SG-LDM: Semantic-Guided LiDAR Generation via Latent-Aligned Diffusion

Supplementary Material

8. Implementation

We build our model on a standard 2D diffusion framework [16], using a 2D U-Net [47] as the autoencoder backbone. Since lidar range images exhibit a wrap-around structure the same as panoramic images, we replace traditional convolutions with circular convolutions [50], following prior lidar diffusion models [20, 45, 77]. Additionally, we employ a lightweight three-layer CNN (the semantic projector) to map the U-Net's latent space to the same resolution as the rescaled semantic map. For inference and lidar translation, we use DDIM [53], a commonly adopted technique for efficient sampling. We adopt the standard linear variance schedule ($\beta_1 = 1 \times 10^{-4}$ to $\beta_T = 0.02$ over T steps) for both lidar generation and translation in our experiments.

9. Range Image and Point Cloud Conversion

Lidar range image leverages spherical projection to convert 3D point clouds into 2D images. Although there are some loss to this conversion, this technique has been shown effective to both the discriminative [33] and generative models [5] for lidar data. Given each 3D point (x, y, z) in the lidar coordinates, we have

- Range:

$$r = \sqrt{x^2 + y^2 + z^2} \tag{8}$$

- Azimuth angle:

$$\theta = \text{atan2}(y, x) \tag{9}$$

- Elevation angle:

$$\phi = \arcsin\left(\frac{z}{r}\right) \tag{10}$$

These angles are then rescaled and quantized to integer image coordinates (u, v). For a 360° sweep horizontally mapped into $u \in [0, 1023]$ and a set of 64 vertical rings mapped to $v \in [0, 63]$, we can apply

1. **Horizontal index**

$$u = \left\lfloor \frac{1024}{2\pi} (\theta + \pi) \right\rfloor \in [0, 1023] \tag{11}$$

so that $\theta = -\pi$ goes to $u = 0$ and $\theta = +\pi$ goes near $u = 1023$.

2. **Vertical index**

$$v = \left\lfloor \frac{64}{\phi_{\max} - \phi_{\min}} (\phi - \phi_{\min}) \right\rfloor \in [0, 63] \tag{12}$$

where $\phi_{\min}, \phi_{\max}$ are the minimum/maximum elevation angles of the lidar (-25° to $+3^\circ$ for both SemanticKITTI and SynLiDAR).

Finally, we can store the measured range r (and possibly intensity and semantic labels) and in the resulting 2D range image at pixel (u, v).

10. Evaluation Metrics

This section discusses the evaluation metrics used in the main body of the paper for assessing the quality of the generated point clouds in terms of both fidelity and diversity. The metrics for data generation can be categorized into two classes: perceptual and statistical.

Perceptual metrics measure the distance between real and generated data by comparing their representations in a perceptual space, which is derived from visual data using a pretrained feature extractor. In this research, we employ three perceptual metrics—FRID, FSVD, and FPVD—which serve as the lidar version of the commonly used Fréchet inception distance (FID).

- **FRID** employs RangeNet++ [33], a range-based lidar representation learning method, to extract features and compute distances. It is used as the primary metric because it evaluates only the regions within the range image, intentionally excluding areas outside where the data is less controlled. This approach reduces the influence of extraneous noise from regions far from the ego vehicle.
- **FSVD** employs MinkowskiNet [8] to extract features by first voxelize the 3D point clouds. This method can cover the entire lidar space. The final feature vector is computed by averaging all non-empty voxel features from every point cloud segment.
- **FPVD** empoloys SPVCNN [55], a point-voxel-based feature extractor which aggregates both point and volumetric features. This method can cover more geometric feature but in the other hand will be impacted more by the noisy points. The final feature vector is computed in the same way as FSVD.

Statistical metrics have been used as evaluation criteria for point cloud generative models since the pioneering work [2]. These metrics rely on distance functions to quantify the similarity between pairs of point clouds. Among these, the Chamfer Distance (CD) has been the prevalent choice in recent studies [45, 66] due to its computational efficiency compared to other measures:

$$CD(X, \hat{X}) = \sum_{x \in X} \min_{y \in \hat{X}} \|x - y\|_2^2 + \sum_{y \in \hat{X}} \min_{x \in X} \|x - y\|_2^2 \tag{13}$$

where $X, \hat{X}$ are the input and the reconstructed point cloud respectively, and x, y are individual points. The Chamfer

Distance is also used for evaluating the quality of the synthetic point clouds generated by the models. Based on this we have two metrics that focus on diversity and fidelity respectively:

- **Jensen-Shannon Divergence (JSD)** measures the similarity between two empirical distribution P_A and P_B based on the KL-divergence.

$$\text{JSD}(P_A||P_B) = \frac{1}{2}D_{KL}(P_A||M) + \frac{1}{2}D_{KL}(P_B||M) \tag{14}$$

  where $M = \frac{1}{2}(P_A + P_B)$ and $D_{KL}(\cdot||\cdot)$ is the KL-divergence of distributions represented by two probability density functions, P_A and P_B.
- **Minimum Matching Distance (MMD)** computes the average minimum distance between two matching point clouds from sets S_g and S_r:

$$\text{MMD}(S_g, S_r) = \frac{1}{|S_r|}\sum_{Y \in S_r} \min_{X \in S_g} CD(X, Y) \tag{15}$$

Statistical metrics were originally designed for object-level point clouds, making them less suited to the more complex, scene-level data we work with. To address this mismatch, we follow the method in [45] by voxelizing the lidar point clouds and computing the metrics based on these voxels. Furthermore, metrics such as MMD are highly sensitive to noise. Since lidar data often includes uncontrollable noisy points, particularly in regions not captured by the range image. As a result, we place less emphasis on these statistical metrics compared to perceptual metrics.

11. Additional Results

In addition to the perceptual and statistical metrics, we further evaluate semantic fidelity by applying the pretrained RangeNet++ to compute both semantic accuracy and mean IoU. We also calculate the mean absolute error (MAE) between the generated and ground-truth lidar range maps.

Method	MAE ↓	Accuracy ↑	mIoU ↑
LiDM	4.31	0.604	0.504
SG-LDM	1.28	0.808	0.696

Table 6. Evaluation of semantic fidelity on SemanticKITTI.