2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Class-Aware Queries for Robust Multi-View 3D Object Detection

Chaeyeon Sung[1,2], Sungmin Woo[1], Sangyoun Lee[1]

***Abstract*— Query-based multi-view 3D object detectors typically rely on a fixed set of learnable queries that jointly predict object categories and locations. However, encoding both semantic and geometric information within a shared query embedding leads to representational conflicts, limiting optimization. While prior works decouple prediction heads to partially address this issue, such decoupling often treats classification and localization as independent tasks, leaving the queries themselves class-agnostic and unaware of the scene's semantic context. In this paper, we present the first 3D object detection framework that constructs class-aware queries using scene-level object class predictions. Specifically, a multi-view image classifier first estimates which object classes are present in the scene, and these predictions are used to generate semantically guided queries for 3D localization within the transformer decoder. This allows our model to initialize each query with class-specific priors, in contrast to conventional uniform query initialization. As a result, queries attend more effectively to relevant regions and objects throughout decoding. Experiments on the nuScenes benchmark show that our method improves mAP by 2.7 points and NDS by 1.5 points over a strong DETR-based baseline. An oracle study further reveals that classification accuracy is a key bottleneck in existing DETR-style detectors, highlighting the benefit of early semantic guidance. The code is publicly available at https://github.com/ssungchae/CaQ3D.**

I. INTRODUCTION

3D object detection is a fundamental task in autonomous driving, enabling the system to recognize object categories and localize them in 3D space. Among the sensing modalities, LiDAR and cameras are the two most widely used for 3D perception. LiDAR sensors provide accurate geometric structure and depth information, but their high cost, hardware complexity, and energy consumption hinder large-scale deployment. In contrast, cameras are lightweight, low-cost, and capture high-resolution appearance cues across wide fields of view. With recent advances in multi-view modeling and depth estimation [17], [20], [23], camera-based 3D object detection has made rapid progress, making it an increasingly practical and competitive direction for autonomous driving.

Transformer-based architectures have recently become the foundation for camera-based 3D object detection. Following the DETR [4] formulation, detection is cast as set prediction with a fixed number of learnable queries. Each query interacts with multi-view image features through attention and is decoded to predict both the object category and its 3D bounding box. The decoder refines predictions layer by layer, and the shared query embedding serves as the central representation, so classification and localization are produced jointly from the same query-based pipeline [1], [2], [24].

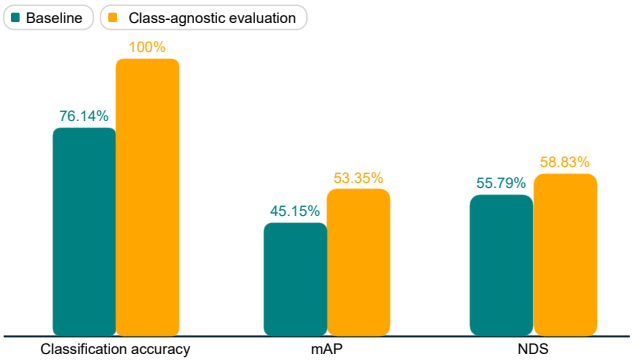

Fig. 1. Class-agnostic evaluation of a camera-only 3D object detection model. By assigning correct class labels to all predicted boxes without modifying their locations, we observe a significant performance boost: mAP increases from 45.15% to 53.35% and NDS from 55.79% to 58.83%. This highlights the critical impact of classification errors on overall detection metrics.

However, using a single representation to serve both objectives can introduce conflicts because classification mainly relies on appearance cues such as color, texture, and fine-grained patterns, whereas localization depends on geometric cues such as shape, scale, and spatial position. This coupling may lead to feature and prediction misalignment, and prior work has suggested that partially disentangling the pathways can help mitigate such conflicts [7]. At the same time, classification and localization are not entirely independent, and certain correlations may exist between them. Thus, the key consideration is how to reduce potential conflicts while still leveraging these correlations.

To better understand how these correlations translate into practical outcomes, we conducted experiments and observed that classification accuracy constitutes a critical bottleneck for detection performance. We therefore conducted an oracle experiment in which all predicted boxes were assigned their correct class labels while their predicted locations were kept unchanged. As shown in Figure 1, this simple adjustment yields a substantial performance boost: mean Average Precision (mAP) increases by 8.2% and nuScenes Detection Score (NDS) by 3.0%. This finding highlights that transformer-based detectors are highly sensitive to classification quality, especially under long-tailed distributions in autonomous driving datasets. Therefore, prior approaches have largely overlooked the critical role of classification, treating it without explicit focus in query design. By explicitly enhancing classification, additional performance gains can be unlocked.

Motivated by these observations, we propose a novel framework that addresses the representational conflicts be-

*Corresponding author: syleee@yonsei.ac.kr
[1]C. Sung, S. Woo, and S. Lee are with Yonsei University, Seoul, South Korea (e-mail: {chaen, smw3250, syleee}@yonsei.ac.kr).
[2]C. Sung is with Hyundai Motor Company, South Korea (e-mail: chaen@hyundai.com).

tween semantic classification and geometric localization in query-based 3D object detection. Unlike conventional approaches that jointly predict object categories and 3D bounding boxes using shared query embeddings, our method explicitly separates the two objectives, as illustrated in Figure 2. In particular, class information is predicted through a dedicated classification branch, and this information is then used as prior guidance when performing geometric localization. This class-guided design avoids the conflicts of a shared representation while effectively integrating classification information into localization, enabling a more robust and accurate detection process.

At the core of our approach is a multi-view scene-level classifier that aggregates information from all camera views to predict which object classes are present in the scene. By incorporating this semantic context early in the detection pipeline, queries receive class-aware guidance from the beginning, which helps semantic and geometric objectives remain connected without being forced into a single entangled representation. Specifically, the predicted scene-level class information is used to generate class-aware queries by combining class embeddings with spatial priors, allowing the transformer decoder to integrate class information with geometric information for more reliable localization, even in cluttered or ambiguous environments. To account for real-world conditions where classifier predictions may be noisy, we adopt a two-stage training strategy: the model is first pretrained with ground-truth scene labels and then fine-tuned with predicted labels to simulate inference conditions. In this process, label noise is injected into the embeddings so that the model becomes robust to potential classification errors made by the scene-level classifier. In addition, perturbed boxes with injected noise are used as a form of augmentation, allowing the model to learn to recover reliable bounding boxes even when the initial box predictions are noisy, thus improving the robustness of the learnable box representation.

To summarize the key innovations of our work, we highlight the following contributions:

- We propose a novel class-guided framework for query-based 3D detection that resolves conflicts between classification and localization by explicitly injecting class guidance into queries.
- We design a class-aware query generation mechanism that combines class embeddings with spatial priors to disentangle class and geometric cues while keeping them complementary.
- We develop a two-stage training schedule that transitions from ground-truth labels to predicted labels, allowing stable optimization before adapting to prediction-based guidance.
- We introduce noised box-level guidance, where noised labels improve robustness to class guidance variations and noised anchors act as geometric perturbations to sharpen localization.

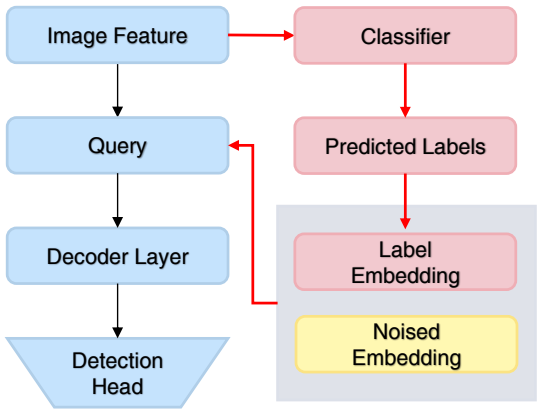

Fig. 2. Classification-guided detection framework. Unlike standard query-based detectors that jointly predict class and box in the decoder, our method first performs scene classification from image features. The predicted categories are converted into class embeddings and injected into the queries before decoding, enabling early incorporation of semantic priors. This class-aware initialization improves both classification robustness and 3D localization accuracy.

II. Related Work

A. Query-based Object Detection

Recent query-based object detectors employ Transformer decoders to generate object predictions from a fixed set of learnable queries. This formulation enables a streamlined detection pipeline, where each query directly outputs a category label and a bounding box. DETR [4] pioneered this approach in 2D object detection, proposing a fully end-to-end model that eliminates region proposals and post-processing steps. Despite its elegant design, DETR suffers from slow convergence and requires long training schedules to reach strong performance [4]. To improve learning efficiency, several works have proposed injecting structured priors into the query representations. DAB-DETR [5] introduces learnable reference points as spatial priors, while DN-DETR [3] improves training stability by incorporating noisy ground-truth targets. DINO [8] further enhances this strategy with contrastive denoising and refined query selection. Building upon DETR-style architectures, the query-based paradigm has been extended to 3D object detection in multi-view settings. PETR [2] introduces 3D positional encoding into 2D image features, while DETR3D [10] projects 3D reference points into multi-view images for attention-based fusion. These methods predict object categories and 3D bounding boxes directly from queries in a single-stage pipeline, following the original DETR philosophy of end-to-end set prediction. In contrast, our approach introduces an explicit class guidance stage prior to query-based detection. A multi-view image classifier first predicts scene-level object categories, which are then used as class-level priors to guide the generation of detection queries. By providing this guidance before spatial localization, the model enables queries to focus on object categories that are likely to be present in the scene, leading to a more informed and context-aware query initialization compared to prior DETR-style 3D detection models.

B. Class-Aware and Decoupled Detection

A large body of work improves detection by injecting priors into queries. Open-vocabulary detectors (ViLD [11], RegionCLIP [12], OV-DETR [13], PromptDet [14]) exploit vision–language pretraining to provide semantic guidance, but they require textual inputs even at inference. In multi-view 3D detection, recent works instead propose vision-only priors: MvACon [32] strengthens feature lifting with attentive contextualization, Moon et al. [29] leverage temporal cues for motion and occlusion robustness, and Zhu et al. [30] learn class prototypes to cope with sparse labels. While effective, these methods mainly reinforce spatial or temporal reasoning under labeled supervision rather than conditioning queries on the overall class distribution. In contrast, our approach derives class-aware priors from a scene-level classifier and injects them at query initialization, requiring no text, prototypes, or pseudo-labels, and directly improving classification reliability under long-tailed distributions.

Many detection frameworks separate classification and regression at the prediction head while sharing intermediate representations. In DETR-style architectures, category prediction and box regression operate on the same query embeddings, which can introduce representational conflicts within the shared feature space. Decoupled DETR [7] mitigates this issue by assigning separate branches to classification and regression, mainly to improve overall detection accuracy. Our perspective differs in emphasis. We argue that classification quality itself often limits detection performance. Instead of further isolating the two objectives, we inject class-aware priors into the queries before decoding so that semantic information is available from the outset. This early conditioning allows attention to be influenced by class cues while geometric refinement proceeds, benefiting both classification and localization. Unlike prototype-based or temporal approaches that rely on additional annotated supervision, our method leverages scene-level classifier predictions to provide semantic guidance. As a result, it avoids requiring extra ground-truth labels and offers a more scalable solution for deployment.

III. Method

A. Overview

The proposed framework addresses representational conflicts between classification and localization in query-based 3D detection by structuring how class guidance is injected into queries. It comprises four components. First, a multi-view classifier predicts scene-level object categories from RGB inputs, providing global class guidance (Sec. III-B). Second, we integrate this class guidance into decoder queries by combining class embeddings with spatial priors, which disentangles class and geometric cues while keeping them complementary (Sec. III-C). Third, we employ a two-stage training schedule that transitions from ground-truth to classifier-predicted class guidance, enabling the detector to stabilize before adapting to prediction-based inputs (Sec. III-D). Finally, we introduce targeted perturbations at the box level, where noised labels disturb the class guidance to improve classification robustness and noised anchors perturb the geometric supervision to sharpen localization and stabilize learnable box representations (Sec. III-E). An overview is shown in Figure 3.

B. Multi-View Scene-Level Classification

We introduce a multi-label classifier that predicts which object categories are present in a scene from multi-view inputs. Each of the V camera views is processed independently, producing a probability vector $p_v \in [0,1]^C$, where C is the number of object categories and $p_v^{(c)}$ indicates the likelihood of class c appearing in view v. To obtain a scene-level prediction, we aggregate the results across all views by taking the category-wise maximum:

$$p = \max_{v=1,\dots,V} p_v, \tag{1}$$

where $p^{(c)}$ represents the final probability that class c exists in the scene. This design ensures that if an object appears in at least one view, it will also be considered at the scene level, allowing the model to handle occlusion and limited field-of-view.

However, in autonomous driving environments, the distribution of object classes is highly imbalanced. Frequent categories such as cars and pedestrians dominate the dataset, whereas rare yet safety-critical classes such as construction vehicles or motorcycles appear only sparsely. To address this issue, we adopt the Class-Balanced focal loss [15], which adjusts the contribution of each class according to its effective number of samples. The loss is defined as

$$\mathcal{L}_{\text{CBFL}}(p,y) = -\frac{1-\beta}{1-\beta^{n_y}} \cdot (1-p)^{\gamma} \cdot \log(p), \tag{2}$$

where p is the predicted probability for the ground-truth class y, and n_y is the number of training samples belonging to class y. The factor $\beta \in [0,1)$ adjusts the contribution of each class based on its effective sample count, giving relatively larger weights to rare classes. In addition, the parameter $\gamma > 0$ emphasizes hard or misclassified examples by assigning them larger loss values. Together, these terms reduce bias toward frequent classes and encourage the classifier to better learn rare categories, which in turn provides stronger class guidance for downstream detection.

C. Class-aware Query Integration

Scene-Level Class Guidance. Let $p \in [0,1]^C$ denote the aggregated scene-level class probability vector predicted by the classifier. We use p as a global signal to guide query initialization. To limit the impact of unreliable low-confidence predictions, we apply a small threshold τ and obtain a refined vector $\tilde{p}$:

$$\tilde{p}^{(c)} = \begin{cases} 0, & \text{if } p^{(c)} < \tau, \\ p^{(c)}, & \text{otherwise}, \end{cases} \tag{3}$$

This operation suppresses negligible probabilities while preserving relative confidence among plausible categories. In practice, τ serves as a simple confidence control mechanism

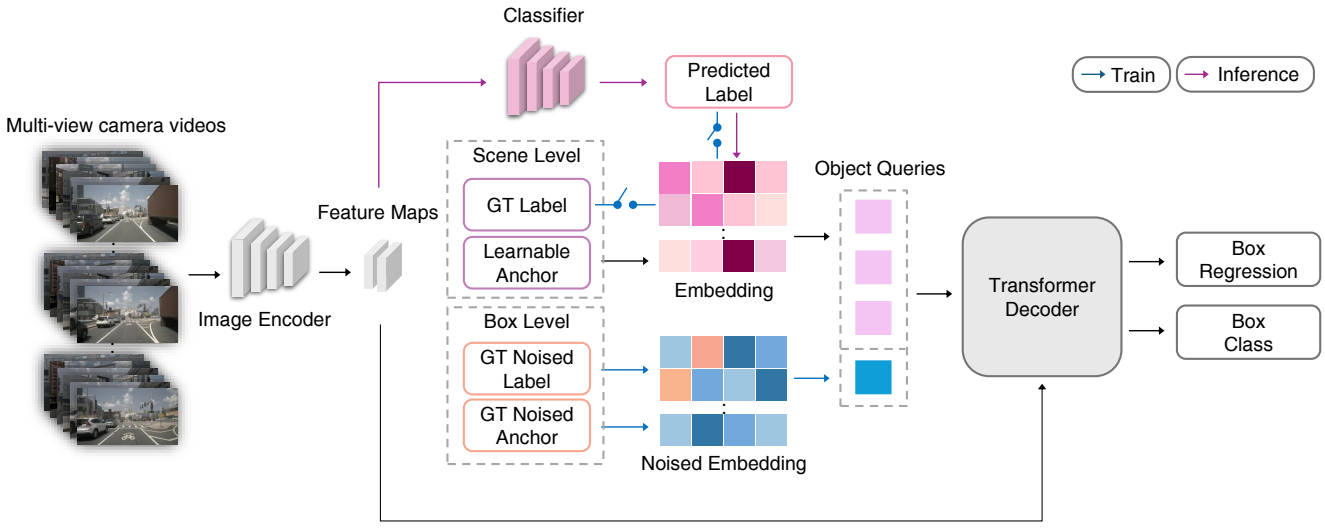

Fig. 3. Overview of the proposed detection framework. Multi-view images are first encoded by a shared image backbone (ResNet50) followed by a Feature Pyramid Network (FPN), producing multi-scale features for each camera. These features are then aggregated and fed into a transformer-based detection head, which predicts 3D bounding boxes and object classes. We incorporate a class-guided query mechanism by injecting label embeddings into the decoder queries based on the predicted scene-level class distribution.

that reduces the influence of noisy predictions without discarding soft class information, thereby preventing unlikely classes from affecting query initialization.

To represent the refined probability vector in the embedding space, we associate each class index with a learnable embedding vector stored in a table $\mathbf{E} \in \mathbb{R}^{C\times d}$. Using $\tilde{p}$ as weights over $\mathbf{E}$, we compute a scene-level semantic embedding:

$$e = \tilde{p}^{\top}\mathbf{E}, \quad \mathbf{E} \in \mathbb{R}^{C\times d}. \tag{4}$$

Here, d denotes the dimension of each class embedding. The resulting vector e provides a compact summary of the categories expected in the scene and forms the semantic basis for constructing class-aware decoder queries.

Class-aware Query Formation. To preserve both the semantic structure summarized in $e \in \mathbb{R}^d$ and the explicit confidence magnitudes in $\tilde{p} \in \mathbb{R}^C$, we concatenate the two signals and project them into the decoder feature space. Specifically, let $[e;\tilde{p}] \in \mathbb{R}^{d+C}$ denote their concatenation. The class-guided query embedding is defined as

$$q^{\text{label}} = f(W_l[e;\tilde{p}]), \quad W_l \in \mathbb{R}^{d_q\times(d+C)}, \tag{5}$$

where W_l is a learnable projection matrix and $f(\cdot)$ denotes a non-linear transformation applied to the concatenated vector in the decoder feature space. By jointly projecting the semantic embedding e and the confidence vector $\tilde{p}$, the model preserves both structured class information and probabilistic cues during query initialization, rather than collapsing them into a single pre-aggregated representation.

The resulting vector acts as a scene-conditioned semantic prior and is shared across all N_q decoder queries. As a consequence, each query begins with an expectation over plausible object categories before spatial reasoning takes place. In practice, each decoder query is represented as a pair consisting of a content feature and a reference box. The semantic prior initializes the query content, while geometric diversity is provided by per-query learnable anchors. Formally, the i-th query is given by

$$(q_i^{(0)}, b_i^{(0)}) = \left(q^{\text{label}}, b_i^{\text{anchor}}\right), \quad i = 1,\ldots,N_q, \tag{6}$$

where b_i^{anchor} denotes the learnable reference box associated with the i-th query. During decoding, the content features interact with image features through attention, while the reference boxes are iteratively refined via box regression.

D. Two-Stage Learning for Robust Class Guidance

To balance stable optimization with robustness to noisy semantic guidance, we adopt a two-stage training schedule. As defined in Eq. (5), the label-conditioned query feature q^{label} is obtained by projecting the concatenation of the semantic embedding e and the refined probability vector $\tilde{p}$. Directly conditioning queries on predicted class priors from the beginning may bias query initialization before reliable geometric representations are formed. During the first stage, $\tilde{p}$ is constructed from ground-truth scene labels. This provides reliable semantic supervision and stabilizes query initialization. After the detector has converged under this clean supervision, we proceed to the second stage, where $\tilde{p}$ is instead obtained from the predictions of the pretrained multi-view scene-level classifier. In this stage, the decoder queries are conditioned on imperfect yet realistic class priors, while the geometric anchors remain unchanged. This transition exposes the detector to classifier noise during training, allowing it to adapt to the semantic uncertainty that arises at inference time. At inference, the detector follows the second-stage setting.

E. Noised Box-Level Guidance

While our framework embeds scene-level class guidance and learnable anchors into the queries, the ultimate goal of

Method	Backbone	Image Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet [17]	ResNet50	256 × 704	0.298	0.379	0.725	0.279	0.589	0.860	0.245
BEVDet4D [18]	ResNet50	256 × 704	0.322	0.457	0.703	0.278	0.495	0.354	0.206
PETRv2 [19]	ResNet50	256 × 704	0.349	0.456	0.700	0.275	0.580	0.437	0.187
BEVDepth [20]	ResNet50	256 × 704	0.351	0.475	0.639	0.267	0.479	0.428	0.198
P2D [31]	ResNet50	256 × 704	0.372	0.500	0.598	0.270	0.438	0.367	0.190
BEVStereo [21]	ResNet50	256 × 704	0.374	0.486	0.631	0.272	0.508	0.384	0.212
BEVPoolv2 [22]	ResNet50	256 × 704	0.406	0.526	0.572	0.275	0.463	0.275	0.188
SOLOFusion [23]	ResNet50	256 × 704	0.427	0.534	0.567	0.274	0.511	0.252	0.181
StreamPETR [24]	ResNet50	256 × 704	0.432	0.540	0.581	0.272	0.413	0.295	0.195
SparseBEV [9]	ResNet50	256 × 704	0.432	0.545	0.606	0.274	0.387	0.251	0.186
Ours	ResNet50	256 × 704	**0.459**	**0.560**	0.574	0.267	0.419	0.248	0.189

TABLE I

PERFORMANCE COMPARISON ON THE NUSCENES VALIDATION SET. ALL METHODS USE THE RESNET50 BACKBONE WITH 256 × 704 RESOLUTION UNLESS NOTED OTHERWISE.

detection remains accurate box-level classification and localization. To achieve this, we introduce noised label guidance and noised anchor guidance at the box level. Both inject perturbations during training, but with different objectives: noised labels improve robustness to noisy class guidance, while noised anchors regularize the geometric initialization so that learnable anchors can converge more effectively. We implement this by randomly perturbing a subset of ground-truth queries in both label and bounding box dimensions. This simulates the classification ambiguities and localization noise that may arise due to imperfections in scene-level priors or sensor inputs.

Noised label guidance. For label perturbation, each ground-truth class label ℓ_i is randomly flipped with probability ρ to simulate classification noise:

$$\tilde{\ell}_i = \begin{cases} \ell_i, & \text{with probability } 1-\rho \\ \text{random class}, & \text{with probability } \rho \end{cases} \tag{7}$$

where $\tilde{\ell}_i$ is the perturbed label, ℓ_i is the original ground-truth label, and ρ controls the noise probability.

Noised Anchor Guidance. Similarly, for the perturbation of the bounding box, each ground-truth 3D box b_i represented as a 10-dimensional vector $b_i = (x, y, z, w, l, h, \sin\theta, \cos\theta, v_x, v_y)$ is perturbed with uniform component noise:

$$\tilde{b}_i = b_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{U}(-\delta \cdot \mathbf{s}_i, \delta \cdot \mathbf{s}_i), \tag{8}$$

where δ is a scalar noise scale and $\mathbf{s}_i$ is a component-wise scaling vector. Position and size dimensions are scaled relative to the box size, while angle and velocity dimensions use fixed scaling values.

In summary, noised labels perturb the class embeddings so that the decoder remains robust under noisy class guidance, whereas noised anchors perturb geometric supervision to act as an augmentation that sharpens localization learning. Together, these complementary perturbations regularize the model from both class guidance and geometric perspectives.

IV. EXPERIMENTS

A. Dataset and Metrics

We evaluate our method on the nuScenes dataset [25], a large-scale benchmark for 3D object detection in autonomous driving. The dataset comprises 1000 driving scenes, each approximately 20 seconds in duration, collected with a full 360-degree sensor suite including six cameras, one lidar, and five radars. Scenes are sampled at 2 Hz, resulting in roughly 1.4 million annotated 3D bounding boxes across ten object categories: car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle, bicycle, traffic cone, and barrier.

The dataset is split into 700 scenes for training, 150 for validation, where our experiments were conducted on the validation set. Following the official evaluation protocol, we report the mean Average Precision (mAP) and five class-wise True Positive (TP) metrics: average translation error (ATE), scale error (ASE), orientation error (AOE), velocity error (AVE), and attribute error (AAE). In addition, we report the nuScenes Detection Score (NDS), which aggregates these metrics into a single composite score for overall system performance.

B. Implementation Details

We adopt SparseBEV [9] as our baseline detector, using a ResNet50 backbone without CBGS (Class-Balanced Grouping and Sampling) [26] and an input resolution of 256×704. The model processes 8-frame multi-view sequences with 900 object queries, and query denoising is applied during training. Detection loss combines focal loss for classification and L1 loss for box regression.

For scene classification, we employ a ResNet50-based classifier that shares the image backbone with the detector. The classifier is first trained independently, pretrained on ImageNet and fine-tuned on the nuScenes dataset for 10-class classification over 9 epochs with single-frame inputs resized to 224×224 using class-balanced focal loss. During detection training, the classifier backbone is frozen, and scene-level

predictions are aggregated across views to guide class-aware query construction.

Following the two-stage learning strategy described in our method, the detector is first trained for 24 epochs using ground-truth scene labels to provide stable class-aware queries. In the second stage, we fine-tune the detector for an additional 3 epochs by replacing ground-truth labels with the classifier's own predictions. This gradual shift improves robustness to imperfect class guidance and better reflects inference conditions.

C. Main Results

We compare our method with recent camera-based 3D object detection models on the nuScenes dataset. As shown in Table I, all methods use a ResNet50 backbone and an input resolution of 256×704 for fair comparison. Our model achieves 45.9% mAP and 56.0% NDS, outperforming SparseBEV [9] on both metrics. Compared with StreamPETR [24] and SOLOFusion [23], our approach yields consistent gains in overall detection accuracy while preserving competitive localization quality. In terms of detailed detection metrics, Ours achieves the lowest mASE of 0.267 and the lowest mAVE of 0.248, reflecting more precise 3D box scale alignment and more reliable velocity estimation.

We further analyze the accuracy–efficiency trade-off. Figure 4 plots NDS against inference speed (FPS) for ResNet50-based camera-only detectors. All models are evaluated on a single NVIDIA TITAN RTX GPU with batch size 1 using full multi-view inputs, and FPS is averaged over repeated forward passes after warm-up to ensure stable measurement. Our method achieves the highest NDS while incurring only a modest reduction in inference speed compared to the baseline detector. Although an additional scene-level classifier is introduced, it shares backbone features with the detector and therefore adds negligible latency at inference time. These results indicate that the proposed class-aware query mechanism improves detection accuracy without sacrificing efficiency.

Figure 5 presents qualitative comparisons in challenging 3D scenes with occlusion and class ambiguity. The baseline often produces boxes that are geometrically plausible yet assigned incorrect class labels. In contrast, our model reduces such semantic inconsistencies while preserving localization precision. These examples illustrate how early class guidance stabilizes semantic predictions without disrupting geometric refinement.

D. Ablation Studies

We conduct controlled ablation experiments to isolate and evaluate the contributions of each component in our proposed framework. Specifically, we analyze the effects of class-aware query construction, two-stage training, and noisy supervision strategies.

Classification Bottleneck in Detection. To assess the impact of class guidance on both classification and localization, we explicitly disentangle the two aspects: classification accuracy (Cls. Acc.) and localization accuracy (Loc. Acc.). Cls. Acc. measures the proportion of correctly classified boxes among those that are already localized accurately, reflecting how well the model assigns semantic labels once geometric alignment is correct. Conversely, Loc. Acc. measures the proportion of accurately localized boxes regardless of their class labels, capturing box regression quality independently of classification. As shown in Table II, the baseline achieves 0.76 in Cls. Acc. and 0.75 in Loc. Acc., while our method improves them to 0.82 and 0.80, respectively. These gains translate into overall detection improvements, with mAP increasing from 0.432 to 0.459 and NDS from 0.545 to 0.560.

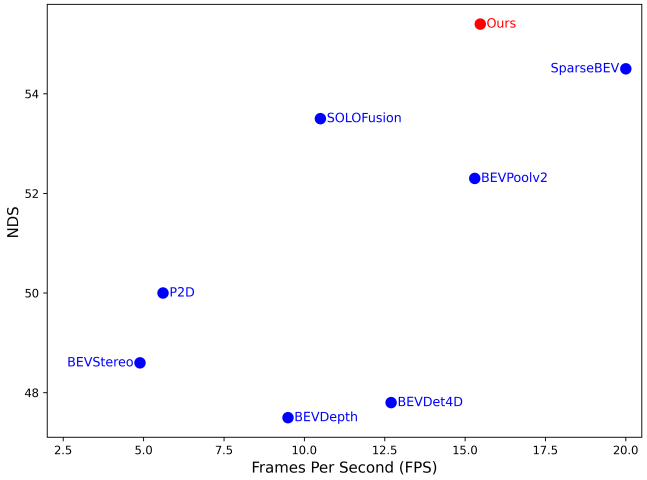

Fig. 4. Comparison of detection accuracy (NDS) and efficiency (FPS) on the nuScenes validation set.

Method	Cls. Acc. ↑	Loc. Acc. ↑	mAP ↑	NDS ↑
Baseline model	0.76	0.75	0.432	0.545
Our models	**0.82**	**0.80**	**0.459**	**0.560**

TABLE II
COMPARISON OF DETECTION-LEVEL CLASSIFICATION AND DETECTION PERFORMANCE ON NUSCENES.

Classifier	Scene-Level Classifier Acc. ↑	mAP ↑	NDS ↑
Swin-B	0.8319	0.457	0.557
ResNet50	0.8048	**0.459**	**0.560**

TABLE III
DETECTION PERFORMANCE WITH DIFFERENT SCENE-LEVEL CLASSIFIERS.

Effect of classification accuracy on detection. To examine how sensitive the detector is to the upstream classifier, we replace the scene-level classifier backbone while keeping the detection architecture fixed. Swin-B achieves higher standalone classification accuracy (0.8319) than ResNet50 (0.8048). However, this improvement in classification does not translate into a noticeable change in detection performance. The two configurations produce similar results, with mAP/NDS of 0.457/0.557 and 0.459/0.560, respectively. This observation suggests that detection performance is not determined solely by classifier capacity. The oracle study shows that accurate class information establishes a clear upper bound, yet simply increasing classification accuracy

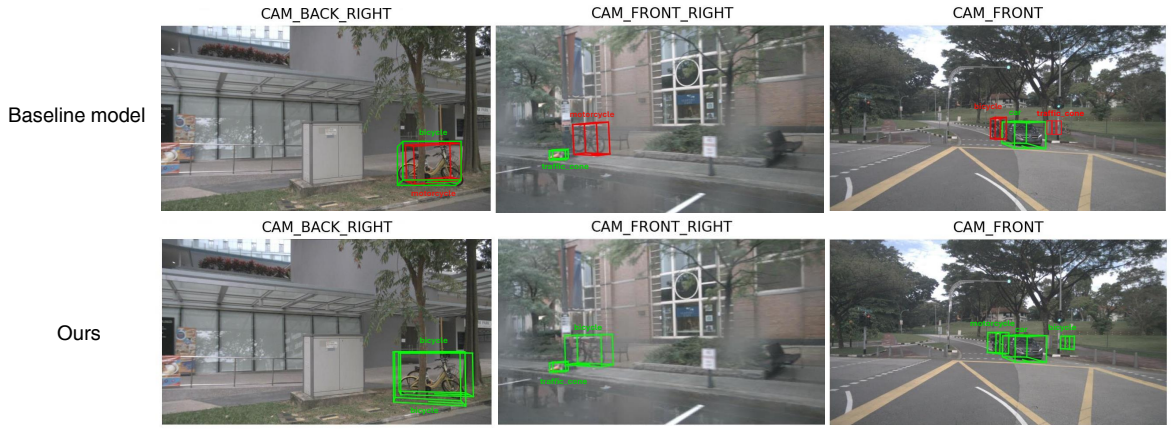

Fig. 5. Qualitative comparison of detection results on the same scene from the baseline model and our proposed method. Red boxes indicate predictions with correct localization but incorrect class labels, while green boxes denote fully correct predictions (both class and location). For improved interpretability, class names are shown instead of numerical class IDs.

Two-Stage Training	Noised Box Guidance	mAP↑	NDS↑
		0.440	0.548
✓		0.443	0.550
	✓	0.448	0.553
✓	✓	**0.459**	**0.560**

TABLE IV

ABLATION ON ROBUSTNESS STRATEGIES. EACH IMPROVES PERFORMANCE, AND THEIR COMBINATION GIVES THE BEST MAP.

Stage	Guidance Source	mAP↑	NDS↑
1-stage	GT Labels	0.468	0.561
	Predicted Labels	0.413	0.534
2-stage	GT Labels	0.450	0.545
	Predicted Labels	0.459	0.560

TABLE V

ROBUSTNESS WITH GT VS. CLASSIFIER-DERIVED LABEL EMBEDDINGS.

does not proportionally improve detection in practice. What appears to matter more is how the predicted class distribution is incorporated into query initialization. By converting scene-level predictions into structured priors rather than directly coupling class logits to box prediction, the proposed design reduces sensitivity to the specific classifier backbone. Overall, the detector exhibits stable behavior across different classifier backbones, indicating that the integration strategy, rather than backbone scale, plays the dominant role.

Robustness Strategies. Table IV reports the effect of two robustness strategies that complement our class-aware query design. Without either strategy, the detector achieves 44.0 mAP and 54.8 NDS. Applying *two-stage learning* improves performance to 44.3 mAP and 55.0 NDS, while applying *noised box-level guidance* further improves it to 44.8 mAP and 55.3 NDS. When both strategies are combined, the detector reaches 45.9 mAP and 56.0 NDS, achieving the best performance in terms of both mAP and NDS.

These improvements can be explained by how each strategy interacts with the proposed query design. Two-stage learning strengthens the *label* pathway by first pretraining with ground-truth scene labels to establish stable semantic priors and then fine-tuning with classifier predictions to adapt to noisy guidance at inference. In contrast, noised box-level guidance regularizes both the *label* and *anchor* pathways by perturbing ground-truth labels and bounding boxes, forcing the model to recover from classification errors and geometric noise. Together, the two strategies enhance the model's robustness and improve detection performance.

Stabilizing Effect of Two-Stage Learning. Table V reports the effect of our two-stage training scheme designed to enhance robustness against noisy class guidance. The key idea is to divide training into two sequential phases: the first stage uses ground-truth (GT) scene labels to establish stable supervision for class-aware queries, and the second stage fine-tunes the detector with predicted labels obtained from the multi-view classifier. This staged design exposes the model to realistic conditions where classifier errors may occur, preparing it to maintain reliable performance during inference when GT labels are unavailable. In the 1-stage setting, training directly with predicted labels yields 0.413 mAP and 0.534 NDS, substantially lower than the GT-labeled counterpart (0.468 mAP, 0.561 NDS). Incorporating predicted labels within the two-stage schedule recovers performance to 0.459 mAP and 0.560 NDS, confirming that GT pre-stabilization effectively reduces sensitivity to classifier errors. For the GT label setting, performance in the two-stage case (0.450 mAP, 0.545 NDS) is slightly lower than the 1-stage case (0.468 mAP, 0.561 NDS). This small drop is expected because the second stage no longer benefits from entirely clean labels. However, this trade-off is not problematic, as GT labels are never provided at inference. The central point is that the model trained with two-stage learning achieves stronger and more stable performance when guided by predicted labels, which reflects the realistic usage scenario.

V. CONCLUSION

We introduced a novel framework for transformer-based object detection that, for the first time, constructs class-

aware queries using scene-level category predictions. Unlike previous DETR-style models that use class-agnostic queries regardless of scene semantics, our approach infuses semantic priors into query generation, allowing the detection process to be guided from the very beginning. This semantic guidance not only improves classification accuracy but also enhances localization by enabling more targeted attention. Our experiments on nuScenes show that class-aware queries yield consistent improvements in both mAP and NDS. Furthermore, our oracle study reveals that classification remains a critical bottleneck in query-based detectors, supporting the need for more semantically informed query designs. We believe this work opens up a promising direction for advancing context-aware object detection in complex scenes.

a) Limitations and Future Work: While the class-aware query framework improves semantic-geometric disentanglement and detection robustness, it relies on accurate scene-level classification, which may be challenging in cluttered or novel environments, and assumes a fixed category set that limits generalization to unseen classes. Future work will explore open-vocabulary extensions and more robust uncertainty modeling.

ACKNOWLEDGMENT

This work was supported in part by the Autonomous Driving Center, Hyundai Motor Company and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. RS-2024-00340745).

REFERENCES

[1] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17830–17839, 2023.

[2] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 531–548, 2022.

[3] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 13619–13627, 2022.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 213–229, 2020.

[5] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[6] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[7] M. Zhang, G. Song, Y. Liu, and H. Li, "Decoupled DETR: Spatially disentangling localization and classification for improved end-to-end object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6601–6610, 2023.

[8] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[9] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "SparseBEV: High-performance sparse 3D object detection from multi-camera videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 18580–18590, 2023.

[10] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learning (CoRL)*, pp. 180–191, 2022.

[11] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[12] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, *et al.*, "RegionCLIP: Region-based language-image pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16793–16803, 2022.

[13] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary DETR with conditional matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 106–122, 2022.

[14] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "PromptDet: Towards open-vocabulary detection using uncurated images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 701–717, 2022.

[15] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9268–9277, 2019.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2980–2988, 2017.

[17] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird's-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[18] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[19] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "PETRv2: A unified framework for 3D perception from multi-camera images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3262–3272, 2023.

[20] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, pp. 1477–1485, 2023.

[21] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, pp. 1486–1494, 2023.

[22] J. Huang and G. Huang, "BEVPoolv2: A cutting-edge implementation of BEVDet toward deployment," *arXiv preprint arXiv:2211.17111*, 2022.

[23] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[24] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3621–3631, 2023.

[25] H. Caesar, V. Bankiti, A. 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11621–11631, 2020.

[26] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3D object detection," *arXiv preprint arXiv:1908.09492*, 2019.

[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11976–11986, 2022.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10012–10022, 2021.

[29] S. Moon, H. Park, J. Lee, and J. Kim, "Learning temporal cues by predicting objects move for multi-camera 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 6607–6613, 2024.

[30] Y. Zhu, L. Hui, H. Yang, J. Qian, J. Xie, and J. Yang, "Learning class prototypes for unified sparse-supervised 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9911–9920, 2025.

[31] S. Kim, Y. Kim, I.-J. Lee, and D. Kum, "Predict to detect: Prediction-guided 3D object detection using sequential images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 18057–18066, 2023.

[32] X. Liu, C. Zheng, M. Qian, N. Xue, C. Chen, Z. Zhang, C. Li, and T. Wu, "Multi-view attentive contextualization for multi-view 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16688–16698, 2024.