EVISCOPE: EVIDENCE-CONDITIONED SAMPLING FOR LONG-RANGE 3D DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As autonomous driving extends to longer ranges, reliable 3D detection depends on using multiple sensing modalities well. Camera–radar fusion, pairing visual semantics with radar's reach, is an important direction. Query-based camera–radar detectors sample features around query positions. On MAN TruckScenes, recall varies substantially across object-level radar-return counts within the same range band. This motivates testing local radar evidence as an additional conditioning signal. We introduce EVISCOPE, evidence-conditioned query sampling: each query adapts its spatial extent and temporal weights to local evidence. The current frame's returns form a non-learned bird's-eye evidence field, which each query reads to predict a sampling radius and frame weights. It adds $5{,}450$ parameters, 0.009% of the detector, with no latency increase in our timing protocol. The module-free recipe exceeds the published mAP values in our comparison. At eight epochs, single runs reach 26.4 / 39.2 mAP / NDS without the module and 26.7 / 39.9 with it. Across seven matched four-epoch pairs, the module changes NDS by $+0.31 \pm 0.36$ and mAP by -0.27 ± 0.22 (mean $\pm$ SE), leaving its incremental benefit unresolved. An exploratory subset gain fails an internally pre-specified test on two further seeds. The sampling radius tracks local evidence in all four probed runs, yet a large penalty from replacing temporal weights also occurs in a control lacking query-local evidence. These results distinguish learned allocation and inference-time dependence from demonstrated detection gains.

1 INTRODUCTION

Camera–radar fusion combines the semantic detail of vision with radar's range and velocity measurements for long-range 3D object detection (Fent et al., 2024; Nabati & Qi, 2021; Kim et al., 2023b; Lin et al., 2024). Detectors of this kind are query-based: they refine a fixed set of object queries over successive decoder layers and sample image and bird's-eye-view (BEV) features from each query's position (Carion et al., 2020; Wang et al., 2021b; Liu et al., 2022; 2023a). Recent work exploits geometric structure through concentric query initialisation with adaptive sampling (Chu et al., 2025), distance-modulated radar aggregation (Wolters et al., 2024), or object-centred Gaussian representations (Bang et al., 2026). We investigate whether explicit local radar evidence provides conditioning information for query sampling beyond query geometry.

Our analysis of the $86{,}252$ evaluated objects of the MAN TruckScenes validation split (Fent et al., 2024) reveals substantial recall differences across radar-return counts within every range band (Figure 1a). Baseline recall increases from 71.4% for objects with no radar return to 89.6% for those with more than fifteen. Range alone therefore does not fully characterise detection difficulty, which motivates local radar evidence as an additional sampling cue.

We propose EVISCOPE, an evidence-conditioned query sampling framework that adapts the spatial sampling extent and the temporal aggregation weights while keeping the query position as the geometric reference. Current-frame radar returns form a parameter-free BEV evidence field of multiscale return counts and radar measurement statistics. Each query reads this field at its decoded centre and combines the descriptor with its own geometry through a lightweight shared network. A spatial branch and a temporal branch then predict a radial sampling scale and a weighting over the eight fused frames, so queries at similar positions but with different evidence sample differently; a training-only density branch supervises the shared representation. The module leaves the feature

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.

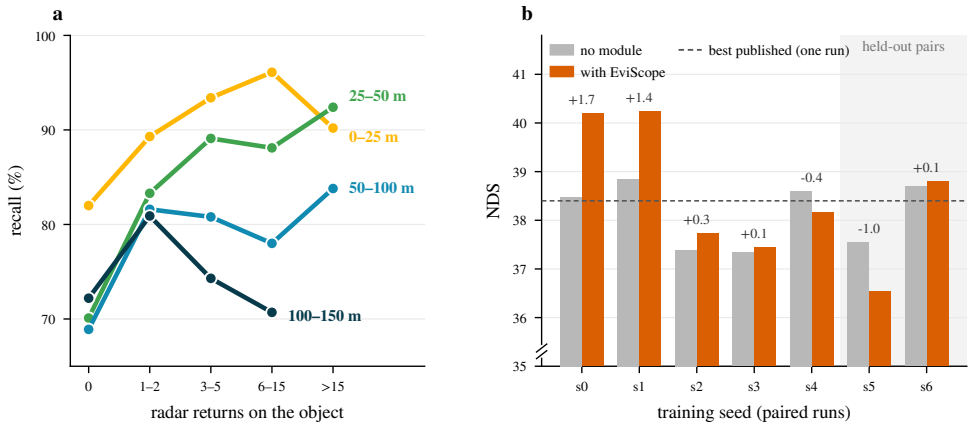


Figure 1: **(a)** Recall of our baseline detector on TruckScenes val by number of radar returns on the object, one curve per range band (Table 19). Recall differs across return-count bins within each band, but is not monotonic in return count; zero-return recall is also not monotonic in range. **(b)** NDS of the same detector with and without EVISCOPE, seven matched seeds, with the paired difference above each pair. The NDS axis starts at 35.

encoder and the detection head unchanged and adds $5{,}450$ parameters, 0.009% of the detector, with no latency increase in our timing protocol.

Our experiments separate the performance of the training configuration from the module's incremental effect. At eight epochs the baseline and EVISCOPE reach 26.4 / 39.2 and 26.7 / 39.9 mAP / NDS, both above the strongest published entry in our comparison (24.1 / 38.4), using a smaller backbone and lower resolution; this is not a comparison at matched compute (Section 4.2). Across seven seed-matched pairs at four epochs the module's paired effect is $+0.31 \pm 0.36$ NDS and -0.27 ± 0.22 mAP (mean $\pm$ SE; Figure 1b). The single eight-epoch pair gives $+0.67$ NDS / $+0.26$ mAP; neither comparison establishes a reproducible benefit. On the seed-0 pair grouped recall rises in 16 of the 19 scored range-by-evidence cells, but the far-range gain tested under a criterion fixed after the first seeds did not replicate on the held-out pairs (Sections 4.3–4.4).

The emitted radial scale tracks local evidence at rank correlations between -0.92 and -0.90 in four probed runs. Replacing the learned temporal weights with uniform ones at inference costs 11.0 ± 2.0 NDS (mean $\pm$ SD), while an evidence-blind control loses 11.75: a substitution penalty alone does not establish the incremental value of query-local evidence (Section 5).

Contributions.

- Grouping the $86{,}252$ evaluated objects of the validation split jointly by range and by radar-return count, we find substantial recall differences across evidence levels within every range band, showing that distance alone does not characterise observation conditions and motivating local radar evidence as a conditioning variable for query sampling (Section 3).
- We propose EVISCOPE, which combines a parameter-free radar evidence field with query geometry to modulate radial sampling extent and temporal aggregation weights jointly, with a training-only density branch supervising the shared representation, at 0.009% of the detector's parameters and no observed latency increase (Section 3.2).
- Seven matched-seed comparisons, an internally pre-specified held-out test and an evidence-blind control separate training-configuration gains from the module's unresolved incremental effect. In this setting, large inference-time substitution penalties do not establish the value of evidence conditioning, and run-to-run variation makes small architectural differences difficult to assess (Sections 4.3–5).

2 Related Work

Long-range camera–radar 3D detection. Camera–radar fusion builds on the camera-only BEV lineage (Philion & Fidler, 2020; Huang et al., 2021; Li et al., 2023; Liu et al., 2022) and is organised around object-level association, BEV fusion, and sparse cross-modal interaction. At the object level, Nabati & Qi (2021) associate radar detections with image objects through a frustum and Kim et al. (2023a) exchange spatial and contextual cues between proposals with cross-attention. In BEV, Kim et al. (2023b) use radar to assist the view transform and absorb the residual misalignment with multi-modal deformable attention, while Lin et al. (2024) align the branches with a radar-specific encoder and cross-modal attention (see also Wolters et al., 2025). At truck range the cost of dense interaction binds, and the entries that reach $150\,\mathrm{m}$ on 4D-radar data (Fent et al., 2024; Palffy et al., 2022; Paek et al., 2022) turn sparse: Wolters et al. (2024) fuse inside sparse frusta with range-adaptive radar aggregation, and Bang et al. (2026) build a sparse representation from primitives at estimated keypoints (see also Li et al., 2025). All of them allocate sampling by position (Section 1).

Modelling the radar return, and reliability-aware fusion. Radar returns are uncertain and the two modalities differ in observation quality (Huang et al., 2025). Long et al. (2025) predict the distribution of returns an object would produce from its class, size and orientation and use it as a post-hoc matching kernel, which presumes at least one hit. Yue et al. (2025) expand radar points into Gaussians from cross-section and velocity priors and gate fusion on camera confidence; Park et al. (2026) supervise a per-pixel reliability map from LiDAR–radar–image feature similarity. Both assume a dependable fallback modality. In our TruckScenes validation analysis, 42.7% of evaluated objects have no annotated radar return.

Query-driven adaptive spatio-temporal sampling. Query-based detectors use their observations through spatial sampling and aggregation across frames. Wang et al. (2021b) project a 3D reference into the multi-view images; Li et al. (2022) fuse past BEV features with temporal self-attention; Lin et al. (2023) propagate sparse instance features recurrently instead of re-sampling every past image. Liu et al. (2023a) generate sampling offsets in space and time from the query feature and mix the sampled features with predicted weights, and Chu et al. (2025) combine polar initialisation with cross-view and cross-modal ray sampling. What these policies condition on is position or the query feature; we examine local radar statistics as an explicit condition on both the spatial scale and the temporal weights.

3 Problem formulation

3.1 Recall by range and radar-return count

For every object the devkit evaluates on TruckScenes val we take the number of radar returns inside its box, which the dataset annotates, and check whether our baseline detector (Section 4.1) places any prediction within $2\,\mathrm{m}$ of its centre. We use recall rather than mAP because mAP on a subset of the ground truth is distorted by false positives that belong to other subsets.

Pooled recall differs across return-count bins. Table 19 in Appendix O gives the numbers behind Figure 1a; its last row pools the ranges. 42.7% of the evaluated objects receive no return at all, and these are recovered at 71.4%, against 89.6% for objects with more than fifteen returns. The pooled relation is monotone across the five bins. On the full trainval annotation set, $774{,}688$ boxes, the zero-return share varies by 57 points across range for cars and by 82 points across classes at 100–$150\,\mathrm{m}$, while the largest spread across any weather or time-of-day tag in a band is 10 points (Appendix L). These are descriptive associations, without adjustment for differences in class or scene composition. At 100–$150\,\mathrm{m}$, 72.6% of cars and 23.4% of trucks carry no return in the current frame.

Differences persist within range bands. In the nearest band, zero-return recall is 82.0%, versus 89.3–96.1% in the nonzero bins. However, recall is not monotonic in return count within any of the four bands. At 100–$150\,\mathrm{m}$, for example, recall in the 6–15 bin is 70.7%, below the zero-return value of 72.2%. These cross-tabulations motivate investigating radar evidence, but do not establish that it predicts detection better than range or adds predictive information after controlling for class, size and other object attributes.

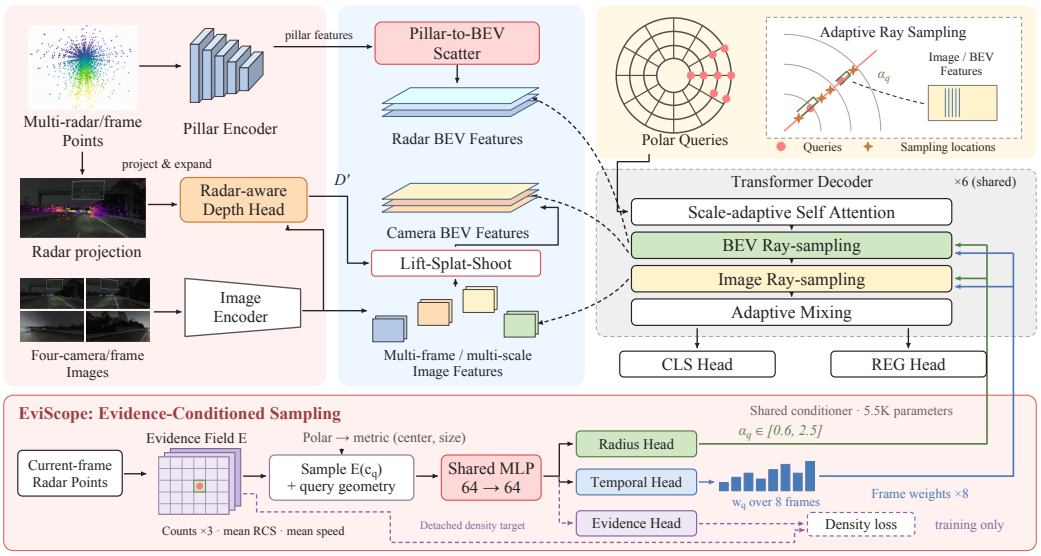


Figure 2: EVISCOPE inside a query-based camera–radar detector. The current frame's returns form a non-learned evidence field; each query reads it at its decoded metric centre, and a small MLP feeds a spatial branch (radial scale, green), a temporal branch (weights over the eight frames, blue) and a density branch (supervised, never read at inference, dashed). The first two set where the shared decoder layer samples for that query; everything downstream is the unmodified detector.

The motivating statistic uses ground-truth boxes, whereas the module below reads a local field at a predicted query centre. The quantities can differ when a query is mislocalised or samples nearby clutter. Their correspondence is not measured here; the object-level association motivates a hypothesis about query conditioning rather than validating that hypothesis.

3.2 EVIDENCE-CONDITIONED QUERY SAMPLING

We test whether current-frame local radar evidence can guide a query's sampling extent and temporal aggregation. Sparse evidence motivates a wider search and a different weighting of past frames, but does not identify which past frames contain useful returns. We implemented this as a small addition to RaCFormer (Chu et al., 2025), shown in Figure 2.

Evidence field. Radar returns for the current keyframe are scattered into a 128×128 bird's-eye grid over the ± 153.6 m detection range. Each cell accumulates point counts at three dilations, mean radar cross-section and mean speed, giving a five-channel field E. The field has no parameters and is recomputed per sample.

Per-query descriptor. The decoder carries queries as rays, $q = [\theta, d, z, \log w, \log l, \log h, \sin\psi, \cos\psi, v_x, v_y]$. We convert each query to its metric centre c_q and linear extent s_q in the same way the detector's own consumers do, read the field there by bilinear sampling, $e_q = E(c_q)$, and concatenate e_q with four geometric terms (normalised range, inverse range, an apparent-size proxy $s_q/(\|c_q\| + 1)$, and $\log(1 + s_q)$). A two-layer MLP with 64 hidden units maps the descriptor to three branches.

Three branches. The *spatial branch* emits a scale $\alpha_q \in [0.6, 2.5]$ applied to the layer's radial sampling extent (cf. Zhu et al., 2021). The *temporal branch* emits a softmax over the $T = 8$ frames, rescaled by T so its mean is one, which reweights the per-frame contributions for that query. It reads current-frame evidence and geometry only, so it learns a policy over frame offsets rather than measuring the evidence in each historical frame. The *density branch* emits a scalar trained to predict the local return density the query sits on, with a binary cross-entropy against a soft target derived from e_q. This is an auxiliary regulariser based on the input counts, not independent supervision; it feeds no prediction and acts on the detector only through the trunk it shares with the other two. All

Table 1: TruckScenes val. **L**: LiDAR, **C**: camera, **R**: radar; bold marks the best camera–radar value in a column, the LiDAR row being a reference. Four-epoch rows are seven-run means (Table 7), eight-epoch rows one run each; the margin over the published entries is present in the module-free baseline rows. *paired* Δ is the within-seed difference, not the difference of the rounded means: the seven-pair mean (± 0.36 NDS, ± 0.22 mAP; Section 4.3) and the single eight-epoch pair. A V2-99 pair is in Appendix D. ‡Error terms not retrieved (Appendix C).

Method	In.	Backbone	Img size	Ep.	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint-V	L	Voxel	—	—	35.3	22.6	0.461	0.405	0.468	3.028	0.261
Far3D	C	V2-99	640×960	—	21.4	10.7	0.883	0.507	0.671	1.352	0.338
RCTrans	C+R	R50	256×864	—	22.9	12.6	1.025	0.507	0.608	0.894	0.331
CRT-Fusion	C+R	R50	256×864	—	28.8	16.9	0.833	0.512	0.444	0.852	0.322
BEVFusion	C+R	R50	256×864	—	30.4	18.2	0.941	0.442	0.389	0.892	0.208
SpaRC	C+R	V2-99	640×960	—	35.4	22.5	0.798	0.449	0.476	0.613	0.248
Horizon3D	C+R	R50	256×864	48	37.4	23.6	0.876	0.410	**0.331**	0.625	0.198
Horizon3D	C+R	V2-99	640×960	48	38.4	24.1	0.833	**0.404**	0.355	**0.561**	0.208
Ours (baseline)	C+R	R50	256×704	4	38.1	25.9	0.655	0.418	0.537	0.747	**0.123**
Ours (EVISCOPE)	C+R	R50	256×704	4	38.4	25.6	**0.650**	0.418	0.519	0.722	0.126
paired Δ				4	$+0.31$	-0.27					
Ours (baseline)‡	C+R	R50	256×704	8	39.2	26.4	—	—	—	—	—
Ours (EVISCOPE)	C+R	R50	256×704	8	**39.9**	**26.7**	0.654	0.408	0.480	0.671	0.129
paired Δ				8	$+0.67$	$+0.26$					

three are initialised so that $\alpha_q = 1$, the frame weights are uniform and the density logit is zero, so training starts from a model that is bit-identical to the baseline.

A control module. An early version of EVISCOPE read the field at the raw polar values (θ, d) instead of the decoded centre. Because these values address one fixed sliver of the grid for every query, the field samples do not represent the radar returns near each decoded centre; the four geometric terms remain query-dependent. The version trained without complaint and scored 26.11 mAP / 39.34 NDS. We keep it as the *evidence-blind control* of Section 5.2: it has the same branches and the same per-query temporal weights, but no correctly aligned query-local evidence. It is an implementation-derived control evaluated at one seed, not a replicated geometry-only ablation. A separate geometry-only row is reported under the shortened one-epoch schedule (Appendix H); it does not resolve the value of evidence under the full schedule.

Cost. One module serves all six shared decoder iterations. It holds $5{,}450$ of the model's 64.0M parameters, 0.009%. Median latency is 965 ms against 997 ms for the baseline build and peak memory is 3.811 GB for both (Appendix J).

4 EXPERIMENTS

4.1 SETUP

We use MAN TruckScenes (Fent et al., 2024) under the official `detection_cvpr_2024` protocol: 12 classes, centre-distance matching at $\{0.5, 1, 2, 4\}$ m, cars scored to 150 m, following the protocol inherited from nuScenes (Caesar et al., 2020). We report the validation split. All models use camera and radar only, with no LiDAR at inference or as depth supervision. The detector is RaCFormer (Chu et al., 2025) with a ResNet-50 backbone (He et al., 2016) at 256×704 input; every row we train uses class-balanced resampling (Zhu et al., 2019), eight temporal frames and four epochs on one node of eight A100 GPUs at an effective batch of eight. We call this configuration without the module the *baseline*, and the same configuration with the module EVISCOPE. Where we report several runs, the runs differ only in the random seed, which sets the initialisation and the data order; the samplers are not deterministic, so two runs at the same seed would also differ.

4.2 THE TRAINING RECIPE

Our ResNet-50 baseline reaches 25.9 mAP over seven runs against 24.1 for the highest published mAP in our comparison, with a standard deviation of 0.45 mAP across our runs. Table 1 places it among the published entries on TruckScenes val, as reported by Bang et al. (2026) (Yin et al., 2021;

Table 2: Recipe ablation on the baseline (no module), ResNet-50. The first row is the seven-run mean of Table 7; the others are single runs. †Without class-balanced resampling an epoch has about a quarter the iterations, so this run was given 12 epochs with the same fraction of training before the multi-frame phase; it still sees about 30% fewer samples and differs from the first row in more than one setting.

configuration	epochs	mAP↑	NDS↑	mAVE↓	runs
CBGS, 8 frames (baseline)	4	25.85	38.13	0.754	7
no CBGS, 8 frames†	12	23.10	35.69	0.713	1
CBGS, 4 frames	4	24.02	36.24	0.844	1
CBGS, 2 frames	4	22.36	33.91	0.957	1

Table 3: mAP / NDS by disjoint range band, official protocol restricted to each band. Published rows are from Bang et al. (2026), Table 9; our rows are the seed-0 models. Bold marks the best value in each column.

	0–25 m		25–50 m		50–100 m		100–150 m	
	mAP	NDS	mAP	NDS	mAP	NDS	mAP	NDS
RCTrans	19.9	28.4	12.5	24.1	7.9	19.9	6.2	15.6
CRT-Fusion	26.2	34.4	16.5	27.3	11.0	26.3	7.0	11.4
BEVFusion	28.8	35.7	16.8	25.2	11.7	20.1	7.4	13.9
Horizon3D (R50)	36.5	43.0	23.3	37.8	**17.0**	**32.7**	8.9	18.4
Ours (baseline)	41.2	44.5	**25.3**	39.0	15.7	29.2	8.6	18.1
Ours (EVISCOPE)	**43.0**	**47.0**	**25.3**	**39.4**	16.8	30.5	**9.4**	**19.2**

Jiang et al., 2024; Li et al., 2025; Kim et al., 2024; Liu et al., 2023b; Wolters et al., 2024). In NDS the four-epoch baseline is at 38.1 against 38.4; repeated runs of the published models would be needed to assess this difference. The error terms are not uniformly better: translation and attribute errors improve over the previous best, scale is close, orientation and velocity are worse.

By range. Table 3 restricts the protocol to disjoint bands. Both seed-0 models exceed the published ResNet-50 entry in the two nearest bands. At 100–150 m, the module arm reaches 9.4 mAP / 19.2 NDS against 8.9 / 18.4, while the baseline reaches 8.6 / 18.1; the published entry leads at 50–100 m. These are single runs.

Training budget. The previous best trains for 48 epochs. Doubling ours to eight, with every other setting fixed, lifts both arms above it on both headline metrics: 26.4 mAP / 39.2 NDS without the module and 26.7 / 39.9 with it. Epoch counts are not a compute-normalised comparison: resampling, iterations per epoch, backbones and input sizes differ, and we do not report matched total iterations or GPU-hours for the published entries. Against the four-epoch mean the module arm gains 1.08 mAP and 1.47 NDS, with orientation and velocity errors falling to 0.480 and 0.671 (Appendix C). These are single runs, and the NDS margin (0.8–1.5) is the size of the spread between four-epoch runs (Section 5.3); the remaining gap to the previous best is in orientation and velocity.

Table 2 examines class-balanced resampling and temporal window length. The no-CBGS run is 2.75 mAP below the baseline mean, but also changes the schedule and sample exposure, so it does not isolate the effect of CBGS. The four- and two-frame runs are lower by 1.83 and 3.49 mAP, respectively, with velocity error rising from 0.754 to 0.957. These single-run results are consistent with benefits from longer temporal context, as also observed in camera-only detectors (Huang & Huang, 2022; Park et al., 2023; Wang et al., 2023). They do not establish whether the benefit comes from radar returns, image context or motion estimation, nor do they quantify each recipe choice's contribution across seeds.

4.3 SEVEN PAIRED RUNS DO NOT SEPARATE THE MODULE FROM ZERO

Figure 1b shows the baseline and EVISCOPE trained seven times each from matched seeds. Two things are visible at once. The seed moves both models: their NDS across seeds correlates at $r = 0.79$, and the spread across seeds (1.5 NDS for the baseline, 3.7 for EVISCOPE) is larger than any within-pair difference after the first two seeds. And the within-pair difference has no consistent sign: the seven paired NDS differences, printed above each pair in Figure 1b and tabulated with the per-

seed scores in Appendix B, average $+0.31$ with a standard error of 0.36 ($t = 0.87$ on six degrees of freedom), and the paired mAP difference is -0.27 ± 0.22. The NDS point estimate is smaller than its standard error.

We report within-seed differences because the seed is a shared source of variation (Picard, 2021; Dodge et al., 2020). With the same seven seeds in both arms, the mean paired difference equals the difference of means; pairing matters for the uncertainty estimate. These estimates do not establish an improvement or equivalence between the models. Recall, a count over $86{,}252$ objects that does not depend on score calibration, is a quieter statistic. Its paired difference over the seven seeds is $+0.12 \pm 0.14$ points ($t = 0.86$).

The first two pairs alone average about $+1.5$ NDS, substantially above the seven-pair estimate and illustrating the sensitivity to which runs are reported.

The module starts at identity and can only act once training has moved it off that point. The reported differences are -3.98 NDS under the shortened one-epoch schedule, $+0.31 \pm 0.36$ at four epochs, and $+0.67$ NDS / $+0.26$ mAP for one pair at eight. Because the one-epoch schedule changes multi-frame activation and the budgets have different numbers of runs, these values do not establish a training-budget trend. For seed 0 specifically, the NDS difference is $+1.71$ at four epochs and $+0.67$ at eight. We had fixed beforehand that one pair cannot settle this and that only a difference above $+1.0$ would justify more pairs (Appendix A); $+0.67$ is recorded as an intermediate result.

4.4 The far-range recall gain does not survive a held-out test

Table 4: Paired recall difference (EviScope minus baseline, points) on objects at 100–150 m with 3–15 returns, the subset where the seed-0 pair gained most (Appendix F). Seeds 0–4 were examined before the test was written; the internally pre-specified criterion was that both held-out differences be positive.

	s0	s1	s2	s3	s4	s5 (held out)	s6 (held out)
Δrecall	$+4.96$	$+0.51$	-2.64	$+3.27$	$+2.32$	-0.67	-0.59

On the seed-0 pair, the module's recall gain was not spread evenly. It was 0.6–1.4 points in every hit bin over all ranges, and 4.3 and 6.2 points for objects at 100–150 m with 3–5 and 6–15 returns (Appendix F). This was an appealing pattern: distant objects with a few returns are where a wider search and a longer temporal window should help most. When we had five pairs, four of them showed a positive difference on this subset, and the mean was $+1.68$ points.

Because we had chosen the subset after looking at the data (Gelman & Loken, 2014; Simmons et al., 2011), we wrote down a confirmation test before computing recall for the two remaining pairs (Nosek et al., 2018). Here "held out" refers to training seeds; all pairs use the same validation objects. The internally recorded protocol is reproduced in Appendix A. The subset was fixed as the two cells merged by count, the matching threshold at 2 m, and the criterion as both held-out differences being positive. Table 4 gives the result. The held-out pairs give -0.67 and -0.59. The seven-pair mean is $+1.02 \pm 0.99$. The gain does not satisfy the pre-specified replication criterion. This limits the interpretation of the exploratory subset, without establishing a zero population effect.

5 Analysis

What the module learns, and what an ablation of it measures, are separate questions. We ask whether it learned what it was meant to learn, what the standard attribution test says about it and whether that test can be trusted, and how much a single run on this benchmark can resolve.

5.1 The module learns the intended allocation

Table 5 reports the allocation the trained model makes. The sampling radius decreases across the local-density bins and increases across the range bins. The temporal window is similar in the first two range bands, rises through 100–150 m, and falls in the final band. The rank correlation of radius with local evidence is $\rho = -0.91$ and with range $+0.51$; these marginal correlations do

Table 5: Forward-hook probe on the trained EVISCOPE model (seed 0), last decoder layer, 270,000 queries over 300 val samples. *radius* multiplies the layer's sampling extent; *window* is $\exp(\mathrm{H})$ of the temporal weights, in frames.

local evidence	radius	range	radius	window
$[0, 0.35)$	0.734	0–25 m	0.627	5.00
$[0.35, 0.7)$	0.629	25–50 m	0.637	4.94
$[0.7, 1.1)$	0.611	50–100 m	0.663	5.22
≥ 1.1	0.602	100–150 m	0.692	5.45
		> 150 m	0.799	5.12

not separate their contributions. The same probe on three further independently trained runs gives -0.90, -0.92 and -0.91. The ordering is in place within the first epoch (Appendix N). These four probes confirm an evidence-associated allocation, while the paired comparisons do not establish an end-to-end benefit. The observed radii fall in $[0.60, 0.84]$, below the initial value of 1: the module shrinks the sampling extent overall and adds an evidence-dependent spread on top of that.

5.2 INFERENCE-TIME DEPENDENCE DOES NOT ESTABLISH INCREMENTAL VALUE

Table 6: Inference-time substitution: each row replaces one modulation on a trained model and re-evaluates the same weights, with no retraining. The first block is mean $\pm$ standard deviation over four ResNet-50 EVISCOPE models; the remaining models are one run each. *recall* is the fraction of evaluated objects matched within 2 m. The further substitutions and the control model of Section 5.2 are in Appendix P.

model	substitution	ΔmAP	ΔNDS	Δrecall
EVISCOPE, 4 runs	radial scale $\rightarrow$ per-frame mean	$+0.00 \pm 0.11$	-0.03 ± 0.18	$+0.06 \pm 0.07$
	temporal weights $\rightarrow$ uniform	-11.4 ± 3.2	-11.0 ± 2.0	-8.2 ± 2.6
EVISCOPE, V2-99	temporal weights $\rightarrow$ uniform	-13.23	-10.45	-9.81
evidence-blind control	temporal weights $\rightarrow$ uniform	-8.71	-11.75	—

A common way to attribute a model's behaviour to one component is to disable that component at inference and measure the drop (Zhou et al., 2018; Morcos et al., 2018; Voita et al., 2019; Michel et al., 2019). We did this for both of the module's outputs. We take a trained EVISCOPE model and, at inference only, replace the radial scale by its per-frame mean, by a shuffled copy of itself, or by a constant, or replace the temporal weights by uniform ones, which is what the baseline uses. Everything else is untouched, and we verify that each intervention fires by counting calls and recording the standard deviation of the radial scale before and after (Appendix K).

The two channels look very different under this test. Table 6 shows the result. Replacing the radial scale by its per-frame mean changes NDS by -0.03 ± 0.18 and recall by $+0.06 \pm 0.07$ points across four models (mean $\pm$ SD). Other radial substitutions also have much smaller observed effects than temporal substitution (Appendix P). Making the temporal weights uniform reduces performance in all four models, by 11.0 ± 2.0 NDS (mean $\pm$ SD), 11.4 mAP and 8.2 recall points on average, and by 10.45 NDS on the V2-99 model, whose end-to-end comparison runs the other way. This contrast measures dependence on the trained outputs, not the benefit of introducing either branch.

Dependence also occurs without query-local evidence. Two further models change the reading. The first is the module trained without the density branch, which removes the auxiliary supervision on the trunk; making its temporal weights uniform still costs 7.34 NDS (Appendix P), so this dependence can arise without density supervision. The second is the evidence-blind control of Section 3.2, whose descriptor lacks correctly aligned local radar information and whose spatial branch, we find, has collapsed to a constant (its standard deviation across queries is 0.000). Making its temporal weights uniform costs 11.75 NDS. This control also depends strongly on its temporal weights; its single run does not establish equivalence to the correctly conditioned module.

What the test measures. The baseline is trained with uniform temporal weights from the start and reaches 38.13 NDS over seven runs. A model trained with learned per-query weights and then evaluated with uniform ones reaches 25–32. This is consistent with co-adaptation to the learned

weights (Hinton et al., 2012); substitution changes the inputs to downstream computation without allowing it to adapt. It does not isolate the value of the evidence signal. Related concerns motivate retraining in feature attribution (Hooker et al., 2019), pruning (Liu et al., 2019) and architecture selection (Wang et al., 2021a). Table 7 estimates the complete module's incremental effect. Isolating the evidence input requires a deliberately constructed geometry-only conditioner trained under the same full schedule and matched seeds, which is not reported here.

Training-time controls. The radial channel could also act during training and be redundant afterwards. Models trained from scratch with every radial scale fixed at 0.6, the value the trained module most often emits, reach 38.02, 37.63 and 37.37 NDS at three seeds, 0.46, 1.22 and 0.02 below their paired baselines. These controls contain no modulation module, so they do not isolate radial adaptation while retaining learned temporal weights. The seed-0 component ablation (Appendix G) is also insufficient to establish each branch's contribution across seeds.

5.3 A SINGLE RUN CANNOT RESOLVE ONE POINT OF NDS

In the configuration of Section 4.1, with the same code, data and initialisation procedure but different seeds, our seven baseline runs span 1.51 NDS (standard deviation 0.67) and our seven EVISCOPE runs span 3.70 (1.39). The two best published entries differ by 1.0 NDS and every published entry is reported as a single run. These numerical comparisons lack the repeated runs needed to resolve small differences statistically.

Radar degradation. Re-scoring the seed-0 pair on identical corrupted inputs tests sensitivity to radar corruption (Appendix Q). Thinning the returns, uniformly or at long range, leaves the within-pair difference between $+0.55$ and $+0.70$ mAP at every rate.

The score is concentrated by scene: highway frames reach 38.1 NDS against 22–25 for city and terminal, so a val mean mixes regimes that differ by more than any effect here (Appendix M).

The observed standard deviation is about twice as large for the module arm. Similar radius–evidence correlations in four probed runs do not locate the source of this variation or establish identical learned policies. The seed-2 EVISCOPE run is two NDS below seed 0 after the first epoch and remains lower at the reported checkpoints (Appendix N).

With a paired standard deviation of 0.96 NDS and an effect of $+0.31$, about 40 pairs would bring the mean to twice its standard error and about 80 would give 80% power (Card et al., 2020), roughly $1{,}200$ GPU-days (Appendix E); we report the seven pairs as they stand.

6 CONCLUSION

We introduced EVISCOPE to test radar evidence as an additional signal for query sampling in camera–radar detection. On TruckScenes, the module-free recipe exceeds the published mAP values in our comparison, while seven matched pairs leave the module's incremental benefit unresolved: $+0.31 \pm 0.36$ NDS and -0.27 ± 0.22 mAP (mean $\pm$ SE). The single eight-epoch pair and four allocation probes do not settle this question, and an exploratory subset gain fails an internally pre-specified test on two further seeds. Large temporal substitution penalties also occur without correctly aligned query-local evidence, so they do not establish the value of that evidence. These findings support evaluating learned allocation, inference-time dependence and paired detection gains separately. A full-budget geometry-only control, a measured link between object-level and query-local evidence, and evaluation beyond this benchmark remain necessary to assess the broader evidence-conditioning hypothesis.

ETHICS STATEMENT

This work uses the public MAN TruckScenes dataset under its research licence; it involves no human subjects and no personal data beyond what the dataset releases with faces and licence plates anonymised. The application domain, perception for autonomous trucks, is safety-relevant. We report sensitivity to simulated radar-sector failure (Appendix Q) and note that the evidence descriptor does not explicitly encode sensor coverage, so that this limitation can be considered before deployment.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the data splits and evaluation settings. Appendix K documents data preparation, hyperparameters, training procedures, and implementation differences between comparison models. Diagnostic definitions and thresholds appear in Sections 3.1, 5.1, and 5.2, with component ablations and additional diagnostics in Appendices G, H, P, and Q. Validation-set results are explicitly identified, and exploratory and held-out analyses are distinguished in Section 4.4 and Appendix A. Training and evaluation code is included in the supplementary material and will be publicly released.

AI USE STATEMENT

In this work, we used generative AI tools only to edit author-written text for grammar, style, and readability. The authors reviewed all AI-assisted revisions and take full responsibility for the final manuscript, including its claims, citations, and experimental reporting.

REFERENCES

Geonho Bang, Geunju Baek, Dongyoung Lee, Wonjun Jeong, and Jun Won Choi. Horizon3D: Sparse radar-camera fusion for long-range 3d perception in autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Houqiang Li, and Yanyong Zhang. RaC-Former: Towards high-quality 3d object detection via query-based radar-camera fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Felix Fent, Fabian Kuttenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, and Markus Lienkamp. MAN TruckScenes: A multimodal dataset for autonomous trucking in diverse conditions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102:460, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

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

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. In *AAAI Conference on Artificial Intelligence*, 2025.

Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3D: Expanding the horizon for surround-view 3d object detection. In *AAAI Conference on Artificial Intelligence*, 2024.

Jisong Kim, Minjae Seong, and Jun Won Choi. CRT-Fusion: Camera, radar, temporal fusion using motion information for 3d object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In *AAAI Conference on Artificial Intelligence*, 2023a.

Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Yiheng Li, Yang Yang, and Zhen Lei. RCTrans: Radar-camera transformer via radar densifier and sequential decoder for 3d object detection. In *AAAI Conference on Artificial Intelligence*, 2025.

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. In *AAAI Conference on Artificial Intelligence*, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

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

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision (ECCV)*, 2022.

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)*, 2023b.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations (ICLR)*, 2019.

Yunfei Long, Abhinav Kumar, Xiaoming Liu, and Daniel Morris. RICCARDO: Radar hit prediction and convolution for camera-radar 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ari S. Morcos, David G. T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations (ICLR)*, 2018.

Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115:2600–2606, 2018.

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 (NeurIPS), Datasets and Benchmarks Track*, 2022.

Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7:4961–4968, 2022.

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 (ECCV)*, 2026.

Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision (ECCV)*, 2020.

David Picard. `torch.manual_seed(3407)` is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359–1366, 2011.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking architecture selection in differentiable nas. In *International Conference on Learning Representations (ICLR)*, 2021a.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning (CoRL)*, 2021b.

Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Felix Fent, and Gerhard Rigoll. SpaRC: Sparse radar-camera fusion for 3d object detection. *arXiv preprint arXiv:2411.19860*, 2024.

Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 (ICLR)*, 2025.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the importance of individual units in cnns via ablation. *arXiv preprint arXiv:1806.02891*, 2018.

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

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.

A INTERNALLY PRE-SPECIFIED HELD-OUT TEST

The far-range low-evidence subset of Section 4.4 was found on the seed-0 pair and examined on seeds 1–4. Before recall was computed for the seed-5 and seed-6 pairs, we fixed the following, reproduced here from our laboratory log with only the wording translated. The held-out units are training seeds evaluated on the same validation split; the description here documents an internal protocol.

> *Origin.* The subset was identified on the seed-0 recall table (Table 10) and has been examined on seeds 1–4; this is exploratory. Recall has never been computed for the seed-5 and seed-6 pairs; they form the held-out set.
>
> *Hypothesis.* EVISCOPE improves recall over the same-seed baseline on objects at 100–150 m with 3–15 radar returns, taking the 3–5 and 6–15 cells merged by object count as one recall rate.
>
> *Criterion.* If the merged recall difference is positive on both held-out pairs, the result is recorded as replicated; if either is ≤ 0, as not replicated. The seven-pair estimate (mean $\pm$ SE) is reported as well, but the conclusion rests on the two held-out pairs only. Overall recall and other cells are not criteria for this test.
>
> *Not permitted.* Changing the cells, the range boundaries, the matching threshold (fixed at 2 m), or dropping any seed. The result is reported either way.

The held-out differences were -0.67 and -0.59 (Table 4). The test was recorded as not replicated.

Training-budget hypothesis. Before the eight-epoch pair of Section 4.3 was trained we recorded the following.

> *Motivation.* The module is initialised at identity and must leave that initialisation before it can act. The paired difference is monotone in budget: -1.66 mAP / -3.98 NDS at one epoch, $+0.31 \pm 0.36$ NDS over seven pairs at four. The published best trains for 48 epochs; we have four.
>
> *Hypothesis.* Raising the budget from four to eight epochs increases the paired ΔNDS above $+0.31$.
>
> *Design.* The two eight-epoch runs are the existing four-epoch configurations of both arms with the epoch count changed from 4 to 8 and nothing else; same hardware, same batch size per device, same seed (0).
>
> *Criterion.* The eight-epoch paired ΔNDS decides. One pair cannot settle the question (run-to-run spread 1.5–3.7 NDS). If $\Delta\text{NDS} > +1.0$, two further pairs are warranted; if ≤ 0, the budget hypothesis is recorded as not supported and no further pairs are trained.
>
> *Not permitted.* Changing the epoch count after seeing the result and reporting selectively; unequal budgets between arms.

The pair gave $+0.67$ NDS / $+0.26$ mAP, between the two thresholds; it is recorded as an intermediate result and no claim about the module rests on it. The motivation above records our hypothesis at the time. The one-epoch experiment also changes the multi-frame schedule, and only one pair was run at eight epochs, so these results do not establish a budget-dependent effect.

B THE SEVEN SEED-MATCHED PAIRS

C TRAINING BUDGET

Table 8 gives the eight-epoch runs of Table 1 in full, next to the seven-run four-epoch means. Both arms are one run. The baseline was evaluated on four GPUs and the module run on eight, on different machines; evaluating one trained model under both set-ups differs by about 0.12 NDS, with the module's machine reading lower. Error terms for the eight-epoch baseline were not retrieved at the time of writing.

Table 7: The ResNet-50 pair at seven seeds, in the configuration of Section 4.1; Δ is EVISCOPE minus baseline within a seed. Seeds 5 and 6 were trained after the held-out test of Section 4.4 was written down.

	s0	s1	s2	s3	s4	s5	s6	mean
mAP								
baseline	25.84	26.48	26.17	25.00	25.86	25.82	25.76	25.85
EVISCOPE	26.47	26.51	25.77	24.42	25.79	24.56	25.52	25.58
Δ	$+0.63$	$+0.03$	-0.40	-0.58	-0.07	-1.26	-0.24	-0.27
NDS								
baseline	38.48	38.85	37.39	37.34	38.60	37.55	38.69	38.13
EVISCOPE	40.19	40.24	37.73	37.44	38.16	36.54	38.80	38.44
Δ	$+1.71$	$+1.39$	$+0.34$	$+0.10$	-0.44	-1.01	$+0.11$	$+0.31$

Table 8: Four- versus eight-epoch training, ResNet-50 at 256×704, with the other training settings fixed. Four-epoch rows are seven-run means (Table 7); eight-epoch rows are single runs evaluated on different GPU configurations, as described above.

model	epochs	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
baseline	4	25.85	38.13	0.655	0.418	0.537	0.747	0.123
baseline	8	26.40	39.24	—	—	—	—	—
EVISCOPE	4	25.58	38.44	0.650	0.418	0.519	0.722	0.126
EVISCOPE	8	26.66	39.91	0.654	0.408	0.480	0.671	0.129

D THE V2-99 PAIR

We trained one pair on the V2-99 backbone. The baseline of that pair is our strongest same-backbone comparison: at 256×704 and four epochs it exceeds the previous best on both headline metrics (Table 1), against a model of the same backbone trained at 640×960 for 48 epochs. Table 9 gives both runs. The within-pair difference, -0.1 mAP and -1.1 NDS, is one pair and falls inside the single-run spread of Section 5.3; the seven-pair estimate of Section 4.3 is the measurement we read for the module.

Table 9: The V2-99 pair at 256×704 and four epochs, seed 0, single runs. Every other setting is that of Section 4.1.

model	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
baseline	26.7	39.0	0.643	0.410	0.506	0.752	0.124
EVISCOPE	26.6	37.9	0.651	0.410	0.530	0.814	0.135

E WHAT IT WOULD TAKE TO RESOLVE THE EFFECT

With a paired standard deviation of 0.96 NDS and an effect of $+0.31$, about 40 pairs would bring the mean to twice its standard error and about 80 would give 80% power (Card et al., 2020), roughly $1{,}200$ GPU-days. That budget is out of proportion to the effect, and we report the seven pairs as they stand.

F EXPLORATORY RECALL GAIN ON THE SEED-0 PAIR

Table 10 is the table on which the far-range subset was found. It is the difference between Table 20 and Table 19 for the seed-0 pair, and it is a single pair; Section 4.4 and Appendix A give the outcome of testing its two largest cells on further pairs.

G COMPONENT ABLATION AT FOUR EPOCHS

Table 11 splits the module at the full schedule, at seed 0. Every row is a single run. The differences among the first three rows are smaller than the run-to-run spread of Section 5.3, and we do not

Table 10: EVISCOPE minus baseline recall, in points, by range band and radar-hit bin, seed-0 pair only, same protocol as Table 19. The >15 cell beyond 100 m holds 23 objects and is omitted. The two bold cells did not replicate on held-out pairs.

range	0 hits	1–2	3–5	6–15	>15
0–25 m	+1.5	−0.6	+0.7	+0.3	+1.7
25–50 m	−0.6	+1.5	−1.1	+0.7	+0.3
50–100 m	+1.1	+1.4	+1.2	+2.1	+1.1
100–150 m	+0.8	+0.5	**+4.3**	**+6.2**	—
all ranges	+0.6	+0.9	+0.8	+1.4	+0.9

interpret them; we include the table because an earlier draft did. The fixed-radius rows are the training-time control discussed in Section 5.2; a third run at seed 2 scored 25.92 mAP / 37.37 NDS.

Table 11: Component ablation at the full four-epoch schedule, seed 0, single runs. The first three rows are cumulative and share every other setting. The last two rows are the baseline with every query's sampling extent fixed at 0.6×, the value the trained module most often emits, and no module.

configuration	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
baseline, no module	25.84	38.48	0.663	0.417	0.522	0.711	0.132
+ radius & window	26.46	38.93	0.662	0.418	0.490	0.750	0.110
+ density head (EVISCOPE)	26.47	40.19	0.636	0.413	0.473	0.665	0.118
control: radius fixed at 0.6×, no module	25.69	38.02	0.660	0.426	0.503	0.780	0.114
same, seed 1	25.45	37.63	0.676	0.434	0.549	0.715	0.136

H SINGLE-EPOCH ABLATION

The single-epoch suite we ran before the four-epoch ablation of Appendix G. All rows use one recipe (CBGS, one epoch, multi-frame from the first step); the column is internally comparable and not comparable to the four-epoch tables. The full module's difference from the baseline and the increment from adding the density head reverse sign between these schedules, while radius plus window improves both reported metrics in both. These single-run results do not support extrapolating component ablations from the shortened schedule.

Table 12: Component ablation under a shortened single-epoch schedule. All rows share one recipe (CBGS, one epoch, multi-frame from the first step) and differ only in the module configuration, so the column is internally comparable; it is not comparable to Table 1, which uses the full four-epoch schedule. *radius* and *window* are the two modulation paths; *evidence input* denotes whether the descriptor sees the evidence field or only its geometric terms.

row	radius	window	evid. head	evidence input	mAP	NDS
baseline (no module)	—	—	—	—	14.20	26.65
radius + window	✓	✓	—	full	**14.95**	**27.58**
radius only	✓	—	—	full	12.13	25.77
window only	—	✓	—	full	10.46	24.33
full module	✓	✓	✓	full	12.54	22.67
full module	✓	✓	✓	geometry only	12.84	24.74

I ABSOLUTE PER-CLASS AP

J EFFICIENCY

Measured on one A100 at batch one with the full eight-frame pipeline, after warm-up, over 50 timed iterations; the two builds differ only in whether the module is attached. Absolute latency includes data preparation on the host and is not comparable to throughput figures reported elsewhere under other protocols; the difference between the two rows is the module's cost.

Table 13: Per-class AP change from adding the module (EVISCOPE minus baseline, same recipe), seed-0 pair only. Per-class differences of this size are within the run-to-run spread of Section 5.3 and we do not read them. Absolute per-class values are in Appendix I.

	car	truck	bus	trail.	o.veh.	ped.	moto.	bike	cone	barr.	anim.	sign
ΔAP	+0.7	+1.9	+1.9	+1.3	−4.1	+2.1	−0.3	+1.2	+2.5	−1.3	0.0	+1.7

Table 14: Per-class AP (%) on TruckScenes val for the rows we train. Rows share the recipe of Table 1; seed 0 unless stated.

	car	truck	bus	trail.	o.veh.	ped.	moto.	bike	cone	barr.	anim.	sign
baseline (R50)	57.8	37.5	0.1	49.0	11.0	10.4	46.6	12.5	43.3	2.9	0.0	39.0
+ radius & window (R50)	58.0	38.0	0.5	49.4	11.3	10.6	46.8	11.3	43.5	9.5	0.0	38.7
EVISCOPE (R50)	58.5	39.4	2.0	50.3	6.9	12.5	46.3	13.7	45.8	1.6	0.0	40.7
baseline (R50, seed 1)	58.2	39.2	5.3	51.3	9.3	9.4	48.3	12.0	41.0	4.5	0.0	39.4
EVISCOPE (R50, seed 1)	58.7	38.3	0.8	51.5	12.0	7.9	47.3	14.8	40.9	6.9	0.0	39.1
baseline (R50, seed 2)	58.0	37.6	1.5	51.0	8.1	10.2	43.7	10.0	40.5	15.1	0.0	38.5
EVISCOPE (R50, seed 2)	58.2	37.4	1.7	49.7	4.2	11.8	50.9	12.7	40.4	1.7	0.0	40.6
baseline (V2-99)	58.2	38.6	0.7	51.1	10.2	12.8	49.3	13.1	44.8	1.1	0.0	40.9
EVISCOPE (V2-99)	59.1	38.3	1.8	51.2	9.6	12.6	46.1	11.8	42.3	5.2	0.0	41.4

K IMPLEMENTATION DETAILS

Detector. RaCFormer (Chu et al., 2025) with a ResNet-50 backbone initialised from nuImages weights, FPN neck, input 256×704, 150 rays $\times$ 6 range clusters $= 900$ queries, six shared decoder iterations with per-layer sampling extents $\{0.08, 0.07, 0.06, 0.05, 0.04, 0.03\}$ of the scene, eight frames (one current, seven past), radar aggregated over five sweeps per frame. Training uses class-balanced resampling (Zhu et al., 2019), an effective batch of 8, AdamW at 2×10^{-4} with cosine decay, four epochs, with multi-frame input enabled from the second epoch. The single-epoch rows of Table 12 enable multi-frame input from the first step and stop after one epoch.

Module. Evidence grid 128×128 over ± 153.6 m; count channels at dilations of 0, 1 and 2 cells, plus mean cross-section and mean speed; descriptor MLP with two hidden layers of 64; radius scale $\alpha_q \in [0.6, 2.5]$ from a sigmoid; frame weights from a softmax over 8 frames scaled by 8; density head trained with binary cross-entropy against $\sigma(4(\bar{e}_q - \tau))$ where $\bar{e}_q$ is the detached local count and τ its batch median.

Activation recomputation. The samplers recompute their activations during the backward pass to save memory. A modulation reached through shared mutable state is re-read at recomputation time, by which point the shared decoder layer has overwritten it with the previous iteration's value, so the backward pass differentiates a function other than the one the forward pass evaluated. We therefore pass the radial scale and the temporal weights to the samplers as explicit tensor arguments, which fixes their values for both passes.

Reading the field at the decoded centre. The decoder carries queries in polar form, $[\theta, d]$ normalised to $[0, 1]$, and the metric centre is recovered by two conversions. The evidence-blind control of Section 3.2 sampled the evidence field at the raw polar values instead, which addresses one fixed sliver of the grid for every query. Its descriptor therefore lacked correctly aligned query-local radar information, yet training converged and the model scored 26.11 mAP / 39.34 NDS at seed 0, between the baseline (38.48) and the corrected module (40.19) at the same seed. When we later hooked its radial head we found that it emits the same value for every query (standard deviation 0.000 across queries), despite retaining geometric inputs. This observation does not identify why the radial head became constant. The probe of Table 5 is the check that the corrected module reads what it is meant to read; we recommend it to anyone reimplementing a module of this kind, because an evidence-blind one trains without complaint.

Single-node training. Every row is trained end-to-end on one node of eight A100 GPUs. We record one run we discarded: a seed-1 model whose first epoch was trained on Turing GPUs (FP32 matrix multiply) and whose remaining epochs were resumed on A100 (TF32) scored 37.92 NDS,

	parameters	of which module	median latency (ms)	peak memory (GB)
baseline	63,998,040	0	997	3.811
EVISCOPE	64,003,490	5,450	965	3.811

against 40.24 for the same seed trained on A100 throughout. When we discarded it we read the gap as a numerics change; the seed-2 pair later showed that a seed alone moves both models by a similar amount (Table 7), so the discarded run may have been an unlucky draw rather than a corrupted one. We keep the rule regardless: no row in this paper mixes hardware or per-GPU batch size within a run. A second discarded run, at 256×864 input, changed both hardware and per-GPU batch size at the same resume point and scored 24.29 / 36.33; it is comparable to no row and is not reported.

Probe. Forward hooks on the last decoder iteration over 300 val samples ($270{,}000$ queries), reporting Spearman rank correlation between the emitted radius scale and, respectively, local return density at the query's decoded centre and the query's range.

Degradation. For each val sample the dropout mask is generated from a random seed derived from the sample token, so the two models in Table 23 receive identical inputs. *random* drops each return independently; *far* drops returns beyond 75 m at the stated rate; *sector* disables two of the six radar sensors.

L DATASET STATISTICS

Table 15 extends the pooled row of Table 19 from the val split to the full trainval annotation set ($774{,}688$ boxes of the twelve official classes). An object is zero-return when no point of the current radar frame falls inside its box. The largest spread across weather or time-of-day tags in any band is 10 points; across range it is 57 points for cars, and across classes at 100–150 m it is 82 points. These marginal summaries do not isolate the effects of weather, class or scene composition.

Table 15: Fraction (%) of annotated objects with no radar return in the current frame, trainval split, against range: by class (top), weather tag (middle) and time-of-day tag (bottom). Cells with fewer than 200 objects are blank; the *all classes* row is the count-weighted mean.

	0–25 m	25–50 m	50–100 m	100–150 m
car	15.7	32.0	51.8	72.6
truck	3.9	8.1	11.1	23.4
bus	0.2	7.7	12.2	15.8
trailer	2.1	6.4	10.1	23.6
other vehicle	0.3	7.9	7.6	24.9
pedestrian	40.1	70.4	86.2	97.9
motorcycle	–	37.8	78.5	93.7
bicycle	34.3	58.3	82.7	–
traffic cone	74.9	89.2	94.8	97.3
barrier	70.9	78.5	71.9	–
animal	–	70.1	80.2	–
traffic sign	56.3	72.3	81.4	87.2
all classes	23.1	40.0	51.8	66.2
clear	22.5	38.9	50.3	64.2
overcast	24.5	43.5	55.3	69.0
rain	22.0	36.3	49.5	66.0
snow	26.0	44.7	48.5	61.8
fog	19.9	38.3	49.3	64.1
morning	21.7	36.9	47.7	61.8
noon	22.7	41.4	54.0	70.0
evening	31.7	43.0	55.7	67.7

M PERFORMANCE BY SCENE TAG

TruckScenes attaches scene-level tags for area, weather, lighting and time of day. Table 16 evaluates the EVISCOPE model (seed 0) on the val frames carrying each tag, under the official protocol restricted to that subset. We show the tags with at least 100 val frames; val has no frame tagged

night, and fog, snow and tunnel carry 40 frames each (two scenes), too few to report. Scene type dominates: highway frames score 38.1 NDS, city and terminal frames 22–25. The weather tags span 28–41 NDS, but tags are not independent of scene type (the rain frames come from a different mix of routes), so the table describes where the model is strong rather than isolating a cause.

Table 16: EVISCOPE (R50, seed 0) on val subsets defined by the official scene tags; n is the number of val frames carrying the tag.

group	tag	n	mAP	NDS	mATE
all	–	2995	26.5	40.2	0.636
area	highway	1960	28.0	38.1	0.645
	city	320	12.4	24.9	0.796
	terminal	320	14.3	22.7	0.849
	rural	315	16.9	25.2	0.790
weather	clear	1560	24.8	35.8	0.677
	overcast	915	28.8	41.0	0.620
	rain	400	19.5	28.0	0.795
lighting	illuminated	1955	29.4	42.1	0.629
	twilight	440	20.1	27.4	0.719
	dark	360	21.9	28.3	0.787
	glare	200	16.1	24.3	0.804
daytime	noon	1515	22.2	34.4	0.692
	morning	1000	34.6	42.3	0.668
	evening	480	22.9	32.7	0.692

Table 17 crosses the same tags with range. The decline with range is present under every tag. Rain and glare cost a roughly constant fraction of the band's mAP at every range (rain 21–35 %, glare 34–40 % below the all-frames value of the band); city frames lose 42 % at 0–25 m and about 72 % beyond. The night-time and tunnel tags are absent or too small to report.

Table 17: EVISCOPE (R50, seed 0), official protocol restricted to the val frames carrying each tag and to objects in each disjoint range band; mAP and NDS (%). Tags with at least 100 val frames.

		0–25 m		25–50 m		50–100 m		100–150 m	
group	tag	mAP	NDS	mAP	NDS	mAP	NDS	mAP	NDS
all	–	43.0	47.0	25.3	39.4	16.8	30.5	9.4	19.2
area	highway	42.5	43.5	35.1	43.2	20.6	31.5	10.6	19.1
	city	24.8	29.7	6.7	14.5	5.0	13.9	2.7	8.4
	terminal	27.0	30.5	12.8	20.0	4.3	10.7	2.7	5.7
	rural	39.9	38.6	16.6	24.2	7.2	17.6	4.9	12.8
weather	clear	40.0	45.4	21.5	31.7	16.2	27.2	11.3	19.8
	overcast	48.7	50.3	38.6	45.7	20.0	33.0	9.7	18.7
	rain	28.0	31.4	19.4	27.6	12.5	18.6	7.4	12.7
lighting	illuminated	46.4	49.2	34.8	44.1	17.3	32.5	10.5	19.8
	twilight	34.7	36.1	15.0	22.9	10.9	19.3	7.5	12.7
	dark	26.8	31.2	22.4	29.0	22.6	28.6	9.5	14.8
	glare	26.8	30.9	16.0	23.2	10.1	16.9	6.2	11.5
daytime	noon	38.6	41.3	19.6	32.6	11.5	23.4	8.1	18.5
	morning	44.9	44.5	31.0	36.5	28.6	34.6	15.3	20.5
	evening	29.3	36.7	20.1	28.3	17.4	26.6	7.7	13.7

N TRAINING TRAJECTORY

Figure 3(a) shows the training loss of the EVISCOPE and baseline runs (seed 0); the module changes neither the level nor the shape of the curve. Figure 3(b) repeats the probe of Table 5 on the model at intermediate iterations over the first 14,410 iterations (1.3 epochs). The correlation between local evidence and radius scale is $\rho = -0.95$ at the first point probed, iteration 2,280, and stays between -0.95 and -0.90 over the window; the final model's value is -0.91 (Section 5.1). The range correlation stays positive and weaker throughout (0.21 to 0.67). The allocation is not a late-training artefact. Table 18 gives the val score at the end of epochs 1, 2 and 4 for seed 0 and for seed 2, the low seed of Table 7; multi-frame input is enabled from epoch 2. Seed 2 is behind from the first

epoch (29.93 against 31.88 NDS) and never closes the gap at the reported checkpoints. These two trajectories do not establish when seed-dependent differences become fixed.

Table 18: EVISCOPE (R50) evaluated at the end of training epochs, seeds 0 and 2.

	seed 0			seed 2		
epoch	1	2	4	1	2	4
mAP	20.03	23.71	26.47	17.81	21.82	25.77
NDS	31.88	36.51	40.19	29.93	34.41	37.73

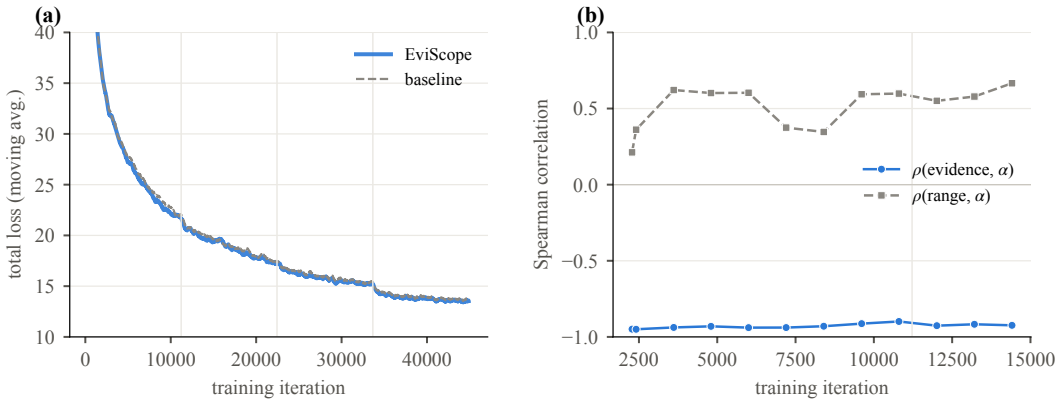


Figure 3: (a) Training loss, moving average over 40 iterations, epoch boundaries marked. (b) Spearman correlation of the emitted radius scale with local evidence and with range, probed on the EVISCOPE model at intermediate iterations of the run.

O RECALL BY RANGE AND EVIDENCE FOR EVISCOPE

Table 19: Recall (%) of our baseline detector (seed 0) on TruckScenes val by range × radar-hit count, official 2 m matching, no score threshold; object counts in parentheses, cells below 30 objects omitted. Hit counts are the official `num_radar_pts` annotation. The last row pools the ranges over all 86,252 evaluated objects.

range	0 hits	1–2	3–5	6–15	>15
0–25 m	82.0 (3,885)	89.3 (4,042)	93.4 (3,455)	96.1 (4,451)	90.2 (1,291)
25–50 m	70.1 (9,950)	83.3 (6,512)	89.1 (3,412)	88.1 (2,542)	92.4 (1,341)
50–100 m	68.9 (14,259)	81.6 (8,688)	80.8 (3,435)	78.0 (2,682)	83.8 (741)
100–150 m	72.2 (8,758)	80.9 (4,243)	74.3 (1,643)	70.7 (899)	—
all ranges	71.4 (36,852)	83.3 (23,485)	85.9 (11,945)	87.5 (10,574)	89.6 (3,396)

Table 19 reports the baseline. Table 20 gives the same cross-tabulation for EVISCOPE, with the number of objects in each cell; Table 10 is the difference of the two. Recall rises in 16 of the 19 cells the table reports and the largest decrease is 1.1 points; the cell beyond 100 m with more than fifteen returns, omitted from the table, decreases by two objects out of 23.

P INFERENCE-TIME SUBSTITUTION, PER RUN

Table 21 gives the absolute scores behind Table 6. Each block is one trained EVISCOPE model, the first four ResNet-50 and the last V2-99; rows within a block share weights and differ only in the substitution applied at inference. Evaluations here use four GPUs rather than eight, which moves an absolute score by up to 0.05 NDS against Table 7; every difference within a block uses the same evaluation configuration. Table 22 gives the two control models of Section 5.2; recall was not computed for them, and the mAP of the radial-mean row was not recorded.

Table 20: EVISCOPE recall (%) by range and by radar returns inside the box (top), and the number of val objects in each cell (bottom); same protocol as Table 19.

	0 hits	1–2	3–5	6–15	>15
0–25 m	83.5	88.7	94.1	96.4	91.9
25–50 m	69.5	84.7	88.0	88.9	92.7
50–100 m	70.0	83.0	82.0	80.2	84.9
100–150 m	73.0	81.4	78.5	77.0	65.2
0–25 m	3,885	4,042	3,455	4,451	1,291
25–50 m	9,950	6,512	3,412	2,542	1,341
50–100 m	14,259	8,688	3,435	2,682	741
100–150 m	8,758	4,243	1,643	899	23

Table 21: Absolute mAP / NDS / recall (%) for every substitution and run.

	run 0			run 2			run 3			run 4			V2-99		
substitution	mAP	NDS	rec.	mAP	NDS	rec.	mAP	NDS	rec.	mAP	NDS	rec.	mAP	NDS	rec.
none	26.48	40.18	80.10	25.77	37.76	79.51	24.38	37.39	79.64	25.79	38.11	79.29	26.55	37.92	80.19
radial mean	26.43	40.01	80.15	25.85	37.98	79.66	24.49	37.22	79.69	25.66	38.12	79.28	26.56	37.86	80.37
radial shuffled	26.44	40.06	80.02	25.70	37.75	79.59	24.43	36.92	79.50	25.70	38.04	79.28	26.53	37.74	80.21
radial constant	26.45	40.14	80.06	25.66	37.55	79.54	24.35	37.41	79.51	25.79	38.18	79.42	26.57	37.86	80.29
temporal uniform	19.60	31.93	75.66	13.68	27.11	71.06	12.07	24.91	69.40	11.37	25.70	69.54	13.32	27.47	70.38

Q RADAR DEGRADATION

Removing radar at evaluation time changes both the detector's features and the evidence field, so it does not isolate the conditioner. Table 23 is the seed-0 pair re-scored on identical corrupted inputs. Thinning the returns, uniformly or preferentially at long range, leaves the within-pair difference between $+0.55$ and $+0.70$ mAP at every rate, including one that costs both models a fifth of their accuracy. Disabling two of the six radar sensors moves it to -0.28. In the light of Table 7 we do not read the sign of a single pair, and the similarity of the differences under thinning does not establish a robustness advantage. A possible limitation is that the descriptor contains no explicit sensor-coverage information: a zero count alone cannot distinguish missing coverage from an observed region with no returns. Whether this ambiguity explains the sector-failure result would require a separate control.

R QUALITATIVE EXAMPLES

Figure 4 shows the radius scale α_q that the trained module emits for the queries of the last decoder layer on three val frames. Queries that land on the carriageway, where returns are dense, contract (dark, small); queries in sparse regions beside the road receive relatively larger scales (light, large). The observed scales remain below the baseline value of one. The pattern is the same in clear, dusk and snow frames.

Table 22: Absolute mAP / NDS for the two control models under inference-time substitution, seed 0. *no density head*: radial and temporal modulation trained without the auxiliary density loss. *evidence-blind control*: the early module that read the evidence field at raw polar coordinates rather than the decoded query centres (Section 3.2).

	no density head		evidence-blind control	
substitution	mAP	NDS	mAP	NDS
none	26.41	38.91	26.02	39.30
radial mean	—	38.81	—	39.25
temporal uniform	19.47	31.57	17.31	27.55

Table 23: Paired radar degradation, four-epoch models. Both models are scored on the same corrupted inputs: the dropout mask is seeded by sample token, so the comparison is per-point paired rather than per-run. Relative drop is measured against each model's own clean score, so the two columns are not affected by the 0.65 mAP the module already holds when nothing is degraded. *sector failure* removes a third of the radar sensors.

	mAP			relative drop	
degradation	EviScope	baseline	Δ	EviScope	baseline
none	26.48	25.83	+0.65	—	—
random drop 25%	26.16	25.46	+0.70	−1.2%	−1.4%
random drop 50%	24.94	24.39	+0.55	−5.8%	−5.6%
random drop 75%	21.11	20.42	+0.69	−20.3%	−20.9%
far-range drop 50%	26.23	25.61	+0.61	−0.9%	−0.8%
sector failure 33%	21.72	22.00	−0.28	−18.0%	−14.8%

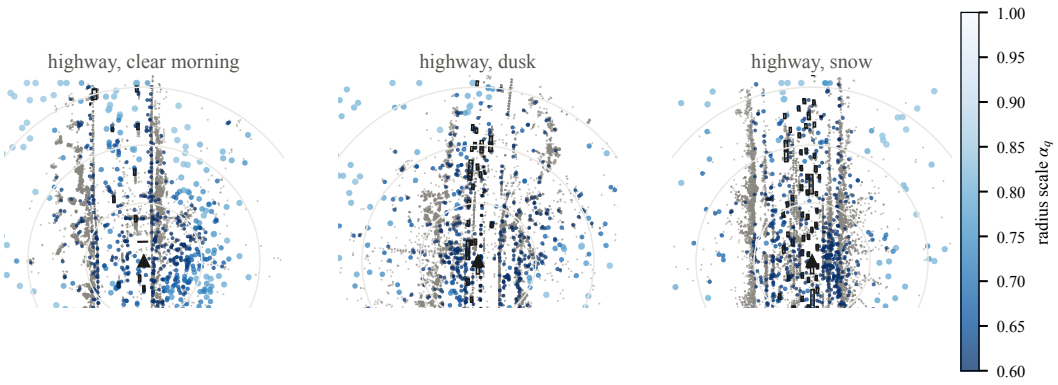


Figure 4: Bird's-eye view; ego at the triangle, forward up, rings at 50, 100 and 150 m. Grey: radar returns. Black: ground-truth boxes. Discs: last-layer queries at their decoded centres, coloured and sized by the emitted radius scale α_q.