Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Limits of continual test-time adaptation for camera–LiDAR 3D detection under simulated view loss

Peiyuan Qu
Hainan University
Mengxing Huang
1005467006@qq.com

Hainan University

Article

Keywords: camera–LiDAR fusion, 3D object detection, continual test-time adaptation, sensor view loss, held out classes

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11669407/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Limits of continual test-time adaptation for camera–LiDAR 3D detection under simulated view loss

Peiyuan Qu[1] and Mengxing Huang[1,*]

[1]College of Information and Communication Engineering, Hainan University, Haikou, China

[*]Correspondence: 1005467006@qq.com

Abstract

Continual test-time adaptation updates a detector during deployment, but improved detection of supervised classes need not preserve objects excluded from current supervision. We evaluated a 672-parameter adapter on three paired nuScenes streams with simulated loss and return of camera views. Relative to a frozen detector, the complete configuration changed mean known-class average precision by +0.055 percentage points and mean held-out recall by −0.233 percentage points. Held-out recall was lower on terminal intact-view inputs in all three online streams. In a separate diagnostic replay, the protection mask covered 22 of 201 eligible held-out object centers and 2278 of 6829 known centers; filtering also reduced known pseudo-label precision from 31.81% to 23.34%. On a five-scene probe with updates disabled, frozen final states improved intact-view held-out recall by 1.40–3.15 percentage points at a small cost to known-class accuracy. Removing one scene eliminated the weakest gain, and a tight candidate budget reversed it. These results show why online outcomes, residual effects of final adapter states, and pseudo-label selectivity should be evaluated separately.

Keywords: camera–LiDAR fusion; 3D object detection; continual test-time adaptation; sensor view loss; held-out classes

Introduction

A camera–LiDAR detector combines visual information with measurements of three-dimensional geometry. When camera views become unavailable, an online update may compensate for changed inputs, but it may also reinforce mistakes that persist after the cameras return. Sensor failure and corruption benchmarks illustrate how strongly the performance of a fixed detector can depend on the failure mode [1–3]. Here we use one precisely defined intervention on nuScenes [4]: selected camera views are zeroed after image preprocessing while LiDAR is retained.

Bird's-eye-view (BEV) representations provide a common spatial frame for fusion [5, 6]. Related work uses voxel, pillar, and center representations for point clouds [7–9], geometric or temporal lifting for images [10, 11], and query-based prediction for object detection [12, 13]. An adapter before fusion can change the strength of downstream responses while keeping most parameters fixed. It cannot reconstruct an unavailable measurement. Its effect on objects excluded from the current label set therefore needs to be measured separately from its effect on familiar classes.

Test-time training and entropy minimization adapt models without labels from the incoming test stream [14, 15]. Related approaches use source hypotheses, augmentations, classifier adjustment or contrastive objectives [16–19]. Continual approaches address accumulated errors with teacher averaging, reliability filtering, regularization or restoration [20–24]. Benchmarking and selection studies caution that apparent benefits can depend on stream construction, reliability criteria and computation [25, 26]. In multimodal perception, cross-modal pseudo-label refinement has been studied for semantic segmentation [27]; it does not follow that two teachers sharing one fused detector provide independent evidence.

Open-set recognition, open-world detection and adaptation under open-world data shift distinguish familiar objects from those outside the current label set [28–30]. Confidence calibration, evidential estimates, ensembles and out-of-distribution scoring provide complementary tools [31–35], but candidate confidence is not a guarantee of correct localization or reliable supervision. Teacher averaging and anti-forgetting constraints [36–38], as well as adaptation of normalization layers [39, 40], offer different ways to update or preserve model state.

We ask whether a 672-parameter BEV adapter improves known detection and retains held-out objects as camera views are removed and then restored. The update combines a quality gate, a fused teacher, candidate-derived protection and source contraction (Figs. 1 and 2). We use three complementary evaluations: matched online streams to measure the actual trajectory, a diagnostic replay to measure the selectivity of the supervision filter, and scene-disjoint frozen replay to test what the final states retain on different inputs. Motorcycle and bicycle are held out of eight-class fine-tuning and online supervision, but were within the ten-class initialization vocabulary.

Results

Continual updates did not consistently preserve held-out recall

Across three paired S2 streams, the complete method M5 changed mean known-class center-distance mean average precision (mAP) from 48.172% to 48.227% and mean held-out recall at three-dimensional intersection-over-union (IoU) 0.25 from 17.885% to 17.652% (Table 1). The paired known changes were +0.023, +0.069 and +0.074 percentage points (pp); the paired recall changes were −0.782, +0.340 and −0.256 pp. The entropy control M2 improved known mAP and reduced held-out recall in every stream (Supplementary Fig. S1). The custom teacher control M3 and restricted SoftRestore control also had lower mean held-out recall than the frozen reference (Table 1). The standard deviations describe variation across schedules; the schedules are not independent training replicates.

Removing protection and its preservation loss jointly yielded 17.898% mean held-out recall, compared with 17.652% for M5 and 17.885% for the frozen reference, with mixed directions across streams. The complete configuration had higher mean held-out recall than variants without quality gating or source contraction, but neither comparison showed a uniform direction (Supplementary Fig. S2). These ablations identify effects of bundled operations rather than isolate each loss or gate component. Across M0 and M5, held-out average precision at IoU 0.25 was about 0.0246%, and recall at IoU 0.5 was about 1.96%, with about 277 candidate-level false positives (FPs) per evaluated frame at IoU 0.5. The absolute results limit the practical significance of small changes at the default candidate ceiling.

In terminal Clean, the continuing online M5 state had lower held-out recall than M0 on the same frames in all three streams (−3.604, −1.709 and −1.796 pp; Supplementary Table S9). On a different, 100-frame diagnostic probe already seen in all three streams, freezing the final states gave +3.390, 0 and 0 pp under Clean input (Supplementary Table S10). That probe also shows an input-only effect: with the same frames, M0 known mAP fell from 56.882% to 48.477% and held-out recall from 23.729% to 10.169% when all six views were removed. In S2, different stages use different scenes, so the stage progression alone cannot isolate the effect of input removal.

The protection mask covered few held-out object centers

The replay recorded 293 successful observations from 300 planned update frames. The mask covered 22 of 201 eligible held-out centers (10.95%) and 2278 of 6829 known centers (33.36%; Fig. 3). The 201 held-out observations include seven centers outside the grid, counted as uncovered. Known candidate precision fell from 4586/14418 (31.81%) before filtering to 2498/10701 (23.34%) after filtering, then reached 2403/9256 (25.96%) after the 50-box cap. When the deleted set was independently rematched, 2201 of 3717 removed candidates matched eligible known objects. Rematching can reassign correspondence, so this count does not quantify how many matches from the original candidate set were lost.

The masks and candidate arrays did not retain the index linkage needed to measure the precision of the final query targets used for gradients. The separate diagnostic trajectories also had higher aggregate held-out recall than the corresponding formal runs by 0.782–1.020 pp. The replay measures the selectivity of the filter at sampled successful updates; together with the ablation it identifies a plausible failure mechanism, without isolating its contribution to formal online performance. The formal update logs show successful M5 updates on roughly 24% of frames and source contractions on roughly 55% of successful updates (Supplementary Table S11).

Frozen final states showed conditional gains on scene-disjoint inputs

On the five-scene, 200-frame frozen probe B, M0 matched 43 of 286 eligible held-out observations under Clean input. The three M5 final states matched 52, 49 and 47, giving recall changes of +3.15, +2.10 and +1.40 pp, respectively, with known mAP changes of −0.63, −0.27 and −0.12 pp (Supplementary Tables S2–S5). The differences were uneven across scenes (Fig. 4a). Scene 0911 accounted for 5 of 9, 3 of 6 and 4 of 4 net additional matches across the three states. In the third state, two extra matches in scene 0275 were offset by two fewer in scene 0928.

Removing one scene at a time and recomputing the pooled metrics yielded recall difference ranges of +1.75 to +4.22, +1.31 to +3.01, and 0.00 to +2.87 pp (Supplementary Table S6). The third state's gain vanished when scene 0911 was omitted. Under miss6, the full-probe differences were +0.70, −0.70 and −1.05 pp, and each could change direction after removing one scene. All states use the same five scenes. These results characterize the sensitivity of residual final-state effects, with no additional online update on the probe.

Under a common per-frame cap of 100 held-out candidates, the Clean recall changes remained +2.45, +2.10 and +1.05 pp; observed candidate FP counts were 99.750–99.785 per frame at the same IoU threshold of 0.25

(Fig. 4b). At a cap of 25, the third state had a recall loss of 0.70 pp relative to M0. This smaller-cap regime limits candidate volume for all states and retains little absolute recall. Known mAP is unchanged across caps within a state because the cap is applied only to the held-out candidate branch. The budget sweep is descriptive: the probe was not used to fit a threshold, and no independent data with matching candidate exports were available for equal-FP calibration (Supplementary Tables S7–S8).

Discussion

The data identify two separate reasons to evaluate known and held-out detections. First, the online objectives can change one metric without improving the other. The entropy control increased known mAP in all three streams while decreasing held-out recall. The full method narrowed the mean recall loss relative to the entropy and teacher controls, but did not outperform the frozen reference on that metric. The protection diagnosis gives a concrete explanation for why its intended selectivity is not assured: the mask covered a larger share of known than held-out centers, and removing candidates within it reduced the precision of the retained known supervision. A one-sided objectness term preserves responses at selected grid cells but does not establish their 3D localization, ranking, or correctness. The independent rematching and separate replay trajectory limit causal interpretation.

Second, online return to intact inputs and the behavior of a fixed final state are distinct outcomes. Source contraction reduces parameter drift by construction, but online terminal recall was below the frozen reference even after all camera views returned. On five other scenes the frozen states recovered additional matches under intact inputs, with a known-class trade-off and sensitivity to a particular scene and to candidate budget. These observations can coexist because state updates and frame sets differ. A recovery claim therefore needs to specify whether adaptation is still active, which data are used and how the output is filtered. The high candidate-level FP counts do not estimate an actual road alarm rate.

This study concerns one inherited detector and candidate generator, one category split, and a software intervention that removes image tensors. It does not test physical camera failure, weather robustness, joint sensor degradation or the full nuScenes validation set. Scene disjointness is defined against the recorded fine-tuning, adaptation and development roles; it does not establish independence from all scenes used for the original ten-class checkpoint. The three S2 schedules can share frames, and the three frozen final states share only five probe scenes. The comparisons are descriptive and do not support statistical equivalence or generalization to other detectors. The additional restoration control is a custom adapter comparison, not an evaluation of a standard public algorithm.

The adapter stores only 672 trainable scalars, but teacher/student computation still passes through the frozen detector. Original paired logs on one RTX 4090 show mean prediction times of 1.812 and 1.813 s/frame for M0 and M5, with complete-path times of 1.864 and 2.774 s/frame and approximate peak allocations of 2.13 and 2.20 GiB. These logs show single-GPU feasibility with added computation. They do not establish a speed advantage, real-time operation or a matched timing comparison against other methods. More broadly, a useful test of online adaptation should measure whether supervision is selective, whether recall survives tighter output budgets and whether performance after input restoration agrees across online and frozen evaluation protocols.

Methods

Source detector and category protocol

The input comprises six camera views and a LiDAR point cloud. The fixed source is a BEVFusion-style detector with a TransFusion-style head and an inherited sparse geometry-anchored depth routing (SGDR) component. LiDAR anchors also enter the camera-depth branch before final fusion. The held-out categories, motorcycle and bicycle, are excluded from the eight-class fine-tuning labels and online supervision, although both were present in the official ten-class initialization vocabulary. The inherited class-agnostic candidate generator is kept fixed across configurations. Evaluation labels are not accessible to the online update.

For each frame x_t, record $\hat{y}_t = f_{W,\theta_t}(x_t; 1, 1)$ with unit modality scales, then optionally compute θ_{t+1} from x_t. The update affects subsequent predictions. The frozen weights W include both encoders, the fusion module, the detection head and normalization parameters and running statistics. An affine adapter for modality m maps F_m to $(\mathbf{1}+\gamma_m)\odot F_m+\beta_m$, with both trainable vectors initialized at zero. The camera and LiDAR branches have 80 and 256 channels, giving $2(80+256) = 672$ trainable scalars. The student θ, exponential moving average (EMA) teacher $\bar{\theta}$, and identity source $\theta^0 = 0$ share the same frozen weights. Both teachers use the fused detection architecture; neither is an independent LiDAR detector.

Quality gating and supervision

Camera quality q_c is the fraction of views passing black-view and overexposure checks. Sharpness is logged but does not determine the gate. LiDAR quality q_l combines point-count, coarse coverage and anchor-related proxies and can exceed one. On the update path, modality scales are $s_m = \mathrm{clip}(q_m/\mu_m, 0, 1.2)$ using source-side quality means μ_m; without these statistics the fallback is $s_m = 0.2 + 0.8q_m$. On every fourth frame an update can be attempted unless both $q_c < 0.25$ and $q_l < 0.20$. The EMA and perturbed student use these scales; the identity-source and recorded prediction paths use unit scales.

Known pseudo-labels come from the EMA teacher using source-side per-class score medians as filters. These medians do not target a specified precision. Up to ten unknown candidates from each teacher path are rasterized into a union BEV mask M_u. A known target is rejected if its center falls in a protected cell. The filtered set is capped at 50 and paired spatially to student queries using a 2 m distance criterion and teacher-index hints. The pairing selects raw prediction tensors; query indices are not assumed to track objects across passes. Separate tensor copies and captured query-label state prevent decoding from modifying the raw training outputs.

Let p^T and p^S be teacher and student distributions obtained by softmax over the heatmap class dimension, $\mathcal{J}_b$ the accepted pairs for batch element b, and c^T, c^S the raw center-head outputs. The consistency losses are

$$\mathcal{L}_{\mathrm{cls}} = B^{-1} \sum_{b=1}^{B} \sum_{j \in \mathcal{J}_b} \sum_{c} p^T_{bjc} \log(p^T_{bjc}/p^S_{bjc}), \tag{1}$$

$$\mathcal{L}_{\mathrm{ctr}} = \mathrm{mean}_{(i,j)\in\mathcal{J}} \left| c^S_i - c^T_j \right|. \tag{2}$$

The first uses batch-size reduction ($B = 1$), with no division by the number of paired queries. The second also averages over center components and does not constrain full 3D box geometry. For dense objectness proxies o^T and o^S, the preservation and total losses are

$$\mathcal{L}_{\mathrm{pres}} = \frac{\sum_v M_u(v)\,\mathrm{ReLU}(o^T(v) - o^S(v))}{\max(\sum_v M_u(v), 1)}, \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{ctr}} + \mathcal{L}_{\mathrm{pres}} + 10^{-3}\,\mathrm{mean}(\theta^2). \tag{4}$$

The student update applies feature-channel dropout (0.05) and additive feature noise (0.01). AdamW [41] uses a 10^{-4} learning rate, 0.01 weight decay and gradient clipping at 1.0, with mixed precision. Nonfinite gradients or a gradient norm above 100 reject the update and restore the adapter state.

After a successful update, $\bar{\theta} \leftarrow 0.99\bar{\theta} + 0.01\theta$. The controller tracks the mean of $(q_c + q_l)/2$ over up to 20 successful controller observations and can act before the window fills. If the mean exceeds 0.93 and squared student drift from identity exceeds 10^{-6}, the student contracts as $\theta \leftarrow 0.85\theta$. This follows EMA averaging and multiplies squared drift by 0.7225 for an isolated contraction.

Streams, configurations and metrics

The source calibration pool contains 806 frames from 20 scenes; 80 frames supply threshold and quality statistics. A separate development pool contains 922 frames from 23 scenes. The 40-scene, 1613-frame confirmation pool is disjoint from these recorded pools. Each S2 stream uses 25 whole scenes ordered by a seed-dependent SHA-256 hash and allocated five per stage. Within a scene, inputs are ordered by timestamp. The stages are Clean, miss1, miss3, miss6 and terminal Clean. At input index n and stream seed s, a random generator with seed $1000003s + n$ selects one, three or all six camera views to zero after preprocessing. LiDAR is retained. Seeds 0, 1 and 2 yield 1009, 1008 and 1008 frames with 639, 588 and 782 eligible held-out observations. Methods share the same realized input schedule within a stream. Adapter states start from identity and persist across stage transitions.

The seven formal configurations are the frozen M0, query-entropy M2, EMA consistency M3, full M5 and three M5 ablations (no quality gate, no protection and preservation, no source contraction). M2 minimizes foreground-query entropy through the same adapters; a foreground query has a maximum heatmap score above 0.1. M3 uses class and center consistency with the same pairing and EMA momentum, unit scales and no quality gate, protection, explicit anchoring or contraction. Disabling the M5 gate removes both permission logic and update-path scales. SoftRestore-Consistency-3D-Adapter is a further custom control on the same S2 schedules. It averages consistency losses from two perturbed student forwards against one unperturbed EMA teacher. After EMA averaging, each successful update has a 0.01 probability of softly mixing the entire student adapter state toward identity by 0.10. Teacher targets are not averaged across augmentations. This control does not reproduce CoTTA's augmentation and restoration implementation.

Known center-distance mAP, denoted $\mathrm{mAP}_{\mathrm{cd}}$, averages eight-class AP over thresholds of 0.5, 1, 2 and 4 m. It is not the official ten-class nuScenes mAP or nuScenes detection score. Eligible ground-truth (GT) boxes in both class groups contain at least one LiDAR point. Class-agnostic held-out evaluation uses one-to-one greedy 3D IoU matching and by default retains at most 500 candidates per frame. The primary metric is recall R_{h} at IoU 0.25; auxiliary metrics are AP at 0.25 and recall and candidate FP per evaluated frame at 0.5. Matchings are computed separately at each IoU. Whole-stream AP pools predictions before computing AP. Summaries are means and sample standard deviations across the three stream schedules. Paired differences use the same frames and are descriptive. For M0 and M5, S2 frame and GT counts, known mAP and held-out recall were cross-checked against archived fixed-evaluator JSON. The B aggregates were recomputed from frozen predictions with the original metric routines.

Mechanism diagnosis and frozen probes

For the mechanism diagnostic (A), up to 20 originally successful updates per stream and stage were selected by a hash of stream seed, stage and sample token. The 293 usable observations span 271 unique tokens. Five planned observations failed with nonfinite gradients and two with excessive gradient norms. The historical random state was unavailable, so statistics are conditional on successful replay updates. The mask is a 180×180 BEV grid with 0.6 m cells over $[-54, 54)$ m along each horizontal axis. A center is covered if its indexed cell exceeds 0.5; out-of-grid centers are uncovered. Candidate precision uses independently matched same-class pairs within 2 m horizontal center distance, sorted by distance and greedily paired one-to-one. Deleted candidates are identified by 0.1 m same-class pre/post correspondence and then separately rematched.

The seen probe contains 100 frames from three scenes encountered in every S2 stream and 59 eligible held-out observations. The scene-disjoint probe (B) consists of five complete validation scenes, selected by hashed scene-token order after excluding S2, calibration, development, the seen probe and recorded eight-class fine-tuning scenes. It contains 200 unique frames and 286 eligible held-out observations; one scene has no held-out GT. Within each probe, a fixed frame-specific camera permutation gives nested missing-view sets across Clean, miss1, miss3 and miss6 (probe seeds 20261101 and 20261102, respectively). M0 and the three formal final M5 student states use the same frames in each condition and perform no updates or normalization-state changes.

For B, we pair by condition, scene and frame token and recompute pooled metrics after omitting each scene from both methods. A zero-held-out scene has undefined recall; its candidates count as unmatched. We also apply common ceilings $K \in \{10, 25, 50, 100, 200, 300, 500\}$ to the stored post-decoding candidates, using stable descending score order and reevaluating recall and FP at IoU 0.25. A ceiling does not force an equal emitted candidate count or an equal observed FP rate. We retain the known prediction branch unchanged and do not select a deployment threshold on B.

Code availability

The detector adaptation, stream construction, evaluation and offline analysis use custom code. The authors retain the code and run configurations. Requests for access, including editorial and peer-review requests, should be directed to the corresponding author. Execution requires the licensed nuScenes data, the source detector checkpoint and the software dependencies of the detector.

Use of generative AI

OpenAI ChatGPT assisted with restructuring the manuscript and drafting English prose and captions from the experimental records. The reported measurements come from those records. The authors are responsible for verifying the manuscript and its interpretation.

Data Availability

The study uses the nuScenes dataset, available from `https://www.nuscenes.org/` under the dataset's access terms. Supplementary Information provides the scene-level, leave-one-scene-out and candidate-budget results. The stream manifests, configurations, run-level metrics, state-provenance records and corrected diagnostic arrays are retained in the study archives and are available on reasonable request from the corresponding author. Access to larger prediction and checkpoint artifacts may depend on the applicable data and model distribution terms.

References

[1] Matos, F., Bernardino, J., Durães, J., & Cunha, J. A Survey on Sensor Failures in Autonomous Vehicles: Challenges and Solutions. *Sensors* **24**, 5108 (2024). DOI: 10.3390/s24165108.

[2] Dong, Y. *et al.* Benchmarking Robustness of 3D Object Detection to Common Corruptions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1022–1032 (2023).

[3] Hendrycks, D. & Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2019).

[4] Caesar, H. *et al.* nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 11621–11631 (2020).

[5] Liu, Z. *et al.* BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2774–2781 (2023).

[6] Bai, X. *et al.* TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1090–1099 (2022).

[7] Zhou, Y. & Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 4490–4499 (2018).

[8] Lang, A. H. *et al.* PointPillars: Fast Encoders for Object Detection from Point Clouds. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 12697–12705 (2019).

[9] Yin, T., Zhou, X., & Krähenbühl, P. Center-Based 3D Object Detection and Tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 11784–11793 (2021).

[10] Philion, J. & Fidler, S. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 194–210 (Springer, 2020). DOI: 10.1007/978-3-030-58568-6_12.

[11] Li, Z. *et al.* BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–18 (Springer, 2022). DOI: 10.1007/978-3-031-20077-9_1.

[12] Carion, N. *et al.* End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 213–229 (Springer, 2020). DOI: 10.1007/978-3-030-58452-8_13.

[13] Zhu, X. *et al.* Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[14] Sun, Y. *et al.* Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the International Conference on Machine Learning (ICML)* **119**, 9229–9248 (PMLR, 2020).

[15] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[16] Liang, J., Hu, D., & Feng, J. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)* **119**, 6028–6039 (PMLR, 2020).

[17] Zhang, M., Levine, S., & Finn, C. MEMO: Test Time Robustness via Adaptation and Augmentation. *Advances in Neural Information Processing Systems* **35**, 38629–38642 (2022).

[18] Iwasawa, Y. & Matsuo, Y. Test-Time Classifier Adjustment Module for Model-Agnostic Domain Generalization. *Advances in Neural Information Processing Systems* **34**, 2427–2440 (2021).

[19] Chen, D., Wang, D., Darrell, T., & Ebrahimi, S. Contrastive Test-Time Adaptation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 295–305 (2022).

[20] Wang, Q., Fink, O., Van Gool, L., & Dai, D. Continual Test-Time Domain Adaptation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 7201–7211 (2022).

[21] Niu, S. *et al.* Efficient Test-Time Model Adaptation without Forgetting. In *Proceedings of the International Conference on Machine Learning (ICML)* **162**, 16888–16905 (PMLR, 2022).

[22] Niu, S. *et al.* Towards Stable Test-Time Adaptation in Dynamic Wild World. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).

[23] Yuan, L., Xie, B., & Li, S. Robust Test-Time Adaptation in Dynamic Scenarios. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 15922–15932 (2023).

[24] Döbler, M., Marsden, R. A., & Yang, B. Robust Mean Teacher for Continual and Gradual Test-Time Adaptation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 7704–7714 (2023).

[25] Lee, J. *et al.* Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2024).

[26] Wang, Z. *et al.* In Search of Lost Online Test-Time Adaptation: A Survey. *International Journal of Computer Vision* **133**, 1106–1139 (2025). DOI: 10.1007/s11263-024-02213-5.

[27] Shin, I. *et al.* MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 16928–16937 (2022).

[28] Bendale, A. & Boult, T. E. Towards Open Set Deep Networks. In *Proceedings of the Conference on Computer Vision*

and Pattern Recognition (CVPR), 1563–1572 (2016).

[29] Joseph, K. J., Khan, S., Khan, F. S., & Balasubramanian, V. N. Towards Open World Object Detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 5830–5840 (2021).

[30] Zhou, Z., Guo, L.-Z., Jia, L.-H., Zhang, D., & Li, Y.-F. ODS: Test-Time Adaptation in the Presence of Open-World Data Shift. In *Proceedings of the International Conference on Machine Learning (ICML)* **202**, 42574–42588 (PMLR, 2023).

[31] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. On Calibration of Modern Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)* **70**, 1321–1330 (PMLR, 2017).

[32] Sensoy, M., Kaplan, L., & Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. *Advances in Neural Information Processing Systems* **31**, 3179–3189 (2018).

[33] Lakshminarayanan, B., Pritzel, A., & Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems* **30**, 6402–6413 (2017).

[34] Liang, S., Li, Y., & Srikant, R. Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2018).

[35] Liu, W., Wang, X., Owens, J. D., & Li, Y. Energy-based Out-of-distribution Detection. *Advances in Neural Information Processing Systems* **33**, 21464–21475 (2020).

[36] Tarvainen, A. & Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems* **30**, 1195–1204 (2017).

[37] Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**, 3521–3526 (2017).

[38] Li, Z. & Hoiem, D. Learning without Forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 614–629 (2016). DOI: 10.1007/978-3-319-46493-0_37.

[39] Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning (ICML)* **37**, 448–456 (PMLR, 2015).

[40] Schneider, S. *et al.* Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems* **33**, 11539–11551 (2020).

[41] Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2019).

Acknowledgements

This work was supported by the Hainan Provincial Key Research and Development Program (grant ZDYF2025GXJS182).

Author contributions

P.Q. developed the methodology and software, conducted the experiments, analyzed and visualized the results, and drafted the manuscript. M.H. supervised the work and acquired funding. P.Q. and M.H. contributed to review and editing.

Additional Information

Competing interests. The authors declare no competing interests.

Supplementary information. Additional metrics, protocol checks and audit details accompany this manuscript in a separate Supplementary Information document.

Tables and figures

Table 1: Results on three matched online streams (S2). Scores are percentages, with sample standard deviations across schedules. Differences are mean paired percentage-point changes from M0. The three ablations alter the named component(s) of M5. SoftRestore is a custom adapter control.

Configuration	Known $\mathrm{mAP}_{\mathrm{cd}}$	Held-out R_{h}@0.25	$\Delta\mathrm{mAP}_{\mathrm{cd}}$	ΔR_{h}
M0 frozen	48.172 ± 2.929	17.885 ± 0.662	—	—
M2 entropy	48.432 ± 3.096	17.108 ± 1.043	$+0.260$	-0.777
M3 EMA consistency	48.221 ± 2.977	17.349 ± 1.179	$+0.049$	-0.536
M5 complete	48.227 ± 2.912	17.652 ± 1.223	$+0.055$	-0.233
M5 without gate	48.234 ± 3.070	17.269 ± 0.656	$+0.062$	-0.615
M5 without protection and preservation	48.158 ± 3.011	17.898 ± 1.560	-0.013	$+0.013$
M5 without source contraction	48.207 ± 2.945	17.530 ± 1.094	$+0.035$	-0.355
SoftRestore	48.167 ± 3.056	16.953 ± 0.341	-0.005	-0.932

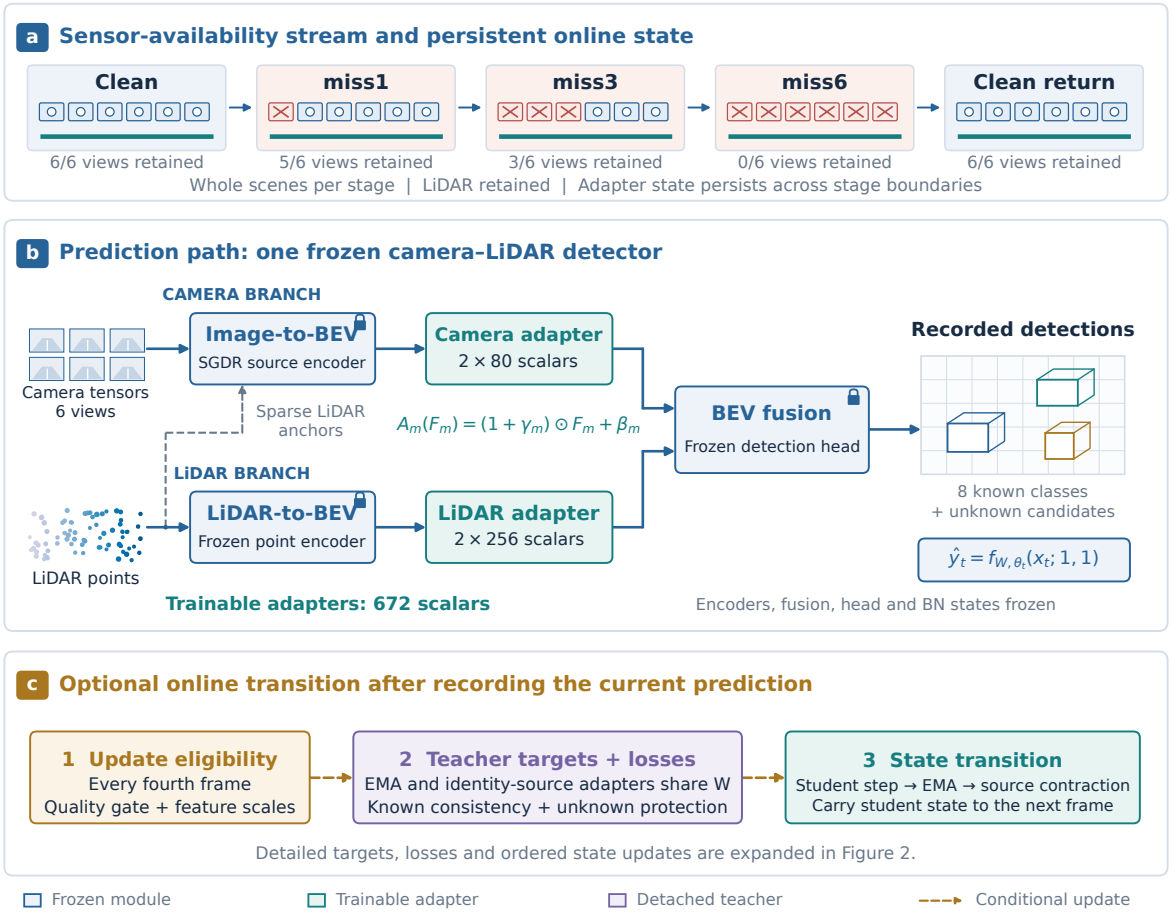

Figure 1: Continual camera–LiDAR test-time adaptation. (a) The input stream passes through five stages with simulated loss and return of camera views; LiDAR remains available. Missing-view icons show counts, and the identities can differ by frame. (b) The camera and LiDAR branches use channel-affine adapters before frozen BEV fusion; the LiDAR anchor branch also informs the inherited camera-depth route (SGDR, sparse geometry-anchored depth routing). (c) A prediction at unit scales is recorded before the optional student update. Teacher and student adapter states share frozen detector weights. Sensor and box drawings are schematics.

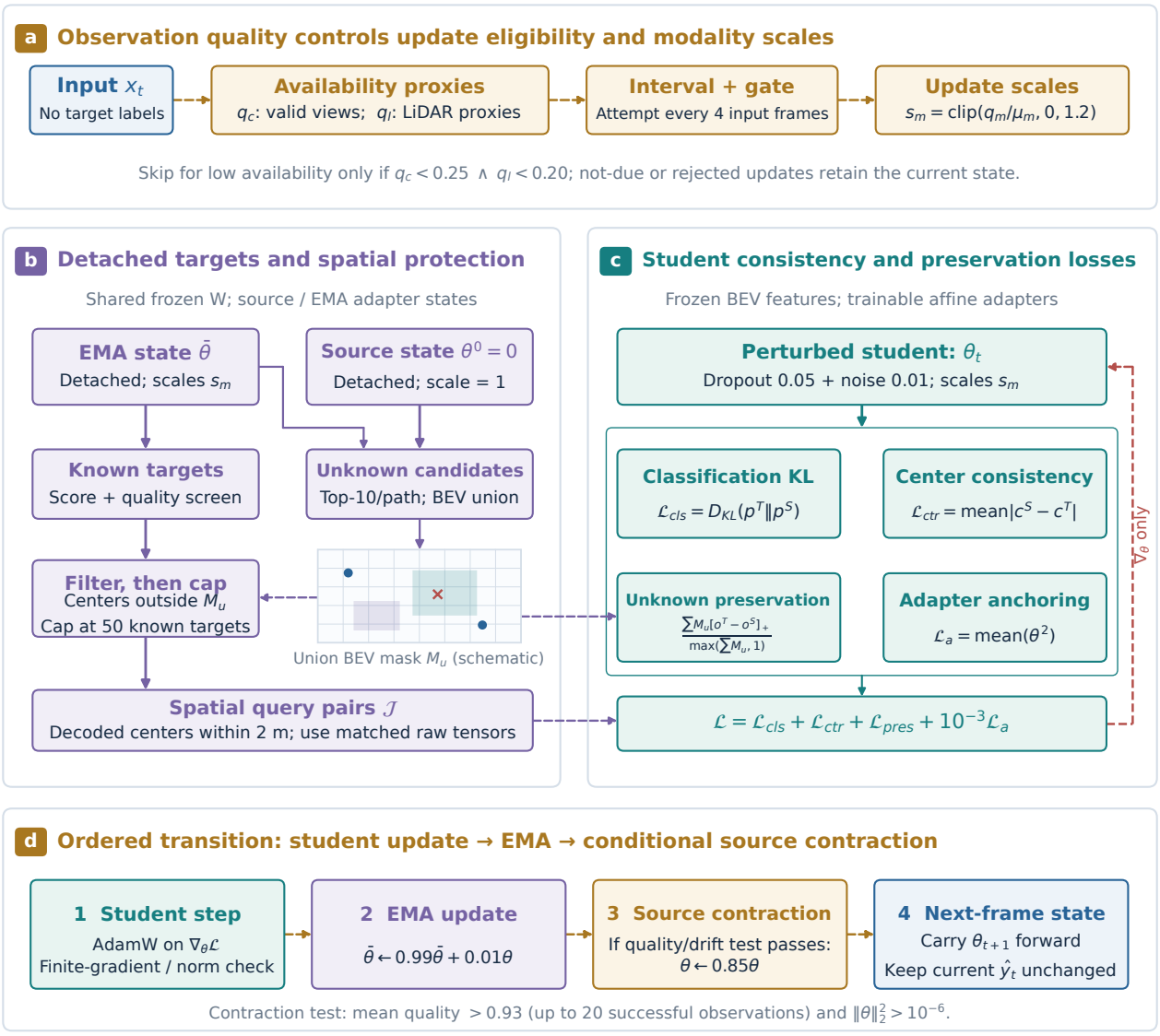

Figure 2: Update path of the complete configuration. (a) Availability checks set update permission and modality scales. (b) Detached EMA and source predictions form the protection mask; known targets are filtered, capped and paired to student queries. *Unknown candidates* in the diagram are class-agnostic proposals. (c) Classification and center consistency, protected objectness and adapter anchoring define a single student objective. (d) EMA averaging follows a successful student step, followed by source contraction if the quality and drift conditions hold. The quality window counts successful controller observations.

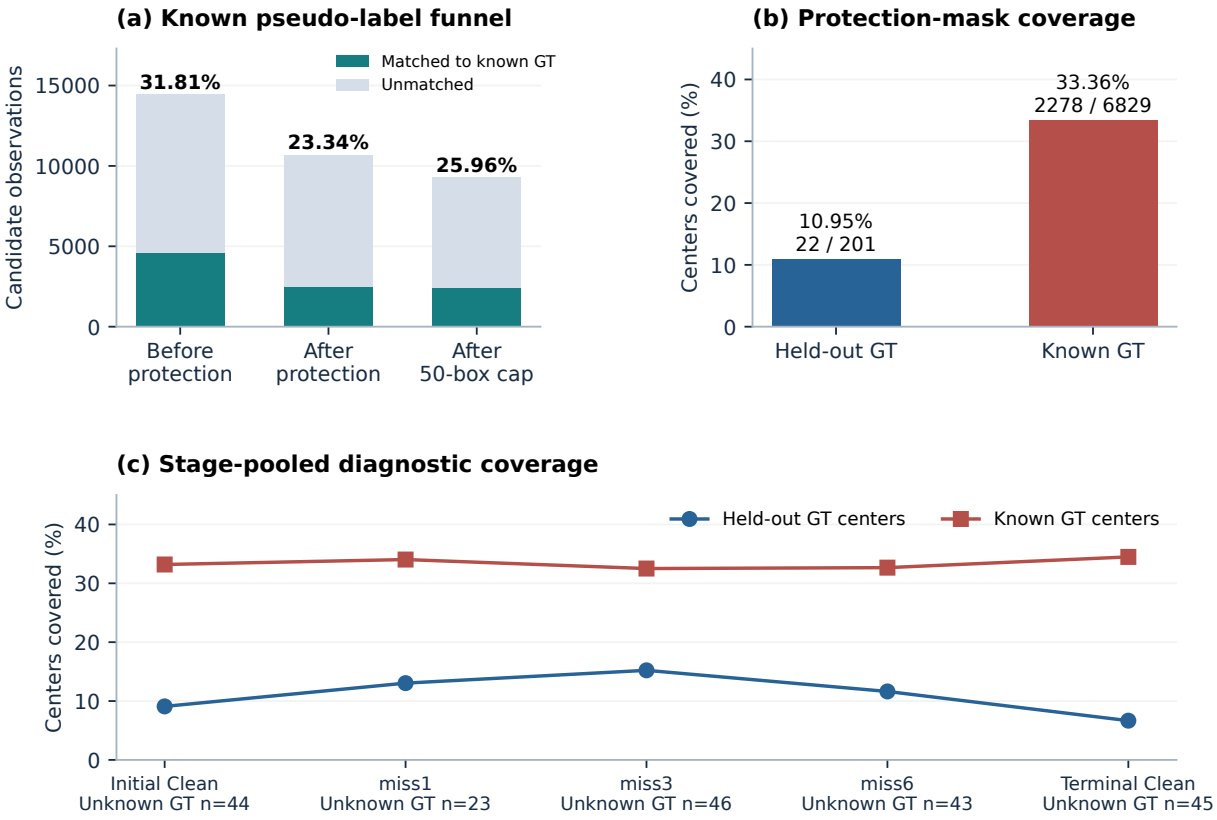

Figure 3: Protection selectivity in the separate mechanism replay (293 observations). (a) Known pseudo-label counts and matching precision at each filtering step. (b) Ground-truth center coverage for known and held-out categories, with separate pooled denominators. (c) Coverage by replay stage; held-out denominators are given below the stage labels. Stage contents differ, so the panels do not show a controlled severity response.

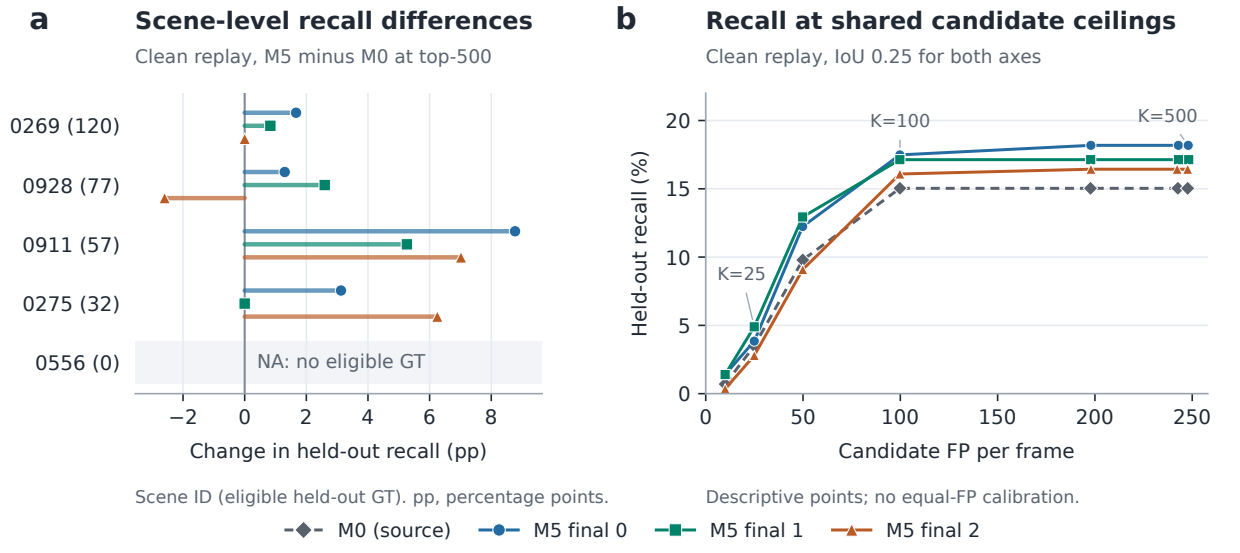

Figure 4: Frozen final-state performance on the intact-view, scene-disjoint probe. (a) Per-scene differences in held-out recall at 3D IoU 0.25 for each of the three final M5 states relative to M0. Recall is undefined in scene 0556, which has no eligible held-out observations. (b) Pooled recall and candidate false positives per frame, both matched at IoU 0.25, under common per-frame candidate ceilings. The 200 frames and 286 eligible held-out observations are shared across states. The displayed points describe candidate-budget sensitivity and do not represent independent equal-FP calibration.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- LimitsofcontinualtesttimeadaptationforcameraLiDAR3D.pdf
- SupplementaryInformation.pdf