

---

# SQS: Enhancing Sparse Perception Models via Query-based Splatting in Autonomous Driving

## *Supplementary Materials*

---

### A Additional Implementation Details

#### A.1 Evaluation Metrics

**3D Object Detection.** Following the evaluation standards described in prior works [2, 1, 4], we employ the NuScenes Detection Score (NDS) and mean Average Precision (mAP) as the main quantitative measures for 3D object detection evaluation.

Unlike conventional approaches based on 3D Intersection over Union (IoU), the mAP here is calculated by measuring the center distance between predictions and corresponding ground truths projected onto the ground plane.

In addition, the nuScenes benchmark specifies five true positive (TP) metrics — ATE, ASE, AOE, AVE, and AAE — which evaluate different error types: translation, scale, orientation, velocity, and attribute errors, respectively. Each TP metric is reported per category, and their averages over all classes are denoted as mATE, mASE, mAOE, mAVE, and mAAE.

Leveraging these TP errors, the TP score is defined as  $TP_{\text{score}} = \max(1 - TP_{\text{error}}, 0)$ . Subsequently, the nuScenes detection score (NDS) is computed as:

$$NDS = \frac{1}{10} \times \left[ 5mAP + \sum TP_{\text{score}} \right], \quad (1)$$

to encapsulate all aspects of the nuScenes detection comprehensively.

**Semantic Occupancy Prediction.** The mIoU and IoU for evaluating the 3D semantic occupancy prediction task are defined as follows:

$$mIoU = \frac{1}{|C'|} \sum_{i \in C'} \frac{TP_i}{TP_i + FP_i + FN_i}, \quad IoU = \frac{TP_{\neq c_0}}{TP_{\neq c_0} + FP_{\neq c_0} + FN_{\neq c_0}}, \quad (2)$$

where  $C'$ ,  $c_0$ ,  $TP$ ,  $FP$ ,  $FN$  denote the set of nonempty classes, the empty class, the number of true positives, false positives and false negative predictions, respectively.

### B Additional Discussions

#### B.1 Fully Utilization of Pre-training Queries

In the limitation section, we acknowledged that the existing framework does not fully harness the potential of pre-trained queries across diverse downstream tasks. As elaborated in Section 3.3 of the main paper, SQS employs a self-supervised splatting module to derive contextual 3D Gaussian representations from sparse queries for holistic scene reconstruction during pre-training. Nonetheless, the roles of these queries vary substantially depending on the downstream objective. For instance, in occupancy estimation, the queries correspond to both foreground and background regions, whereas in object detection, only foreground-related queries are actively optimized. In the

Table 1: **Ablation on the number of Gaussians.** The latency and memory are tested with batch size one during inference.

Number of Gaussians	Latency	Memory	mIoU	IoU
7500	<b>180</b> ms	<b>4615</b> M	15.2	26.6
12500	198 ms	4650 M	17.8	28.1
25600	210 ms	4680 M	18.0	<b>28.5</b>
51200	259 ms	4796 M	<b>18.4</b>	28.4
144000	372 ms	5635 M	18.2	28.3

current SQS formulation, the query interaction module adaptively learns from pre-trained queries without explicitly accounting for such functional divergence.

This inconsistency poses difficulties in transferring knowledge from pre-training to task-specific fine-tuning, as the model lacks an explicit mechanism to distinguish and exploit queries with different semantic significance. A potential research avenue is to incorporate explicit semantic supervision—such as pseudo-label assignment or clustering-based semantic grouping—into the pre-training phase to categorize queries by their semantic attributes. Doing so would enable more adaptive and task-aware query interactions in downstream applications, allowing selective utilization or weighting of queries based on semantic relevance. Such an enhancement could promote more effective knowledge transfer and improve performance, particularly in tasks that demand fine-grained semantic reasoning.

Introducing this semantic distinction during pre-training would allow queries to encode more discriminative and task-relevant features, boosting their transferability and downstream task adaptability. For instance, in the object detection head, only queries pre-trained with ‘foreground’ semantics would participate in the detection decoder, leading to improved sample efficiency and interpretability.

## C Additional Experimental Results

### C.1 Impact of the Number of Gaussians

Since SQS can be integrated into any SPM, memory usage and latency become increasingly critical factors. In this context, we also investigate the impact of the number of Gaussians during the pre-training phase, specifically to analyze the associated latency and memory overhead introduced by SQS. The corresponding results are presented in Tab. 1. From these results, it can be observed that the number of Gaussians exhibits a proportional relationship with both latency and memory consumption. However, during the fine-tuning stage, increasing the number of Gaussians does not lead to a commensurate improvement in performance. This phenomenon may be attributed to the enlarged disparity between the pre-training reconstruction objective and the fine-tuning perceptual task as the number of Gaussians increases. Therefore, to maintain a balance between performance and resource usage, we set the default number of Gaussians to 25,600.

### C.2 Computational Overhead and Possible Solutions

As illustrated in Tab. 2 for the occupancy prediction task with the quarter of training data during the pre-training and fine-tuning stage. Although our proposed SQS is flexible to enhance any SPMs, the plug-in mechanism introduces additional computational overhead (an increase of 60% latency) and memory consumption (an increase of 78% memory). To address this, we propose several feasible strategies:

1) **Tight Computation Application.** In scenarios with tight computational constraints, users can opt to load only the pre-trained image encoder (backbone and neck) and omit the query interaction module during fine-tuning (SQS<sup>†</sup> in the Tab. 2). This setup incurs no extra overhead compared to conventional models while still benefiting from the representational improvements gained through pre-training.

Table 2: **Variants of SQS to alleviate the computational and memory overhead.** † means a tight computation application, which omits the query interaction module during the fine-tuning stage. ‡ share the image encoder between the Gaussian queries (for query interaction) and downstream task queries.

Methods	IoU	mIoU	Latency	Memory
GaussianFormer	25.8	15.2	350ms	6100 M
GaussianFormer + SQS	28.5	18.0	560ms (↑ 60%)	10880 M (↑ 78%)
GaussianFormer + SQS†	28.2	17.5	350ms (↑ 0%)	6100 M (↑ 0%)
GaussianFormer + SQS‡	28.3	17.8	452ms (↑ 30%)	8250 M (↑ 35%)

2) **Partial Module Sharing.** It is also feasible to share the image encoder between the Gaussian queries (for query interaction) and downstream task queries, only enabling the Gaussian Transformer Decoder during fine-tuning. As detailed in the Tab. 2 for SQS‡, this results in a modest overhead (30% increase of latency and 35% increase of memory) but provides a favorable accuracy-resource trade-off.

3) **Future Optimization.** We are actively exploring model compression and distillation techniques that can further reduce memory consumption and computation, without sacrificing performance.

## D Additional Visualization Results

As illustrated in Fig. 1 and Fig. 2, we visualize an additional scene to qualitatively assess the efficacy of our method. From the Fig. 1, we found that our method is able to obtain 3D object detection results with less false positives and false negatives when compared to the SparseBEV [3] baseline. We also visualize the 3D semantic occupancy prediction results in Fig. 2. The results demonstrate the effectiveness of our method on this task.

## E Broader Impacts

The development of SQS offers significant potential for both positive and negative societal impacts, particularly in the context of autonomous driving technology and related fields.

For potential positive societal impacts, (1) **Enhanced Safety:** By improving the accuracy of 3D perception tasks like occupancy prediction and 3D object detection, SQS can contribute to safer autonomous vehicles, reducing the risk of accidents through better decision-making. (2) **Increased Efficiency:** SQS’s design may lead to more efficient autonomous driving systems, potentially reducing energy consumption and improving traffic flow. (3) **Advancements in Related Fields:** Techniques from SQS, such as query-based 3DGS reconstruction, can be applied to robotics, virtual/augmented reality, and 3D content creation, fostering innovation in these areas

For potential negative societal impacts, (1) **Privacy Concerns:** Autonomous vehicles equipped with advanced perception systems may raise privacy concerns related to data collection and surveillance. (2) **Reliability and Safety Concerns:** While SQS aims to improve safety, failures or errors in the system could still lead to accidents and harm, especially in corner cases. Ensuring the reliability and robustness of the system is critical. (3) **Fairness Considerations:** Deployment of autonomous driving technologies could unfairly impact specific groups if the systems are not designed and tested to account for diverse populations and environments.

## References

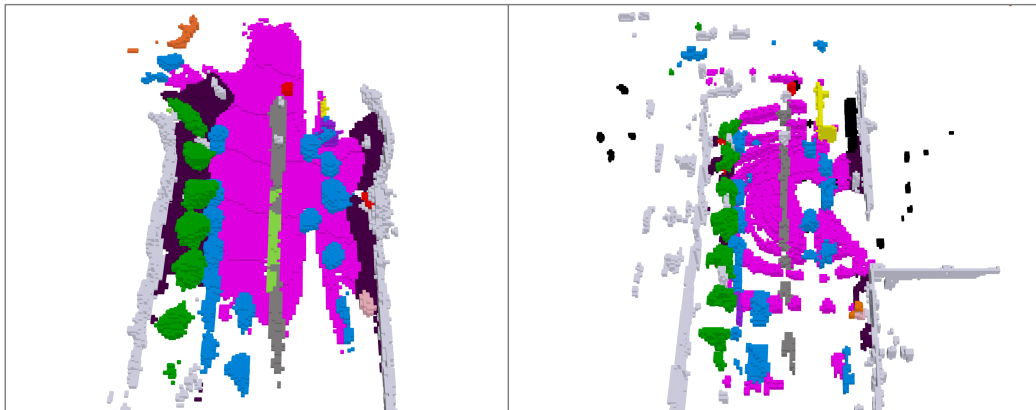
- [1] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, abs/2112.11790, 2021.
- [2] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022.
- [3] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023.
- [4] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15238–15250, 2024.



Figure 1: **Visualization of 3D object detection results.** The top and bottom two rows denote detection results of SparseBEV and ours, respectively. The false negatives and false positives are annotated with red dotted circles and green dotted circles, respectively.



Multi-view Image Inputs



Occupancy Prediction

Occupancy Ground Truth



Figure 2: **Visualization of semantic occupancy prediction results.** Our method could predict the occupancy accurately.