REPURPOSING INFLUENCE DIAGNOSTICS FOR LOW-BIT 3D OBJECT DETECTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D object detection is essential for instance-level scene understanding, yet deploying computationally intensive detectors on resource-constrained devices often requires low-bit quantization. However, reducing bit width remains challenging because 3D detectors exhibit activation distributions in which high-magnitude outliers are not uniformly informative. While some outliers carry task-relevant information, others arise from incidental or nuisance patterns. Such uninformative outliers can unnecessarily broaden the activation range, leading to coarse discretization and feature distortion under low-bit quantization. To address this issue, we repurpose influence diagnostics for quantization, where influence reflects how strongly an individual activation affects task-relevance. Our approach builds on the statistical perspective that *informative outliers* are crucial for preserving the properties of the data distribution. This perspective aligns with the classical notion of *influential observations*, motivating us to identify informative outliers by assessing their impact on task-relevant information. To make this influence task-aware, we use activation gradients to capture the relevance of individual activations to the task objective. Because these gradients contain both dominant task-relevant patterns and incidental variations, we model their shared structure with a low-rank representation that retains the principal patterns to be preserved. We then assess activation influence by measuring how strongly each activation affects the resulting gradient structure. Experiments on the nuScenes benchmark demonstrate improvements across W8A8, W4A16, W4A8, and W4A4 settings, with larger gains at lower bit widths. These results suggest that activation influence provides an effective criterion for identifying informative outliers in low-bit quantization.

1 INTRODUCTION

3D object detection enables instance-level scene understanding for video analytics and robotics, but high-capacity detectors built on modern architectures (He et al., 2016; Vaswani et al., 2017) often incur substantial computational and memory costs, limiting their deployment on resource-constrained devices. Quantization is therefore widely used to enable efficient low-bit computation (Chen et al., 2020; Stäcker et al., 2021), yet reducing activation precision inevitably incurs information loss (Qin et al., 2020), making the preservation of task-relevant information a central challenge.

To preserve task-relevant information, it is necessary to identify which activations are task-relevant and characterize how their values are distributed. This is particularly challenging in real-world 3D vision, where task-relevant signals are often entangled with background clutter, sensor noise, occlusion, and irregular geometric structures (Zhang et al., 2023; Zhou et al., 2024). Due to this property, 3D detection models often exhibit activation distributions, in which high-magnitude outliers are not uniformly informative: some encode task-relevant cues, whereas others arise from incidental or nuisance patterns. These uninformative outliers can unnecessarily expand the activation range, leading to coarse discretization under low-bit quantization and consequently distorting information in the feature representations (Kim et al., 2026). Conventional approaches (Xiao et al., 2023; Li et al., 2025; Lin et al., 2024) mitigate this issue by shifting outlier magnitudes from activations to weights. However, under long-tailed activation distributions, rare uninformative outliers can dominate the channel-wise scaling factors. This can reduce the effective quantization resolution of typical activations while amplifying the corresponding weight channels, thereby increasing weight quantization error. Distinguishing informative outliers from uninformative ones is therefore critical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

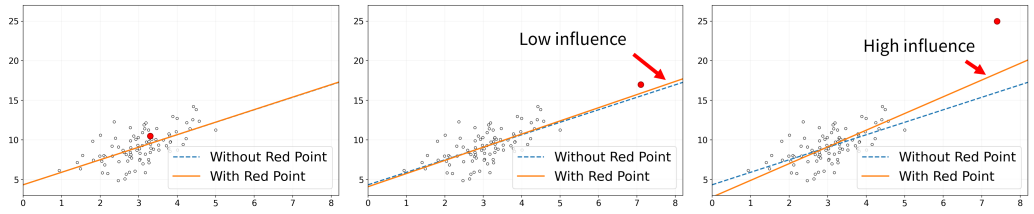


Figure 1: **Categorization of activation observations.** Outlying activations differ in importance according to their influence on the dominant representation structure. We first identify influential activations and then extract task-relevant outliers within this set.

for overcoming this limitation, yet a clear criterion for what constitutes an informative outlier has yet to be established, leaving their reliable identification unresolved in low-bit quantization.

To distinguish informative outliers, we draw inspiration from the classical statistical view of outliers (Grubbs, 1969). In statistics, an outlier is typically defined as an observation that deviates markedly from the overall pattern of the data. The extent of this deviation can be further characterized using *influence diagnostics* (Hoaglin & Welsch, 1978; Cook, 2000), which provide a criterion for identifying observations that exert a disproportionate influence on statistical estimates. This influence of an observation arises from two complementary factors: leverage, reflecting how much the observation deviates from the rest of the data (Hoaglin & Welsch, 1978), and residual discrepancy, reflecting how poorly the observation is explained by the fitted model (Cook & Weisberg, 1982). As illustrated in Figure 1, a highly influential observation can perturb the resulting statistical estimates.

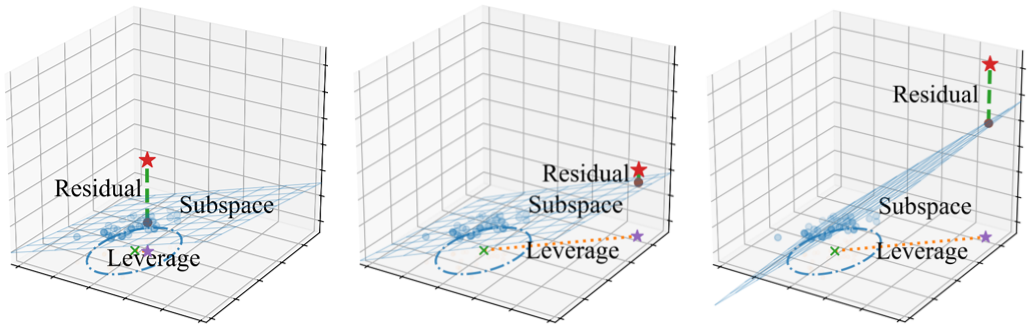
Motivated by this insight, we measure activation influence with respect to the task objective to identify informative outliers. To obtain a task-aware influence measure, we first construct an objective-aware representation that captures how activations affect the task objective. Specifically, we use layer-wise gradients to quantify the sensitivity of the task objective to individual activations (Selvaraju et al., 2017). We then extract the dominant gradient structure via a low-rank representation, which captures principal task-relevant patterns, and use its preservation as a proxy for retaining task-relevant information (Papyan, 2020). Influence diagnostics on the resulting structure identify activations that contribute most to preserving the dominant patterns. The identified activations then guide existing outlier-aware quantization methods (Xiao et al., 2023; Lin et al., 2024). Experiments on nuScenes show improvements across W8A8, W4A16, W4A8, and W4A4 settings, with larger gains under more aggressive quantization. The results demonstrate that activation influence serves as an effective criterion for identifying informative outliers.

2 Related Work

2.1 Low-Bit 3D Object Detection

Long-tailed activation distributions in 3D object detection models induce wide activation ranges, causing coarse quantization of dense regions and degraded low-bit performance. As 3D quantization research has progressed, the sources of this difficulty have become clearer. LiDAR-PTQ (Zhou et al., 2024) identifies sparse point clouds, large voxelized ranges, and foreground–background imbalance as key obstacles to PTQ, addressing them with sparsity-aware calibration, task-guided loss, and adaptive rounding. Inlier-Centric PTQ (Kim et al., 2026) offers a modality-agnostic perspective, attributing degradation to anomalous, long-tailed activations induced by cluttered backgrounds, irregular object structures, and redundant features. By calibrating with inlier activations, it shows that robust activation statistics are crucial across 3D detection modalities. These works indicate that effective 3D quantization requires not only minimizing reconstruction error but also using calibration statistics robust to sparse, imbalanced, and long-tailed representations.

Building on these works, we analyze activations from a regression-diagnostic perspective. Rather than treating all outliers uniformly, we identify important outliers and narrow activation ranges by measuring the influence of layer-wise activation vectors on the low-rank structure.



Low Leverage, Large Residual | High Leverage, Small Residual | High Leverage, Large Residual

Figure 2: **Various leverage and residual cases in linear regression.** *Leverage* indicates deviation in the input subspace, whereas *Residual* indicates deviation from the fitted subspace.

2.2 Regression Diagnostics

Classical regression diagnostics characterize outliers as observations that deviate markedly from the rest of the data and use statistical criteria to assess their significance (Grubbs, 1969). Regression diagnostics extend this view by analyzing how individual observations affect model estimation. Hoaglin & Welsch (1978) proposed using the diagonal elements of the hat matrix to identify high-leverage points, highlighting leverage as a measure of strong influence on model fitting. Cook & Weisberg (1982) studied influential observations through residual-based diagnostics and proposed Cook's distance (Cook, 2000) to quantify individual influence by jointly accounting for residual and leverage. Inspired by these perspectives, we treat activation vectors as observations and measure their influence on the dominant representation structure. This provides a criterion for selecting important outliers and mitigating quantization degradation from long-tailed activations.

3 Preliminaries

Leverage and Residual. Regression diagnostics characterize each observation relative to the fitted model through two complementary quantities: *leverage* (Hoaglin & Welsch, 1978) and *residual* (Cook & Weisberg, 1982). Leverage measures how unusual an observation is with respect to the geometry of the design space, whereas the residual measures its deviation from the fitted response.

Consider the linear regression model $y_i = \mathbf{z}_i^\top \boldsymbol{\beta} + \varepsilon_i$, where $\mathbf{z}_i \in \mathbb{R}^p$ is the predictor vector of the i-th observation, $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the regression coefficients, and ε_i is the error term. Given the full-column-rank design matrix $\mathbf{Z} = [\mathbf{z}_1; \cdots; \mathbf{z}_N] \in \mathbb{R}^{N \times p}$, the ordinary least-squares estimate yields the fitted response $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, where

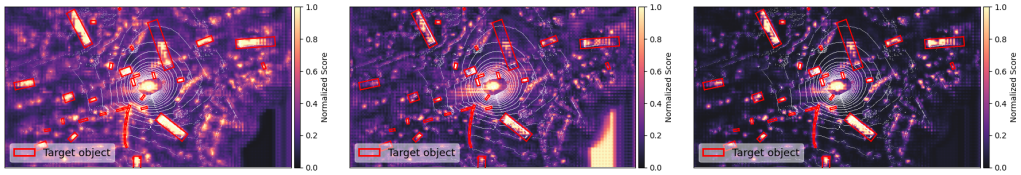
$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top, \qquad v_i = \mathbf{H}_{ii}, \qquad e_i = y_i - \hat{y}_i. \tag{1}$$

Here, $\mathbf{H}$ is the hat matrix, v_i is the leverage of the i-th observation, and e_i is its residual. A large v_i indicates that the observation is strongly represented in the fitted design subspace, whereas a large $|e_i|$ indicates substantial disagreement between the observed and fitted responses. Together, leverage and residual provide complementary diagnostics of observation influence, as illustrated in Figure 2.

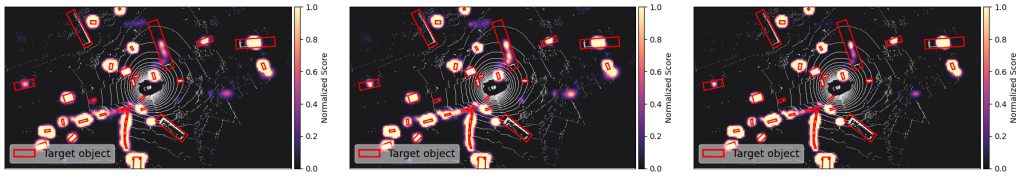
Influential Observation. An observation is considered *influential* if its presence or removal substantially changes the fitted model. Neither leverage nor residual alone is sufficient: a high-leverage observation may still be well fitted, whereas a large-residual observation with low leverage may have limited impact on the fit. Influence diagnostics therefore combine these two complementary quantities. Cook's distance is a representative measure of observation influence in linear regression (Cook, 2000). For the i-th observation, it can be expressed as

$$D_i = \underbrace{\frac{e_i^2}{s^2(1-v_i)}}_{\text{residual}} \underbrace{\frac{v_i}{p(1-v_i)}}_{\text{leverage}}, \tag{2}$$

where e_i is the residual, s^2 is the estimated residual variance, p is the number of regression parameters, and $v_i = \mathbf{H}_{ii}$ is the leverage. Thus, D_i becomes large when an observation exhibits both a large



(a) **Feature Domain.**



(b) **Gradient Domain.**

Figure 3: **Visualization of Activation Scores at a Decoder Layer.** Activation scores are visualized in the egocentric bird's-eye view, with bounding boxes overlaid to indicate target objects. Feature-domain measurements assign high scores to salient activation values, including those that may not be relevant to the detection task. In contrast, gradient-domain measurements reflect task sensitivity, providing precise guidance for identifying influential activations in 3D object detection.

residual and high leverage. This decomposition motivates our low-rank formulation, where we construct analogous residual and leverage components to characterize activation influence with respect to the fitted low-rank structure. These components are detailed in Sections 4.2 and 4.3.

4 METHOD

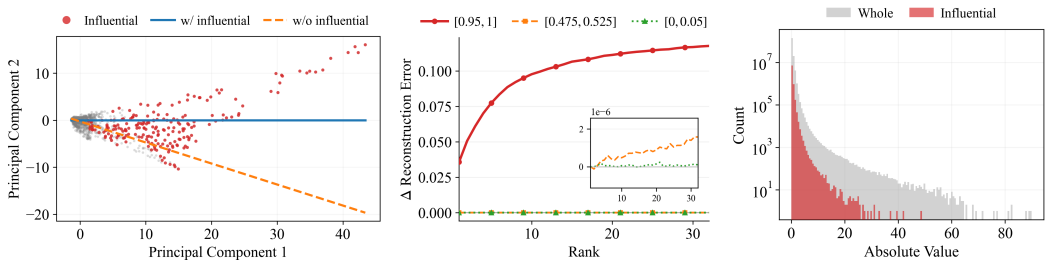
We extend the notion of influential observations to define an activation set for quantization calibration and identify such activations through a layer-wise influence measurement framework. We first formulate the influence scoring problem in Section 4.1. We then introduce two complementary activation-level measures (see Figure 3): a leverage score in Section 4.2, which measures the contribution to the layer-wise subspace, and a residual score in Section 4.3, which captures reconstruction error with respect to the dominant structure. In Section 4.4, we combine these measures to define an influence and construct the corresponding influential activation set. Finally, in Section 4.5, we demonstrate how the influential activation set can be incorporated into static quantization frameworks, using SmoothQuant (Xiao et al., 2023) to address activation outliers and AWQ (Lin et al., 2024) to identify important weights based on activation statistics.

4.1 PROBLEM FORMULATION

To assess how individual activations influence the dominant gradient structure, we use a low-rank reconstruction of activation gradients as the fitted reference. Adapting influence diagnostics from regression to this low-rank setting (Cook, 2000), we construct an influence surrogate from the shared structure, avoiding the prohibitive cost of refitting it for each leave-one-out assessment. Accordingly, we characterize each activation by two complementary aspects: its deviation from the low-rank reconstruction and its contribution to the fitted low-rank subspace. These aspects respectively parallel the residual and leverage in regression diagnostics. Building on this analogy, we formulate a layer-wise influence measure that identifies activations exhibiting both large reconstruction deviation and high subspace influence.

Influence. Consider a matrix $\mathbf{G} \in \mathbb{R}^{N\times C}$ (e.g., gradients) at a given layer, where $\mathbf{G}_i \in \mathbb{R}^C$ denotes the i-th vector. Let $\hat{\mathbf{G}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$ denote the rank-k approximation of $\mathbf{G}$, where $\hat{\mathbf{U}} \in \mathbb{R}^{N\times k}$, $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{k\times k}$, and $\hat{\mathbf{V}} \in \mathbb{R}^{C\times k}$ are obtained from the truncated singular value decomposition of $\mathbf{G}$. We then define an influence estimator $f : (\mathbb{R}^C, \mathbb{R}^C, \mathbb{R}^k) \to \mathbb{R}$ that assigns an influence $\mathbf{c}_i$ to each gradient vector $\mathbf{G}_i$ as

$$\mathbf{c}_i = f(\mathbf{G}_i, \hat{\mathbf{G}}_i, \hat{\mathbf{U}}_i) = f_r(\mathbf{G}_i, \hat{\mathbf{G}}_i)\, f_l(\hat{\mathbf{U}}_i), \tag{3}$$



(a) **Influential Outliers** (b) **Structural Sensitivity** (c) **Activation Histogram**

Figure 4: **Characteristics of Our Influential Activation Set.** We define $\tau \in [0, 1]$ as the percentile rank of activation influence and select the top 5% ($\tau \in [0.95, 1]$) as influential. (a) **Influential Outliers**: removing influential activations substantially changes the fitted gradient-space trend. (b) **Structural Sensitivity**: removing the top 5% increases low-rank reconstruction error far more than equally sized median ($\tau \in [0.475, 0.525]$) or bottom ($\tau \in [0, 0.05]$) subsets. (c) **Activation Histogram**: influential activations exhibit structural importance beyond magnitude.

where $f_r : (\mathbb{R}^C, \mathbb{R}^C) \to \mathbb{R}$ denotes the residual (Section 4.3) quantifying the reconstruction discrepancy between $\mathbf{G}_i$ and $\hat{\mathbf{G}}_i$. $f_l : \mathbb{R}^k \to \mathbb{R}$ denotes the leverage (Section 4.2) measuring the contribution of $\hat{\mathbf{G}}_i$ to the low-rank subspace. The former captures how far an activation departs from the low-rank structure, whereas the latter captures how strongly it contributes to the low-rank subspace, mirroring the residual–leverage decomposition.

4.2 Leverage Score

Let $\mathbf{G} \in \mathbb{R}^{N \times C}$ denote the gradient matrix, where N and C are the numbers of activations and channels, respectively, and let $\hat{\mathbf{G}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$ denote its rank-k approximation, with $\hat{\mathbf{U}} \in \mathbb{R}^{N \times k}$. Following the leverage component of Cook's distance (Cook, 2000), we quantify the relative importance of each activation to the fitted low-rank structure as

$$f_l(\hat{\mathbf{U}}_i) = \frac{v_i}{k(1 - v_i)}, \quad \text{where } v_i = \mathbf{H}_{ii} = (\hat{\mathbf{U}}\hat{\mathbf{U}}^\top)_{ii} = \|\hat{\mathbf{U}}_i\|_2^2. \tag{4}$$

Here, $\mathbf{H} = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top$ is the projection matrix onto the rank-k left-singular subspace induced by the gradients. Accordingly, v_i measures how strongly the i-th activation is represented in this shared low-rank structure. Since $0 \le v_i \le 1$ and $\sum_i v_i = k$, a larger v_i indicates a greater relative contribution to the fitted structure. Following Cook's distance, the factor $1/(1 - v_i)$ further emphasizes activations with high leverage, while the normalization by k accounts for the dimension of the fitted subspace.

4.3 Residual Score

Complementary to the leverage score, we quantify how strongly each activation deviates from the fitted low-rank structure. Analogous to the studentized residual in Cook's distance, we normalize the reconstruction error by the pooled residual scale s^2 and the leverage-dependent factor $(1 - v_i)$:

$$f_r(\mathbf{G}_i, \hat{\mathbf{G}}_i) = \frac{\left\|\mathbf{G}_i - \hat{\mathbf{G}}_i\right\|_2^2}{Cs^2(1 - v_i)}, \quad \text{where } s^2 = \frac{\left\|\mathbf{G} - \hat{\mathbf{G}}\right\|_F^2}{C(N - k)}. \tag{5}$$

The quantity s^2 represents the pooled reconstruction-residual energy, normalized by the number of channels C and the residual subspace dimension $N - k$, and thus serves as a low-rank analogue of the residual variance estimate in linear regression. The factor $(1 - v_i)$ further adjusts the reconstruction error according to the activation's leverage, following the studentization used in Cook's distance.

4.4 Influential Activation Set

Activations with large $\mathbf{c}_i$ exhibit high sensitivity to the task objective, as perturbing these activations is more likely to induce substantial changes in the loss (Selvaraju et al., 2017). We transform the influence score $\mathbf{c}_i$ to percentile rank $[0, 1]$ such that $\mathbf{c}_i > 0.95$ corresponds to the top 5% most influential activations. We then define the influential activation set as $\mathcal{I} = \{\, i \mid \mathbf{c}_i > \tau \,\}$.

The resulting set $\mathcal{I}$ prioritizes activations that are most relevant to the task objective, thereby preserving the low-rank statistics of task-relevant gradients, as illustrated in Figure 4. Combined with the preceding structural influence analysis, this selection enables us to isolate task-relevant influential activations, including informative outliers, while excluding incidental tail activations.

4.5 Example Application: Static Quantization

We apply the influential activation set to static quantization through SmoothQuant (Xiao et al., 2023) and AWQ (Lin et al., 2024). For SmoothQuant, we compute the smoothing scale from the influential set rather than the full activation range, while for AWQ, we use it to determine channel-wise scaling factors and weight quantization parameters during calibration.

With SmoothQuant. We compute a smoothing scale $\mathbf{s} \in \mathbb{R}^C$ via $\mathcal{I}$, migrate it into the weights, and then optimize the quantization parameters by minimizing the reconstruction error. Specifically, we compute $\mathbf{s}$ from the weights $\mathbf{W} \in \mathbb{R}^{C \times C_{\text{out}}}$ and the influential activation set $\mathcal{I}$ as

$$\mathbf{s} = \frac{\text{pow}(\mathbf{m}^{\mathcal{I}}, \alpha)}{\text{pow}(\mathbf{m}^{\mathbf{W}}, \beta)}, \qquad m_b^{\mathcal{I}} = \max_{j \in \mathcal{D}} \left(\max_{i \in \mathcal{I}} |\mathbf{X}_{i,b}^j| \right), \qquad m_b^{\mathbf{W}} = \max_k |\mathbf{W}_{b,k}|. \tag{6}$$

Here, $\mathbf{m}^{\mathcal{I}}, \mathbf{m}^{\mathbf{W}} \in \mathbb{R}^C$ denote the channel-wise maximum vectors of the influential activations and weights, respectively. $\mathcal{D} = \{1, \ldots, J\}$ is a calibration data index set. α and β control the degree of scale migration between activations and weights. Using $\mathbf{s}$, we obtain the smoothed activations $\tilde{\mathbf{X}}$ and the migrated weights $\tilde{\mathbf{W}}$, and then optimize the quantization parameters $\theta = \{\theta_{\mathbf{X}}, \theta_{\mathbf{W}}\}$ consisting of activations $\theta_{\mathbf{X}}$ and weights $\theta_{\mathbf{W}}$ by minimizing

$$\theta^* = \arg\min_{\theta} \left\| \tilde{\mathbf{X}}\tilde{\mathbf{W}} - Q(\tilde{\mathbf{X}}; \theta_{\mathbf{X}}) Q(\tilde{\mathbf{W}}; \theta_{\mathbf{W}}) \right\|_F^2, \tag{7}$$

where $Q(\cdot; \theta)$ denotes the Q-DQ operation of fake quantization (Nagel et al., 2021). The subsequent reconstruction-error minimization is still performed over the smoothed activations and migrated weights, preserving compatibility with conventional static quantization while shaping a quantization-friendly activation distribution.

With AWQ. We further apply the influential activation set to activation-aware weight quantization (AWQ) (Lin et al., 2024). In conventional AWQ, channel-wise scaling factors are determined from calibration activations to preserve activation-sensitive weight channels after quantization. Instead of using the full calibration activations, we use the influential activation set $\mathcal{I}$ to determine the scaling factors during AWQ calibration. Using the channel-wise maximum $m_b^{\mathcal{I}}$ defined in Equation (6), we parameterize the AWQ scaling factor as $\mathbf{s} = \text{pow}\left(\mathbf{m}^{\mathcal{I}}, \alpha\right)$. Then the channel-wise scaling corresponds to an equivalent transformation of the linear operation,

$$\mathbf{X}_i^j \mathbf{W} = \left(\mathbf{X}_i^j \, \text{diag}(\mathbf{s})^{-1} \right) \left(\text{diag}(\mathbf{s}) \mathbf{W} \right), \tag{8}$$

which redistributes the quantization difficulty across channels. We then select the AWQ scaling and weight quantization parameters by grid search using the influential activations:

$$(\alpha^\star, \theta_{\mathbf{W}}^\star) = \arg\min_{\alpha, \theta_{\mathbf{W}}} \sum_{j \in \mathcal{D}} \sum_{i \in \mathcal{I}} \left\| \mathbf{X}_{i,:}^j \mathbf{W} - \left(\mathbf{X}_{i,:}^j \, \text{diag}(\mathbf{s})^{-1} \right) Q(\text{diag}(\mathbf{s}) \mathbf{W}; \theta_{\mathbf{W}}) \right\|_2^2. \tag{9}$$

The minimization in Equation (9) is performed by the discrete grid search of AWQ. $\mathcal{I}$ is used only during calibration to guide the selection of the channel-wise scaling and weight quantization parameters. Once these parameters are determined, the resulting quantized weights are shared across all activations at inference time.

5 Experiments

In this section, we validate whether our influence-based selection identifies structurally important activations and improves low-bit quantization across 3D detector architectures and sensing modalities. We first assess its effectiveness for calibration by comparing low-bit detection performance against prior static PTQ methods and alternative activation-sampling criteria in Section 5.1. We then analyze

Table 1: **Quantitative Comparison to the Prior Static Quantization Approaches.**

Detector	Bit	Method	Detection Errors (↓)					Performance (↑)	
			mATE	mASE	mAOE	mAVE	mAAE	mAP	NDS
BEVDet	FP32	-	0.715	0.280	0.598	0.779	0.230	30.9	39.4
	W8A8	SmoothQuant	0.720	0.282	**0.606**	0.788	0.233	30.2	38.8
		InlierQ	0.746	0.307	0.616	0.916	0.234	29.7	36.7
		Ours-Smooth	0.716	**0.281**	0.609	**0.781**	**0.232**	**30.7**	**39.2**
	W4A16	AWQ	0.728	0.285	**0.627**	0.847	0.232	28.9	37.3
		Ours-AWQ	**0.720**	**0.282**	0.629	**0.813**	**0.225**	**29.7**	**38.1**
	W4A8	SmoothQuant	0.721	0.283	0.627	0.817	**0.226**	28.4	37.4
		InlierQ	0.761	0.310	0.631	0.919	0.234	27.4	35.2
		Ours-Smooth	**0.713**	**0.279**	**0.624**	**0.814**	0.227	**28.9**	**37.9**
	W4A4	SmoothQuant	0.772	0.291	0.629	**0.817**	0.243	25.7	35.3
		InlierQ	0.753	0.303	0.687	1.043	0.253	23.6	31.8
		Ours-Smooth	**0.752**	**0.286**	**0.620**	0.838	**0.239**	**27.1**	**36.2**
CenterPoint	FP32	-	0.291	0.255	0.336	0.287	0.187	55.8	64.4
	W8A8	SmoothQuant	0.292	**0.255**	0.336	**0.290**	0.187	**55.8**	**64.3**
		LiDAR-PTQ	0.299	0.258	0.353	0.294	**0.182**	54.4	63.4
		InlierQ	0.297	0.257	0.346	0.293	**0.182**	54.7	63.6
		Ours-Smooth	**0.291**	**0.255**	**0.335**	0.291	0.184	**55.8**	**64.3**
	W4A16	AWQ	0.300	**0.256**	0.369	**0.300**	0.190	54.3	63.0
		Ours-AWQ	**0.298**	**0.256**	**0.353**	**0.300**	**0.187**	**55.2**	**63.7**
	W4A8	SmoothQuant	**0.297**	**0.255**	0.356	**0.298**	0.193	53.9	62.9
		LiDAR-PTQ	0.309	0.260	0.390	0.309	0.187	50.7	60.8
		InlierQ	0.304	0.259	0.376	0.304	0.192	52.2	61.7
		Ours-Smooth	0.302	**0.255**	**0.345**	0.317	**0.185**	**54.2**	**63.1**
	W4A4	SmoothQuant	0.336	0.267	0.438	0.388	0.207	41.6	54.4
		LiDAR-PTQ	0.326	0.272	0.431	0.366	0.192	39.5	53.9
		InlierQ	0.319	0.261	0.401	0.343	**0.189**	46.6	58.1
		Ours-Smooth	**0.303**	**0.259**	**0.353**	**0.315**	0.190	**52.2**	**61.8**

the contributions of individual components and the effect of selection strictness in Section 5.2. Further implementation details are provided in Section A, covering the experimental setup, calibration and quantization configurations, and our gradient computation. SVD Rank analysis for influence is examined in Section B, and calibration time is reported in Section C.

5.1 Quantitative Comparison

Static Quantization. In Table 1, we compare our method with representative quantization baselines for 3D object detection across camera- and LiDAR-based detectors under W8A8, W4A16, W4A8, and W4A4 settings. To isolate the effect of our activation selection strategy, we consider static PTQ baselines including LiDAR-PTQ (Zhou et al., 2024), InlierQ (Kim et al., 2026), AWQ (Lin et al., 2024), and SmoothQuant (Xiao et al., 2023). All methods use per-channel symmetric quantization for weights and per-tensor affine quantization for activations. We consider FP32/FP16, INT8, and INT4 precision; Wn_1An_2 denotes n_1-bit weight and n_2-bit activation precision. For both BEVDet and CenterPoint, W8A8 closely preserves FP32 performance: BEVDet retains 99.4% of FP32 mAP and 99.5% of NDS, while CenterPoint retains 100.0% and 99.8%, respectively. The magnitude of the improvement varies across detectors and precision settings, with larger gains generally observed under low-bit activation quantization. At W4A4, our method improves the strongest baseline from 25.7/35.3 to 27.1/36.2 mAP/NDS on BEVDet and from 46.6/58.1 to 52.2/61.8 on CenterPoint, corresponding to relative mAP gains of 5.4% and 12.0%. This trend is consistent with the sensitivity of coarse activation quantization to long-tailed outliers that expand the quantization range and reduce effective resolution. Our influence-based criterion also benefits weight-only quantization. At W4A16, where activations remain in FP16, using our influential activation set for AWQ calibration yields relative mAP/NDS gains of 2.8%/2.1% on BEVDet and 1.7%/1.1% on CenterPoint. Together, these results support the effectiveness of influence-based calibration across the evaluated detectors and precision settings, with the largest benefits arising under aggressive activation quantization.

Activation Sampling. In Table 2, we compare activation selection criteria using the top-k% activations for calibration. Random sampling provides a score-free baseline, Feature Norm ranks activations by magnitude, and Fisher Diagonal uses a diagonal Fisher approximation. Magnitude alone is unreliable: under W4A4 on BEVDet, Feature Norm reduces mAP/NDS from 24.7/33.7 with random sampling to 19.9/29.3, whereas Fisher Diagonal incorporates task-related gradients and reaches 25.5/33.9. By additionally accounting for structural importance and residual discrepancy, our influential activation set improves mAP over Fisher Diagonal in every evaluated detector–precision

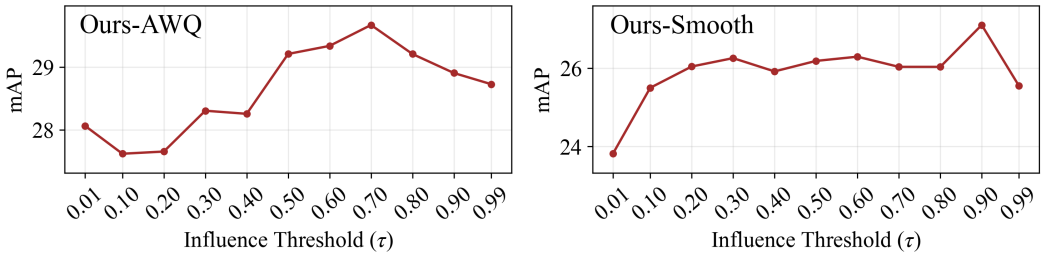
Table 2: **Quantitative Comparison of Activation Samplings.**

(a) Ours-AWQ, W4A16 Quantization

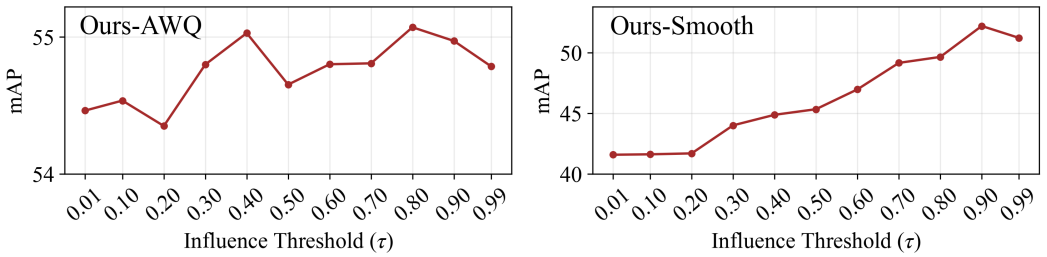
Detector	Method	Detection Errors (↓)					Performance (↑)	
		mATE	mASE	mAOE	mAVE	mAAE	mAP	NDS
BEVDet	Random	0.738	0.286	0.644	0.854	0.218	27.8	36.5
	Feature Norm	0.731	0.285	0.641	0.827	**0.224**	28.3	37.1
	Fisher Diagonal	**0.720**	**0.282**	**0.626**	0.816	0.227	29.5	**38.1**
	Ours	**0.720**	**0.282**	0.629	**0.813**	0.225	**29.7**	**38.1**
CenterPoint	Random	**0.298**	0.256	0.363	0.302	0.187	54.4	63.1
	Feature Norm	0.299	**0.255**	0.358	**0.300**	**0.186**	54.6	63.3
	Fisher Diagonal	0.300	**0.255**	0.358	0.303	0.188	54.8	63.4
	Ours	**0.298**	0.256	**0.353**	**0.300**	0.187	**55.2**	**63.7**

(b) Ours-Smooth, W4A4 Quantization

Detector	Method	Detection Errors (↓)					Performance (↑)	
		mATE	mASE	mAOE	mAVE	mAAE	mAP	NDS
BEVDet	Random	0.800	0.302	0.660	0.867	0.240	24.7	33.7
	Feature Norm	0.771	0.329	0.691	1.034	0.278	19.9	29.3
	Fisher Diagonal	0.808	**0.286**	0.649	0.895	0.247	25.5	33.9
	Ours	**0.752**	**0.286**	**0.620**	**0.838**	**0.239**	**27.1**	**36.2**
CenterPoint	Random	0.427	0.372	0.882	0.976	0.310	33.6	37.1
	Feature Norm	0.328	0.266	0.412	0.376	0.208	41.9	55.1
	Fisher Diagonal	0.306	**0.256**	**0.345**	0.331	0.186	51.2	61.3
	Ours	**0.303**	0.259	0.353	**0.315**	0.190	**52.2**	**61.8**



(a) **BEVDet**



(b) **CenterPoint**

Figure 5: **Effect of influence thresholds.** Ours-AWQ is evaluated under W4A16, while Ours-Smooth is evaluated under W4A4. Increasing the influence threshold generally improves task selectivity and mAP, although the optimal threshold varies across detectors and quantization methods.

combination while matching or improving NDS. Under W4A16, our method achieves 29.7/38.1 mAP/NDS on BEVDet and 55.2/63.7 on CenterPoint, with mAP gains of 0.2 and 0.4 points over Fisher Diagonal. Gains are larger under W4A4: mAP/NDS increases from 25.5/33.9 to 27.1/36.2 on BEVDet and from 51.2/61.3 to 52.2/61.8 on CenterPoint, corresponding to relative mAP improvements of 6.3% and 2.0%. Although their magnitude varies across detectors and precision settings, these consistent gains support the additional calibration value of structural influence beyond activation magnitude or local gradient information.

Decision Strictness. In Figure 5, we vary the influence threshold τ, where larger values restrict calibration to higher-influence activations. Increasing τ generally improves low-bit performance, with Ours-Smooth on CenterPoint increasing from approximately 41.7 mAP at $\tau = 0.01$ to 51.2 at $\tau = 0.99$. On BEVDet, Ours-AWQ and Ours-Smooth peak at $\tau = 0.70$ and $\tau = 0.90$, respectively, before declining at the strictest threshold, suggesting that overly narrow selection can discard useful

Table 3: **Ablation Study.** We analyze activation selection (Influential and Non-influential), and score components (Residual and Leverage) in our formulation. The gray-shaded rows denote the default setting, from which each factor is varied for ablation.

(a) Ours-AWQ, W4A16 Quantization

Method	Selection		mAP	Score		mAP
	Influential	Non-influential		Residual	Leverage	
BEVDet	-	✓	28.1	-	✓	29.2
	✓	✓	28.9	✓	-	29.3
	✓	-	**29.7**	✓	✓	**29.7**
CenterPoint	-	✓	41.9	-	✓	51.1
	✓	✓	54.3	✓	-	54.8
	✓	-	**55.2**	✓	✓	**55.2**

(b) Ours-Smooth, W4A4 Quantization

Method	Selection		mAP	Score		mAP
	Influential	Non-influential		Residual	Leverage	
BEVDet	-	✓	21.1	-	✓	26.9
	✓	✓	25.7	✓	-	26.1
	✓	-	**27.1**	✓	✓	**27.1**
CenterPoint	-	✓	41.1	-	✓	51.0
	✓	✓	41.6	✓	-	50.7
	✓	-	**52.2**	✓	✓	**52.2**

activations. In contrast, Ours-AWQ on CenterPoint remains relatively stable, with mAP ranging from approximately 54.3 to 55.2. Notably, for $\tau > 0.5$, both methods mostly outperform their respective baselines across the evaluated settings. Overall, emphasizing high-influence activations is beneficial, while the optimal strictness depends on the detector.

5.2 Ablation Study

We ablate the activation selection strategy and individual score components in Table 3 to examine their contributions to quantization performance. The *Selection* analysis compares calibration using only influential activations, only non-influential activations, or their union. Across all settings, the influential-only set consistently achieves the highest mAP. Under W4A16, it reaches 29.7 mAP on BEVDet and 55.2 on CenterPoint, compared with 28.1 and 41.9 using only non-influential activations, and 28.9 and 54.3 when both sets are used. Under W4A4, it achieves 27.1 mAP on BEVDet and 52.2 on CenterPoint, compared with 21.1 and 41.1 for non-influential only selection, and 25.7 and 41.6 when both sets are included. These results suggest that using influential activations to address informative outliers provides a cleaner calibration signal. The *Score* analysis evaluates the residual term f_r and leverage term f_l in Equation (3). Combining both consistently yields the highest mAP across detectors and quantization settings. Under W4A16, the combined score achieves 29.7 mAP on BEVDet and 55.2 on CenterPoint, compared with 29.3 and 54.8 using residual alone, and 29.2 and 51.1 using leverage alone. Under W4A4, it reaches 27.1 mAP on BEVDet and 52.2 on CenterPoint, compared with 26.1 and 50.7 using residual alone, and 26.9 and 51.0 using leverage alone. These gains indicate that residual discrepancy and leverage capture complementary aspects of activation influence, making their combination an effective criterion for identifying informative outliers.

6 Conclusion

In this paper, we introduced an influence diagnostics perspective for identifying informative outliers in low-bit 3D object detection. We characterize activation influence with respect to a task-aware low-rank gradient structure by jointly considering residual and leverage, allowing us to distinguish structurally important activations from incidental or task-irrelevant activations. The resulting influential activation set guides existing static quantization methods, including SmoothQuant and AWQ, toward calibration statistics that better preserve task-relevant information. Experiments on both camera- and LiDAR-based 3D detectors show improvements across quantization settings, with larger gains under aggressive 4-bit quantization regimes. These results suggest that activation influence provides a practical and principled criterion for identifying informative outliers and improving low-bit 3D object detection across camera and LiDAR modalities.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing for clarity, readability, grammar, and style, as well as to check potential formatting or style issues with respect to the submission guidelines. All AI-assisted suggestions were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed influence-based activation selection and its integration with static quantization in Section 4. The experimental protocol and comparisons are described in Section 5. Additional implementation details, including the detector configurations, dataset and calibration setup, gradient computation, quantization settings, and optimization protocol, are provided in Appendix A. We further report sensitivity to the SVD rank in Appendix B and calibration cost in Appendix C. These details are provided to facilitate reproduction of our experimental results, and the code will be made publicly available upon publication.

REFERENCES

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 *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020. 12

Peng Chen, Jing Liu, Bohan Zhuang, Mingkui Tan, and Chunhua Shen. Aqd: Towards accurate fully-quantized object detection. *arXiv preprint arXiv:2007.06919*, 2020. 1

R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 42(1):65–68, 2000. 2, 3, 4, 5

R Dennis Cook and Sanford Weisberg. Residuals and influence in regression. 1982. 2, 3

Frank E Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1): 1–21, 1969. 2, 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1, 12

David C Hoaglin and Roy E Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978. 2, 3

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 12

Minsu Kim, Dongyeun Lee, Jaemyung Yu, Giseop Kim, Junmo Kim, et al. Inlier-centric post-training quantization for object detection models. In *International Conference on Learning Representations*, volume 2026, pp. 141556–141567, 2026. 1, 2, 7, 12

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank component for 4-bit diffusion models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 97831–97858, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f34f0630c33be15b8c89426bb8056798-Paper-Conference.pdf. 1

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH. 12

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100, 2024. 1, 2, 4, 6, 7

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021. 6

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020. 2

Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2247–2256. IEEE, 2020. 1

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. 2, 5

Lukas Stäcker, Juncong Fei, Philipp Heidenreich, Frank Bonarens, Jason Rambach, Didier Stricker, and Christoph Stiller. Deployment of deep neural networks for object detection on edge ai devices with runtime optimization. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1015–1022. IEEE, 2021. 1

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

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023. 1, 2, 4, 6, 7

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021. 12

Yifan Zhang, Zhen Dong, Huanrui Yang, Ming Lu, Cheng-Ching Tseng, Yuan Du, Kurt Keutzer, Li Du, and Shanghang Zhang. Qd-bev: Quantization-aware view-guided distillation for multi-view 3d object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3802–3812. IEEE, 2023. 1

Sifan Zhou, Liang Li, Xinyu Zhang, Bo Zhang, Shipeng Bai, Miao Sun, Ziyu Zhao, Xiaobo Lu, and Xiangxiang Chu. Lidar-ptq: Post-training quantization for point cloud 3d object detection. In *International Conference on Learning Representations*, volume 2024, pp. 51272–51286, 2024. 1, 2, 7

Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499. IEEE, 2018. 12

A Implementation Details

Base Detectors. To assess the generality of our method across architectures and sensing modalities, we evaluate our approach on 3D detectors with different input modalities and backbone architectures: CenterPoint (Yin et al., 2021) with a VoxelNet (Zhou & Tuzel, 2018) point-cloud backbone, and BEVDet (Huang et al., 2021) with a ResNet-50 (He et al., 2016) image backbone.

Dataset. We conduct experiments on the nuScenes benchmark (Caesar et al., 2020), which provides LiDAR point clouds and multi-view camera inputs across 10 object categories. For calibration, we sample 64 examples from a driving sequence in the training split with data augmentation disabled. We report evaluation results on the full validation split.

Experimental Configuration. For a fair comparison, our method and all baselines use the same detector architectures, layer-wise quantization settings, and training configurations within the block-wise quantization framework of BRECQ (Li et al., 2021), which optimizes quantization parameters independently for each block. Task-related gradients are obtained by backpropagating the Heatmap Top-K NLL loss (Kim et al., 2026). For BEVDet, the first and last convolution layers are quantized to INT8, whereas those of CenterPoint remain in FP16 for stable detection performance. For Ours-AWQ and Ours-Smooth, we fix the influence thresholds to $\tau = 0.7$ and $\tau = 0.9$, respectively, and set the SVD rank to 16 for BEVDet and 32 for CenterPoint. These hyperparameters are tuned using only a calibration set sampled from the training split. Calibration takes approximately 30 minutes for each detector on a single NVIDIA RTX 4090 GPU. In Figure 3, we visualize influence score maps from a CenterPoint decoder layer by projecting them onto the bird's-eye-view LiDAR point cloud. To facilitate reproducibility, we will make our code publicly available upon publication.

B Rank Analysis for Influence Estimation

Table 4: **Effect of the SVD rank k on influence estimation.** We report mAP across fixed ranks $k \in \{4, 16, 32, 64\}$ for the low-rank structure used to compute activation influence.

(a) **Ours-Smooth**

Model	Rank	mAP
CenterPoint	4	50.5
CenterPoint	16	50.7
CenterPoint	32	**52.2**
CenterPoint	64	51.2
BEVDet	4	26.9
BEVDet	16	**27.1**
BEVDet	32	26.7
BEVDet	64	25.7

(b) **Ours-AWQ**

Model	Rank	mAP
CenterPoint	4	54.9
CenterPoint	16	55.0
CenterPoint	32	**55.2**
CenterPoint	64	55.1
BEVDet	4	29.6
BEVDet	16	**29.7**
BEVDet	32	29.2
BEVDet	64	28.9

Our influence formulation relies on a low-rank approximation as the fitted structure with respect to which influential observations are identified. Therefore, the rank k should be sufficiently large to capture the dominant task-relevant structure, while remaining sufficiently low-rank so that atypical directions are not simply absorbed into the fitted subspace. We empirically examine this trade-off through quantization performance in Table 4.

For Ours-Smooth, CenterPoint performs best at $k = 32$, reaching 52.2 mAP, while performance decreases to 51.2 at $k = 64$. BEVDet instead performs best at a smaller rank, reaching 27.1 mAP at $k = 16$, and gradually degrades as the rank increases further. A similar pattern is observed for Ours-AWQ. CenterPoint remains relatively stable across ranks, with its best performance of 55.2 mAP at $k = 32$, whereas BEVDet achieves its best performance of 29.7 mAP at $k = 16$ and degrades at larger ranks. Overall, the two quantization schemes exhibit similar rank-dependent behavior for each detector: CenterPoint favors an intermediate rank, while BEVDet performs better with a smaller rank. These results suggest that maintaining an appropriately low-rank fitted structure is important for effective influence estimation, and that simply increasing the rank does not necessarily improve quantization performance.

C CALIBRATION TIME

Table 5: Breakdown of calibration cost compared with InlierQ. The gradient-collection cost is nearly identical between the two methods, while our additional overhead mainly comes from SVD and block optimization. All reported costs correspond to one-time calibration and do not affect inference.

	SVD		Gradient		Optimization		Overall	
Method	Time	Mem.	Time	Mem.	Time	Mem.	Total Time	Peak Mem.
InlierQ	–	–	6.7 min	11.5 GiB	20.1 min	0.5 GiB	26.8 min	11.5 GiB
Ours	2.0 min	0.4 GiB	6.5 min	11.6 GiB	21.8 min	2.4 GiB	30.3 min	11.6 GiB

We compare the calibration cost of Ours-Smooth with InlierQ on CenterPoint in Table 5. Gradient collection incurs nearly identical cost for both methods, requiring 6.5 minutes and 11.6 GiB for ours compared with 6.7 minutes and 11.5 GiB for InlierQ. Our additional overhead mainly comes from SVD-based influence estimation and block-wise optimization: SVD requires 2.0 minutes and 0.4 GiB, while optimization takes 21.8 minutes compared with 20.1 minutes for InlierQ. Overall, our calibration takes 30.3 minutes, only 3.5 minutes (13.1%) longer than InlierQ, while peak memory increases by just 0.1 GiB (0.9%), from 11.5 to 11.6 GiB. Importantly, these costs are incurred only once during calibration and introduce no additional overhead at inference time.