DECOUPLING TO FUSE: LEARNING SHARED AND MODALITY-SPECIFIC REPRESENTATIONS FOR MULTI-SENSOR 3D OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fusing information from cameras, LiDAR, and 4D radar is an important approach to improving 3D object detection in autonomous driving. However, multimodal features contain a mixture of shared information, modality-specific cues, and background responses. A key challenge is to exploit cross-modal consistency while preserving useful differences and guiding their fusion according to object relevance. We introduce ObjDec, a multi-sensor fusion architecture that learns shared and modality-specific representations and uses object-related information to guide their fusion. Specifically, ObjDec applies constraints to aligned local representations within foreground regions, encouraging cross-modal consistency in shared features while preserving differences among modality-specific features. Based on these representations, foreground gating and modality control regulate each modality's feature contribution, while object context further modulates cross-modal attention queries and fused outputs. Experiments on K-Radar dataset demonstrate that ObjDec achieves state-of-the-art performance, improves 8.04% AP_{3D}@IoU=0.3 (88.36%)and 7.77% AP_{BEV}@IoU=0.5 (88.10%)at a confidence threshold of 0.3, with only approximately 1.7% more parameters and fast inference. Component ablations and analyses of learned representations and spatial gating further support the effectiveness of the proposed design. Our code will be made publicly available.

1 INTRODUCTION

Accurate 3D object detection is essential for autonomous driving. Cameras, LiDAR, and 4D radar observe a scene through different sensing mechanisms, providing visual appearance, geometric structure, and radar-return information. Combining these observations can compensate for limitations of individual sensors and provide a richer basis for detecting objects. Realizing this potential requires integrating features with different sensing characteristics and spatial sampling patterns into a joint representation that supports detection, motivating continued research on multi-sensor fusion (Liu et al., 2023; Park et al., 2025b; Yue et al., 2025).

Existing approaches have advanced multi-sensor detection through spatial interaction and adaptive fusion. Establishing correspondence in a common spatial domain allows features from different sensors to exchange information about the same regions. Attention and gating mechanisms further adjust their contributions, while foreground modeling and reliability estimation help account for differences in input quality. Together, these developments provide an effective foundation for combining sensor observations under varying scene and sensing conditions (Chae et al., 2024; Huang et al., 2025; Chae et al., 2025; Paek & Kong, 2025; Park et al., 2026).

Nevertheless, spatial correspondence and adaptive weighting do not by themselves distinguish the information that can be shared across sensors from the cues specific to each modality. Even within aligned local regions, these two types of information remain mixed in the input features, although they play different roles in fusion: shared information can provide cross-modal support, while modality-specific cues can preserve complementary observations. Learning consistency alone may suppress useful differences, whereas emphasizing differences alone may overlook evidence shared across sensors. This motivates approaching fusion through representation decoupling: explicitly

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.

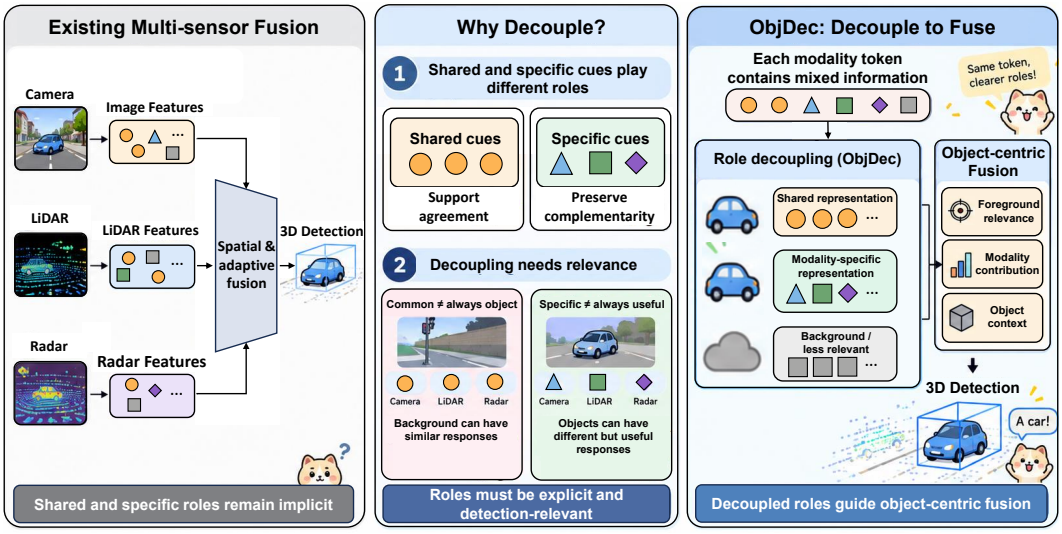

Figure 1: **Decoupling before fusion.** Spatial interaction and adaptive weighting can leave shared and modality-specific information entangled. ObjDec explicitly learns both representations and uses object relevance to guide their fusion.

learning shared and modality-specific representations and using both to construct the fused detection features. Such a formulation builds on advances in learning and retaining common and distinct information for multimodal prediction (Liang et al., 2023; Liu et al., 2025; Qian et al., 2026; Wang et al., 2026).

For dense 3D detection, this decoupling also needs a clear connection to detection targets. Shared responses may originate from either objects or background, and modality-specific differences may reflect useful observations or noise. Separating the two types of information therefore does not by itself ensure their relevance to detection. We address this issue by grounding representation constraints in object regions and using object-related information to guide the contribution of the learned features to fusion. Object guidance thus gives the learning and use of decoupled representations a detection-specific focus.

We introduce **ObjDec**, a multi-sensor 3D detection architecture centered on shared and modality-specific representation decoupling. Within spatially corresponding bird's-eye-view (BEV) patches, separate mappings construct the two representations from each sensor token, and both participate in building the fused detection features. Object-region supervision encourages cross-modal agreement in shared representations, preserves modality-specific differences, and limits similarity between the two branches. The learned representations further determine feature updates, modality contributions, and object context, allowing their information to shape fusion inputs and cross-modal interaction. Foreground gating and context modulation implement this connection between decoupled representations and fusion computation. The representation mappings operate at every patch; object regions specify where the representation constraints are applied during training. Ground-truth boxes provide training supervision only, while inference predicts all fusion-control signals from input features without prior detections or weather labels.

We evaluate ObjDec primarily on K-Radar. At a confidence-filtering threshold of 0.3 on v1, ObjDec achieves 88.36% $\mathrm{AP}_{\mathrm{3D}}$ at IoU 0.3 and 88.10% $\mathrm{AP}_{\mathrm{BEV}}$ at IoU 0.5. Compared with the archived results associated with the released fusion baseline under the same filtering setting, these scores improve by 8.04 and 7.77 percentage points, respectively, with approximately 1.7% more parameters. Component ablations assess the contribution of each design, while representation visualizations and spatial gate analyses examine the learned feature organization and fusion behavior. We also evaluate our method on K-Radar v2 and train and test the architecture on V2X-Radar-V, examining its applicability beyond the primary category and sensor setting under the specified dataset protocols (Paek et al., 2022; Paek & Kong, 2025; Yang et al., 2025).

Our main contributions are as follows:

- **A fusion architecture based on representation decoupling.** We develop a multi-sensor 3D detection architecture that explicitly learns shared and modality-specific representations in corresponding local regions and uses both to construct fused detection features.
- **Object-guided learning and use of decoupled representations.** We supervise representation decoupling within object regions and connect the learned representations to foreground gating, modality contribution control, and object context, guiding their use in detection-oriented fusion.
- **Evaluation of detection performance and model behavior.** We evaluate detection performance and component contributions on K-Radar, analyze representations and spatial gating, and examine the architecture under extended categories and target-dataset training on V2X-Radar-V.

2 Related Work

2.1 Multi-Sensor Fusion for 3D Object Detection

Multi-sensor detectors combine observations through spatial correspondence and feature interaction. BEVFusion (Liu et al., 2023) brings camera and LiDAR features into a unified bird's-eye-view (BEV) space, while TransFusion (Bai et al., 2022) and CMT (Yan et al., 2023) establish cross-modal interactions through LiDAR-guided queries and 3D positional encoding, respectively. For radar–camera fusion, RCBEVDet (Lin et al., 2024) combines radar BEV encoding with deformable cross-attention for feature alignment. These designs offer different ways to establish correspondence and integrate sensor features.

Recent studies further address modality contributions and varying input quality. MoME (Park et al., 2025b) routes queries among camera, LiDAR, and joint expert decoders. CCF (Wu et al., 2026) combines separate query supervision, geometric priors, and complementary masking to address modality imbalance in cross-domain detection. RobuRCDet (Yue et al., 2025) uses Gaussian expansion of radar points and camera-confidence-guided fusion to accommodate different sensor disturbances.

On K-Radar, 3D-LRF (Chae et al., 2024) and L4DR (Huang et al., 2025) investigate spatial interaction, foreground denoising, and gated fusion. DLRFusion (Chae et al., 2025) separately encodes Doppler, radar power, and LiDAR features for iterative interaction. WCBR (Li et al., 2026) adjusts branch contributions according to environmental conditions, while SRF (Mukhatbekov et al., 2026) fuses stereo images and radar in 3D voxel space. For three-sensor fusion, ASF (Paek & Kong, 2025) uses unified canonical projection and cross-sensor attention within corresponding patches, whereas RAF (Park et al., 2026) modulates camera features using explicitly supervised image reliability maps. Building on ASF's projection and local interaction mechanisms, ObjDec makes shared and modality-specific representation decoupling an explicit part of the local fusion architecture.

2.2 Shared and Modality-Specific Representation Learning

Explicitly organizing shared and specific information has an established foundation. Domain Separation Networks (Bousmalis et al., 2016) distinguish shared and domain-specific representations for domain adaptation. MISA (Hazarika et al., 2020) learns modality-invariant and modality-specific representations through similarity, difference, and reconstruction constraints for affect analysis. FactorCL (Liang et al., 2023) retains task-relevant shared and unique information, while DeCUR (Wang et al., 2024) learns common and unique representations through cross-modal and intra-modal correlation objectives. These studies support learning consistency while preserving useful differences.

Recent work further explores how these representations are selected, aligned, and used. Liu et al. (2025) use a causal formulation to further separate modality-specific information into unique and redundant components. MultiLoReFT (Tonekaboni et al., 2026) learns shared and modality-specific subspaces through low-rank representation fine-tuning. DecAlign (Qian et al., 2026) uses prototype-guided optimal transport and distribution matching for modality-specific and common representations, respectively. For detection, SPFD (Wang et al., 2026) separates RGB–event shared and private features through frequency-domain coherence and uses them in both its encoder and decoder. LMD (Park et al., 2025a) instead provides post-hoc decomposition of modality contributions in pretrained fusion networks.

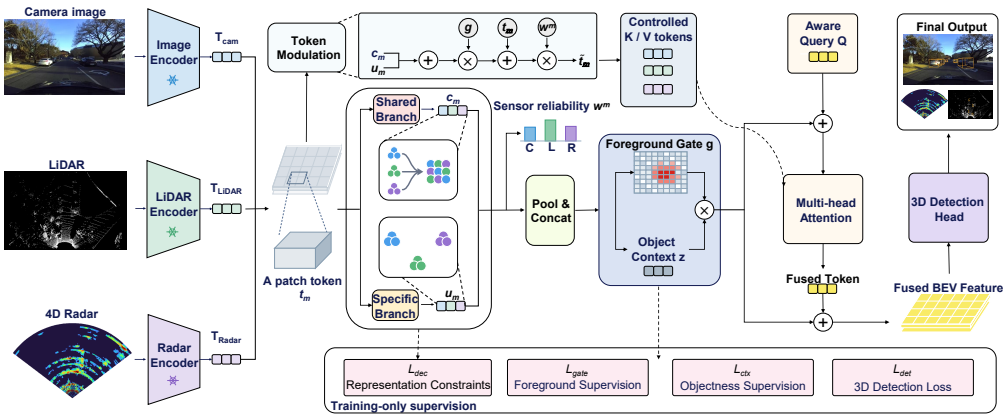

Figure 2: **ObjDec architecture.** Within aligned BEV patches, each modality token produces shared and modality-specific representations. These representations predict foreground gates, modality contribution weights, and object context: the first two control token updates and scaling, while the context modulates attention queries and fused outputs. GT supervises training losses only; inference uses feature-predicted controls.

Building on this line of work, ObjDec uses shared and modality-specific representation decoupling to organize local fusion for multi-sensor 3D detection. Its design connects spatially corresponding sensor patches, representation constraints supervised within object regions, and direct use of the decoupled representations in token updates and cross-modal interaction. Object guidance gives the learning and use of these representations a detection-specific focus, while representation decoupling provides the organizing principle of the fusion architecture.

3 METHOD

3.1 OVERVIEW

We introduce ObjDec, a multi-sensor 3D detection architecture centered on shared and modality-specific representation decoupling. As illustrated in Figure 2, modality-specific encoders first transform camera, LiDAR, and 4D radar observations into spatially corresponding bird's-eye-view (BEV) features. We use unified projection to map features with different channel dimensions into patch tokens for corresponding local regions, and retain local cross-sensor attention and subsequent feature transformation as the interaction backbone (Paek & Kong, 2025). Let $\mathcal{M}$ denote the participating modalities, with $M = |\mathcal{M}|$, and let $t_m^p \in \mathbb{R}^d$ be the input token of modality m at patch p. We omit the batch dimension for clarity.

Within each corresponding patch, separate mappings learn shared and modality-specific representations with distinct training constraints. Both representations contribute to fusion through residual token updates and the prediction of modality contributions and object context. Object-region supervision and predicted foreground gates give their learning and use a detection-specific focus. The gates and modality weights regulate input-token updates and scaling, while object context modulates attention queries and fused outputs. Local interaction is followed by feature transformation and spatial reassembly to produce fused BEV features for the 3D detection head. Control signals are predicted from input features during both training and inference; ground-truth (GT) boxes provide supervision targets for representation constraints and object-related predictions only during training.

3.2 SHARED AND MODALITY-SPECIFIC REPRESENTATION DECOUPLING

Two independent MLPs per modality map each token to shared and modality-specific representations:

$$c_m^p = \phi_m^c(t_m^p), \qquad u_m^p = \phi_m^u(t_m^p), \quad c_m^p, u_m^p \in \mathbb{R}^d. \tag{1}$$

Parameters are separate across modalities; shared agreement is learned through supervision. We define $\mathcal{P}_{\text{fg}}$ as patches whose centers lie inside margin-expanded GT BEV box footprints. Both mappings operate everywhere, while the following constraints apply only within this foreground set:

$$\begin{aligned} \mathcal{L}_c &= \mathbb{E}_{p\in\mathcal{P}_{\text{fg}},\, m<n}\left[1-\cos(c_m^p, c_n^p)\right], \\ \mathcal{L}_u &= \mathbb{E}_{p\in\mathcal{P}_{\text{fg}},\, m<n}\left[\cos(u_m^p, u_n^p)-\delta\right]_+, \\ \mathcal{L}_{\text{sep}} &= \mathbb{E}_{p\in\mathcal{P}_{\text{fg}},\, m}\left[|\cos(c_m^p, u_m^p)|\right]. \end{aligned} \tag{2}$$

Here, cos is cosine similarity, $[a]_+ = \max(a, 0)$, and δ is a similarity margin. Expectations average the indicated patches and modalities or unordered modality pairs. $\mathcal{L}_c$ encourages cross-modal shared agreement, $\mathcal{L}_u$ penalizes excessive similarity of specific features, and $\mathcal{L}_{\text{sep}}$ separates the two branches within each modality. These soft constraints organize foreground representations alongside the detection objective.

3.3 Representation-Driven Fusion

The joint descriptor $h^p = [\bar{c}^p; \bar{u}_{\text{abs}}^p]$ concatenates the mean shared feature $\bar{c}^p = M^{-1}\sum_m c_m^p$ and mean specific-feature magnitude $\bar{u}_{\text{abs}}^p = M^{-1}\sum_m |u_m^p|$, with elementwise absolute values.

Input-token modulation. The descriptor predicts a foreground gate; each modality score also uses the original token, its representations, and its deviation from the shared mean:

$$\begin{aligned} g^p &= \sigma(f_g(h^p)), \\ s_m^p &= f_m([t_m^p; c_m^p; u_m^p; |c_m^p - \bar{c}^p|]), \\ \alpha_m^p &= \frac{\exp(s_m^p/\tau)}{\sum_{n\in\mathcal{M}}\exp(s_n^p/\tau)}. \end{aligned} \tag{3}$$

Here, f_g, f_m are learned heads, σ is sigmoid, and $\tau > 0$ is temperature. The continuous gate g^p is shared across modalities at each patch; α_m^p denotes relative modality contribution. Neither requires prior detections. Bounded scaling and a representation residual then update each token:

$$\begin{aligned} w_m^p &= \text{clip}(1 + \eta g^p(M\alpha_m^p - 1),\, w_{\min}, w_{\max}), \\ \widetilde{t}_m^p &= w_m^p\left[t_m^p + \lambda g^p(c_m^p + u_m^p)\right]. \end{aligned} \tag{4}$$

The coefficients η, λ control scaling and residual strength; $[w_{\min}, w_{\max}]$ contains one. Uniform contributions give unit scaling, and $g^p \to 0$ recovers the original token. Thus, gating controls update strength without a binary background mask.

Object-context-guided interaction. In the primary single-class setting, a separate objectness head f_{obj} and learned mapping W_z produce context:

$$r^p = \sigma(f_{\text{obj}}(h^p)), \qquad z^p = \tanh(W_z r^p), \quad z^p \in \mathbb{R}^d. \tag{5}$$

The gate controls update strength, while z^p supplies a context vector. Multi-class settings use foreground class probabilities instead (Appendix A.2). Let $Q \in \mathbb{R}^{N_q\times d}$ be learned queries shared across patches and $\widetilde{T}^p \in \mathbb{R}^{M\times d}$ stack the controlled modality tokens. Context modulates queries and fused outputs:

$$\begin{aligned} Q'^p &= Q + \beta g^p \mathbf{1}_{N_q}(z^p)^\top, \\ F^p &= \text{MHA}(Q'^p, \widetilde{T}^p, \widetilde{T}^p), \\ F'^p &= F^p + \gamma g^p \mathbf{1}_{N_q}(z^p)^\top. \end{aligned} \tag{6}$$

MHA uses learned query/key/value projections; $\mathbf{1}_{N_q}$ broadcasts context across query positions, with residual strengths β, γ. The output F'^p is transformed and spatially reassembled into fused BEV features.

3.4 Training Objectives

In the primary setting, foreground indicators $y^p = \mathbf{1}[p \in \mathcal{P}_{\mathrm{fg}}]$ supervise both the foreground gate g^p and objectness prediction r^p. Their losses, $\mathcal{L}_g$ and $\mathcal{L}_{\mathrm{ctx}}$, use binary cross-entropy with positive-class weighting over foreground and background patches in a training batch. The three representation constraints in Equation equation 2 are evaluated within the foreground set. The training forward pass also uses the feature-predicted g^p and r^p; labels y^p enter only the loss computation and do not replace these predictions in fusion.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{aux}}\big(\lambda_c\mathcal{L}_c + \lambda_u\mathcal{L}_u + \lambda_{\mathrm{sep}}\mathcal{L}_{\mathrm{sep}} + \lambda_g\mathcal{L}_g + \lambda_{\mathrm{ctx}}\mathcal{L}_{\mathrm{ctx}}\big). \tag{7}$$

Here, $\mathcal{L}_{\mathrm{det}}$ includes detection classification, box regression, and orientation losses. The coefficient λ_{aux} controls the overall weight of the added supervision, while the remaining coefficients balance its components. The forward modulation is differentiable, allowing detection and auxiliary objectives to jointly optimize the representation branches and control heads.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate primarily on K-Radar (Paek et al., 2022), using v1 for Sedan detection in the narrow driving corridor and v2 for Sedan and Bus/Truck detection over a wider region. The filtered test sets contain 10,065 and 13,727 frames, respectively. We report 3D and BEV AP at a confidence filter of 0.3; v2 means average the two classes. Published baselines retain their source protocols, so labels, preprocessing, and evaluators are not uniformly matched. Spatial ranges, evaluator details, and checkpoint provenance are given in Appendix B.

Implementation. K-Radar ObjDec models use frozen pretrained modality encoders and AdamW, with an 11-epoch budget and batch size 2. The v1 model and ablations use a preliminary 1,000-frame evaluation for checkpoint selection; v2 uses final-epoch weights. Training settings and selection details are provided in Appendices A and B.

4.2 Detection Performance on K-Radar

Main results. In Table 1, ObjDec achieves state-of-the-art performance. Our method achieves 88.36 $\mathrm{AP}_{3D}@0.3$ and 88.10 $\mathrm{AP}_{BEV}@0.5$, improving over ASF (Paek & Kong, 2025) by 8.04 and 7.77 percentage points under the reported confidence setting. At $\mathrm{AP}_{3D}@0.5$, the gain is smaller: 67.19 to 67.50. Weather-wise results show complementary strengths: rainy-weather 3D@0.5 increases from 67.33 to 74.92, whereas sleet decreases from 67.28 to 57.60. Appendix C reports stricter IoUs and confidence sensitivity.

Extended setting. Table 3 evaluates our method on the K-Radar v2. It obtains 43.05 mean $\mathrm{AP}_{3D}@0.5$, versus 41.57 for ASF, with Bus/Truck improving from 31.06 to 33.97. Mean BEV@0.5 remains lower at 60.93 versus 62.45, and the official L4DR checkpoint (Huang et al., 2025) remains stronger in mean 3D@0.5. Fixed-weight sensor-availability results are provided in Appendix F.

4.3 Ablations and Efficiency

Component contributions. Table 2 tests representation supervision and fusion control. Removing decoupling supervision, while retaining both representation branches and all forward paths, reduces $\mathrm{AP}_{3D}@0.3$ from 88.36 to 80.23. Removing foreground gating, modality control, or object context also reduces both reported 3D metrics. These results support jointly learning decoupled represen-

Table 1: **K-Radar v1 detection AP (%) across weather conditions.** C/L/R: camera/LiDAR/radar; Nor./Ove./Sle./L.s./H.s.: normal/overcast/sleet/light snow/heavy snow. Dark/light blue mark first/second among listed values, including ties. ASF and L4DR use our reproduced model. Other methods retain reported protocols (Section 4.1). Dashes: unreported. †: camera-assisted weather routing.

Method	Sensors	IoU	Metric	Total	Nor.	Ove.	Fog	Rain	Sle.	L.s.	H.s.
RTNH (Paek et al., 2022)	R	0.3	BEV	41.10	41.00	44.60	45.40	32.90	50.60	81.50	56.30
			3D	37.40	37.60	42.00	41.20	29.20	49.10	63.90	43.10
		0.5	BEV	36.00	35.80	41.90	44.80	30.20	34.50	63.90	55.10
			3D	14.10	19.70	20.50	15.90	13.00	13.50	21.00	6.36
RTNH (Paek et al., 2022)	L	0.3	BEV	76.50	76.50	88.20	86.30	77.30	55.30	81.10	59.50
			3D	72.70	73.10	76.50	84.80	64.50	53.40	80.30	52.90
		0.5	BEV	66.30	65.40	87.40	83.80	73.70	48.80	78.50	48.10
			3D	37.80	39.80	46.30	59.80	28.20	31.40	50.70	24.60
3D-LRF (Chae et al., 2024)	L+R	0.3	BEV	84.00	83.70	89.20	95.40	78.30	60.70	88.90	74.90
			3D	74.80	81.20	87.20	86.10	73.80	49.50	87.90	67.20
		0.5	BEV	73.60	72.30	88.40	86.60	76.60	47.50	79.60	64.10
			3D	45.20	45.30	55.80	51.80	38.30	23.40	60.20	36.90
L4DR (Huang et al., 2025)	L+R	0.3	BEV	79.49	86.30	89.62	89.91	81.07	62.31	89.14	61.32
			3D	77.96	77.73	80.04	88.56	79.16	60.11	78.91	51.91
		0.5	BEV	77.54	76.81	88.56	89.72	78.22	59.30	80.93	53.81
			3D	53.50	53.01	64.06	73.23	53.77	46.21	52.41	36.98
RAF on L4DR (Park et al., 2026)	C+L+R	0.3	BEV	—	—	—	—	—	—	—	—
			3D	—	—	—	—	—	—	—	—
		0.5	BEV	82.00	82.00	88.20	93.90	82.60	63.00	88.50	60.40
			3D	57.40	59.40	67.60	79.30	53.30	45.90	56.30	37.80
AW-MoE (Lin et al., 2026)	L+R†	0.3	BEV	88.20	87.70	**94.50**	96.70	88.80	**81.00**	**95.40**	70.20
			3D	83.90	84.20	90.00	**95.30**	84.40	72.90	**90.20**	64.00
		0.5	BEV	84.20	82.80	**91.60**	**96.30**	85.30	**75.00**	**94.70**	66.40
			3D	61.50	59.00	67.20	85.70	63.50	43.30	70.10	53.10
AW-MoE-LRC (Lin et al., 2026)	C+L+R	0.3	BEV	—	—	—	—	—	—	—	—
			3D	84.30	84.70	**91.00**	**95.30**	84.00	72.90	89.60	63.70
		0.5	BEV	—	—	—	—	—	—	—	—
			3D	61.80	60.20	70.40	**85.80**	63.40	43.70	69.90	52.80
ASF (Paek & Kong, 2025)	C+L+R	0.3	BEV	80.78	88.40	90.28	90.86	81.22	80.48	89.53	71.96
			3D	80.31	79.57	89.89	90.67	80.97	80.20	80.89	**71.71**
		0.5	BEV	80.33	79.63	90.19	90.78	81.13	71.04	80.82	**71.04**
			3D	67.19	64.56	79.71	79.57	67.33	**67.28**	77.63	**61.61**
ObjDec (ours)	C+L+R	0.3	BEV	**88.84**	**88.44**	90.47	**99.51**	**89.22**	80.72	89.50	**79.26**
			3D	**88.36**	**87.66**	90.39	90.57	**88.90**	**80.42**	89.28	71.41
		0.5	BEV	**88.10**	**87.41**	90.30	90.54	**89.02**	70.24	88.89	70.51
			3D	**67.50**	**65.98**	**80.26**	79.71	**74.92**	57.60	**78.33**	60.51

Table 2: **Component ablations on K-Radar v1.** AP (%) with C+L+R at confidence 0.3. Ablation definitions and checkpoint selection follow Appendices D and B.

Variant	AP_{3D} @0.3	AP_{3D} @0.5	AP_{BEV} @0.5
Full ObjDec	**88.36**	**67.50**	**88.10**
w/o modality contribution control	79.62	64.38	79.91
w/o object context	80.14	66.23	79.96
w/o decoupling supervision	80.23	66.16	80.04
w/o foreground gating	80.00	64.80	80.12

tations and using them to guide fusion. Ablation definitions and additional metrics are detailed in Appendix D.

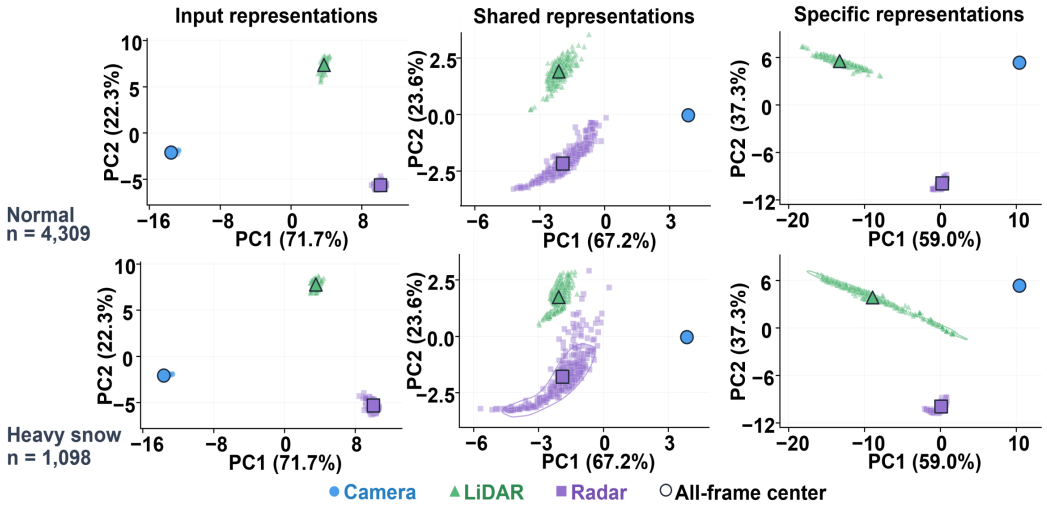

Figure 3: **Frame-level PCA of input, shared, and specific features.** Points are frame-mean features over GT foreground patches: camera (blue circles), LiDAR (green triangles), and radar (purple squares). Outlined markers and contours denote full-group centroids and approximately 80% density regions. Columns use separate full-test PCA bases shared across rows; axis percentages indicate explained variance. Fitting details are in Appendix E.1.

Efficiency. ObjDec uses 79.47M parameters versus ASF's 78.13M, an increase of approximately 1.7%. Paired batch-one measurements on an RTX A6000 give eager-inference times of 84.69 and 84.00 ms, respectively, excluding disk loading. Appendix H provides the timing protocol and optimized-runtime measurements.

4.4 Representation and Qualitative Analysis

Decoupled representations. Figure 3 shows different distribution structures for shared and modality-specific features, with residual modality structure in the shared branch. Across seven weather groups, mean cosine similarity at matched foreground locations in the original feature space ranges from 0.951–0.959 for shared features and 0.446–0.549 for specific features. Together with the decoupling-supervision ablation, these results support the intended feature organization. High cosine alone does not establish semantic correspondence; Appendix E provides all-weather statistics and mismatch controls.

Table 3: **Extended K-Radar v2 evaluation.** Mean AP on the two-class K-Radar v2 setting.Complete class-wise and BEV results are provided in Appendix C.

Method	Sensors	AP_{3D} @0.3	AP_{3D} @0.5	AP_{3D} @0.7	AP_{BEV} @0.5
ASF	C+L+R	64.77	39.70	9.55	58.77
L4DR	L+R	65.53	**43.84**	9.82	60.73
ObjDec	C+L+R	**66.91**	43.05	**11.11**	**60.93**

Spatial gating. Figure 4 relates predicted gates to objects and final detections. Across all 10,065 test frames, foreground/background gate means are 0.270/0.118. Replacing gates with frame-wise means or spatially permuting them reduces detection performance, supporting the importance of their spatial arrangement. These fixed-weight interventions are distinguished from retrained ablations in Appendix D.3; additional scenes and localization failures are provided in Appendix E.

4.5 Evaluation on V2X-Radar-V

Setup. We train on V2X-Radar-V (Yang et al., 2025) with 8,391 training frames and deduplicated validation/test sets of 1,487/1,486 frames. Models use 0.16 m grids, 80 epochs, and effective batch size 8; checkpoints are selected by validation mean 3D AP. We report Moderate AP_{R40} at IoUs 0.7/0.5/0.5 for Vehicle/Pedestrian/Cyclist. The C+L+R comparison shares encoders and detection

Table 4: **V2X-Radar-V detection AP (%).**

(a) Published validation references from the dataset paper (Yang et al., 2025)

Method	Sensors	Vehicle 3D	Pedestrian 3D	Cyclist 3D	Mean 3D
PointPillars (Lang et al., 2019)	L	68.80	38.16	65.24	57.40
CenterPoint (Yin et al., 2021)	L	72.19	50.59	75.26	66.01
PV-RCNN (Shi et al., 2020)	L	79.38	58.83	78.01	72.07
SECOND (Yan et al., 2018)	L	79.65	58.79	79.46	72.63
BEVDepth (Li et al., 2023)	C	15.47	8.51	9.46	11.15
RPFA-Net (Xu et al., 2021)	R	30.44	10.37	11.98	17.60

(b) Local reproduced test AP_{R40} . Means average three classes.

Method	Sensors	Vehicle 3D	Pedestrian 3D	Cyclist 3D	Mean 3D	Mean BEV
ASF	C+L+R	83.54	74.17	82.18	79.96	85.72
ObjDec	C+L+R	84.06	75.94	84.13	81.38	86.27
L4DR	L+R	83.17	73.25	78.31	78.24	82.56
ObjDec-LR	L+R	83.87	74.84	80.24	79.65	83.81

heads;ASF omits ObjDec's added branches and controls, and uses no SCL. L4DR and ObjDec-LR retain their native architectures. Table 4(a) provides published references under a different protocol; comparisons below use only panel (b). Adaptation details are in Appendix G.

Results. ObjDec-LR achieves mean 3D/BEV AP of 79.65/83.81, exceeding L4DR by 1.41/1.25 points, with improvements across all three classes. With C+L+R, ObjDec obtains 81.38/86.27, improving over ASF by 1.42/0.55 points. The three-sensor 3D gain is modest and concentrated in Cyclist; Vehicle and Pedestrian remain slightly lower. These results support applicability after target-dataset training. Appendix G includes complete BEV results, resolution studies, and additional VoD experiments.

5 Conclusion

We presented ObjDec, a multi-sensor 3D detection architecture centered on learning and using shared and modality-specific representations in corresponding local regions. Object-region supervision gives representation decoupling a detection-specific focus, while the learned representations contribute directly to token updates and guide cross-modal interaction through foreground gating, modality contribution control, and object context. Experiments on K-Radar, together with component ablations and representation analyses, support the effectiveness of this design under the reported evaluation settings. Training and evaluation on V2X-Radar-V further demonstrate its applicability to different object categories and sensor combinations. These findings support representation decoupling as a useful organizing principle for local sensor fusion, with object guidance connecting the learned representations to detection.

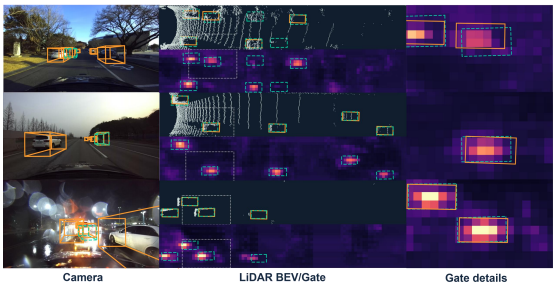

Figure 4: **Predicted gates and detections.** Camera views, aligned LiDAR BEV/gate maps, and gate detail windows. Orange solid/green dashed boxes denote predictions/GT.

AI USE STATEMENT

In this work, Generative AI tools were not used to assist with manuscript editing, refinement of tables and figures, generate experimental results or data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies multi-sensor 3D object detection using publicly available autonomous-driving benchmarks. It does not involve human-subject experiments, collection of personal data, or release of new datasets. While improved perception may benefit autonomous-driving systems, the models evaluated in this work are research prototypes and have not been validated for safety-critical deployment.

REPRODUCIBILITY STATEMENT

Model configurations, training settings, loss definitions, and optimization procedures are provided in the main paper and Appendix A. Evaluation protocols, post-processing settings, and result provenance are described in Appendix B. Additional class-wise and weather-wise results, ablations, representation analyses, and efficiency measurements are provided in Appendices C–H.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.11496.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/45fbc6d3e05ebd93369ce542e8f2322d-Abstract.html.

Yujeong Chae, Hyeonseong Kim, and Kuk-Jin Yoon. Towards robust 3D object detection with LiDAR and 4D radar fusion in various weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15172, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chae_Towards_Robust_3D_Object_Detection_with_LiDAR_and_4D_Radar_CVPR_2024_paper.html.

Yujeong Chae, Heejun Park, Hyeonseong Kim, and Kuk-Jin Yoon. Doppler-aware LiDAR-RADAR fusion for weather-robust 3D detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27197–27208, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chae_Doppler-Aware_LiDAR-RADAR_Fusion_for_Weather-Robust_3D_Detection_ICCV_2025_paper.html.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*, 2020. URL https://arxiv.org/abs/2005.03545.

Xun Huang, Ziyu Xu, Hai Wu, Jinlong Wang, Qiming Xia, Yan Xia, Jonathan Li, Kyle Gao, Chenglu Wen, and Cheng Wang. L4DR: LiDAR-4DRadar fusion for weather-robust 3D object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3806–3814, 2025. doi: 10.1609/aaai.v39i4.32397. URL https://ojs.aaai.org/index.php/AAAI/article/view/32397.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1812.05784.

Hongsheng Li, Lingfeng Zhang, Zexian Yang, Liang Li, Rong Yin, Xiaoshuai Hao, and Wenbo Ding. Weather-conditioned branch routing for robust LiDAR-Radar 3D object detection, 2026. URL https://arxiv.org/abs/2604.05405.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1477–1485, 2023. doi: 10.1609/aaai.v37i2.25233. URL https://ojs.aaai.org/index.php/AAAI/article/view/25233.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/6818dcc65fdf3cbd4b05770fb957803e-Abstract-Conference.html.

Hongwei Lin, Xun Huang, Chenglu Wen, and Cheng Wang. AW-MoE: All-weather mixture of experts for robust multi-modal 3D object detection, 2026. URL https://arxiv.org/abs/2603.16261.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-camera fusion in bird's eye view for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2403.16440.

Ye Liu, Zihan Ji, and Hongmin Cai. Plug-and-play feature causality decomposition for multimodal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/55123f38c9f4acf417335cff41be6e27-Abstract-Conference.html.

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. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. URL https://arxiv.org/abs/2205.13542.

Batyrbek Mukhatbekov, Dong-Hee Paek, Woo-Jin Jung, and Seung-Hyun Kong. SRF: Stereo-Radar Fusion for 3D Object Detection in Adverse Weather Conditions. In *IEEE Intelligent Vehicles Symposium*, 2026. URL https://ave.kaist.ac.kr/2026/02/12/srf-stereo-radar-fusion-for-3d-object-detection-in-adverse-weather-conditions/.

Dong-Hee Paek and Seung-Hyun Kong. Availability-aware sensor fusion via unified canonical space. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/80bd5c815cdb033ac23eb27605adaaba-Abstract-Conference.html.

Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-Radar: 4D radar object detection for autonomous driving in various weather conditions. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.08171.

Heejun Park, Jaeseok Jeong, and Kuk-Jin Yoon. RAF: Reliability-aware fusion of camera, LiDAR, and 4D RADAR for robust 3D object detection in adverse weather. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.04587.

Jaehyun Park, Konyul Park, Daehun Kim, Junseo Park, and Jun Won Choi. Layer-Wise Modality Decomposition for Interpretable Multimodal Sensor Fusion. In *Advances in Neural Information Processing Systems*, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d3e0aaa57c0d639f32a520cda39aec6d-Abstract-Conference.html.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6729,

2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Park_Resilient_Sensor_Fusion_Under_Adverse_Sensor_Failures_via_Multi-Modal_Expert_CVPR_2025_paper.html.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f7f5f501282771c96bb3fedcc96bedfe-Paper-Conference.pdf.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10529–10538, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shi_PV-RCNN_Point-Voxel_Feature_Set_Abstraction_for_3D_Object_Detection_CVPR_2020_paper.html.

Sana Tonekaboni, Viktoria Schuster, and Caroline Uhler. MultiLoReFT: Decoupling Shared and Modality-Specific Subspaces in Multimodal Learning via Low-Rank Representation Fine-Tuning, 2026. URL https://arxiv.org/abs/2607.16789.

Keyao Wang, Shuai Liu, Hengda Shi, Lukui Shi, and Haiyong Chen. Beyond duality: A hybrid framework of leveraging shared and private features for RGB-Event object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Wang_Beyond_Duality_A_Hybrid_Framework_of_Leveraging_Shared_and_Private_CVPR_2026_paper.pdf.

Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/04236.pdf.

Yuchen Wu, Kun Wang, Yining Pan, and Na Zhao. CCF: Complementary Collaborative Fusion for Domain Generalized Multi-Modal 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.23276.

Baowei Xu, Xinyu Zhang, Li Wang, Xiaomei Hu, Zhiwei Li, Shuyue Pan, Jun Li, and Yongqiang Deng. RPFA-Net: A 4D RaDAR pillar feature attention network for 3D object detection. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 3061–3066, 2021. doi: 10.1109/ITSC48978.2021.9564754. URL https://doi.org/10.1109/ITSC48978.2021.9564754.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18268–18278, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yan_Cross_Modal_Transformer_Towards_Fast_and_Robust_3D_Object_Detection_ICCV_2023_paper.html.

Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. doi: 10.3390/s18103337. URL https://www.mdpi.com/1424-8220/18/10/3337.

Lei Yang, Xinyu Zhang, Jun Li, Chen Wang, Jiaqi Ma, Zhiying Song, Tong Zhao, Ziying Song, Li Wang, Mo Zhou, Yang Shen, Kai Wu, and Chen Lv. V2X-Radar: A multi-modal dataset with 4D radar for cooperative perception. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/a501f3238029713afdad57ce7924667a-Abstract-Datasets_and_Benchmarks_Track.html.

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. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yin_Center-Based_3D_Object_Detection_and_Tracking_CVPR_2021_paper.html.

Jingtong Yue, Zhiwei Lin, Xin Lin, Xiaoyu Zhou, Xiangtai Li, Lu Qi, Yongtao Wang, and Ming-Hsuan Yang. RobuRCDet: Enhancing robustness of radar-camera fusion in bird's eye view for 3D object detection. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/21dabaacda3edba8bb281da45d7cbc17-Abstract-Conference.html.

A Implementation and Training Details

A.1 Representation dimensions and foreground targets

This section specifies tensor dimensions and supervision targets omitted from Section 3. Network settings are listed in Table 5. Both the shared and modality-specific branches follow LayerNorm–Linear–GELU–Linear–LayerNorm, with separate parameters for each modality. The foreground and modality-scoring heads use a hidden dimension of 128, while the single-class objectness head uses 160. The final modality-scoring layer is initialized to zero, giving equal initial modality contributions. For v1, the final foreground layer has zero weights and a bias of -1.2. Context projections have no bias and use an initial weight standard deviation of 0.015.

For a BEV patch center $p = (p_x, p_y)$ and a ground-truth box $b = (x_b, y_b, l_b, w_b, \theta_b)$, we rotate their relative displacement into the box coordinate system and assign the foreground target according to Equation equation 8. The margin ϵ extends each local box axis by 0.7 m . Membership is determined by the patch center rather than patch–box area overlap. Binary foreground labels form the union of valid box footprints. For multiclass context, overlapping labels are overwritten in GT iteration order, retaining the last assigned class. Padded boxes and boxes with nonpositive dimensions are ignored.

$$\begin{aligned} a_{pb} &= (p_x - x_b)\cos\theta_b + (p_y - y_b)\sin\theta_b, \\ b_{pb} &= -(p_x - x_b)\sin\theta_b + (p_y - y_b)\cos\theta_b, \\ y^p &= \mathbf{1}[\exists b : \ |a_{pb}| \le l_b/2 + \epsilon, \quad |b_{pb}| \le w_b/2 + \epsilon] \, . \end{aligned} \tag{8}$$

Equation equation 8 constructs training targets only. The representation branches and predicted gate are evaluated at every patch, and fusion uses the predicted values during both training and inference. GT outlines in the visualizations are added after prediction. Training labels, predicted gates, and reference overlays therefore belong to distinct processing steps.

A.2 Auxiliary-loss normalization and boundary cases

Let N_+ and N_- denote foreground and background patch counts in the current batch. Foreground gating uses binary cross-entropy with positive weight ω_+ defined in Equation equation 9, averaged over all batch patches. The upper bound is $k = 30$ for the gate and $k = 4$ for single-class objectness on v1. The representation constraints in Section 3, Equation equation 2, are evaluated only at foreground locations. If a batch has no selected foreground, no valid GT, or fewer than two modalities on this training path, the implementation skips the added auxiliary-loss block while retaining the corresponding detection-training path.

$$\omega_+ = \mathrm{clip}\left(\frac{N_-}{\max(N_+, 1)}, 1, k \right) . \tag{9}$$

On V2X, object context is obtained from three-class softmax probabilities through a bias-free linear projection and tanh. Classification cross-entropy is evaluated only on foreground patches, with inverse-frequency class weights computed from foreground counts in the current batch and clipped

Algorithm 1 Forward computation and training supervision

Input: available sensor observations, model variant, and optional training GT.

1. Encode X and project BEV features into spatially corresponding patch tokens t.
2. Compute shared c and modality-specific u for every available patch and modality.
3. Predict foreground gate g and modality contributions from tokens and their shared/specific representations.
4. If the variant uses object context, predict context z from its context head.
5. Construct controlled tokens and queries using the main-text forward equations.
6. Apply local cross-modal attention, the enabled context output residual, and PFT.
7. Reassemble the BEV feature and obtain predictions from the detection head.
8. If training:
9. Compute the detector's training losses using GT.
10. If SCL is enabled, add detection losses on the six controlled-token subsets.
11. Construct foreground/context targets from valid GT boxes using Equation equation 8.
12. If valid foreground exists and at least two modalities are available:
13. Add foreground representation losses, gate loss, and enabled context loss.
14. Return predictions and the total loss with Table 6 coefficients.
15. Return predictions.

The pseudocode specifies data dependencies rather than low-level operator scheduling. Steps 1–7 are shared by training and inference; GT enters supervision from step 9 onward. The target foreground set in step 11 does not replace the predicted gate in step 3. Step 10 follows the SCL implementation in Section A.3 and is skipped on V2X. Losses follow main-text Equation equation 7.

according to the configuration. The separate gate retains foreground/background supervision. Table 6 lists forward-control coefficients and auxiliary-loss weights; effective loss weights are the products of the outer and inner coefficients.

A.3 Optimization and sensor-combination supervision

K-Radar fusion models use pretrained single-modality encoders. Their parameters are frozen, while BatchNorm statistics follow the original training mode. The configuration uses AdamW, an initial learning rate of 10^{-3}, a minimum learning rate of 10^{-4}, weight decay of 0.01, batch size 2, and the cosine schedule of the original training entry point, with an 11-epoch limit. Checkpoint selection is described in Section B.2. On V2X, the shared backbone initializes its camera ResNet-50 (**?**) from ImageNet and all other parameters from untrained initialization. All encoders are trained, using AdamW with an initial learning rate of 10^{-3} and weight decay of 0.01. The 80-epoch schedule uses micro-batches of two and four accumulation steps, giving an effective batch size of eight, followed by cosine decay by optimizer update to 10^{-5} after one warmup epoch.

K-Radar uses ASF's sensor-combination supervision (SCL) (Paek & Kong, 2025). For three-sensor inputs, detection losses from the six one- and two-sensor branches are summed without division by the number of combinations. Subsets are selected from the controlled tokens already computed using the full input and share the modulated query. Gate, modality contribution, and context are not independently recomputed for each subset. The subset branches do not receive the additional context output residual used by the main fusion path. SCL is disabled on V2X. This implementation provides detection supervision on feature combinations and differs from the fixed-weight missing-sensor inference in Section F.

Table 5: **Representation and fusion dimensions.** Input channels are ordered C/L/R. V2X-LR omits the camera branch; both representation branches preserve token dimensionality.

Item	K-Radar v1	K-Radar v2	V2X , 2×2
Input BEV channels, C/L/R	256 / 512 / 768	256 / 512 / 768	64 / 64 / 64
Common projection channels	256	256	64
Token dimension	256	256	128
Shared/specific MLP hidden dimension	256	256	128
Patch size, BEV cells	2×2	2×2	2×2
Queries per patch	32	32	4
Attention heads	16	16	8
Reassembled fusion channels	2048	2048	128
Gate / modality-score hidden dimensions	128 / 128	128 / 128	128 / 128
Object-context hidden dimension	160	—	160
Context target and activation	Objectness, sigmoid	—	3 classes, softmax
Query/output context modulation	Enabled	—	Enabled
Sensor-combination supervision	Enabled	Enabled	Disabled

Table 6: **Control and supervision coefficients.** Symbols follow the method equations; effective loss weights multiply outer and inner coefficients. Strong omits the context path. Foreground-class weighting is described in Section A.2.

Item	v1	v2	V2X
FG margin ϵ, m	0.7	0.6	0.7
Specific similarity margin δ	0.1	0.1	0.1
Modality scaling strength η	0.75	0.65	0.75
Residual strength λ	0.08	0.08	0.08
Temperature τ	1.0	1.0	1.0
Modality scale clipping interval	[0.4, 2.1]	[0.45, 2.0]	[0.4, 2.1]
Gate range	[0, 1]	[0, 1]	[0, 1]
Gate initial bias	-1.2	-1.5	-1.2
Query / output context strength, β/γ	0.16 / 0.06	—	0.16 / 0.06
Outer auxiliary weight, λ_{aux}	0.12	0.10	0.12
Inner common / specific / separation weights	0.12 / 0.015 / 0.10	0.12 / 0.015 / 0.10	0.12 / 0.015 / 0.10
Inner gate / context weights	0.15 / 0.30	0.12 / —	0.15 / 0.30
Effective common / specific / separation weights	0.0144 / 0.0018 / 0.012	0.012 / 0.0015 / 0.010	0.0144 / 0.0018 / 0.012
Effective gate / context weights	0.018 / 0.036	0.012 / —	0.018 / 0.036
Gate positive-weight cap	30	30	30
Objectness / foreground-class weight cap	4	—	4
SCL coefficient	1.0	1.0	0

B Evaluation Protocols and Model Selection

B.1 AP computation and postprocessing

Additional evaluation settings are listed in Table 7. K-Radar v1 uses the legacy AP aggregation, selecting every fourth value from 41 precision samples, for 11 values in total. The revised v2 implementation averages all 41 values and uses its corresponding height-center convention. Both are distinguished from the standard AP_R40 aggregation used on V2X. K-Radar's reported confidence threshold is an additional prediction-export filter applied after the detection head's score prefilter. Thus, confidence 0.0 does not remove the head-level prefilter.

B.2 Result provenance and checkpoint selection

The K-Radar tables distinguish archived results associated with official weights, local evaluation of official weights, and local training. On v2, we uses pretrained encoders, while local L4DR trains its native L+R network from scratch. These two local runs match the training data, 11 epochs, and effective batch size, while pretraining cost, sensor inputs, and architecture differ.

The full v1 model and ablated variants undergo checkpoint selection using a 1,000-sample preliminary evaluation before full-set evaluation. The local v2 training comparison uses final-epoch weights.

On V2X, the deduplicated validation set is evaluated at epoch 1 and every fifth epoch thereafter. The best checkpoint is selected by mean Moderate 3D AP at the strict class-specific IoUs, retaining the earlier epoch in a tie. After training, best and epoch-80 last checkpoints are evaluated on both raw and deduplicated validation and test sets. Test performance is not used for selection. Small numerical differences can occur between training-time validation and final independent reevaluation; their provenance is retained, and reevaluated scores are not substituted into training curves.

C Complete Results on K-Radar

C.1 Confidence thresholds and stricter matching

Table 9 extends the main comparison to IoU 0.7 and the additional confidence setting. At confidence 0.0, ObjDec and the official ASF archive obtain 88.06 and 87.34 for 3D@0.3, a smaller difference than at confidence 0.3 in the main text; their 3D@0.5 scores are 72.83 and 72.95. Rankings and margins for fixed weights can change with export thresholds and IoU, so operating points are reported separately. ObjDec exceeds official ASF at the stricter 3D@0.7 criterion under both confidence settings, while its BEV@0.7 is slightly lower.

C.2 Weather groups and aggregation

Tables 11 and 12 provide the complete weather breakdown. Total AP is computed from the precision–recall sequence over the entire set and cannot be reconstructed by averaging weather APs. A dash marks a class–weather group with no corresponding GT. In particular, v2 Fog contains no Bus/Truck objects and is excluded from weather averages or gain/loss counts for that class.

On v1, rainy-weather 3D@0.5 is 74.92 versus 67.33 for ASF, whereas sleet yields 57.60 versus 67.28. On v2, performance differences also vary across class–weather groups: at IoU 0.3, ObjDec obtains 94.46 versus 92.26 for ASF on Sedan in fog.

D Component Interventions and Additional Metrics

D.1 What each ablation changes

This section specifies the interventions in main-text Table 2. Disabling modality contribution control sets the scaling strength in Equation equation 4 to zero, giving unit modality scaling while retaining representation residuals, predicted gating, and context. Disabling object context removes its supervision and its injection into queries and attention outputs. Disabling decoupling supervision sets only

Table 7: **Additional evaluation settings.** Dataset columns specify evaluation protocols, not matched training conditions. The origin of the v1 1,000-sample selection subset remains to be documented in Section B.2.

Setting	K-Radar v1	K-Radar v2	V2X-Radar-V local adaptation
Evaluated classes	Sedan	Sedan; Bus/Truck	Vehicle; Pedestrian; Cyclist
Test frames	10,065	13,727	1,486 deduplicated; 1,501 raw
Validation frames	Selection subset: B.2	Final-epoch local comparison	1,487 deduplicated; 1,498 raw
x / y / z range, m	[0,72] / [−6.4,6.4] / [−2,6]	[0,72] / [−16,16] / [−2,7.6]	[0,102.4) / [−51.2,51.2) / [−5,3)
Reported difficulty	Project legacy class evaluation	Project revised class evaluation	Moderate
Primary class IoUs	0.3 / 0.5 / 0.7 for Sedan	0.3 / 0.5 / 0.7 for both classes	0.7 / 0.5 / 0.5 by class
Additional loose IoUs	—	—	0.5 / 0.25 / 0.25 by class, in raw reports
AP aggregation	11 values from 41 precision samples	Mean of 41 precision values	AP_R40
3D evaluator z-center parameter	1.0	0.5	Native V2X evaluation convention
Head score prefilter	0.1	0.1	0.001
Additional export confidence filter	0.3; 0.0 supplement	0.3; 0.0 for availability supplement	No K-Radar-style additional 0.3 filter
Effective rotated-NMS IoU	0.01	0.01 for aligned comparison	0.1
Max detections in local V2X postprocessing	—	—	500 per class

Table 8: **v2 evaluator cross-check.** Two-class mean AP at confidence 0.3. Within each checkpoint, predictions, GT, and postprocessing are fixed; only AP aggregation and the 3D height-center convention change.

Checkpoint	Precision values averaged	z-center	Mean 3D@0.3	Mean 3D@0.5
ObjDec	11	1.0	62.01	39.46
ObjDec	11	0.5	63.26	43.66
ObjDec	41	1.0	63.93	40.04
ObjDec	41	0.5	66.91	43.05
L4DR	11	1.0	66.44	46.37
L4DR	11	0.5	70.86	47.75
L4DR	41	1.0	67.43	45.27
L4DR	41	0.5	69.48	48.23

Table 9: **Complete v1 AP and confidence settings.** Each pair uses one checkpoint with different export confidence filters. ASF values come from official checkpoint logs.

Method	conf	3D@0.3	3D@0.5	3D@0.7	BEV@0.3	BEV@0.5	BEV@0.7
ASF	0.3	80.31	67.19	18.85	80.78	80.33	62.85
ASF	0.0	87.34	72.95	18.85	88.59	86.97	62.85
ObjDec	0.3	88.36	67.50	22.04	88.84	88.10	62.63
ObjDec	0.0	88.06	72.83	22.02	88.84	87.18	62.63

Table 10: **Class-wise v2 AP at confidence 0.3.** Strong omits object context. Checkpoint sources and training differences follow Appendix B.2.

Method	Class	3D@0.3	3D@0.5	3D@0.7	BEV@0.3	BEV@0.5	BEV@0.7
ASF	Sedan	74.98	52.09	11.85	77.85	71.70	43.21
ASF	Bus/Truck	58.13	31.06	7.93	67.85	53.21	20.85
L4DR	Sedan	74.49	54.28	14.08	75.13	71.59	48.20
L4DR	Bus/Truck	56.58	33.39	5.56	61.46	49.86	22.50
ObjDec	Sedan	74.58	52.14	12.31	77.32	71.27	44.26
ObjDec	Bus/Truck	59.24	33.97	9.92	65.14	50.59	23.13

Table 11: **Weather-wise v1 3D AP.** Sedan at confidence 0.3; Total is evaluated over the combined set.

Method	IoU	Total	Normal	Over-cast	Fog	Rain	Sleet	Light snow	Heavy snow
ASF	0.3	80.31	79.57	89.89	90.67	80.97	80.20	80.89	71.71
ObjDec	0.3	88.36	87.66	90.39	90.57	88.90	80.42	89.28	71.41
ASF	0.5	67.19	64.56	79.71	79.57	67.33	67.28	77.63	61.61
ObjDec	0.5	67.50	65.98	80.26	79.71	74.92	57.60	78.33	60.51

Table 12: **Weather-wise v2 3D AP at confidence 0.3.** Dashes indicate class–weather groups with no GT.

(a) IoU 0.3

Method	Class	Total	Normal	Over-cast	Fog	Rain	Sleet	Light snow	Heavy snow
ASF	Sedan	74.98	74.14	82.35	92.26	66.37	70.43	88.24	63.61
L4DR	Sedan	74.49	73.12	80.06	92.33	76.27	54.55	82.74	51.99
ObjDec	Sedan	74.58	73.51	84.30	94.46	68.11	72.52	89.56	63.57
ASF	Bus/Truck	58.13	53.28	75.33	—	7.89	59.55	88.26	68.20
L4DR	Bus/Truck	56.58	54.60	82.07	—	7.81	59.89	89.16	54.27
ObjDec	Bus/Truck	59.24	51.63	73.82	—	10.08	57.47	89.56	72.34

(b) IoU 0.5

Method	Class	Total	Normal	Over-cast	Fog	Rain	Sleet	Light snow	Heavy snow
ASF	Sedan	52.09	49.63	53.38	80.55	46.85	44.05	59.25	49.52
L4DR	Sedan	54.28	52.33	49.24	83.44	54.95	41.22	59.28	38.72
ObjDec	Sedan	52.14	49.39	59.62	80.30	48.30	49.13	60.75	50.49
ASF	Bus/Truck	31.06	23.93	47.04	—	2.48	38.83	70.81	32.55
L4DR	Bus/Truck	33.39	32.18	55.89	—	6.60	43.05	69.00	17.00
ObjDec	Bus/Truck	33.97	23.94	49.65	—	5.05	38.48	75.09	39.37

Table 13: **Additional ablation AP on v1.** Confidence is 0.3; metrics already in Table 2 are omitted.

Variant	3D@0.7	BEV@0.7	BEV@0.3
Full ObjDec	22.04	62.63	88.84
w/o modality contribution control	11.08	61.36	80.38
w/o object context	17.10	54.91	80.68
w/o decoupling supervision	16.49	55.71	80.65
w/o foreground gate	19.41	55.49	80.65

Table 14: **Fixed-weight inference interventions on v1.** AP (%) uses all 10,065 frames, C+L+R, and confidence 0.3. Deltas use the paired baseline. Shuffle seeds vary inference permutations, not training; all settings reuse the same encoded inputs.

	3D AP			BEV AP		
Setting	@0.3	@0.5	@0.7	@0.3	@0.5	@0.7
Paired baseline	88.3599	67.4838	22.0603	88.8370	88.1066	62.6532
Frame-mean gate	53.9015	40.2109	7.9593	54.2623	53.1635	37.6025
Shuffled gate (260923)	21.3775	17.5284	4.7356	21.5984	21.0439	16.6446
Shuffled gate (260924)	21.2514	17.4692	6.2176	21.5240	20.9864	16.5998
Shuffled gate (260925)	21.2448	12.9107	3.9612	21.4998	20.9850	16.6030
No query increment	88.3492	67.4711	22.1068	88.8342	88.0989	62.7320
No output increment	88.3599	67.4833	22.0565	88.8358	88.1066	62.6516
No query/output increments	88.3453	67.4742	22.1096	88.8342	88.0980	62.7320
Mean of three shuffles	21.2913	15.9694	4.9715	21.5407	21.0051	16.6158

the shared-alignment, modality-specific similarity, and inter-branch separation loss weights to zero; both representation branches still generate features and participate in forward control. Disabling the foreground gate fixes its value to one and removes its supervision, retaining the other branches.

These definitions distinguish supervision changes, gating amplitude, and the forward paths themselves. In particular, removing decoupling supervision does not remove the shared/specific networks and is not equivalent to Concat.

D.2 Strict 3D matching and BEV results

Table 13 reports 3D@0.7, BEV@0.3, and BEV@0.7, which are omitted from the main ablation table. Disabling modality contribution control reduces 3D@0.7 from 22.04 to 11.08, while BEV@0.7 remains at 61.36, close to the full model's 62.63.

D.3 Fixed-weight spatial-gate and context interventions

Table 14 complements the retrained ablations with inference-time interventions on the final K-Radar v1 checkpoint. Each setting uses the same encoded features per frame. Replacing the gate by its frame-wise mean preserves its mean amplitude but removes spatial variation; permuting its values preserves the per-frame distribution while changing correspondence with local features. The three fixed seeds are repeated permutations, not independently trained models. GT is not used to generate the modified signals.

Frame-mean gates reduce 3D@0.3/BEV@0.5 from 88.36/88.11 to 53.90/53.16, while the three permutations average 21.29/21.01. This supports the trained model's dependence on spatial correspondence.

Jointly removing the query and output context increments changes 3D@0.3 and BEV@0.5 by -0.0146 and -0.0087 percentage points, respectively; all six overall AP changes have magnitudes below 0.08 points. Removing these residual increments retains cross-modal attention and the trained representations.

E Representation Statistics and Additional Visualizations

E.1 Statistical units and weighted PCA

For frame i, modality m, and location p, let x_{im}^p denote a representation and $\mathcal{P}_i$ the GT foreground locations. Equation equation 10 defines the frame-mean vector μ_{im} and the mean matched-location cosine s_i^{mn}. Weather-group cosines average s_i^{mn} equally over frames. This differs both from taking the cosine of frame-mean vectors and from pooling all patches with equal patch weights. All 10,065 analyzed frames contain valid foreground, totaling 642,649 foreground spatial locations. Table 15 separates modality pairs.

$$\mu_{im} = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} x_{im}^p, \qquad s_i^{mn} = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \cos(x_{im}^p, x_{in}^p). \tag{10}$$

Input, Shared, and Specific each have a separately fitted PCA basis. Within each representation type, all modalities and weather groups share that basis. PCA is fitted to μ_{im}, weighting each vector by $1/(|\mathcal{W}|N_w M)$, where N_w is its weather-group frame count and $M = 3$. The leading two eigenvectors of the weighted centered covariance are used, without per-vector normalization or channel standardization. Each weather group and modality therefore contributes equally to fitting. Different representation columns have different bases and scales, so absolute Euclidean distances are not compared across columns.

The seven-weather scatter plot displays up to 400 frames per weather group, including all 383 overcast frames. Displayed frame indices are shared across modalities and representation panels; centroids and density contours use all frames. With the random seed and PCA bases fixed, we do not refit for visual separation, add jitter, or clip outliers. Figure 5 expands the main PCA visualization to all weather groups.

E.2 High-dimensional similarity distributions and weather variation

Table 15 shows that Specific features are not approximately orthogonal for every modality pair: C–R cosines remain around 0.65 across the seven weather groups, while pairs involving LiDAR vary more across groups. This is compatible with the relative organization encouraged by the two branches and does not require modality-specific features to contain no information in common.

Figure 6 and Table 16 further show frame-level similarity distributions. The L–R Shared medians are 0.9632 in normal weather and 0.9662 in heavy snow. For Specific, the 5th–95th percentile range changes from 0.2113–0.4948 to 0.2578–0.7363. Violin width represents empirical density; thick and thin intervals show the 25th–75th and 5th–95th percentiles. These describe across-frame variability, not confidence intervals.

As an aggregation-sensitivity check, equal weighting of sequences gives mean cross-modal Specific cosines of approximately 0.440/0.584 in normal/heavy-snow conditions, compared with 0.446/0.537 under equal frame weighting. Shared features retain high consistency, while the precise Specific difference depends on aggregation weights.

Frame-level correspondence controls. Figure 7 compares cosines of foreground-mean vectors for the same frame and for different-sequence frames within the same weather group, using ten fixed permutation seeds. Matched scores use the same retained anchors. Before centering, Shared C–L, C–R, and L–R cosines are 0.979214/0.970201/0.987697 for matched pairs and 0.979205/0.970225/0.987136 for mismatches. After subtracting each modality's weather-group mean, the matched/mismatched L–R values are 0.148740/-0.014689, while the camera-related matched values remain close to zero.

E.3 Construction of gate and detection visualizations

Main-text Figure **??** presents three scenes with camera predictions, complete LiDAR BEV and gate maps, and gate detail windows. Foreground/background gate means are 0.213/0.128, 0.259/0.115, and 0.288/0.118 from top to bottom. Full maps cover forward $x \in [0, 72]$ m and lateral $y \in [-6.4, 6.4]$ m; detail windows span 14×8.4 m. The gate uses the native 16×90 array and a common

Table 15: **Full-test cosine similarity by modality pair.** Original-space matched-patch cosines are averaged within each frame, then equally over frames in each weather group.

Weather	Frames	Seqs.	FG patches	Shared C–L	Shared C–R	Shared L–R	Specific C–L	Specific C–R	Specific L–R
Normal	4309	17	317581	0.9455	0.9584	0.9621	0.3429	0.6545	0.3410
Overcast	383	2	33918	0.9520	0.9574	0.9676	0.4211	0.6501	0.4111
Fog	1049	6	43909	0.9391	0.9505	0.9647	0.4659	0.6612	0.4572
Rain	1317	8	125257	0.9431	0.9565	0.9586	0.3855	0.6511	0.3767
Sleet	1106	7	41665	0.9584	0.9478	0.9638	0.5112	0.6529	0.4840
Light snow	803	4	36196	0.9470	0.9582	0.9622	0.4787	0.6590	0.4643
Heavy snow	1098	7	44123	0.9525	0.9481	0.9601	0.4954	0.6458	0.4693

Table 16: **L–R similarity quantiles.** Each observation averages matched-patch cosines within a frame; quantiles describe across-frame variability.

Weather	Branch	Pair	Median	5th percentile	95th percentile
normal	common	LiDAR–4D Radar	0.9632	0.9433	0.9776
normal	unique	LiDAR–4D Radar	0.3384	0.2113	0.4948
heavysnow	common	LiDAR–4D Radar	0.9662	0.9141	0.9806
heavysnow	unique	LiDAR–4D Radar	0.4538	0.2578	0.7363

[0,1] scale without smoothing. Camera panels show the selected target's GT, while BEV and gate panels retain all GT references within the displayed region. GT overlays and regional statistics are added after feature-based prediction; no prior detector provides the gate.

Figure 8 separately retains the seven-weather gate examples selected from 21 complete-gate candidates for scene readability, object arrangement, and local responses. During that export, GT fields were withheld at the fusion-module input and used only afterward. One historical candidate with unexplained repeated-inference differences was excluded. Table 17 summarizes these seven appendix examples, not the three main-text scenes. Figure 9 adds the full-test analysis: equally weighted frame-level foreground/background gate means are 0.2701/0.1176, and foreground means exceed background means in 99.06% of frames. Selected examples and full-test statistics have different sampling units and are reported separately.

Figure 10 compares strictly loaded released ASF and final ObjDec checkpoints on identical test frames, using FP32, batch size 1, the same detection postprocessing, and a score> 0.3 filter. Input-tensor hashes and GT were checked for all 20 candidates in the paired visualization pool. Original camera images, sequence-specific calibration, intrinsics, and distortion are retained. Both methods use identical camera crops, complete BEV ranges, and 12×6 m BEV detail windows. The selected objects' maximum geometric 3D IoUs among retained predictions change from 0.572/0.000/0.000 to 0.719/0.476/0.741 from top to bottom.

Figure 11 adds heavy- and light-snow examples: the selected objects' maximum 3D IoUs change from 0.095 to 0.728 and from 0.354 to 0.223, respectively. The latter also has a locally high gate response, illustrating that foreground-related activation is sufficient to determine final box accuracy.

F Sensor Availability and Corrupted Inputs

F.1 Two input interventions

When a modality is removed, only the remaining sensor features are supplied to fusion, and gating, modality contributions, and context are recomputed for that set. Corrupted inputs instead retain the corresponding branch. C* replaces the raw RGB image with a black frame before the original normalization and camera encoder. L* uses a no-return convention: height-compressed sparse LiDAR

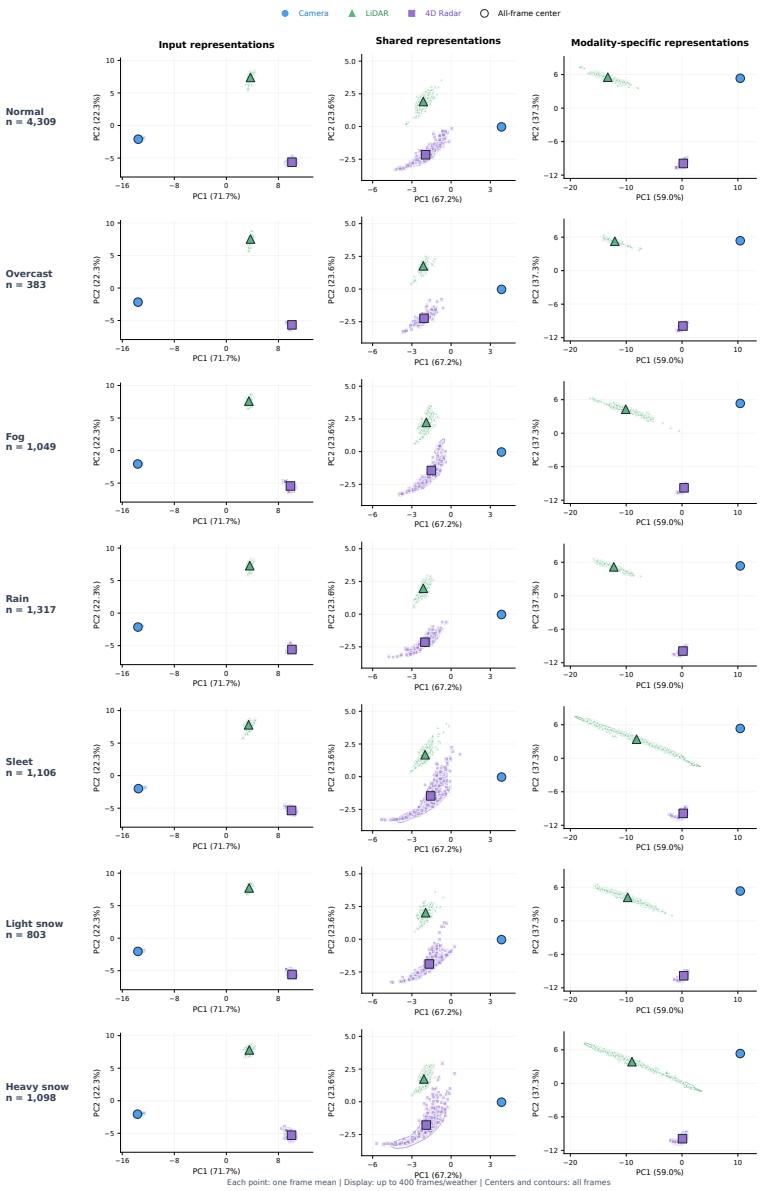

Figure 5: **All-weather frame-level PCA.** Symbols follow Figure 3. Up to 400 frames are displayed per weather group; n, centroids, and contours use all frames. Each column shares one basis and axis range across weather groups; columns are fitted independently.

Table 17: **Statistics for the seven selected gate scenes.** Each frame contains 1,440 patches. GT regions use a 0.7 m margin after prediction. These are individual examples, not weather-group averages.

Weather	Frame ID	GT objects	FG patches	FG gate mean	BG gate mean	FG − BG
Normal	seq20_rdr00628	1	28	0.3154	0.1208	0.1946
Overcast	seq13_rdr00146	2	60	0.2995	0.1112	0.1883
Fog	seq39_rdr00582	1	40	0.2942	0.1078	0.1864
Rain	seq25_rdr00154	4	142	0.2882	0.1181	0.1702
Sleet	seq50_rdr00456	2	65	0.2913	0.1179	0.1734
Light snow	seq43_rdr00169	1	35	0.3711	0.1103	0.2608
Heavy snow	seq55_rdr00385	1	35	0.2817	0.1186	0.1631

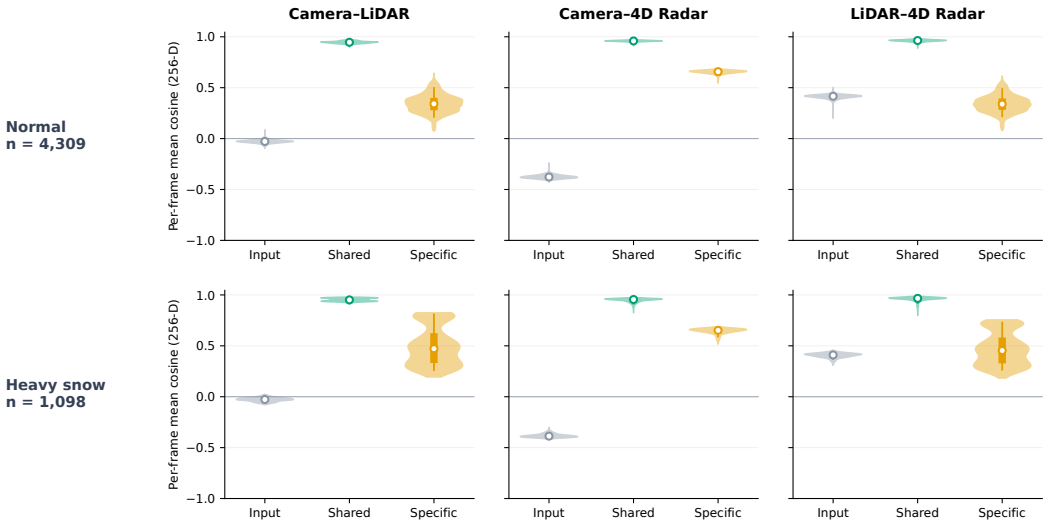

Figure 6: **Original-space cosine distributions.** Each observation averages matched-foreground-patch cosines within one frame. Gray/green/yellow denote input/shared/specific features. Violins show density; white points mark medians, and thick/thin intervals span the 25th–75th/5th–95th percentiles. All normal and heavy-snow frames are included.

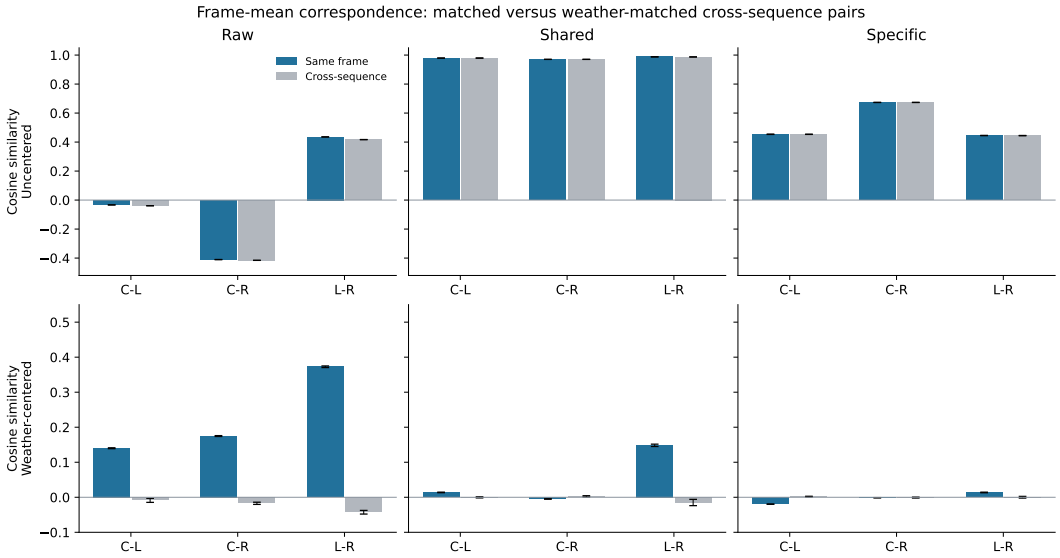

Figure 7: **Frame-level correspondence controls.** Cosines of foreground-mean vectors compare same-frame pairs with different-sequence pairs from the same weather group. Centering subtracts each modality's weather-group mean. Error variation reflects ten mismatch permutations, not training seeds. The statistical unit differs from Figure 6.

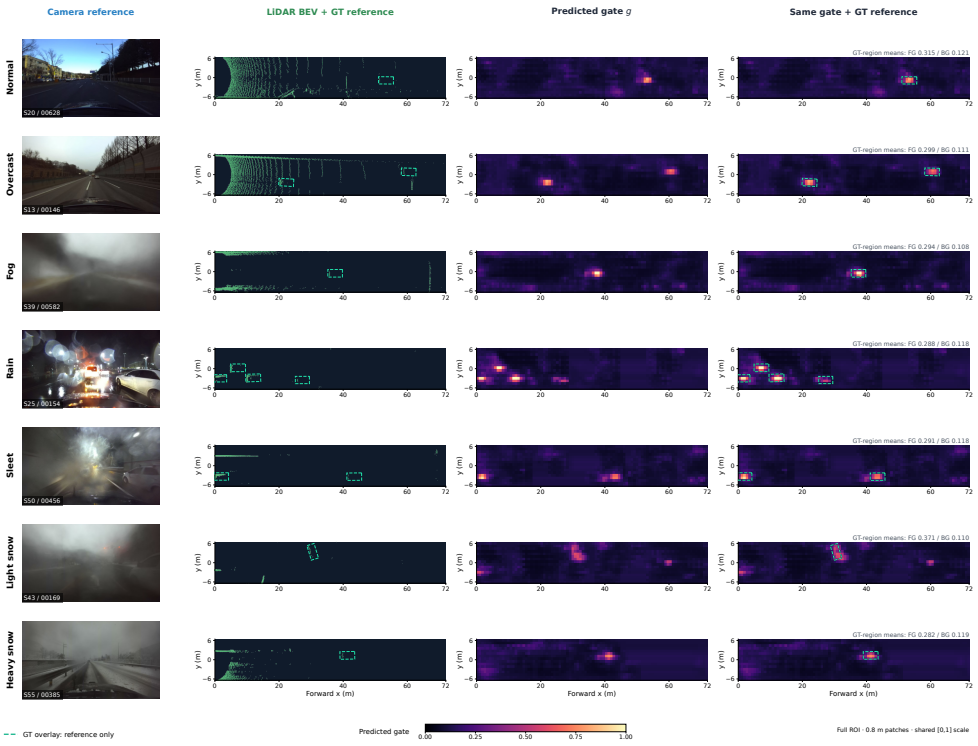

Figure 8: **Gate examples across seven weather groups.** Columns show camera, LiDAR BEV, predicted gates, and gates with post-hoc GT overlays. All maps share the native 0.8 m grid and $[0, 1]$ scale without smoothing. Statistics for these selected scenes appear in Table 17.

Table 18: **v1 input availability with fixed weights.** AP on 10,065 frames using one ObjDec checkpoint. NR denotes an unverified summary. The LR/conf=0.3 row uses audited saved predictions.

Input	conf	3D@0.3	3D@0.5	BEV@0.5	Δ 3D@0.3 vs CLR
C+L+R	0.3	88.36	67.50	88.10	+0.00
C+L+R	0.0	88.06	72.83	87.18	+0.00
L+R	0.3	88.06	71.68	85.45	-0.30
L+R	0.0	NR	NR	NR	NR
C+L	0.3	86.77	72.58	86.35	-1.59
C+L	0.0	86.16	69.74	83.76	-1.90
C+R	0.3	57.62	34.59	55.52	-30.74
C+R	0.0	63.65	38.48	60.28	-24.41
L	0.3	79.58	60.05	77.96	-8.77
L	0.0	85.46	59.23	77.96	-2.60
R	0.3	49.81	26.64	46.80	-38.55
R	0.0	61.75	30.86	52.60	-26.31
C	0.3	0.86	0.16	0.21	-87.49
C	0.0	0.32	0.03	0.14	-87.74
C*	0.3	0.00	0.00	0.00	-88.36
C*	0.0	0.17	0.01	0.07	-87.89
C*+L+R	0.3	88.36	67.41	88.09	+0.00
C*+L+R	0.0	88.06	72.75	87.17	+0.00
C+L*+R	0.3	50.26	34.83	49.27	-38.09
C+L*+R	0.0	63.05	38.48	61.12	-25.01

features are zeroed before the original dense BEV backbone. This avoids sparse operators that do not support empty inputs; its precise interpretation is a no-return simulation at that feature interface.

All ten settings use one fixed checkpoint within each dataset version, without retraining or checkpoint selection per combination. Tables 18 and 19 report the full v1 and v2 , respectively. Unlike the SCL path in Section A.3, which reuses full-input controlled tokens, missing-sensor inference recomputes fusion control from the features actually available.

F.2 Threshold effects, sensor dependence, and the ASF reference

Beyond the complete-input and two-sensor combinations discussed in the main text, Tables 18 and 19 add single-sensor inputs, corrupted inputs, and the second confidence setting. On v2, C+L*+R obtains mean 3D@0.3 of 27.99 at confidence 0.3 and 34.70 at confidence 0.0. Removing LiDAR, C+R, yields 34.25 and 34.73 at those thresholds. Retaining a corrupted branch and removing that branch are not always equivalent across operating points.

Black-frame C*+L+R remains close to complete-input performance, while camera-only predictions are weak. This combination establishes limited sensitivity to that particular camera-input change; a low camera contribution could also produce such behavior. The substantial decline without LiDAR indicates continued sensor dependence. If AP retention is reported, its denominator is complete-input AP for the same checkpoint and threshold; it is not interpreted as object recall.

G Cross-Dataset Architecture Adaptation

G.1 Data processing and the common backbone

We use 8,391 training frames from the public package. The raw validation/test sets contain 1,498/1,501 frames and become 1,487/1,486 after deduplication. Car, Truck, and Bus are merged into Vehicle; the other classes are Pedestrian and Cyclist. ROI, difficulty, and IoU settings are listed

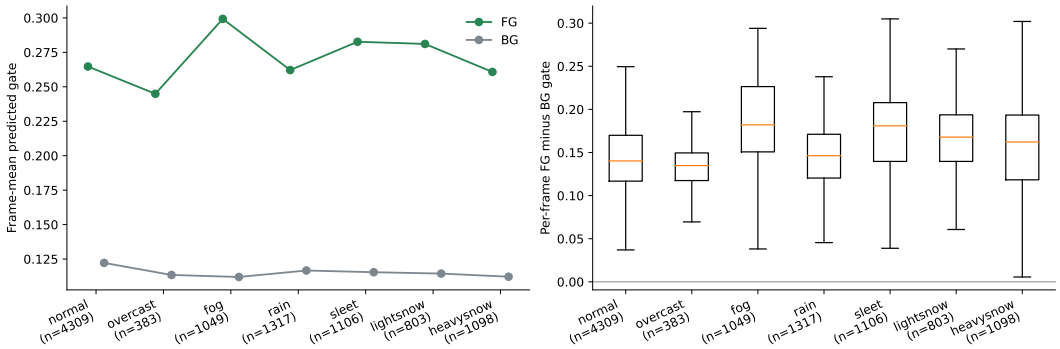

Figure 9: **Full-test gate statistics.** Left: weather-wise foreground/background means with equal frame weights. Right: distributions of per-frame foreground-minus-background means; outliers are hidden. GT defines regions only after gate prediction. The analysis includes all 10,065 test frames.

Table 19: **v2 input availability with fixed weights.** Two-class mean AP on 13,727 frames. Deltas use full inputs at the same confidence; stars denote corruptions defined in Appendix F.

Input	conf	Mean 3D@0.3	Mean 3D@0.5	Mean BEV@0.5	Δ mean 3D@0.3 vs CLR
C+L+R	0.3	66.91	43.05	60.93	+0.00
C+L+R	0.0	69.41	43.81	63.23	+0.00
L+R	0.3	66.02	41.81	59.93	-0.89
L+R	0.0	68.96	42.77	62.74	-0.46
C+L	0.3	64.08	41.17	57.80	-2.83
C+L	0.0	65.13	41.16	58.53	-4.28
C+R	0.3	34.25	16.99	29.97	-32.66
C+R	0.0	34.73	16.82	29.89	-34.68
L	0.3	59.39	32.36	53.45	-7.53
L	0.0	62.33	33.47	54.75	-7.08
R	0.3	30.51	13.59	26.28	-36.40
R	0.0	31.63	13.45	26.70	-37.79
C	0.3	0.01	0.00	0.01	-66.90
C	0.0	0.01	0.00	0.01	-69.40
C*	0.3	0.01	0.01	0.01	-66.90
C*	0.0	0.01	0.01	0.00	-69.40
C*+L+R	0.3	66.93	42.98	60.91	+0.01
C*+L+R	0.0	69.40	43.73	63.19	-0.01
C+L*+R	0.3	27.99	14.11	25.01	-38.92
C+L*+R	0.0	34.70	15.51	30.76	-34.71

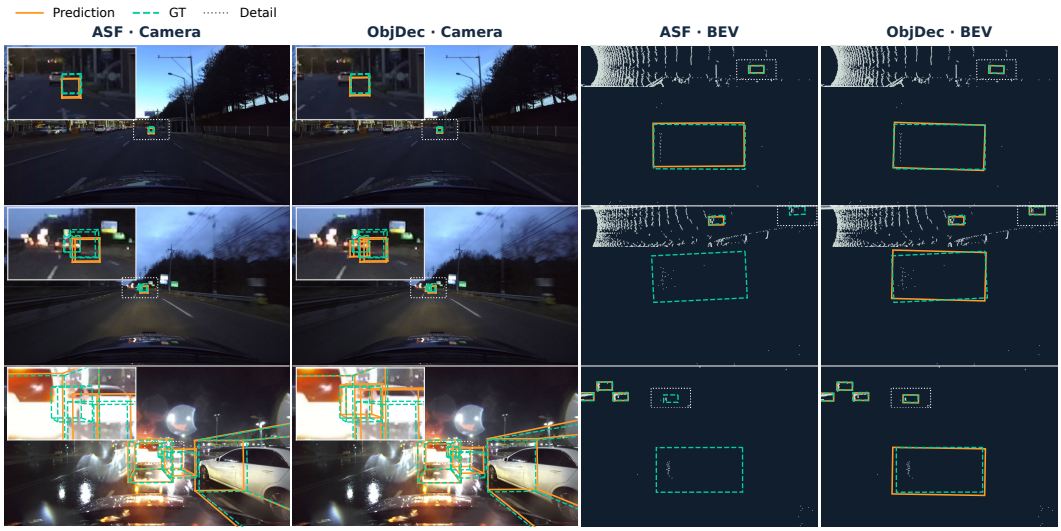

Figure 10: **Paired ASF/ObjDec detections.** Columns show camera projections and LiDAR BEV views with matched detail windows. Orange solid boxes indicate predictions (score > 0.3); green dashed boxes indicate GT. Inputs, postprocessing, and crops are identical. White rectangles identify enlarged regions; additional failures appear in Figure 11.

in Table 7. Local adaptation additionally includes ROI-based GT filtering, calibrated 2D box projection for difficulty filtering, and an intersection-area numerical correction.

The camera branch combines ResNet-50 multiscale features with a learned depth distribution and LSS (**?**) projection to BEV, using 288×512 input images. LiDAR and radar use independent PillarVFE, scatter, and BEV convolutional stems, with four and six point fields, respectively. Each pillar contains at most 32 points, and at most 32,000 pillars are retained. Each modality outputs 64 channels. Fusion produces 128 channels before the common multiscale BEV backbone; the detection head receives 384 channels and uses anchors determined from training-set statistics.

Common encoder and detection-head tensors are copied from the same untrained initialization, without warm-starting from a trained target-dataset checkpoint. Augmentations include global rotation within ± 0.392699 radians, scaling within [0.95,1.05], and flipping with probability 0.5. Camera geometry and point clouds share the same spatial transform. Other training settings are provided in Section A.3.

G.2 Fusion structures and spatial resolution

ASF uses common projection, local attention, and PFT. ObjDec jointly incorporates the representation, gating, modality-contribution, and context paths from Section 3 into fusion. The three-sensor group shares encoders and a detection head. ObjDec-LR is trained with only LiDAR and radar branches from the outset; comparisons with L4DR retain their native architectural differences.

The 0.4 m and 0.16 m settings use BEV grids of 256×256 and 640×640, respectively, with detection-head stride two relative to BEV. The earlier 4×4 patch setting uses 16 queries, and the 2×2 setting uses four; both output 128 channels. Physical patch coverage changes from 1.6 m or 0.8 m at a 0.4 m grid to 0.32 m at a 0.16 m grid. Grid and patch changes therefore alter both computation and the size of local fusion regions relative to objects.

G.3 Supplementary results under the full training budget

The main Table 4 presents published references and local fusion comparisons. Table 20 retains the earlier 0.4 m group, selected epochs, and corresponding validation/test scores omitted from the main table. Table 21 adds per-class BEV and last-epoch results for completed 0.16 m runs. At 0.4

Table 20: **Spatial granularity under an 80-epoch budget.** Moderate AP_{R40}; best checkpoints are selected by validation. Validation and test scores are reported separately.

Method	Sensors	Grid (m)	Best epoch	Val mean 3D	Test Vehicle	Test Pedestrian	Test Cyclist	Test mean 3D	Test mean BEV
ASF	C+L+R	0.4	75	72.76	78.57	58.93	79.02	72.17	78.21
ObjDec 4×4	C+L+R	0.4	80	71.31	78.82	58.75	79.09	72.22	78.76
ObjDec 2×2	C+L+R	0.4	80	74.19	80.73	62.36	78.99	74.03	80.26
ASF	C+L+R	0.16	80	81.03	84.54	76.17	83.18	81.30	85.72
ObjDec 2×2	C+L+R	0.16	70	81.41	84.06	75.94	84.13	81.38	86.27
L4DR	L+R	0.16	80	77.19	83.17	73.25	78.31	78.24	82.56
ObjDec-LR 2×2	L+R	0.16	75	79.12	83.87	74.84	80.24	79.65	83.81

m, changing ObjDec from 4×4 to 2×2 patches increases test Pedestrian 3D AP from 58.75 to 62.36 and mean 3D AP from 72.22 to 74.03. This change also modifies the query count.

All five 0.16 m runs have completed 80 training epochs and final best/last evaluations. Validation-selected three-sensor ObjDec at epoch 70 obtains mean test 3D/BEV AP of 81.38/86.27, compared with 81.55/86.35 at epoch 80. ObjDec-LR obtains 79.65/83.81 at its selected epoch 75 and 79.75/83.80 at epoch 80. Last-checkpoint test differences do not change the validation-based selection rule. ASF and L4DR both select epoch 80. Relative to L4DR, ObjDec-LR improves class-wise 3D AP by 0.71/1.58/1.93 points and BEV AP by 0.29/1.73/1.73 points. Relative to ASF, three-sensor ObjDec improves Cyclist 3D and BEV AP by 0.95/1.21 points, with small decreases in Vehicle 3D and BEV AP.

G.4 Additional architecture adaptation on VoD

Setup. We supplement the main V2X-Radar-V evaluation by training a LiDAR–4D-radar version within a native PointPillars (Lang et al., 2019) detection backbone on VoD (**?**), using 5,139 training and 1,296 validation frames. Inputs comprise one LiDAR frame and five radar frames. Both recent ObjDec runs use a 0.16 m grid, a 320×320 BEV, 2×2 patches, and four queries. Each modality has two 64-channel 3×3 Conv–BN–ReLU layers for local encoding, followed by token LayerNorm. Both models are trained from scratch for 80 epochs with seed 666, FP32, Adam OneCycle, and a peak learning rate of 0.003. A microbatch of eight with two accumulation steps gives an effective batch size of 16. Full validation is performed every five epochs; checkpoints are selected by official EAA mean AP and used unchanged for DC and other metrics. The score and NMS thresholds remain 0.1 and 0.01.

Results. Table 22 (a,b) use official VoD region/class filtering and 11-point 3D AP, with IoU thresholds of 0.5/0.25/0.25 for Car/Pedestrian/Cyclist. EAA and DC denote the Entire Annotated Area and Driving Corridor; both report 3D detection. Table 23 additionally reports KITTI AP_R40 for the same checkpoints.

H Efficiency Measurement and Scope

H.1 Paired timing protocol

Timing uses an RTX A6000, batch size 1, four PyTorch CPU threads, FP32 tensors, and no autocast. The original environment's matmul/cuDNN TF32 settings remain enabled; cuDNN deterministic is enabled and benchmark is disabled. The PyTorch version is 1.10.1+cu113. We sample 120 test indices uniformly, using the first 20 for warmup and the remaining 100 for timing, and repeat the same frames in two rounds.

Each frame's CPU input is copied for the eager and Graph paths, alternating their order and synchronizing each measurement. Model order is ObjDec, ASF, ASF, ObjDec. Timing covers the

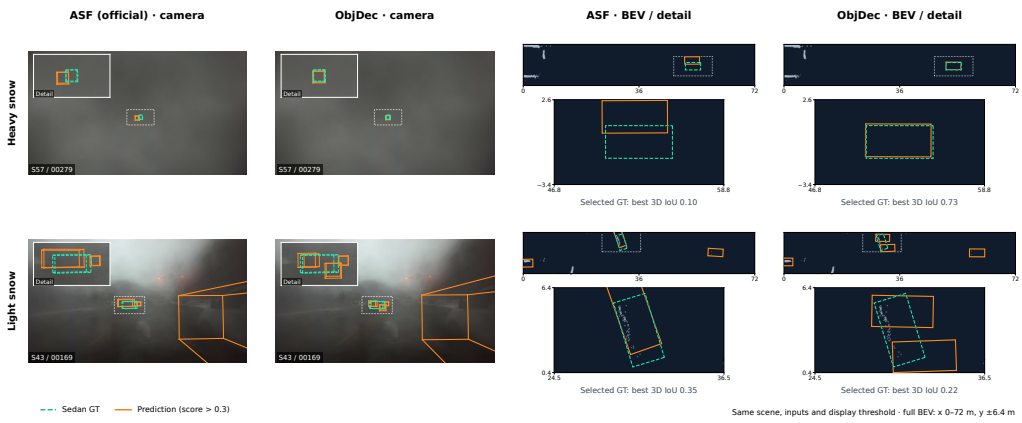

Figure 11: **Improvement and localization counterexample.** ASF and ObjDec use identical inputs and crops. Green dashed/orange solid boxes denote GT/predictions (score > 0.3). Selected-object maximum 3D IoU improves in heavy snow but decreases in light snow despite a high foreground gate response; these overlaps are not AP.

Table 21: **BEV and last-checkpoint results at 0.16 m.** Deduplicated test AP; best is validation-selected, and last is epoch 80.

Method	Checkpoint	Epoch	BEV Vehicle	BEV Pedestrian	BEV Cyclist	Mean BEV	Mean 3D
ASF	best	80	90.95	80.41	85.80	85.72	79.96
ASF	last	80	90.95	80.41	85.80	85.72	79.96
ObjDec 2×2	best	70	90.91	80.88	87.02	86.27	81.38
ObjDec 2×2	last	80	90.84	81.16	87.07	86.35	81.55
L4DR	best	80	89.74	76.60	81.35	82.56	78.24
L4DR	last	80	89.74	76.60	81.35	82.56	78.24
ObjDec-LR 2×2	best	75	90.02	78.34	83.08	83.81	79.65
ObjDec-LR 2×2	last	80	90.10	78.21	83.09	83.80	79.75

Table 22: **VoD region-based 3D AP.** L+R without synthetic fog; official 11-point AP at class IoUs 0.5/0.25/0.25. Local rows use 1,296 validation frames and the same checkpoint for EAA/DC. Reported rows come from L4DR (Huang et al., 2025); training differences are detailed in Section G.4.

(a) Entire Annotated Area (EAA)

Method	Epoch	Car	Pedestrian	Cyclist	Mean
ObjDec	80	66.93	65.28	83.89	72.03
L4DR	99	68.69	65.35	78.98	71.00

(b) Driving Corridor (DC)

Method	Epoch	Car	Pedestrian	Cyclist	Mean
ObjDec	80	89.44	73.02	88.83	83.76
L4DR	99	90.51	75.13	88.90	84.84

Table 23: **VoD KITTI-difficulty AP_{R40}.** Equal class means for the EAA-selected checkpoints in Table 22, at IoUs 0.5/0.25/0.25. Difficulty filtering and AP aggregation differ from official EAA/DC.

Configuration	Epoch	Metric	Easy mean	Moderate mean	Hard mean
ObjDec	75	3D	83.43	78.64	72.23
ObjDec	75	BEV	85.47	80.55	74.56

Table 24: **Latency distributions.** Each model/path uses 100 frames repeated twice, excluding disk loading. FPS is reciprocal mean latency; standard deviations describe timing variation.

Model	Execution	Mean ms	Median ms	P95 ms	Std ms	1000 / mean ms
ASF	Eager	84.00	82.63	93.26	5.34	11.91
ASF	CUDA Graph	59.30	59.10	64.82	2.83	16.86
ObjDec	Eager	84.69	84.50	89.98	3.01	11.81
ObjDec	CUDA Graph	59.57	59.23	65.93	2.92	16.79

network call, including point-cloud preprocessing, host-to-device transfer, encoding, fusion, decoding, NMS, and GT recall computation in the original evaluation path. Disk access, collation, initial graph capture, initialization, and result verification are excluded. Table 24 adds median, P95, and timing standard deviation.

H.2 Graph replay and numerical checks

CUDA Graph captures only the repeated camera-Swin and fusion computations. Each frame supplies new images and BEV features and executes the captured operators; calibration is read per frame, and predictions are not cached across frames. Inputs with different shapes, sensor sets, or visualization requirements fall back to eager execution according to the implementation conditions.

The paired runs checked 240 forwards for each model. Under identical captured inputs, all capture checks were exact. Across the full paths, all checked tensors were exactly equal in 218/240 cases for each model.