电子科技大学
University of Electronic Science and Technology of China

Text2LiDAR: Text-guided LiDAR Point Cloud Generation via Equirectangular Transformer

Abstract

- LiDAR 点云数据可以用于机器人视觉理解以及自动驾驶等方面，然而 LiDAR 设备昂贵并且在天气等多样性环境下收集点云数据难度较大
- 本文提出了文本控制的多模态 LiDAR 点云数据生成模型，即 Text2LiDAR Denoising Network，包含 Equirectangular Transformer Network（EA + REA）、Control-signal Embedding Injector 以及 Frequency Modulator
- 在 nuScenes 的基础上构建了 nuLiDARtext Dataset

Background

LiDAR Point Cloud data 的收集目前面临两个问题：

1. LiDAR 设备价格昂贵

2. 特殊环境下（如极端天气等）LiDAR 数据收集困难

因此可靠的 LiDAR 数据生成模型具有非常重要的研究意义

Previous Work

1. CARLA：基于 LiDAR 数据的物理意义，模拟图像处理过程，即 Physics-based（受限于物理模型，生成效果不好）

2. Lidarsim：在 Physics-based 模型的基础上加入了 Learning-based，但模型需要提前扫描背景

3. Pure Learning-based approaches，但是不能很好地拟合非线性的分布

4. 使用 UNet 和 Diffusion Model 进行扩散生成的模型

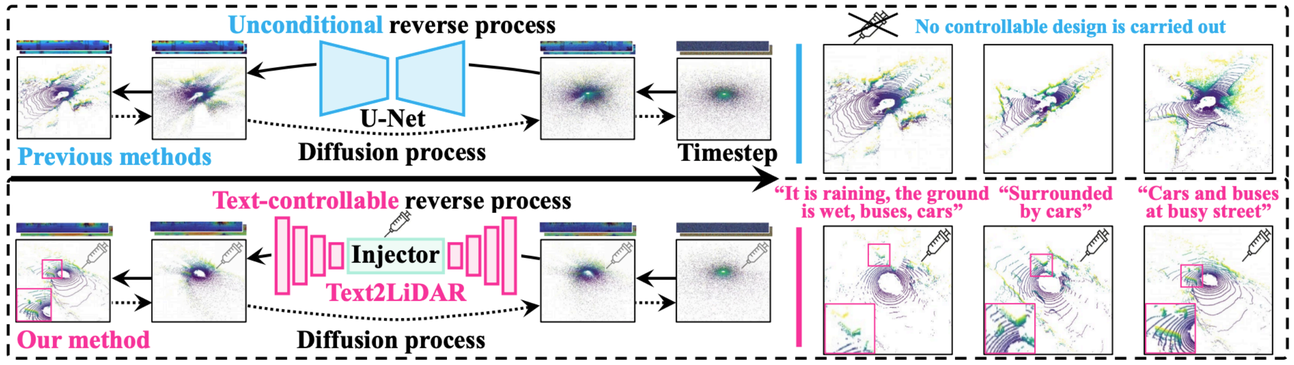


Figure 1: Schematic comparison of our Text2LiDAR and the existing diffusion-based generation framework [50, 89] without text guidance.

Previous Work

之前工作使用的卷积结构在处理 Point Cloud data 时存在的缺陷：

1. 卷积会破坏 Equirectangular Image 环形结构（Circular Structure）像素间的连续性
2. 卷积的 scalability 较差，很难适应 control signal 的融合
3. 现有工作没有考虑点云结构和高频细节之间的关系，MLP 操作会对点云图像中的高频细节产生平滑效果

Previous Work

文本条件引导的 LiDAR Point Cloud 数据生成面临的困难：

1. 没有为 Equirectangular Image 以及文本多模态融合而特殊改进设计的模型结构
2. 缺少可靠的 Text-LiDAR Point Cloud 数据对供模型进行对比学习

Main Contributions

1. 提出 Text2LiDAR Transformer 模型：

   (1) EA（Equirectangular Attention）

   (2) REA（Reverse Equirectangular Attention）

   (3) CEI（Context Embedding Injector）

   (4) FM（Frequency Modulator）

2. 提出新的数据集 nuLiDARtext

   对 nuScenes 进行更多针对 LiDAR 点云数据的特殊修改，最终得到在 850 个场景下的 34,149 个 Text-LiDAR Point Cloud 数据对

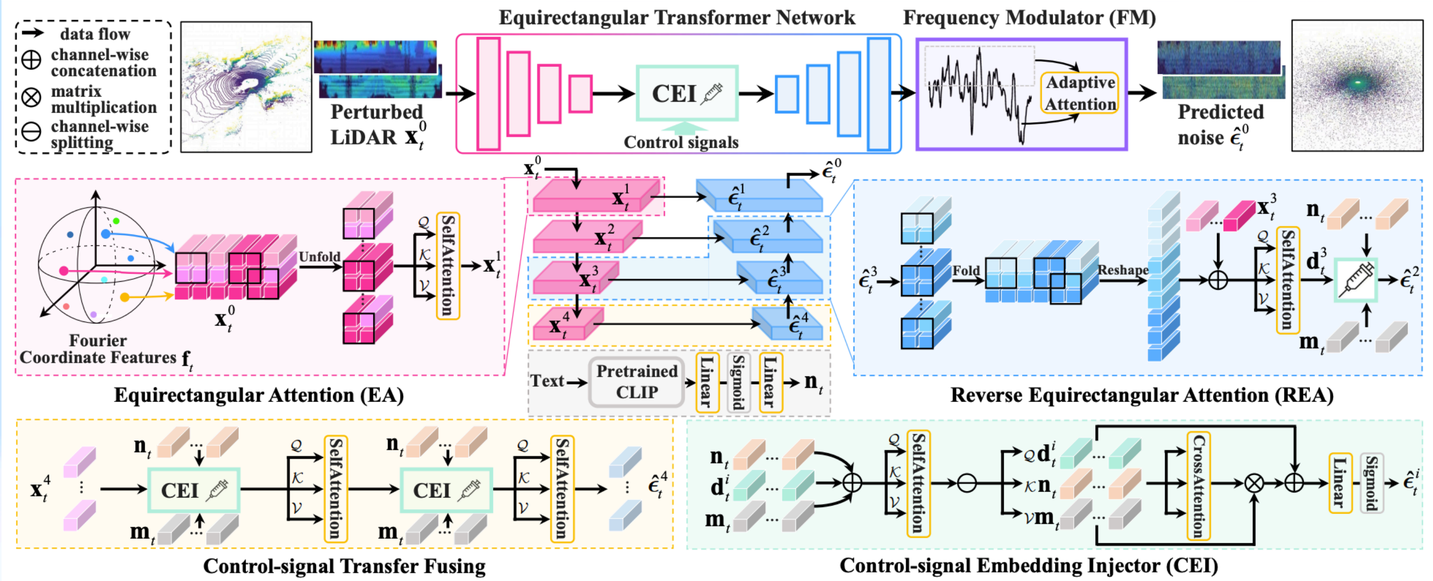


Figure 2: The architecture of the designed Text2LiDAR, where the designed equirectangular transformer is composed of stacked EA (encoding stage) and REA (decoding stage). The feature sequence will start interacting with the control signal at the 4th layer and be fed into a 4-layer decoder composed of REA. During decoding, the feature sequence continuously fuses the control signal through CEI. Finally, after frequency modulation, we can get the predicted noise.

Preliminary

This section introduces the formulation of the denoising diffusion probabilistic model (DDPM) and the loss function. As shown in Figure 1, the DDPM employs a forward diffusion process to gradually destroy the data sample $\mathbf{x}$ by adding noise as evolving the timestep $t \in [0, 1]$ until it becomes pure Gaussian noise. It also contains a backward reverse process, which aims at predicting the noise in each timestep and converting the pure Gaussian noise back into the data $\mathbf{x}$. To be more specific, at the timestep t, we can obtain the noised sample $\mathbf{x}_t$ through $q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\alpha_t\mathbf{x}, \sigma_t^2\mathbf{I})$, where $\mathbf{x}_t$ can be re-parameterized as: $\mathbf{x}_t = \alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}_t$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon}_t$ is the noise that vary with timestep t. α_t and σ_t are hyperparameters that depend on timestep t following the α-cosine schedule [50], we set $\alpha_t = \texttt{cos}(\pi t/2)$, $\sigma_t = \texttt{sin}(\pi t/2)$. Under the assumption $\alpha_t^2 + \sigma_t^2 = 1$, the process of obtaining the intermediate noised sample x_s can be described as $q(\mathbf{x}_t|\mathbf{x}_s) = \mathcal{N}(\alpha_{t|s}\mathbf{x}_s, \sigma_{t|s}^2\mathbf{I})$, where $0 \leq s < t \leq 1$, $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s}^2 = \sigma_t - \alpha_{t|s}\sigma_s$. The corresponding reverse process can be described as:

$$p(\mathbf{x}_s|\mathbf{x}_t) = q(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}). \tag{1}$$

After obtaining the noised sample $\mathbf{x}_t$, we need to design a denoiser $\texttt{Text2LiDAR}_{\boldsymbol{\varphi}}$ to predict the noise $\hat{\boldsymbol{\epsilon}}_t = \texttt{Text2LiDAR}_{\boldsymbol{\varphi}}(\mathbf{x}_t, t)$ at each timestep t. Then, the denoised results can be obtained through Equation 1. Completing the entire denoising process for each timestep t, we can yield the final generated result. We use the mean squared error (MSE) loss function for the training process:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[||\boldsymbol{\epsilon}_t - \texttt{Text2LiDAR}_{\boldsymbol{\varphi}}(\mathbf{x}_t, t)||_2^2], \tag{2}$$

where $\boldsymbol{\varphi}$ means the learnable parameters. As is customary [50], our denoiser is also conditioned on t. After training, we can obtain the final generated results by recursively evaluating $p(\mathbf{x}_s|\mathbf{x}_t)$ through the process for $t = 1 \rightarrow 0$.

Text2LiDAR Denoising Network: Overall Architecture

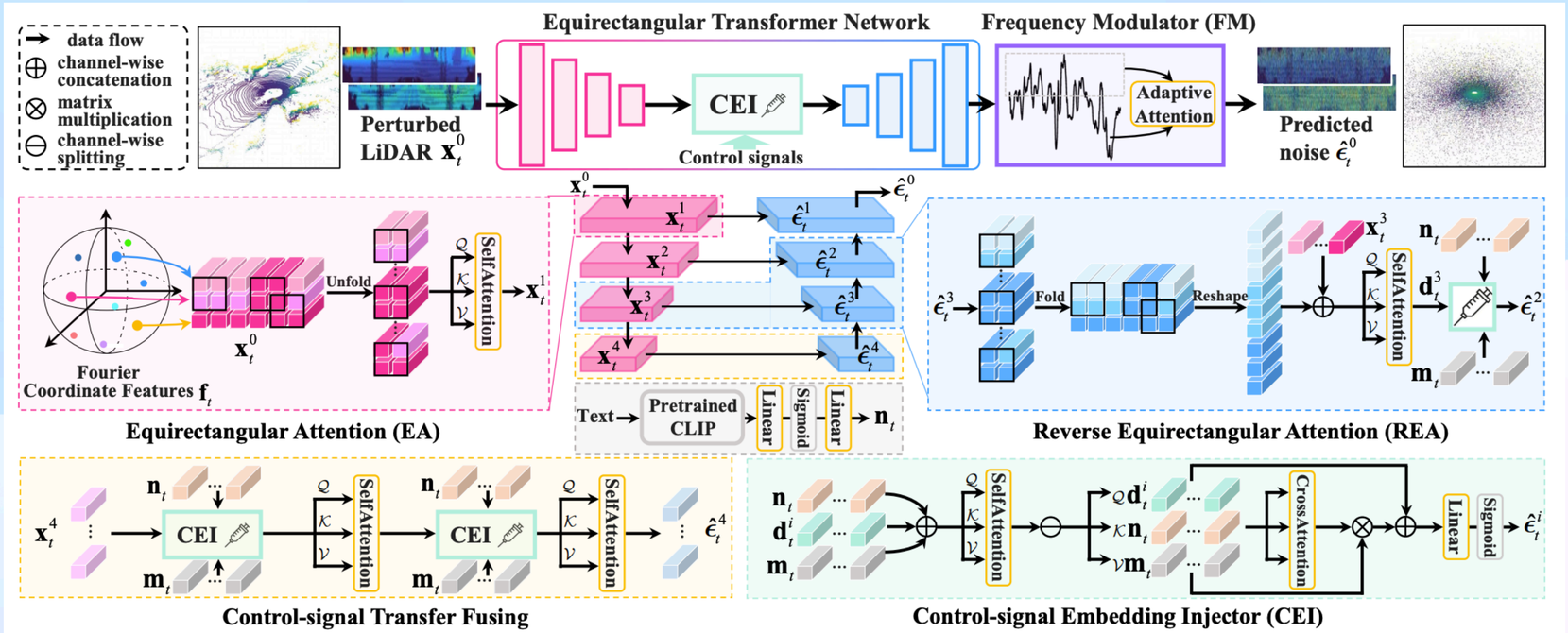
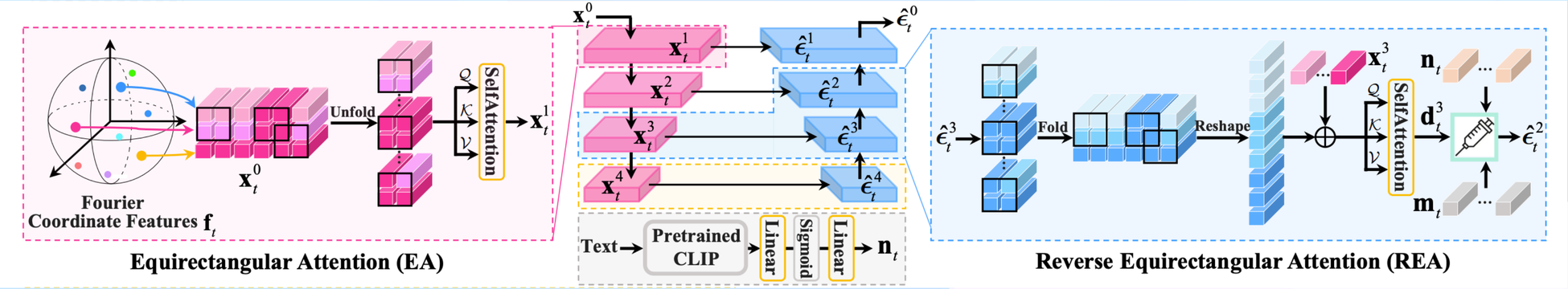


Figure 2: The architecture of the designed Text2LiDAR, where the designed equirectangular transformer is composed of stacked EA (encoding stage) and REA (decoding stage). The feature sequence will start interacting with the control signal at the 4th layer and be fed into a 4-layer decoder composed of REA. During decoding, the feature sequence continuously fuses the control signal through CEI. Finally, after frequency modulation, we can get the predicted noise.

Text2LiDAR Denoising Network: EA



$$\mathbf{x}_t^{i+1} = \mathbf{EA}(\mathbf{x}_t^i) = \mathbf{SelfAttention}(\mathbf{Unfold}(\mathbf{x}_t^i \oplus \mathbf{f}_t)), \qquad (3)$$

1. 由于 Equirectangular Image 的特殊 Circular Structure，使用 Self Attention 代替卷积
2. 使用 Fourier Feature 作为 Positional Embedding，引入 Elevation（仰角） 和 Azimuth（方位角）
3. 由于 Equirectangular Image 具有会把特征拉长（Elongated Nature）的特点，存在 Scale Variation 问题，提出 Mutually Overlapping Unfolding，在不同阶段将特征分成不同的 Scales

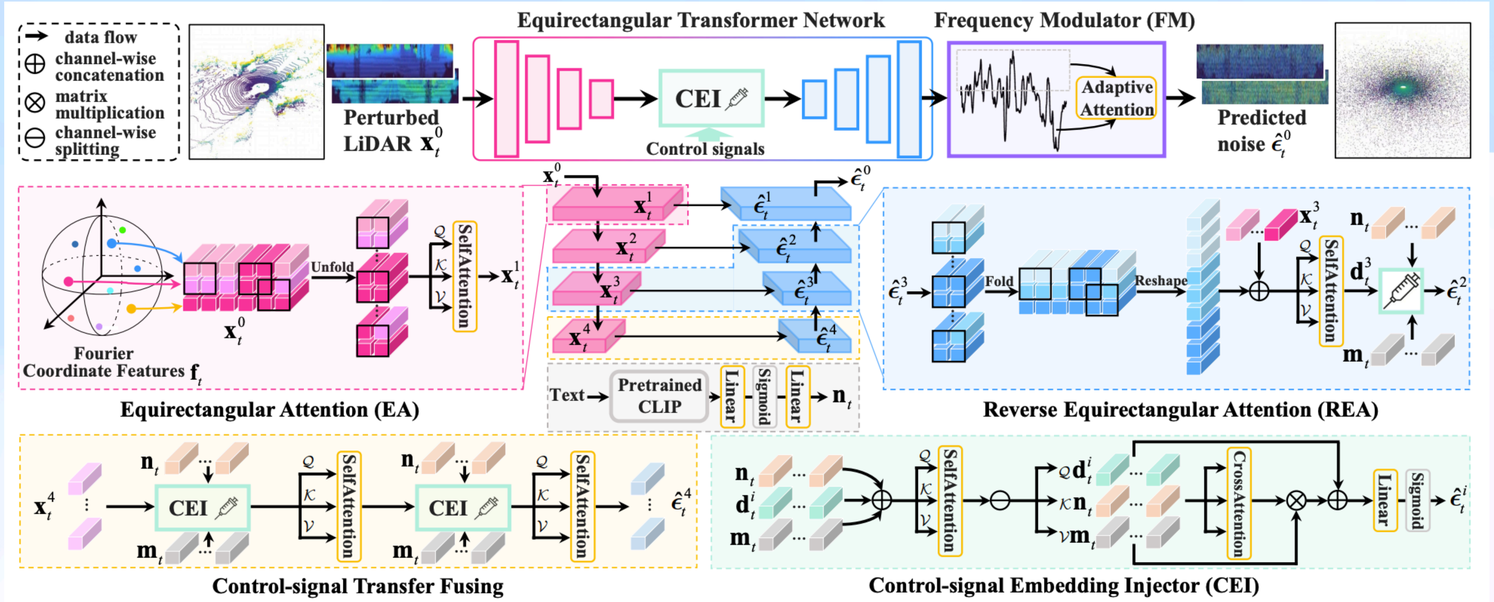
Text2LiDAR Denoising Network: REA

1. 在 Decoding 正式开始之前，对于 Encoder 的输出进行额外的控制信号初始化融合（Initial Fusion of Control Signal），即对 x_t^4 进行 Control-signal Transfer Fusing 得到 $\hat{\epsilon}_t^4$

2. Fold 操作

3. 为了更好的恢复原始点云图像细节，在 Decoding 阶段加入了对应 Encoder 层输出的残差

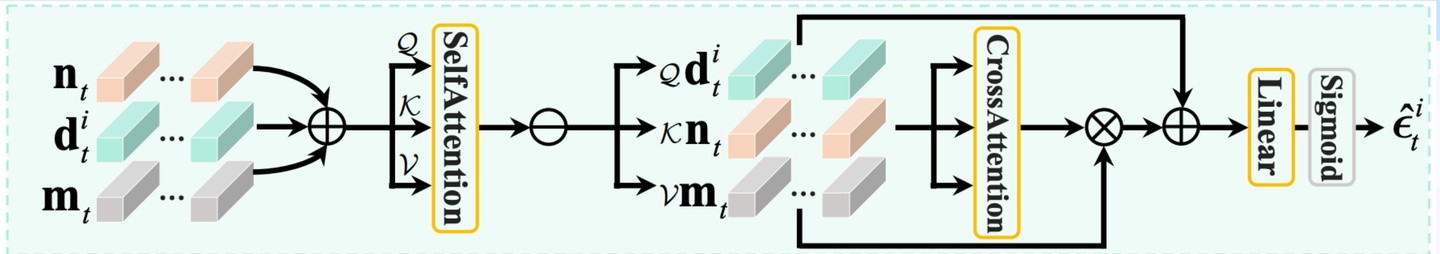
4. CEI（Control Signal Embedding Injector）



$$\hat{\boldsymbol{\epsilon}}_t^i = \texttt{REA}(\mathbf{x}_t^{i+1}, \hat{\boldsymbol{\epsilon}}_t^{i+1}, \mathbf{n}_t, \mathbf{m}_t) = \texttt{CEI}(\texttt{SelfAttention}(\mathbf{x}_t^{i+1} \oplus \texttt{Fold}(\hat{\boldsymbol{\epsilon}}_t^{i+1})), \mathbf{n}_t, \mathbf{m}_t), \tag{4}$$

Text2LiDAR Denoising Network: Control-signal Embedding Injector

1. Self Attention（文本条件嵌入、数据输入以及时间步嵌入）
2. Cross Attention，条件分别为文本嵌入以及时间步嵌入
3. 为时间步的 V 矩阵 $\mathbf{v}m_t$ 额外执行一次矩阵乘法，保证时间步的引导
4. 对原始的数据输入额外执行一次残差 concatenation



Control-signal Embedding Injector (CEI)

$$\hat{\epsilon}_t^i = \texttt{CrossAttention}(\texttt{Split}(\texttt{SelfAttention}(\mathbf{n}_t \oplus \mathbf{d}_t^i \oplus \mathbf{m}_t))) \otimes \mathcal{V}\mathbf{m}_t \oplus \mathcal{Q}\mathbf{d}_t^i, \quad (5)$$

Text2LiDAR Denoising Network: Frequency Modulator

1. DWT（离散小波变换）：将输入分别沿水平方向和竖直方向根据信号频率的高低分成四个子带：LL，HL，LH，HH

2. FG（Frequency Gating）：由卷积组成，通过训练来确定保存哪些频率的信号、舍弃哪些频率的信号

3. IDWT（Inverse Discrete Wavelet Transform）：将四个字带重新恢复为完整信号

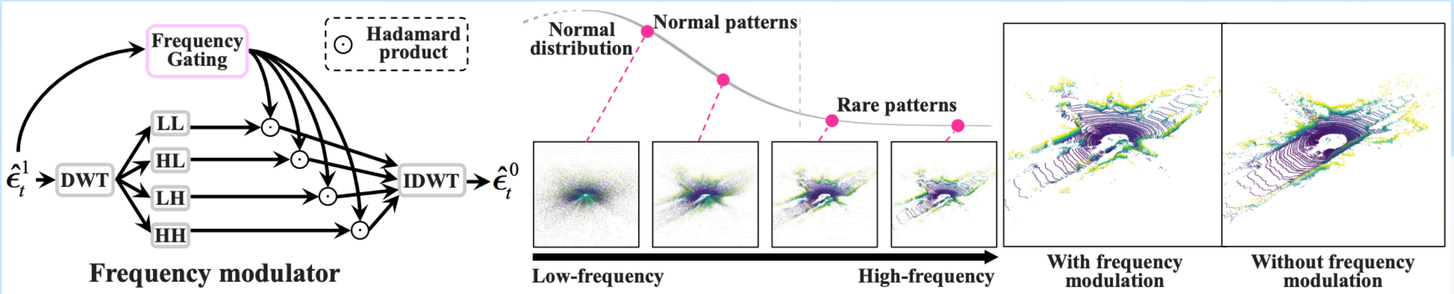


Figure 3: The architecture of the frequency modulator.

$$\hat{\epsilon}_t^0 = \mathtt{FM}(\hat{\epsilon}_t^1) = \mathtt{IDWT}(\mathtt{DWT}(\hat{\epsilon}_t^1) \odot \mathtt{FG}(\hat{\epsilon}_t^1)).$$

FM 的目标是将输入分解为多角度高频小波带进行调制，引导模型自适应地适应不同的频率，缓解等距柱状图像的过渡平滑度

nuLiDARtext

- nuScenes 数据集中的文本描述用于描述短时间内一段连续的场景，而不是专门用于 LiDAR 数据配对
- 存在拼写错误、语义歧义、连续状态的描述和干扰词等问题。nuLiDARtext 调整了 nuScenes 提供的 850 个场景中 34,149 帧的描述，包括添加、删除、修改和标准化等操作
- 缩写：将 “ped” 更正为 “pedestrians”，可以获得更有效的文本嵌入
- 语义歧义：nuScenes 中存在同时提及 “turn left” 和 “turn right” 的实例
- 干扰词：将“waiting at the intersection”修改为“at the intersection”，因为“waiting”是一个连续的状态描述，可能会稀释单帧生成的有效信息。

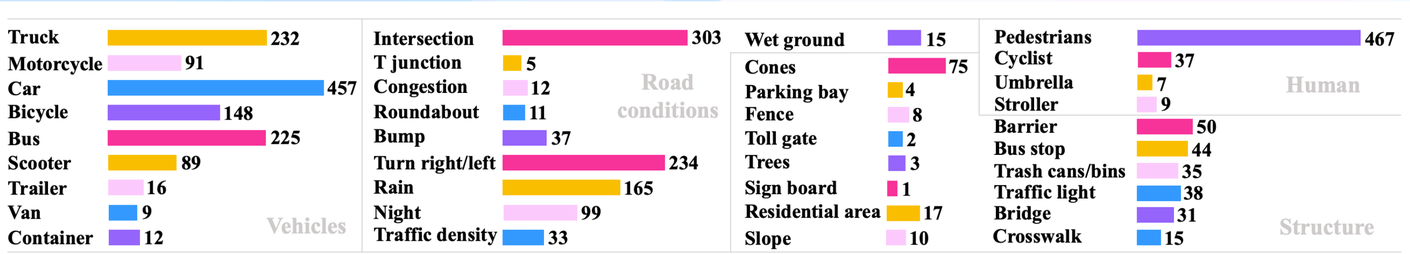


Figure 4: The number of occurrences of text in 850 scenes.