Semantics-Guided Multimodal Masked Autoencoder Pretraining for 3D BEV Object Detection

Prabuddhi Wariyapperuma[1], Rajitha de Silva[1], Marc Hanheide[1], Thomas Bohné[2] and Leonardo Guevara[1,*]

***Abstract*— Accurate 3D bird's-eye view (BEV) object detection is essential for autonomous driving, and depends strongly on effective multimodal representations from complementary sensors such as cameras and LiDAR. Multimodal masked autoencoders have shown strong potential for learning such representations for downstream 3D BEV object detection. However, existing methods typically apply uniform random masking to camera and LiDAR inputs, treating all regions equally, and learn representations only through masked reconstruction. We propose a semantics-guided multimodal masked autoencoder framework that introduces semantic information during pretraining through two separate components: (i) semantics-guided LiDAR voxel masking, which preserves semantically important LiDAR regions more strongly, and (ii) an auxiliary point-wise LiDAR semantic decoder branch that injects semantic guidance in addition to reconstruction. On BEVFusion 3D object detection, our semantics-guided pretraining strategy improves performance on the nuScenes mini validation set compared to the standard UniM^{2}AE baseline: semantics-guided LiDAR voxel masking yields +1.49% mean Average Precision (mAP) and +1.66% nuScenes Detection Score (NDS), while decoder-side point semantic supervision yields +1.39% mAP and +3.22% NDS over the baseline.**

I. Introduction

Reliable 3D object detection is essential for autonomous driving because planning and safety-critical decision-making depend on accurately localising surrounding actors and obstacles in three-dimensional space. BEV representation has emerged as a strong formulation for this problem because it reduces perspective distortion and provides a unified top-down view of scene layout, making spatial reasoning and sensor fusion more reliable and effective. In particular, multimodal fusion of camera-LiDAR inputs has become central to high-performance 3D BEV object detection, as the two modalities provide complementary semantic and geometric information [1].

Building on this, multimodal masked autoencoders have emerged as a promising way to learn transferable camera-LiDAR representations for downstream 3D detection, because masked reconstruction encourages the model to learn meaningful multimodal structure from incomplete inputs. Multimodal Masked Autoencoders with Unified 3D Representation (UniM^{2}AE) follows this paradigm by randomly masking image patches and LiDAR voxels before fusing the two modalities in a unified 3D volume space for reconstruction and downstream detection [2]. The existing works demonstrate the promise of multimodal masked autoencoders, but they also raise an important question: is uniform random masking together with reconstruction-only learning sufficient for learning the most effective multimodal representations?

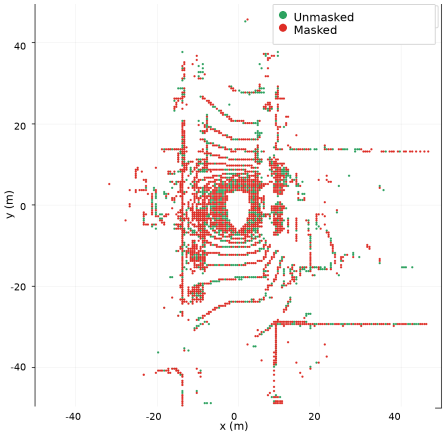

Fig. 1: Proposed semantics-guided LiDAR voxel masking policy on a sample from the nuScenes mini validation set.

Existing masked autoencoders largely follow a simple design: randomly hiding part of the input and learning representations by reconstructing the missing content [2], [3]. More recent work suggests that the masking policy itself can substantially influence what the model learns. In the LiDAR domain, Occupancy-MAE [4] introduces range-aware masking to reflect distance-dependent sparsity, while I2P-MAE [5] shows that importance-aware masking can preserve semantically important 3D tokens more effectively than uniform random masking. These works highlight the value of preserving semantically important regions during masking. Similar evidence also appears in the image domain, where SemMAE [6] demonstrates that semantic-guided masking improves representation learning over uniform random masking. Motivated by these findings, our first contribution introduces semantics-guided masking for LiDAR voxels in multimodal camera-LiDAR masked autoencoders, as illustrated in Fig. 1.

Beyond importance-aware masking, I2P-MAE also introduces semantic targets during pretraining, indicating that semantic guidance can improve masked representation learning beyond pure reconstruction [5]. Inspired by this, our second contribution introduces auxiliary point-wise LiDAR semantic supervision beyond reconstruction. More broadly, neither semantics-guided masking nor auxiliary semantic

This work was supported by the Engineering and Physical Sciences Research Council and AgriFoRwArdS CDT [EP/S023917/1].

[1] University of Lincoln, Lincoln Centre for Autonomous Systems, Lincoln, UK. [2] University of Cambridge, Institute for Manufacturing, Department of Engineering, Cambridge, UK.

* Corresponding author: lguevara@lincoln.ac.uk

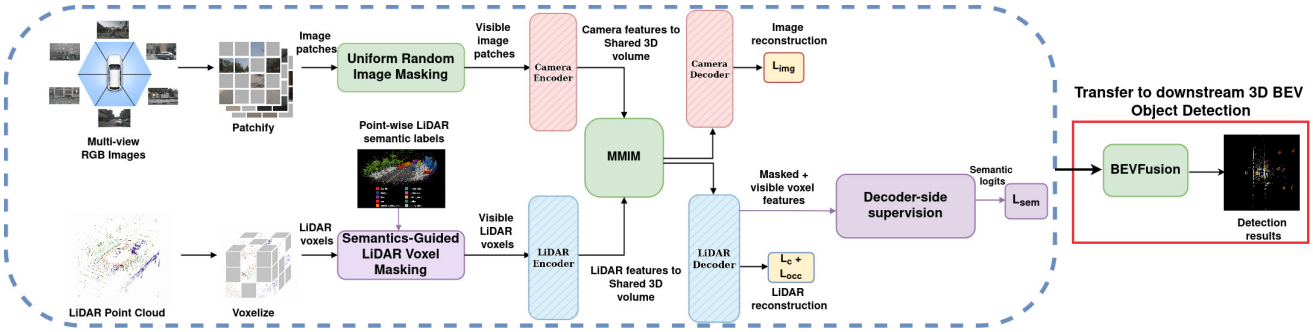

Fig. 2: Overview of our semantics-guided multimodal masked autoencoder pretraining framework for downstream 3D BEV object detection. The purple blocks denote the components newly introduced in our method: semantics-guided LiDAR voxel masking and decoder-side point semantic supervision. The dotted blue boundary marks the pretraining-only part of the framework, indicating that all components inside it are used only during pretraining.

supervision has yet been explored in multimodal camera-LiDAR masked autoencoders for 3D BEV detection, where LiDAR voxels are still generally treated as equally maskable units, learning remains dominated by reconstruction, and the impact of semantic information on downstream 3D BEV object detection has not yet been studied.

To address this gap, we propose a semantics-guided multimodal masked autoencoder pretraining framework for downstream 3D BEV object detection. The main contributions of this paper are as follows:

- We introduce a semantics-guided LiDAR voxel masking strategy for multimodal masked autoencoder pretraining that preserves semantically important LiDAR regions more strongly than uniform random masking.
- We introduce auxiliary point-wise LiDAR semantic supervision during pretraining to complement reconstruction with explicit semantic guidance for LiDAR features.
- We show that introducing semantic information during multimodal masked autoencoder pretraining improves downstream 3D BEV object detection on the nuScenes mini validation set compared with the standard UniM^{2}AE baseline.

II. Method

Our approach introduces semantic guidance to multimodal masked autoencoder pretraining with semantic guided masking and auxiliary semantic supervision. Figure 2 shows an overview of the proposed framework.

A. Baseline Multimodal Masked Autoencoder Pretraining

We adopted UniM^{2}AE [2] as the baseline, since it is a strong camera-LiDAR masked autoencoder. The LiDAR branch first voxelises the input point cloud into LiDAR voxel tokens, while the camera branch divides the multi-view images into non-overlapping image patch tokens. Uniform random masking is then applied independently to the LiDAR voxels and image patch tokens, and only the resulting visible tokens are processed by the LiDAR and camera encoders, respectively. The visible camera and LiDAR features are projected into a shared 3D volume, fused through the learnable Multimodal 3D Interaction Module (MMIM), and then mapped back to modality-specific decoder inputs. For masked positions, learned mask-token embeddings are inserted at the corresponding voxel or patch locations before decoding to represent the missing content. In both branches, the decoder receives the encoded visible features together with learned mask tokens at masked positions to reconstruct the original camera and LiDAR inputs. The baseline pretraining objective combines masked image and LiDAR voxel reconstruction losses:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{img}} + \mathcal{L}_c + \mathcal{L}_{\text{occ}} \tag{1}$$

where $\mathcal{L}_{\text{img}}$ is the mean squared error for masked image reconstruction, $\mathcal{L}_c$ is the Chamfer distance loss for LiDAR point-set reconstruction, and $\mathcal{L}_{\text{occ}}$ is the occupancy loss for predicting whether a voxel is empty or occupied.

After pretraining, the decoders are discarded and the pretrained encoders and fusion components are transferred to BEVFusion [1] for downstream 3D BEV object detection. In this work, we keep this downstream fine-tuning stage unchanged and introduce semantic information only during pretraining through (i) semantics-guided LiDAR voxel masking and (ii) auxiliary point-wise LiDAR semantic supervision.

B. Semantics-Guided LiDAR Voxel Masking

The original UniM^{2}AE applies uniform random masking to LiDAR voxels, treating all voxels equally. In this work, we investigated whether different semantic classes affect masked LiDAR reconstruction differently, and whether this influences downstream 3D BEV object detection. To do this, we used LiDAR semantic labels to guide voxel masking during pretraining.

a) Semantic Class Importance Analysis: We analysed the importance of each semantic class for masked LiDAR reconstruction. For a target semantic class c, we define the set of target voxels as $\mathcal{V}^{(c)} = \{v \mid n_v^{(c)} \geq \tau_c\}$, where $n_v^{(c)}$ denotes the number of points of class c in voxel v, and τ_c is a class-specific threshold. A voxel was therefore assigned to $\mathcal{V}^{(c)}$ if it contained at least τ_c points from class c. To ensure a fair comparison with the baseline, we kept the overall LiDAR masking ratio fixed at the baseline value ρ, so that only the

TABLE I: Downstream 3D BEV object detection results on the nuScenes mini validation set.

Method	mAP (%) ↑	NDS (%) ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
Baseline (Uniform Random Masking)	24.72	31.41	0.5682	0.5373	1.2709	0.6455	0.3436
Semantics-Guided LiDAR Voxel Masking	**26.21**	33.07	0.5054	0.5089	1.1032	0.6555	0.3342
Decoder-Side Point Semantic Supervision	26.11	**34.63**	**0.4802**	**0.5033**	1.1235	**0.5614**	**0.2976**
Post-MMIM Point Semantic Supervision (Ablation)	26.09	32.96	0.4821	0.5203	**1.0509**	0.6845	0.3217

semantic composition of the masked voxels changed. We first masked all voxels in $\mathcal{V}^{(c)}$ and then masked the remaining occupied voxels at random until the final LiDAR masking ratio matched the baseline masking ratio ρ.

We then evaluated each class-specific masking setting using voxel-level LiDAR reconstruction metrics in Eq. (1), namely the Chamfer distance from predicted points to ground-truth points, the Chamfer distance from ground-truth points to predicted points, and voxel occupancy accuracy. Classes that caused larger degradation in these metrics when masked were treated as more important during pretraining.

b) Importance-Based Masking Policy: Based on the class-importance analysis above, we used the resulting ranking to determine how the fixed LiDAR masking ratio should be distributed across semantic groups during pretraining. We first grouped all non-empty LiDAR voxels into four categories according to their semantic importance level: high, medium, low, and background. We then kept the same baseline masking ratio ρ, but redistributed the masked voxels across these groups so that more important semantic groups were protected more strongly, while less important and background groups were masked more heavily. For voxels containing multiple semantic classes, each voxel was assigned to the most important class present. Within each group, voxels were sampled uniformly at random. In this way, the original UniM^{2}AE pretraining protocol was preserved, while semantic structure was introduced into the LiDAR masking policy.

C. Auxiliary Point-wise LiDAR Semantic Supervision

While the reconstruction losses in Section II-A train the model to recover masked LiDAR voxels, they do not directly encourage the learned LiDAR representations to predict semantic classes. To address this, we introduce an auxiliary point-wise LiDAR semantic supervision branch during pretraining, while keeping the downstream BEVFusion fine-tuning stage unchanged.

We attach the semantic supervision branch to the decoder side, so that it can supervise points from both masked and unmasked voxels involved in LiDAR reconstruction. Let Ω_{dec} denote the set of LiDAR points with semantic labels that map to decoder voxels used for LiDAR reconstruction. For each point $p \in \Omega_{\mathrm{dec}}$, we gather the decoder-side voxel feature of its corresponding voxel, denoted by $\mathbf{f}_p^{\mathrm{dec}} \in \mathbb{R}^{128}$. Since multiple points can lie inside the same voxel and therefore share the same voxel-level feature, we concatenate this feature with a 3D local point offset $\Delta\mathbf{p} \in \mathbb{R}^3$, representing the point position relative to the voxel center, to form the point-wise semantic input $\mathbf{z}_p = \left[\mathbf{f}_p^{\mathrm{dec}}\,;\,\Delta\mathbf{p}\right]$.

A lightweight multi-layer perceptron (MLP) semantic head H_{sem} takes $\mathbf{z}_p$ as input and predicts per-point semantic logits as $\hat{\mathbf{y}}_p = H_{\mathrm{sem}}(\mathbf{z}_p)$, where $\hat{\mathbf{y}}_p$ denotes the predicted logit vector for point p, and y_p denotes its ground-truth semantic label. The auxiliary semantic loss is defined as the average cross-entropy over valid labeled points, $\mathcal{L}_{\mathrm{sem}} = \frac{1}{|\Omega_{\mathrm{dec}}|}\sum_{p\in\Omega_{\mathrm{dec}}} \mathrm{CE}(\hat{\mathbf{y}}_p, y_p)$. The overall pretraining objective is then extended as $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}}$, where $\mathcal{L}_{\mathrm{base}}$ is the baseline reconstruction loss defined in Section II-A, and λ_{sem} controls the contribution of the auxiliary semantic term. In this way, the decoder-side LiDAR features are trained not only for reconstruction, but also for semantic prediction.

III. Experimental Evaluation

A. Experimental Setup

We conducted our experiments on nuScenes mini, a subset of the nuScenes trainval split containing 10 scenes. nuScenes is a multimodal autonomous driving dataset with 360° surround sensing using six RGB cameras and LiDAR [7]. We used the nuScenes-lidarseg annotations, which provide 32 point-wise semantic labels for keyframe LiDAR points. Following the standard nuScenes 3D detection task, we evaluated downstream 3D BEV object detection on the 10 benchmark object categories and reported mAP, NDS, and the true-positive error metrics: mean Average Translation Error (mATE), Scale (mASE), Orientation (mAOE), Velocity (mAVE), and Attribute (mAAE).

B. Results and Discussion

Table I reports downstream 3D BEV object detection results on the nuScenes mini validation set for four settings: the uniform random masking baseline from Section II-A, the semantics-guided LiDAR voxel masking strategy from Section II-B, our decoder-side point semantic supervision method from Section II-C, and a post-MMIM semantic supervision method used only as an ablation. mAP and NDS are reported in %, where higher is better, while the true-positive error metrics are reported in their native units, where lower is better.

a) Random masking baseline: The baseline follows the original UniM^{2}AE pretraining pipeline with uniform random LiDAR voxel masking at a fixed masking ratio of $\rho = 0.7$ (70%). This setting achieved 24.72% mAP and 31.41% NDS, and serves as the reference point for evaluating the effect of introducing semantic information during pretraining.

b) Semantics-guided LiDAR voxel masking: Semantics-guided LiDAR voxel masking keeps the same overall LiDAR masking ratio of 70% but redistributes the masked voxels according to semantic class importance, using $\tau_c = 1$ for the class-specific voxel threshold. As shown in Table I, this

TABLE II: Semantic class importance analysis used to construct the Semantics-Guided LiDAR Voxel Masking policy in Section II-B.

Detection class	Chamfer distance GT to Pred ↓	Chamfer distance Pred to GT ↓	Occupancy accuracy ↑	Mean rank	Importance level	Masking weight
car	0.181647	0.436208	0.977032	7.5	High	0.75
pedestrian	0.180714	0.438906	0.977074	7.5	High	0.75
construction_vehicle	0.180739	0.437317	0.976535	7.5	High	0.75
motorcycle	0.179613	0.436273	0.976579	5.0	Medium	0.95
truck	0.181256	0.434774	0.977436	5.5	Medium	0.95
bus	0.180607	0.437096	0.977038	6.0	Medium	0.95
traffic_cone	0.181641	0.435008	0.976332	6.5	Medium	0.95
barrier	0.182271	0.434331	0.977185	6.5	Medium	0.95
trailer	0.178449	0.433590	0.976840	1.0	Low	1.05
bicycle	0.179253	0.434264	0.977334	2.0	Low	1.05
background	–	–	–	–	Background	1.20

improves the baseline by +1.49% mAP and +1.66% NDS. It also yields lower true-positive error metrics than the baseline, with only a slight increase in mAVE. Overall, these results show that preserving semantically important regions during pretraining improves downstream detection.

Table II summarises the semantic class importance analysis used to construct the final masking policy. Since the downstream task is 3D BEV object detection on the 10 nuScenes detection categories, we mapped the 32 raw semantic labels to the corresponding detection classes wherever possible, while all remaining labels were treated as background. We then ranked the classes using voxel-level LiDAR reconstruction metrics, where lower Chamfer distances indicate better reconstruction and higher occupancy accuracy indicates better voxel prediction. Classes that caused larger degradation in reconstruction when masked were treated as more important and were therefore protected more strongly in the final masking policy. Based on this ranking, each class was assigned an importance level and a masking weight that controls how strongly that class is masked under the fixed overall LiDAR masking ratio.

c) Decoder-side point semantic supervision: As shown in Table I, the best decoder-side point semantic supervision setting achieved 26.11% mAP and 34.63% NDS with $\lambda_{\text{sem}} = 0.25$, improving over the random masking baseline by +1.39% mAP and +3.22% NDS. Compared with semantic masking, it achieved slightly lower mAP (26.11% vs. 26.21%) but higher NDS (34.63% vs. 33.07%), indicating stronger overall detection quality. It also achieved lower mATE, mASE, mAVE, and mAAE than both the baseline and semantic masking method, while mAOE remained slightly higher than semantic masking method. Overall, these results show that auxiliary point-wise semantic supervision improves downstream 3D BEV object detection and gives the best overall NDS among all evaluated settings.

d) Post-MMIM semantic supervision (ablation): As an ablation, we also evaluated a post-MMIM semantic supervision method in which the semantic decoder is attached after MMIM and before the decoder, using post-MMIM LiDAR voxel features together with local point offsets for semantic prediction. As shown in Table I, the best post-MMIM setting achieved 26.09% mAP and 32.96% NDS with $\lambda_{\text{sem}} = 0.25$, improving over the random masking baseline by +1.37% mAP and +1.55% NDS. Although this confirms that semantic supervision is beneficial even before the decoder, it remained below the method introduced in Section II-C, especially in NDS. This observation is intuitive: cascading the semantic and LiDAR reconstruction decoders enables supervision over both masked and visible voxels, whereas the post-MMIM branch primarily supervises only visible voxels.

IV. Conclusion

In this paper, we presented a semantics-guided multi-modal masked autoencoder pretraining framework that can improve downstream 3D BEV object detection. Building on the UniM^{2}AE baseline, we introduced semantic information only during pretraining through two separate extensions: semantics-guided LiDAR voxel masking, which preserves semantically important LiDAR regions more strongly, and auxiliary point-wise LiDAR semantic supervision, which complements reconstruction with explicit semantic guidance for LiDAR features. Experiments on the nuScenes mini validation set demonstrated that both extensions improved performance over the uniform random masking baseline. Semantics-guided LiDAR voxel masking improved performance by +1.49% mAP and +1.66% NDS, achieving the best mAP, while decoder-side point semantic supervision improved performance by +1.39% mAP and +3.22% NDS, achieving the best overall NDS and outperforming the post-MMIM method. Taken together, these results indicate that uniform random masking and reconstruction alone are not sufficient to learn the most effective multimodal representations, and that incorporating semantic information during pretraining leads to better downstream detection.

For future work, an important next step is to combine semantics-guided LiDAR voxel masking and auxiliary point-wise LiDAR semantic supervision within a single pretraining framework, to evaluate their joint effect after transfer to downstream detection, and to explore the use of predicted semantic labels from a semantic segmentation model instead of ground-truth semantic annotations to guide the masking stage. Another important direction is to introduce semantic information into the image branch through semantics-aware image masking or auxiliary semantic supervision for image features. It would also be valuable to evaluate the proposed ideas on larger datasets beyond nuScenes mini and to study their generalisation more broadly across additional datasets.

REFERENCES

[1] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781, IEEE, 2023.

[2] J. Zou, T. Huang, G. Yang, Z. Guo, T. Luo, C.-M. Feng, and W. Zuo, "Unim 2 ae: Multi-modal masked autoencoders with unified 3d representation for 3d perception in autonomous driving," in *European Conference on Computer Vision*, pp. 296–313, Springer, 2024.

[3] G. Hess, J. Jaxing, E. Svensson, D. Hagerman, C. Petersson, and L. Svensson, "Masked autoencoder for self-supervised pre-training on lidar point clouds," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 350–359, 2023.

[4] C. Min, L. Xiao, D. Zhao, Y. Nie, and B. Dai, "Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 7, pp. 5150–5162, 2023.

[5] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21769–21780, 2023.

[6] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng, "Semmae: Semantic-guided masking for learning masked autoencoders," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14290–14302, 2022.

[7] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. 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.