Pyramid Diffusion for Fine 3D Large Scene Generation - Supplementary Material

This supplementary document details our evaluation setup, hyperparameters setting, data pre-processing, and more experimental results. The supplementary material is organized as follows:

A Evaluation Setup

In all evaluations of our paper, we consistently randomly sample 1,000 distinct scenes to assess the generation quality. The specific methodologies for F3D (Fréchet 3D Distance) and MMD (Maximum Mean Discrepancy) are as follows.

A.1 F3D

It is an evaluation metric adapted from the 2D Fréchet Inception Distance (FID) [5] to evaluate the quality of generated 3D scenes. Implementing F3D aims to complement semantic segmentation by capturing the richness and diversity of generated scenes, which semantic segmentation might overlook. F3D ensures that the generated scenes maintain complexity and reflect the similarity between generated 3D scenes and real-world structures.

Our F3D is calculated following the next steps. Initially, we pre-train a 3D CNN-based autoencoder, which is subsequently utilized to extract high-dimensional features from the generated 3D scenes. The F3D is then computed akin to FID, leveraging the extracted features to evaluate the discrepancies between the generated and real scenes. Mathematically, F3D is represented as:

$$\text{F3D} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|^2 + \mathbf{Tr}\left(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_r - 2(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_r)^{1/2}\right) \tag{1}$$

where $\boldsymbol{\mu}_g$, $\boldsymbol{\mu}_r$ are the feature means, and $\boldsymbol{\Sigma}_g$, $\boldsymbol{\Sigma}_r$ are the feature covariances of the generated and real scenes.

A.2 MMD

We incorporate the Maximum Mean Discrepancy (MMD) as a key statistical measure to quantify the disparity between the distributions of generated and real-world scenes. Following a method akin to our F3D approach, we initially extract high-dimensional features from the 3D scenes using a 3D CNN architecture which is used in F3D. Subsequently, we employ a Gaussian kernel, expressed as

$$k(\boldsymbol{f}, \boldsymbol{f}') = \exp\left(-\frac{\|\boldsymbol{f} - \boldsymbol{f}'\|^2}{2\sigma^2}\right) \tag{2}$$

where σ is the kernel width to map these features into a higher-dimensional space for MMD calculation. The bandwidth σ is determined using the median heuristic, a robust method to estimate the scale of data in the feature space. The MMD formula is given by

$$\mathbf{MMD}^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} k(\boldsymbol{f}_i, \cdot) - \frac{1}{m} \sum_{j=1}^{m} k(\boldsymbol{f}'_j, \cdot) \right\|^2 \tag{3}$$

$\boldsymbol{f}$ and $\boldsymbol{f}'$ are the extracted features from the generated and real dataset, respectively. The utilization of MMD, especially with the Gaussian kernel, not only captures the overall statistical distribution but also considers finer details in the feature space, rendering it an indispensable tool in our evaluation protocol.

B Hyperparameters Setting

B.1 Pyramid Discrete Diffusion Models

Training Hyperparameters. In the main experiment of our Pyramid Discrete Diffusion, a total of four diffusion models, namely PDD (s_1), PDD (s_2), PDD (s_3), and PDD (s_4), are utilized. Each model is trained on four NVIDIA A100 GPUs with batch sizes set to 128, 32, 16, and 8, respectively. A unified learning rate of 10^{-3} is applied and the AdamW optimizer is used to train each model for 800 epochs. Additionally, during training, data augmentation techniques, including flipping and rotation, are employed to enhance the robustness of the models.

Cross-dataset Hyperparameters. In our paper, we demonstrate the capability of our method for cross-dataset generation. The models labeled with *FT* are those where the generation model trains on the CarlaSC dataset [12] and undergoes finetuning using the SemanticKITTI dataset [2]. Specifically, we use the PDD (s_4) model for finetuning, which has already been trained for 800 epochs on CarlaSC. This model is then further trained for 200 epochs using the SemanticKITTI dataset, with all other hyperparameters remaining unchanged. Consequently, the entire model completes a total of 1,000 epochs of training.

Sampling Hyperparameters. In our paper, all the sampled scenes we demonstrate are generated using 100 diffusion steps.

B.2 Evaluation Models

We train six different models across three distinct networks in our evaluation process. These include 3D CNNs for evaluating Fréchet 3D Distance (F3D) and

Maximum Mean Discrepancy (MMD), SparseUNet [4], and PointNet++ [10] for semantic segmentation evaluation.

3D CNN Hyperparameters. As mentioned in Section A of our paper, the evaluation of F3D and MMD relies on a well-trained 3D CNN network. We train two separate 3D CNNs for the CarlaSC and SemanticKITTI datasets to perform their respective F3D and MMD evaluations but maintain identical training parameters for both networks. The 3D CNNs are trained as an autoencoder, with the primary aim of feature extraction. The loss function employed in this training is a balanced cross-entropy, a form of reconstruction loss. We set the batch size to 16 and employed SGD as the optimizer with a cosine scheduler. The networks are trained for 30 epochs at a 10^{-2} learning rate. We adhere to the original dataset splits for training, using the training sets as delineated in the original dataset distributions.

SparseUNet Hyperparameters. SparseUNet [4], designed for voxel-based semantic segmentation, is trained with two separate models for the CarlaSC and SemanticKITTI datasets. All training parameters of these SparseUNet models are consistent with those of the 3D CNNs. This includes a batch size 16, utilizing SGD as the optimizer, employing a cosine scheduler, and training for 30 epochs at a learning rate of 10^{-2}. The training adheres to the original dataset divisions for the respective training sets.

PointNet++ Hyperparameters. Similarly to our previous approach, we train a separate semantic segmentation model based on PointNet++ (SSG) for each dataset. Each model is trained with a batch size of 16. We randomly sample 16,384 points for each scene, and if the number of points is insufficient, we supplement them by replication. We set the optimizer to Adam and use a step scheduler with a step size 5. The models are trained for 30 epochs at a learning rate of 5×10^{-3}.

B.3 Baseline

Unconditional Generation. In our paper's unconditional generation comparison experiments, we include two baseline models: the Discrete Diffusion model (DD) [1] and the Latent Diffusion model [7]. For the DD, following the descriptions in [7], we train generation models with a scale of $256 \times 256 \times 16$ for both the CarlaSC and SemanticKITTI datasets. Regarding the Latent Diffusion model, we utilize the pre-trained model published in the work [7] as our baseline. However, since the original work trains the model at a scale of $128 \times 128 \times 8$, during the generation phase, we use the Scale Adaptive Function to upsample the generated scenes to match our scale of $256 \times 256 \times 16$.

Conditional Generation. In our paper's conditional generation experiment sections, we use a discrete diffusion model conditioned on point clouds as our baseline. The scene is divided into a voxel-based representation based on scale, and each voxel is assigned a binary value (0 or 1). A voxel is set to 1, indicating the presence of points if it contains one or more points from the point cloud.

We recognize that this comparison might not be entirely fair due to scale differences. However, the final upscaling and subsequent visualization of the gen-

Table a: Simplified Semantic Labels for the CarlaSC Dataset After Merging. The table lists the 10 consolidated classes used in our experiments, with *0* denoting unclassified elements not shown here.

Index	Label	Index	Label
1	Building	6	Road
2	Fences	7	Ground
3	Other	8	Sidewalk
4	Pedestrian	9	Vegetation
5	Pole	10	Vehicle

erated scenes allow for a clear discernment of quality differences. It's important to note that training in the original method at the larger scale of 256 × 256 × 16 would require substantial computational resources, specifically over 16 days on 4 A100 GPUs. Despite this potential bias, the visual evaluation method, after upscaling, effectively demonstrates the quality distinction between the generated scenes, which is crucial for our evaluation. This approach is chosen as the most feasible solution given our resource constraints.

C Datasets Pre-processing

Our paper's experiments utilize two outdoor scene datasets: CarlaSC [12] and SemanticKITTI [2]. CarlaSC is a dataset collected through simulated road scenes and is primarily used in our main experiments. On the other hand, SemanticKITTI, gathered from real-world scenes, is employed in experiments focusing on cross-dataset applications. Due to the different origins of these datasets and their varied labels, we undertake specific processing steps to facilitate better experimentation. The details of these processing steps are as follows.

C.1 CarlaSC

CarlaSC [12], primarily employed in our main experiments, is a synthetic dataset featuring outdoor road point cloud scenes. Originally comprising 23 semantic labels, these merge according to the dataset's official guidelines to simplify the categorization process. The dataset encompasses 11 semantic classes, as detailed in Table a, with *0* representing the *unclassified* category. This dataset includes 18 scenes for training, 3 for validation, and 3 for testing. In our experiments, we utilize a high-scale version of CarlaSC, where each scene has a scale of 256 × 256 × 16 voxels, covering a physical space of 25.6 meters both in front of and behind the radar scanner and extending up to a height of 3 meters.

C.2 SemanticKITTI

SemanticKITTI [2], utilized for cross-dataset validation, showcases diverse environments, including inner-city traffic, residential areas, highways, and country-

Table b: Conversion of SemanticKITTI Labels to Correspond with CarlaSC's 11 Categories. Labels listed as *remove* are absent in CarlaSC, while those marked with - are omitted from semantic segmentation according to the original settings.

Index	Original Labels	Mapped index	Mapped Labels	Ratio	Index	Original Labels	Mapped index	Mapped Labels	Ratio
0	unlabeled	0	Unclassified	1.8×10^{-2}	51	fence	2	fence	7×10^{-2}
1	outlier	-	-	2×10^{-4}	52	other-structure	3	other	2×10^{-3}
10	car	10	vehicle	4.1×10^{-2}	60	lane-marking	6	road	4×10^{-5}
11	bicycle	remove	-	1×10^{-4}	70	vegetation	9	vegetation	2.6×10^{-1}
13	bus	remove	-	3×10^{-5}	71	trunk	9	vegetation	6×10^{-3}
15	motorcycle	remove	-	3×10^{-4}	72	terrain	7	ground	8×10^{-2}
16	on-rails	remove	-	0	80	pole	5	pole	3×10^{-3}
18	truck	remove	-	2×10^{-3}	81	traffic-sign	5	pole	6×10^{-4}
20	other-vehicle	remove	-	2×10^{-3}	99	other-object	3	other	1×10^{-2}
30	person	4	pedestrian	2×10^{-4}	252	moving-car	-	-	2×10^{-3}
31	bicyclist	remove	-	1×10^{-8}	253	moving-bicyclist	-	-	1×10^{-4}
32	motorcyclist	remove	-	5×10^{-9}	254	moving-person	-	-	2×10^{-4}
40	road	6	road	2×10^{-1}	255	moving-motorcyclist	-	-	3×10^{-5}
44	parking	7	ground	1.5×10^{-2}	256	moving-on-rails	-	-	0
48	sidewalk	8	sidewalk	1.4×10^{-1}	257	moving-bus	-	-	1×10^{-4}
49	other-ground	7	ground	4×10^{-3}	258	moving-truck	-	-	1×10^{-4}
50	building	1	building	1.3×10^{-1}	259	moving-other	-	-	4×10^{-5}

side roads. After voxelization, it comprises 22 sequences: sequences 00 to 10 (excluding 08) for training, 08 for validation, and 11 to 20 for testing. However, due to the absence of semantic labels for the test set in the official SemanticKITTI release, our assessment of scene generation depends on the validation set. The chosen scope extends 51.2 meters ahead of the vehicle, 25.6 meters laterally, and 6.4 meters in height, yielding a voxel scale of $256 \times 256 \times 32$. Originally featuring 28 categories, SemanticKITTI undergoes a class remapping or removal process to align with the 11 categories found in CarlaSC, detailed in Table b.

In this alignment, we have removed certain categories from SemanticKITTI. This decision is primarily based on the original dataset's configuration for semantic segmentation, such as moving objects, and the absence of corresponding labels in the CarlaSC dataset, like bicycles and motorcycles. Notably, the removed categories represent only a minor portion of the total number of labels. This ensures the labels conform to the relationships in Table a. Additionally, we omit the upper 16 voxels in height from SemanticKITTI to synchronize with CarlaSC's height representation, a decision justified by the rarity of objects above this limit and the need for consistency in dataset comparison.

D Additional Experiment Results

Due to the limitations on the length of the main text, we have reserved the primary experimental results for inclusion in the main body of the paper. In this section, we provide supplementary experimental outcomes not mentioned in the main text and more showcases of our visualization effects.

D.1 Generation Quality

In this section, we present an array of qualitative comparative results. Figures a and c showcase the outcomes of our *Unconditional Generation*, from which we can discern that scenes generated using the PDD method achieve greater semantic accuracy and encompass more details within the generated scenes. Figures d and b illustrate the results of our *Conditional Generation*. The outcomes reveal that the scenes generated by our conditional generation method closely resemble the ground truth. In contrast, while the approach using point clouds as a condition for scene restoration maintains structural similarity, it exhibits numerous inaccuracies in label correctness.

D.2 Applications

Our PDD method extends to two applications: cross-dataset and infinite scene generation. Figures e, g, f and h provide additional visualizations of cross-dataset scene generation on the SemanticKITTI dataset. Specifically, Figure e and g illustrate the distinctions in generation effects between DD and our method. Our method consistently generates more coherent semantic scenes, encompassing a wider range of objects and richer details. For Figure f and h, we initially downsample the ground truth to $64 \times 64 \times 8$ and then restore it to $256 \times 256 \times 16$. These results demonstrate our method's capability to reconstruct scenes closely aligned with the ground truth accurately.

Additionally, Figure i demonstrates the capability of our method to generate infinite scenes. Overall, the scenes generated by our method show high diversity with minimal repetition. We note certain artifacts, such as road interruptions, which can be attributed to two main factors. First, our approach, in line with prior works [8], assumes a local dependency for infinite scene generation (see Equation 6 in the paper). Though efficient, this approach may result in artifacts due to its tendency to overlook long-range dependencies. This could be addressed by sequential models like transformers, which is beyond the scope of this paper. Furthermore, the training dataset [2,12] consists of radar-scanned road segments rather than complete regional scenes, which potentially causes artifacts. The extrapolation beyond the scanned scene segments poses a significant challenge, resulting in disruptions within the generated infinite scenes. We anticipate that, with enhanced datasets and the incorporation of sequential models, our method has the potential to generate better infinite 3D scenes.

D.3 Ablation Study

Pyramid Diffusion. Due to space constraints, only a subset of experimental results are provided in Table 3. We now present the full experimental outcomes in Table c. The full results show that the performance for both conditional and unconditional generation improves incrementally with adding scales. However, within the four-scale pyramid, the increase is only marginal.

E Additional Discussion

Differences from 2D Approaches. While inspired by coarse-to-fine approaches in 2D image processing [6,9,11], directly applying them to 3D presents signifi-

Table c: Complete comparison of different diffusion pyramids on 3D semantic scene generation.

Cascaded	**Condition**	**mIoU(V)**	**MA(V)**	**mIoU(P)**	**MA(P)**	**F3D(↓)**	**MMD(↓)**
s_4	×	40.0	63.7	25.5	38.7	1.36	0.60
$s_1 \rightarrow s_4$	×	67.0	85.4	32.1	51.3	0.32	0.24
$s_1 \rightarrow s_2 \rightarrow s_4$	×	**68.0**	**85.7**	**33.9**	**52.1**	**0.32**	**0.20**
$s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4$	×	**68.0**	85.6	33.4	52.0	**0.32**	0.23
$s_1 \rightarrow s_4$	✓	52.5	77.2	27.9	43.1	0.36	0.28
$s_1 \rightarrow s_2 \rightarrow s_4$	✓	55.8	78.7	**29.8**	**46.6**	**0.34**	**0.27**
$s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4$	✓	**55.9**	**79.5**	29.6	45.8	**0.34**	0.28

cant challenges due to the added dimension, resulting in more complex data and increased computational demands. Furthermore, we focus on generating large-scale outdoor 3D scenes rather than the more prevalent generation of individual 3D objects. Outdoor scenes imply a higher level of diversity, comprising numerous objects, all of which require semantic coherence within the generated environments.

Another consideration is the relative scarcity of high-quality 3D datasets compared to the more mature field of 2D images. These constraints pose challenges for diffusion models in scene generation. To address this, we adopt a multi-scale approach. Initially, diffusion models train efficiently on small-scale data, ensuring diverse and semantically accurate scene generation. We then employ conditional generation techniques to refine the scenes progressively. Diffusion models excel under conditions' guidance, allowing for high-quality scene generation.

The flexibility offered by our pyramid approach ensures the diversity and quality of the generated scenes and facilitates cross-dataset generation. Additionally, the concept of our proposed Scene Subdivision Module aids in the realization of infinite scene generation, allowing for the seamless stitching and extension of scenes beyond fixed boundaries.

In conclusion, by tailoring the diffusion process to the unique demands of 3D data and leveraging conditional inputs for refinement, our method effectively bridges the gap between 2D inspiration and 3D application, unlocking new possibilities in scene generation with efficiency and adaptability.

Limitations. Despite the notable advantages of our method in both unconditional and conditional generation compared to other methods, as well as its extension to cross-dataset and infinite scene generation, it is subject to limitations primarily stemming from the scale and collecting methods of the current outdoor 3D scene datasets [2, 3, 12]. Consequently, the scenes generated by our method are constrained to the largest scale of $256 \times 256 \times 16$, although our method possesses the theoretical capability to generate larger scale. Additionally, incomplete object generation may occur in the generated scenes. Despite our efforts to mitigate this limitation through infinite scene generation, the quality of the generated results is still influenced by the characteristics of the training dataset.

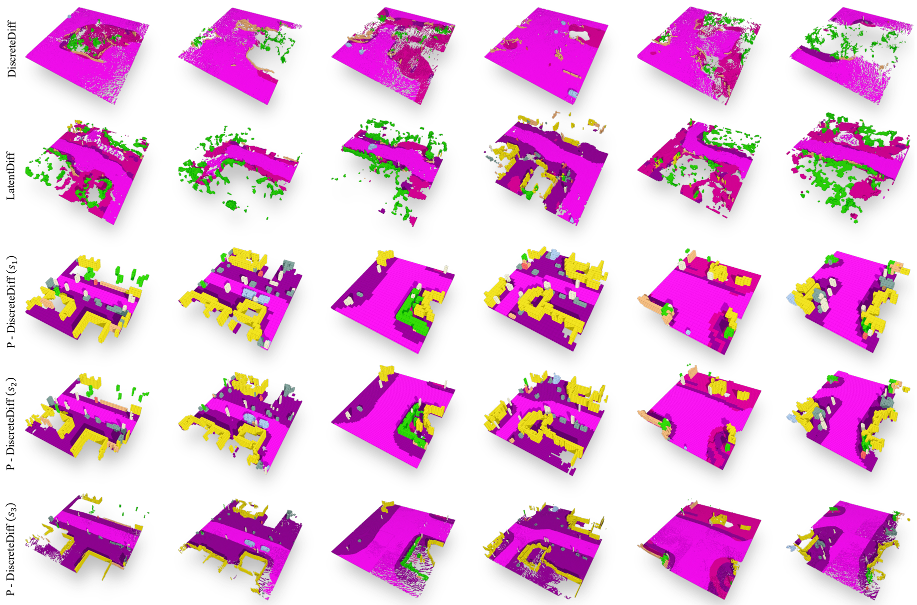

Fig. a: Additional visualization of unconditional generation results on CarlaSC. Our method produces more diverse scenes compared to the two baseline models [1, 7].

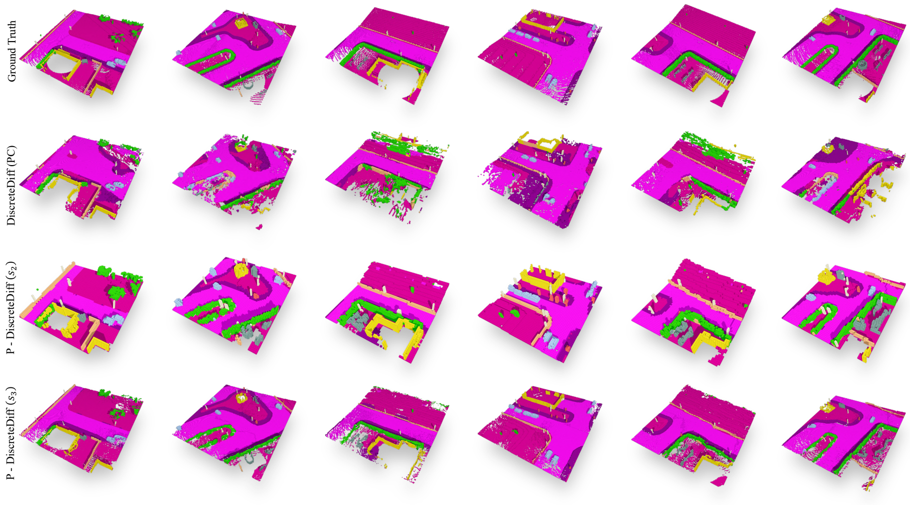

Fig. b: Additional visualization of conditional generation results on CarlaSC. *PC* stands for point cloud condition.

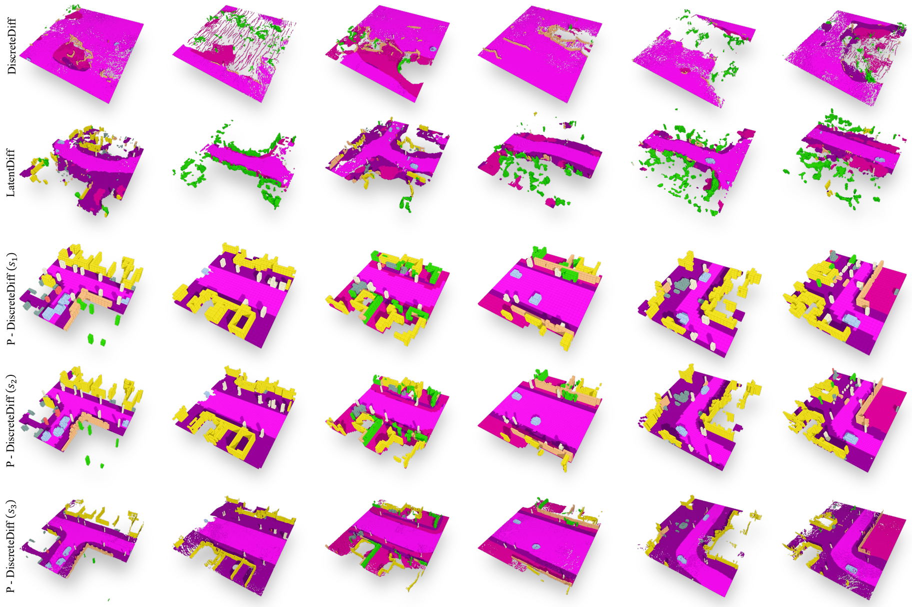

Fig. c: Additional visualization of unconditional generation results on CarlaSC. Our method produces more diverse scenes compared to the two baseline models [1, 7].

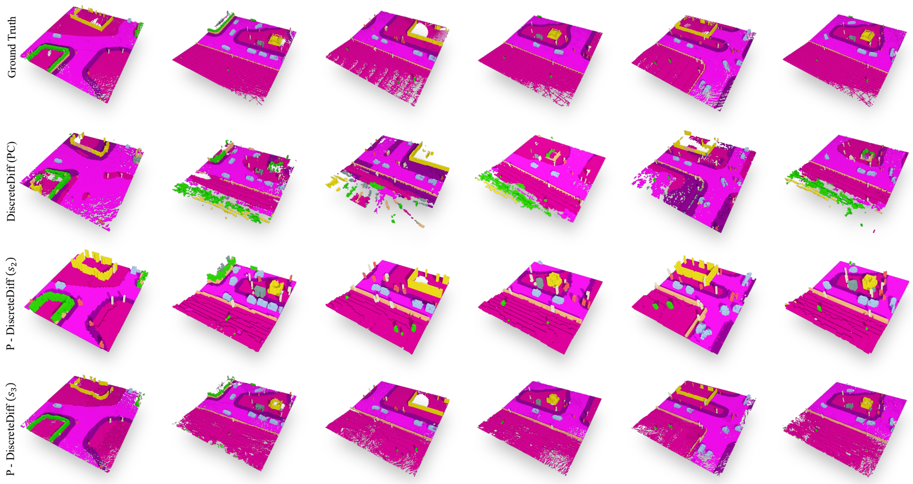

Fig. d: Additional visualization of conditional generation results on CarlaSC. *PC* stands for point cloud condition.

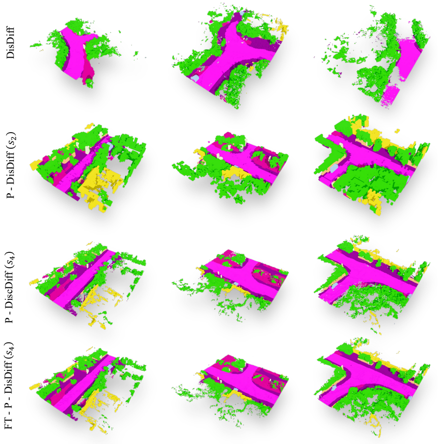

Fig. e: Additional SemanticKITTI unconditional generation. *FT* stands for finetuning pre-trained model from CarlaSC.

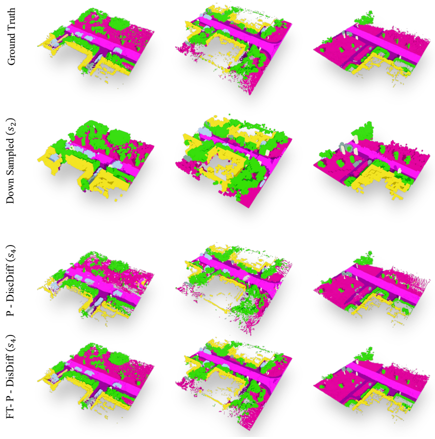

Fig. f: Additional SemanticKITTI conditional generation. Our proposed PDD achieves results close to the groundtruth. Note that *FT* stands for finetuning from CarlaSC models.

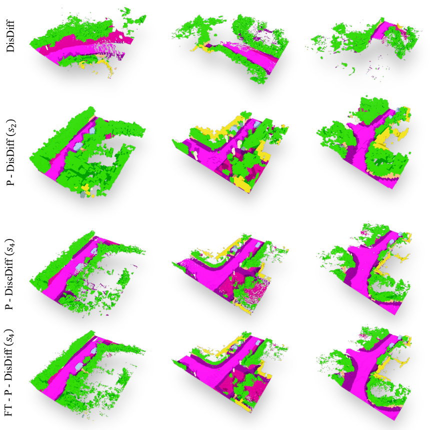

Fig. g: Additional SemanticKITTI unconditional generation. *FT* stands for finetuning pre-trained model from CarlaSC.

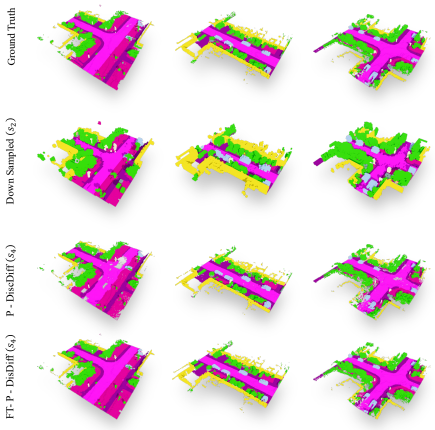

Fig. h: Additional SemanticKITTI conditional generation. Our proposed PDD achieves results close to the groundtruth. Note that *FT* stands for finetuning from CarlaSC models.

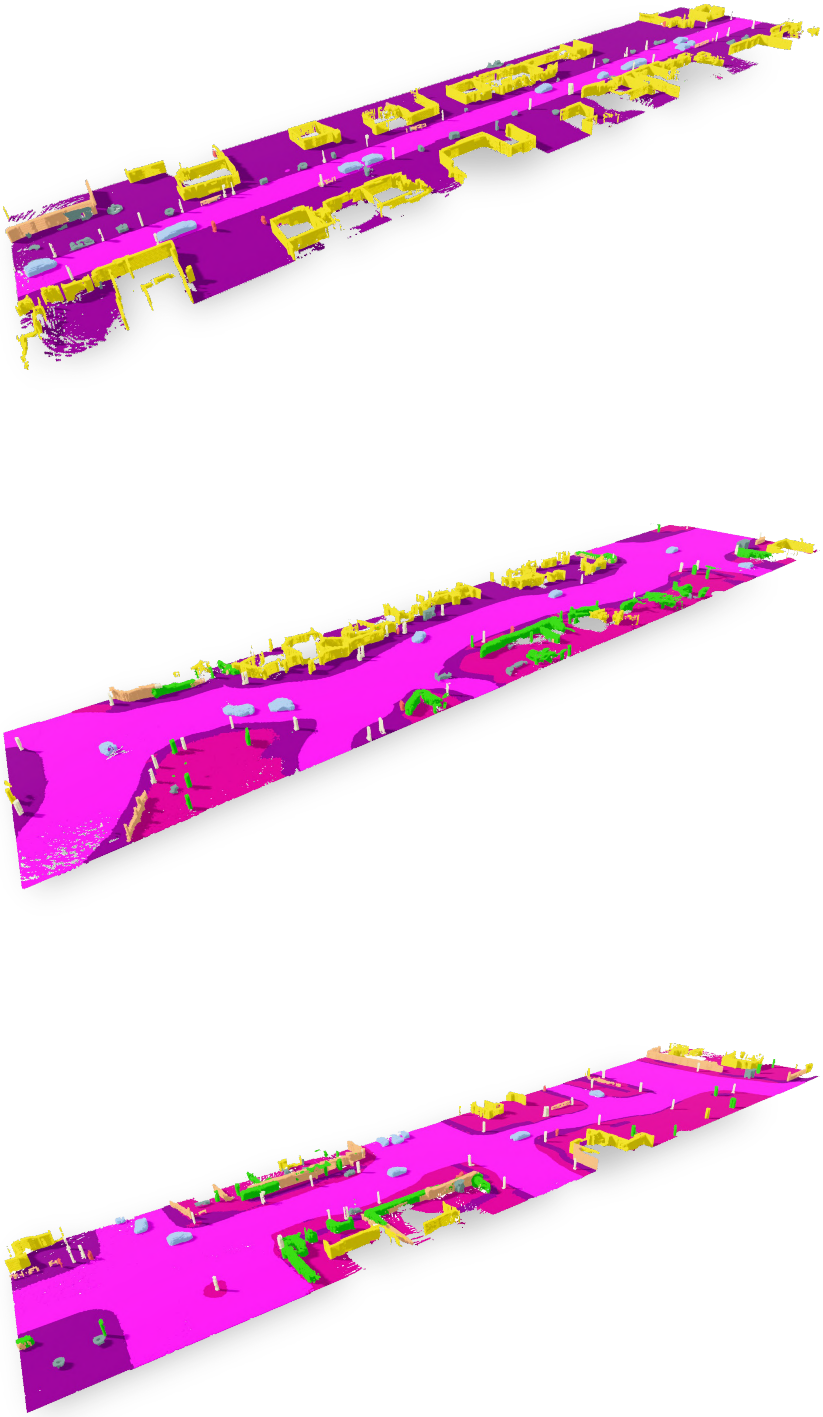

Fig. i: Infinite Scene Generation. Using PDD, we generate three different scenes. Our method produces infinite and consistent urban landscapes, seamlessly blending diverse urban elements to create a coherent and realistic cityscape.

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: NeurIPS (2021)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV (2019)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
4. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR (2018)
5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
7. Lee, J., Im, W., Lee, S., Yoon, S.E.: Diffusion probabilistic models for scene-scale 3d categorical data. arXiv preprint arXiv:2301.00527 (2023)
8. Lin, C.H., Lee, H.Y., Menapace, W., Chai, M., Siarohin, A., Yang, M.H., Tulyakov, S.: Infinicity: Infinite-scale city synthesis. arXiv preprint arXiv:2301.09637 (2023)
9. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
10. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
11. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. arXiv preprint arXiv:1503.03585 (2015)
12. Wilson, J., Song, J., Fu, Y., Zhang, A., Capodieci, A., Jayakumar, P., Barton, K., Ghaffari, M.: Motionsc: Data set and network for real-time semantic mapping in dynamic environments. IEEE Robotics and Automation Letters **7**(3) (2022)