CLIP-BEVFormer: Enhancing Multi-View Image-Based BEV Detector with Ground Truth Flow

Chenbin Pan[1]* Burhaneddin Yaman[2] Senem Velipasalar[1] Liu Ren[2]
[1]Syracuse University
[2]Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
{cpan14, svelipas}@syr.edu, {burhaneddin.yaman,liu.ren}@us.bosch.com

Abstract

Autonomous driving stands as a pivotal domain in computer vision, shaping the future of transportation. Within this paradigm, the backbone of the system plays a crucial role in interpreting the complex environment. However, a notable challenge has been the loss of clear supervision when it comes to Bird's Eye View elements. To address this limitation, we introduce CLIP-BEVFormer, a novel approach that leverages the power of contrastive learning techniques to enhance the multi-view image-derived BEV backbones with ground truth information flow. We conduct extensive experiments on the challenging nuScenes dataset and showcase significant and consistent improvements over the SOTA. Specifically, CLIP-BEVFormer achieves an impressive 8.5% and 9.2% enhancement in terms of NDS and mAP, respectively, over the previous best BEV model on the 3D object detection task.

1. Introduction

Autonomous Driving Systems (ADS) have witnessed rapid advancements, revolutionizing the landscape of transportation [6, 9, 12, 13, 15, 17–19]. The current state-of-the-art vision based ADS approaches heavily rely on Bird's Eye View (BEV) representations extracted from multi-view images to perceive and understand the surrounding environment. The BEV detector, which acts as the backbone of ADS, transforms multi-view images into a top-down view representation known as the BEV feature map [11, 16, 20, 21, 32]. The effectiveness of the BEV detector significantly influences the success of perception tasks in autonomous driving, underscoring its pivotal role in enhancing the system's situational awareness [13, 15, 16].

However, in current BEV models such as BEVFormer [16], the BEV encoding process and its information embedding quality are ensured by decoding specific 3D object information through a transformer decoder [16]. This dependency on the decoder's capability lacks a compelling force to align the produced BEV with the ground truth BEV. Additionally, decoding is executed using a set of initialized queries [4, 16]. In particular, through cross-attention and self-attention mechanisms [25], the final informed queries align with ground-truth instances through a matching algorithm [4]. The decoding process resembles a black box and lacks interpretability, leaving uncertainty about which query corresponds to predicting which ground truth instance before decoding. Hence, there is a lack of a ground-truth perspective on the decoding process, such as how each ground truth instance interacts with others and the BEV feature map in the real scenarios.

In order to address the aforementioned limitations, we introduce CLIP-BEVFormer framework, which is composed of Ground Truth BEV (GT-BEV) module and Ground Truth Query Interaction (GT-QI) module.

In the absence of supervision on the BEV in previous models, the representation of object classes and boundaries is not guaranteed to align with the expressive nature of the ground truth BEV. This limitation imposes constraints on the perception capabilities of the model, leading to a reduced discriminative capacity. Hence, we introduce the GT-BEV module, a crucial component of our CLIP-BEVFormer framework, to enhance the quality of BEV generation. GT-BEV employs a ground-truth information flow (GT-flow) guidance during the BEV encoding phase. In particular, we implement a contrastive learning technique, as presented in CLIP [23], to align the produced BEV with the ground truth BEV, ensuring explicit arrangement of BEV elements based on their class label, location, and boundary. This explicit element arrangement, guided by ground truth information, serves to enhance the perceptual abilities of the model, allowing for improved detection and differentiation of various objects on the BEV map.

The intricate interactions among ground truth instances on the BEV map have remained unexplored by previous models. The recovery of ground truth information from

*Work done while interned at Bosch Research North America.

empty queries through self-attention and cross-attention mechanisms is limited in its interpretability, functioning as a black box exploration process and offering supervision only at the endpoint. To address this limitation, we introduce the Ground Truth Query Interaction (GT-QI) module, an integral component of our CLIP-BEVFormer framework. The GT-QI module injects GT-flow into the decoder during training, enriching the query pool and providing valuable learning insights into the decoding process. The incorporation of GT-flow into the decoding process facilitates interactions and communication among ground truth instances and between ground truth instances and the BEV map. By incorporating ground truth queries, the expanded query pool not only enhances the model's robustness but also augments its ability to utilize queries for detecting various objects within the source map.

Our innovative CLIP-BEVFormer, aiming to enhance the image-based BEV transformer with GT-flow guidance, improves both the BEV encoder process and the perception decoder process. It is a novel training framework that can be applied to any transformer and image based BEV detectors. Moreover, our method doesn't introduce any additional parameters and computations during inference stage, thus it maintains the efficiency of the original model.

Our main contributions can be summarized as follows:

- **CLIP-BEVFormer Framework:** We propose CLIP-BEVFormer, a pioneering training framework that enhances the BEV detector by integrating ground truth flow guidance into both BEV encoding and perception decoding processes.
- **Superior Performance:** Through extensive experiments on challenging nuScenes dataset [3], we show that CLIP-BEVFormer consistently outperforms counterpart BEV detector methods on various tasks.
- **Generalization and Robustness:** Our model exhibits superior generalization and robustness, particularly in long-tail cases and scenarios involving sensor failures, ensuring heightened safety in autonomous driving.
- **Flexibility and Efficiency:** CLIP-BEVFormer showcase the flexibility of not necessitating a language model for training, relying on a simple MLP layer for substantial performance gains. CLIP-BEVFormer does not incur any extra computation time during inference time as its novel components are only employed during training time.

2. Related Works

2.1. Bird's Eye View Feature Generation

BEV feature generation has recently gained a lot of interest for various downstream tasks as its holistic representation of the scene has been a success for various downstream tasks. Early works [8, 20, 21] leverages ConvNets and inverse perspective mapping (IPM) for mapping features from perspective view to BEV view. More recently, transformers based architectures have been extensively studied for BEV feature generation [5, 11, 16, 33, 38]. While some of these works focuses on only spatial feature transformation [5, 33], more recent works also incorporates temporal information for BEV generation. In particular, BEVFormer [16] proposes a spatiotemporal transformer which employs spatial cross-attention to aggregate spatial features from multi-view camera images and temporal self-attention to fuse history BEV features.

BEV training is conducted by attaching downstream task heads to the generated BEV representation. 3D perception tasks such as 3D object detection and segmentation are two of the main downstream tasks for applications of BEV representations. The evaluation performance on downstream tasks serves as an indicator for the quality of BEV formation technique.

2.2. Vision-Language Models

Vision-language models (VLMs) have shown great promise for learning good representations for variety of downstream tasks [1, 14, 23, 24, 35, 37]. The success of VLMs have been driven by training transformers on large scale image-text pairs data collected from web using contrastive learning [14, 23]. Among VLMs, CLIP [23] which have been trained on 400 million pairs of data have shown remarkable zero-shot generalization on various image recognition tasks. The contrastive learning mechanism which maps image and text pairs to a joint embedding space has been pivotal for CLIP success. While VLMs generally show good performance on downstream tasks without any fine-tuning, prompting and fine-tuning techniques have also been employed for adaptation of VLMs to a new downstream task [10, 41].

While VLMs have been extensively explored and used in various domains, its application to BEV generation process is yet to be explored. A recent work has explored using CLIP for BEV retrieval [2]. However, unlike our proposed approach, this work has no impact on BEV generation.

2.3. Contrastive Learning in Computer Vision

Contrastive learning has gained prominence in computer vision for self-supervised representation learning. Techniques like SimCLR [7] and CLIP [23] demonstrate the power of contrastive learning in capturing meaningful representations from diverse data modalities. These approaches have primarily been applied to image and text domains.

However, the application of contrastive learning in the perception domain, particularly in the context of Bird's Eye View (BEV) detection, has been limited. Previous models [26, 30, 36, 39, 40], including CLIP, often employ dual pathways for multi-modality incorporation and feature transfer learning. This dual-path approach necessitates the application of both pathways during inference, potentially

impacting efficiency. In contrast to previous methodologies, our approach harnesses contrastive learning as a guiding mechanism to facilitate model parameter learning.

3. Methodology

We present the technical details of our proposed CLIP-BEVFormer illustrated in Fig. 1. The core innovation of CLIP-BEVFormer lies in its provision of a ground truth perspective for both the Bird's Eye View (BEV) encoding and perception query decoding processes, achieved through the integration of the Ground Truth BEV (GT-BEV) module and the Ground Truth Query Interaction (GT-QI) module, respectively. We begin by examining the architecture of previous BEV detectors in Sec. 3.1. Subsequently, we delve into the specifics of the GT-BEV and GT-QI modules in Sec. 3.2 and Sec. 3.3, respectively. The training loss of our framework is outlined in Sec. 3.4.

3.1. Preliminary

In the realm of image-based Bird's Eye View (BEV) detectors, the processing pipeline traditionally involves the utilization of multi-view camera images $\mathcal{X}_{views}$. These images undergo initial processing through an image backbone, followed by a BEV encoder to amalgamate pertinent image features into a unified top-down view BEV feature map $z_{bev} \in \mathbb{R}^{H_b \times W_b \times C}$. Based on transformer architecture, a set of decoder queries q_{dec} is initialized to extract information from the BEV feature map [16]. These queries are responsible for decoding perceptual information specific to objects within the receptive view based on the acquired information. The informed decoder queries are then directed to a perception head $Head$, generating detection results through a matching algorithm $Match$ [4]. Each query serves as a prediction for a matched instance or an empty instance, determined through the matching algorithm [4]. Following the matching process, the final perception loss is computed by applying a perceptual loss function between the query predictions and the ground truth instances. More formally, the process can be expressed as Eq. (1):

$$\begin{aligned} z_{bev} &= BEVEnc(\mathcal{X}_{views}), \\ q^{'}_{dec} &= Dec(z_{bev}, q_{dec}), \\ q_{pred} &= Match(Head(q^{'}_{dec}), y), \\ \mathcal{L} &= \mathcal{L}_{perc}(q_{pred}, y), \end{aligned} \tag{1}$$

where $BEVEnc$ and Dec denote the BEV encoder and perception decoder respectively, y denotes the ground truth of the perception task, and q_{pred} denotes the query predictions after matching.

3.2. Ground Truth BEV

In the absence of supervision on the BEV in previous models, the accurate representation of object classes and boundaries is not guaranteed to align with the expressive nature of the ground truth BEV. This discrepancy imposes constraints on the perception capabilities of the model, leading to inherent limitations. To mitigate this shortfall, we introduce the Ground Truth BEV (GT-BEV) module. The core objective of the GT-BEV is to align the generated BEV representation with the GT-BEV, ensuring an explicit arrangement of BEV elements based on their class label, location, and boundary.

In GT-BEV, we first employ a ground truth encoder $GTEnc$ to represent the class label c^i and ground-truth bounding box p^i information of the i^{th} instance on the BEV map as expressed in Eq. (2):

$$\beta^i = GTEnc(c^i, p^i), \tag{2}$$

where β^i, which has the same feature dimension C as the BEV feature, denotes the encoded ground truth feature of the i^{th} instance on the BEV map. Notably, in our experiments (as detailed in Sec. 4), we demonstrate the effectiveness of applying either large language models (LLM) or a simple Multi-Layer Perceptrons (MLP) layer as $GTEnc$ in achieving comparable good results.

Subsequently, for each instance, to enhance the clarity of its boundaries on the BEV map, we crop the area within its ground truth bounding box from the BEV feature map. We apply a pooling operation to the cropped tensor, serving as the representation for the corresponding object α^i, as shown in Eq. (3):

$$\alpha^i = Pool(Crop(z_{bev}, p^i)). \tag{3}$$

Finally, to pull the BEV and GT-BEV embeddings closer, we employ the contrastive learning procedure [23] to optimize the element relationship and distances inside the BEV feature space, as formulated in Eq. (4).

$$\begin{aligned} \mathcal{M} &= \lambda \cdot \frac{\alpha}{\|\alpha\|_2} \otimes \frac{\beta}{\|\beta\|_2}, \\ \mathcal{L}_{GT-BEV} &= \frac{\mathcal{L}_{CE}(\mathcal{M}, \mathcal{I}) + \mathcal{L}_{CE}(\mathcal{I}, \mathcal{M})}{2}, \end{aligned} \tag{4}$$

where $\mathcal{M}$ and $\mathcal{I}$ denote the produced and target similarity matrices between object BEV feature and object ground truth feature, respectively, λ is the logit scale learned during contrastive learning, $\otimes$ represents the matrix multiplication, $\mathcal{L}_{CE}$ is the cross entropy loss employed for the optimization of the similarity matrix, and $\mathcal{L}_{GT-BEV}$ is the final loss produced from the GT-BEV module. The detailed principle and explanation of the contrastive learning process is explained in CLIP [23]. The BEV object relationship is guided by the ground truth similarity, with training loss optimized through the similarity matrix between BEV features and GT features. This involves averaging cross-entropy loss along both the BEV and GT axes, ensuring effective alignment.

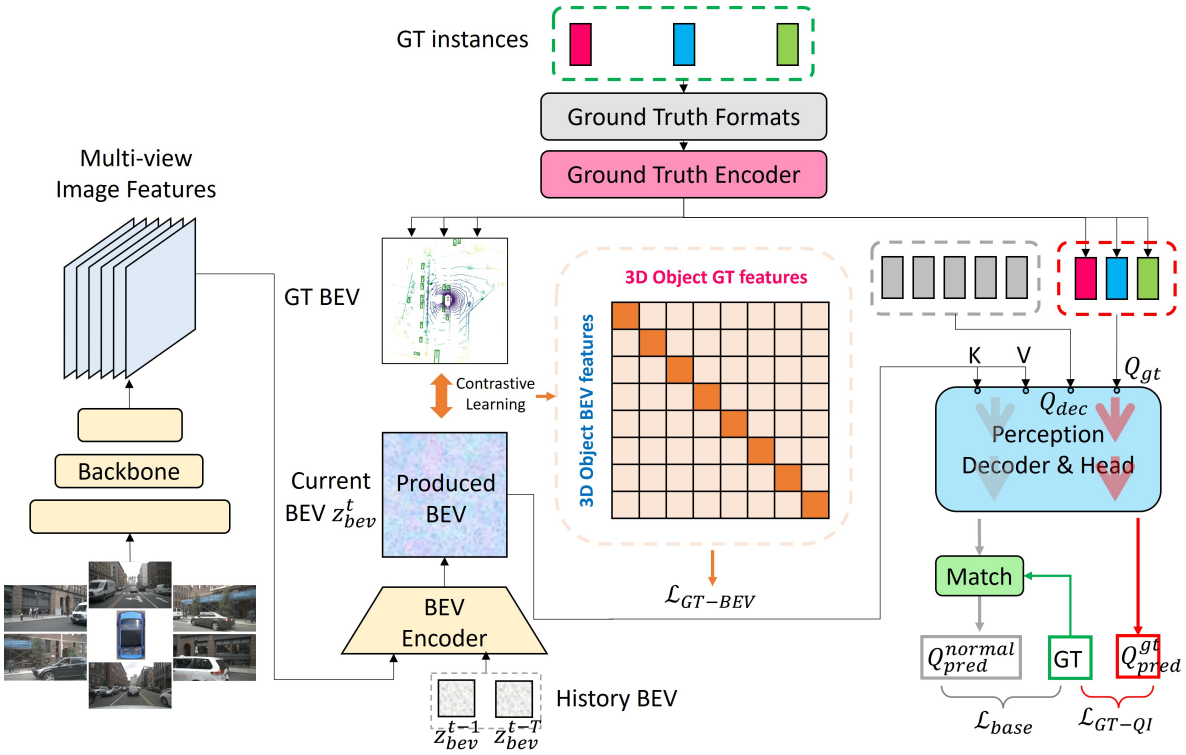

Figure 1. **The overview of CLIP-BEVFormer.** The architecture integrates two key modules: the Ground Truth BEV (GT-BEV) module employs a contrastive learning framework inspired by CLIP [23] to enrich the quality of BEV representations, while the Ground Truth Query Interaction (GT-QI) module introduces ground truth flow guidance into perception decoding processes. This integration leads to superior 3D object detection performance, as demonstrated in our extensive experiments on the challenging nuScenes dataset [3].

By orchestrating a meticulous arrangement of similarities in the feature space, our method provides explicit guidance for BEV generation. This explicit element arrangement, coupled with GT guidance (label, position, and clear boundary), serves to elevate the perceptual capabilities of the model, enabling improved detection and differentiation of various objects on the BEV map.

3.3. Ground Truth Query Interaction

A previously unexplored aspect in prior models lies in understanding how ground truth instances interact with each other on the BEV map. The recovery of ground truth information from empty queries through self-attention and cross-attention mechanisms, as employed in previous models, is inherently limited. This process operates as a black-box exploration, providing supervision solely at the endpoint, thus lacking a comprehensive understanding of ground truth decoding. To address this limitation, we introduce the Ground Truth Query Interaction (GT-QI) module, a novel addition to our CLIP-BEVFormer framework. The primary objective of the GT-QI module is to inspire the decoder parameter learning process by performing interactions among ground truth instances.

In the GT-QI module, the encoded ground truth features β from the Ground Truth Encoder (GTEnc) are introduced into the query pool of the decoder Dec, undergoing the same processes and modules as normal queries. This mimics real information communication in actual scenarios through self-attention (SA) mechanisms within the ground truth queries and global environment communication through cross-attention (CA) with the BEV map. Further processing of the ground truth query information is performed using a Feedforward Neural Network (FFN). It is worth noting that although the procedure and module weights are shared, the attention during decoding is executed in a parallel manner, preventing information leakage. The ground truth flow only influences and inspires the learning of decoder parameters. Since the ground truth query is initially targeted at a specific instance, no matching procedure is required for the ground truth query flow. After decoding, the same head $Head$ and perception loss $\mathcal{L}_{perc}$ are applied to the processed ground truth queries. The above process can be formulated as Eq. (5):

$$\begin{aligned} q_{gt} &= \beta, \quad q_{gt}^{'} = Dec(q_{gt}), \\ q_{gt}^{pred} &= Head(q_{gt}^{'}), \\ \mathcal{L}_{GT-QI} &= \mathcal{L}_{perc}(q_{gt}^{pred}, y), \end{aligned} \tag{5}$$

where q_{gt} denotes the initial GT queries, $q_{gt}^{'}$ and q_{gt}^{pred} indicate the GT queries after processing of perception decoder and perception head respectively, and $\mathcal{L}_{GT-QI}$ is the final loss generated by GT-QI module.

With ground truth flow injected during the decoding phase of training, our GT-QI module enables the modules to gain insights from both ground truth inter-instance interaction and ground truth instance-BEV communication. The enlarged query pool, injected with GT queries, not only enhances the robustness of the model but also augments its ability for detecting various objects on the source map.

3.4. Loss

The training loss formulation for our CLIP-BEVFormer encompasses three key components, each designed to optimize specific aspects of the model's performance: baseline BEV detector loss $\mathcal{L}_{base}$, Ground Truth BEV supervision loss $\mathcal{L}_{GT-BEV}$, perception loss of Ground Truth Query Interaction $\mathcal{L}_{GT-QI}$, as expressed in Eq. (6):

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{GT-BEV} + \mathcal{L}_{GT-QI}. \tag{6}$$

The ground truth flow, applied exclusively during the training phase via GT-BEV and GT-QI, introduces no additional parameters or computations during inference. This ensures that the efficiency of the original model is maintained at the inference stage.

4. Experiments

4.1. Implementation Details

Dataset. Our experiments are conducted on public nuScenes dataset[3], which is a commonly used dataset in autonomous driving area. It contains 1000 scenes, each with a ∼20s duration, and the keyframes are annotated at 2Hz. Each scene is captured with 6 cameras covering the entire 360 ° field-of-view. Overall, the nuScenes dataset contains 1.4M 3D annotated bounding boxes of 10 object categories.

Metrics. We focus our experiments on the 3D object detection task within the nuScenes dataset. This task involves placing a 3D bounding box around objects belonging to 10 distinct categories while estimating various attributes and the current velocity vector. For the evaluation of performance, we employ a set of metrics, including mean Average Precision (mAP), Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), Average Attribute Error (AAE), and nuScenes detection score (NDS), as defined in [3].

Experiment Set-up. We apply the typical BEV detector, BEVFormer[16], as our baseline model. To maintain consistency, we adopt the same training hyperparameters as outlined in [16]. This choice allows for a fair comparison and highlights the improvements brought by our proposed CLIP-BEVFormer framework. For training with language model as GT encoder, we freeze the whole model except the last projection layer, while for training with MLP as GT encoder, we train the whole layer. More details regarding the experiment set-up can be found in supplementary materials.

4.2. 3D Detection Results

We compare our proposed CLIP-BEVFormer against the baseline BEVFormer and other state-of-the-art BEV detectors listed in Tab. 1. We assess the performance across different model configurations, specifically applying CLIP-BEVFormer to both tiny and base variants of BEVFormer. Additionally, we explore the impact of employing either a language model (LM) in the pretrained CLIP [23] or a simple MLP layer as the ground truth (GT) encoder.

As illustrated in the Tab. 1, all four variants of CLIP-BEVFormer consistently outperform the baseline models accross all metrics except a slight degradation for tiny configuration in mAAE metric. The improvements are more evident in both nuScenes detection scores (NDS) and mean Average Precision (mAP). Specificially for tiny configuration, in comparison with baseline BEVFormer, CLIP-BEVFormer with MLP achieves 8.5% and 9.2% improvement in terms of NDS and mAP, respectively. Similarly, CLIP-BEVFormer with LM demonstrates 9.3% and 8.8% improvement in NDS and mAP, respectively, compared to counterpart BEVFormer. For base configuration, CLIP-BEVformer shows similar consistent improvement over the baseline, with our MLP variant showing best results.

The observed enhancements across diverse model configurations emphasize the effectiveness of injecting GT-flow for inspiring both BEV encoding and perception decoding processes in BEV detectors. The comparable performances achieved by both MLP and LM variants indicate that our framework is robust and less sensitive to the choice of the ground truth encoder. This flexibility makes CLIP-BEVFormer adaptable and easily deployable for integration with various detectors. The consistent improvements across all variants signify the robustness of CLIP-BEVFormer, demonstrating its capability to enhance 3D object detection tasks under different model complexities and encoder choices.

4.3. Long-tail Detection Results

To assess the effectiveness and generalization ability of our proposed method in handling long-tail cases, we present per-class detection results in Tab. 2. The nuScenes dataset exhibits a considerable class imbalance, where some classes represent a small portion (1%) and others a large portion (43%) of the dataset, as detailed in the Tab. 2. We evaluate the performance of CLIP-BEVFormer across specific object classes to highlight its ability to address challenges

Method	GT Enc	Modality	Backbone	NDS ↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
SSN [42]	-	Lidar	PointPollar	49.8	36.6	-	-	-	-	-
CenterPoint-Voxel [34]	-	Lidar	VoxelNet	64.8	56.4	-	-	-	-	-
FCOS3D [27]	-	Cam	R101	41.5	34.3	0.725	0.263	0.422	1.292	0.153
PGD [28]	-	Cam	R101	42.8	36.9	0.683	0.260	0.439	1.268	0.185
DETR3D [29]	-	Cam	R101	42.5	34.6	0.773	0.268	0.383	0.842	0.216
BEVerse [38]	-	Cam	Swin	46.6	32.1	0.681	0.278	0.466	0.328	0.190
+Ours	LM	Cam	Swin	**48.3**	**34.2**	**0.665**	**0.270**	**0.456**	**0.318**	**0.170**
BEVFormer-tiny[16]	-	Cam	R50	35.5	25.1	0.898	0.293	0.651	0.657	0.216
+Ours	MLP	Cam	R50	**38.5**	**27.4**	**0.869**	**0.283**	**0.607**	**0.542**	0.220
+Ours	LM	Cam	R50	**38.8**	**27.3**	**0.856**	**0.282**	**0.583**	**0.538**	0.228
BEVFormer-base[16]	-	Cam	R101	51.7	41.6	0.673	0.274	0.372	0.394	0.198
+Ours	MLP	Cam	R101	**56.2**	**46.7**	**0.605**	**0.253**	**0.331**	**0.336**	**0.187**
+Ours	LM	Cam	R101	**55.1**	**44.1**	**0.641**	**0.253**	**0.319**	**0.307**	**0.172**
BEVformerV2 [31]	-	Cam	R50	42.6	35.1	0.753	0.286	0.466	0.807	**0.186**
+Ours	LM	Cam	R50	**44.1**	**37.0**	**0.729**	**0.281**	**0.438**	**0.791**	0.204

Table 1. **3D Detection results on nuScenes validation set.** Comparison of 3D object detection performance across various state-of-the-art BEV detectors. CLIP-BEVFormer, applied to both tiny and base variants of BEVFormer, demonstrates consistent improvements over baseline models. Results are reported for different ground truth encoder choices, including Language Model (LM) and Multi-Layer Perceptron (MLP). Similar improvements achieved through LM and MLP encoders underscores the generalizability of proposed framework.

associated with less common occurrences.

As shown in the Tab. 2, CLIP-BEVFormer demonstrates an overall improvement in 3D detection performance across all classes, which indicates that our method positively contributes to the perception ability of the model, showcasing its robustness and adaptability, irrespective of the class distribution. Notably, for long-tail classes such as construction vehicle, bus, motorcycle, bicycle, and trailer, in which each of these categories account for approximately 1% of the dataset, CLIP-BEVFormer demonstrates substantial improvements. In particular, our CLIP-BEVFormer-base with simple MLP as ground truth encoder improves detection performance of construction vehicle, bus, motorcycle, bicycle, and trailer classes by 46.5%, 14.4%, 15.6%, 10.5%, and 26.7%, respectively, in comparison with the baseline BEVformer.

The considerable enhancements observed in long-tail classes underscore the enhanced learning ability and reduced sensitivity to data imbalance of CLIP-BEVFormer. By addressing challenges associated with less common object occurrences, our method showcases its potential for real-world deployment, where imbalanced class distributions are common. The ability to improve detection accuracy in long-tail scenarios further establishes CLIP-BEVFormer as a robust and reliable solution for 3D object detection tasks in autonomous driving systems.

4.4. Robustness Results

In real-world deployment scenarios, autonomous driving systems must contend with potential sensor failures arising from hardware malfunctions, adverse weather conditions, or physical obstructions. Evaluating the detector's performance under such conditions is crucial for assessing its robustness. Thus, we design and conduct first sensor failure robustness study. To simulate sensor failures, we conducted experiments involving the random masking of one camera view during inference, mimicking scenarios where a camera might malfunction.

The robustness evaluation metrics include NDS, mAP, mATE, mASE, mAOE, mAVE, mAAE, and the mAP on the long-tail classes. As depicted in Tab. 3, CLIP-BEVFormer consistently outperforms the baseline models in various configurations. Specifically, for tiny configuration, CLIP-BEVFormer with the language model (LM) as the ground truth encoder achieves 12.3% NDS, 15.7% mAP, and 27.5% long-tail mAP improvements over the baseline BEVFormer, demonstrating its superior performance and enhanced robustness under the simulated sensor failure scenario.

4.5. Ablation Study

GT encoder and GT input format. We investigate the ground truth input format along with the ground truth encoder. Three distinct input formats, namely digit, semantic, and scene, are explored. For the digit format, we directly utilize ground truth digit numbers, encompassing the one-hot format of the class label and the normalized 3D bounding box for each instance. In the semantic format, a semantic sentence template, such as "This is a {class-label}. Its 3D bounding box is {3D-bbox location}," is employed as the GT input. For the scene format, we augment the seman-

	GT Enc	CV	BUS	MOT	BIC	TR	TUK	CONE	BAR	PED	CAR
Total Num		650	657	748	857	1114	4215	6591	10263	11564	27727
Percentage		1.0%	1.0%	1.2%	1.3%	1.7%	6.5%	10.2%	15.9%	18.0%	43.1%
BEVFormer-tiny [16]	-	5.8	23.3	21.4	20.3	6.6	19.2	38.4	37.9	33.2	45.7
CLIP-BEVFormer-tiny	MLP	**7.1**	**28.0**	**26.1**	**21.6**	**8.1**	**20.9**	**41.1**	**40.0**	**33.9**	**46.8**
CLIP-BEVFormer-tiny	LM	**6.0**	**28.3**	**25.2**	**22.2**	**8.7**	**22.2**	**40.5**	**39.3**	**34.3**	**46.6**
BEVFormer-base [16]	-	12.9	44.4	42.9	39.8	17.2	37.0	58.4	52.5	49.4	61.8
CLIP-BEVFormer-base	MLP	**18.9**	**50.8**	**49.6**	**44.0**	**21.8**	**40.9**	**63.1**	**55.7**	**55.2**	**66.6**
CLIP-BEVFormer-base	LM	**14.0**	**46.9**	**46.6**	**41.1**	**19.6**	**37.9**	**62.6**	**56.4**	**52.1**	**64.6**

Table 2. **Per-class 3D detection results on NuScenes validation set.** Evaluation of per-class 3D object detection results, showcasing the performance of CLIP-BEVFormer and baseline BEVFormer on long-tail cases. The table provides insights into the distribution, total numbers, and percentages of instances across specific object classes. CLIP-BEVFormer exhibits overall enhancements, particularly more pronounced improvement in classes with lower(∼1-2%) occurrence frequencies, highlighting its efficacy in addressing imbalanced class distributions. The CV, BUS, MOT, BIC, TR, TUK, CONE, BAR, PED, and CAR denote the construction vehicle, bus, motorcycle, bicycle, trailer, truck, traffic cone, barrier, pedestrian, and car, respectively.

Figure 2. **Visualization results on nuScenes validation set.** We demonstrate qualitative detection performance on both camera and BEV images. As can be seen in BEV images, CLIP-BEVFormer demonstrates improved alignment with ground truth detections.

tic input with a scene description provided in the dataset, i.e. "This is a {class-label}. Its 3D bounding box is {3D-bbox location}. Its scene description is {scene}.". For the ground truth encoder, we investigate different configurations, including the language model in pretrained CLIP [23], GPT-2 [22], and a simple MLP layer.

As depicted in Tab. 4, all configurations outperform the baselines, demonstrating that GT flow provides crucial guidance to the detector. Notably, the results reveal that this guidance is less sensitive to the GT encoder and the GT input format. Regardless of the specific configuration, the performance remains consistently improved across both tiny and base variants, emphasizing the robust and flexible nature of our proposed CLIP-BEVFormer.

Effectiveness of each GT guidance. To assess the impact of each Ground Truth (GT) guidance component in CLIP-BEVFormer, we conduct ablation studies on both the tiny and base variants, as presented in Tab. 5. The GT encoder

Method	GT Enc	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	Long-Tail↑
BEVFormer-tiny [16]	-	33.3	21.7	0.908	0.295	0.656	0.681	0.218	13.1
CLIP-BEVFormer-tiny	MLP	**36.6**	**24.3**	**0.881**	**0.284**	**0.617**	**0.550**	0.225	**16.2**
CLIP-BEVFormer-tiny	LM	**37.4**	**25.1**	**0.873**	**0.283**	**0.589**	**0.540**	0.230	**16.7**
BEVFormer-base [16]	-	49.6	38.0	0.686	0.275	0.379	0.395	0.201	28.3
CLIP-BEVFormer-base	MLP	**53.4**	**42.1**	**0.612**	**0.273**	**0.348**	**0.341**	**0.192**	**31.7**
CLIP-BEVFormer-base	LM	**52.0**	**40.2**	**0.663**	**0.263**	**0.355**	**0.341**	**0.183**	**30.5**

Table 3. **Robustness evaluation results.** The robustness study experiments is conducted by random masking of one camera view during inference to simulate practical deployment scenarios with potential sensor failures. Proposed CLIP-BEVFormer shows improved robustness performance over the BEVFormer baseline for both tiny and base configurations.

GT Enc	GT Format	Tiny		Base	
		NDS↑	mAP↑	NDS↑	mAP↑
-	-	35.5	25.1	51.7	41.6
MLP	digit	**38.5**	**27.4**	**56.2**	**46.7**
CLIP-LM	digit	**38.4**	**27.6**	**55.9**	**45.3**
CLIP-LM	semantic	**38.8**	**27.3**	**55.1**	**44.1**
GPT2	semantic	**37.1**	**26.3**	**54.8**	**45.1**
CLIP-LM	scene	**38.2**	**27.7**	**55.3**	**45.0**

Table 4. **Ablation study on ground truth encoder and ground truth input format.** CLIP-BEVFormer consistently shows improvement over the baseline across variety of ground truth encoders and input formats.

GT Enc	GT Guidance	Tiny		Base	
		NDS↑	mAP↑	NDS↑	mAP↑
-	-	35.5	25.1	51.7	41.6
MLP	BEV	**37.3**	**26.9**	**55.2**	**43.2**
LM	BEV	**37.6**	**26.9**	**54.3**	**42.8**
MLP	BEV & Dec	**38.5**	**27.4**	**56.2**	**46.7**
LM	BEV & Dec	**38.8**	**27.3**	**55.1**	**44.1**

Table 5. **Ablation study for effectiveness of GT-BEV and GT-QI.** Results show the importance of both ground truth guidance components in achieving improved detection accuracy.

configurations include a Multilayer Perceptron (MLP) and a Language Model (LM). The ablation study explores the effectiveness of GT guidance solely for BEV encoding (BEV) and both BEV encoding and perception decoding (BEV & Dec). Results indicate that incorporating only GT-BEV guidance yields notable improvements over the baseline, enhancing NDS by 5.9% and 5.0%, and mAP by 7.2% and 2.9%, with LM as GT encoder, for both tiny and base models, respectively. Further enhancement is observed when integrating both GT guidance components, resulting in superior performance. This consistency across various configurations emphasizes the robust and effective nature of our proposed GT guidance in enhancing both BEV encoding and perception decoding processes.

4.6. Qualitative Results

The visual comparative results presented in Fig. 2 illustrates the enhanced performance of CLIP-BEVFormer in generating better Bird's Eye View (BEV) representations compared to the baseline BEVFormer. The visualized detection results shows that the output of CLIP-BEVFormer aligns more closely with the Ground Truth BEV, indicating an improvement in the precision of BEV generation. The qualitative results affirm that our method effectively enhances both the quality of BEV representation and the accuracy of 3D object detection.

5. Conclusion and Future Work

Conclusion: In this study, we introduce CLIP-BEVFormer, a pioneering framework aimed at advancing multi-view image-based Bird's Eye View (BEV) detectors. Comprising the GT-BEV module and GT-QI module, CLIP-BEVFormer leverages ground truth flow guidance in both the BEV encoding and perception decoding processes. The GT-BEV module orchestrates explicit arrangements for BEV elements, driven by class labels, locations, and boundaries, effectively elevating the BEV to approximate ground truth structures. Simultaneously, the GT-QI module enriches the decoder query pool and inspires the perception learning process. Our extensive experiments on the challenging nuScenes dataset demonstrate the consistent superiority of CLIP-BEVFormer over various tasks.

Limitations and Future Work: The current study establishes a strong foundation for enhancing BEV detectors through CLIP-BEVFormer. In the future work, we will investigate its application to different sensor modalities such as LiDAR. Additionally, the framework will be generalized to encompass a broader spectrum of autonomous driving tasks, including object tracking, scene segmentation, and motion prediction. Continuous refinement and adaptation will be pursued to uphold the robustness and versatility of CLIP-BEVFormer in real-world scenarios.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Anonymous. BEV-CLIP: Multi-modal BEV retrieval methodology for complex scene in autonomous driving. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. under review. 2

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 4, 5

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 3

[5] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*, pages 550–567. Springer, 2022. 2

[6] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[8] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 2

[9] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[11] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15273–15282, 2021. 1, 2

[12] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 1

[13] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[15] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023. 1

[16] 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 *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 3, 5, 6, 7, 8

[17] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 1

[18] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023.

[19] Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*, 2023. 1

[20] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 1, 2

[21] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 1, 2

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 7

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 7

[24] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[26] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[27] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 6

[28] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 6

[29] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 6

[30] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 2

[31] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 6

[32] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 1

[33] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15536–15545, 2021. 2

[34] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 6

[35] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[36] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2

[37] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 2

[38] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2, 6

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[42] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 581–597. Springer, 2020. 6