OMT-DISTILL: OBJECT-MEASURE TRANSPORT FOR HETEROGENEOUS 3D DETECTION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation enables LiDAR-only 3D detectors to learn from multi-modal teachers. However, differences in encoders, decoders, and detection heads complicate feature alignment and the association of teacher and student predictions. Predicted 3D boxes and class labels have common geometric and semantic meanings, allowing the two prediction sets to be represented as confidence-weighted object measures. We propose Object-Measure Transport (OMT), which combines teacher-location supervision with soft transport over native student objects. Teacher-location supervision extends quality and ordering guidance to responses omitted by student candidate selection. Unbalanced transport distributes supervision through many-to-many object associations, accommodating differences in prediction coverage and confidence. On the nuScenes validation set, OMT-Distill achieves 61.07 mAP and 67.92 NDS with a CenterPoint-Voxel student. OMT-Distill improves detection accuracy across four heterogeneous teacher–student pairs while retaining each student's original LiDAR-only inference architecture.

1 INTRODUCTION

Heterogeneous distillation (Hinton et al., 2015) for 3D object detection aims to transfer knowledge between teacher and student networks with different architectures. We focus on using multimodal teachers trained on images and point clouds to improve LiDAR-only students while preserving the students' original inference architectures (Zhang et al., 2026; Yang et al., 2025; Yin et al., 2021; Lang et al., 2019). However, differences in encoders, decoders, and detection heads make it difficult to align intermediate features and associate predicted objects across the two networks.

Existing methods transfer spatial features and dense responses (Zhou et al., 2023; Zhao et al., 2024b), or establish correspondences through prediction heads and object queries (Wang et al., 2024a; Chen et al., 2023; Wang et al., 2024b). Feature similarity alone does not ensure that the transferred information benefits the student's detection task. Predicted class labels and decoded 3D boxes describe object categories and physical geometry, providing a common basis for comparing teacher and student predictions.

To test whether teacher object information can improve student detections, we replace selected prediction components at evaluation time while keeping both models fixed (Section 3). Teacher candidate locations, classes, and confidence ordering are combined with the student's box regression and post-processing. For a distilled CenterPoint-Voxel student, these substitutions increase nuScenes validation mAP from 60.06 to 67.24. This result motivates learning teacher object information through distillation while preserving the student's original LiDAR-only inference architecture.

Learning teacher object information requires deciding which student responses to supervise and how to distribute supervision across student predictions. Teacher-identified locations may be omitted by student candidate selection, so supervision restricted to selected candidates cannot directly reach the responses at those locations. Teacher and student predictions can also differ in number, location, and confidence, requiring associations that determine how teacher supervision is distributed among student objects.

We propose *Object-Measure Transport* (OMT), which combines *teacher-location supervision* and *soft transport over native objects* (Figure 1). Teacher-location supervision reads student responses at

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.

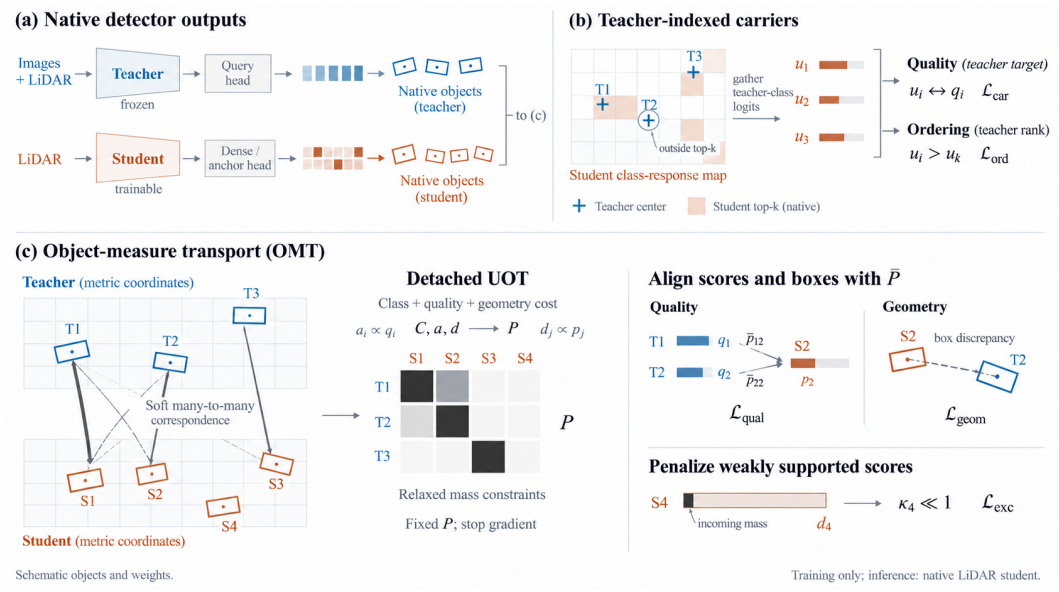


Figure 1: **Teacher-location supervision and native-object transport.** (a) Native heads decode teacher and student objects. (b) Teacher-location supervision reads student logits at teacher centers, including positions outside the native top-k set, for quality and teacher-score ordering supervision. (c) Soft native-object transport uses a detached unbalanced plan. The resulting losses supervise native student scores and boxes and suppress weakly supported predictions. Blue and copper denote teacher and student quantities. Objects and weights are schematic.

teacher centers to learn quality and ordering, including responses outside the selected set. Soft transport distributes quality, geometry, and support supervision among native student objects through many-to-many associations. It uses relaxed mass constraints to accommodate differences in coverage and confidence between teacher and student predictions. Both paths preserve native detector architectures and LiDAR-only inference.

Combining OMT with feature, response, and candidate supervision yields the complete OMT-Distill framework. Our contributions are: **(1)** A controlled diagnostic shows that teacher candidate locations, classes, and confidence ordering can improve detection results while retaining student box regression and post-processing. **(2)** OMT combines teacher-location supervision with unbalanced transport over confidence-weighted object measures, preserving native detector architectures and LiDAR-only inference. **(3)** Controlled comparisons show gains from OMT beyond feature, response, and candidate supervision. OMT-Distill improves detection across all four evaluated heterogeneous teacher–student pairs.

2 Related Work

Feature and response transfer. Cross-modal BEV distillation transfers spatial features and detection responses despite different sensory inputs (Zhou et al., 2023; Chen et al., 2023; Wang et al., 2023; Zhao et al., 2024b). SimDistill equips a camera student with simulated multimodal processing (Zhao et al., 2024a); VeXKD constructs a fusion teacher whose learned BEV queries guide feature masks for a CenterHead student (Ji et al., 2024). FGD and MGD transfer feature structure through attention and masked generation (Yang et al., 2022b;c), while BCKD addresses the classification protocol used for dense-response distillation (Yang et al., 2023).

Object and prediction interfaces. CrossKD feeds intermediate student-head features through a teacher-head suffix (Wang et al., 2024a); KD-DETR creates consistent distillation points for different detection transformers (Wang et al., 2024b). BEVDistill matches teacher boxes to student queries (Chen et al., 2023), and SparseKD studies efficient and cross-stage 3D distillation, with

decoded-box matching in its released implementation (Yang et al., 2022a). OMT combines teacher-location supervision with soft transport over candidates selected by the student's native center or anchor head. These two paths transfer object knowledge without a shared head or distillation-query interface.

Transport of prediction distributions. Optimal transport has been used for model compression and output distillation (Lohit & Jones, 2022; Lv et al., 2024). More directly, local-prediction KD transports confidence-weighted keypoint distributions for 6D pose (Guo et al., 2023), and uncertainty-aware KD adapts prediction transfer to teacher uncertainty (Ousalah et al., 2025). Unbalanced transport also constructs detached prediction-to-GT assignments in object detection (De Plaen et al., 2023). Building on established unbalanced scaling (Cuturi, 2013; Chizat et al., 2018), OMT uses a detached plan to supervise native object quality, geometry, and support, alongside direct supervision at teacher locations.

3 What Can Be Transferred Through Objects?

We first test whether teacher object judgments retain value when used by the student's detection pipeline. Hierarchical Representation-Response-Candidate Alignment (HRCA) is our distillation baseline combining feature, response, and candidate supervision without OMT. We fix a BEVDilation teacher and an HRCA-trained CenterPoint-Voxel student. The reference checkpoint O0 is the HRCA student at 60.06 mAP and 67.53 NDS.

O1 uses teacher locations and classes to read student scores and box regression. O2 adds teacher score ordering, reassigning the same student score multiset according to the teacher's global rank. O3 uses raw teacher scores (O3-raw) or calibrated teacher scores (O3-cal), retaining student box regression. O4 combines teacher boxes with student scores; O5-cal supplies complete teacher objects with calibrated scores. O1–O5 replace the student candidate pool and retain its suppression rule, while O6 uses native teacher output. Calibration is label-free, five-fold, and scene-disjoint. GT is used only for evaluation. Appendix C.1 specifies the operations and all replacement/union variants.

Table 1: **Object interchange on nuScenes validation.** Scores are on a 0–100 scale. G. rec. (gap recovery) denotes the percentage of the O0–O6 performance gap recovered, computed separately for mAP and NDS. O1–O5 replace the student candidate pool and retain student Scale-NMS. These are evaluation-time diagnostics, not distilled student results.

ID	Condition	mAP	NDS	mAP G. rec. (%)	NDS G. rec. (%)
O0	HRCA student	60.06	67.53	0.0	0.0
O1	Teacher locations + classes	62.17	68.91	18.2	19.6
O2	O1 + teacher score ordering	67.24	71.18	62.1	51.9
O3-raw	O1 + raw teacher scores	64.63	70.34	39.5	39.9
O3-cal	O1 + calibrated teacher scores	70.17	72.75	87.4	74.3
O4	Teacher boxes + student scores	62.36	69.93	19.9	34.1
O5-cal	Complete teacher objects	70.48	74.03	90.1	92.5
O6	Native teacher output	71.63	74.56	100.0	100.0

Teacher locations and classes (O1) increase mAP from 60.06 to 62.17. Adding teacher score ordering (O2) further raises it to 67.24 (Table 1). Raw teacher scores (O3-raw) reach 64.63 mAP; calibrated teacher scores (O3-cal) raise this to 70.17 while retaining student box regression. Complete teacher objects (O5-cal) further increase mAP from 70.17 to 70.48.

Teacher locations and relative object judgments therefore carry useful knowledge under the student's detection pipeline. Direct interchange bypasses learning, providing a diagnostic reference for that knowledge. OMT makes it available as teacher-location supervision and soft transport over native student objects. Its batchwise quality targets are distinct from the diagnostic calibration.

4 OBJECT-MEASURE TRANSPORT DISTILLATION

4.1 TRAINING SETUP AND OBJECT INTERFACE

We train a LiDAR student with a frozen multimodal teacher using $\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{HRCA}} + \mathcal{L}_{\text{OMT}}$. The student's encoder, head, task loss, and inference path remain native. The Clean KD base aligns projected BEV features and dense class responses. HRCA extends it with teacher-guided object patches, center responses, candidate ranking, and suppression. OMT adds teacher-location supervision and soft transport over native objects without a learned projection.

For each scene, let teacher objects be $\{(b_i^T, c_i^T, s_i^T)\}_{i=1}^n$ and native student objects be $\{(b_j^S, c_j^S, p_j)\}_{j=1}^m$. Here b is a decoded physical box, c a class, and s_i^T, p_j a score. Compatible class semantics and known coordinate transforms are required; internal features and head parameterizations need not match. Teacher objects use pre-NMS native queries. Students select the top 64 logits per task group, at most 384 objects over six groups before validity filtering. CenterHead selects class–cell entries; Anchor3DHead selects class–anchor entries. These readouts serve distillation and do not replace native inference selection.

4.2 HRCA DISTILLATION BASE

HRCA supervises features, responses, and candidates. For feature supervision, a learned 1×1 projection matches channel dimensions, and spatial resampling aligns the student and teacher BEV grids. Teacher-weighted SmoothL1 and unweighted cosine losses compare the aligned features. Dense-response supervision uses teacher-weighted squared errors between student and teacher class probabilities. These feature and response losses form the Clean KD base.

Candidate supervision adds guidance around teacher-detected objects. Following UniDistill's physical-instance sampling (Zhou et al., 2023), it compares student and teacher features at nine physical locations per teacher box using score-weighted SmoothL1 and cosine losses. Student class probabilities at teacher centers are also regressed to teacher scores. A margin loss ranks these center responses above strong responses outside class-specific teacher-center neighborhoods. A separate squared-probability penalty suppresses high responses outside these neighborhoods.

The three groups define

$$\mathcal{L}_{\text{HRCA}} = \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{resp}} + \mathcal{L}_{\text{cand}}, \tag{1}$$

where each group includes its component weights, specified in Appendix A.6 with the sampling and reduction details. OMT complements this base with percentile-based quality targets, relative ordering among teacher objects, and soft associations to native student predictions.

4.3 TEACHER-LOCATION SUPERVISION

A teacher object may identify a useful location absent from the student's top-k set. We therefore also read the student response for the teacher's class at the cell containing that teacher center. CenterHead gathers the class channel; Anchor3DHead takes the highest-logit anchor of that class at the cell. We call this readout a *carrier*, with logit ℓ_i and probability $u_i = \sigma(\ell_i)$. Carriers receive direct supervision; they are distinct from the native objects in the transport plan.

Raw teacher and student scores can have different scales. Let F_c be the bounded empirical percentile of valid teacher scores of class c in the current distributed batch, and set $q_i = F_{c_i^T}(s_i^T)^2$. This target expresses class-relative teacher quality. Carrier quality uses

$$\mathcal{L}_{\text{car}} = \langle \text{BCEWithLogits}(\ell_i, q_i) \rangle_{0.25+q_i}, \qquad \langle f \rangle_w = \frac{\sum_k w_k f_k}{\max(\sum_k w_k, 10^{-8})}. \tag{2}$$

To transfer ordering separately, let o_i be the bounded percentile of the raw teacher score within the scene, pooled across classes. For unordered pairs with $|o_i - o_k| \geq 0.05$, use

$$\mathcal{L}_{\text{ord}} = \left\langle [\min(0.5|o_i - o_k|, 0.25) - \text{sign}(o_i - o_k)(u_i - u_k)]_+ \right\rangle_{|o_i - o_k|}. \tag{3}$$

Thus classwise quality and scene-level ordering convey different information.

4.4 SOFT TRANSPORT OVER NATIVE OBJECTS

The student may propose several alternatives around a teacher object, while other predictions receive little teacher support. We express these sets as marked measures $\mu_T = \sum_i a_i \delta_{(b_i^T, c_i^T)}$ and $\mu_S = \sum_j d_j \delta_{(b_j^S, c_j^S)}$. Masses $a_i \propto \max(q_i, 10^{-4})$ and $d_j \propto \max(\mathrm{sg}(p_j), 10^{-4})$ each sum to one; sg stops gradients. Associations use the cost

$$C_{ij} = \frac{\|(x_i^T, y_i^T) - (x_j^S, y_j^S)\|_2^2}{4^2} + (q_i - \mathrm{sg}(p_j))^2 + 0.25 g(b_i^T, b_j^S) + 12\mathbf{1}[c_i^T \neq c_j^S], \tag{4}$$

where g combines scaled SmoothL1 position, log-size, and velocity errors with periodic yaw error. A finite class-mismatch penalty discourages cross-class associations.

We allow the coupling to deviate from the prescribed masses because teacher and student support can disagree. Entropic unbalanced scaling (Cuturi, 2013; Chizat et al., 2018) approximately solves

$$\min_{P \geq 0} \langle P, C \rangle + \varepsilon \sum_{ij} P_{ij}(\log P_{ij} - 1) + \rho \mathrm{KL}(P\mathbf{1} \| a) + \rho \mathrm{KL}(P^\top \mathbf{1} \| d), \tag{5}$$

with generalized KL, $\varepsilon = 0.2$, $\rho = 1$, and 30 log-domain iterations. The plan permits soft many-to-many associations and total transported mass different from one. Recomputed per scene from current predictions, the plan and its inputs are detached. It supplies fixed weights for the student-output losses below; their gradients update student scores and boxes with the associations held constant.

4.5 SUPERVISION FROM THE TRANSPORT PLAN

Normalize the plan by its total mass, $\bar{P} = P / \max(\sum_{ij} P_{ij}, 10^{-8})$, to supervise native objects:

$$\mathcal{L}_{\mathrm{qual}} = \sum_{ij} \bar{P}_{ij} h(p_j - q_i), \qquad \mathcal{L}_{\mathrm{geom}} = \sum_{ij} \bar{P}_{ij} g(b_i^T, b_j^S), \tag{6}$$

where h is unit-transition SmoothL1. These terms update selected student scores and decoded boxes. To account for predictions receiving little teacher support, retain the unnormalized plan and define

$$\kappa_j = \mathrm{clip}\left(\frac{\sum_i P_{ij}}{\max(d_j, 10^{-8})}, 0, 1 \right), \qquad e_j = \mathrm{sg}(1 - \kappa_j), \qquad \mathcal{L}_{\mathrm{exc}} = \langle p_j^2 \rangle_{e_j}. \tag{7}$$

This suppresses weakly supported native responses, including classes absent from the teacher set. Since support reflects teacher coverage, teacher omissions can also suppress valid student predictions.

The two paths combine as

$$\mathcal{L}_{\mathrm{OMT}} = \mathcal{L}_{\mathrm{car}} + 2\mathcal{L}_{\mathrm{ord}} + 1.25\mathcal{L}_{\mathrm{qual}} + 0.5\mathcal{L}_{\mathrm{geom}} + \mathcal{L}_{\mathrm{exc}}. \tag{8}$$

Terms are averaged over valid scenes on each worker. Top-k and carrier indices are discrete. OMT introduces no trainable module; HRCA retains its training adapter. At inference only the native LiDAR student and its post-processing are executed.

5 EXPERIMENTS

5.1 MATCHED EVALUATION PROTOCOL

We evaluate the ten nuScenes categories on the standard train/validation split (Caesar et al., 2020), reporting official epoch-20 mAP and NDS on a 0–100 scale, with absolute differences expressed in points. All accuracy comparisons use single seed-0 runs with nondeterministic execution.

The primary pair is BEVDilation and CenterPoint-Voxel. Training uses class-balanced sampling, ten LiDAR sweeps, 20 epochs, and eight GPUs with four samples per GPU. Student configuration, data, and optimization are fixed within controlled comparisons.

Table 2: **Distillation under fixed detector interfaces.** BEVDilation → CenterPoint-Voxel, nuScenes validation. Scores are on a 0–100 scale; differences are relative to task-only. †: adaptations on Clean KD. CRKD retains feature, affinity, and class transfer; BCKD retains class transfer.

Method	mAP ↑	NDS ↑	Δ mAP	Δ NDS
Task-only	59.156	66.886	+0.000	+0.000
Clean KD	59.156	66.949	+0.000	+0.063
UniDistill (Zhou et al., 2023)†	59.096	66.987	-0.060	+0.101
DistillBEV (Wang et al., 2023)†	59.206	67.199	+0.050	+0.313
CRKD (Zhao et al., 2024b)†	58.778	66.816	-0.378	-0.070
FGD (Yang et al., 2022b)†	59.154	66.903	-0.002	+0.017
BCKD (Yang et al., 2023)†	59.202	66.970	+0.046	+0.084
MGD (Yang et al., 2022c)†	59.127	67.054	-0.029	+0.168
Clean KD + OMT	59.665	67.245	+0.509	+0.359
HRCA	60.057	67.534	+0.901	+0.648
OMT-Distill	**61.069**	**67.920**	**+1.913**	**+1.034**

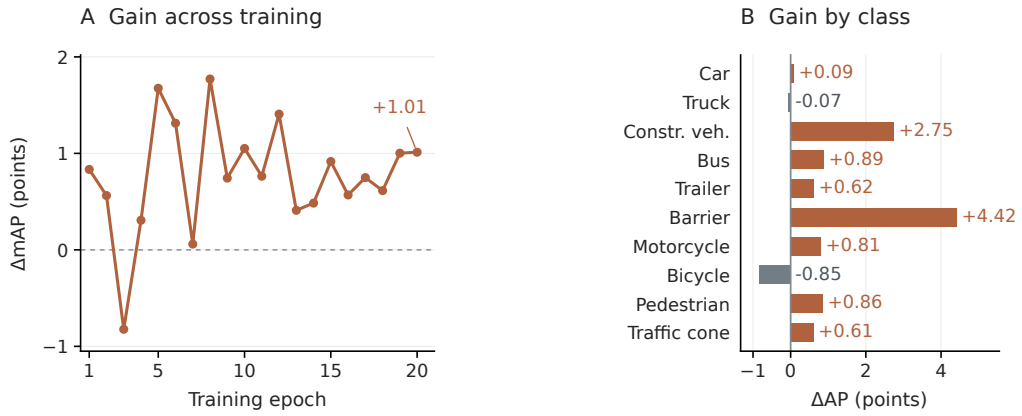


Figure 2: **OMT gains across training and classes.** (A) HRCA+OMT minus HRCA mAP at every recorded epoch, without smoothing. (B) The same comparison for all ten class APs at epoch 20.

5.2 Gains from object distillation

We test whether OMT adds useful object knowledge beyond existing distillation. Six literature baselines augment Clean KD through available native interfaces (Table 2). Their retained mechanisms, response adaptations, and pre-training weight calibration are specified in Appendix B.3.

On Clean KD, OMT adds 0.51 mAP points and 0.30 NDS points. It exceeds the BCKD classification-response adaptation on this base by 0.46 mAP points and 0.28 NDS points, providing a task-side comparison. With stronger candidate supervision, OMT adds 1.01 mAP points and 0.39 NDS points over HRCA.

OMT-Distill gains 1.91 mAP points and 1.03 NDS points over task-only. It exceeds our DistillBEV adaptation by 1.86 mAP points and 0.72 NDS points under the same protocol.

With checkpoints fixed, paired resampling of all 150 validation scenes gives 95% intervals of [0.40, 1.66] mAP points and [0.01, 0.72] NDS points for OMT beyond HRCA (2,000 draws). These quantify validation-scene sampling uncertainty, not variation across training seeds (protocol in Appendix B.2). OMT improves mAP over HRCA in 19 of 20 recorded epochs, decreasing at epoch 3 (Figure 2A). Class gains are uneven, including a 0.85-point bicycle AP decrease (Figure 2B). AP improves at all four matching tolerances, while orientation and velocity errors increase (Appendix B.4).

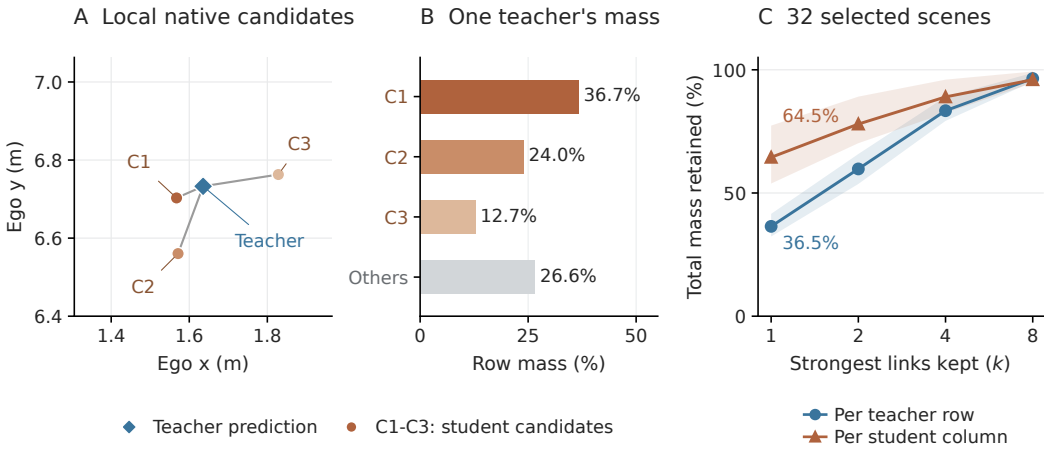


Figure 3: **Mass sharing in native-object transport.** (A) C1–C3 are all native candidates in a ± 0.30 m crop around a barrier teacher center. (B) Their shares of its full outgoing mass; "Others" sums the 381 uncropped candidates. (C) Fraction of total mass retained by the k strongest entries per row or column. Lines show means and bands interquartile ranges over 32 selected scenes, using plans reconstructed from checkpoint outputs.

5.3 Transfer across detector designs and efficiency

Table 3: **Transfer across detector architectures.** CP: CenterPoint; PP: PointPillars; BEVD: BEVDilation; DI++: DeepInteraction++. Center/Anchor denote native Center-Head/Anchor3DHead; Voxel/Pillar denote input encoding. Scores are on a 0–100 scale; differences are relative to matched task-only controls. The two voxel rows share the task-only run.

Student / teacher	Head	Task-only		OMT-Distill		Difference	
		mAP	NDS	mAP	NDS	mAP	NDS
CP-Voxel / BEVD	Center	59.156	66.886	61.069	67.920	+1.913	+1.034
CP-Pillar / BEVD	Center	49.947	60.608	52.062	61.655	+2.115	+1.047
PP-Anchor / BEVD	Anchor	37.463	53.732	50.464	60.044	+13.001	+6.312
CP-Voxel / DI++	Center	59.156	66.886	61.413	68.320	+2.257	+1.434

The same HRCA+OMT recipe improves both additional center and anchor students, including all ten class APs (Table 3). The large anchor gain starts from a weaker 37.46 mAP control.

DeepInteraction++ changes image–LiDAR interaction during both encoding and prediction; with the voxel student, the same objective weights also transfer successfully. These matched comparisons extend the complete recipe across detector designs, while the primary pair isolates the OMT increment.

OMT adds 13.10% training step time and 0.05% allocated memory over HRCA, without learned parameters; the complete recipe adds 54.82% step time over task-only. Training and inference measurements are detailed in Appendix B.5.

5.4 Where object supervision acts

We examine how the two paths supervise student outputs, then compare the trained students' detections. Diagnostic protocols and supplementary analyses are provided in Appendix C.

Soft transport shares supervision among native objects. Keeping the strongest link per teacher row retains only 36.5% of total mass on average; per student column, it retains 64.5% (Figure 3). Supervision thus reaches multiple native alternatives in both directions.

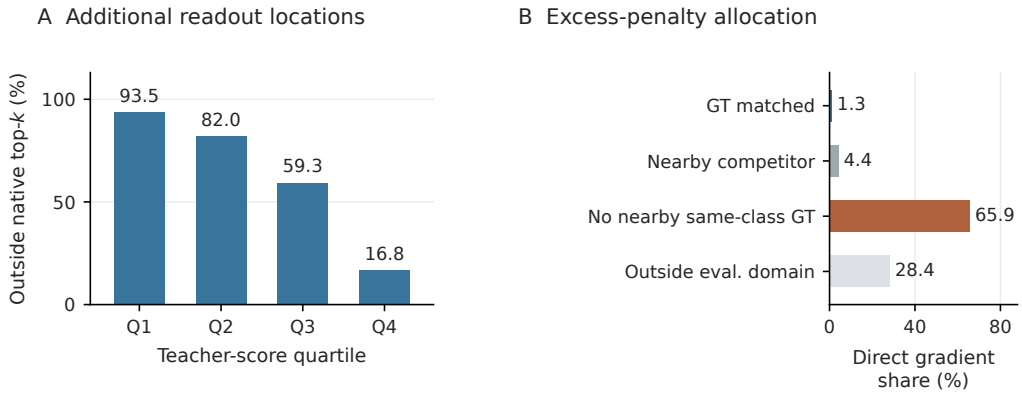


Figure 4: **Where the two supervision paths reach.** (A) Teacher-indexed class/cell addresses outside native top-k, grouped by class-relative teacher-score quartile (Q1 lowest). (B) Excess-loss direct logit-gradient shares with the plan fixed, averaged equally over 32 selected states. Candidate categories use official filtering and same-class matching/proximity at 2 m, before NMS; evaluation-domain exclusions remain separate. Neither panel measures the isolated training benefit of a path.

Teacher-location supervision extends coverage. In 32 selected states, 3,695 of 5,863 distinct teacher-indexed addresses (63.0%) lie outside native top-k (Figure 4A). Even among addresses whose teacher queries approach same-class GT within 2 m, 592/1,916 (30.9%) are outside. Teacher-indexed supervision thus reaches responses omitted by native candidate selection.

Support directs suppression. Among in-domain native candidates with $\kappa \leq 0.1$, 3,448/3,462 lack same-class GT within 2 m. Across all support levels, candidates lacking nearby same-class GT receive 65.9% of the direct excess-loss gradient, while 28.4% lies outside the evaluation domain (Figure 4B). Three low-support candidates still match GT missed by nearby teacher queries, exposing omission risk.

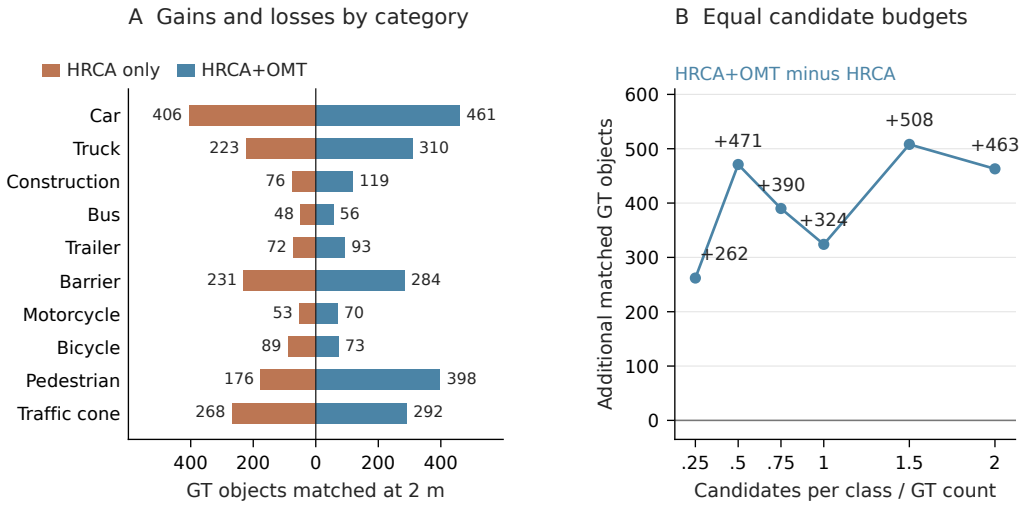


Figure 5: **Object gains beyond a change in prediction count.** (A) GT matches gained and lost by class. (B) Additional matches at equal per-class budgets (prediction count / GT count), selecting each model's highest scores globally within class. Both panels use all 6,019 validation frames, official score-ordered matching at 2 m, and native NMS outputs with scores at least 0.1.

Object gains persist at equal prediction counts. Across the full validation set, OMT gains 2,156 GT matches and loses 1,642 at 2 m (Figure 5A): a net gain of 514, or 0.422 percentage points in GT recall (paired scene-bootstrap 95% interval [0.273, 0.574]). All six equal per-class prediction bud-

gets retain gains (panel B); at 121,861 predictions per model, OMT matches 324 more GT objects. These gains jointly reflect localization, classification, and ordering.

Net matches increase in 94 scenes, decrease in 36, and tie in 20. GT with 1–5 current-keyframe LiDAR points gains 0.673 percentage points in recall [0.435, 0.933]. Gains remain uneven across distance, point-count, and class strata, including losses and intervals crossing zero.

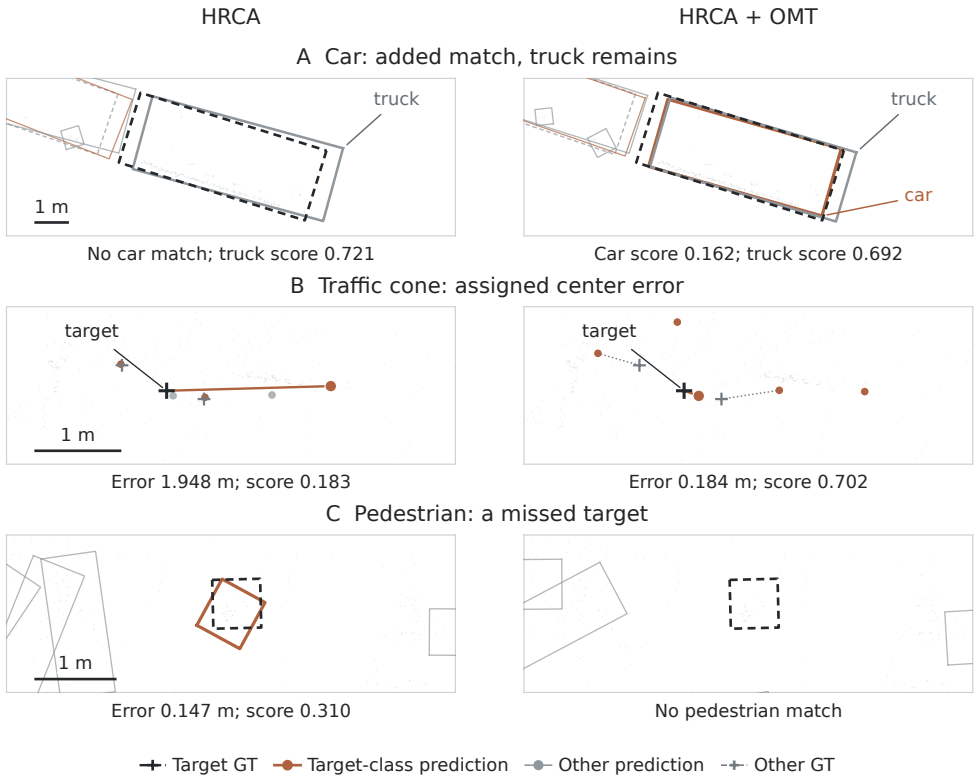


Figure 6: **Object-level changes from adding OMT to HRCA.** Each pair shares GT, LiDAR points, crop, and scale. (A,C) BEV boxes. (B) Centers and actual assignments: the solid link identifies the target's match; dotted links show neighboring GT assignments. Copper denotes each row's target class. Matches use whole-frame, same-class, distance-first one-to-one assignment within 2 m, with prediction scores at least 0.1. These selected cases do not estimate outcome frequencies.

In Figure 6, the cone's assigned error decreases while those of neighboring GT objects increase. Such local changes coexist with a modest 3.4 mm mean center-error improvement over jointly matched validation GT.

6 Limitations and Conclusion

OMT distills 3D object knowledge from multimodal teachers into heterogeneous LiDAR students through teacher-location supervision and soft transport over native objects. The two paths supervise responses at teacher centers and candidates generated by the student. Controlled experiments show benefits beyond feature/response distillation and the stronger HRCA base; paired detection gains persist at equal prediction counts. Additional students and a second teacher extend OMT-Distill across detector designs, with native student inference preserved.

The evidence covers nuScenes and single training runs; scene-bootstrap intervals quantify uncertainty conditional on the checkpoints. Teacher omissions can propagate through support-based suppression, and gains in geometry and ordering remain uneven. Direct object interchange also retains substantially more utility than the trained student realizes. The present results demonstrate useful transfer with native student inference; learning more of that available object knowledge remains open.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and drafting, literature lookup, interpretation and checking of archived experimental records, analysis code, and preparation of tables, figures, and typesetting. Accuracy endpoints originate from recorded experiments; additional analyses use their archived logs and predictions. The authors are responsible for the scientific claims and the final disclosure of AI use.

ETHICS STATEMENT

This work studies 3D object detection using the existing nuScenes benchmark. Improved benchmark accuracy does not establish the reliability required for safety-critical deployment. Errors in detecting rare, distant, or partially observed road users remain relevant, and teacher supervision can propagate teacher errors. Deployment would require additional validation beyond the experiments reported here.

REPRODUCIBILITY STATEMENT

The method description and appendix specify the object representation, transport solver, loss terms, and training settings. The experiments use fixed-epoch comparisons and distinguish native inference from training-only teacher computation. Reproduction requires the corresponding teacher implementations and checkpoints, nuScenes data and preprocessing, and the student training environment. The appendix describes the oracle calibration and candidate contracts separately from the trainable objective, and specifies the fixed-checkpoint scene-bootstrap sampling and metric reconstruction.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection. In *ICLR*, 2023. URL https://arxiv.org/abs/2211.09386.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, volume 26, 2013.

Henri De Plaen, Pierre-François De Plaen, Johan A. K. Suykens, Marc Proesmans, Tinne Tuytelaars, and Luc Van Gool. Unbalanced optimal transport: A unified framework for object detection. In *CVPR*, pp. 3198–3207, 2023.

Shuxuan Guo, Yinlin Hu, Jose M. Alvarez, and Mathieu Salzmann. Knowledge distillation for 6D pose estimation by aligning distributions of local predictions. In *CVPR*, pp. 18633–18642, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Yuzhe Ji, Yijie Chen, Liuqing Yang, Rui Ding, Meng Yang, and Xinhu Zheng. VeXKD: The versatile integration of cross-modal fusion and knowledge distillation for 3D perception. In *NeurIPS*, volume 37, 2024.

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

Suhas Lohit and Michael Jones. Model compression using optimal transport. In *WACV*, 2022.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals Kullback–Leibler divergence for knowledge distillation. In *NeurIPS*, volume 37, 2024.

Nassim Ali Ousalah, Anis Kacem, Enjie Ghorbel, Emmanuel Koumandakis, and Djamila Aouada. Uncertainty-aware knowledge distillation for compact and efficient 6DoF pose estimation. In *IROS*, pp. 15398–15405, 2025. doi: 10.1109/IROS60139.2025.11245982.

Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. CrossKD: Cross-head knowledge distillation for object detection. In *CVPR*, pp. 16520–16530, 2024a.

Yu Wang, Xin Li, Shengzhao Weng, Gang Zhang, Haixiao Yue, Haocheng Feng, Junyu Han, and Errui Ding. KD-DETR: Knowledge distillation for detection transformer with consistent distillation points sampling. In *CVPR*, pp. 16016–16025, 2024b.

Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation. In *ICCV*, 2023.

Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3D object detection with knowledge distillation. In *NeurIPS*, volume 35, 2022a.

Longrong Yang, Xianpan Zhou, Xuewei Li, Liang Qiao, Zheyang Li, Ziwei Yang, Gaoang Wang, and Xi Li. Bridging cross-task protocol inconsistency for distillation in dense object detection. In *ICCV*, 2023.

Zeyu Yang, Nan Song, Wei Li, Xiatian Zhu, Li Zhang, and Philip H. S. Torr. DeepInteraction++: Multi-modality interaction for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. URL https://arxiv.org/abs/2408.05075.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, 2022b.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *ECCV*, 2022c.

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

Guowen Zhang, Chenhang He, Liyi Chen, and Lei Zhang. BEVDilation: LiDAR-centric multi-modal fusion for 3D object detection. In *AAAI*, volume 40, pp. 12448–12456, 2026. doi: 10.1609/aaai.v40i15.38238.

Haimei Zhao, Qiming Zhang, Shanshan Zhao, Zhe Chen, Jing Zhang, and Dacheng Tao. SimDistill: Simulated multi-modal distillation for BEV 3D object detection. In *AAAI*, 2024a.

Lingjun Zhao, Jingyu Song, and Katherine A. Skinner. CRKD: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *CVPR*, 2024b.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. UniDistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view. In *CVPR*, 2023.

A Implementation and Reproduction Details

A.1 Detector configurations

Table 4 records the evaluated configurations.

Table 4: **Detector paths in the evaluated setting.** Both teachers use images and LiDAR; all students use LiDAR only. The shared object objective does not require these internal paths to correspond. CP: CenterPoint; PP: PointPillars.

Detector	Representation and fusion	Native prediction interface
BEVDilation	Visual guidance in sparse voxels; cascaded BEV dilation	DAL query head with image context
DeepInteraction++	Separate image/LiDAR streams; bidirectional interaction	Iterative multimodal query decoder
CP-Voxel	Sparse voxels; SECOND/FPN	Dense center responses
CP-Pillar	Pillars and scatter; SECOND/FPN	Dense center responses
PP-Anchor	Pillars and scatter; SECOND/FPN	Class–anchor responses

Teachers. BEVDilation (Zhang et al., 2026) produces 32-channel image BEV features with ResNet-50, CustomFPN, and LSS. Its modified SparseEncoder uses image/LiDAR foreground estimates to expand sparse coordinates and Mamba to propagate features; new coordinates start from a learned feature. The resulting 192-channel representation and image features enter four Cascade_SBDB layers. The 256-channel output and image context feed DALHead, with 200 proposals and one 256-channel query decoder layer. DeepInteraction++ (Yang et al., 2025) uses Swin/FPN and HardSimpleVFE/SparseEncoder/SECOND/SECONDFPN. Two FusionTransformerv4 layers include point-side polar interaction; its 200-proposal, 128-channel decoder retains separate modality representations through four predictive interaction rounds.

Students and interfaces. All three students use SECOND/SECONDFPN after voxel or pillar encoding. Their head inputs have 512, 384, and 384 channels on grids of 180×180, 136×136, and 216×216, respectively. Native assignment, box coding, and suppression are preserved within each task-only/distilled pair. DeepInteraction++ and CenterPoint-Voxel share LiDAR stem components: heterogeneity concerns the complete representation and prediction paths, rather than every operator. OMT reads native class responses and physical boxes; HRCA additionally requires the feature projections and spatial alignment specified below.

A.2 Optimization and coordinate conventions

The training environment uses Python 3.8.20, PyTorch 1.12.1 (CUDA 11.6 build), MMCV 1.6.0, MMDetection 2.25.1, and MMDetection3D 1.0.0rc4; the system CUDA compiler is 11.8. AdamW uses initial learning rate 10^{-4}, weight decay 0.01, and gradient clipping at norm 35. The twenty-epoch schedule has one cycle, learning-rate target ratios $(10, 10^{-4})$, momentum ratios $(0.85/0.95, 1)$, and a 40% increasing phase. Batch and data controls follow Section 5.

The primary geometry spans $[-54, -54, -5, 54, 54, 3]$ meters. CenterPoint-Voxel uses $[0.075, 0.075, 0.2]$ m voxels and output stride 8; CenterPoint-Pillar uses horizontal size $108/544$ m on a 544×544 grid; PointPillars-Anchor uses $[0.25, 0.25, 8]$ m pillars on a 432×432 grid. HRCA projects student features with a bias-enabled 1×1 convolution to the teacher width and bilinearly resizes to the teacher grid with `align_corners=False`.

For DeepInteraction++, recorded calibration maps inputs into the teacher LiDAR frame and decoded centers, heading vectors, and velocities back into the student frame; dimensions are unchanged. Teacher BEV features and dense logits are bilinearly warped into the student frame. Camera projections include inverse BEV augmentation. Distillation runs every iteration with a frozen-teacher microbatch of four; outputs are concatenated before forming the OMT batch.

A.3 NATIVE OBJECT AND TEACHER-LOCATION READOUTS

Teacher objects use the final `num_proposals` native queries. Centers follow the teacher coder's voxel size, stride, and origin; dimensions are exponentiated, yaw is $\mathrm{atan2}(\mathrm{rot}_0, \mathrm{rot}_1)$, and bottom height is center height minus half the vertical dimension. A query retains its assigned class. Its score multiplies that class's final sigmoid response by the dense initialization prior (one if absent). OMT applies finite and spatial-validity checks, without the HRCA score threshold, object cap, or NMS.

CenterHead carriers gather the teacher class channel at $\lfloor ((x_i^T, y_i^T) - \mathrm{origin}_{xy})/(\mathrm{voxel\,size}_{xy} \cdot \mathrm{stride}) \rfloor$. Native objects select the highest 64 logits from each task's complete class–spatial tensor and gather regression at those cells. Teacher centers must lie in the student's feature-map domain; native boxes also satisfy the coder's post-center range. Anchor3DHead carriers take the maximum teacher-class logit over anchors at that cell. Native selection spans class–anchor entries in each group, followed by the native delta decoder, discrete direction argmax, and yaw-period correction. There is no surrogate gradient for direction selection. Finite outputs and student x, y bounds are checked. Both readers support one prediction level and six groups: car; truck/construction vehicle; bus/trailer; barrier; motorcycle/bicycle; pedestrian/traffic cone. Native p_j is the sigmoid of the selected logit.

A.4 QUALITY TARGETS AND REDUCTION

Pool valid teacher scores classwise across all scenes and distributed workers. For class c with N_c scores, let R_i count scores less than or equal to s_i^T (the right insertion rank). Then

$$q_i = \left[\mathrm{clip}\left(\frac{R_i - 1/2}{N_{c_i^T}}, \frac{1}{2N_{c_i^T}}, 1 - \frac{1}{2N_{c_i^T}} \right) \right]^2 . \tag{9}$$

Ties share a percentile and singleton classes receive $q_i = 0.25$. Teacher finite-value and spatial checks precede pooling; the subsequent carrier-finiteness filter does not recompute targets. The ordinal target o_i in Equation 3 instead ranks raw scores across classes among surviving teacher–carrier pairs within each scene, using the same bounded right-rank convention. Each worker averages over valid scenes, skipping empty ones; distributed gradient averaging then combines workers.

A.5 GEOMETRY AND TRANSPORT NUMERICS

Write $b = (x, y, z, d_x, d_y, d_z, \psi, v_x, v_y)$ with bottom height z. Let $h(t) = t^2/2$ for $|t| < 1$ and $|t| - 1/2$ otherwise, and $H(v) = |v|^{-1} \sum_k h(v_k)$. The exact discrepancy is

$$\begin{aligned} g(b^T, b^S) = \frac{1}{6} \Bigg[2H\left(\frac{(x^S, y^S) - (x^T, y^T)}{4} \right) + h\left(\frac{z^S - z^T}{2} \right) \\ + H(\log d^S - \log d^T) + \frac{1 - \cos(\psi^S - \psi^T)}{2} + H\left(\frac{v^S - v^T}{5} \right) \Bigg]. \end{aligned} \tag{10}$$

Dimensions are clamped below by 10^{-3} before taking logarithms. For box tensors with fewer than nine entries, velocity contributes zero and the denominator remains six. A CenterHead reader without a velocity head supplies two zero entries, so omission depends on tensor dimension.

The generalized divergence in Equation 5 is $\mathrm{KL}(x\|y) = \sum_k [x_k \log(x_k/y_k) - x_k + y_k]$. With $K = \exp(-C/\varepsilon)$, $P = \mathrm{diag}(r) K \,\mathrm{diag}(t)$, and $\tau = \rho/(\rho + \varepsilon)$, initialize $\log r = \log t = 0$ and alternate

$$\log r_i \leftarrow \tau[\log a_i - \mathrm{LSE}_j(-C_{ij}/\varepsilon + \log t_j)] , \tag{11}$$

$$\log t_j \leftarrow \tau[\log d_j - \mathrm{LSE}_i(-C_{ij}/\varepsilon + \log r_i)] . \tag{12}$$

The kernel uses entropy relative to counting measure, without an $a_i d_j$ factor. Plan inputs are detached float32 tensors; after 30 iterations, $\log P$ is clamped to $[-60, 20]$ before exponentiation, and non-finite plans raise an error. Teacher and student masses are normalized independently. Quality and geometry use $\bar{P}$, whereas coverage and excess suppression use P. This approximate solver costs $O(nm)$ storage and $O(30nm)$ arithmetic per scene, excluding readout and the quadratic carrier-ordering loss.

A.6 EXACT HRCA BASE AND DEPLOYMENT

Feature and response alignment. For projected student features F^S and detached teacher features F^T, separately min–max normalize teacher channel-RMS energy and maximum class heatmap probability into $[0.05, 1]$ per scene. Their pointwise maximum weights feature SmoothL1, normalized by the weight sum and channel count. Feature cosine loss is the unweighted spatial and batch mean of one minus channel-vector cosine similarity. Resize teacher probabilities T to student heatmap resolution S. Positive response loss is $\sum(S-T)^2T^2/\max(\sum T^2, 10^{-6})$; dense loss is the mean of $(S-T)^2W$, where $W_{bxy} = 0.05 + (\max_c T_{bcxy})^2$ is normalized by its global mean. Anchor3DHead class logits take the maximum over anchors at each cell.

Candidate guidance. The native teacher coder's filtered decode retains scores at least 0.05 and at most 80 objects, without an additional NMS call. Each object supplies nine yaw-rotated samples (center, four side midpoints, four corners) using bilinear `align_corners=True` sampling. Coordinates use the physical extent and are clamped to the feature grid. Patch SmoothL1 and cosine errors are normalized with raw teacher-score weights. Support loss is mean squared error between student probabilities at teacher centers and teacher scores clamped to $[0.05, 1]$. A class-specific square mask of radius two surrounds each rounded teacher center. Ranking compares every positive center with the 64 strongest responses outside the mask at margin 0.15; false-peak suppression averages squares of the 128 strongest unsupported responses. Candidate losses average over applicable scenes. The weights are

$$\begin{aligned}\mathcal{L}_{\text{HRCA}} = {} & 2\mathcal{L}_{\text{feat,L1}} + \mathcal{L}_{\text{feat,cos}} + \mathcal{L}_{\text{resp,pos}} + 0.25\mathcal{L}_{\text{resp,dense}} \\ & + 2\mathcal{L}_{\text{patch,L1}} + \mathcal{L}_{\text{patch,cos}} + \mathcal{L}_{\text{support}} + 2\mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{false}}.\end{aligned} \tag{13}$$

The teacher stays in evaluation mode and is excluded from exported weights. Training checkpoints retain the HRCA adapter; inference executes only the native LiDAR student, decoder, and post-processing.

B EXTENDED EXPERIMENTAL RESULTS

B.1 TRAINING CONTROLS AND TRAJECTORIES

All 16 formal accuracy runs train for twenty epochs and report the shared epoch-20 endpoint. Clean KD+OMT, the six external ports, PointPillars-Anchor OMT, and DeepInteraction++ OMT have validation records only for epochs 16–20; the other runs have records for 1–20. The primary OMT run resumes after epoch 14 following a repository refactor with a matched configuration; Clean KD+OMT resumes after epoch 16 following an evaluation-hook update. These are continuations of single runs, not independent repetitions.

Figure 7 supplements the late-epoch comparison with every available evaluation, including early regressions. Epochs share training history and validation data, so the curves do not estimate training-seed uncertainty or wall-clock convergence. Clean KD is the common base of the module control: HRCA adds candidate guidance. Subtracting the two conditional OMT effects in Table 2 gives a descriptive interaction of $+0.50$ mAP points and $+0.09$ NDS points, from one run per condition.

B.2 PAIRED SCENE BOOTSTRAP

Cached Task-only, HRCA, and HRCA+OMT predictions cover the same 6,019 frames from 150 scenes. The archived formatter maps predictions from camera-timestamp ego coordinates to global coordinates. Ground truth uses official nuScenes class-range, point-count, and bicycle-rack filters. Replayed mAP and NDS agree with official cache evaluations to six decimal places; main tables retain the logged epoch-20 endpoints.

For each of 2,000 draws (seed 202709), sample 150 scenes with replacement, including all their frames and using identical scene multiplicities for all checkpoints. Repeated frames receive distinct identities while preserving prediction order for score ties. Recompute globally pooled, score-ordered AP, true-positive errors, and NDS, rather than averaging per-scene AP. Table 5 uses linear 2.5th and 97.5th percentiles of paired differences. We check accumulation against the official devkit, including

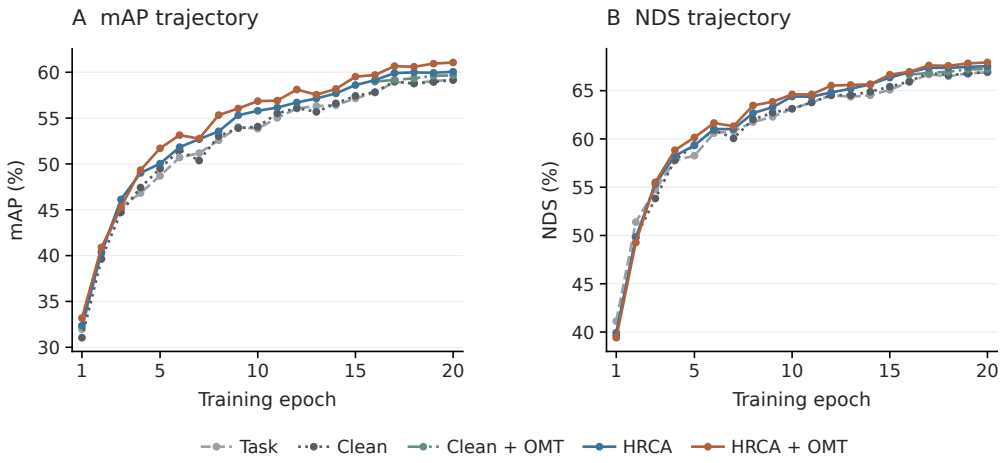


Figure 7: **Complete recorded validation trajectories.** Points are unsmoothed evaluations from single twenty-epoch runs. Clean KD+OMT has evaluations only at epochs 16–20; other controls have evaluations at every epoch. Early regressions are retained.

Table 5: **Paired scene bootstrap.** Differences are percentage points, with 2.5th–97.5th percentiles from 2,000 paired resamples of 150 validation scenes with replacement. Checkpoints are fixed; intervals do not quantify training-seed variability.

Fixed-checkpoint comparison	Δ mAP [95% interval]	Δ NDS [95% interval]
HRCA+OMT minus HRCA	+1.01 [+0.40, +1.66]	+0.39 [+0.01, +0.72]
HRCA+OMT minus Task-only	+1.91 [+1.38, +2.50]	+1.03 [+0.65, +1.44]

duplicated scenes and tied scores. Intervals quantify validation-scene variation at fixed weights, not training variation or the factorial interaction.

B.3 External baseline interfaces and calibration

Feature and response interfaces. The six ports share the frozen teacher, complete voxel student, training recipe, and Clean KD weights. Student sparse-encoder output pairs with teacher middle output; student backbone stage 0 with teacher stage 1; and the final BEV head inputs with each other. All paired maps are 180×180, with channel projections. DistillBEV, FGD, and MGD use three independent stage pairs; UniDistill uses the encoder pair for nine-point transfer and the head-input pair for relations. Responses align the ten native classes.

Table 6 records retained mechanisms and missing interfaces. CRKD's camera/radar transfer is unavailable in the LiDAR student; the query teacher lacks corresponding dense box fields for UniDistill/CRKD response regression and BCKD localization. CRKD pools positive-count normalization across all classes with uniform task weights, replacing reference task-wise normalization and weighting. BCKD sums binary class losses and averages over BEV locations and batch, replacing its original head-specific normalization. These are adapted comparisons under the fixed teacher/student constraint.

Weight calibration. Four scales per method are screened with fixed-batch backward passes, then three candidates run for 50 optimization steps on eight GPUs. Relative to selected weights in Table 7, candidates are $\{0.5, 1, 1.5\}$ for UniDistill, DistillBEV, and CRKD; $\{1, 1.5, 2\}$ for FGD; $\{0.5, 1, 2\}$ for BCKD; and $\{0.25, 0.5, 1\}$ for MGD. Relative component weights and MGD's 0.65 masking coefficient stay fixed. Probes start from a common epoch-20 HRCA checkpoint but optimize Clean KD plus the baseline objective, disabling HRCA's candidate terms. Final models train from scratch.

Table 6: **Baseline mechanisms under fixed detector interfaces.** All six methods augment Clean KD for BEVDilation–CenterPoint. Missing modality and dense box branches are distinguished from feature-pairing and response-normalization choices.

Method	Retained mechanism	Interface adaptation
UniDistill	Nine-point feature and relation transfer at separate stages; Gaussian class responses.	Encoder and head-input feature pairs; dense box-response regression unavailable.
DistillBEV	Three real BEV stages; foreground, background and teacher false-positive regions; spatial attention.	Encoder, backbone, and head-input pairs with independent adapters; channel mask disabled.
CRKD	Range/velocity-aware feature masks, affinity relations and class responses.	Modality-specific CSRD and dense box responses unavailable; global class normalization and uniform task weights.
FGD	Foreground/background features, spatial/channel attention and relation/context terms.	Three BEV pairs replace the 2D FPN correspondence, with independent relation modules.
BCKD	Binary class-response distillation on aligned dense heatmap logits.	Dense localization responses unavailable; losses summed over classes and averaged over BEV locations and batch.
MGD	Spatial masks, independent stage generators and teacher reconstruction.	Three BEV pairs replace the 2D FPN correspondence; independent generators and mask drop ratio 0.65.

Table 7: **Baseline weight calibration.** Selected weights are used for twenty-epoch training. The last column reports additional-objective/task gradient-norm ratios on student parameters, measured by a fixed-batch backward pass after each selected 50-step probe. Ratios describe calibration, not training-loss shares or detection quality.

Method	Selected weights	Grad. ratio (%)
UniDistill	Feature 3.2; relation 1.6; class response 0.32	2.39
DistillBEV	FG 3.6×10^{-6}; BG 2.4×10^{-5}; spatial 1.5×10^{-6}; FP 3.6×10^{-5}	4.64
CRKD	Mask 6.0; affinity 0.15; response 0.6	4.12
FGD	$\alpha = 1.6\times10^{-5}$; $\beta = 8\times10^{-6}$; $\gamma = 1.6\times10^{-5}$; $\lambda = 8\times10^{-8}$	5.03
BCKD	Loss weight 75; $\beta = 1$	3.40
MGD	$\alpha = 6.2\times10^{-7}$; $\lambda = 0.65$	0.35

Selection considers additional-objective/task gradient-norm ratios on student parameters, reconstruction residuals, auxiliary gradients, and task-loss stability. Selected ratios span 0.35–5.03%. MGD uses its largest tested dynamic weight and still has the lowest ratio. Calibration checks that the objective acts on the student; it does not establish validation-optimal weights for the original methods.

B.4 Distance tolerances, classes, and box errors

Table 8 averages classes at each matching tolerance; Tables 10 and 11 average the four distance APs within each class. These log fields are rounded and may not exactly average to the separately logged official mAP used for overall comparisons. Matching tolerance measures prediction–GT center error, rather than object distance from the sensor. Table 9 preserves all five true-positive errors, including the increased scale and orientation errors of PointPillars-Anchor despite its aggregate accuracy gain.

B.5 Efficiency protocol

Training uses eight A800-SXM4-40GB GPUs, batch four and six data workers per GPU, with 20 warmup and 50 measured steps. Timing includes transfer to GPU, forward/backward passes, reduc-

Table 8: **AP by center-distance matching tolerance.** AP averages ten classes at 0.5, 1, 2, and 4 m using rounded logged class APs. Values are multiplied by 100; differences are percentage points. Tolerances measure prediction–GT center distance, not object range from the sensor.

Run	$AP_{0.5}$	$AP_{1.0}$	$AP_{2.0}$	$AP_{4.0}$
Task-only	47.941	57.957	63.887	66.840
HRCA	48.453	58.785	65.009	67.982
OMT-Distill	49.412	59.675	66.384	68.802
Δ vs HRCA	+0.959	+0.890	+1.375	+0.820

Table 9: **True-positive errors at epoch 20.** Native scales (lower is better): mATE in metres, mAOE in radians, mAVE in metres per second; mASE and mAAE are unitless. Ported baselines use Clean KD; remaining rows follow the main-table names and recipes.

Run	mATE	mASE	mAOE	mAVE	mAAE
Task-only	0.2863	0.2518	0.2993	0.2434	0.1885
Clean KD	0.2844	0.2521	0.3037	0.2355	0.1872
UniDistill	0.2834	0.2503	0.3031	0.2284	0.1908
DistillBEV	0.2850	0.2498	0.2934	0.2270	0.1851
CRKD	0.2896	0.2528	0.2885	0.2386	0.1878
FGD	0.2857	0.2508	0.3096	0.2330	0.1882
BCKD	0.2843	0.2508	0.2969	0.2403	0.1907
MGD	0.2875	0.2518	0.2975	0.2302	0.1840
Clean KD + OMT	0.2811	0.2503	0.3004	0.2379	0.1891
HRCA	0.2864	0.2523	0.2956	0.2318	0.1834
OMT-Distill	0.2853	0.2524	0.3007	0.2398	0.1833
CP-Pillar / Task	0.3147	0.2593	0.3707	0.2909	0.2010
CP-Pillar / OMT	0.3172	0.2588	0.3848	0.2819	0.1949
PP-Anchor / Task	0.3542	0.2705	0.4233	0.2511	0.2009
PP-Anchor / OMT	0.3400	0.2858	0.4535	0.2535	0.1860
CP-Voxel / DI++ OMT	0.2833	0.2518	0.2920	0.2323	0.1793

tions, clipping, and AdamW, excluding host data loading. Task-only has $8{,}938{,}262$ trainable parameters; HRCA and OMT each have $9{,}069{,}590$, including the $131{,}328$-parameter HRCA adapter. The frozen teacher has $80{,}416{,}327$ parameters. Memory is peak allocated CUDA memory; step timings are not measured twenty-epoch runtimes.

Inference uses one A800, batch one, 50 warmup and 300 fixed sequential samples, including transfer, native forward, decoding, and NMS. The checkpoints use the same student-only model and parameter shapes. Their mean decoded-box counts are 112.05 and 128.48, allowing checkpoint-dependent post-processing time despite identical architecture. Dense FLOP counters do not cover the sparse/custom operators faithfully.

B.6 Teacher construction control

Unlike the fixed-teacher baselines, this VeXKD-style control constructs learned BEV fusion queries for feature-mask transfer (Ji et al., 2024). It therefore changes the supervision source.

One construction learns a query module with the original detector frozen; another adapts active teacher inputs and detection components. Both retain the cascaded BEV backbone and DAL interface. Their selected teacher checkpoints are epochs 10 and 7, respectively, of ten-epoch construction runs. Student transfer uses low/high BEV features, independent query-guided deformable masks, and masked Attention Transfer, with teacher detection replay for the first two epochs. Students train from scratch for twenty epochs under the matched recipe.

Table 10: **Class AP for objective controls.** AP is multiplied by 100; the final column reports changes in percentage points. Each AP averages four rounded distance-threshold logs, so their aggregate can differ slightly from official mAP.

Class	Task	Clean	Clean+OMT	HRCA	Full OMT	Δ vs HRCA
Car	85.445	85.445	85.580	85.553	85.638	+0.085
Truck	55.938	57.845	57.140	58.648	58.575	-0.073
Construction	17.908	18.608	18.038	18.823	21.568	+2.745
Bus	71.275	71.053	70.825	70.335	71.225	+0.890
Trailer	36.858	38.523	37.845	37.688	38.305	+0.618
Barrier	71.945	68.430	70.908	69.518	73.938	+4.420
Motorcycle	60.728	60.023	62.620	62.680	63.488	+0.808
Bicycle	39.938	39.873	40.353	44.090	43.240	-0.850
Pedestrian	84.458	84.185	84.768	84.593	85.450	+0.858
Traffic cone	67.073	67.583	68.568	68.648	69.258	+0.610

Table 11: **Class AP for additional students.** Both use BEVDilation. AP is multiplied by 100; differences are percentage points against matched task-only controls. All ten classes improve in these two seed-0 runs.

Class	CenterPoint-Pillar			PointPillars-Anchor		
	Task	OMT	Δ	Task	OMT	Δ
Car	84.333	84.443	+0.110	80.805	84.285	+3.480
Truck	52.808	53.985	+1.178	46.243	50.920	+4.678
Construction	9.430	13.798	+4.368	9.535	15.190	+5.655
Bus	67.438	67.460	+0.023	61.195	64.460	+3.265
Trailer	33.415	35.405	+1.990	31.090	35.818	+4.728
Barrier	58.675	61.793	+3.118	31.273	57.618	+26.345
Motorcycle	45.883	48.923	+3.040	26.113	48.160	+22.048
Bicycle	13.745	18.445	+4.700	2.205	18.695	+16.490
Pedestrian	77.505	78.515	+1.010	65.830	77.810	+11.980
Traffic cone	56.245	57.863	+1.618	20.343	51.690	+31.348

The initial two runs retained mask-state drift after freezing (Table 13, †). The strict variant removes mask gradients and optimizer-state updates from epoch 3. The encoder-aligned variant additionally replaces the teacher's downstream BEV encoder with SECOND/SECONDFPN initialized from the task-only student and widens the DAL input to 512 channels, retaining the strict mask protocol. Its lower teacher accuracy confounds any effect of encoder alignment. These controls characterize the tested interface constructions; they do not establish failure of the original VeXKD setting or feature distillation in general.

C Diagnostic Protocols and Additional Analyses

C.1 Object interchange: protocol and additional conditions

The diagnostic fixes the BEVDilation epoch-10 teacher and HRCA CenterPoint-Voxel epoch-20 student over all 6,019 validation frames. Its 200 native DAL queries undergo physical-range filtering; teacher centers index the 180×180 student grid over $[-54, 54]$ m by floor, reading regression without spatial search. Following the archived diagnostic protocol, O1–O4 exclude centers outside the indexable domain instead of clipping them. O5 imports complete teacher objects, including non-indexable objects permitted by the student's post-center range. These are the earlier interchange rules, separate from the final distillation implementation.

O2 permutes the full multiset of student scores read at valid teacher indices according to teacher ordering, including cross-class order. O3-cal and O5-cal use deterministic five-fold scene partitions, with each calibration map fit outside its held-out fold. For each class and sample, the target distri-

Table 12: **Training and native-student inference costs.** Training: 20 warmup and 50 measured steps; inference: 50 warmup and 300 measured samples. Memory is peak allocated CUDA memory per GPU; host loading is excluded. The two workloads are not directly comparable.

Variant	Mean (ms)	P95 (ms)	Samples/s	Peak MiB
Training: 8 A800 GPUs, batch 32				
Task-only	1410.650	1503.409	22.685	7613.5
HRCA	1931.003	2042.462	16.572	9289.0
OMT-Distill	2183.998	2263.420	14.652	9293.5
Student-only inference: 1 A800 GPU, batch 1				
Task-only	156.718	166.535	6.381	272.3
OMT-Distill	157.291	167.056	6.358	272.3

Table 13: **Teacher-construction controls.** VeXKD-style teachers distill to native CenterPoint-Voxel students. Teacher scores use the supplied checkpoint; student scores use epoch 20. Teachers vary across rows, unlike the fixed-teacher comparison in Table 2. †: initial runs before the strict post-freeze mask-state repair.

Construction and student control	Teacher		Student	
	mAP	NDS	mAP	NDS
Frozen detector; query-mask KD†	71.715	74.430	58.847	66.896
Adapted teacher; query-mask KD†	71.923	74.578	58.976	67.022
Adapted teacher; strict mask freeze	71.923	74.578	58.792	66.814
Aligned BEV encoder; strict freeze	68.172	70.139	58.983	66.903

bution uses the student's top k native pre-NMS scores, where k is that sample's teacher query count for the class. An empirical quantile map aligns teacher scores to this budget-matched distribution without GT or task metrics. This cross-fitting differs from training-batch quality normalization.

Replacement discards the student pre-NMS pool; union concatenates it with imported objects before the same student Scale-NMS. Table 14 adds the conditions omitted from the main table: union variants and complete objects with raw confidence. The calibrated union underperforms replacement here, without ruling out other fusion mechanisms. O5 and O6 also differ in calibration and filtering, so their difference does not isolate an NMS cost.

Table 14: **Additional object-interchange conditions.** These eight rows complement the eight conditions in Table 1. Scores are multiplied by 100; recovery uses the same student–teacher gap. Raw/calibrated confidence and replacement/union pools remain distinct.

Object fields	mAP	NDS	mAP recovered (%)	NDS recovered (%)
o1 index union	59.91	67.46	-1.3	-1.1
o2 rank union	65.88	70.44	50.4	41.3
o3 score raw union	62.26	68.71	19.1	16.7
o3 score cal union	69.50	72.36	81.6	68.7
o4 geometry union	59.99	67.72	-0.6	2.6
o5 object raw replace	64.97	71.59	42.5	57.8
o5 object raw union	62.38	68.97	20.1	20.4
o5 object cal union	69.84	73.30	84.6	82.1

C.2 Diagnostic states and reconstructed transport

Mechanism analyses reuse one archived frame from each of 32 validation scenes, with saved points for eight frames. This convenience sample does not estimate validation-wide mechanism frequen-

cies. Plans use the formal cost and detached solver, with classwise teacher percentiles pooled over the 32 states to match the formal batch size; this does not reconstruct their original training batches.

Figure 3 selects the highest-quality teacher in the first archived frame. The crop contains every native center within ± 0.30 m in both coordinates, independent of transport mass or detection outcome. Local barrier IDs C1–C3 order the candidates by that teacher's row mass; they are not class or track identifiers. Row shares normalize over all 384 columns, with *Others* pooling the 381 outside the crop. For scene s with $M_s = \sum_{ij} P^s_{ij}$, concentration is

$$R^s_T(k) = \frac{\sum_i \sum_{j \in \mathrm{Top}_k(P^s_{i,:})} P^s_{ij}}{M_s}, \qquad R^s_S(k) = \frac{\sum_j \sum_{i \in \mathrm{Top}_k(P^s_{:,j})} P^s_{ij}}{M_s}.$$

All rows and columns contribute. Curves report equal-weight scene means and 25th–75th percentiles at $k \in \{1, 2, 4, 8\}$. They characterize the reconstructed plans; no hard assignment is solved or trained.

C.3 Matched objects and equal prediction budgets

Full-validation analyses use 6,019 frames, 150 scenes, and 121,861 officially eligible GT objects. Predictions map to global coordinates with camera-keyframe ego poses; official range filtering uses the LiDAR ego pose. GT follows class-range, point-visibility, and bicycle-rack filters. Cached predictions have native NMS and score threshold 0.1; discarded candidates are unavailable.

For each model, descending-score predictions match the closest unmatched same-class GT within the chosen XY tolerance, following official evaluation. Gained/lost objects are GT identities matched only by HRCA+OMT/HRCA, without matching boxes between the models. Table 15 extends the main 2 m analysis to all four official tolerances. Intervals reuse the 2,000 paired scene draws in Appendix B.2.

Table 15: **Paired GT changes across matching tolerances.** All validation frames and the original prediction cache (score at least 0.1). The interval is a paired scene-bootstrap 95% interval for the net change in GT recall, in percentage points.

Tolerance (m)	Gained	Lost	Net	Recall change [95% CI]
0.5	4,578	3,938	+640	+0.525 [0.298, 0.746]
1.0	3,030	2,451	+579	+0.475 [0.287, 0.674]
2.0	2,156	1,642	+514	+0.422 [0.273, 0.574]
4.0	1,745	1,355	+390	+0.320 [0.187, 0.464]

A sensitivity check with distance-ordered greedy assignment at 2 m gives 2,144 gains and 1,633 losses (net 511). Among the 112,276 GT matched by both models under score ordering, mean center error changes from 0.27202 to 0.26864 m: the paired mean is -3.39 mm (95% interval $[-5.53, -1.41]$ mm), with median -0.42 mm. The large local correction in Figure 6 therefore exceeds the modest average localization change.

For Figure 5B, each model retains its highest $K_c = \lfloor r N_c \rfloor$ scores per class over validation, where N_c is the class GT count and $r \in \{0.25, 0.5, 0.75, 1, 1.5, 2\}$. Both caches contain enough predictions for every budget. Identical counts precede score-ordered matching at 2 m; locations, labels, and scores remain model-specific. The comparison controls prediction count but jointly reflects localization, classification, and ordering.

C.4 Paired responses at shared teacher locations

The 32-state comparison pairs the native class channel and integer BEV cell, covering 6,255 of 6,326 teacher rows (98.88%) and 5,792 distinct shared addresses; 71 rows lack an HRCA record. Duplicates have identical responses. Statistics keep teacher-row weights within frames and weight frames equally.

Because the archives have nonidentical teacher query sets, both students are compared against the fixed OMT teacher state. Classwise quality uses its full 32-frame pool, and ordering uses all teacher

rows per frame, before restricting to shared addresses. BCE and weighted ordering hinge follow Section 4; inversion fractions use pairs with teacher-rank gaps at least 0.05. Targets express teacher-relative quality, not GT correctness.

Table 16: **Carrier responses at shared native addresses.** Means over the same 32 selected frames; lower is better for each statistic. The last column counts frames with a lower value under HRCA+OMT.

Statistic	HRCA	HRCA+OMT	Frames improved
Teacher-relative quality BCE	0.6399	0.6247	26 / 32
Weighted ordering hinge	0.0817	0.0795	22 / 32
Reversed pairs (%)	21.497	21.402	16 / 32

Table 16 shows more consistent quality alignment than reversal correction: BCE falls in 26/32 frames, but inversions fall in only 16/32. These selected states and shared addresses do not give unbiased validation-wide estimates. A stricter archive check requiring teacher centers within 0.01 m, scores within 0.001, and equal class/cell retains 5,593 rows. Recomputing targets on this subset yields inversion fractions of 25.182% and 25.080%, also indicating little reversal change.

C.5 Coverage of the two readout paths

The 32 states contain 6,326 teacher rows and 12,288 native candidates (64 per task group per state). Carrier identity is the frame, class, and integer teacher-center cell, rather than decoded-box proximity. For 2,533 rows, class, logit, and all nine decoded box components exactly match a native candidate; the other 3,793 logits lie strictly below their group's 64th selected logit. Deduplication gives 2,168 addresses inside top-k and 3,695 outside, totaling 5,863. The 5,792 paired addresses above are the subset shared with HRCA. For the quartiles in Figure 4A, a repeated address takes its highest-quality teacher row. Nearby-GT support requires a same-class teacher query strictly within 2 m, without one-to-one assignment.

An additional geometric check applies official GT visibility/range/rack filters and candidate range/rack filters, leaving 1,284 GT rather than the 1,611 used before full filtering in qualitative views. A GT is covered if any eligible same-class box center is strictly within 2 m, without extra score thresholding or NMS. Carriers and native candidates cover 1,131 GT in common, 119 only through carriers, 11 only through native candidates, and 23 through neither. Carrier regression is read only for this diagnostic: the carrier loss supervises class quality and ordering. These counts indicate complementary access, not additional trained detections or an isolated branch effect.

C.6 SUPPORT-PENALTY ALLOCATION

Reconstruct P and κ_j from the same states. In-domain native candidates match nearest available same-class GT strictly within 2 m, in descending score order. An unmatched candidate near same-class GT is a *nearby competitor*; other in-domain candidates have no nearby same-class GT. Range/rack exclusions form a separate category. These are pre-NMS training candidates without an extra score filter.

Table 17: **Native candidates and excess-penalty allocation at 2 m.** Counts pool the 32 states; gradient shares average within-frame fractions equally. All support levels contribute to the gradient column.

Candidate category	$\kappa \leq 0.1$	$\kappa > 0.1$	Gradient (%)
GT matched	3	1,130	1.34
Nearby competitor	11	2,909	4.39
No nearby same-class GT	3,448	1,850	65.90
Outside evaluation domain	1,691	1,246	28.37
Total	5,153	7,135	100.00

Table 17 complements the derivative shares in Figure 4B with candidate counts. The $\kappa \leq 0.1$ split is diagnostic; training uses continuous $e_j = 1 - \kappa_j$. Holding the plan and e_j fixed, for $p_j = \sigma(z_j)$,

$$\frac{\partial \mathcal{L}_{\text{exc}}}{\partial z_j} = \frac{2e_j p_j^2(1-p_j)}{\max(\sum_k e_k, 10^{-8})}.$$

Sum magnitudes by category within each frame, divide by that frame's total, then average frame shares equally. Teacher omissions include 11 native matches without a same-class teacher query within 2 m, including all three low-support matches. Their within-frame shares of derivative magnitude average 0.20%. This direct output derivative does not measure shared-parameter effects or post-training benefit/harm.

A composition check stratifies in-domain candidates by class, score (edges $0, 0.1, 0.2, 0.4, 0.6, 0.8, 1$), and 10 m distance bins. Requiring ten candidates in each support group retains 41 overlapping strata and 4,595 candidates. Common pooled-count weights yield no-nearby-GT fractions of 99.70% at low support and 69.44% at higher support. Limited overlap and within-bin differences prevent a causal or calibrated interpretation.

C.7 SCENE AND OBSERVATION STRATA

Removing any one validation scene leaves 460–533 net gains at 2 m. The equally weighted scene-mean recall change is 0.441 percentage points [0.284, 0.614] (median 0.293 percentage points), distinct from the main GT-weighted statistic. Tables 18 and 19 retain all predefined bins and classes. Intervals use shared scene draws with resampled GT denominators.

Table 18: **All predefined observation strata.** Full-validation score-ordered matches at 2 m; intervals are paired scene-bootstrap 95% intervals for recall changes in percentage points, conditional on fixed checkpoints.

GT stratum	GT count	Gained	Lost	Net	Recall change [95% CI]
Distance from LiDAR ego (m)					
0–10	15,769	172	108	+64	+0.406 [0.115, 0.673]
10–20	36,649	470	375	+95	+0.259 [0.026, 0.500]
20–30	34,418	639	461	+178	+0.517 [0.224, 0.850]
30–40	22,737	485	358	+127	+0.559 [0.273, 0.865]
40–50	12,288	390	340	+50	+0.407 [-0.054, 0.841]
Current-keyframe LiDAR points					
0	3,468	145	136	+9	+0.260 [-0.861, 1.368]
1–5	41,289	1,106	828	+278	+0.673 [0.435, 0.933]
6–20	32,900	468	385	+83	+0.252 [-0.012, 0.529]
21–50	17,544	243	171	+72	+0.410 [0.023, 0.891]
51+	26,660	194	122	+72	+0.270 [0.097, 0.464]

Table 19: **Class composition of net match changes.** HRCA+OMT minus HRCA at 2 m, retaining negative and zero changes. Each five-column group partitions the same validation GT; a dash indicates no eligible GT. Pooled counts and recall intervals are in Table 18.

	Distance (m)					Current-keyframe LiDAR points				
Class	0–10	10–20	20–30	30–40	40–50	0	1–5	6–20	21–50	51+
Car	+1	+2	+35	-11	+28	0	+30	+6	+16	+3
Truck	+19	-5	+46	+16	+11	0	+29	+7	+26	+25
Construction vehicle	+3	+3	+13	+15	+9	+2	-1	+16	+6	+20
Bus	+1	+3	-3	+4	+3	0	+4	0	+1	+3
Trailer	0	0	+20	+2	-1	-1	-2	+2	+5	+17
Barrier	0	+38	+15	–	–	+2	+51	0	-1	+1
Motorcycle	+1	+2	+6	+8	–	+3	+5	+5	+4	0
Bicycle	+4	-14	-10	+4	–	+1	-11	-7	+1	0
Pedestrian	+18	+73	+42	+89	–	-1	+147	+61	+12	+3
Traffic cone	+17	-7	+14	–	–	+3	+26	-7	+2	0
All	+64	+95	+178	+127	+50	+9	+278	+83	+72	+72

The 40–50 m, zero-point, and 6–20-point intervals include zero. Point counts describe the keyframe rather than ten-sweep input; radar-visible GT can have zero LiDAR points. Class range limits change bin composition. These intervals establish neither simultaneous subgroup significance, training-seed stability, nor a causal effect of sparsity.

D ADDITIONAL QUALITATIVE EXAMPLES

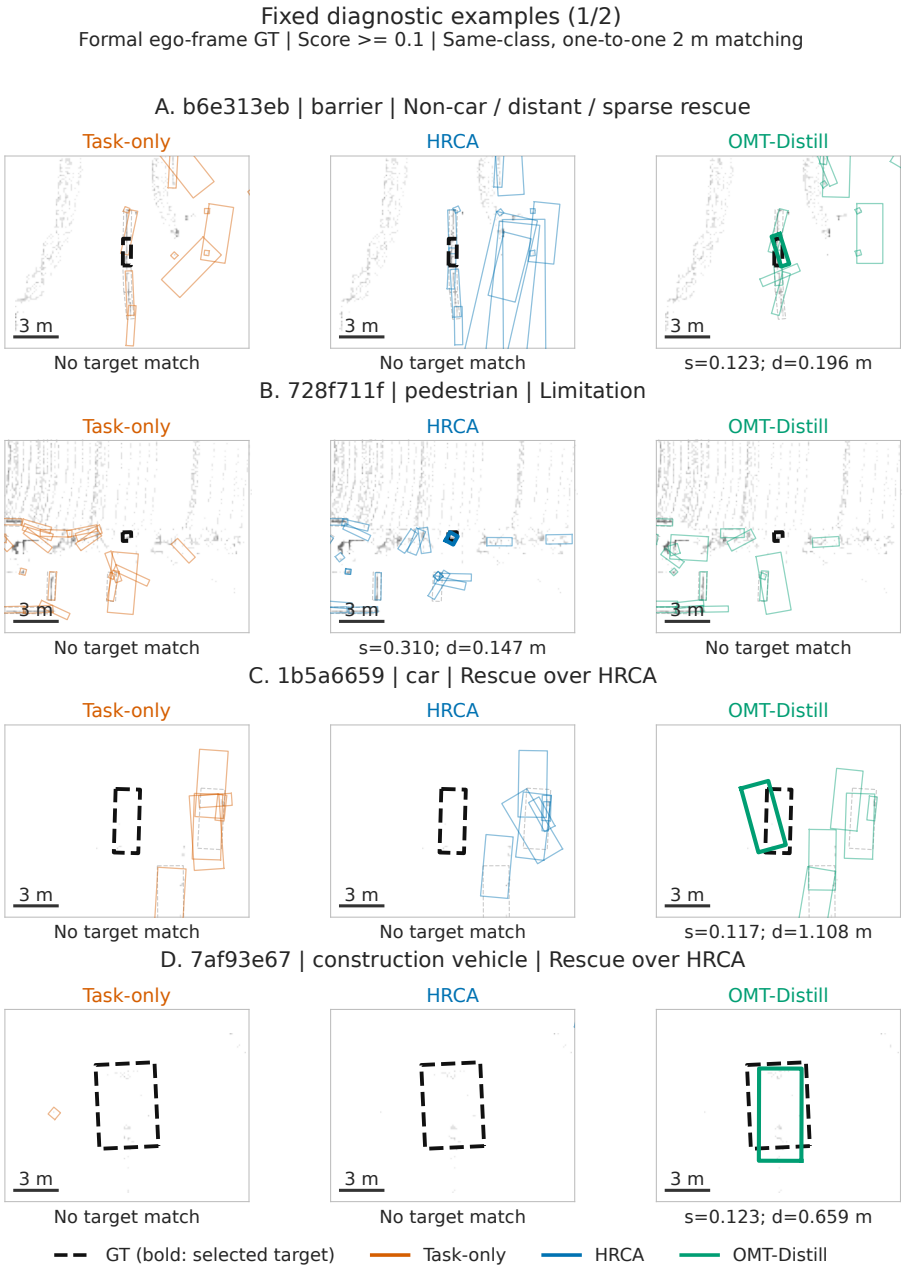


Figure 8: **Additional examples (1/2).** The first four of eight fixed frames with saved points. GT, predictions, and points share CAM_FRONT-keyframe ego coordinates; each row uses the same crop and score threshold 0.1 across methods. Same-class pairs are sorted by XY distance and greedily matched one-to-one within 2 m; bold predictions are assigned matches. This differs from official score-ordered evaluation. The complete qualitative set contains 1,611 GT and 4,833 target–method records. Selected targets illustrate behavior without estimating its frequency.

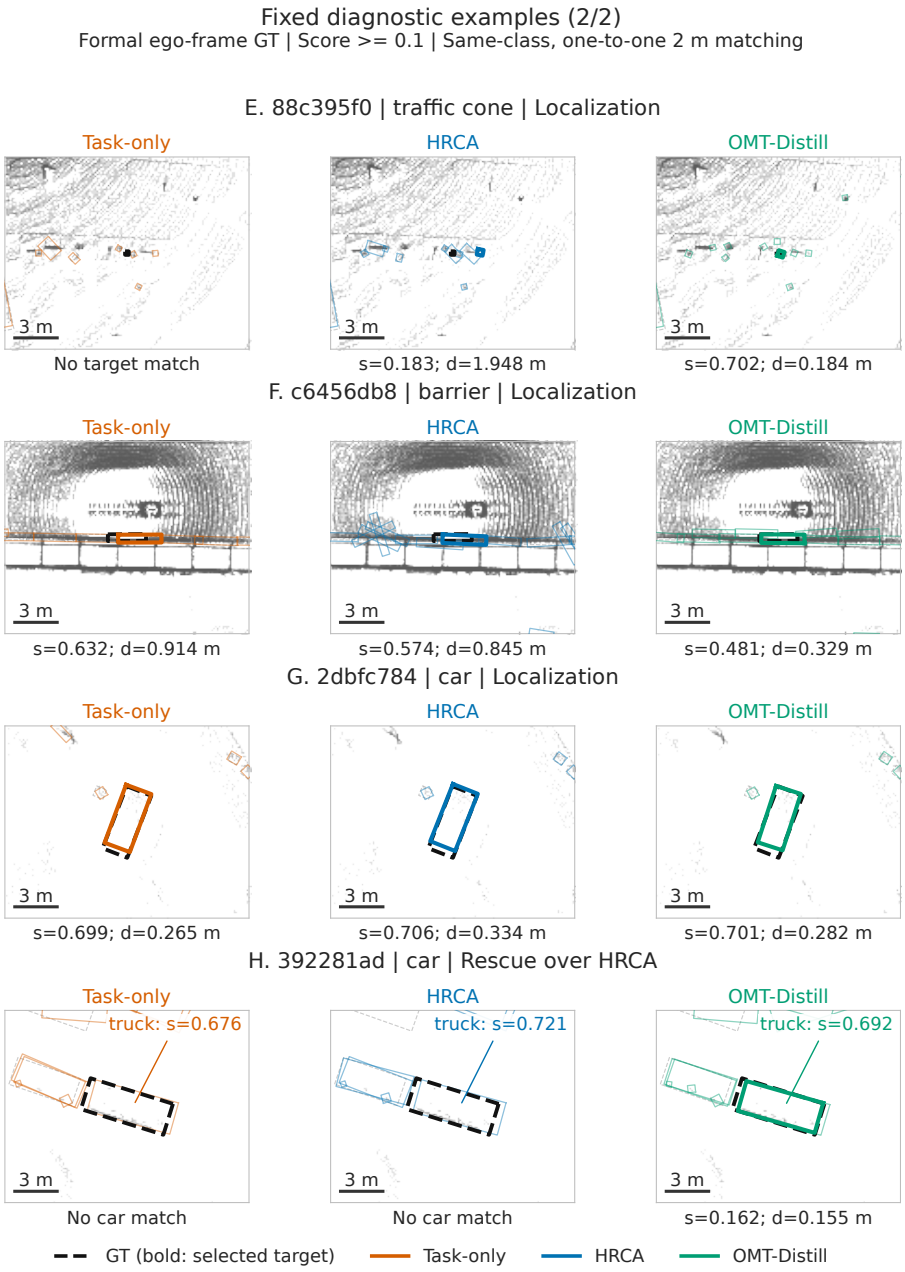


Figure 9: **Additional examples (2/2).** The remaining four fixed frames, retaining the same coordinate, threshold, and matching conventions as Figure 8. All eight frames are retained, including unfavorable cases; target roles follow common-frame matches.