This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Q-TempFusion: Quantization-Aware Temporal Multi-Sensor Fusion on Bird's-Eye View Representation

Pinrui Yu[1*], Zhenglun Kong[1*], Pu Zhao[1], Peiyan Dong[1], Hao Tang[2], Fei Sun[3], Xue Lin[1], Yanzhi Wang[1]
[1]Northeastern University, [2]CVL, ETH Zurich, [3]Meta Inc.
{yu.pin, kong.zhe, zhao.pu, dong.pe, xue.lin, yanz.wang}@northeastern.edu
hao.tang@vision.ee.ethz.ch, fsun@feisun.org

Abstract

Recent advancements in bird's-eye view (BEV) perception models have highlighted the superior performance of LiDAR-camera fusion systems over single-modality approaches, garnering considerable interest in the field. Despite the progress, the integration of temporal information, a technique that has considerably benefitted camera-only BEV models, remains underexplored for LiDAR-camera fusion. This paper presents Q-TempFusion, a novel approach for temporal multi-sensor fusion designed to enhance the BEV model's inference speed while keeping high predictive performance compared with the current state-of-the-art. Moreover, we are the first to make the multi-modality BEV model profiling on hardware devices. To address the challenges of substantial memory demands and non-trivial latency that hinder deployment in on-vehicle systems, particularly when temporal dynamics are incorporated into complex multi-sensor models, we introduce an activation-aware quantization framework to generate the fully 8-bit quantized Q-TempFusion model based on the profiling result, which can be directly deployed to target devices with negligible detection performance degradation. Our experiments show that our Q-TempFusion (8-bit) achieves 70.3% mAP and 72.7% NDS with 3×∼18× FPS improvement over leading multi-modality baselines and the Q-TempFusion (32-bit) achieves 72.1% mAP and 74.8% NDS, comparable to SOTA multi-modality approaches. The results suggest that Q-TempFusion is a promising step toward real-time multi-sensor BEV applications, setting a new benchmark for efficient and reliable perception.

1. Introduction

Three-dimensional (3D) perception derived from a bird's-eye view (BEV) is pivotal for a plethora of applications, ranging from autonomous vehicles to robotic navigation systems. Enhancing the accuracy of such perception systems requires the integration of multiple sensors, each contributing a unique modality. Cameras provide two-dimensional imagery replete with rich semantic details, whereas LiDAR sensors furnish three-dimensional point clouds with precise spatial measurements. Consequently, a myriad of methodologies have been developed to synergistically process these heterogeneous sensor signals.

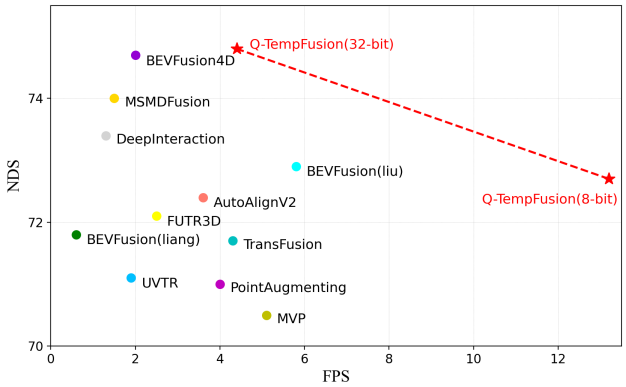


Figure 1. Comparison with existing methods. Our method achieves a better accuracy-speed trade-off on nuScenes test set than other BEV 3D detection methods. Latency results are obtained with NVIDIA V100.

Contemporary research in this domain can be broadly categorized into three distinct streams: camera-only, LiDAR-only, and camera-LiDAR fusion-based approaches. Camera-only methods [17, 35, 40, 69] are constrained by the absence of depth information in 2D images, and LiDAR-only approaches lack the semantic details inherent in visual data. In contrast, camera-LiDAR fusion techniques typically outperform their unimodal counterparts. They achieve superior detection performance by harmonizing features from both modalities with techniques such as camera-to-LiDAR projection [22, 53] to map visual features onto LiDAR points, LiDAR point augmentation [54, 55, 64] to enrich LiDAR point clouds with image-derived features, and

*Equal contributions

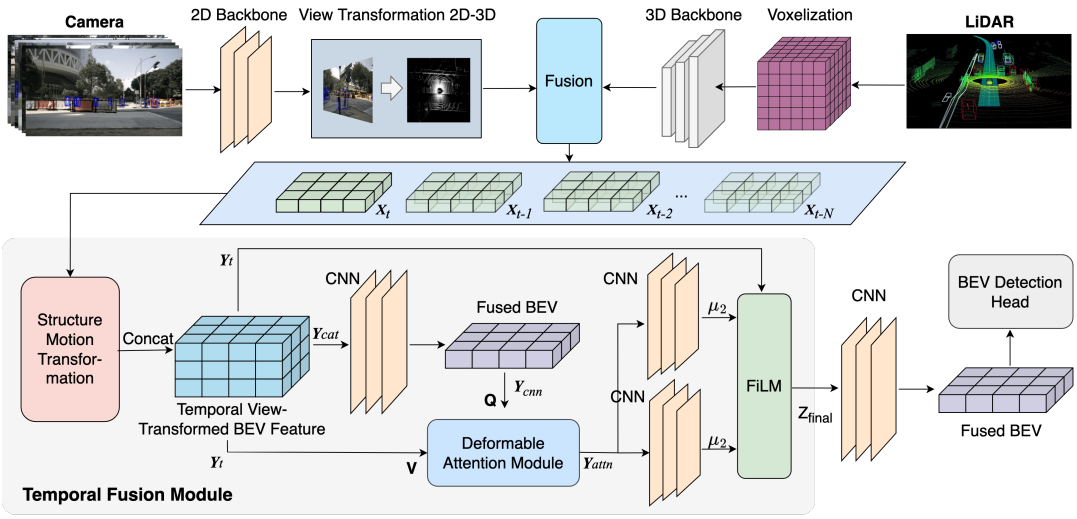


Figure 2. The architecture of the proposed Q-TempFusion framework with the pre-fusion mode. The features extracted from the camera and Lidar branches are fused and transformed into a unified BEV feature. The *Temporal Fusion Module* (TFM) aggregates the current and past BEV representation. The unified representation is then passed to a sparse convolution-based BEV decoder.

BEV space fusion [36, 43] to execute sensor fusion within a unified BEV spatial domain. These strategies underscore the multifaceted advantages of multimodal feature fusion.

Additionally, incorporating temporal information from consecutive frames into the BEV framework has recently gained interest. However, most BEV temporal integration research focuses on camera-only methods, with limited exploration of spatiotemporal dynamics in camera-LiDAR fusion techniques due to the complexity and computational intensity of these models. [67] first studied spatiotemporal integration in multimodal models using a vision transformer (ViT) framework, but its global self-attention mechanism significantly increased computational load. To mitigate this, [5] introduced the Lidar-Guided View Transformer with deformable attention for efficient spatiotemporal fusion. Despite advancements in reducing inference latency, real-time inference for streaming applications remains challenging.

There are many factors that limit the improvement of the inference speed for 3D detectors. (i) Despite the superior performance of 3D detectors employing ViT-based backbones [42] over CNN-based backbones [30, 57], a significant drawback of transformers is their high computational and memory complexity [26, 27, 31]. This complexity limits their scalability in various scenarios, particularly during deployments on resource-constrained edge devices, where local computing capabilities are limited. Moreover, to ensure safety for users of 3D detector applications, it often requires that the 3D detectors can achieve fast inference with available detection results in real-time. (ii) There is a notable absence of latency profiling studies for multi-modality BEV models in the existing literature, resulting in a lack of detailed information and insights that could be instrumental for targeted efficiency enhancements at the operator level. (iii) Model quantization is a promising approach to address the above-mentioned issues. Researchers [10, 23, 28, 44, 47, 56, 68] have developed quantization techniques to reduce both the computation and communication requirements of transformers. However, most of these techniques are based on training-free Post-Training Quantization (PTQ), which was originally proposed for natural language processing (NLP) oriented Transformers [3, 16, 51, 58, 63, 66] and may lead to accuracy drops when directly applied to ViTs. On the other hand, ViT can adapt Quantization-Aware Training (QAT) to minimize the quantization error incurred in PTQ. (iv) PTQ is not stable for classical transformer architecture in the training stage of multi-modality BEV models [34, 67]. Thus, it is essential yet challenging, to achieve inference efficiency on resource-limited edge devices for 3D detection.

To deliver substantial latency reduction and attain leading performance in 3D object detection tasks, we propose Q-TempFusion to achieve multi-sensor fusion on the BEV representation with temporal information and quantization (see Figure 1). Our framework makes use of temporal features from both cameras and LiDAR to achieve high detection accuracy with a temporal fusion module (TFM). We depart from the classical transformer framework presented in [67], opting instead for an architecture that is more amenable to QAT to facilitate the spatiotemporal fusion process. Furthermore, this paper introduces a latency profiling of multi-modality BEV models on hardware devices, aiming to shed light on potential avenues for model quantization. Leveraging the insights gleaned from profiling and the observed data distribution in ViTs, we employ an activation-aware full-quantization technique to quantize the proposed Q-TempFusion model.

Our contributions are summarized as follows:

- We propose Q-TempFusion, a quantized light-weight multi-frame and multi-sensor fusion framework for 3D object detection. To make the best use of temporal information to improve detection accuracy, a novel TFM is designed.
- Our work pioneers the model latency profiling of multi-sensor BEV models on hardware devices. The findings provide foundational insights that are expected to facilitate further advancements.
- Informed by the outcomes of our latency profiling, we introduce a novel activation-aware QAT framework specifically designed to optimize the inference efficiency and effectiveness of multi-sensor BEV detectors that incorporate temporal information.
- Our extensive experiments demonstrate that our 8-bit model achieves 70.3% mAP and 72.7% NDS with 3×∼18× FPS improvement over the leading multi-modality baselines. And our 32-bit model has comparable performance with SOTA multi-modality methodologies in 3D detection task with 72.1% mAP and 74.8% NDS, while This enhancement in speed facilitates its application in real-time inference scenarios, showcasing its practical advantages.

2. Background and Related Work

2.1. Camera-LiDAR Fusion Perception

2D camera images are rich in semantic content, yet lack precise depth information, whereas 3D LiDAR data offer exact depth information but are limited in semantic information for street environments [13–15] and traffic identification [48–50]. To leverage the strengths of both the camera and the LiDAR sensors, the researchers explore fusion methods to make use of the camera and the LiDAR data in a single model. Depending on their different feature fusion mechanisms, recent camera-LiDAR fusion methods can be categorized into three classes, including camera-to-LiDAR projection, LiDAR point augmentation, and BEV space fusion. The camera-to-LiDAR projection [22, 53] projects camera features to LiDAR points and uses projected LiDAR features to make predictions. The LiDAR point augmentation adds semantic labels [54], CNN features [55], or virtual points from 2D images [18, 64] to LiDAR points for data augmentation. The BEV space fusion [36,43] performs camera-LiDAR sensor fusion in a shared BEV space, treating foreground and background, as well as geometric and semantic information, equally. By employing fusion methods, researchers can take advantage of both modalities in a complementary manner. Although multi-modality fusion BEV models exhibit superior performance, their prohibitive inference latency precludes practical real-time applications.

2.2. Temporal BEV Models

Temporal information has become an integral feature of camera-only BEV models to improve the model's understanding of dynamic environments over time [21,34,41,61]. On the contrary, the integration of temporal information into camera-LiDAR fusion BEV models remains less explored, with only a handful of studies [5, 67] addressing this complex but potentially more informative approach. Challenges related to computational demand and the high-dimensional nature of the combined data have limited advancements in this area. Furthermore, while the inclusion of temporal information yields only a marginal performance improvement, it significantly escalates the model's complexity and inference latency. This trade-off undermines the practicality in real-time streaming applications, rendering the slight gains in performance not justifiable. Although [5] has begun to bridge this gap by introducing more efficient deformable attention-based mechanisms, the inference latency of current temporal multi-sensor fusion models remains prohibitively high, posing challenges for their deployment in real-time applications.

2.3. Quantization

Quantization can reduce the computation and memory consumption of DNNs through low-bit representations for weight and activation tensors. Fixed-point representations, such as INT8, are particularly effective in reducing energy consumption, as they enable more efficient fixed-point operations than floating-point ones [20]. There are two main types of quantization methods: QAT and PTQ. Although NLP-oriented transformers often use PTQ [52] due to their large model sizes (usually over 350M), lack of open-source large-scale datasets, and limited computational resources, ViT's smaller model size and publicly available dataset make it suitable for QAT. PTQ can lead to suboptimal model performance or significant accuracy reduction without fine-tuning. For instance, previous PTQ approaches [39, 44] quantize ViT models to 8 bits, leading to an accuracy decrease of 1.2% to 1.8%. [33] proposes a QAT method for ViTs with information-rectified and similarity-aware strategies. However, it lacks an analysis of the data distribution in the network, leading to a constrained accuracy. We aim to develop a fully quantized multi-modality BEV framework, employing QAT across ViT and 3D CNN backbones.

3. Temporal and Modality Fusion with Detailed Profiling

In Figure 2, we illustrate the overall architecture of the proposed Q-TempFusion framework. There are two

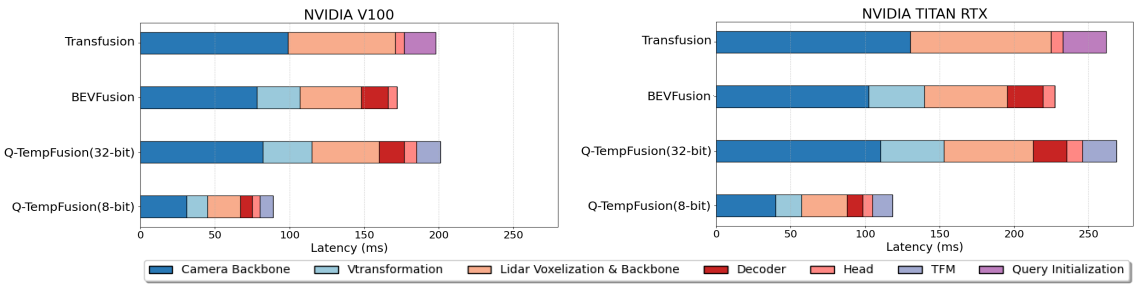


Figure 3. Device speed profiling. Results are obtained on NVIDIA TITAN RTX and NVIDIA TESLA V100. The ondevice speed for frequently used backbone and various operators is reported.

branches to deal with the multi-view input images (from cameras) and 3D point clouds (from LiDAR sensors), separately. We combine camera and LiDAR BEV features from two branches and transform them into a unified current BEV representation to preserve both geometric and semantic information. The details of this stage are shown in Appendix A. After we obtain the current unified BEV feature, a TFM is developed to aggregate historical and current BEV features. The unified representation is then passed to a sparse convolution-based BEV decoder.

Our work is pioneering in fusing temporal information from both LiDAR and camera data and is the first to apply quantization to a multi-sensor framework for real-time inference. The TFM we propose skillfully leverages historical data with a deformable attention mechanism, offering improved performance at manageable computational costs.

3.1. Temporal Fusion Module

Our work with temporal fusion uses time-varying BEV representations to capture the dynamic behavior of objects in a scene over time. It involves analyzing the changes in the BEV representation of a scene over a sequence of frames to detect and track moving objects. We present the details of the TFM in Figure 2. First, we send the current aggregated features $\boldsymbol{X}_t$ and historical BEV features $\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-N}$ to the *Structure Motion Transformation* (SMT) to obtain aligned temporal BEV features $\boldsymbol{Y}_t, \cdots, \boldsymbol{Y}_{t-N}$, drawing inspiration from [34]. Next, we concatenate the aligned temporal BEV features to have $\boldsymbol{Y}_{cat} = \mathit{Concat}(\boldsymbol{Y}_t, \cdots, \boldsymbol{Y}_{t-N})$ and feed them into a convolution-based module, resulting in the extracted temporal fusion feature $\boldsymbol{Y}_{cnn}$. We then use $\boldsymbol{Y}_t$ as the value and $\boldsymbol{Y}_{cnn}$ as the query, which we input into a deformable attention module to obtain $\boldsymbol{Y}_{attn}$. Eventually, the *Feature-wise Linear modulation* (FiLM) is presented to incorporate $\boldsymbol{Y}_{attn}$, and the current motion-transformed feature $\boldsymbol{Y}_t$.

Structure motion transformation. The efficacy of fusing historical BEV features is notably contingent on the alignment of features from different frames. Misalignment, often resulting from movements like translation or rotation of LiDAR sensors or camera platforms, can detrimentally affect fusion performance. Consequently, it is important to compensate for ego-motion and ensure proper alignment of BEV features across various frames to optimize the fusion process. To achieve this, we design the SMT with a learnable temporal motion transformation matrix $\boldsymbol{M}_t, \cdots, \boldsymbol{M}_{t-N}$. The SMT outputs transformed/aligned temporal BEV features $\boldsymbol{Y}_t, \cdots, \boldsymbol{Y}_{t-N}$ and $\boldsymbol{Y}_{t-k} = \boldsymbol{M}_{t-k}\boldsymbol{X}_{t-k},\ k = 0, \cdots, N$.

Deformable attention module. The TFM includes a deformable cross-attention module to capture spatiotemporal dynamics. It operates on a query feature map $\boldsymbol{Y}_{cnn}$ and a value feature map $\boldsymbol{Y}_t$ by computing their dot product to obtain a similarity matrix $\boldsymbol{S} = \boldsymbol{Y}_{cnn}^{\mathrm{T}} \cdot \boldsymbol{Y}_t$. This matrix captures the similarity between each location in the query feature map and the value feature map. The *Softmax* function is then applied to $\boldsymbol{S}$ along the spatial dimensions, and the result is dot-producted with the value feature map $\boldsymbol{Y}_t$ to obtain the attended value feature map $\boldsymbol{W} = \mathit{Softmax}(\boldsymbol{S}) \cdot \boldsymbol{Y}_t$. Finally, a deformable convolution operation is performed on $\boldsymbol{W}$ using the query feature map $\boldsymbol{Y}_{cnn}$ as the deformation signal, resulting in the output $\boldsymbol{Y}_{attn} = \mathit{Deformable_Conv}(\boldsymbol{Y}_{cnn}, \boldsymbol{W})$. This operation enables the network to adjust its receptive field per query location, effectively handling objects of different scales and aspect ratios.

Feature-wise linear modulation. Following the fusion technique in [46], we design a FiLM to adaptively learn the weight of $\boldsymbol{Y}_{attn}$ with respect to the current BEV feature $\boldsymbol{Y}_t$, ultimately producing our final fused BEV feature $\boldsymbol{Z}_{final} = \mu_1 \boldsymbol{Y}_t + \mu_2$, where $\mu_1 = \mathrm{Conv}_1(\boldsymbol{Y}_{attn})$ and $\mu_2 = \mathrm{Conv}_2(\boldsymbol{Y}_{attn})$.

3.2. Model Profiling on Hardware Devices

In our study, we delve into detailed profiling of our Q-TempFusion as well as some other SOTA multi-sensor models on high-performance hardware devices, specifically the Titan and V100 GPUs, as shown in Figure 3. This profiling exercise aims to meticulously measure the latency of each operator within the model. We aim to gain in-depth knowledge of our model's computational behavior and performance limitations on advanced hardware by isolating

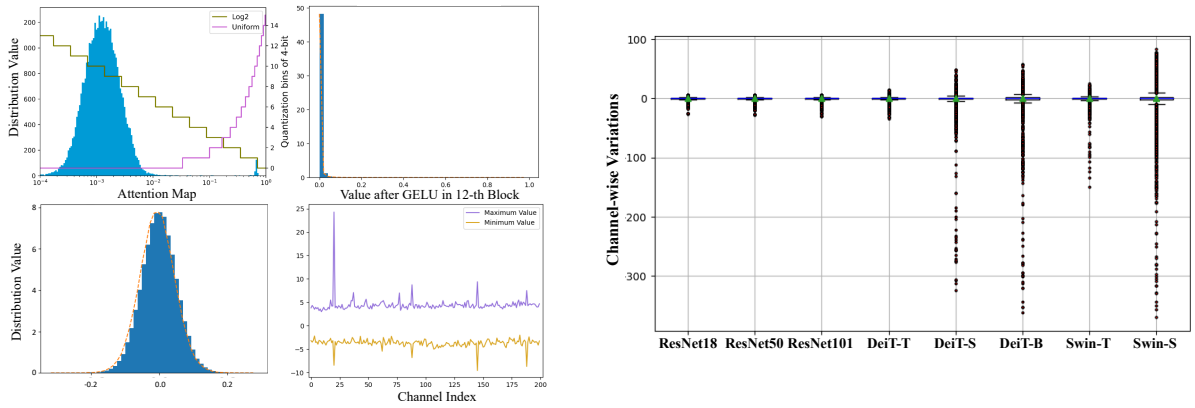


Figure 4. Data distribution. Left: (1) The long-tail distribution of attention maps visualizes 4-bit quantized bins of uniform and log2 quantization. The X-axis is in log-scale, where log2 covers more bins than uniform for small values. (2) The long-tail distribution of activation after GELU. (3) Normal distribution of the linear transform in the 6_{th} block. (4) Channel-wise minimum and maximum values of the last LayerNorm inputs in full precision Swin-T. Right: Channel-wise ranges of the last LayerNorm inputs in representative models.

and measuring individual latencies. This understanding is important for guiding optimization and ensuring efficient deployment, as well as benefiting future research.

Our empirical analysis highlights that in our model, the camera backbone using the Swin Transformer (Swin-T) contributes to 40% - 50% of the total inference latency. This considerable proportion is largely due to the self-attention mechanisms of Swin-T, which are resource-intensive but vital for nuanced feature representation. Meanwhile, the LiDAR backbone employing VoxelNet accounts for about 25% - 30% of the total latency. Given that these two components together comprise roughly 65% - 80% of total latency, our paper proposes targeted quantization strategies for these parts. Furthermore, in Figure 3, we can see that other multi-sensor fusion BEV models also exhibit similar profiling characteristics (with backbones typically incurring the majority of computational overhead). Therefore, our quantization strategy with a high potential for generalizability can be seamlessly adapted to optimize other models, promising widespread improvements across various implementations. This approach is particularly pertinent for real-time applications where inference speed is paramount, and it sets the stage for efficient deployment of multi-sensor BEV models in practical scenarios.

4. Activation-Aware Full-Quantization for BEV

We implement activation-aware full-quantization for the two branches, substantially speeding up computation while preserving high detection accuracy.

4.1. Quantization for 2D Branch

The data distributions of CNNs typically follow normal distributions due to their linear operations and BatchNorm layer [7]. Thus, we adopt *uniform quantization* [23] for the CNN-based architecture in Q-TempFusion.

For ViT-based structures, we make a few key observations about their data distributions, as shown in Figure 4, which are different from CNNs. In ViTs, weights typically follow normal distributions. However, activations have long-tailed distributions (for attention maps and activation value after GELU) and systematic channel-wise outliers before LayerNorms.

Based on these observations, we implement (i) uniform quantization on weight, (ii) log2 quantization in attention maps, and (iii) uniform quantization on activation values after GELU, clipped to the range [-1, 10]. (iv) We introduce the Power-of-Two Ratio (PTR) for LayerNorm quantization to manage channel-wise outliers and use outlier-aware training for predicting channel indexes and regularized scales, eliminating the need for dynamic data selection during inference.

Log2 quantization for attention maps. Given the quantization bit-width b, the quantizer $Q(X|b)$ can be formulated as a function mapping a floating-point number $X \in R$ to the nearest quantization bin. Since the output range $(0, 1)$ of Softmax makes the $\log 2$ quantization calibration-free and log2 can convert the MatMul to more hardware-efficient BitShift, we propose:

$$\begin{aligned} \text{Attn}_Q &= Q(\text{Attn}|b) = \text{clip}(\lfloor -\log_2(\text{Attn}) \rceil, 0, 2^b - 1), \\ \text{Attn}_Q \cdot V_Q &= 2^{-\text{Attn}_Q} \cdot V_Q = V_Q >> \text{Attn}_Q \\ &= \frac{1}{2^M} \cdot (V_Q << (M - \text{Attn}_Q)), \end{aligned} \tag{1}$$

where $M = 2^b - 1$, V_Q represents the quantized V in

self-attention and Attn_Q represents the quantized attention maps. Note that the direct right shift V_Q with Attn_Q may result in a severe truncation error. $M-\text{Attn}_Q$ with scaling $\frac{1}{2^M}$ can prevent the error by performing a left-shift operation.

Uniform quantization for activations after GELU. For the activations after GELU, GELU truncates activation values to be concentrated around 0, while the rest keep the original normal distribution. To allocate enough precision bits, we *clip* the activation value to [-1, 10] after GELU with uniform quantization. Moreover, we adapt GELU to support integer-only inference and regularize the quantization error as [19].

PTR quantization for systematic channel-wise outliers. The input of LayerNorm contains many systematic and channel-wise outliers. To address channel-wise outliers, traditional solutions employ group-wise [51] or channel-wise [28] quantization, which necessitates computing mean and variance in the floating-point domain, leading to high hardware overhead. To mitigate this, we propose a PTR for LayerNorm quantization, which regulates channel-wise outliers using different ratios rather than different quantization parameters.

Given the input activation $X \in B \times L \times C$, and the PTR $r \in N^C$, the quantized activation X_Q can be formulated as:

$$X_Q = Q(X|b) = \text{clip}(\lfloor \frac{X}{2^r s} \rceil + z, 0, 2^b - 1), \tag{2}$$

where

$$s = \frac{\max(X) - \min(X)}{2^R(2^b - 1)}, \quad z = \text{clip}(\lfloor -\frac{\min(X)}{\max(X)} \rceil, 0, 2^b - 1). \tag{3}$$

The hyperparameter $R = \max(r) = 4$ satisfies different inter-channel variations between models. We predict each outlier's channel index i and PTR r using the outlier-aware training algorithm in Sec. 4.3, which obviates the need for complex control logic during inference to support data-adaptive selection.

4.2. Quantization for 3D Branch

The inherent sparsity and irregular distribution of LiDAR point clouds make it challenging for the quantization of 3D LiDAR-based detectors. There is an imbalance between foreground instances and a large redundant background area. These small foreground instances and the large perception ranges in 3D detection require the quantized model to have minimal information loss to maintain detection performance. Therefore, these challenges hinder the direct application of quantization methods developed for 2D vision tasks to 3D point cloud tasks.

For the 3D backbone model, we adopt uniform signed symmetric quantization to achieve significant acceleration. We quantize both the weights and the activations of the model. Since activation quantization is input-dependent, it often requires a few batches of calibration data for the estimation of the dynamic ranges to converge. To approximate the real-valued input, we perform the de-quantization step $\hat{x} = (x_{\text{int}} - p) \cdot q$, where $\hat{x}$ is the de-quantized floating-point value with an error introduced during the quantization process, p is zero-point. q denotes the quantization scale factor, which reflects the proportional relationship between floating-point values and integers. Following [1], we employ a max-min calibrator equipped with a lightweight grid search to appropriately initialize the quantization parameters for both weights and activations. The quantization range is defined as $x_{\text{max}} = \max(|x|), x_{\text{min}} = -x_{\text{max}}$ to cover the whole dynamic range of the floating point value x. This leads to no clipping error. To obtain a fine-grained scale factor and avoid the influence of outliers on rounding errors, a lightweight grid search is incorporated to further optimize the quantization parameters. Please refer to Appendix B for more quantization details.

4.3. Quantization-Aware Training Flow of Q-TempFusion

Outlier-aware training. This algorithm quantizes the 2D branch with three steps: • Initialize the PTR by detecting the outlier of the full-precision model with 3σ method [12]. • Search for the channel index i and the PTR r of each outlier with the l_2 minimization. • Fix the index i and r obtained in step 2 and fine-tune the model. We also utilize the distillation loss DGD in Q-ViT [33] to transfer the knowledge of the larger teacher model into the quantized model, e.g., transform the knowledge from SwinV2-T to 8-bit Swin-T. The detailed algorithm is shown in Appendix C.

Overall quantization-aware training flow. First, we utilize 100 epochs ($\frac{1}{3}$ of the train-from-scratch) to quantize the full precision backbone model into 8-bit with the outlier-aware training method on the ImageNet dataset. Because we have determined the index and scale of outliers in the backbone pre-train stage, after the backbone quantization, we just integrate the 8-bit image backbone into Q-TempFusion and quantize the whole detection system in a standard finetuned way (8 epochs) with nuScenes Dataset.

Inference with Int-LayerNorm. During inference, with the gathered i and r of systematic outliers, we extract the layer-wise parameters s & z, and compute μ & σ in the integer domain as below. Meanwhile, thanks to the BitShift, PTR r can be calculated with the quantized Layernorm efficiently (refer to [25]).

$$\begin{aligned} \widehat{X_Q} &= (X_Q - z) << r, \\ \mu(X) &\approx \mu(2^r s \cdot (X_Q - z)) = s \cdot \mu(\widehat{X_Q}), \\ \sigma(X) &\approx \sigma(2^r s \cdot (X_Q - z)) = s \cdot \sigma(\widehat{X_Q}). \end{aligned} \tag{4}$$

Table 1. Performance comparison on nuScenes test set (NVIDIA V100). 'L' is LiDAR, 'C' for camera, and 'T' for temporal information.

Method	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Tra.	Bar.	Motor.	Bike	Ped.	T.C.	FPS
PointPainting [54]	LC	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4	4.8
3D-CVF [65]	LC	52.7	62.3	83.0	45.0	15.9	48.8	49.6	65.9	51.2	30.4	74.2	62.9	-
PointAugmenting [55]	LC	66.8	71.0	87.5	57.3	28.0	65.2	60.7	72.6	74.3	50.9	87.9	83.6	4.0
MVP [64]	LC	66.4	70.5	86.8	58.1	32.1	70.2	61.0	73.9	78.5	52.9	87.1	83.8	5.1
TransFusion [2]	LC	68.9	71.7	87.1	60.0	33.1	68.3	60.8	78.1	73.6	52.9	88.4	86.7	4.3
FusionPainting [59]	LC	68.1	71.6	87.1	60.8	30.0	68.5	61.7	71.8	74.7	53.5	88.3	85.0	-
AutoAlignV2 [9]	LC	68.4	72.4	87.0	59.0	33.1	69.3	59.3	-	72.9	52.1	87.6	-	3.6
VFF [32]	LC	68.4	72.4	86.8	58.1	32.1	70.2	61.0	73.9	78.5	52.9	87.1	83.8	-
FUTR3D [8]	LC	69.4	72.1	-	-	-	-	-	-	-	-	-	-	2.5
BEVFusion [36]	LC	69.2	71.8	88.1	60.9	34.4	69.3	62.1	78.2	72.2	52.2	89.2	85.2	0.6
BEVFusion [43]	LC	70.2	72.9	88.6	60.1	39.3	69.8	63.8	80.0	74.1	51.0	89.2	86.5	5.8
UVTR [29]	LC	67.1	71.1	87.5	56.0	33.8	67.5	59.5	73.0	73.4	54.8	86.3	79.6	1.9
DeepInteraction [62]	LC	70.8	73.4	87.9	60.2	37.5	70.8	63.8	80.4	75.4	54.5	90.3	87.0	1.3
MSMDFusion [24]	LC	71.5	74.0	88.4	61.0	35.2	71.4	64.2	80.7	76.9	58.3	90.6	88.1	1.5
LIFT [67]	LCT	65.1	70.2	87.7	55.1	29.4	62.4	59.3	69.3	70.8	47.7	86.1	83.2	-
CMT [60]	LCT	72.0	74.1	-	-	-	-	-	-	-	-	-	-	-
BEVFusion4D [5]	LCT	**73.3**	74.7	89.7	**65.6**	**41.1**	**72.9**	**66.0**	81.0	79.5	**58.6**	**90.9**	87.7	2.0
Q-TempFusion (8-bit)	LCT	70.3	72.7	88.2	59.8	38.5	69.5	61.7	79.0	72.6	51.7	88.7	84.9	**13.2**
Q-TempFusion (32-bit)	LCT	72.1	**74.8**	**89.9**	64.9	40.9	70.7	63.3	**81.6**	**79.7**	57.1	89.5	**88.2**	4.4

5. Experiments

In this section, we show our experimental results with detailed analysis to demonstrate the efficacy of the proposed Q-TempFusion framework. Our approach is evaluated in comparison to other SOTA multi-sensor methods, focusing on both detection precision and inference latency. Furthermore, we conduct a series of ablation studies to assess the impact of TFM, quantization method, and the optimal length of historical frames for aggregation.

5.1. Dataset and Experiment Settings

Dataset. The nuScenes dataset [4] is a widely used dataset designed for autonomous driving research. It consists of nearly 1000 scenes divided into three sets: 700 for training, 150 for validation, and 150 for testing. Each scene is captured by six cameras and one LiDAR sensor, with 20-second recordings. The cameras capture RGB images at 12 frames per second, providing a 360-degree horizontal field of view. The LiDAR sensor has 32 beams and records at 20 frames per second. Keyframes are provided at 2Hz to ensure synchronization between the cameras and LiDAR. For object detection, the previous nine frames are transformed into the current point cloud and annotated at 2Hz. The dataset supports ten detection categories and is evaluated using mean Average Precision (mAP) and nuScenes detection scores (NDS).

Experimental settings. Our implementation is based on the public code base mmdetection3d [11]. For the camera branch, we use Swin-T [42] with FPN [37] as the feature extractor and initialize it from the instance segmentation model Cascade Mask R-CNN [6] pretrained on COCO [38] and then nuImage [4], which is previously adopted by BEVFusion [43]. For the LiDAR branch, we adopt VoxelNet [70] as our LiDAR backbone. The multi-view aggregated image feature dimension is set to 80 and the LiDAR feature dimension is set to 256. The voxel size of 0.075 and the image size of 256×704 are adopted by default like [43] to make a fair comparison. Our model is trained with a batch size of 4 on 8 Titan GPUs. It is trained for a total of 8 epochs with CBGS [71]. We adopt the AdamW [45] optimizer with a learning rate of $2\mathrm{e}^{-4}$ and a weight decay of 10^{-2} for optimization. For the QAT part, we use 100 epochs to quantize the full precision 2D backbone model Swin-T into 8-bit with the training receipt and the same distillation as [33] on the ImageNet dataset. We apply symmetric channel-wise quantization on weights and asymmetric layer-wise quantization on activations. The hyperparameter R is set to 4. The training settings follow Q-ViT [33] with its DGD distillation. Because we execute the three-step outlier-aware training in a fine-tuned way, the training hours of the entire pipeline can be 37.5% of train-from-scratch of full precision models. Our model is benchmarked against SOTA detection baselines. Detailed information on these baseline models is provided in Appendix D.

Table 2. Abaltion study of TFM on nuScenes validation dataset.

Method	mAP	NDS	FPS
Q-TempFusion (32-bit without TFM)	69.7	72.3	5.2
Q-TempFusion (32-bit with TFM)	70.9	73.4	4.4
Q-TempFusion (8-bit without TFM)	68.3	70.9	13.8
Q-TempFusion (8-bit with TFM)	69.4	71.9	13.2

5.2. Comparison with State-of-the-Art Methods

We compare the proposed Q-TempFusion with SOTA approaches on the nuScenes test set and present the results in Table 1. The comparison demonstrates that our 32-bit

Table 3. Impact of Component-wise Quantization on Model Performance and Latency on nuScenes validation dataset. In this table, "Camera-B" refers to the camera backbone, "LiDAR-B" signifies the LiDAR backbone, and "Vtrans" represents the view transformation component in the framework.

Camera-B	LiDAR-B	Vtrans	TFM	mAP	NDS	FPS
✔	✗	✗	✗	71.0	73.5	7.4
✗	✔	✗	✗	70.9	73.2	6.1
✗	✗	✔	✗	71.2	73.6	5.0
✗	✗	✗	✔	71.4	73.7	4.7
✔	✔	✔	✔	69.4	71.9	13.2

Q-TempFusion outperforms all previous methods, achieving new SOTA performances with 72.1% mAP, 74.8% NDS and 4.4 FPS on the nuScenes test dataset. The 8-bit Q-TempFusion model demonstrates comparable accuracy, achieving 70.3% mAP and 72.7% NDS, akin to the leading multi-modality baselines. Notably, it attains a close-to-real-time inference speed of 13.2 FPS, thereby surpassing all existing SOTA multi-modality methods in latency by 3×∼18× improvements. Specifically, our approach consistently outperforms SOTA methods in most object detection categories including car, truck, construction vehicle (C.V.), barrier (Bar.), Motorcycle (Motor.) and traffic cone (T.C.). The experimental results show the significance of temporal fusion for both LiDARs and cameras in improving the model's performance.

5.3. Ablation Studies

We conduct every experiment with 8 epochs and ablate the effect of the TFM and our quantization design.

Temporal fusion module. In Table 2, we compare with the baseline without TFM, which is obtained by replacing our TFM with a naive temporal feature concatenation. Without the TFM, the accuracy performance drops by 1.2% mAP and 1.1% NDS. Moreover, we notice that even without TFM in our model, our result is still better than the previous BEVFusion in Table 1, demonstrating the effectiveness of temporal fusion for both cameras and LiDAR features.

Component-wise quantization. In Table 3, we explored the impact of quantizing different main components of Q-TempFusion, including camera and LiDAR backbones, view transformation component and TFM. On the basis of the profiling results, the decoder and prediction head have a negligible impact on the total latency of the model. Therefore, these two components are excluded here. This ablation study aims to assess the balance between computational efficiency, and performance accuracy. By applying quantization to each component separately, we identify which parts of the model can tolerate lower precision with minimal accuracy loss. Our experiments reveal that quantization applied to the LiDAR backbone results in a more pronounced reduction in accuracy compared to other model components. In contrast, our view transformation component and TFM demonstrate a higher resilience to activation-aware QAT. Additionally, quantization of the camera and LiDAR backbones leads to more substantial improvements in FPS over view transformation part and TFM. This observation aligns with our profiling results, which indicates that the two backbones account for the majority of the model's total latency.

Table 4. Comparision of Q-TempFusion 32-bit & 8-bit between different lengths of historical information on nuScenes validation dataset. "N = 0" refers to the absence of historical information. "N = 5" denotes the use of features from the last five historical frames combined with the current feature.

32-bit	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5
mAP	68.5	69.5	70.1	70.5	70.8	70.9
NDS	71.0	71.8	72.5	73.0	73.3	73.4
8-bit	**N = 0**	**N = 1**	**N = 2**	**N = 3**	**N = 4**	**N = 5**
mAP	67.9	68.4	68.8	69.1	69.3	69.4
NDS	70.4	71.0	71.4	71.8	71.9	71.9

Temporal fusion length. In Table 4, we evaluated the impact of varying temporal lengths on model performance. The baseline model (without temporal information, $\mathrm{N} = 0$) was established using the current feature $\boldsymbol{Y}_t$ instead of the historical aligned feature $\boldsymbol{Y}_{cat}$. We then incrementally increased the number of historical frames (from $\mathrm{N} = 1$ to $\mathrm{N} = 5$) for subsequent baselines. The results in Table 4 indicate a progressive improvement in model performance with the inclusion of additional historical frames. However, beyond $\mathrm{N} = 4$, no significant performance gains are observed. Moreover, the performance enhancement becomes less significant as N increases from 0 to 5, highlighting the diminishing returns of older temporal information compared with more recent data.

6. Conclusion

This paper presents Q-TempFusion, an efficient spatiotemporal multi-modality framework for 3D object detection. We propose the temporal fusion module based on a deformable attention mechanism, which efficiently aggregates historical information while maintaining a low requirement for memory and computation resources. To inform the quantization strategy for Q-TempFusion, we conduct a model profiling on various multi-modality BEV approaches. Our experiments demonstrate that Q-TempFusion (8-bit) matches the accuracy of leading methods and significantly increases FPS by 3×∼18×, demonstrating marked efficiency improvements, while Q-TempFusion (32-bit) achieves comparable SOTA performance in multi-modality 3D detection. Our model can set a new standard for **real-time** BEV 3D detection in multi-sensor fusion models, sparking new directions in future research.

References

[1] Anonymous. LiDAR-PTQ: Post-training quantization for point cloud 3d object detection. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. under review. 6

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022. 7

[3] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. *EMNLP*, 2021. 2

[4] 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. 7

[5] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, Wenbo Ding, and Jiu-Yang Zhao. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. *ArXiv*, abs/2303.17099, 2023. 2, 3, 7, 12

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation, 2019. IEEE transactions on pattern analysis and machine intelligence. 7

[7] Sung-En Chang, Yanyu Li, Mengshu Sun, Runbin Shi, Hayden K-H So, Xuehai Qian, Yanzhi Wang, and Xue Lin. Mix and match: A novel fpga-centric deep neural network quantization framework. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 208–220. IEEE, 2021. 5

[8] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection, 2023. 7

[9] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. *arXiv preprint arXiv:2207.10316*, 2022. 7

[10] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3009–3018. IEEE, 2019. 2

[11] MMDetection3D Contributors. Mmdetection3d: Openmmlab next-generation platform for general 3d object detection, 2020. https://github.com/open-mmlab/mmdetection3d. 7

[12] Wikipedia contributors. 68–95–99.7 rule. In *Wikipedia, The Free Encyclopedia.*, pages 0–1, 2022. 6

[13] Han-Cheng Dan, Zhetao Huang, Bingjie Lu, and Mengyu Li. Image-driven prediction system: Automatic extraction of aggregate gradation of pavement core samples integrating deep learning and interactive image processing framework. *Construction and Building Materials*, 453:139056, 2024. 3

[14] Han-Cheng Dan, Bingjie Lu, and Mengyu Li. Evaluation of asphalt pavement texture using multiview stereo reconstruction based on deep learning. *Construction and Building Materials*, 412:134837, 2024. 3

[15] Han-Cheng Dan, Peng Yan, Jiawei Tan, Yinchao Zhou, and Bingjie Lu. Multiple distresses detection for asphalt pavement using improved you only look once algorithm based on convolutional neural network. *International Journal of Pavement Engineering*, 25(1):2308169, 2024. 3

[16] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 33:28089–28154, 2021. 2

[17] Peiyan Dong, Zhenglun Kong, Xin Meng, Pinrui Yu, Yifan Gong, Geng Yuan, Hao Tang, and Yanzhi Wang. Hotbev: Hardware-oriented transformer-based multi-view 3d detector for bev perception. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[18] Peiyan Dong, Zhenglun Kong, Xin Meng, Peng Zhang, Hao Tang, Yanzhi Wang, and Chih-Hsien Chou. Speeddetr: Speed-aware transformers for end-to-end object detection. 2023. 3

[19] Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, et al. Heatvit: Hardware-efficient adaptive token pruning for vision transformers. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 442–455. IEEE, 2023. 6

[20] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14. IEEE, 2014. 3

[21] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection, 2022. arXiv. 3

[22] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 35–52. Springer, 2020. 1, 3

[23] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018. 2, 5

[24] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21643–21652, 2023. 7

[25] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR, 2021. 6

[26] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning. *arXiv preprint arXiv:2112.13890*, 2021. 2

[27] Zhenglun Kong, Haoyu Ma, Geng Yuan, Mengshu Sun, Yanyue Xie, Peiyan Dong, Xin Meng, Xuan Shen, Hao Tang, Minghai Qin, et al. Peeling the onion: Hierarchical reduction of data redundancy for efficient vision transformer training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8360–8368, 2023. 2

[28] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2810–2819, 2019. 2, 6

[29] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *arXiv preprint arXiv:2206.00630*, 2022. 7

[30] Yuqi Li, Qingqing Long, Yihang Zhou, Ning Cao, Shuai Liu, Fang Zheng, Zhihong Zhu, Zhiyuan Ning, Meng Xiao, Xuezhi Wang, et al. Comae: Comprehensive attribute exploration for zero-shot hashing. *arXiv preprint arXiv:2402.16424*, 2024. 2

[31] Yuqi Li, Yao Lu, Zeyu Dong, Chuanguang Yang, Yihao Chen, and Jianping Gou. Sglp: A similarity guided fast layer partition pruning for compressing large deep models. *arXiv preprint arXiv:2410.14720*, 2024. 2

[32] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1120–1129, 2022. 7

[33] Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-vit: Accurate and fully quantized low-bit vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2022. 3, 6, 7

[34] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022. 2, 3, 4, 12

[35] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 1

[36] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. 2022. 2, 3, 7, 12

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[39] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Fully quantized vision transformer without retraining. *International Joint Conferences on Artificial Intelligence Organization*, 34:28092–28103, 2022. 3

[40] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 531–548. Springer, 2022. 1

[41] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 3

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 2, 7

[43] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. 2023. 2, 3, 7

[44] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 28092–28103, 2021. 2, 3

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. arXiv preprint arXiv:1711.05101. 7

[46] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 4

[47] Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. Fully quantized transformer for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1–14, 2020. 2

[48] Fei Shen, Xiaoyu Du, Liyan Zhang, and Jinhui Tang. Triplet contrastive learning for unsupervised vehicle re-identification. *arXiv preprint arXiv:2301.09498*, 2023. 3

[49] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31th ACM International Conference on Multimedia*, 2023. 3

[50] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing*, 2023. 3

[51] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821, 2020. 2, 6

[52] Xuan Shen, Peiyan Dong, Lei Lu, Zhenglun Kong, Zhengang Li, Ming Lin, Chao Wu, and Yanzhi Wang. Agilequant: Activation-guided quantization for faster inference of llms on the edge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18944–18951, 2024. 3

[53] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019. 1, 3

[54] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 1, 3, 7

[55] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021. 1, 3, 7

[56] Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In *International Conference on Machine Learning*, pages 9847–9856. PMLR, 2020. 2

[57] Zihao Wei, Yidong Huang, Yuang Chen, Chenhao Zheng, and Jingnan Gao. A-esrgan: Training real-world blind super-resolution with attention u-net discriminators. In *Pacific Rim International Conference on Artificial Intelligence*, pages 16–27. Springer, 2023. 2

[58] Guangxuan Xiao, Ji Lin, Mickael Seznec, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. *Advances in Neural Information Processing Systems*, 34:27318–27313, 2022. 2

[59] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3047–3054. IEEE, 2021. 7

[60] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer via coordinates encoding for 3d object dectection. *arXiv preprint arXiv:2301.01283*, 2023. 7

[61] Chenyu Yang, Yuntao Chen, Haofei Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Y. Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17830–17839, 2022. 3

[62] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *arXiv preprint arXiv:2208.11112*, 2022. 7, 13

[63] Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183, 2022. 2

[64] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021. 1, 3, 7

[65] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 720–736. Springer, 2020. 7

[66] Ali Hadi Zadeh, Isak Edo, Omar Mohamed Awad, and Andreas Moshovos. Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 811–824. IEEE, 2020. 2

[67] Yihan Zeng, Da Zhang, Chunwei Wang, Zhenwei Miao, Ting Liu, Xin Zhan, Dayang Hao, and Chao Ma. Lift: Learning 4d lidar image fusion transformer for 3d object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17151–17160, 2022. 2, 3, 7

[68] Zihan Zhao, Yuncong Liu, Lu Chen, Qi Liu, Rao Ma, and Kai Yu. An investigation on different underlying quantization schemes for pre-trained language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 359–371. Springer, 2020. 2

[69] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6945–6956, 2023. 1

[70] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 7

[71] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 7