

SemLT3D: Semantic-Guided Expert Distillation for Camera-only Long-Tailed 3D Object Detection

Supplementary Material

The supplementary material provides additional experiments, analyses, and qualitative results to further support our main findings:

- **A. Extended Ablation Studies:** We examine the impact of different configurations of the proposed language-guided mixture-of-experts (LMoE) module discussed in Sec. 3.1, as well as the performance of the model with various weights for the knowledge distillation loss \mathcal{L}_{KD} (which is set to 0.5 as stated in Sec. 4.2).
- **B. Data Analysis:** We present the category distributions in both the nuScenes and Argoverse 2 (AV2) benchmarks, highlighting the pronounced long-tail imbalance. Besides, we show more samples in tail classes to illustrate challenges of inter-class diversity and intra-class ambiguity.
- **C. Additional Quantitative Analysis:** Building on the insights from our data analysis, we conduct deeper evaluations on both nuScenes and AV2 benchmarks, including comparisons against additional baselines on subsets of categories that exhibit strong intra-class diversity and intra-class ambiguity. We also report per-class performance to highlight the effectiveness of **SemLT3D** across individual categories.
- **D. Additional Qualitative Results:** We provide more qualitative examples across diverse tail and head categories, along with visual illustrations of the enhanced feature representations obtained after integrating our proposed components.
- **E. Additional Discussion:** We further discuss the performance of our method compared to other long-tailed recognition methods in both 2D and LiDAR-based settings, analyze performance trade-offs, and examine our approach to occlusion handling.

Table S-I. More Configurations of LMoE ablation on nuScenes val split.

Experts	Top-k	mAP	NDS	Many	Medium	Few
2	1	27.58	38.52	50.66	31.0	3.90
4	1	28.56	40.26	49.50	33.20	5.90
4	2	29.59	40.94	50.92	34.55	6.03
4	4	28.62	40.31	50.26	33.8	4.62
8	1	28.24	40.4	47.84	31.76	8.60
8	2	28.54	39.24	47.96	32.06	8.21
8	4	27.88	39.85	47.70	33.20	6.90

A. Extended Ablation Studies

Ablations on expert count and routing for LMoE. Table S-I presents additional ablations on our language-guided MoE design from Table 5 in our main paper. We observe a consistent trend: increasing the number of experts improves performance on tail classes, as the model can partition the semantic space into finer groups, making it easier to model subtle variations and mitigate intra-class diversity. However, using too many experts leads to reduced performance on medium-shot and many-shot classes, likely due to insufficient expert activation and diminished feature sharing. On the other hand, the top- k routing analysis shows a consistent and intuitive trend in which top-2 gating performs more reliably than top-1 across all expert configurations, while expanding the selection to top-4 fails to provide additional benefits and can even degrade overall performance. This limitation arises because activating too many experts weakens their ability to specialize, preventing each expert from developing a clear and meaningful role. These observations align with prior MoE studies such as [4, 8, 9], where top-2 gating has been widely adopted due to its stability and strong empirical behavior. Guided by these trends, we adopt a configuration of 4 experts with top- $k = 2$ as our default setting.

Table S-II. Ablation on weight of distillation loss on nuScenes val split.

λ_{kd}	mAP	NDS	Many	Medium	Few
0	28.24	39.32	51.60	30.70	4.02
0.2	28.70	40.45	50.40	32.42	5.02
0.5	29.59	40.94	50.92	34.55	6.03
1	29.02	40.2	49.82	35.51	5.82

Ablation study on the weight of knowledge distillation loss \mathcal{L}_{KD} . To clarify the full training setup, we summarize the overall objective that integrates all loss components introduced in the main paper. The training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}}, \quad (1)$$

Following prior works [2, 6], we set $\lambda_{\text{cls}} = 2$, $\lambda_{\text{reg}} = 0.25$, and $\lambda_{\text{contrast}} = 1$ by default. To determine an appropriate weight for the knowledge distillation module, we conduct an ablation study on λ_{KD} , as shown in Table S-II. We observe that increasing λ_{KD} generally improves performance

up to $\lambda_{KD} = 0.5$, after which performance begins to degrade. We attribute this drop to noise introduced by imperfect 2D semantic cues, particularly in challenging cropped regions such as heavily occluded objects.

B. Data Analysis

Category distribution. We examine the category distribution in two representative street-scene understanding benchmarks: nuScenes and Argoverse 2 (AV2). As visualized in Figure S-I and Figure S-II, both datasets exhibit a pronounced long-tailed distribution, with a small number of head classes dominating the data while many tail classes are severely underrepresented. This imbalance highlights trends commonly observed in large-scale real-world datasets and underscores the necessity of developing effective methods that can robustly address long-tailed recognition challenges.

Visualization of challenging samples in tail classes: Beyond the distribution itself, our qualitative analysis reveals that long-tail difficulties arise from two major factors: *inter-class ambiguity* and *intra-class diversity*.

As illustrated in Figure S-III, nuScenes contains strong inter-class ambiguity among tail classes (e.g., Police Officer vs. Construction Worker) and between head and tail classes (e.g., Car vs. Police Vehicle). Simultaneously, many long-tail categories exhibit substantial intra-class diversity (e.g., Debris, Trailer), further complicating recognition.

A similar pattern appears in the AV2 dataset, illustrated in Figure S-IV. We observe inter-class ambiguity between long-tail classes (e.g., Wheelchair vs. Wheeled Rider, Stop Sign vs. Sign) and significant intra-class diversity in the long-tail categories such as Large Vehicle and Vehicle Trailer.

C. Additional Quantitative Analysis

In-depth comparisons with baselines on tail categories.

To further quantify the challenges posed by long-tailed street-scene data, we evaluate performance specifically on categories characterized by inter-class ambiguity and intra-class diversity.

As shown in Figure S-V (left), within the same superclass (e.g., *human*), existing baselines such as RayDN, StreamPETR, and SparveBEV achieve competitive results on many-shot (head) categories like Adult. However, their performance drops substantially on few-shot (tail) categories such as Child, Construction Worker, and Police Officer, where fine-grained distinctions are required. This consistent degradation across baselines indicates that current models inadequately capture subtle semantic and visual cues necessary to resolve inter-class ambiguity in long-tail scenarios. In contrast, our method is

explicitly designed to enhance class disambiguation in low-data regimes, yielding significantly more robust detections on these challenging categories.

Furthermore, we analyze categories with high intra-class diversity, as in Figure S-V (right). Tail classes such as Trailer and Debris exhibit substantial visual variation despite having very few training examples. The combination of few-shot scarcity and large appearance variation makes these categories particularly difficult for prior methods, leading to consistently low recall and unstable predictions. Our approach demonstrates clear improvements on these categories, highlighting its ability to model broad intra-class diversity even under limited supervision.

Per-class evaluation on nuScenes [1] and AV2 [7].

Figures S-VI and S-VII present per-class mAP comparisons between our SemLT3D and the runner-up method for each benchmark.

On nuScenes (Figure S-VI), SemLT3D is compared against the runner-up StreamPETR. Our approach consistently improves detection performance across categories that suffer from substantial inter-class ambiguity and intra-class diversity, with particularly pronounced gains on difficult tail and medium-shot classes such as Stroller, Construction Worker, and Police Officer.

A similar pattern emerges on AV2 (Figure S-VII), where SemLT3D outperforms the runner-up Far3D, especially on long-tail categories including Truck and Stroller. Although slight performance reductions appear in a small number of head classes, largely attributable to the model’s emphasis on long-tail robustness, SemLT3D remains highly competitive overall. Importantly, as summarized in Table 1 of the main paper, our framework retains both computational efficiency and real-time inference speed (FPS), introducing no additional computational or memory overhead and preserving a deployment-friendly architecture.

D. Additional Visualizations

More qualitative results. Figure S-VIII presents additional qualitative comparisons between SemLT3D and baseline model across various tail categories where inter-class ambiguity and intra-class diversity are particularly pronounced. The results show that SemLT3D is able to consistently localize and classify tail instances, even when they are heavily occluded or extremely small, demonstrating its enhanced ability to learn robust and discriminative features under challenging long-tail conditions.

Improvement in feature learning. To further demonstrate the improvement in feature representation after integrating the proposed method, we visualize example feature maps for several tail categories in Figure S-IX. The visualizations show that our model produces more focused and discriminative activations on tail instances, while baseline models exhibit diffuse or ambiguous patterns. As a result, SemLT3D

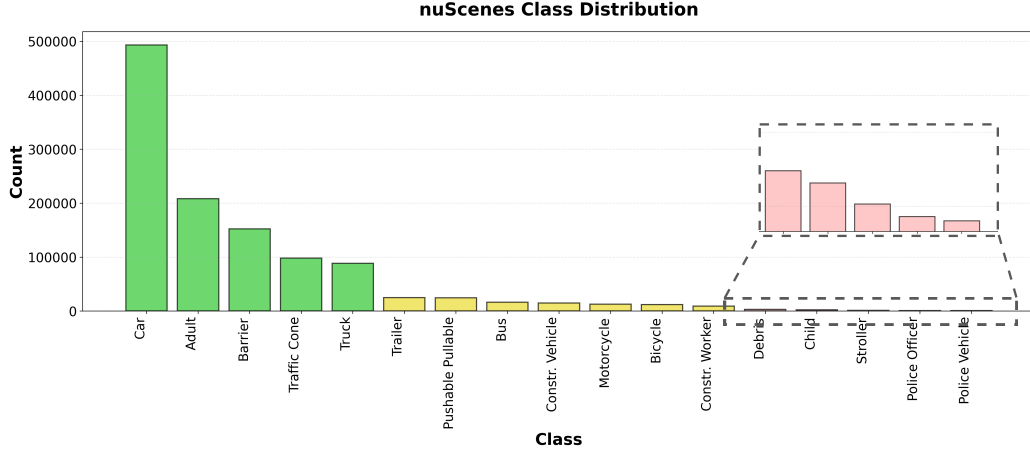


Figure S-I. Per-class distribution on the nuScenes [1] validation set. The bar chart shows the number of instances per category, illustrating the pronounced long-tail imbalance. ■ corresponds to Many-shot classes, ■ indicates Medium-shot classes, and ■ represents Few-shot classes.

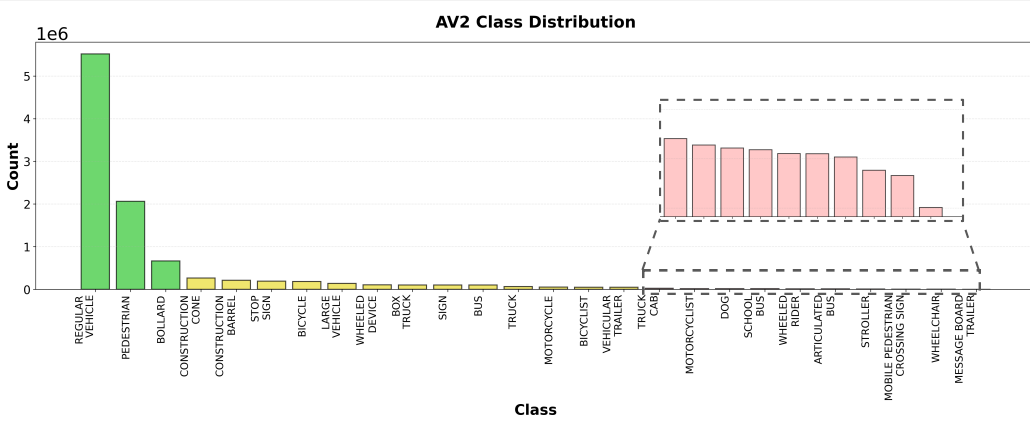


Figure S-II. Per-class distribution on the AV2 [7] validation set. The bar chart shows the number of instances per category, illustrating the pronounced long-tail imbalance. ■ corresponds to Many-shot classes, ■ indicates Medium-shot classes, and ■ represents Few-shot classes.

is better able to highlight and localize these challenging categories, whereas the baselines often fail to detect them.

E. Additional Discussion.

Table S-III. SemLT3D vs. StreamPETR with long-tail techniques.

	Loss/Method	All	Many	Medium	Few
R50	StreamPETR				
	Focal Loss	26.97	53.32	28.53	3.22
	Equal. Loss	25.65	51.1	28.21	1.42
	Cls. Bal. Loss	26.25	52.76	28.34	1.93
	CBGS	26.20	46.42	26.55	4.80
	Sem. Hierarchy	27.43	49.64	28.32	4.20
	SemLT3D (ours)	29.59	50.92	34.55	6.03

Performance comparison with standard long-tail techniques. We first note that all baselines, including StreamPETR, already employ Focal Loss [5] to mitigate class imbalance. To provide a more comprehensive comparison, we retrain the runner-up StreamPETR with several widely used long-tailed recognition techniques, as summarized in

Table S-III. While some of these approaches yield modest improvements on tail categories, they often come at the cost of degraded performance on many-shot classes. In contrast, SemLT3D consistently outperforms all variants by substantial margins across all category groups, demonstrating that our semantic-guided framework addresses long-tailed challenges more effectively than conventional re-balancing or re-weighting strategies.

Table S-IV. SemLT3D vs. StreamPETR with ViT backbone.

	Method	All	Many	Medium	Few
ViT	StreamPETR	38.23	58.4	40.06	14.55
	Ours	41.1	62.08	40.62	20.77

Table S-V. Performance and computational cost at inference.

	Method	mAP	FPS	Params
R50	StreamPETR_2048	26.97	32.36	37M
	StreamPETR_4096	26.97	31.54	43M
	Ours (w/o LMoE)	28.30	32.36	37M
	Ours (w LMoE)	29.59	29.40	41M

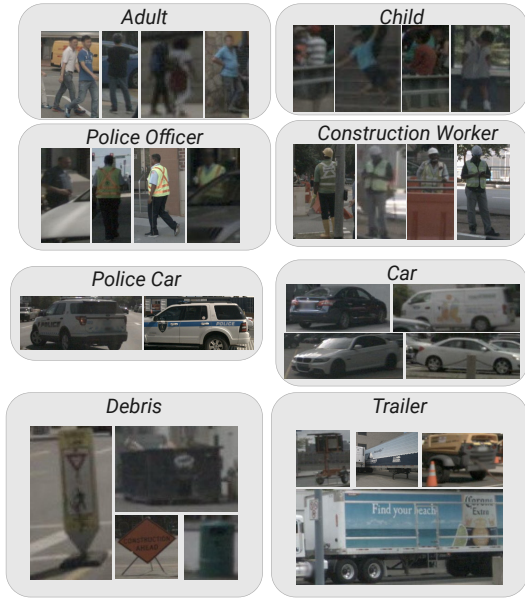


Figure S-III. More samples for data analysis of challenging (inter-class diversity and intra-class ambiguity) samples in tail categories in nuScenes benchmarks [1].

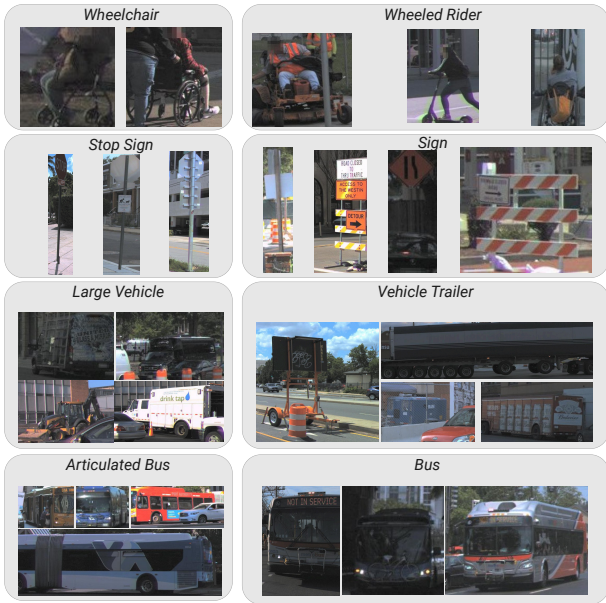


Figure S-IV. More samples for data analysis of challenging (inter-class diversity and intra-class ambiguity) samples in AV2 benchmarks [7].

Performance trade-off analysis. As shown in Table S-IV and Table S-V, SemLT3D achieves a favorable balance between detection accuracy and computational efficiency. With the ResNet-50 backbone, our full framework (with

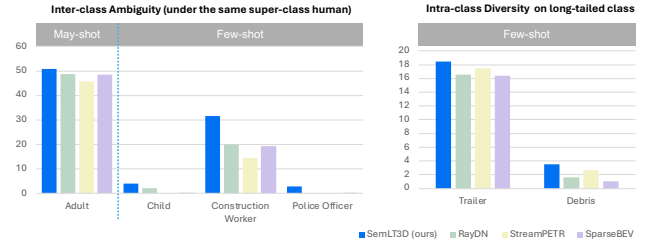


Figure S-V. Additional category mAP comparisons on the inter-class ambiguity (left) and intra-class diversity (right). The results are reported on nuScenes dataset.

LMoE) introduces only 4M additional parameters (a 10% increase) over the baseline StreamPETR, while delivering a 2.62 mAP improvement and maintaining real-time inference at 29.40 FPS. Notably, simply scaling the baseline’s feed-forward dimension from 2048 to 4096 adds 6M parameters yet yields no accuracy gain, confirming that our improvements stem from the proposed semantic-guided architecture rather than increased model capacity.

Furthermore, the non-LMoE variant of SemLT3D, which relies solely on the SPD distillation and contrastive alignment modules, achieves a 1.33 mAP improvement with no additional parameters or latency overhead compared to the baseline. This demonstrates that a significant portion of our performance gains can be obtained at zero inference cost, as the distillation branch is discarded after training.

When upgrading to a ViT backbone, SemLT3D further improves to 41.1 mAP, outperforming StreamPETR by 2.87 mAP across all category groups, including a 6.22 mAP gain on few-shot classes. This result indicates that stronger visual backbones can better leverage the semantic supervision provided by our framework, effectively eliminating the minor trade-off on many-shot classes observed under the more constrained ResNet-50 setting. Overall, these results suggest that SemLT3D offers a practical and scalable solution: practitioners can choose the non-LMoE variant for a cost-free accuracy boost, or enable LMoE for further gains with only a marginal increase in latency.

Occlusion handling discussions. Occlusion in multi-view 2D crops mainly occurs in two cases: (i) The object is visible to at least one view: SPD distills from all cameras, so clean views provide reliable object semantics while partially occluded views mainly add context. (ii) The object is occluded in all views: heavy occlusion shifts embeddings toward surrounding context, but the target object is still preserved in higher ranks. To prove this, we use similarity-based CLIP probe (ViT-B/16) on projected 2D boxes and observe that the true object appears in Top-5/Top-10 predictions with 75.42/87.22% hit rate, indicating that information of occluded target object semantics are largely preserved. To evaluate capability of SemLT3D in handling occlusion,

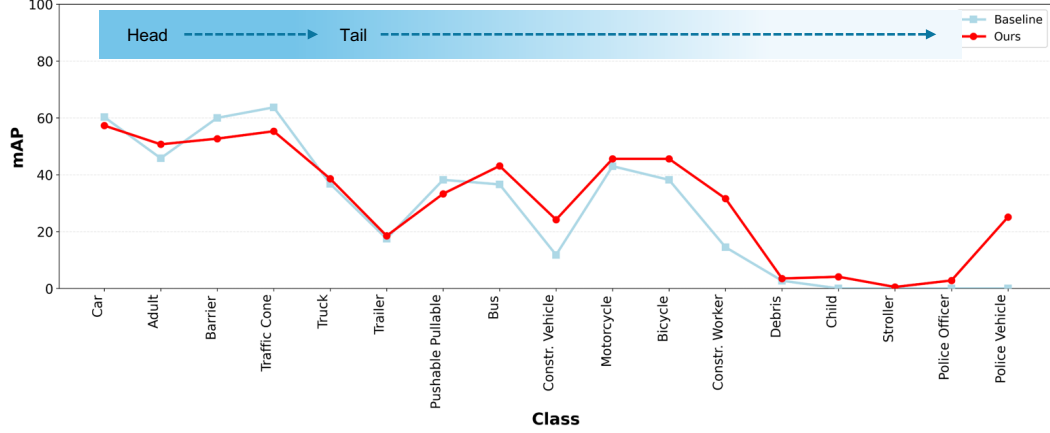


Figure S-VI. Per-class performance comparison nuScenes [1] validation set. The line plots compare the per-class mAP between our SemLT3D — and the baseline, the runner-up StreamPETR [6] —.

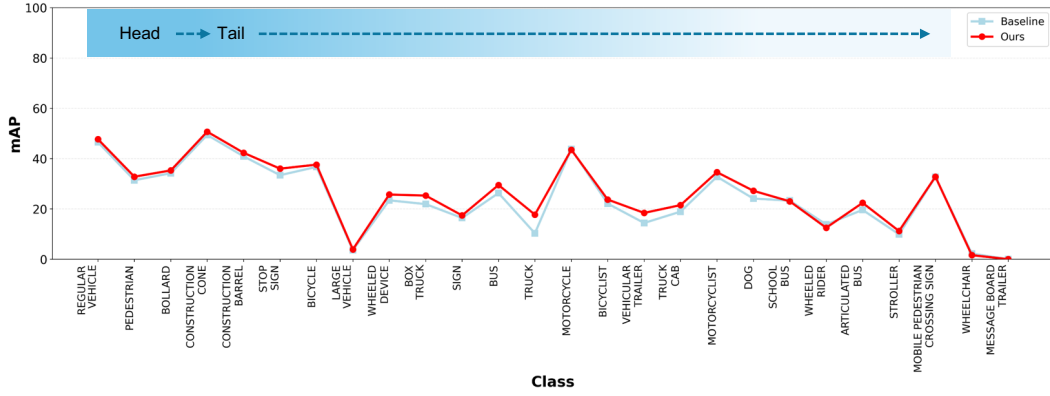


Figure S-VII. Per-class performance comparison AV2 [7] validation set. The line plots compare the per-class mAP between our SemLT3D — and the baseline, the runner-up Far3D [3] —.

we stratify objects by visibility using provided annotations and compare Recall against StreamPETR in Table S-VI.

Table S-VI. Recall score under various visibility levels.

	Method	[0,40]	[40,60]	[60,80]	[80,100]
R50	StreamPETR	46.13	55.28	57.71	59.01
	Ours	56.92	64.33	65.39	71.03

References

- [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nusenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 3, 4
- [2] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23 (120):1–39, 2022. 0
- [3] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2561–2569, 2024. 4
- [4] Dmitry Lepikhin, HyounJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv*

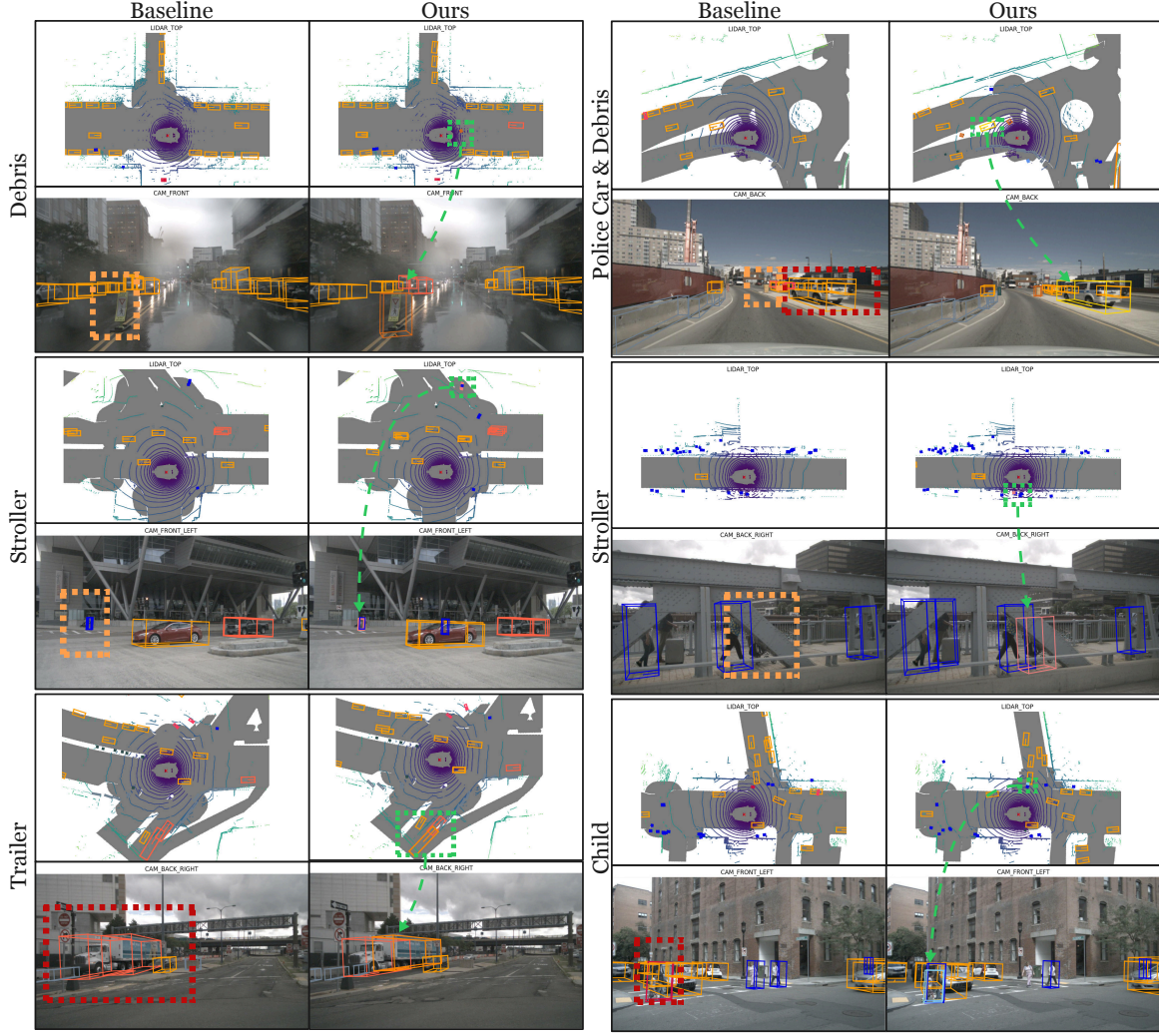


Figure S-VIII. Qualitative comparison between baseline (Left) and our **SemLT3D** (Right). The **orange** dashed box highlight failure cases caused by intra-class diversity, while the **red** dashed box indicate failure cases arising from inter-class ambiguity. The **green** dashed lines and boxes denote true positive predictions in the LiDAR and camera views

preprint arXiv:2006.16668, 2020. 0

- [5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2
- [6] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xianguyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3621–3631, 2023. 0, 4
- [7] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Process-*

ing Systems Track on Datasets and Benchmarks (*NeurIPS Datasets and Benchmarks 2021*), 2021. 1, 2, 3, 4

- [8] Chang-Bin Zhang, Yujie Zhong, and Kai Han. Mr. detr: Instructive multi-route training for detection transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9933–9943, 2025. 0
- [9] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022. 0

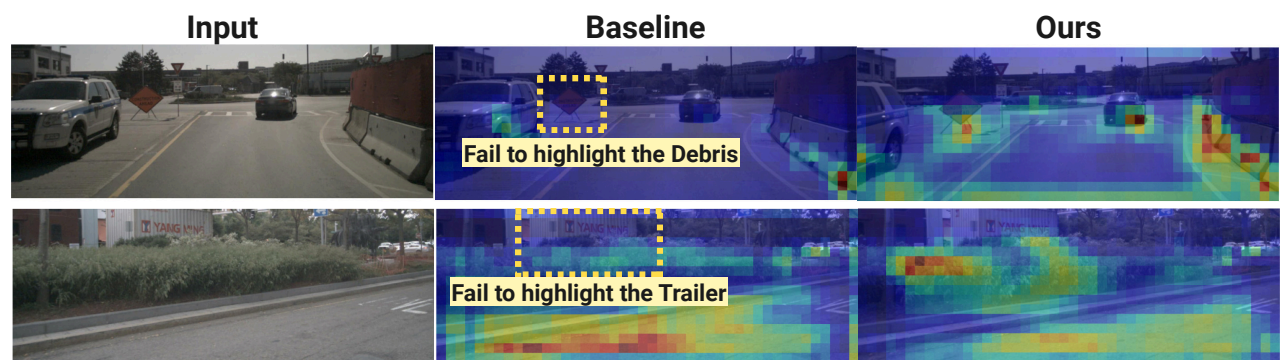


Figure S-IX. Visualization of feature map extracted from the trained backbone between our baseline and proposed SemLT3D. The result show that by introduced SemLT3D method help to enhance the responses of model on tail-class.