LMF-3DDET: EFFICIENT LIDAR–CAMERA DETECTION VIA RESPONSE-GUIDED SAMPLING AND LINEAR FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

LiDAR–camera fusion enhances 3D object detection by combining geometric and semantic information, but redundant inputs and costly cross-modal interactions hinder real-time deployment. Aggressive input reduction can also discard informative observations, making it difficult to improve efficiency while maintaining detection accuracy. To address this challenge, we propose LMF-3DDet, a lightweight framework that jointly addresses input selection, feature fusion, and knowledge transfer. First, response-guided sampling uses image detection heatmaps and a geometric dilation margin to select informative LiDAR observations while retaining nearby geometric context. Second, kernelized linear cross-attention fuses the selected point tokens with image features without constructing a dense pairwise attention matrix, reducing the cost of cross-modal interaction. Third, uncertainty-aware distillation guides the lightweight student using a full-input teacher, emphasizing difficult examples without adding teacher computation at inference. Experiments on nuScenes and Waymo demonstrate a favorable accuracy–efficiency trade-off with real-time throughput, while ablation studies validate the benefits of the sampling, fusion, and distillation components.

1 INTRODUCTION

LiDAR and cameras provide complementary information for autonomous-driving perception. LiDAR supplies metric geometry, while images supply appearance and semantic context. Their fusion introduces a resource-allocation problem: processing all available measurements can be expensive, yet discarding measurements can remove the few returns that represent a distant or partially visible object. An efficient detector should allocate computation to informative observations while making the resulting coverage loss measurable.

Existing architectures address this problem at different stages. TransFusion uses object queries and image attention for fusion (Bai et al., 2022); BEVFusion aggregates modalities in a shared bird's-eye-view (BEV) representation (Liu et al., 2023); CMT performs cross-modal token interaction (Yan et al., 2023); and UniTR shares a transformer backbone across modalities (Wang et al., 2023). These designs show that efficient fusion is an established research direction. The question considered here is more specific: how should image-guided input reduction, inexpensive token interaction, and training-only supervision be combined in one lightweight student?

LMF-3DDet addresses three stages of this pipeline. Response-guided sampling (RGS) selects LiDAR points using projected image heatmaps and a geometric dilation margin. Kernelized cross-attention then fuses the retained LiDAR tokens with image tokens without materializing their pairwise attention matrix. Finally, a frozen teacher supervises student predictions during training. These components act on different costs: RGS changes the query count, linear attention changes interaction scaling, and distillation changes the training signal while leaving the deployed architecture fixed.

Our contributions are threefold. First, we introduce a response-guided sampling strategy that combines image semantics with a 3D dilation margin to reduce the LiDAR token budget. Second, we integrate this sampler with established kernelized cross-attention and uncertainty-aware teacher supervision in a lightweight detection framework. The contribution is the coordinated system design, rather than a new linear-attention identity. Third, experiments on nuScenes and Waymo demonstrate

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.

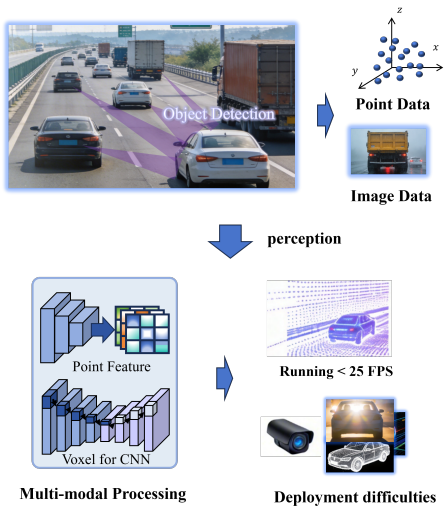


Figure 1: LiDAR–camera perception and the motivation for efficient cross-modal feature processing.

high detection accuracy at approximately 45 FPS, supported by sampling, attention, and distillation studies and a module-level latency breakdown.

2 Related Work

Fusion and efficient 3D detection. CenterPoint represents objects by BEV centers and regresses their geometric attributes (Yin et al., 2021). TransFusion combines a query-based detector with LiDAR–camera interaction (Bai et al., 2022), whereas BEVFusion uses a shared BEV feature space (Liu et al., 2023). CMT and UniTR further demonstrate direct token fusion and efficient shared representations, respectively (Yan et al., 2023; Wang et al., 2023). GAFusion uses LiDAR guidance and adaptive fusion (Li et al., 2024); EVT improves view transformation through adaptive sampling and projection (Lee et al., 2025). These approaches motivate our focus on reducing the input token budget before cross-modal interaction, while retaining a lightweight deployed student.

Linear attention. Katharopoulos et al. (2020) express attention through positive kernel feature maps and reassociate matrix products to obtain linear token scaling. BAFusion already applies focused linear attention to LiDAR–camera fusion (Liu et al., 2024). We adopt the ELU-plus-one kernel for one-way image-to-LiDAR information transfer. Our contribution lies in combining this established operator with response-guided query reduction and training-only distillation, rather than introducing a new attention formulation.

Distillation. Knowledge transfer has also been studied for compact 3D detectors, including structured feature transfer in PointDistiller (Zhang et al., 2023). Here, the student and teacher can have different output structures, so prediction matching must precede distillation. The increasing uncertainty weight is a hard-example heuristic rather than a calibrated teacher-confidence weight. The comparison with uniform distillation in Section 4 evaluates the resulting accuracy gain.

3 Method

3.1 Inputs and Student Pipeline

Let $P = \{p_i\}_{i=1}^{N}$ be a LiDAR scan in the reference sensor frame, with additional point attributes, and let (I, T, K) denote a synchronized single-camera image, the LiDAR-to-camera transform, and camera intrinsics. The experiments use one camera view. Image resizing and cropping are incorporated into the projection. A ResNet-18 image backbone supplies image features F; the RGS heatmap prediction branch supplies classwise center responses A and remains frozen during student training.

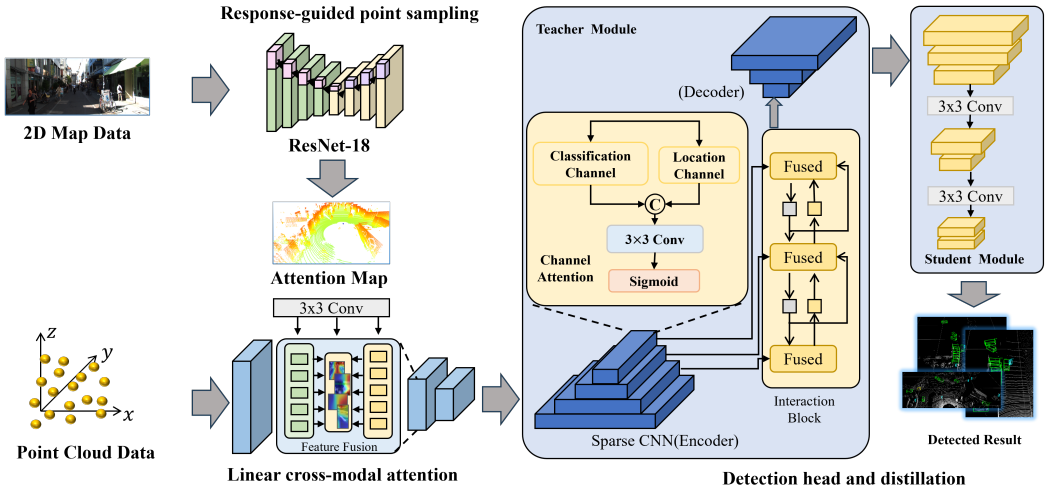


Figure 2: Overview of LMF-3DDet. The deployed student uses response-guided sampling, cross-modal fusion, and a lightweight detection head. The teacher branch supplies supervision during training and is removed at inference. Encoder and interaction blocks are illustrated at the architectural level.

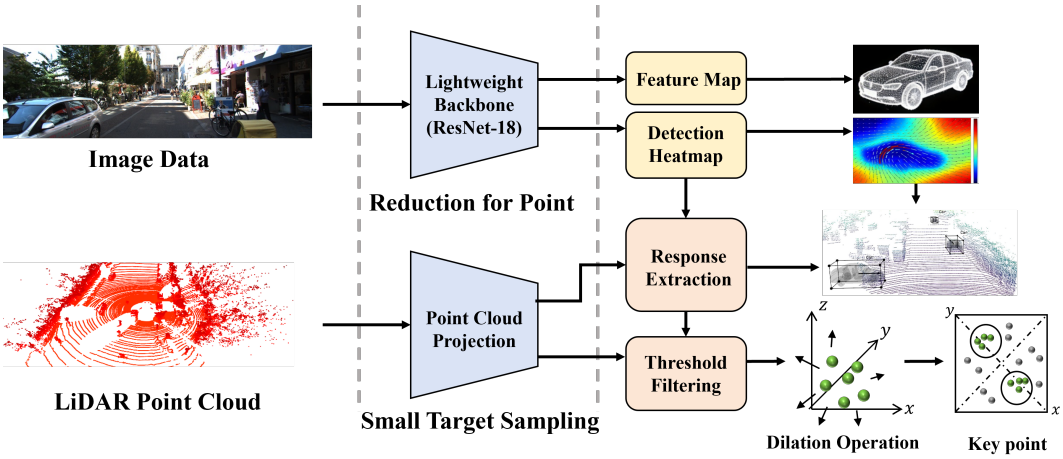


Figure 3: Response-guided sampling. Image responses select LiDAR observations, and a 3D dilation margin retains neighboring points around the selected seeds.

RGS selects original point indices, the point encoder produces M query tokens with width D, and the image features produce $L = H'W'$ key/value tokens. Raw point count and encoded token count are distinct when voxelization merges points.

The retained LiDAR features are augmented with position embeddings and serve as queries for cross-modal fusion. The fusion width is $D = 256$. Fused features are projected to a BEV grid by height-wise max pooling and processed by a lightweight convolutional detection head. The head predicts object centers, box geometry, and velocity. This organization separates the cost of selecting input evidence from the cost of fusing it, allowing the number of LiDAR queries to be reduced before attention.

3.2 Response-guided sampling

For a calibrated image view, a LiDAR point p_i is projected to the response map using the camera transform, intrinsics, and image preprocessing transform. Only projections with positive camera

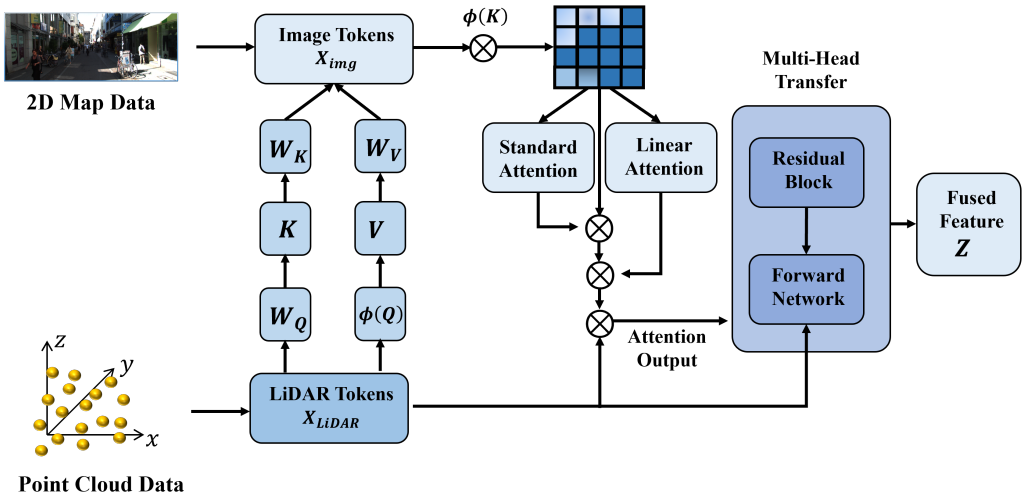


Figure 4: Cross-modal attention. Standard and linear attention denote alternative interaction operators. The normalized kernel, dimensions, and computational costs are defined in Section 3.3.

depth and valid image coordinates can be queried. Let $\pi(p_i)$ denote the response-map coordinates and A_k the response for class k. The image-guided score is

$$s_i = \max_{k=1,\dots,K} A_k(\pi(p_i)). \tag{1}$$

The selection threshold is $\tau = 0.3$. With $S = \{i : s_i > \tau\}$ over valid projections, the retained indices are

$$\mathcal{I} = S \cup \{i : \exists j \in S,\ \|p_i - p_j\|_2 \le r\}, \qquad r = 1 \text{ m}. \tag{2}$$

Dilation adds existing LiDAR observations within the radius; it does not generate synthetic points. Set union retains each point only once. The geometric margin complements the image response: a point need not itself exceed the threshold if it lies near a selected observation. Thus, the sampler combines semantic selection with local geometric coverage.

The experiments use the single-camera setting illustrated in Figures 3 and 5; there is no cross-view response aggregation. Points without a valid image projection cannot become image-response seeds. Equation 2 describes the geometric expansion around valid seeds. The realized token count depends on the response map and scene geometry, rather than being a fixed-size sample. Table 3 measures the resulting token count, detection accuracy, and throughput.

Frozen response prediction and gradient flow. The RGS heatmap prediction branch is frozen during student training. Its responses supply the sampling scores, and no new heatmap supervision term is introduced into the student objective. Thresholding and set membership are discrete operations, so the retention mask is treated as $\mathrm{stopgrad}(\mathbb{1}[i \in \mathcal{I}])$. Detection and distillation optimize the trainable student path downstream of selection; they do not update the frozen response predictor through the hard mask. This separates a fixed sampling signal from the differentiable fusion and detection network, and removes the need to claim that point selection itself is end-to-end differentiable.

3.3 Kernelized Cross-Modal Attention

For one head, let $Q \in \mathbb{R}^{M\times d}$ and $K, V \in \mathbb{R}^{L\times d}$ be projected LiDAR queries and image keys/values. With $\phi(x) = \mathrm{ELU}(x) + 1$, define

$$B = \phi(K)^\top V \in \mathbb{R}^{d\times d}, \quad a = \phi(K)^\top \mathbf{1}_L \in \mathbb{R}^d, \quad z_i = \frac{\phi(q_i)^\top B}{\phi(q_i)^\top a + \epsilon}, \tag{3}$$

where $\epsilon > 0$ stabilizes the denominator. The scalar denominator is broadcast over the d value channels of each query. Reductions should accumulate in float32 under mixed precision. Outputs

from $h = 8$ heads, each with $d = 32$, are concatenated and linearly projected to $D = hd = 256$, followed by a residual connection and a two-layer GELU feed-forward block with another residual connection.

Equation 3 is the established kernel-attention reassociation (Katharopoulos et al., 2020). The ELU-plus-one kernel is not generally equal to, or a guaranteed approximation of, the exponential kernel in softmax attention. Changing kernels changes the model and may change accuracy.

Forming B costs $\mathcal{O}(Ld^2)$, and applying it to all queries costs $\mathcal{O}(Md^2)$. The interaction cost for h heads is therefore $\mathcal{O}(h(M + L)d^2)$, compared with $\mathcal{O}(hMLd)$ for dense attention. The auxiliary summary is $\mathcal{O}(hd^2)$, rather than an $\mathcal{O}(hML)$ attention matrix. These memory statements exclude stored inputs, outputs, and training activations. Dense projections also cost $\mathcal{O}((M+L)D^2)$, and the point encoder, image backbone, BEV network, and feed-forward blocks contribute additional work. Linear token scaling describes the attention interaction with fixed width, not an end-to-end latency guarantee.

3.4 Complete detection representation and supervision

For BEV cell (u, v) with physical step sizes (Δ_x, Δ_y) and origin $(x_{\min}, y_{\min})$, the student predicts K center probabilities, eight geometric channels, and optionally two velocity channels:

$$b = (\delta_x, \delta_y, z, \log w, \log l, \log h, \sin\theta, \cos\theta), \qquad \nu = (v_x, v_y). \tag{4}$$

The decoded center is $(x_{\min}+(u+\delta_x)\Delta_x,\ y_{\min}+(v+\delta_y)\Delta_y,\ z)$. Dimensions are exponentiated and yaw is decoded with $\mathrm{atan2}$. Including z makes the eight-channel geometry representation sufficient to reconstruct a full 3D box. A common coordinate frame and box-origin convention are required for labels, student predictions, teacher targets, and benchmark export.

Let $Y_{uvk} \in [0, 1]$ be a Gaussian center target and $\hat{p}_{uvk} \in (0, 1)$ its predicted probability. The Gaussian focal loss is

$$\ell(\hat{p}, Y) = \begin{cases} -(1-\hat{p})^{\gamma} \log \hat{p}, & Y = 1, \\ -(1-Y)^{\beta} \hat{p}^{\gamma} \log(1-\hat{p}), & 0 \le Y < 1, \end{cases} \tag{5}$$

with $\gamma = 2$ and $\beta = 4$, and $\mathcal{L}_{\text{heat}} = \sum_{uvk} \ell / \max(1, N_{\text{pos}})$. Unlike box regression, this loss includes background locations and Gaussian shoulders. Probabilities are clamped away from zero and one for numerical evaluation. Geometry uses SmoothL1 at assigned positive centers; velocity uses only centers with valid velocity annotations. Each loss is normalized by its own valid count, with denominator at least one. The detection objective is

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{heat}} + 2\mathcal{L}_{\text{box}} + 0.2\mathcal{L}_{\text{vel}}. \tag{6}$$

Velocity loss applies only to annotations with valid velocity targets. Geometry and velocity are supervised at assigned positive centers, whereas the heatmap loss includes background and Gaussian shoulders. Normalization by $\max(1, N_{\text{pos}})$ handles frames without positive centers.

3.5 Uncertainty-aware distillation

A pretrained TransFusion detector (Bai et al., 2022) processes the full input and supervises the lightweight student. Its uncertainty head shares detection features and uses two fully connected layers to predict positive regression variances under a diagonal Gaussian likelihood. Writing the regression target and mean as b^* and $\hat{b}$, with variance $v_j > 0$ in channel j, the negative log-likelihood has the form

$$\mathcal{L}_{\text{unc}} = \frac{1}{2} \sum_j \left(\frac{(b_j^* - \hat{b}_j)^2}{v_j} + \log v_j \right), \tag{7}$$

up to a constant. This objective trains the uncertainty head together with the teacher's detection objective. The teacher is frozen during student training. Let $\sigma_i \ge 0$ denote the instance-level uncertainty supplied for a matched object.

Teacher queries and student BEV cells have different index spaces. Training aligns intermediate teacher and student representations and applies loss constraints to corresponding outputs. Accordingly, let $\mathcal{E}$ denote a set of corresponding teacher–student predictions, with p_i^T, p_i^S representing class

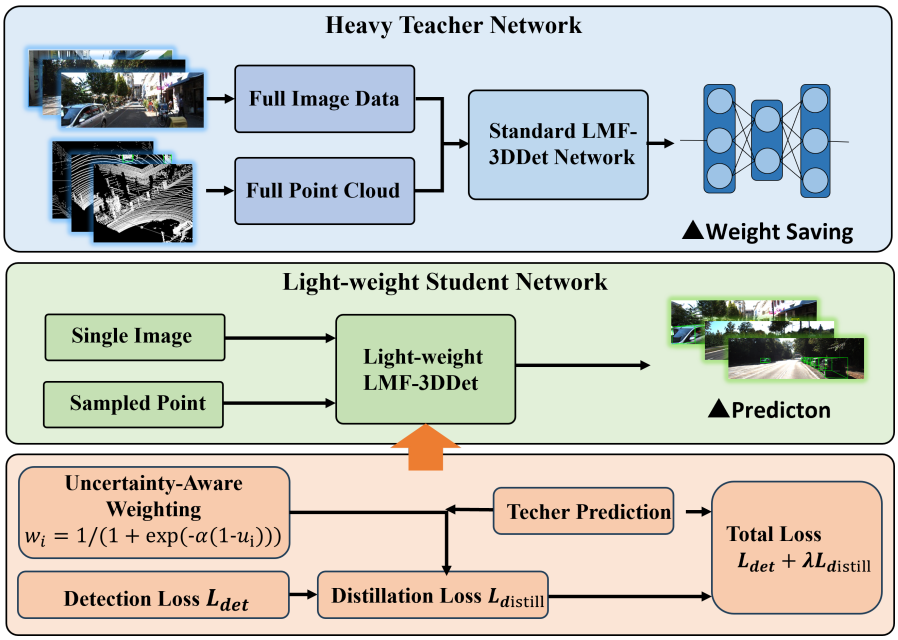


Figure 5: Teacher–student distillation. The full-input teacher supplies prediction and uncertainty targets during training; inference retains only the lightweight student. The illustration shows the image input as a single view.

distributions and b_i^T, b_i^S representing boxes in the same coordinate convention. The index i refers to a correspondence, not the same raw output index in both networks. Distillation compares only corresponding predictions:

$$\mathcal{L}_{\mathrm{KD}} = \sum_{i \in \mathcal{E}} w_i \left[D_{\mathrm{KL}}(p_i^T \| p_i^S) + \mathrm{SmoothL1}(b_i^S, b_i^T) \right]. \tag{8}$$

The correspondence is established through the training alignment, rather than assuming that a teacher query number equals a student cell number. Intermediate-layer matching supplies representation-level alignment, while the classification and box losses constrain prediction agreement. The formula above describes the prediction-level transfer term.

The uncertainty-dependent weight is

$$w_i = \mathrm{sigmoid}\big(\alpha(1 - e^{-\sigma_i})\big), \qquad \alpha = 5. \tag{9}$$

For $\sigma_i \geq 0$, the weight increases from $1/2$ toward $\mathrm{sigmoid}(5) \approx 0.9933$. Thus, the largest relative weight is less than twice the smallest. This bounded emphasis gives difficult examples greater influence while retaining supervision from confident examples. It is a difficulty weight rather than a teacher-confidence weight: uncertain targets are emphasized, not suppressed. Supplementary Section A gives the derivative and range.

The student objective combines detection and knowledge transfer:

$$\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{det}} + \lambda \mathcal{L}_{\mathrm{KD}}, \qquad \lambda = 1. \tag{10}$$

Teacher predictions and weights are held fixed when updating the student. During inference, the teacher and its uncertainty head are discarded. Knowledge transfer therefore improves the training signal without adding a teacher branch to the deployed model. Because a difficult target may also be unreliable, uncertainty emphasis is a trade-off; its detection benefit is evaluated against uniform weighting in Table 5.

4 EXPERIMENTS

4.1 DATASETS AND IMPLEMENTATION

We evaluate LMF-3DDet on nuScenes (Caesar et al., 2020) and Waymo Open Dataset (Sun et al., 2020). We report nuScenes test mAP and NDS for the main comparison, and Waymo validation mAP and mAPH at LEVEL_1. Component studies use nuScenes validation. Ablation rows can differ in enabled model components; the distillation table's FPS differences reflect these configurations, not the training loss.

Experiments use an NVIDIA RTX 4090 GPU with mixed precision. The student uses a single camera view, a frozen RGS response branch, an ImageNet-pretrained ResNet-18 image backbone, $\tau = 0.3$, a 1 m sampling radius, and eight attention heads with total width 256. We train for 60 epochs on nuScenes and 30 on Waymo using AdamW with batch size eight. The learning rate follows a cosine schedule from 10^{-4} to 10^{-6} after a five-epoch warm-up. Inference throughput is measured at batch size one. Supplementary Section B summarizes the remaining stated hyperparameters.

Table 1: nuScenes test comparison, ordered by publication year. †: published results; FPS follows the cited source (—: not reported). Bold highlights our model. Speeds are not a uniform benchmark.

Method	Venue	FPS ↑	mAP (%) ↑	NDS (%) ↑
PointPillars (Lang et al., 2019)	CVPR'19	62.1	45.1	52.4
CenterPoint (Yin et al., 2021)	CVPR'21	28.4	60.2	65.3
Fast-CLOCs (Pang et al., 2022)	WACV'22	42.1	62.3	65.4
DeepFusion (Li et al., 2022b)	CVPR'22	10.5	70.1	72.5
Focals Conv (Chen et al., 2022)	CVPR'22	38.5	64.1	67.2
TransFusion (Bai et al., 2022)	CVPR'22	15.2	68.9	71.2
BEVFormer (Li et al., 2022c)	ECCV'22	14.2	68.5	71.0
UVTR (Li et al., 2022a)	NeurIPS'22	12.6	69.1	71.5
BEVFusion (Liu et al., 2023)	ICRA'23	12.8	69.2	71.8
SparseFusion (Xie et al., 2023)	ICCV'23	18.3	70.2	72.6
GAFusion† (Li et al., 2024)	CVPR'24	—	73.6	74.9
EVT† (Lee et al., 2025)	ICCV'25	4.9	72.6	75.3
LMF-3DDet (ours)	—	**45.3**	**70.4**	**72.9**

Table 2: Waymo validation comparison (LEVEL_1), ordered by year. †: published LiDAR-only results; ‡: mean of three published class scores. FPS follows the source (—: not reported); speeds are not a uniform benchmark. Bold highlights our model.

Method	Venue	FPS ↑	mAP (%) ↑	mAPH (%) ↑
PointPillars (Lang et al., 2019)	CVPR'19	58.4	52.1	50.8
CenterPoint (Yin et al., 2021)	CVPR'21	26.8	66.5	65.2
Fast-CLOCs (Pang et al., 2022)	WACV'22	40.2	68.7	67.4
DeepFusion (Li et al., 2022b)	CVPR'22	10.1	76.5	75.8
Focals Conv (Chen et al., 2022)	CVPR'22	36.9	70.3	69.1
TransFusion (Bai et al., 2022)	CVPR'22	14.8	74.2	73.1
BEVFormer (Li et al., 2022c)	ECCV'22	13.5	72.3	71.0
UVTR (Li et al., 2022a)	NeurIPS'22	12.1	73.4	72.2
BEVFusion (Liu et al., 2023)	ICRA'23	12.3	75.0	74.2
SparseFusion (Xie et al., 2023)	ICCV'23	17.2	75.1	74.3
SAFDNet† (Zhang et al., 2024)	CVPR'24	20.2	81.8	79.9
UniMamba†‡ (Jin et al., 2025)	CVPR'25	—	82.3	80.2
LMF-3DDet (ours)	—	**44.8**	**76.8**	**76.1**

4.2 BENCHMARK RESULTS

On nuScenes (Table 1), LMF-3DDet achieves 70.4% mAP and 72.9% NDS at 45.3 FPS with a single camera. Recent multi-camera methods GAFusion and EVT report 73.6/74.9 and 72.6/75.3 mAP/NDS, respectively. They achieve higher accuracy, while LMF-3DDet offers a lightweight

single-camera operating point. These literature results use different input and timing protocols, so the comparison supports a discussion of accuracy and efficiency rather than a controlled speed ranking.

On Waymo (Table 2), LMF-3DDet achieves 76.8% mAP and 76.1% mAPH at 44.8 FPS, exceeding the listed DeepFusion result by 0.3 points on both accuracy metrics. Recent LiDAR-only methods SAFDNet and UniMamba achieve higher LEVEL_1 accuracy. SAFDNet reports 81.8/79.9 mAP/mAPH at 20.2 FPS; UniMamba yields 82.3/80.2 after averaging its vehicle, pedestrian, and cyclist LEVEL_1 scores. These results place LMF-3DDet at a high-throughput operating point, with scope to improve absolute accuracy. Published speeds provide context rather than a controlled speed ranking.

4.3 Component studies

Table 3: Sampling strategies on nuScenes validation. Tokens denotes the input token budget; RGS reduces the budget while preserving detection accuracy.

Sampling method	Tokens(k)	mAP (%)	NDS (%)	FPS
Random (30k)	30	68.5	70.3	32.1
Farthest point	30	68.9	70.7	28.7
Random (10k)	10	66.2	68.5	48.2
Ours	11.5	**69.7**	**71.3**	**44.5**

Response-guided sampling. Table 3 shows that RGS uses 11.5k tokens compared with 30k for the larger random baseline, a 61.7% reduction. At the same time, mAP increases from 68.5 to 69.7, NDS increases from 70.3 to 71.3, and throughput rises from 32.1 to 44.5 FPS. Against the more aggressive 10k random setting, RGS improves mAP by 3.5 points with 1.5k additional tokens. Together, these operating points demonstrate a favorable balance between selective input reduction and detection quality. The comparison changes both selection policy and token budget. Compared with farthest-point sampling, RGS achieves 0.8 points higher mAP and 1.55$\times$ throughput with fewer tokens. The 10k random setting is faster, but its 3.5-point mAP deficit shows the cost of aggressive indiscriminate thinning. RGS therefore favors detection quality over the highest possible sampling-induced throughput.

Table 4: Attention comparison on nuScenes validation. Complexity is per head with M LiDAR tokens, L image tokens, and width d, excluding projections. FPS measures end-to-end throughput.

Attention type	Complexity	mAP (%)	FPS
Standard (quadratic)	$\mathcal{O}(MLd)$	69.9	32.5
Linear (Ours)	$\mathcal{O}((M+L)d^2)$	**69.7**	**44.5**

Linear attention. Table 4 compares standard and kernelized attention. Linear attention increases throughput from 32.5 to 44.5 FPS, a 1.37$\times$ end-to-end improvement, while mAP changes from 69.9 to 69.7. The resulting mAP ratio is 99.7%. Thus, replacing dense token interaction yields a substantial throughput gain with a 0.2-point accuracy cost. Complexity is linear in token count at fixed width; the measured gain is end-to-end, not attention-only. Converting throughput to average per-frame time gives approximately 30.8 ms for standard attention and 22.5 ms for the linear-attention configuration, a 27.0% reduction. These values are calculated from the reported FPS, not separately profiled attention timings. The small accuracy change suggests that dense pairwise interactions can be reduced at the evaluated operating point.

Uncertainty-aware distillation. Table 5 reports 68.2, 69.8, and 70.4 mAP for no distillation, uniform distillation, and uncertainty-aware distillation, respectively. Uncertainty-aware transfer exceeds uniform transfer by 0.6 mAP and 0.8 NDS points, with similar measured throughput (45.3 versus 45.1 FPS). The uncertainty-aware configuration provides the best accuracy among the evaluated distillation settings. The complete uncertainty-aware setting is 2.2 mAP points above the no-distillation row. Because the teacher is removed at inference, the 32.5-to-45.3 FPS difference comes from the ablated model configurations, not the distillation loss itself. This supports choosing uncertainty-aware transfer for detection quality; its training-only loss does not accelerate inference.

Table 5: Distillation comparison on nuScenes validation. Rows correspond to ablated model configurations. FPS differences arise from these configurations, not from the training-only distillation loss.

Distillation variant	mAP (%)	NDS (%)	FPS
No distillation	68.2	70.1	32.5
Vanilla distillation (uniform)	69.8	72.1	45.1
Uncertainty-aware (Ours)	**70.4**	**72.9**	**45.3**

Table 6: Runtime breakdown on an RTX 4090 at batch size one. Module latencies sum to 22.1 ms per frame.

Module	Latency (ms)	Proportion (%)
Voxelization	2.4	10.8
Image backbone	6.2	28.0
RGS sampling	1.1	5.0
Linear attention	5.8	26.2
Detection head	4.3	19.5
NMS & postproc.	2.3	10.5
Total	22.1	100

4.4 Runtime analysis

Table 6 gives the module-level runtime at batch size one. The listed latencies sum to 22.1 ms per frame, corresponding to approximately 45.25 FPS and agreeing with the reported 45.3 FPS to the precision of the timings. The image backbone and linear attention take 6.2 and 5.8 ms, respectively, while RGS takes 1.1 ms. Input selection is a small fraction of the measured computation; image processing and fusion remain the main optimization targets. Together, the backbone and attention account for 12.0 of the 22.1 ms, or approximately 54.3% of the total. Consequently, further reductions in sampling cost alone have limited impact: even eliminating the listed 1.1 ms would reduce the total by only about 5.0%, assuming all other stages stayed fixed. This breakdown motivates optimizing feature extraction and fusion jointly.

5 Limitations

RGS depends on image responses and calibration, so degraded images or projection errors can remove useful LiDAR observations. A dilation margin reduces sensitivity to local selection errors but does not guarantee foreground coverage. Kernel attention compresses image information into a fixed-width summary, and uncertainty weighting can emphasize unreliable teacher targets. The present evaluation establishes aggregate detection accuracy and GPU throughput; class-specific robustness, embedded-device latency, and energy use require dedicated measurements. Repeated-seed evaluation and matched-token-budget controls would further characterize small accuracy differences and separate the effects of sampling policy from input size.

6 Conclusion

LiDAR–camera detection faces a tension between computational efficiency and preserving informative observations. LMF-3DDet addresses this problem through coordinated sampling, fusion, and distillation. Response-guided sampling concentrates computation on informative LiDAR observations, with geometric dilation retaining surrounding context. Kernelized linear cross-attention reduces the cost of feature interaction, while uncertainty-aware distillation strengthens the lightweight student without retaining the teacher at inference. Experiments on nuScenes and Waymo demonstrate a favorable accuracy–efficiency trade-off and real-time throughput. Component studies validate selective sampling over aggressive random thinning, efficient linear fusion with a small accuracy cost, and improved student accuracy through uncertainty-aware supervision.

AI USE STATEMENT

Generative AI assisted with proofreading and polishing the manuscript, checking literature, and revising mathematical presentation and result analysis sentences.

REPRODUCIBILITY STATEMENT

Section 3 describes the sampling, attention, detection, and distillation formulations. Section 4.1 and Supplementary Section B summarize the training and evaluation settings, and Supplementary Section A derives the attention complexity and uncertainty-weight properties.

REFERENCES

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 *CVPR*, pp. 1090–1099, 2022.

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 *CVPR*, pp. 11621–11631, 2020.

Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3D object detection. In *CVPR*, pp. 5428–5437, 2022.

Xin Jin, Haisheng Su, Kai Liu, Cong Ma, Wei Wu, Fei Hui, and Junchi Yan. UniMamba: Unified spatial-channel representation learning with group-efficient mamba for LiDAR-based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.12009.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, volume 119, pp. 5156–5165, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *CVPR*, pp. 12697–12705, 2019.

Yongjin Lee, Hyeon-Mun Jeong, Yurim Jeon, and Sanghyun Kim. EVT: Efficient view transformation for multi-modal 3D object detection. In *ICCV*, pp. 26632–26642, 2025.

Xiaotian Li, Baojie Fan, Jiandong Tian, and Huijie Fan. GAFusion: Adaptive fusing LiDAR and camera with multiple guidance for 3D object detection. In *CVPR*, pp. 21209–21218, 2024.

Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3D object detection. In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection. In *CVPR*, pp. 17182–17191, 2022b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pp. 1–18, 2022c. doi: 10.1007/978-3-031-20077-9_1.

Min Liu, Yuanjun Jia, Youhao Lyu, Qi Dong, and Yanyu Yang. BAFusion: Bidirectional attention fusion for 3D object detection based on LiDAR and camera. *Sensors*, 24(14):4718, 2024. doi: 10.3390/s24144718.

Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023. URL https://arxiv.org/abs/2205.13542.

Su Pang, Daniel Morris, and Hayder Radha. Fast-CLOCs: Fast camera-lidar object candidates fusion for 3D object detection. In *WACV*, 2022.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pp. 2446–2454, 2020.

Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, and Liwei Wang. UniTR: A unified and efficient multi-modal transformer for bird's-eye-view representation. In *ICCV*, pp. 6792–6802, 2023. URL https://arxiv.org/abs/2308.07732.

Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection. In *ICCV*, pp. 17591–17602, 2023.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3D object detection. In *ICCV*, pp. 18268–18278, 2023. URL https://arxiv.org/abs/2301.01283.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *CVPR*, pp. 11784–11793, 2021.

Gang Zhang, Junnan Chen, Guohuan Gao, Jianmin Li, Si Liu, and Xiaolin Hu. SAFDNet: A simple and effective network for fully sparse 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2403.05817.

Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. PointDistiller: Structured knowledge distillation towards efficient and compact 3D detection. In *CVPR*, pp. 21791–21801, 2023.