a common data structure used in robotics, uses hierarchical subdivisions of occupied space only, taking advantage of the natural sparsity of 3d scenes. However, octrees alone lack regular neighborhood structure, making standard convolutions intractable. We therefore convert octree leaves into a dual octree graph, where face-adjacent nodes across depths are connected (see Figure 2 for architectural details) [13, 24]. This dual graph forms a semi-regular structure with a finite set of edge types, enabling efficient graph convolutions, with orientation-specific kernels shared across all nodes. Dual octree graph CNNs can thus aggregate multi-scale context while maintaining sparsity. Downsampling and upsampling are implemented via grouping and splitting sibling nodes, ensuring consistent multi-resolution processing [24].

B. Patch-Based Variational Autoencoder

Unlike 3D shapes, real-world scenes often exhibit occupancy distribution ill conditioned for octree construction:

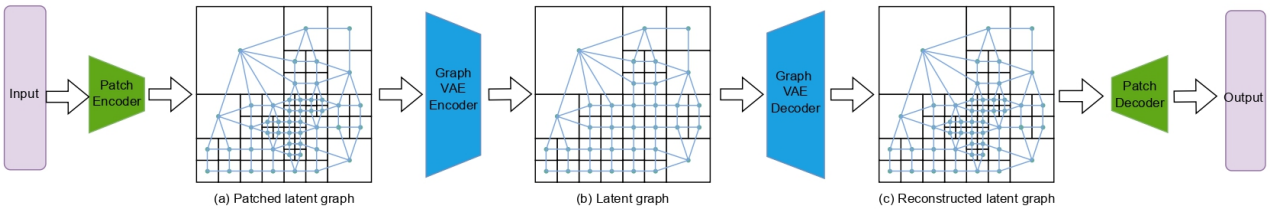

Fig. 3: The Patch-VAE pipeline. A semantic voxel representation of the scene is given as input. Next, the patch encoder does spatial compression over every non-empty patch in the semantic voxel map and forms an octree in the compressed space. Then, it is converted into a dual octree graph. The VAE encoder outputs a latent representation at a shallower depth graph. The VAE decoder utilizes a shared MLP head to predict the split signal to each node at the finest depth, and reconstructs the latent graph. Finally, the patch decoder converts the latent into a semantic voxel map.

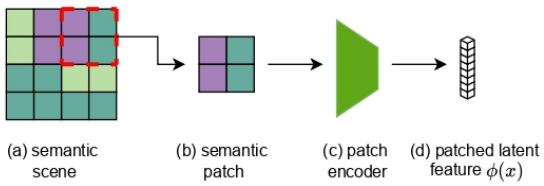

Fig. 4: Given a prespecified patch size, a shared patch encoder operates on every single patch from the semantic scene, and output a latent vector for each patch.

geometry is concentrated near floors and walls, while large volumes remain empty. Moreover, the overall geometry is often quite flat, especially in outdoors (e.g. 256*256*32). Encoding such data directly results in wasted computation, ill-conditioned octree geometries and unbalanced information flow.

We therefore introduce a patch-based VAE, which compresses local cuboid patches into compact latent vectors while discarding empty regions. More specifically, each voxel patch is processed by a shared encoder composed of convolutions followed by an MLP, yielding a latent vector (Figure 4). Patches consisting only of empty voxels are skipped, such that the latents remain as sparse as possible. The resulting octree is therefore, of shallower depth, denser, and more geometry-alligned for dual octree graph network.

The latent vectors are then embedded into a dual octree graph, where a GraphVAE aggregates multi-scale context. The encoder compresses the latent graph to a shallower depth, while the decoder expands it back to its original depth, i.e., per-patch latents. During decoding, before each deconvolution step, a shared MLP head predict the split signal for each node at the depth d (split signal is described in section III-C). Deconvolution only operates on nodes that are split. Finally, each latent is decoded by a shared deconvolutional head into voxel-wise semantic probabilities.

The VAE is optimized with three terms:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{octree}} + \beta\,\mathcal{L}_{\text{KL}}, \tag{1}$$

where $\mathcal{L}_{\text{sem}}$ is cross-entropy for semantic reconstruction, $\mathcal{L}_{\text{octree}}$ is binary cross-entropy on split signals across octree levels, and $\mathcal{L}_{\text{KL}}$ is a KL penalty with weight β. This encourages the VAE to produce structure-aware, generative-ready latents.

C. Two-Stage Diffusion for Structure and Semantics

We disentangle generation into two diffusion stages: structure diffusion and latent diffusion.

Structure diffusion predicts binary split signals of the octree, progressively constructing the occupancy structure. We model each node as either subdivided or terminal, recursively generating a dual octree graph [27]. This ensures global topology is established before semantic details are added. To be more specific, beginning with a dense voxel grid of dimensions $2^D \times 2^D \times 2^D$, where each voxel holds a binary split signal (0 or 1) predicted by the structure diffusion model. Voxels labeled "1" are recursively subdivided into eight octants, and a dual octree graph at depth $D{+}1$ can be constructed with face-adjacent relations. Repeating this process by assigning 0/1 to nodes and subdividing those labeled "1" yields a dual octree graph at depth $D{+}2$. Unlike previous methods [27], which trains a unified network across multiple depths, our patch-based encoder design allows us to operate at shallower depths, reducing memory and computation.

Latent semantic diffusion operates on the generated structure (i.e., a dual octree graph), and predicts latent codes for each octree node. Here, a dual octree graph U-Nets serve as the backbone of the diffusion model, propagating multi-scale context to produce coherent semantic latents on each node. The semantic latents are then decoded via the graph VAE decoder, and then reconstructed by the patch decoder, yielding a full voxel-level semantic scene. This factorization, structure first, semantics second, ensures valid and sparse geometry while capturing semantic diversity.

D. Completion via Postconditioned Blending

To extend generation into completion, we incorporate partial observations as conditions during sampling using postconditioned blended diffusion.

During structure diffusion, known occupancies from LiDAR scans are forward-noised into the sampling trajectory. At each denoising step, the model's prediction is blended with the noised reference according to a binary mask, preserving observed voxels while freely synthesizing unobserved regions. This enables both inpainting and outpainting

without retraining. At each reverse step, the model's denoised prediction $\tilde{x}_{t-1}$ is interpolated with a noised reference input x_{t-1}^{ref}:

$$\hat{x}_{t-1} \;=\; (1-m) \odot \tilde{x}_{t-1} \;+\; m \odot x_{t-1}^{\text{ref}}, \qquad (2)$$

where $\odot$ denotes the element-wise product. At each denoised step, with a binary mask m, the entire regions are freely synthesized, while observed voxels are copied exactly from x_{t-1}^{ref}. Thus, the conditioned reference will gradually be injected into the denoising process, and will be kept intact at the sampling result.

Once the structure is generated, partial semantic information can be injected at the latent level. Known latents are assigned to observed nodes in the dual octree graph and preserved during denoising. This anchors the semantics to known regions while allowing generative refinement elsewhere. To our best knowledge, blended diffusion has not been implemented in graph networks. Our empirical results shows it does work in this semi-regular graph.

Together, these steps form a training-free pipeline: LiDAR scans are first converted into partial occupancy masks, which guide structure diffusion; the resulting octree is then augmented with partial semantic latents, guiding latent diffusion. The final output is a completed semantic scene that is both structure-consistent and observation-faithful.

Feature	SceneSense	Octfusion	SemCity	Ours
Indoor	✓	✗	✗	✓
Outdoor	✗	✗	✓	✓
Objects	✗	✓	✗	✗
Sparse representation	✗	✓	✓	✓
Scene Extension	✓	✗	✓	✓
Scene Completion	✓	✓	✗	✓
Scene Generation	✓	✗	✓	✓
Semantics	✗	✗	✓	✓

TABLE I: Comparison of the capabilities of SceneSense, Octfusion, SemCity, and our approach.

IV. Experiments

We evaluate our dual octree graph latent diffusion model on indoor and outdoor scene generation, completion, and extension tasks. For outdoor scenes, we train on the SemanticKITTI [3] training set and report results on the validation set. For indoor scenes, we use the Replica dataset [21]. Since Replica does not provide an official train/validation split, we randomly partition the scenes into a 90/10 split and keep this split fixed across all experiments.

SemanticKITTI contains 20 semantic categories, whereas Replica contains 92. Each voxel in SemanticKITTI corresponds to a $0.2\,\text{m}$ cube, while in Replica each voxel represents $0.05\,\text{m}$. We adopt a voxel grid of size $256\times256\times32$ for outdoor experiments and $128\times128\times32$ for indoor experiments. Patch sizes are set to $1\times4\times4$ for SemanticKITTI and $1\times2\times2$ for Replica to balance resolution with memory efficiency.All experiments were primarily conducted on NVIDIA GeForce RTX 4070 Ti and RTX 4060 Ti GPUs (16GB and 8GB VRAM, respectively).

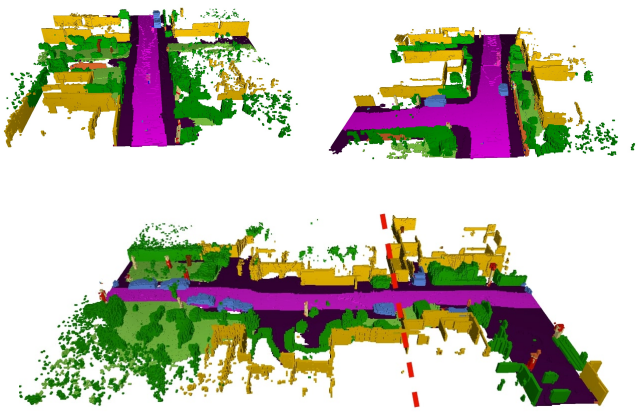

Fig. 5: Scene generation results via our two-stage pipeline. **Top:** Two example semantic scene generations **Bottom:** Example semantic scene extension. Left of the dotted line is an input semantic scene, to the right is an extension of the scene via outpainting.

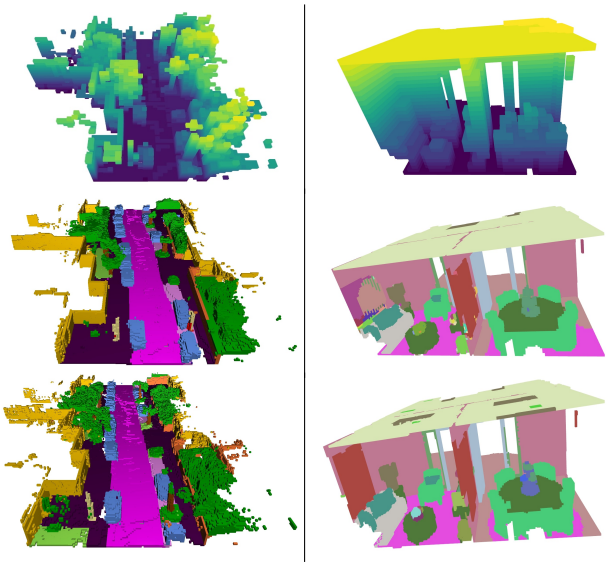

Fig. 6: **Latent Semantic Generation.** Each column shows an example scene. **Top:** Input downsampled occupancy structure. **Middle:** Semantic scene generated via latent semantic diffusion and VAE decoding, conditioned on the structure. **Bottom:** Ground-truth semantic map.

A. Scene Generation

We first evaluate unconditional scene generation on both outdoor and indoor datasets. To quantify distributional similarity between generated and real scenes, we compute three complementary metrics comparing the distribution of generated samples with that of the validation set:

Fréchet Inception Distance (FID) computes the Fréchet distance between the Gaussian approximations of real and generated feature distributions. Lower FID indicates closer alignment in feature space. **Kernel Inception Distance (KID)** is an unbiased estimate of the squared Maximum Mean Discrepancy (MMD) between Inception features, com-

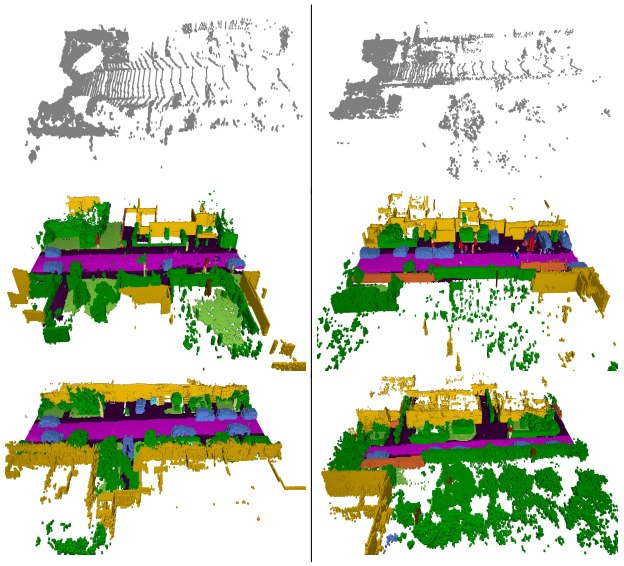

Fig. 7: Semantic Scene Completion. Each column shows an example scene. **Top:** Input a single LiDAR scan. **Middle:** Generated semantic map. **Bottom:** Ground-truth semantic map.

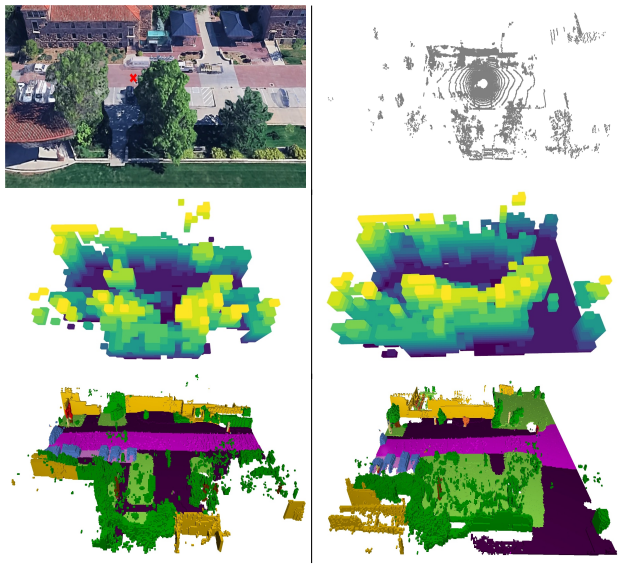

Fig. 8: Zero Shot transfer to CU-Multi Dataset. **Top Left:** satellite view of the scene, where the red crossing denotes the position of the robot. **Top Right:** input LiDAR scan from scene via CU-Multi dataset.(Note that LiDAR scan and satellite image is collected on different days.) **Middle:** inpainted latent structure of the scene. **Bottom:** output semantically labeled scene from two-stage generation with latent structural inpainting.

Method	FID ↓	KID ↓	MMD ↓
2-stage Generation (Ours)	0.2107	8.8700	0.1856
Semantic Generation (Ours)	0.0289	0.7622	0.0265
Semantic Completion (Ours)	0.0353	0.6980	0.0216
SemCity (Baseline)	0.0940	2.2256	0.0773

TABLE II: Distributional metrics (FID, KID, and MMD) between generated scenes and the SemanticKITTI validation set. Note that KID is scaled by $\times 10^3$. We compare our two-stage generation pipeline, latent-semantic generation, and semantic scene completion against the SemCity baseline.

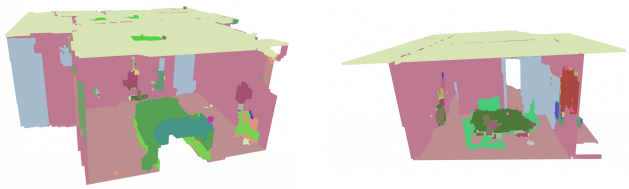

Fig. 9: Generations of indoor scenes

puted with a degree-3 polynomial kernel. **Maximum Mean Discrepancy (MMD)** is additionally computed directly on our VAE latent features using an Radial Basis Function (RBF) kernel. This captures distributional differences in the model's native representation space.

Together, these metrics provide a robust measure of both global distributional alignment and task-specific latent consistency. Specifically, we train a bird's-eye-view VAE to compute metrics in outdoor generation, and train a 3D VAE for indoor generation. We present these results alongside comparable methods in tables II and III. Qualitative results (Figures 5 and 9) demonstrate that the generated samples capture both the global structure (roads, buildings, room layouts) and fine-grained semantics (cars, pedestrians, furniture). We modify the voxel-based diffusion model [18] and triplane latent diffusion model [12] to compare in indoor semantic generation in table III.

To evaluate the latent semantic diffusion in isolation, we run stage-2 latent semantic diffusion using the ground-truth coarse occupancy as input, bypassing the structure generation stage. Given the ground truth down-sampled structure, the model achieves competitive metrics, indicating its strong capability to generate semantics that remain plausible and consistent with the data distribution. Since we operate in a $32\times32\times16$ downsampled occupancy grid (each voxel corresponds to $8\times8\times2$ neighboring voxels), semantically correct predictions may be penalized in IoU if they shift by several cell in original space. Figure 7 illustrates this "mask dilation" effect: while predictions align structurally with the ground truth, they are not perfectly voxel-to-voxel aligned.

Results in tables IV and V show that our model nonetheless produces coherent semantics across major classes such as road, sidewalk, vegetation, floor and wall. Given ground-truth coarse occupancy in both SemanticKITTI and Replica validation dataset, we generate semantics via latent diffusion. Results show strong performance on major scene classes relevant for navigation and planning.

B. Semantic Scene Completion

Next, we test semantic scene completion from sparse partial LiDAR scans. We downsample a single input scan and project it into the voxelized structure space, then apply

Method	FID ↓	KID ↓	MMD ↓
2-stage Generation (Ours)	0.8671	0.5245	0.2837
Semantic Generation (Ours)	0.0486	0.1193	0.0668
SceneSense (Baseline)	31.0950	89.7884	47.8399
SemCity (Baseline)	50.2789	187.2540	99.5552

TABLE III: FID, KID, and MMD between pure generation and the Replica validation set. Note that FID, KID and MMD values are scaled by 10^4, 10^5 and 10^4 for respectively. We compare our 2-stage generation with triplane- and voxel-based representations [12, 18].

Class	Road	Sidewalk	Building	Vegetation	mIoU
IoU (%)	67.6	37.0	28.1	33.5	16.9

TABLE IV: Latent Semantic Generation on SemanticKITTI Given Ground-Truth Coarse Occupancy.

structure diffusion to inpaint the missing geometry. Semantic diffusion is then applied on top of the completed structure to recover semantic occupancy. As shown in Figure 7 and table II, our blended diffusion produces coherent completions that respect local context while maintaining global consistency. Conditioned on a single LiDAR scan, our model generates scenes of comparable quality to those produced when conditioned on the complete oracle structure. Outpainting further extends the scene beyond the observed region, producing plausible continuations (Figure 5).

C. Zero-Shot Generalization

To assess the generalization ability of our model, we apply our trained structure and latent semantic diffusion models to new domains without any retraining or fine-tuning. We test on LiDAR scans from CU-Multi [1], a dataset collected in outdoor areas of a campus, which differs from SemanticKITTI in LiDAR configuration (e.g. height and tilt angle) and scene layout. Given a single LiDAR scan from the dataset, our model successfully completes missing structure and generates a series of plausible semantic maps aligned with real-world layouts, despite the domain shift (Figure 7). This highlights the robustness of our dual octree graph latent diffusion framework and its potential for real-world deployment.

V. Limitations and Future Work

Although our two-stage framework successfully disentangles structure generation from latent semantic diffusion, we observe that the overall generation quality is still bottlenecked by the structure generation stage. In our current implementation, structure generation is performed with 3D CNNs at a coarse occupancy level. While more advanced or domain-specific models could improve this stage, generating plausible structures without semantic guidance remains a challenging problem.

Another limitation arises from mask handling during post-conditioned generation. Because masks must be defined in the downsampled latent space, a single masked voxel corresponds to a larger region in the original space. This leads to slight mask dilation when mapped back to the high-resolution voxel grid, causing imperfect alignment between the conditioned input and the final generation, which is reported as a common limitation for latent-space postconditioning methods [8].

Class	Floor	Wall	Door	Table	mIoU
IoU (%)	98.57	97.89	72.49	71.58	32.22

TABLE V: Latent Semantic Generation on Replica Given Ground-Truth Coarse Occupancy.

For future work, we plan to explore classifier-free guidance to more explicitly unify indoor and outdoor scene generation within a single model. A major challenge is that existing datasets are typically domain-specific and define different semantic taxonomies, making joint training difficult. One promising direction is to adopt open-vocabulary semantic representations, allowing cross-domain training without label-space conflicts. Finally, while this work is evaluated offline, our method is efficient enough to be deployed in online settings. We plan to integrate it into indoor and outdoor robotic navigation systems to enable real-time semantic completion and scene reasoning.

VI. Conclusion

We introduced a dual octree graph latent diffusion framework that is capable of generating, completing, and extending 3D semantic scenes in a single architecture. Experiments on SemanticKITTI, Replica, and zero-shot inference highlight its flexibility, and generation quality. Overall, our results suggest that completion-through-generation in a locality-preserving structured latent space is a viable direction for robotic perception.

Acknowledgment

The authors acknowledge the use of GPT-4o (OpenAI) for refining the literature review in Section II and methods in Section III. The generated content was reviewed and edited by the authors to ensure technical accuracy and consistency with the overall manuscript.

References

[1] Doncey Albin et al. *CU-Multi: A Dataset for Multi-Robot Data Association*. 2025. arXiv: 2505.17576 [cs.RO].

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. "Blended Latent Diffusion". In: *ACM Trans. Graph.* 42.4 (July 2023). ISSN: 0730-0301.

[3] J. Behley et al. "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences". In: *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*. 2019.

[4] André Brock et al. "Generative and Discriminative Voxel Modeling with Convolutional Neural Networks". In: *CoRR* abs/1608.04236 (2016). arXiv: 1608.04236.

[5] Anh-Quan Cao and Raoul de Charette. "MonoScene: Monocular 3D Semantic Scene Completion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 3991–4001.

[6] Helin Cao and Sven Behnke. "DiffSSC: Semantic LiDAR Scan Completion using Denoising Diffusion Probabilistic Models". In: *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2024), pp. 2185–2192.

[7] Gene Chou, Yuval Bahat, and Felix Heide. "DiffusionSDF: Conditional Generative Modeling of Signed Distance Functions". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 2262–2272.

[8] Ciprian Corneanu, Raghudeep Gadde, and Aleix M. Martinez. "LatentPaint: Image Inpainting in Latent Space With Diffusion Models". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2024, pp. 4334–4343.

[9] Ankit Dhiman et al. "Reflecting Reality: Enabling Diffusion Models to Produce Faithful Mirror Reflections". In: *International Conference on 3D Vision 2025*. 2025.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[11] Armin Hornung et al. "OctoMap: An efficient probabilistic 3D mapping framework based on octrees". In: *Autonomous robots* 34.3 (2013), pp. 189–206.

[12] Jumin Lee et al. "SemCity: Semantic Scene Generation with Triplane Diffusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 28337–28347.

[13] Alejandro León, Juan Carlos Torres, and Francisco Velasco. "Volume octree with an implicitly defined dual grid". In: *Computers & Graphics* 32.4 (2008), pp. 393–401. ISSN: 0097-8493.

[14] Yiming Li et al. "VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 9087–9098.

[15] Bencheng Liao et al. "DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving". In: *CoRR* abs/2411.15139 (2024).

[16] Andreas Lugmayr et al. "RePaint: Inpainting Using Denoising Diffusion Probabilistic Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 11461–11471.

[17] Alec Reed et al. "Online Diffusion-Based 3D Occupancy Prediction at the Frontier with Probabilistic Map Reconciliation". In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 2025, pp. 2846–2852.

[18] Alec Reed et al. "SceneSense: Diffusion Models for 3D Occupancy Synthesis from Partial Observation". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2024, pp. 7383–7390.

[19] Gernot Riegler, Ali O. Ulusoy, and Andreas Geiger. "OctNet: Learning Deep 3D Representations at High Resolutions". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 6620–6629.

[20] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10674–10685.

[21] Julian Straub et al. *The Replica Dataset: A Digital Replica of Indoor Spaces*. 2019. arXiv: 1906.05797 [cs.CV].

[22] Jianyuan Wang et al. "VGGT: Visual Geometry Grounded Transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2025.

[23] Peng-Shuai Wang, Yang Liu, and Xin Tong. "Dual octree graph networks for learning adaptive volumetric shape representations". In: *ACM Trans. Graph.* 41.4 (July 2022). ISSN: 0730-0301.

[24] Peng-Shuai Wang, Yang Liu, and Xin Tong. "Dual octree graph networks for learning adaptive volumetric shape representations". In: *ACM Transactions on Graphics (TOG)* 41.4 (2022), pp. 1–15.

[25] Peng-Shuai Wang et al. "O-CNN: octree-based convolutional neural networks for 3D shape analysis". In: *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301.

[26] Yi Wei et al. "SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving". In: Oct. 2023, pp. 21672–21683.

[27] Bojun Xiong et al. "OctFusion: Octree-based Diffusion Models for 3D Shape Generation". In: *Computer Graphics Forum* 44 (2024).