OMNICORE: HETEROGENEOUS CONDITIONING FOR CONTROLLABLE 3D SCENE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable outdoor 3D scene generation is central to autonomous driving and urban simulation, yet existing methods remain limited in scalability and flexibility. They typically depend on dense semantic annotations or rigid conditioning signals that are rarely available in practice. We introduce OmniCore, to our knowledge the first unified framework for controllable 3D scene generation in which roadmaps define the primary road layout, while natural language and heightmaps further specify its semantic layout and vertical structure. This design enables both high-fidelity 3D semantic scene synthesis and region-specific editing. OmniCore features a novel windowed spectral-attention block (WSAB), which integrates local spatial attention with scene-adaptive spectral filtering within each window to improve fine-grained geometric continuity across multimodal representations. To facilitate systematic evaluation, we present OmniControl-3D, a large-scale benchmark for controllable 3D scene generation that enriches SemanticKITTI, KITTI-360, and nuScenes with diverse annotations, including roadmaps, natural language descriptions, and heightmaps, each paired with 3D semantic scenes. Extensive experiments show that OmniCore achieves state-of-the-art performance across all conditioning configurations, demonstrating strong flexibility and spatial precision for controllable 3D scene generation. In addition, OmniCore can serve as a scalable data engine, generating diverse and semantically consistent 3D urban scenes that improve downstream 3D perception, as demonstrated in 3D semantic scene completion. A subset of OmniControl-3D is available here; the full benchmark and code will be released.

1 INTRODUCTION

Controllable 3D outdoor scene generation has applications in autonomous driving (Pan et al., 2021; Li et al., 2022; Lin et al., 2024) and large-scale urban simulation (Bougie & Watanabe, 2025; Zheng et al., 2025), enabling scalable generation of diverse, realistic, and user-controllable environments. Achieving such generation requires conditioning signals that are intuitive for users, expressive enough to capture scene layout, semantics, and geometry, and scalable to complex scenes.

Recent 3D outdoor scene generation approaches have achieved notable progress, enabling diverse scene synthesis conditioned on strong structural and semantic priors. For instance, UrbanDiff (Zhang et al., 2024b) uses annotated Bird's-Eye View (BEV) maps with object annotations to generate urban scenes with geometry and semantics. To reduce reliance on dense BEV inputs, CymbaDiff (Liang et al., 2025) leverages freehand-like sketches and pseudo-labeled satellite annotations (PSA) as high-fidelity priors for 3D outdoor semantic scene generation. PrITTI (Tze et al., 2025) further improves editability through a primitive-based representation that couples vectorized objects with rasterized ground surfaces, but still depends on accurate dense semantic scene labels. While these priors improve controllability and editability, they rely on fixed modalities, are hard to obtain in user-controllable scenarios, and are costly to scale.

These limitations highlight the need for flexible and readily obtainable conditioning modalities that support high-fidelity, semantically rich generation without oracle-level supervision. Natural language offers an expressive and intuitive interface for controllable scene generation, allowing users to specify free-form semantic intent without dense annotations. It has been widely adopted in 2D image synthesis (Xiao et al., 2025; Bernal-Berdun et al., 2025; Thakral et al., 2025; Han et al., 2025b; Srivatsan et al., 2025), and recently extended to 3D generation across individual objects (Chen et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025b; Yang et al., 2025a), indoor environments (Li et al., 2024; Öcal et al., 2024; Deng et al., 2025; Han et al., 2025a; Bai et al., 2025; Bokhovkin et al., 2025), and outdoor scenes (Zhang et al., 2024c; Qu et al., 2025; Yang et al., 2025b; Liu et al., 2025b). However, existing language-conditioned 3D semantic scene generation methods offer limited fine-grained control: they struggle to translate detailed instructions into precise, region-specific semantic edits while preserving the surrounding scene structure. This hinders iterative improvements and weakens natural language as a practical interface for high-fidelity, user-controllable 3D scene synthesis.

To address these challenges, we need a multi-modal generative model that supports flexible and readily available conditions, including intuitive natural language, for high-fidelity 3D scene synthesis and region-specific editing. However, this introduces two coupled challenges: (i) multi-modal fusion may cause substantial computational overhead and abstraction-level information loss; and (ii) propagating salient, condition-specific cues into 3D generative modeling requires preserving fine local geometric continuity to further improve global structural consistency. Recent methods (Li et al., 2023b; Tze et al., 2025) often rely on window attention (Liu et al., 2021) to avoid quadratic attention costs caused by Transformers (Vaswani et al., 2017), which unfortunately hinders geometric continuity across 3D grids. Although subsequent hybrid spatial-frequency designs (Zeng et al., 2024; Tran & Xin, 2025) employ global FFT to bridge this gap, they still struggle to preserve fine-grained local details within complex volumetric environments.

Motivated by these limitations, we propose **OmniCore**, to our knowledge the first unified framework for controllable 3D scene generation conditioned on three complementary signals: roadmaps (horizontal layout), heightmaps (vertical structure), and natural language (semantic placement). The roadmap, a simple and readily obtainable signal, defines the global road layout; natural language specifies where each semantic category is placed within it; and a grid-aligned heightmap recovers the vertical structure discarded by top-down conditions such as BEV maps and sketches (Zhang et al., 2024b; Liang et al., 2025), providing explicit control over vertical geometry that existing methods lack. OmniCore grounds these sparse, abstract inputs into fine-grained voxel-level control, enabling high-fidelity 3D semantic scene synthesis and region-specific editing without disturbing surrounding structure. Because each signal governs a single factor, it can be edited independently of the others. We further introduce the Windowed Spectral-Attention Block (WSAB), which couples local spatial attention with a scene-adaptive spectral filter within each window to improve fine-grained geometric continuity across volumetric grids. Finally, OmniCore can serve as a scalable data engine, generating diverse and semantically consistent 3D urban scenes from readily obtainable conditions.

To enable systematic evaluation, we further present **OmniControl-3D**, a large-scale multimodal benchmark for controllable 3D scene generation. It comprises three subsets, OmniControl-3D-SemanticKITTI, -KITTI-360, and -nuScenes, each pairing 3D ground-truth (GT) scenes with roadmaps, heightmaps, and textual descriptions to support flexible conditioning. Importantly, all conditioning signals are inexpensive to obtain at scale: roadmaps can be collected from public mapping services (e.g. Google Maps, OpenStreetMap), textual descriptions are generated by vision-language models (e.g. GPT-4 (Achiam et al., 2023)) over the scenes, and heightmaps are derived deterministically from the underlying volumes. This pipeline eliminates manual annotation of conditioning and bypasses the multi-stage pseudo-labeling pipelines required to obtain dense conditioning signals, substantially reducing the human and computational cost of dataset construction.

In summary, our main contributions include:

- We introduce OmniCore, the first unified framework for controllable 3D scene generation that supports flexible conditioning on roadmaps (horizontal layout), heightmaps (vertical structure), and natural language (semantic placement) for region-specific editing and high-fidelity 3D semantic scene synthesis. OmniCore achieves superior performance across diverse conditioning settings with precise user-driven control.
- We propose a Windowed Spectral-Attention Block (WSAB), which couples local spatial attention with a scene-adaptive spectral filter within each window to emphasize the frequency components most salient to the underlying scene structure, to further improve fine-grained geometric continuity across volumetric grids.
- We present OmniControl-3D, a large-scale multimodal benchmark for controllable 3D outdoor scene generation, comprising OmniControl-3D-SemanticKITTI, -KITTI-360, and -

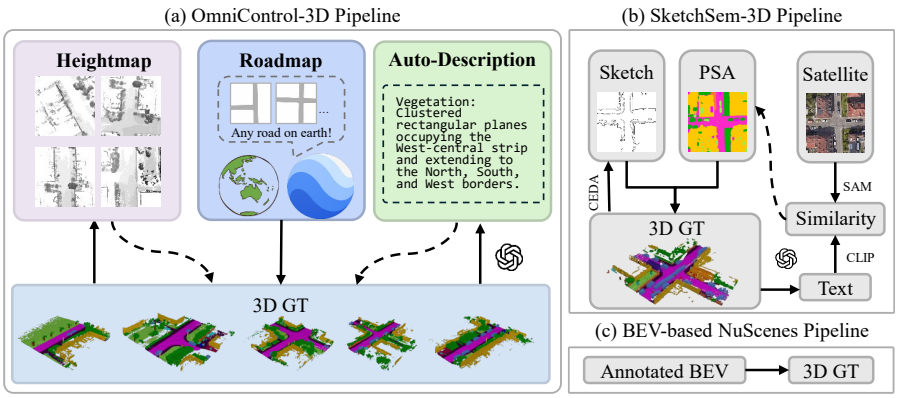

Figure 1: Data construction pipelines for (a) OmniControl-3D, (b) SketchSem-3D, and (c) BEV-based nuScenes. CEDA: Canny Edge Detection Algorithm; SAM: Segment Anything Model.

nuScenes, which pair 3D semantic scenes with roadmaps, heightmaps, and textual descriptions.

Table 1: Data statistics of existing datasets and our OmniControl-3D. Cond.: condition, R: roadmap; T: text; H: heightmap.

Dataset	Samples	Cond. Source	Input Modalities	Conditioning	Classes	3D GT Size
BEV-based nuScenes (Zhang et al., 2024b)	34149	GT annotation	BEV	Static	17	$192 \times 192 \times 16$
SketchSem-3D-SemanticKITTI (Liang et al., 2025)	58987	Pseudo-label	Sketch / PSA	Static	20	$256 \times 256 \times 32$
SketchSem-3D-KITTI-360 (Liang et al., 2025)	36057	Pseudo-label	Sketch / PSA	Static	19	$256 \times 256 \times 32$
OmniControl-3D-SemanticKITTI	92460	annotation-free	R/T/H	Dynamic	20	$256 \times 256 \times 32$
OmniControl-3D-KITTI-360	53640	annotation-free	R/T/H	Dynamic	19	$256 \times 256 \times 32$
OmniControl-3D-nuScenes	68242	annotation-free	R/T/H	Dynamic	11	$256 \times 256 \times 32$

2 Related Work

Window-Based Attention and Spectral Methods. Transformers (Vaswani et al., 2017) have become a dominant architecture, with applications spanning 2D image generation (Rombach et al., 2022; Tian et al., 2024; Jia et al., 2025), segmentation (Perera et al., 2024; Wu et al., 2024a; Zhou et al., 2025), and 3D spatial reasoning tasks such as semantic scene completion (Wang & Tong, 2024; Yu et al., 2024; Xue et al., 2025; Guo et al., 2025; Pham et al., 2025), object-level completion (Cai et al., 2024; Cui et al., 2024; Zhang et al., 2024a; Wu et al., 2024b; Rong et al., 2024), and semantic scene synthesis (Fan et al., 2024; Jia et al., 2024; Wu et al., 2024c; Bokhovkin et al., 2025). However, the quadratic complexity of standard self-attention is prohibitive at volumetric resolutions. Swin Transformer (Liu et al., 2021) addresses this through local-window attention with shifted windows for cross-window exchange, and later variants further improve the paradigm through stabilized high-resolution training (Liu et al., 2022), cross-shaped partitions (Dong et al., 2022), and neighborhood attention (Hassani et al., 2023). Despite these advances, window-based attention architectures still struggle to preserve geometric continuity across 3D volumetric grids, leading to spatial fragmentation or boundary artifacts at window seams (Wang et al., 2023).

The Fast Fourier Transform (FFT) (Nussbaumer, 1981; Almeida, 1994) reduces the cost of the Discrete Fourier Transform from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, enabling efficient spectral feature extraction in imagery (Reddy & Chatterji, 1996; Kruse et al., 2017; Hu et al., 2025), radar (Singh et al., 2024; Chen et al., 2025a), and learned representations (Xiao et al., 2024; Liu et al., 2025a; Liu & Tang, 2025). Its adoption in vision has progressed from classical spectral filtering (Chellappa & Bagdazian, 1984; Beaudoin & Beauchemin, 2002), to Fourier-based input pre-processing (Yang & Soatto, 2020; Xu et al., 2021), to frequency-domain layers integrated within network architectures (Nguyen et al., 2022; Buchholz & Jug, 2022; Chu et al., 2023; Rao et al., 2021; Luo et al., 2024). Recently, several methods (Zhou et al., 2022; Zeng et al., 2024; Tran & Xin, 2025) combine Fourier transforms with window-based attention to mitigate spatial fragmentation and boundary artifacts at window seams. By mixing spatial tokens in the spectral domain, they expand interaction beyond local receptive fields. These global spectral operations, however, carry inherent drawbacks: (i) foreground and background components interfere in a shared frequency space, weakening

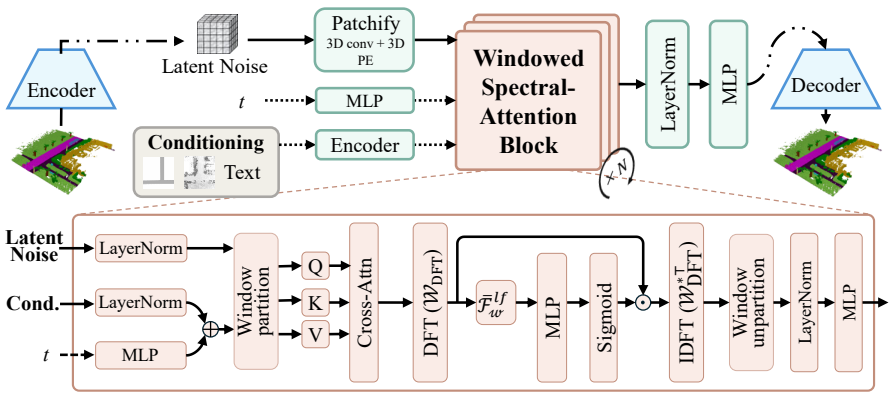

Figure 2: Architecture of the OmniCore diffusion generation network. The latent noise is patchified and fused with time and multimodal conditioning features, then processed by stacked WSABs to model local spatial structure and learnable frequency-domain interactions. The decoder reconstructs the denoised latent into a 3D semantic scene. DFT: Discrete Fourier Transform; IDFT: Inverse Discrete Fourier Transform.

instance-level refinement; (ii) they discard spatial localization, inducing Gibbs-type ringing at sharp discontinuities and suppressing spatially localized periodic structures; (iii) the $\mathcal{O}(N \log N)$ cost of global FFT erodes the $\mathcal{O}(nw^3)$ efficiency of window attention, where n is the number of windows and w is the window size, creating a scalability bottleneck at volumetric resolutions.

3D Semantic Scene Generation. 3D semantic scene generation aims to synthesize geometrically plausible and semantically meaningful 3D environments, with applications in autonomous driving (Pan et al., 2021; Li et al., 2022; Yin et al., 2024; Lin et al., 2024) and virtual simulation (Bougie & Watanabe, 2025; Zheng et al., 2025). Recent progress spans both indoor (Wu et al., 2023; 2024c; Li et al., 2024; Öcal et al., 2024; Deng et al., 2025; Han et al., 2025a; Bai et al., 2025; Bokhovkin et al., 2025) and outdoor (Zhang et al., 2024c;b; Liu et al., 2024; Lee et al., 2024; Bian et al., 2024; Qu et al., 2025; Yang et al., 2025b; Liu et al., 2025b; Nunes et al., 2026) settings, with the field shifting from unconstrained synthesis toward controllable generation. Early work such as STPD (Wu et al., 2023) introduced multimodal guidance via sketches and text for 3D shape generation. Subsequent methods extended this paradigm to urban scenes: UrbanDiff (Zhang et al., 2024b) utilizes BEV maps with annotated objects as conditions for 3D semantic scene generation, while CymbaDiff (Liang et al., 2025) combines freehand-style sketches with PSAs (pseudo-labels) for more flexible control. Departing from purely grid-based representations, PrITTI (Tze et al., 2025) further improves editability through a primitive-based representation that couples vectorized objects with rasterized ground surfaces, but still depends on accurate dense semantic scene labels.

3 OmniControl-3D

Existing conditioning modalities are either costly to annotate or require multi-stage pseudo-labeling, as illustrated in Figure 1. For instance, human-annotated BEV maps require substantial labeling cost, while sketch and PSA demand multi-stage pipelines of off-the-shelf models and post-processing that increases latency and error propagation. To this end, we introduce **OmniControl-3D**, a large-scale multimodal benchmark for controllable 3D scene generation. It comprises three subsets, OmniControl-3D-SemanticKITTI, -KITTI-360, and -nuScenes, each pairing 3D ground-truth scenes with roadmaps, heightmaps, and textual descriptions to support flexible conditioning. All conditioning signals are annotation-free and inexpensive to obtain at scale.

3.1 Data Sourcing and Formatting

We construct the three subsets from the 3D voxelized ground truth of SemanticKITTI (Behley et al., 2019), SSCBench-KITTI-360 (Li et al., 2023a), and SSCBench-nuScenes (Li et al., 2023a), respectively. Within each subset, each scene is paired with a roadmap, a textual description, a heightmap, and its 3D voxelized ground truth. As shown in Figure 1, roadmaps are queried from public mapping services (e.g., Google Maps, OpenStreetMap) using GPS coordinates from KITTI (Geiger

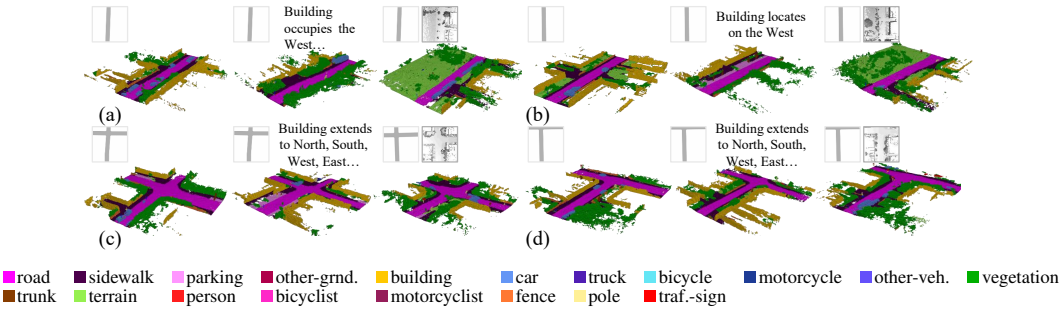

Figure 3: Qualitative results of OmniCore under different conditioning. Columns left-to-right of each group (a–d) correspond to roadmap-only, roadmap-text, and roadmap-heightmap conditioning.

et al., 2012), KITTI-360 (Liao et al., 2022), and nuScenes (Caesar et al., 2020). Textual descriptions are generated by GPT-4 (Achiam et al., 2023) for selected semantic categories in each 3D ground-truth scene, using carefully designed prompts that enforce a consistent format and vocabulary. Heightmaps record the maximum vertical extent at each spatial coordinate in the 3D ground truth, providing a compact geometric representation. Roadmaps and heightmaps share a spatial resolution of 256×256, matching the orthographic BEV resolution of the 3D ground truth. Textual descriptions cover seven semantic categories: road, sidewalk, parking, building, vegetation, car, and fence. Storage formats and label encoding follow SemanticKITTI (Behley et al., 2019). Further details are provided in Appendix A.1.

3.2 Data Statistics and Evaluation Metrics

Table 1 compares OmniControl-3D with three representative 3D scene generation benchmarks: the BEV-based nuScenes dataset (Zhang et al., 2024b) and the SketchSem-3D-SemanticKITTI and -KITTI-360 datasets (Liang et al., 2025), whose conditioning modalities are either predefined or costly. OmniControl-3D provides substantially larger paired corpora than prior benchmarks (92,460 vs. 58,987 on SemanticKITTI; 53,640 vs. 36,057 on KITTI-360) and nuScenes (68,242 samples). This establishes OmniControl-3D as a large-scale, multimodal foundation for high-fidelity 3D scene synthesis at the scale and diversity required by autonomous-driving research. To assess the quality and diversity of generated scenes, we use two standard metrics: Fréchet Inception Distance (FID) (Paschalidou et al., 2021) and Maximum Mean Discrepancy (MMD) (Zhang et al., 2024b), which jointly quantify distributional alignment and feature-level realism between generated and ground-truth scenes. Formal definitions and implementation details are provided in Appendix A.2.

4 Method

Problem Formulation. We denote the roadmaps, text prompts, and heightmaps as $\mathbf{R} \in \mathbb{R}^{L \times W}$, $\mathbf{T} \in \mathbb{R}^{C_t}$, and $\mathbf{H} \in \mathbb{R}^{L \times W}$, and project them jointly into a structured voxel grid $\mathbf{V} \in \mathbb{R}^{L \times W \times H}$, where L, W, and H denote the length, width, and height of the scene. The goal is to predict a discrete label for each voxel, $c \in \{0, 1, \ldots, C-1\}$, where $c = 0$ denotes empty space and $c \in \{1, \ldots, C-1\}$ denotes one of the $C-1$ semantic categories. This formulation casts scene generation as joint occupancy and semantic prediction over volumetric grids, unifying geometric and semantic labeling in a single output space.

4.1 Geometry-Aware Semantic Variational Autoencoder

The OmniCore denoising network operates in the latent space of a geometry-aware semantic VAE tailored for 3D scenes with discrete category labels, which provides a compact yet expressive representation of 3D semantic scenes. Given a voxelized scene $\mathbf{V} \in \mathbb{R}^{L \times W \times H}$, the encoder $\mathcal{E}$ maps $\mathbf{V}$ to a latent posterior, from which we sample $\mathbf{z} \sim \mathcal{E}(\mathbf{V})$. The decoder $\mathcal{D}$ then reconstructs the scene as $\hat{\mathbf{V}} = \mathcal{D}(\mathbf{z})$. With a downsampling factor f, the encoder produces a compact latent representation that retains the salient geometric and semantic structure. The training objective of the VAE is:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{lov}} \mathcal{L}_{\text{Lov}} + \lambda_{\text{kl}} \mathcal{L}_{\text{KL}}, \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ (Zhang & Sabuncu, 2018) supervises per-voxel classification, $\mathcal{L}_{\text{Lov}}$ (Berman et al., 2018) directly optimizes intersection-over-union and encourages sharper semantic boundaries, and

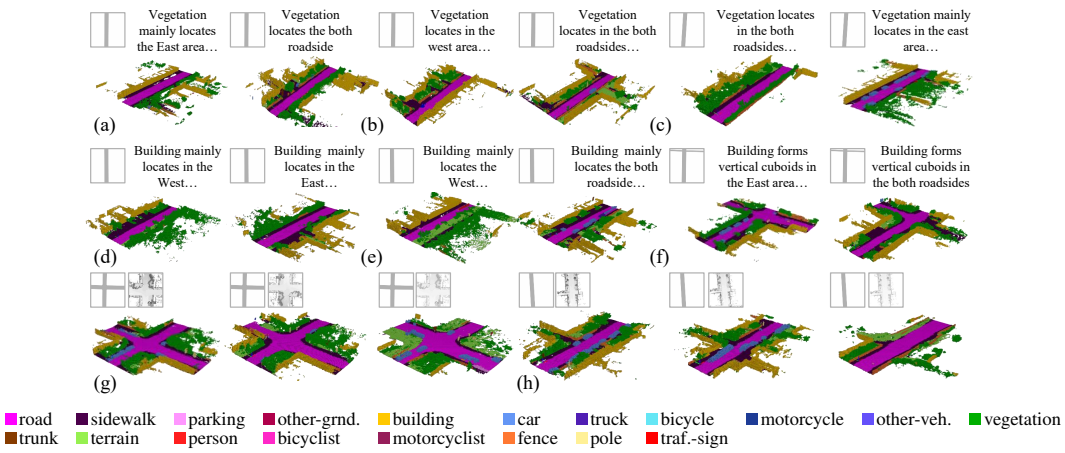

Figure 4: Editable semantic scene generation on OmniControl-3D-SemanticKITTI. The first two rows show text-driven editing of vegetation (a–c) and buildings (d–f). The third row shows heightmap editing in two groups (g–h), where columns left-to-right of each group correspond to heightmap scales of $1.0\times$, $0.8\times$ and $0.5\times$.

$\mathcal{L}_{\text{KL}}$ (Rombach et al., 2022) regularizes the latent posterior toward a standard Gaussian to yield a well-structured latent space for downstream diffusion. λ_{lov} and λ_{kl} balance the contributions of each loss component. Architecturally, $\mathcal{E}$ comprises two down-sampling stages, each with two Conv–BN–ReLU blocks followed by a strided Conv–BN–ReLU block that halves the spatial resolution; $\mathcal{D}$ mirrors this design with transposed convolutions to restore the original resolution. OmniCore training objective is provided in Appendix A.3.

4.2 OmniCore Denoising Network

OmniCore is a conditional diffusion backbone for voxelized semantic scenes. As shown in Figure 2, the latent noise is patchified via 3D convolution with 3D positional encoding (PE), while the diffusion timestep t and conditioning signals are embedded by an MLP and a dedicated condition encoder, respectively. The tokens then pass through a stack of Windowed Spectral-Attention Blocks, which couple local spatial attention with frequency-domain context within each window to sharpen fine-grained geometric continuity across the grid by emphasizing structurally salient frequencies.

Conditioning and Tokenization. Since OmniCore accepts flexible condition combinations, roadmaps-texts, or roadmap-heightmap, we design two fusion pathways, one for each configuration, whose outputs are then unified through a shared voxel tokenization step. For *roadmap–text* conditioning, we apply FiLM-style (Perez et al., 2018) modulation, regarding the roadmap as the structural backbone and the text prompt as a global semantic modulator. A lightweight 2D convolutional encoder extracts roadmap features $\mathbf{F}_r \in \mathbb{R}^{H\times W\times C}$. The text embedding from OpenCLIP (Cherti et al., 2023) is projected by a two-layer MLP into channel-wise FiLM parameters $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^C$, which modulate the roadmap features as $\mathbf{F}_r \odot (\mathbf{1} + \boldsymbol{\gamma}) + \boldsymbol{\beta}$. By having the roadmap provide local spatial structure and the text globally rescale each semantic channel, this design enables free-form prompts to change scene semantics without perturbing the geometric layout.

For *roadmap-heightmap* conditioning, both signals are spatially aligned on the 2D ground-plane grid, we concatenate them along the channel dimension and apply a lightweight 2D convolutional encoder to produce fused features $\mathbf{F}_{RH} \in \mathbb{R}^{H\times W\times C}$. For *roadmap-only* conditioning, the encoder takes $\mathbf{R}$ as a single-channel input, producing $\mathbf{F}_R$ with the same dimensions. For *latent noise tokenization*, the noisy latent $\mathbf{V} \in \mathbb{R}^{H\times W\times D\times C_{\text{in}}}$ is partitioned into non-overlapping patches of size $p \times p \times \frac{p}{2}$ via a 3D convolution, yielding a token sequence $\mathbf{E} \in \mathbb{R}^{L\times d}$ with $L = 2HWD/p^3$, where the factor of two arises from the anisotropic patch size along the vertical axis. We then augment $\mathbf{E}$ with fixed 3D sinusoidal positional encodings, obtained by channel-wise concatenation of three axis-wise 1D sinusoidal encodings (Vaswani et al., 2017), similar to DiT (Peebles & Xie, 2023).

Windowed Spectral-Attention Block. To reduce the quadratic complexity of vanilla self-attention, we restrict attention to non-overlapping local windows. Although computationally efficient, this lim-

Table 2: Semantic scene generation and ablation results. C: condition; SK: sketch; PSA: pseudo-labeled satellite image annotations; R: roadmap; T: text; H: heightmap; W-CA: windowed cross-attention; "-" indicates results not available. Our model consistently outperforms all baselines in both FID and MMD across diverse conditioning signals and datasets.

(a) Comparison under multiple conditionings.

Method	C	FID ↓	MMD ↓
SemanticKITTI			
SSD (Lee et al., 2023)	-	112.82	-
SemCity (Lee et al., 2024)	-	56.55	-
CymbaDiff (Liang et al., 2025)	SK+PSA	40.67	0.04
PrITTI (Tze et al., 2025)	Scene Label	71.03	-
OmniControl-3D-SemanticKITTI			
3D DiT (Mo et al., 2023)	R	83.21	0.09
CymbaDiff (Liang et al., 2025)	R	112.51	0.10
OmniCore (ours)	R	**38.27**	**0.04**
3D DiT (Mo et al., 2023)	R+T	78.55	0.08
CymbaDiff (Liang et al., 2025)	R+T	107.62	0.10
OmniCore (ours)	R+T	**30.84**	**0.04**
3D DiT (Mo et al., 2023)	R+H	42.93	0.06
CymbaDiff (Liang et al., 2025)	R+H	53.86	0.06
OmniCore (ours)	R+H	**15.19**	**0.03**
OmniCore (ours)	R+H+T	**14.35**	**0.03**

(b) Generalization on unseen datasets.

Method	C	FID ↓	MMD ↓
OmniControl-3D-KITTI-360			
3D DiT (Mo et al., 2023)	R	156.74	0.11
CymbaDiff (Liang et al., 2025)	R	234.83	0.13
OmniCore (ours)	R	**80.19**	**0.08**
OmniControl-3D-nuScenes			
3D DiT (Mo et al., 2023)	R	175.87	0.11
CymbaDiff (Liang et al., 2025)	R	280.13	0.14
OmniCore (ours)	R	**123.29**	**0.08**

(c) Ablation study of WSAB.

Method	C	FID ↓	MMD ↓
OmniControl-3D-SemanticKITTI			
Vanilla W-CA	R	73.82	0.06
Global-scope WSAB	R	63.18	0.06
WSAB + shifted window	R	47.07	0.05
WSAB (ours)	R	**38.27**	**0.04**

its cross-window information exchange and may weaken fine-grained geometric continuity across volumetric grids. Spectral modeling offers a complementary way to enhance structural coherence beyond local spatial attention. However, since 3D scenes exhibit non-stationary geometric statistics, a single global spectral transform provides only a coarse and spatially insensitive approximation. We therefore introduce scene-adaptive spectral attention within local windows, jointly encoding spatial geometry and frequency-domain context in a unified representation. Such coupling strengthens low-frequency structural cues while attenuating high-frequency noise, biasing the model toward coherent geometric layouts and reducing scattered voxel predictions.

Prior to windowing, we apply LayerNorm to latent volume and condition features, then modulate the condition by a timestep embedding $\phi(t) \in \mathbb{R}^d$ produced by a two-layer MLP: $\widetilde{\mathbf{X}} = \mathrm{LN}(\mathbf{X})$, $\widetilde{\mathbf{C}} = \mathrm{LN}(\mathbf{C}) + \phi(t)$. The latent tokens $\widetilde{\mathbf{X}}$ are then reshaped into a volumetric form and partitioned into non-overlapping $M \times M \times M$ windows, with each window denoted $\widetilde{\mathbf{X}}_w \in \mathbb{R}^{M^3 \times d}$. The condition feature $\widetilde{\mathbf{C}}$ is similarly lifted into a 3D volumetric form and partitioned identically, with each window denoted $\widetilde{\mathbf{C}}_w \in \mathbb{R}^{M^3 \times d}$, all subsequent operations are performed independently within each window. Within each window, $\widetilde{\mathbf{X}}_w$ serves as queries and $\widetilde{\mathbf{C}}_w$ as keys and values for n_h-head cross-attention. For each head $h = 1, \ldots, n_h$,

$$\mathbf{Z}_w^{(h)} = \mathrm{Softmax}\left(\frac{(\widetilde{\mathbf{X}}_w \mathbf{W}_Q^{(h)})(\widetilde{\mathbf{C}}_w \mathbf{W}_K^{(h)})^\top}{\sqrt{d_h}}\right)(\widetilde{\mathbf{C}}_w \mathbf{W}_V^{(h)}), \tag{2}$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{d \times d_h}$ are per-head projections. The head outputs are then concatenated and projected back to d dimensions by $\mathbf{Z}_w = \left[\mathbf{Z}_w^{(1)} \,\|\, \cdots \,\|\, \mathbf{Z}_w^{(n_h)}\right] \mathbf{W}_O$, $\mathbf{W}_O \in \mathbb{R}^{d \times d}$ is the output projection and $[\cdot \,\|\, \cdot]$ denotes channel-wise concatenation. Subsequently, we apply a learnable 3D DFT on $\mathbf{Z}_w \in \mathbb{R}^{M \times M \times M \times d}$ along each spatial axis,

$$\mathcal{F}_w = \mathrm{DFT3D}(\mathbf{Z}_w; \mathbf{W}_{\mathrm{DFT}}) \in \mathbb{C}^{M \times M \times M \times d}, \tag{3}$$

where $\mathbf{W}_{\mathrm{DFT}} \in \mathbb{C}^{M \times M}$ is a learnable transform matrix initialized to the canonical normalized DFT basis. To keep the transform invertible, we parametrize it on the unitary group $U(M)$ via a Cayley map that sends an unconstrained trainable matrix to a unitary factor applied to the fixed DFT base point. $\mathbf{W}_{\mathrm{DFT}} \mathbf{W}_{\mathrm{DFT}}^{*\top} = \mathbf{I}$ holds throughout training, $\mathbf{W}_{\mathrm{DFT}}^{*\top}$ in Eq. 4 is the exact inverse of Eq. 3.

To enable content-adaptive spectral modulation, we define a low-frequency band $\mathcal{K}_{\mathrm{lf}} = \{\mathbf{k} : \|\tilde{\mathbf{k}}\|_2 \leq 1\}$, where the folded frequency $\tilde{k}_i = \min(k_i, M - k_i)$ accounts for conjugate symmetry. We

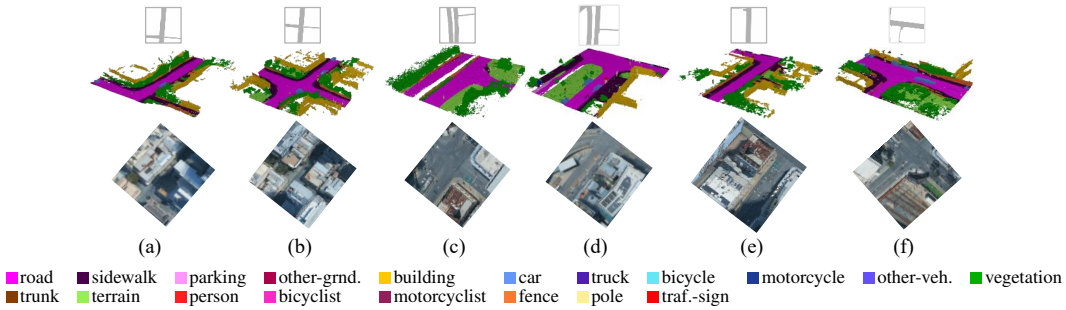

Figure 5: Zero-shot generation from in-the-wild roadmaps. Six locations (a–f) are randomly sampled from public mapping services worldwide, none seen during training. Each column shows the input roadmap (top), the scene generated by OmniCore (middle), and the corresponding satellite image (bottom) provided for reference.

Table 3: Ablation results on OmniControl-3D-SemanticKITTI. FS = full spectrum, LF = low frequency; R: roadmaps, T: texts, H: heightmaps.

(a) Ablation study of learnable W_{DFT} + LF.

Method	Cond.	FID ↓	MMD ↓
fixed W_{DFT} + FS gate	R	68.23	0.13
learnable W_{DFT} + FS gate	R	61.78	0.06
fixed W_{DFT} + LF	R	54.44	0.06
learnable W_{DFT} + LF (ours)	R	**38.27**	**0.04**

(b) Ablation study of fusion design.

Method	FID ↓ (R+T)	MMD ↓ (R+T)	FID ↓ (R+H)	MMD ↓ (R+H)
Concatenation	56.25	0.06	21.21	0.03
Cross-attention	44.52	0.05	33.54	0.04
Fusion (ours)	**30.84**	**0.04**	**15.19**	**0.03**

summarize this band by the descriptor $\bar{\mathcal{F}}_w^{\text{lf}} = |\mathcal{K}_{\text{lf}}|^{-1} \sum_{\mathbf{k}\in\mathcal{K}_{\text{lf}}} |\mathcal{F}_w[\mathbf{k},:]| \in \mathbb{R}^d$, which a bottleneck MLP maps to a channel-wise gate $\mathbf{g}_w = \sigma(\text{ReLU}(\bar{\mathcal{F}}_w^{\text{lf}}\mathbf{W}_1)\mathbf{W}_2)$, with $\mathbf{W}_1 \in \mathbb{R}^{d\times d/4}$, $\mathbf{W}_2 \in \mathbb{R}^{d/4\times d}$, and $\sigma(\cdot)$ the sigmoid. The gate is applied only outside $\mathcal{K}_{\text{lf}}$:

$$\mathbf{X}'_w = \text{Re}\big(\text{IDFT3D}(\mathbf{G}_w \odot \mathcal{F}_w;\, \mathbf{W}_{\text{DFT}}^{*\top})\big)\,, \quad \mathbf{G}_w[\mathbf{k},:] = \begin{cases} \mathbf{1}, & \mathbf{k}\in\mathcal{K}_{\text{lf}}, \\ \mathbf{g}_w, & \mathbf{k}\notin\mathcal{K}_{\text{lf}}, \end{cases} \tag{4}$$

The gate thus preserves the low-frequency band and adaptively attenuates the rest, retaining coarse structure while suppressing high-frequency noise. This band-dependent form is essential, as a frequency-shared gate would commute with the unitary transform and reduce to channel-wise rescaling. The real part is taken since the learnable basis need not preserve Hermitian symmetry. The windows are then merged and refined by a LayerNorm–MLP residual block.

5 Experiments

Implementation Details. We train all models on the OmniControl-3D-SemanticKITTI training split and evaluate on the validation splits of OmniControl-3D-SemanticKITTI, OmniControl-3D-KITTI-360, and OmniControl-3D-nuScenes. Following UrbanDiffusion (Zhang et al., 2024b), we train a dedicated encoder that jointly encodes geometry and semantics of 3D semantic voxels into a latent space, on which we compute 3D Fréchet Inception Distance (3D FID) and Maximum Mean Discrepancy (MMD) to assess generation fidelity and distributional alignment with ground-truth scenes. Additional details are provided in the Appendix A.4.

5.1 Main Results

Roadmap-Guided Scene Generation. We evaluate OmniCore for 3D semantic scene generation against four recent baselines with diverse conditioning designs: 3D DiT (Mo et al., 2023), SemCity (Lee et al., 2024), CymbaDiff (Liang et al., 2025), and PrITTI (Tze et al., 2025). As shown in Table 2 (a), OmniCore consistently outperforms all baselines on FID and MMD, indicating superior perceptual fidelity and tighter distributional alignment with ground-truth. Qualitative comparisons are provided in the left column of each group in Figure 3.

To assess cross-dataset generalization, we train all models exclusively on OmniControl-3D-SemanticKITTI and evaluate them zero-shot on OmniControl-3D-KITTI-360 and OmniControl-

3D-nuScenes, restricting evaluation to the overlapping semantic classes (16 and 9, respectively). As shown in Table 2 (b), OmniCore maintains leading performance on both target datasets despite substantial domain shift. Qualitative results in Appendix A.5 further show that it produces structurally coherent and semantically meaningful scenes. Beyond benchmark datasets, OmniCore also generalizes zero-shot to roadmaps from unseen real-world locations, generating diverse, high-quality 3D urban scenes (Figure 5; more in Appendix A.6). This demonstrates that OmniCore can serve as a scalable data engine for downstream 3D perception tasks such as 3D semantic scene completion. Detailed downstream experiments are provided in Appendix A.7.

Language-Guided Scene Generation. Under roadmap-text conditioning, OmniCore outperforms 3D DiT (Mo et al., 2023) and CymbaDiff (Liang et al., 2025) on both FID and MMD and improves over its roadmap-only counterpart (Table 2(a)), indicating that text complements the geometric layout. Figure 3 shows that the generated scenes (middle column of each group) satisfy both the road geometry imposed by the roadmap and the spatial constraints specified by the text prompt.

We further investigate OmniCore's text-driven editing capability by holding the roadmap fixed and varying only the textual prompt. As shown in the first two rows of Figure 4, the controlled category exhibits clearly different distributions across prompts while the underlying road structure remains stable, demonstrating that natural language serves as a flexible editing interface for controllable 3D semantic scene generation.

Heightmap-Guided Scene Generation. Under roadmap-heightmap conditioning, OmniCore outperforms all baselines on both FID and MMD and improves over its roadmap-only and roadmap-text counterparts (Table 2(a)), indicating that heightmaps add vertical guidance beyond the ground-plane layout. As shown in Figure 3, compared with the roadmap-only (left) and roadmap-text (middle) settings, roadmap-heightmap results (right) better preserve road topology and exhibit stronger vertical fidelity, especially for tall structures such as buildings. More qualitative results are in Appendix A.5.

To assess heightmap-based editing, we fix the roadmap and scale the heightmap by factors of 0.8 and 0.5, compressing the vertical extent of occupied regions. As shown in the last row of Figure 4, the generated scenes exhibit correspondingly reduced vertical structures while the road topology remains intact, demonstrating that heightmaps serve as a flexible interface for geometric control.

5.2 Ablation Study

We ablate the key components of OmniCore on the OmniControl-3D-SemanticKITTI validation split to assess their individual contributions. As reported in Table 2 (c), replacing WSAB with vanilla windowed cross-attention raises FID from 38.27 to 73.82, showing that spectral modulation contributes beyond attention alone. Applying the modulation globally yields an FID of 63.18, and adding shifted windows to WSAB degrades it to 47.07. These results indicate that the gain stems from confining the spectral gate to a fixed local support, rather than from the frequency-domain operation alone. Within WSAB, Table 3 (a) shows that both band-limiting the gate and making $\mathbf{W}_{\mathrm{DFT}}$ learnable contribute to performance, reducing FID from 61.78 to 54.44 and to 38.27, while the fixed W_{DFT} + FS gate lags at 68.23. This suggests that content-adaptive modulation is necessary to capture scene-dependent frequency structures. In addition, Table 3 (b) reveals that our method further improves upon concatenation and cross-attention by treating the roadmap as a dominant signal to further preserve its spatial structure. Additional ablation results are provided in Appendix A.8.

6 Conclusion

In this work, we proposed OmniCore, the first unified framework for controllable 3D scene generation that supports flexible conditioning on roadmaps, natural language, and heightmaps, enabling region-specific editing and high-fidelity 3D semantic scene synthesis, such as building or vegetation distribution. OmniCore achieves state-of-the-art performance across all conditioning configurations, with strong flexibility and spatial precision for user-controllable 3D scene generation. Beyond generation, it serves as a scalable data engine for downstream perception tasks, such as 3D semantic scene completion. To facilitate systematic evaluation, we further introduced OmniControl-3D, a large-scale multi-modal benchmark comprising three subsets, OmniControl-3D-SemanticKITTI, -KITTI-360, and -nuScenes, each providing 3D semantic scenes paired with roadmaps, textual descriptions, and heightmaps. OmniCore and OmniControl-3D establish a unified foundation for controllable 3D scene generation and a scalable pipeline for advancing 3D perception in the wild.

AI USE STATEMENT

We used GPT-4 to generate the textual scene descriptions in OmniControl-3D from ground-truth 3D scenes using prompt templates; all generated descriptions follow a constrained format and vocabulary and were checked by the authors. GPT-4 was also used to re-caption generated scenes for the text-alignment metric (Appendix A.10). We also used large language models to polish the grammar and wording of the manuscript. We did not use generative AI tools to propose research ideas, design the method or experiments, implement the method, or interpret results. The authors reviewed all AI-assisted content and take full responsibility for the paper.

ETHICS STATEMENT

OmniControl-3D is built upon SemanticKITTI (Behley et al., 2019), SSCBench-KITTI-360, and SSCBench-nuScenes (Li et al., 2023a) and follows their non-commercial licenses (CC BY-NC-SA). We will release the conditioning signals (roadmaps, heightmaps, and textual descriptions) paired with 3D semantic scenes, but not raw sensor data. Roadmaps are released under the ODbL with attribution to OpenStreetMap contributors. The data contain no personally identifiable information, and the benchmark and code are intended for research use only.

REPRODUCIBILITY STATEMENT

A subset of OmniControl-3D is provided via an anonymous link; the full benchmark and code will be released upon acceptance. Dataset construction and model details are described in Sec. 3 and Sec. 4, with implementation details, metrics, and evaluation protocols in Appendix A.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Luis B Almeida. The fractional fourier transform and time-frequency representations. *IEEE Transactions on signal processing*, 42(11):3084–3091, 1994.

Tongyuan Bai, Wangyuanfan Bai, Dong Chen, Tieru Wu, Manyi Li, and Rui Ma. Freescene: Mixed graph diffusion for 3d scene synthesis from free prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5893–5903, 2025.

N. Beaudoin and S.S. Beauchemin. An accurate discrete fourier transform for image processing. In *2002 International Conference on Pattern Recognition*, volume 3, pp. 935–939 vol.3, 2002. doi: 10.1109/ICPR.2002.1048189.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4413–4421, 2018.

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2724–2733, 2025.

Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. *arXiv preprint arXiv:2410.18084*, 2024.

Aleksey Bokhovkin, Quan Meng, Shubham Tulsiani, and Angela Dai. Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 628–639, 2025.

Nicolas Bougie and Narimawa Watanabe. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 215–229, 2025.

Tim-Oliver Buchholz and Florian Jug. Fourier image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1846–1854, 2022.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Pingping Cai, Deja Scott, Xiaoguang Li, and Song Wang. Orthogonal dictionary guided shape completion network for point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 864–872, 2024.

R. Chellappa and R. Bagdazian. Fourier coding of image boundaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1):102–105, 1984. doi: 10.1109/TPAMI.1984.4767482.

Xingyu Chen, Zihao Feng, Kun Qian, and Xinyu Zhang. Radio frequency ray tracing with neural object representation for enhanced rf modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21339–21348, 2025a.

Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, and Xingang Pan. Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28371–28382, 2025b.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Tianyi Chu, Jiafu Chen, Jiakai Sun, Shuobin Lian, Zhizhong Wang, Zhiwen Zuo, Lei Zhao, Wei Xing, and Dongming Lu. Rethinking fast fourier convolution in image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23195–23205, 2023.

Ruikai Cui, Weizhe Liu, Weixuan Sun, Senbo Wang, Taizhang Shang, Yang Li, Xibin Song, Han Yan, Zhennan Wu, Shenzhou Chen, et al. Neusdfusion: A spatial-aware generative model for 3d shape completion, reconstruction, and generation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Wei Deng, Mengshi Qi, and Huadong Ma. Global-local tree search in vlms for 3d indoor scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8975–8984, 2025.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12124–12134, 2022.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems*, 37:40212–40229, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Xiyue Guo, Jiarui Hu, Junjie Hu, Hujun Bao, and Guofeng Zhang. Sgformer: Satellite-ground fusion for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11929–11938, 2025.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. Roomtour3d: Geometry-aware video-instruction tuning for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27586–27596, 2025a.

Woojung Han, Yeonkyung Lee, Chanyoung Kim, Kwanghyun Park, and Seong Jae Hwang. Spatial transport optimization by repositioning attention map for training-free text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18401–18410, 2025b.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6185–6194, 2023.

Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya Wen, Meng Wu, Cong Hu, Bingliang Hu, et al. beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30839–30849, 2025.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pp. 289–310. Springer, 2024.

Weinan Jia, Mengqi Huang, Nan Chen, Lei Zhang, and Zhendong Mao. Dˆ 2it: Dynamic diffusion transformer for accurate image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12860–12870, 2025.

Jakob Kruse, Carsten Rother, and Uwe Schmidt. Learning to push the limits of efficient fft-based image deconvolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4586–4594, 2017.

Jumin Lee, Woobin Im, Sebin Lee, and Sung-Eui Yoon. Diffusion probabilistic models for scene-scale 3d categorical data. *arXiv preprint arXiv:2301.00527*, 2023.

Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28337–28347, 2024.

Ruihuang Li, Zhengqiang Zhang, Chenhang He, Zhiyuan Ma, Vishal M Patel, and Lei Zhang. Dense multimodal alignment for open-vocabulary 3d scene understanding. In *European Conference on Computer Vision*, pp. 416–434. Springer, 2024.

Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: Monocular 3d semantic scene completion benchmark in street views. 2023a.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9087–9098, 2023b.

Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17182–17191, 2022.

Li Liang, Bo Miao, Xinyu Wang, Naveed Akhtar, Jordan Vice, and Ajmal Mian. Cymbadiff: Structured spatial diffusion for sketch-based 3d semantic urban scene generation. *arXiv preprint arXiv:2510.13245*, 2025.

Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):3292–3310, 2022.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14928–14937, 2024.

Ninghe Liu, Kexuan Liu, Yixin Yang, Yifan Peng, and Liangcai Cao. Propagation-adaptive 4k computer-generated holography using physics-constrained spatial and fourier neural operator. *Nature Communications*, 16(1):7761, 2025a.

Xiaoyi Liu and Hao Tang. Difffno: Diffusion fourier neural operator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 150–160, 2025.

Yuheng Liu, Xinke Li, Xueting Li, Lu Qi, Chongshou Li, and Ming-Hsuan Yang. Pyramid diffusion for fine 3d large scene generation. In *European Conference on Computer Vision*, pp. 71–87. Springer, 2024.

Yuheng Liu, Xinke Li, Yuning Zhang, Lu Qi, Xin Li, Wenping Wang, Chongshou Li, Xueting Li, and Ming-Hsuan Yang. Controllable 3d outdoor scene generation via scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28052–28062, 2025b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.

Xiaoyu Luo, Huan Wang, Te Han, and Ying Zhang. Fft-trans: Enhancing robustness in mechanical fault diagnosis with fourier transform-based transformer under noisy conditions. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024.

Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *Advances in neural information processing systems*, 36:67960–67971, 2023.

Tan Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. *Advances in Neural Information Processing Systems*, 35:29319–29335, 2022.

Lucas Nunes, Rodrigo Marcuzzi, Jens Behley, and Cyrill Stachniss. Towards generating realistic 3d semantic training data for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier transform and convolution algorithms*, pp. 80–111. Springer, 1981.

Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoğlu, and Theo Gevers. Sceneteller: Language-to-3d scene generation. In *European Conference on Computer Vision*, pp. 362–378. Springer, 2024.

Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7463–7472, 2021.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in neural information processing systems*, 34:12013–12026, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4988, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Duc-Hai Pham, Duc-Dung Nguyen, Anh Pham, Tuan Ho, Phong Nguyen, Khoi Nguyen, and Rang Nguyen. Semi-supervised 3d semantic scene completion with 2d vision foundation model guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6514–6522, 2025.

Wentao Qu, Guofeng Mei, Yang Wu, Yongshun Gong, Xiaoshui Huang, and Liang Xiao. A self-conditioned representation guided diffusion model for realistic text-to-lidar scene generation. *arXiv preprint arXiv:2511.19004*, 2025.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.

B Srinivasa Reddy and Biswanath N Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, 5(8):1266–1271, 1996.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Yi Rong, Haoran Zhou, Lixin Yuan, Cheng Mei, Jiahao Wang, and Tong Lu. Cra-pcn: Point cloud completion with intra-and inter-level cross-resolution transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4676–4685, 2024.

Tourangbam Harishore Singh, Po-Tsang Huang, Kung-Shuo Kao, Chih-Shiang Cheng, Kuei-Ann Wen, and Li-Chun Wang. Energy-efficient sparse fft and compressed transpose memory for mmwave fmcw radar sensor system. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23765–23774, 2025.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9121–9130, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Nhat Thanh Tran and Jack Xin. Fourier-mixed window attention for efficient and robust long sequence time-series forecasting. *Frontiers in Applied Mathematics and Statistics*, 11:1600136, 2025.

Christina Ourania Tze, Daniel Dauner, Yiyi Liao, Dzmitry Tsishkou, and Andreas Geiger. Pritti: Primitive-based generation of controllable and editable 3d semantic urban scenes. *arXiv preprint arXiv:2506.19117*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. Dsvt: Dynamic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13520–13529, 2023.

Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5722–5730, 2024.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6030–6038, 2024a.

Xianzu Wu, Xianfeng Wu, Tianyu Luan, Yajing Bai, Zhongyuan Lai, and Junsong Yuan. Fsc: Few-point shape completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26077–26087, 2024b.

Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8929–8939, 2023.

Zijie Wu, Mingtao Feng, Yaonan Wang, He Xie, Weisheng Dong, Bo Miao, and Ajmal Mian. External knowledge enhanced 3d scene generation from sketch. In *European Conference on Computer Vision*, pp. 286–304. Springer, 2024c.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Zipeng Xiao, Siqi Kou, Zhongkai Hao, Bokai Lin, and Zhijie Deng. Amortized fourier neural operators. *Advances in Neural Information Processing Systems*, 37:115001–115020, 2024.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14383–14392, 2021.

Yujie Xue, Huilong Pi, Jiapeng Zhang, Yunchuan Qin, Zhuo Tang, Kenli Li, and Ruihui Li. Sdformer: Vision-based 3d semantic scene completion via sam-assisted dual-channel voxel transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26837–26847, 2025.

Xingyi Yang, Songhua Liu, and Xinchao Wang. Hash3d: Training-free acceleration for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21481–21491, 2025a.

Xuemeng Yang, Hao Zou, Xin Kong, Tianxin Huang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Semantic segmentation-assisted scene completion for lidar point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3555–3562. IEEE, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4085–4095, 2020.

Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *arXiv preprint arXiv:2506.13558*, 2025b.

Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14905–14915, 2024.

Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:1531–1555, 2024.

Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying N Wu, and Dongfang Liu. Visual fourier prompt tuning. *Advances in Neural Information Processing Systems*, 37:5552–5585, 2024.

Jingwen Zhang, Zikun Zhou, Guangming Lu, Jiandong Tian, and Wenjie Pei. Robust 3d tracking with quality-aware shape completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 7160–7168, 2024a.

Junge Zhang, Qihang Zhang, Li Zhang, Ramana Rao Kompella, Gaowen Liu, and Bolei Zhou. Urban scene diffusion through semantic occupancy map. *arXiv preprint arXiv:2403.11697*, 2024b.

Songchun Zhang, Yibo Zhang, Quan Zheng, Rui Ma, Wei Hua, Hujun Bao, Weiwei Xu, and Changqing Zou. 3d-scenedreamer: Text-driven 3d-consistent scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10170–10180, 2024c.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Yu Zheng, Fengli Xu, Yuming Lin, Paolo Santi, Carlo Ratti, Qi R Wang, and Yong Li. Urban planning in the era of large language models. *Nature computational science*, 5(9):727–736, 2025.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20852–20862, 2025.

A Appendix

A.1 OmniControl-3D Data Format

Roadmaps are stored as `.png` images and heightmaps as binary (`.bin`) files with integer values in $[0, 31]$. Textual descriptions are stored in `.txt` format. The 3D ground truth follows the SemanticKITTI (Behley et al., 2019) volumetric label format, where each voxel is encoded as a 16-bit unsigned integer corresponding to its semantic class.

A.2 Evaluation Metrics

To comprehensively assess the quality and diversity of the generated 3D semantic scenes, we adopt two widely used metrics: the Fréchet Inception Distance (FID) (Paschalidou et al., 2021) and the Maximum Mean Discrepancy (MMD) (Zhang et al., 2024b). These two metrics are complementary: FID captures distributional similarity under a Gaussian assumption, while MMD provides a non-parametric measure of feature-level discrepancy, together yielding a comprehensive evaluation of generative performance.

Fréchet Inception Distance. FID quantifies the similarity between the distributions of generated and real samples in a latent feature space by modeling each as a multivariate Gaussian. It is formally defined as

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2\,(\Sigma_r\Sigma_g)^{\frac{1}{2}}\Big)\,, \tag{5}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) denote the means and covariance matrices of the feature distributions extracted from real and generated samples, respectively. A lower FID indicates that the generated distribution is closer to the real one.

Maximum Mean Discrepancy. MMD is a non-parametric, kernel-based metric that measures the distance between two probability distributions without assuming any specific parametric form. This property makes it particularly suitable for evaluating generative models under flexible distributional assumptions. In our setting, MMD is computed with a Gaussian kernel over features extracted from the same latent space as FID, and is defined as

$$\mathrm{MMD}^2(X, Y) = \mathbb{E}_{x,x'}[k(x, x')] + \mathbb{E}_{y,y'}[k(y, y')] - 2\,\mathbb{E}_{x,y}[k(x, y)]\,, \tag{6}$$

where $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ are the sets of latent features extracted from real and generated 3D scenes, respectively, $k(\cdot, \cdot)$ denotes the Gaussian kernel, and x, x' (resp. y, y') are independent samples drawn from X (resp. Y). A smaller MMD value indicates a closer match between the two distributions.

A.3 OmniCore Training Objective

Following the latent diffusion model (LDM) framework (Rombach et al., 2022), we train the OmniCore denoising network by minimizing the expected mean squared error between the predicted and ground-truth noise:

$$\mathcal{L}_{\mathrm{LDM}} = \mathbb{E}_{x,\,\epsilon\sim\mathcal{N}(0,\mathbf{I}),\,t}\Big[\,\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\,\Big]\,, \tag{7}$$

where $t \in \{1, \ldots, T\}$ is uniformly sampled from the diffusion time steps, x_t denotes the noisy latent obtained by progressively perturbing the clean input x with Gaussian noise according to a predefined schedule, and $\epsilon_\theta(\cdot, \cdot)$ is the denoising network parameterized by θ, which is trained to predict the noise component ϵ injected at step t.

A.4 Additional Implementation Details

All experiments were conducted on two NVIDIA A100 GPUs, with peak memory consumption of $\sim$30 GiB and $\sim$40 GiB per GPU during VAE and OmniCore denoising network training, respectively. Two separate VAEs were trained: one dedicated to *large-scale categories* only (road, parking, sidewalk, other-ground, building, fence, vegetation, terrain) over 92 epochs, and the other additionally incorporating *small categories* (car, bicycle, motorcycle, truck, other-vehicle, person, bicyclist, motorcyclist, trunk, pole, traffic-sign) over 43 epochs. Both VAEs and the OmniCore denoising network were optimized with AdamW using a batch size of 20, where VAE training employed a WarmupCosineLR scheduler to ensure stable convergence. The OmniCore denoiser was trained for 600 epochs with a fixed learning rate of 1×10^{-4}, a weight decay of 1×10^{-6}, and 100 denoising steps.

A.5 Qualitative Results on 3D Semantic Scene Generation

To further demonstrate the effectiveness of our proposed framework, we present additional qualitative results in Figures 6, which show representative examples randomly selected from the OmniControl-3D-SemanticKITTI validation set. Besides, the city-scale 3D semantic scene generation results are presented in Figures 7. These examples illustrate that our model can generate semantically meaningful and spatially coherent 3D scenes across different datasets.

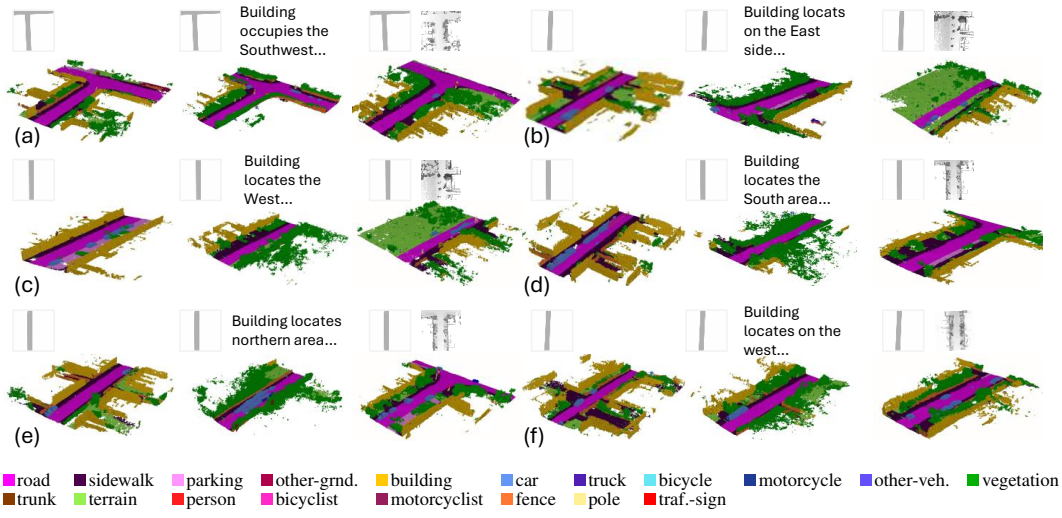

Figure 6: Qualitative results of OmniCore under different conditioning configurations. Columns left-to-right of each group (a–f) correspond to roadmap-only, roadmap-text, and roadmap-heightmap conditioning.

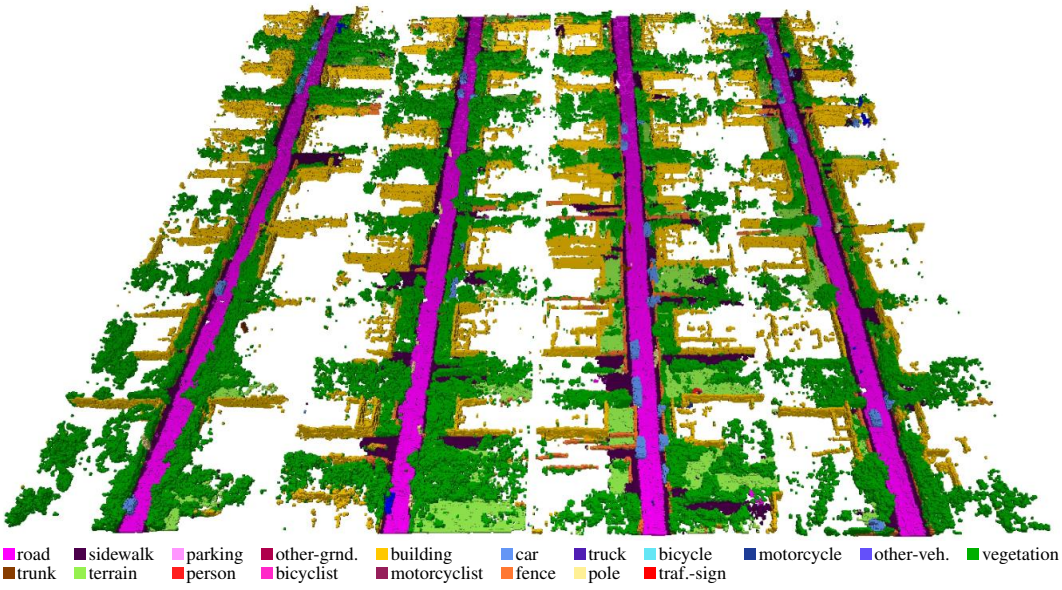

Figure 7: Qualitative result of city-scale 3D semantic scene generation.

A.6 Zero-Shot Generalization to Unseen Datasets and Real-World Scenarios

To further demonstrate the effectiveness and generalization of our OmniCore, we present additional qualitative results in Figures 8, 9, and 10. Our model is trained *exclusively* on OmniControl-3D-SemanticKITTI and directly evaluated on these datasets or real-world data in a **zero-shot** manner, without any fine-tuning.

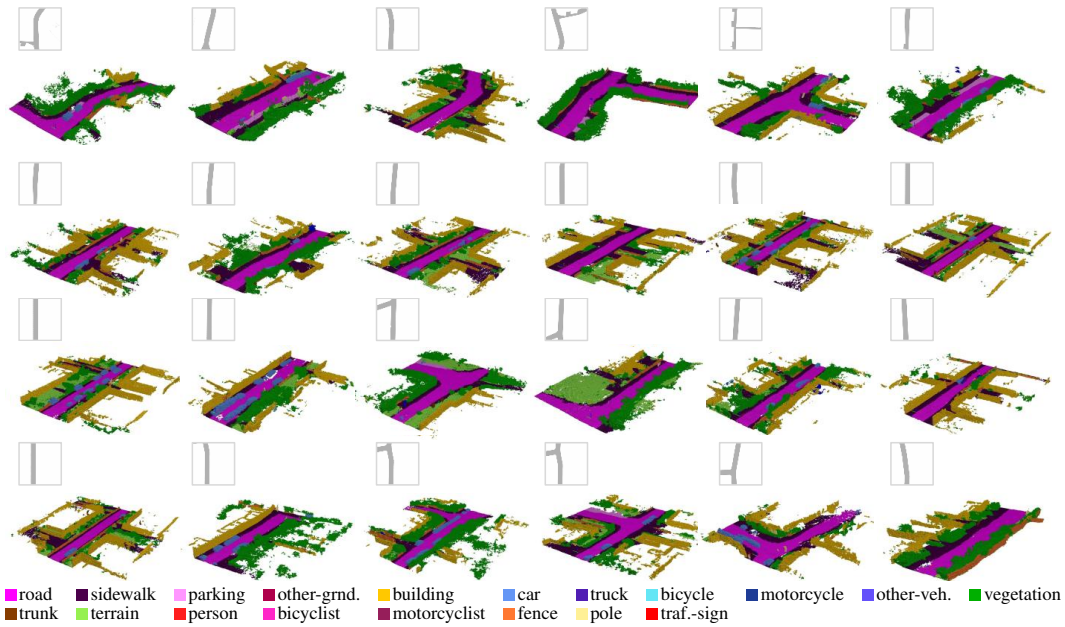

Figure 8: Qualitative results on OmniControl-3D-KITTI-360.

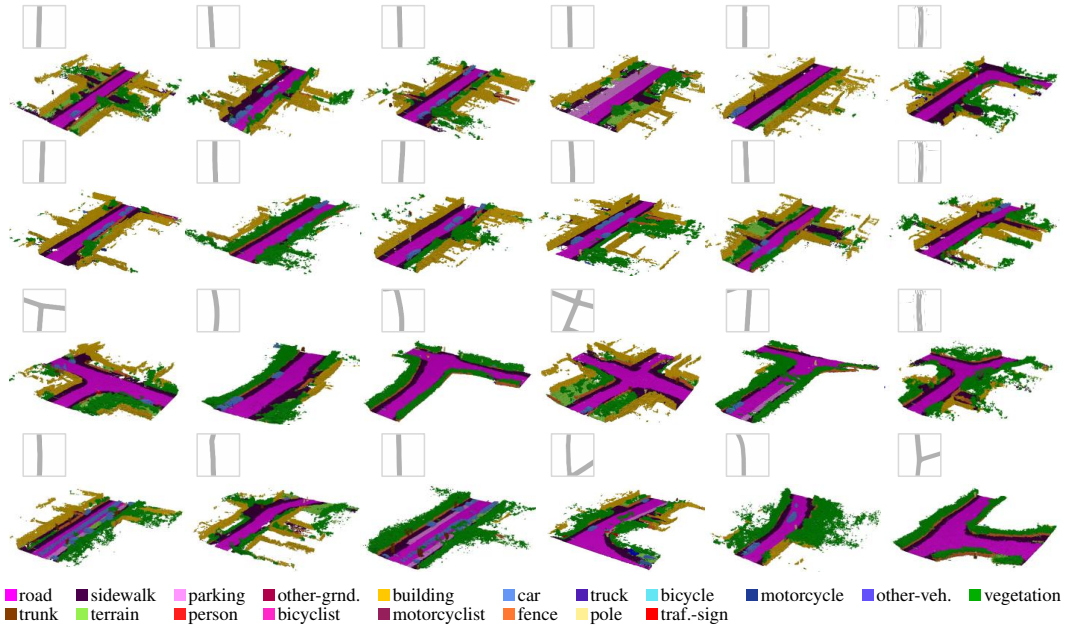

Figure 9: Qualitative results on OmniControl-3D-nuScenes.

Specifically, Figures 8 and 9 present representative samples randomly drawn from the validation sets of OmniControl-3D-KITTI-360 and OmniControl-3D-nuScenes, respectively, while Figure 10 showcases results conditioned on roadmaps randomly collected from public mapping services worldwide. These results demonstrate that our model generates semantically meaningful and spatially coherent 3D scenes from arbitrary roadmap inputs, exhibiting strong cross-dataset generalization and robust performance in real-world scenarios.

A.7 Downstream Evaluation

To verify that our generated scenes benefit downstream tasks, we evaluate three training settings for 3D semantic scene completion with SSA-SC (Yang et al., 2021) as the downstream model, as shown

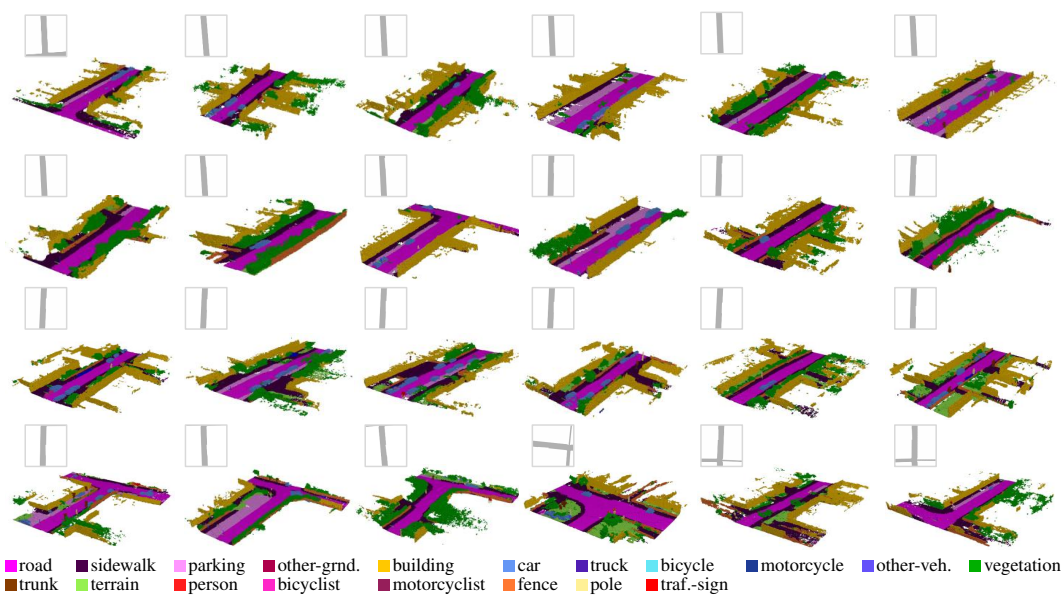

Figure 10: Qualitative results on in-the-wild roadmaps.

Table 4: Semantic scene completion with SSA-SC under three training sources, all evaluated on the real SemanticKITTI validation split.

Training source	IoU ↑	mIoU ↑
Real-only (released)	58.3	24.5
Syn-only	55.3	21.5
Real+Syn	**59.8**	**25.6**

Table 5: Boundary continuity under roadmap-only conditioning on the OmniControl-3D-SemanticKITTI validation set.

Method	BDR	BDR (offset 32)	ΔBDR
Ground Truth	1.0182	1.0112	0.0070
Vanilla W-CA	1.1103	1.0836	0.0267
WSAB + shifted window	1.0599	1.0458	0.0141
WSAB (ours)	**1.0373**	**1.0278**	**0.0095**

in Table 4. *Real-only* is the officially released SSA-SC checkpoint. *Syn-only* is trained from scratch on 3,834 scenes sampled from OmniControl-3D-SemanticKITTI, exactly matching the size of the original SemanticKITTI training set. *Real+Syn* is trained on the 3,834 real frames together with the 3,834 generated scenes. All settings share identical backbones, optimizers, schedules, epochs, and random seeds, use the official SSA-SC configuration with batch size 2 for 54 epochs, and are evaluated on the same untouched real validation split (seq-08).

Table 6: Ablation on the text encoder.

Text encoder	Cond.	FID ↓	MMD ↓
BERT-base	R+T	31.47	0.04
CLIP ViT-B/32	R+T	32.56	0.04
OpenCLIP ViT-B/32 (default)	R+T	**30.84**	**0.04**

Table 7: Ablation on the roadmap–heightmap encoder.

Condition encoder	Cond.	FID ↓	MMD ↓
ResNet-18	R+H	24.32	0.04
Lightweight 2D CNN (default)	R+H	**15.19**	**0.03**

A.8 Ablation Results on the Condition Embedding Architecture

We provide two tables below that ablate both condition-embedding components on OmniControl-3D-SemanticKITTI under identical training settings, replacing only the corresponding encoder in each case: (i) the text encoder, verifying the choice of pretrained language model, and (ii) the roadmap-heightmap encoder, verifying the effectiveness of the designed lightweight 2D CNN.

For the text encoder in Table 6, the three variants perform closely despite different pretraining, showing that our framework is insensitive to this choice. For the roadmap-heightmap condition encoder in Table 7, our lightweight 2D CNN clearly outperforms ResNet-18, whose aggressive stem downsampling erodes the thin road boundaries and per-coordinate heights that these sparse, low-

texture signals depend on. This confirms that the encoder should be matched to the spatial precision each condition carries, which is what our lightweight design provides.

A.9 Boundary Continuity Analysis

To assess boundary continuity directly, we evaluate all variants on the OmniControl-3D-SemanticKITTI validation set (815 samples) under roadmap-only conditioning, varying only the block. With a downsampling factor of 4, latent patch size $4 \times 4 \times 2$, and window size $M = 4$, each window spans a 64-voxel period in the decoded $256 \times 256 \times 32$ grid, placing boundary planes at $x, y \in \{64, 128, 192\}$; the vertical axis holds a single window and is excluded. The Boundary Discontinuity Ratio (BDR) is the semantic-label disagreement rate between non-empty voxel pairs straddling such a plane, normalised by neighbouring interior planes, so $\mathrm{BDR} = 1$ means boundary and interior are indistinguishable. Shifting the planes by half a period gives a control, and $\Delta\mathrm{BDR}$ isolates the partition-specific excess (Table 5). Against a ground-truth reference of 0.0070, WSAB reduces this excess from 0.0267 to 0.0095, ahead of shifted windows at 0.0141.

Table 8: Condition alignment on OmniControl-3D-SemanticKITTI. Each metric is compared against a shuffled-pairing chance level and an oracle computed on ground-truth scenes.

(a) Roadmap IoU.

Setting	Roadmap IoU ↑
Shuffled pairing (chance)	0.103
R-only	0.485
GT scene (oracle)	0.503

(b) Text Jaccard index.

Setting	Text Jaccard ↑
Shuffled pairing (chance)	0.132
R-only	0.211
R+T	0.595
GT scene (oracle)	0.850

(c) Heightmap MAE.

Setting	Heightmap MAE ↓
Shuffled pairing (chance)	5.344
R-only	4.095
R+H	2.146
GT scene (oracle)	0.0

A.10 Condition Alignment Evaluation

To directly assess how faithfully a generated scene follows its conditioning signal, we introduce one alignment metric per conditioning modality: roadmap IoU for roadmaps (R), heightmap MAE for heightmaps (H), and text Jaccard index for text (T). We report each metric under three settings. *Shuffled pairing* computes each metric on randomly shuffled condition-scene pairs and thus reflects the chance level; *R-only* (and likewise R+T and R+H) evaluates the scenes generated by our method; and *GT scene (oracle)* evaluates each ground-truth scene against its own condition, providing an upper bound for each metric. All results are reported on OmniControl-3D-SemanticKITTI, as shown in Table 8.

The evaluation metrics are defined below. *Roadmap IoU* in Table 8 (a): IoU between the road mask in the generated scene's BEV and the input roadmap. The oracle stays well below 1.0 because roadmaps are fixed-width vector renderings, whereas ground-truth road voxels are limited by LiDAR coverage and occlusion; it therefore marks the attainable ceiling rather than a perfect reference. *Text Jaccard* in Table 8 (b): We re-caption the generated scene with the same prompt template and compute the Jaccard index between the new caption and the input text over shape and location terms, where a term counts as matched if it or a synonym appears. *Heightmap MAE* in Table 8 (c): We extract a heightmap from the generated scene and compute the mean absolute error, in voxel units, against the input heightmap.

Table 8 shows that all three conditions exert direct control over the generated scenes. The roadmap IoU of our generations (0.485) is close to the oracle (0.503) and far above chance (0.103), indicating that the road layout is reproduced nearly as faithfully as the ground-truth scene permits. For text, R-only already exceeds chance (0.211 vs. 0.132), since layouts are partly predictable from road structure alone. Adding the text condition under an identical roadmap nearly triples the similarity to 0.595, closing much of the gap to the oracle (0.850). This isolates the contribution of text to scene semantics. For heightmaps, roadmap-only conditioning reduces the error below chance (4.095 vs. 5.344), and adding the heightmap halves it again to 2.146.

A.11 VAE Results

Our OmniCore denoising network operates in the latent space of a VAE, whose reconstruction quality directly determines the upper bound of semantic scene generation. While sharing the same architecture, the VAE is trained under two separate configurations on SemanticKITTI: one dedicated to

Table 9: VAE reconstruction performance on the SemanticKITTI validation set. VAE-large is trained on large-scale categories, while VAE-full additionally incorporates small categories.

Model	Epochs	Batch size	IoU	mIoU
VAE-large	92	20	98.6	97.4

Model	Epochs	Batch size	IoU	mIoU
VAE-full	43	20	94.8	92.9

large-scale categories only, and the other additionally incorporating small categories. Table 9 reports the reconstruction performance of both configurations on the OmniControl-3D-SemanticKITTI validation set.

A.12 Broader Impacts

OmniCore is inherently neutral and intended for beneficial applications such as urban simulation and autonomous driving. As with other generative models, however, it could pose societal risks if misused, for example, in unauthorized mass surveillance.

A.13 Limitations

While OmniCore supports flexible conditioning on automatically extracted roadmaps, natural language, and heightmaps for controllable 3D scene generation, its practical utility for interactive scene authoring could be further enhanced by incorporating human-provided conditioning signals such as hand-drawn roadmaps. Future work will explore human-in-the-loop conditioning and extend OmniCore to a broader range of user-supplied control signals.

A.14 Licenses

Licenses of SemanticKITTI, SSCBench KITTI-360, SSCBench nuScenes. The SemanticKITTI and SSCBench nuScenes datasets are licensed under CC BY-NC-SA 4.0, while SSCBench KITTI-360 is released under CC BY-NC-SA 3.0.

Terms of Use and License of OmniControl-3D dataset. The roadmaps in our released dataset are under the Open Database License (ODbL) with attribution to OpenStreetMap contributors. All components of the OmniControl-3D dataset other than the roadmaps are released under CC BY-NC-SA 4.0, following the licenses of their source datasets.