

OpenRiO: A Community Reproduction of RiO-DETR with Practical Accuracy and Efficiency Improvements

Zhicheng Du Lan Ma*

Tsinghua Shenzhen International Graduate School, Tsinghua University

duzc21@mails.tsinghua.edu.cn malan@sz.tsinghua.edu.cn

Technical report, October 16, 2026

Abstract

We present OpenRiO, an independent community reproduction of RiO-DETR that turns competitive checkpoints into a baseline for further improvement. Selected single-view EMA results on DIOR-R reach 71.891, 75.020, 76.448, and 76.199 AP50 for Nano through Large, against the paper’s 71.92, 74.44, 75.73, and 76.11. Fresh inference and independent polygon matching verify these endpoints. The reproduction process reveals why a score alone is an incomplete reference: normalization state can repair or damage a model, and a continuation recipe that helps one checkpoint need not help another. We retain these conditional outcomes as reusable experimental guidance. Building on fixed reproduced models, we distinguish available candidate geometry from its use in the final predictions. Cross-view rescoring improves AP75 by 1.61 and 1.51 points with selected boxes fixed in two Nano lineages; coordinate-only uniform consensus adds 1.16, 0.67, 0.65, and 0.57 points over agreement-selected anchors across four paths. These studies retain their own checkpoints and controls. Separately, output-equivalent CUDA Graph execution reduces median latency from 34.17 to 30.36 ms in a declared shared-device measurement. OpenRiO contributes verified community artifacts, evidence about reproduction decisions, and controlled accuracy and execution improvements. Extra-Large and additional-dataset reproduction continue. We credit the original detector and existing techniques, and disclose implementation differences, adaptive test use, duplicate-image sensitivity, and missing independent training seeds.

1 Introduction

An open detector becomes a useful baseline when others can recover its behavior and improve it without repeating an opaque search. A benchmark score is necessary, but it does not specify the training history, checkpoint state, evaluation conventions, or cost of obtaining that score. This problem is particularly concrete for RiO-

DETR [11]: its public repository provides a foundational implementation while explaining that the full research implementation cannot be released [10]. The community needs an independently inspectable implementation with competitive performance and a clear account of its differences.

We develop **OpenRiO** around two goals. First, we reproduce the paper’s performance as closely as our open implementation and documented recipes permit. Nano approaches the reported reference; Small, Medium, and Large exceed their references. Fresh full-set inference and independently implemented metrics verify the selected states. These results establish numerical competitiveness under disclosed conditions, not identity to unavailable code or robustness across independent training seeds. Extra-Large and additional datasets remain in progress.

Second, we use the reproduced models as fixed starting points for improvement. The reproduction itself shows why this separation matters. Changing normalization buffers can recover a large deficit in one lineage but harm another. A matched learning-rate continuation helps Small and Medium but not the tested Nano endpoint. A plateau therefore describes a state under a recipe, rather than proving that the architecture has exhausted its capacity. Recording the successful interventions together with their counterexamples makes the resulting baseline more reusable than reporting only its best score.

The same discipline guides optimization. In frozen predictions, a well-localized candidate can lose to a more confident, poorly localized one. Changing only scores tests this ordering problem; changing only coordinates tests geometric selection. Both yield complete-set gains under their specified controls. We then test transfer across scales and to one YOLO detector, while separately evaluating prediction-preserving execution. The result is a connected account of building, understanding, and improving an open detector, with neither borrowed techniques nor exploratory trials presented as architectural inventions.

Our contributions are: (i) competitive, independently evaluated community checkpoints with explicit state and protocol provenance; (ii) controlled reproduction lessons

*Corresponding author.

that identify when a training or state intervention helps and where it fails; and (iii) measured inference-accuracy and execution improvements on fixed baselines. The original detector’s methodological credit remains with its authors. The community contribution is the open implementation, documented recipes, analysis, and verified practical utility. The preview packages inference, evaluation and the positive optimization procedures now; portable multi-stage training and independent-seed validation remain explicit release milestones.

2 Context and Attribution

Reproduction as an experimental foundation.

RiO-DETR addresses real-time oriented detection [11]. OpenRiO independently studies a reconstruction and a community variant, preserving the original method’s credit. Shared tools such as MMRotate [36] and transparent implementations such as CleanRL [13] motivate the inspectability of configurations, evaluation, and training records. Training-recipe studies [27] also illustrate why improved optimization of an established architecture should be distinguished from architectural intervention. A related protocol audit in training-free 3D retrieval shows how normalization and aggregation can confound comparisons of local descriptors [7]. In oriented detection, we apply this principle by identifying checkpoint state and evaluation conventions before attributing a gain to an intervention.

Existing community implementations and deployment.

Independent RiO-inspired components already appear in the public RiO-RTDETR Oriented implementation [20], including rotated attention and periodic angle handling. Its inspected release provides training configurations, while we found no accompanying paper-level reproduction table or trained release artifacts; this is an evidence boundary, not a claim that the implementation fails. LibreYOLO integrates the authors’ RT-DETRv2-OBb checkpoints for inference and validation [16]. TensorRT-DETR already provides oriented inference with CUDA Graph execution and several memory strategies [4]. OpenRiO therefore does not claim the first community implementation or first such deployment support. Its contribution concerns verified reproduction results, experimental lessons, and measured improvements under explicit controls and cost scopes.

Quality and aggregation.

Localization-aware confidence is established prior art; VarifocalNet explicitly learns an IoU-aware classification score [31]. Weighted boxes fusion aggregates predictions from detectors [25]. Our study does not introduce the distinction between classification and localization, or the general idea of combining boxes. It examines frozen, oriented predictions through controls that separate ranking, suppression, and

representative selection. A candidate medoid retains an observed geometry, whereas coordinate fusion can construct a new one. This distinction motivates an experiment, not an automatic novelty claim.

Studying capabilities before adding components.

Highly Compressed Tokenizer Can Generate Without Training investigates properties of a pretrained tokenizer and uses them for test-time optimization [2]. We adopt the methodological order of observation, intervention, and a minimal procedure. No equivalence between detector queries and that work’s latent tokens is assumed. The frozen inference rules in our primary comparisons do not update model parameters. Parameter adaptation under explicit domain shift, related to methods such as Tent [26], is a separate study detailed in Appendix D.

3 What Makes a Reproduced Baseline Useful?

OpenRiO asks two connected questions: can an open implementation attain the paper’s performance, and what can be improved once that implementation is fixed? Answering the second requires a precise answer to the first. We identify a baseline by its architecture, learned parameters, non-parameter state, input pipeline, and evaluation protocol. A checkpoint filename or a close AP value specifies only part of this object.

A plateau belongs to a training state and a recipe.

A small improvement after more epochs does not distinguish insufficient optimization from an exhausted representation. Our same-parent continuations ask a narrower, answerable question: does a specified intervention improve this state at a matched horizon? A learning-rate change helps Small and Medium but not the tested Nano endpoint. A buffer correction can help one lineage while damaging another. These contrasts support conditional recipes; they do not establish capacity or hardware as the cause of a remaining gap. Matching convergence opportunities before interpreting model differences is also central to the experiments of He et al. [9].

Available geometry and useful predictions are different.

An existing candidate can overlap an object well and still be ranked below a worse candidate, suppressed, or redundant with another prediction. Fixed-box rescoring tests ordering; coordinate-only replacement tests localization while retaining the same output identities. Both can improve AP75 in the tested pools. This supports two available opportunities, rather than a single explanation that the detector simply needs more proposals. Error analysis of detection outputs has substantial precedent [3]; our controls connect such observations to deployable, label-free inference rules. Ground-

truth-assisted candidate selection remains an offline diagnostic.

Useful optimization must name what it preserves.

Execution changes are tested for prediction equivalence before timing. Accuracy changes are tested against the same source model and candidate budget before attribution. The graph-execution result and the multi-view accuracy result therefore represent separate improvements, not an unmeasured combined operating point. Likewise, changing router values or bounding parameter drift establishes an internal property, not useful conditional computation or improved target accuracy. These distinctions determine the controls in the remaining sections.

4 OpenRiO: Reproduction and Evaluation

4.1 Model and training identity

The Nano community variant uses a public five-dimensional query design and has 4,012,456 parameters. The Compact Nano reconstruction uses a four-dimensional query and a compact positional mapping, with 4,027,432 parameters. Their selected single-view AP50 values are 71.8908 and approximately 70.0025. Small and Medium use the Compact query design with 8,267,680 and 19,210,054 parameters. These are identifiable community implementations, not interchangeable labels for the author’s exact models. Neither matching a parameter count nor approaching a score establishes equivalence to unreleased code.

Continuation experiments investigate training duration, initialization, learning rate, batch size, normalization state, moving-average weights, and rotation augmentation. A fixed-parent seed study measures continuation variability; it is not independent training from scratch. Existing bounded continuation and paired Compact experiments remain part of the reproduction audit. The community reference is not declared globally converged merely because its score is close to the paper.

Why state is part of the baseline. In one recorded lineage, changing two batch-normalization buffers while holding learned parameters fixed repairs approximately 5.91 AP50. Related calibration interventions do not restore a declining Compact lineage. This contrast across checkpoint histories makes a narrower point than attributing all reproduction gaps to normalization: checkpoint state can be a decisive variable, and the effective repair depends on the training history. A reproducible reference must specify this state rather than identifying a model only by its learned weights.

4.2 Evaluation and provenance

We evaluate all 11,738 DIOR-R test images and report rotated VOC 2007 eleven-point AP50 and AP75 separately. These quantities are not COCO AP averaged over IoU thresholds. The current test annotations contain no difficult objects marked for exclusion. Predictions are capped at 300 detections per image. Five-view experiments use the original image, a horizontal flip, and rotations by 90, 180, and 270 degrees. All boxes are mapped back to the original image before comparison.

Independent XML parsing, polygon intersection, matching, and AP computation reproduce five complete prediction sets to a maximum absolute discrepancy of approximately 2×10^{-6} AP points. Spread-index re-inference checks reproduce cached outputs exactly for 16 images in each of two model paths. These are checks on specified paths, not proof that every implementation detail matches the author’s private code.

The data audit finds 23 test images whose bytes also occur under different training identifiers. All pairs occur in the locally retained source archives, and the split lists and test annotations match those archives. Standard comparisons retain all test images. A separate 11,715-image sensitivity analysis excludes these duplicates. We have not independently authenticated the archives against the publisher or exhausted near-duplicate detection.

Selection limits. Training and inference recipes have been selected with feedback from this test set, including a historical 518-image diagnostic subset. Complete-set evaluation prevents subset-only reporting, but does not create an untouched validation set. No ground truth enters prediction-time scoring or representative selection; ground truth does enter experimental selection and the offline analyses. Independent training seeds, frozen-rule transfer, and geographically independent evaluation remain distinct requirements.

5 Lessons from Reproducing RiO-DETR

Numerical reproduction and optimization evidence differ. Table 1 establishes proximity to the paper’s targets under our declared protocol. It does not explain how the remaining gap arises. A normalization-state repair can identify a faulty checkpoint path, but cannot establish that a correctly evaluated model has finished learning. Conversely, calibrating every model is not a justified remedy: an earlier selected Small checkpoint falls from 74.776 to 74.520 AP50 after the tested calibration. Uncalibrated EMA remains the primary reproduction convention; raw weights and calibrated states are separately labeled alternatives.

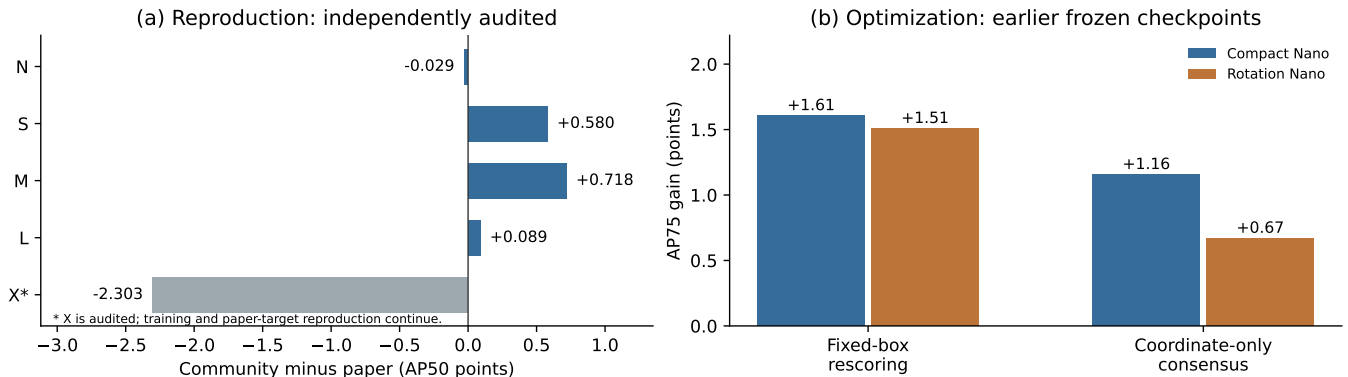


Figure 1: **A competitive open baseline, followed by controlled improvement.** Left: independently audited single-view EMA endpoints relative to the paper; Extra-Large remains in progress. Right: AP75 differences from frozen-pool interventions in two earlier Nano lineages. Ranking changes keep boxes fixed; uniform geometric selection keeps selected identities, scores and counts fixed. The two interventions use different immediate controls and their bars must not be added. The right panel does not measure improvements on the latest left-panel checkpoints.

Table 1: Selected community reproduction results at the October 15 evidence cut. Full-test, single-model, single-view, uncalibrated EMA AP50; no test-time augmentation. These are selected checkpoints from one training ancestry per path, not independent-seed means. Fresh inference and independent metric checks verify each selected endpoint.

Scale	Paper	Community	Difference
Nano	71.920	71.891	-0.029
Small	74.440	75.020	+0.580
Medium	75.730	76.448	+0.718
Large	76.110	76.199	+0.089
Extra-Large*	77.430	75.127	-2.303

A matched intervention can reopen a plateau.

For Small and Medium, we branch from the same within-scale checkpoint and change the learning rate from 2.5×10^{-7} to 10^{-6} . Within each pair, batch size, seed, rotation augmentation, frozen normalization, optimizer initialization, and EMA rule are held fixed. Small uses batch 8; Medium uses batch 4. At 24 additional epochs, Small reaches 74.776 versus 74.396 AP50, a 0.380-point difference. At the common observed horizon of 12 epochs, Medium reaches 75.949 versus 75.822, a 0.128-point difference. Figure 2 shows both trajectories relative to their shared starting scores. These selected, single-seed comparisons support the higher rate at these checkpoints and horizons. They do not establish final convergence, a universally optimal rate, or a causal relationship between model size and improvement.

Nano separates a training plateau from a state correction. The established rotation tail improved by only 0.0078 AP50 in its latest 24-epoch window. This justified stopping repetition of that recipe, not declaring

a global optimum. A subsequent matched 24-epoch comparison starts from the same 71.8562 checkpoint, holds batch eight, seed, augmentation, frozen normalization, and primary EMA decay 0.999 fixed, and compares learning rates 2.5×10^{-7} and 10^{-6} . The endpoints are 71.8543 and 71.8030 before calibration. The lower-rate path supplies the selected intermediate primary EMA score of 71.8908 in Table 1. Thus, the higher-rate advantage previously observed in Small and Medium does not extend to this Nano endpoint.

Calibrating two normalization layers from 256 training-side images changes the higher-rate endpoint from 71.8030 to 71.9783, without changing learned parameters. This is a separately labeled post-training state correction, not the primary uncalibrated reproduction result. Fresh full-set inference gives 71.97825 AP50; independent polygon matching gives 71.97825, with an AP75 discrepancy below 0.00023 points. Three registered random subsets of 256 training images yield 71.8519, 71.9160, and 71.9594 AP50, with mean 71.9091. Every subset improves the same 71.8030 parent, but the spread is 0.1075 points and none reaches the original deterministic subset’s 71.9783. The evidence supports a repeatable direction in this small subset check, with sensitivity to calibration images. It does not establish subset-independent magnitude or identify the deterministic subset as optimal. These subset seeds are not independent detector-training seeds.

A usable stopping decision. The table freezes independently audited endpoints. Extra-Large retains 75.126961 training-run AP50; fresh inference gives 75.126405, and independent polygon matching gives 75.126386. Its latest 24-epoch window ends at 75.066902 EMA AP50 and does not improve the selected best. One flat window is insufficient for the registered consecutive-window stopping rule. The asterisk marks an ongoing re-

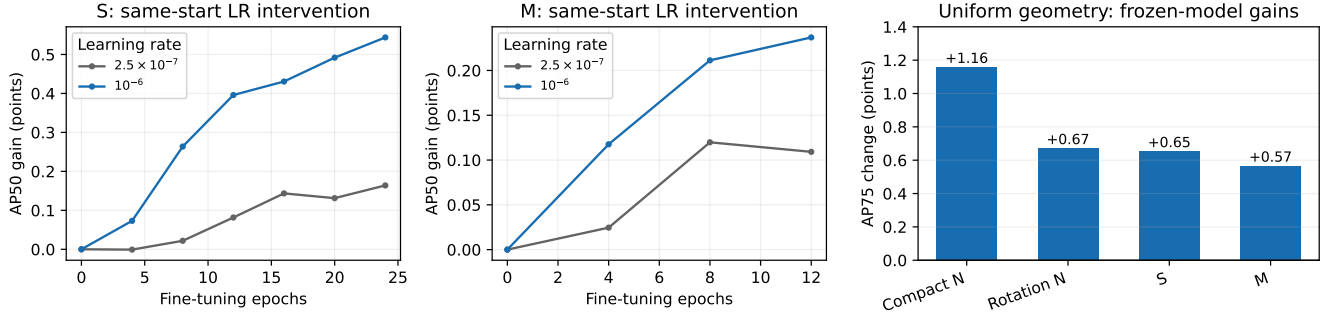


Figure 2: **The useful intervention depends on the experimental state.** Left and middle: matched learning-rate continuations of Small and Medium, plotted from their respective common starting checkpoints. The different observed horizons must not be interpreted as a controlled comparison of scale. Right: full-test AP75 gains from uniform coordinate selection relative to fixed agreement-selected anchors. Scores, labels, and output identities are held constant. The positive effect extends to all four tested community paths. The inference checkpoints predate the selected training results and are held fixed within each comparison.

production; its weights are not among the four released primary references.

Both registered Small learning-rate paths have stopped after two consecutive windows meeting the stated plateau rule. The higher-rate path retains 75.0200 AP50; its final two 24-epoch windows do not improve the selected best. This is operational saturation of those recipes, not proof of a global optimum. Medium and Large also reach their empirical stopping rules, retaining 76.4483 and 76.1990. Extra-Large with quarter-turn training reaches 75.1270 and remains 2.3030 points below the paper reference. Its continuation and FAIR1M training remain resource priorities. All five selected research states now pass fresh full-set inference and independently implemented polygon matching and AP. Proximity to the paper alone is not a stopping condition. The ordinary Extra-Large continuation was stopped after a flat full window and declining subsequent snapshots, with its committed checkpoint retained. This is a resource decision about that recipe, not convergence of every Extra-Large route.

The earlier relocated release candidate passes full-set evaluation for Nano, calibrated Nano, Small and its earlier 76.1754 Medium. The October 14 source endpoints and exported tensor-only states have explicit, separately checked identities. Source-checkpoint evaluation, export identity, and relocated runtime checks are distinct receipts. Multi-view fusion, score calibration, ensembles and parameter adaptation belong to the subsequent optimization objective and never substitute for the reproduction entries in Table 1.

The cross-size contrast changes the question from whether the smallest model has exhausted its capacity to which part of the optimization and checkpoint state still limits the observed score. The matched endpoint comparison supports a state-dependent recipe recommendation. It does not identify loss curvature, representation

capacity, or hardware as the cause. The earlier matched-horizon curves in Figure 2 retain their original checkpoints and horizons; the updated selected-score table is not a replacement for those controlled comparisons.

A batch has more than one measure. Resource-constrained reproduction preserves a nominal image batch more easily than it preserves the training objective. In FAIR1M, independently normalizing each micro-batch weights sparse and dense images differently from normalizing their combined targets; physical batch size separately controls normalization statistics. We verify these distinctions with the actual detection losses and a fixed-image gradient probe, then register paired accuracy experiments (Appendix N). This observation refines the reproduction recipe rather than introducing a new detector. It also motivates a protocol check: our FAIR validation crops and the original logs contain different numbers of targets, so their AP values are not yet an established paper-gap comparison.

Dataset coverage is an explicit work in progress. The preview prioritizes independently checked DIOR-R Nano through Large artifacts. FAIR1M-2.0 Medium and Extra-Large remain in their registered 36-epoch schedules. At the October 15 cut, Medium’s epoch-thirty-two EMA obtains 40.4801 AP50 after merging predictions over 8,287 original validation scenes and 34.8772 under native patch evaluation of 71,217 patches. Extra-Large’s epoch-fourteen EMA obtains 35.7347 and 25.1252, respectively. These are different protocols, whose equivalence to the paper’s 43.6 and 47.4 references has not been established. We therefore do not interpret either measurement as a verified paper gap or a converged FAIR1M result. Training excludes 54 audited byte-identical training/validation originals. Fixed-recipe completion triggers full raw/EMA verification before the reg-

istered state-preserving continuation branches. DOTA official-server results remain pending; local export is not a substitute for the official metric.

The October 15 protocol census confirms different crop and target counts from the original logs; see Appendix N.

6 Improving the Reproduced Detector

The community checkpoints allow us to improve inference without changing the trained detector. We first expose complementary geometric hypotheses with multiple views, then use their agreement for scoring and representative selection. The controls distinguish improvements due to more candidates from improvements in how an existing pool is used. The following formulations define the two simple inference rules; supporting statements and proofs are collected in Appendix C. Let the inverse-mapped candidate pool be

$$\mathcal{B}(x) = \bigcup_{v=1}^V \{(b_i, s_i, c_i, v_i)\}_{i \in \mathcal{I}_v}, \quad V = 5. \quad (1)$$

The network and cached pool are fixed throughout each paired comparison. For candidate i , define cross-view support

$$u_i = \frac{1}{V-1} \sum_{v \neq v_i} \max_{j: v_j=v, c_j=c_i} \text{IoU}(b_i, b_j), \quad (2)$$

where the maximum over an empty set is zero. Our current score is $s'_i = s_i(0.75 + 0.25u_i)$. The coefficient is frozen for the registered controls; it is not refit for individual classes or images. Support is agreement among predictions, not a calibrated probability of correctness.

6.1 A restricted change that diagnoses ordering

Let S be the identities selected by classwise rotated NMS at IoU 0.5 followed by a 300-detection cap using the original scores. The fixed-set intervention returns (b_i, s'_i, c_i) for exactly $i \in S$. It leaves every coordinate, category, and count unchanged. Its AP difference is attributable to ordering within this fixed set, including any change in score-ordered duplicate matching. A second variant applies suppression using s' and may select different identities; its additional gain cannot be called a score-only effect on the original set.

A strictly increasing classwise score transformation would preserve AP on this fixed set if the implemented matching and ranking preserve the same total order, including ties. The gain therefore depends on candidate-dependent reordering, rather than score rescaling alone (Appendix C). Appendix D gives a numerical counterexample when precision conversion violates this condition.

This does not establish that support is the best possible quality estimator.

6.2 Selecting an observed geometric representative

For an anchor, construct a same-class group G containing that anchor and its highest-overlap match from each other view when IoU is at least 0.55. Groups with fewer than three members retain the anchor. Otherwise, normalized original scores define weights $w_j = s_j / \sum_{k \in G} s_k$. The chosen representative is

$$\hat{b} = \arg \max_{b_i: i \in G} \sum_{j \in G} w_j \text{IoU}(b_i, b_j). \quad (3)$$

The finite support distribution $Q_G = \sum_{j \in G} w_j \delta_{b_j}$ makes this an empirical risk minimizer for loss $\ell(b, z) = 1 - \text{IoU}(b, z)$. It avoids averaging angle parameterizations by returning an observed box. It does not establish that the scores are posterior probabilities or that the group corresponds to a real object.

Agreement is a surrogate for localization quality, not a guarantee of it. An elementary conditional risk bound in Appendix C makes the required assumption explicit: candidate agreement must approximate the target geometric risk. Shared view errors or groups containing adjacent objects can violate it. The coordinate-only and random controls below test the utility of structured selection without assuming an AP-improvement guarantee.

6.3 Controls that can reject the explanation

The initial representative experiment replaces candidate geometry before suppression, retaining agreement scores. It therefore couples geometry with selection. The fixed-identity experiment instead freezes the agreement-selected anchors, classes, scores, order, and number of outputs. It replaces coordinates using (i) score-weighted representatives, (ii) uniform representatives, or (iii) a seeded random member of the same group. It does not rerun NMS or top- k after replacement, even if two anchors acquire identical coordinates.

Positive AP75 changes for both structured representatives relative to the random control would support useful geometric structure in the groups. If the uniform choice suffices, score weights are unnecessary. If only the original joint procedure improves, the claim must concern selection interaction rather than pure coordinate repair. Random selection uses one registered seed, so its variability is not exhausted by this control. These alternatives are determined by complete-set results.

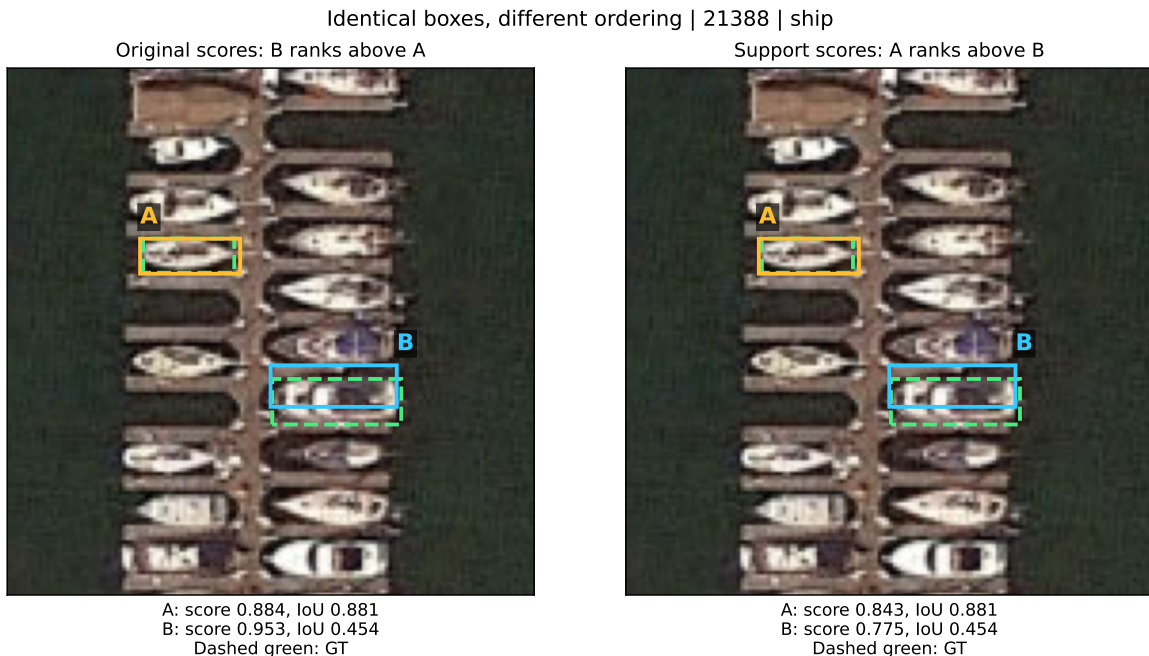


Figure 3: **An ordering intervention without a geometric change.** The displayed candidate boxes and their IoUs to same-class annotations are identical in both panels. Candidate-dependent support reverses their score order. The example is selected by an explicit full-set scan for a visible, positive-overlap localization contrast; it is explanatory rather than a random sample. Only the selected boxes and associated annotations are highlighted.

6.4 A ceiling imposed by the candidate pool

For class-aware coverage at threshold τ , define

$$C_\tau = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbf{1}[\exists i : c_i = c_g, \text{IoU}(b_i, g) \geq \tau]. \quad (4)$$

Any procedure that returns only members of this fixed pool has recall at most C_τ : every matched ground truth must have at least one qualifying pool member. Coverage need not be attainable because multiple objects can share a candidate. Scoring and medoid selection cannot exceed this ceiling without changing the pool. Our new full-set coverage analysis uses cached top-300 predictions per view, and is consequently distinct from the historical all-decoder-query subset diagnostic.

7 Accuracy Improvements and Practical Performance

7.1 From a visible error to an identifiable intervention

The visual analyses in RiO-DETR [11] connect internal sampling, successive refinement, and feature organization to their design hypotheses. Here the intervention is at prediction time: we ask what changes when a frozen detector’s outputs are reordered or geometrically re-

lected. Figure 5 makes the two controls explicit. Annotations enter evaluation and post-hoc case selection only.

Agreement is useful geometry, but not an oracle.

A new full-set diagnostic examines the frozen Medium geometry experiment. Among 80,618 changed output identities meeting a fixed score and size filter and retaining the same nearest same-class annotation, uniform consensus moves 2,601 outputs above IoU 0.75 and 1,558 below it. Random selection from the same candidate groups instead gives 2,127 upward and 2,509 downward crossings. The average IoU changes are +0.00575 and −0.00109. These are overlapping output identities, not independent objects or true-positive counts. Correct one-to-one matching and global score ordering give the complete-set AP75 changes of +0.566 and −1.509 points (Figure 6).

The controlled difference supports geometric information in cross-view agreement beyond merely choosing another candidate. The downward crossings show why it is insufficient to equate agreement with ground-truth quality. Selected ship and airplane examples visualize both outcomes in Appendix K; the examples illustrate the mechanism, while the full-set comparisons establish its measured scope. They do not prove which latent representation or view error caused a specific failure.



Figure 4: **A geometric intervention and its failure mode.** Original and replacement predictions retain the same score, category, and anchor identity. The two rows illustrate upward and downward crossings of IoU 0.75 against the same annotation. Candidates are selected by the largest IoU change among eligible full-set cases, with the rule and image identities recorded. These deliberately strong illustrations explain what changes; they do not estimate how often the method helps.

7.2 Complete-set interventions

Table 2 separates scoring an unchanged detection set from selecting different identities. Both Nano lineages exhibit a fixed-set improvement. The original geometric-representative procedure improves strict-IoU performance relative to agreement selection with almost unchanged AP50. These are related implementations, not evidence across unrelated detector families. The rotation-lineage multi-view cache uses a 71.7993 single-view source, not the newer 71.8562 checkpoint. The inference comparisons below retain their own frozen sources; they are not inference measurements on every latest training result in Table 1.

7.3 Coordinate-only controls

Compact coordinate-only differences (AP50 / AP75) from its fixed-identity agreement control are weighted: +0.0555 / +0.9156, uniform: +0.0152 / +1.1585, random: -0.8335 / -0.9819. These point estimates isolate coordinate replacement under the specified groups and selection identities; their uncertainty is assessed separately.

Rotation coordinate-only differences (AP50 / AP75) from its fixed-identity agreement control are weighted: +0.0257 / +0.4566, uniform: -0.0809 / +0.6715, random: -1.0553 / -1.5600. These point estimates isolate coordinate replacement under the specified groups and selection identities; their uncertainty is assessed separately.

Two interventions answer different questions

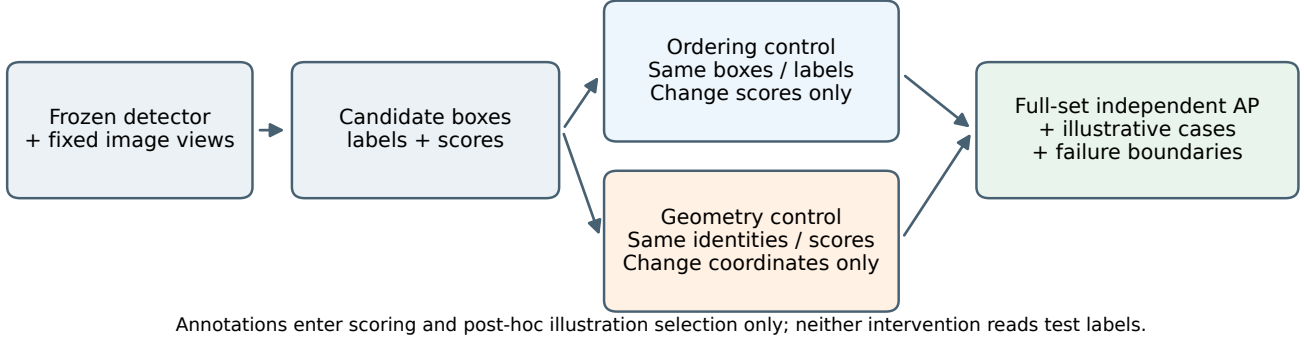


Figure 5: **The intervention determines the explanation.** Score-only and coordinate-only comparisons constrain different parts of a prediction. This schematic describes the experimental controls, not an attention map or a newly trained architecture. Full-set AP remains the performance criterion.

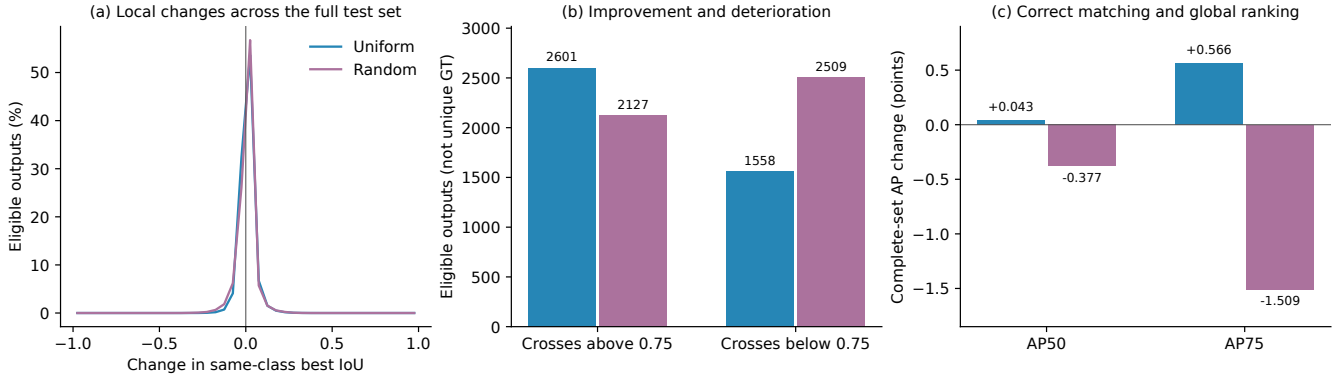


Figure 6: **A local geometric change and its global metric consequence.** The scan covers all 11,738 DIOR-R test images. Panels (a,b) condition on changed outputs with score at least 0.3, shortest side at least eight pixels, and unchanged nearest same-class annotation. Their counts are neither recall nor AP. Panel (c) uses every retained output in the complete-set evaluator; scores, labels, identities, and output counts are fixed in both interventions. The random control uses one registered seed.

7.4 Full-set candidate diagnosis

Compact: the analysis contains 124,443 valid objects, including 75,315 small objects. At confidence 0.3 and IoU 0.5, 20,663 of 25,770 small-object misses (80.2%) have no geometric cover in the original-view cache. Class-aware small-object coverage at IoU 0.75 rises from 40.52% to 56.94% in the five-view pool. This is an upper-bound candidate statistic, not matched recall or an all-query result.

Rotation: the analysis contains 124,443 valid objects, including 75,315 small objects. At confidence 0.3 and IoU 0.5, 18,907 of 23,604 small-object misses (80.1%) have no geometric cover in the original-view cache. Class-aware small-object coverage at IoU 0.75 rises from 43.19% to 58.73% in the five-view pool. This is an upper-bound candidate statistic, not matched recall or an all-query result.

The XML object count is 124,445, while the evaluated

valid-rectangle count is 124,443; two degenerate annotations are excluded by rectangle parsing. The same valid-object set is used by the independently checked metrics and these diagnostics.

7.5 Paired image-resampling analysis

Compact, fixed set rank minus geometry nms: AP75 difference +1.6082, paired image-bootstrap 95% percentile interval [+1.1488, +1.7913].

Compact, weighted geometry only minus agreement rank: AP75 difference +0.9156, paired image-bootstrap 95% percentile interval [+0.4027, +1.1263].

Compact, weighted geometry only minus random geometry only: AP75 difference +1.8975, paired image-bootstrap 95% percentile interval [+1.3531, +2.5707].

Rotation, fixed set rank minus geometry nms: AP75 difference +1.5104, paired image-bootstrap 95% percentile interval [+1.1091, +1.7819].

Table 2: Full-test results using identical frozen five-view pools within each lineage. Compact and the rotation lineage use different checkpoints. All values are AP points.

Inference	Compact		Rotation lineage	
	AP50	AP75	AP50	AP75
Ordinary NMS	72.903	45.516	74.247	46.996
Fixed-box scoring	73.427	47.124	74.710	48.506
Agreement selection	73.541	47.660	74.822	49.044
Joint representative	73.535	48.332	74.817	49.409

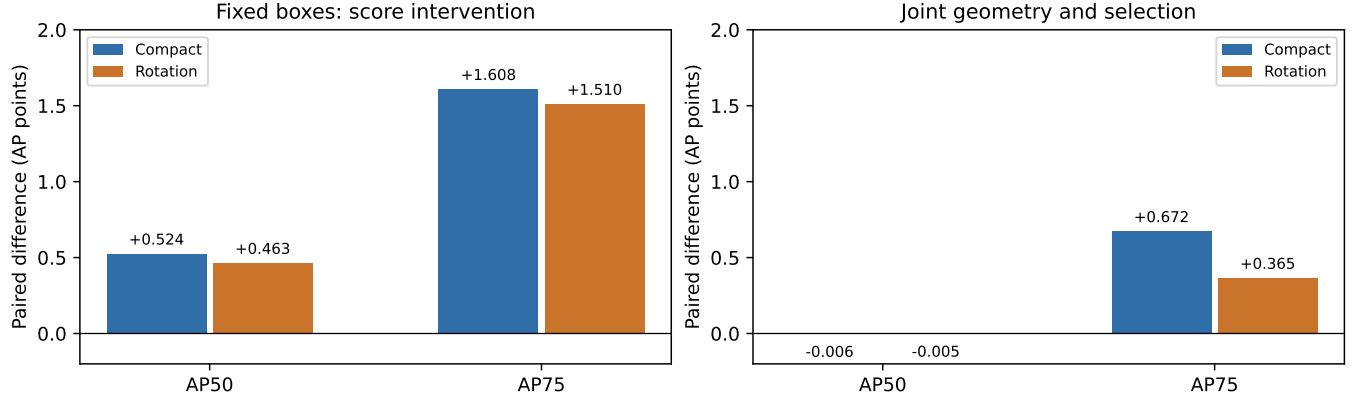


Figure 7: Paired changes on all 11,738 images. Fixed-set scoring is compared with ordinary five-view NMS; the joint geometric representative is compared with agreement-based selection. Different controls answer different questions. Point estimates alone do not establish significance. No geometric-only claim is inferred from the joint comparison.

Rotation, weighted geometry only minus agreement rank: AP75 difference +0.4566, paired image-bootstrap 95% percentile interval [+0.1158, +0.8345].

Rotation, weighted geometry only minus random geometry only: AP75 difference +2.0166, paired image-bootstrap 95% percentile interval [+1.5120, +2.6670].

Intervals are pointwise, exploratory, and conditional on the chosen checkpoints and test set. They do not correct for method selection, multiple comparisons, training variability, or correlations between images.

All six coordinate-only prediction sets also pass the independent polygon-intersection and AP recomputation. The maximum absolute discrepancy is 0.0000536 AP points.

Compact, uniform versus random representative: AP75 difference +2.1404, paired 95% interval [+1.5777, +2.7842].

Rotation, uniform versus random representative: AP75 difference +2.2316, paired 95% interval [+1.6156, +2.8250].

The uniform representative has at least the weighted representative’s AP75 in both lineages. It is a simpler candidate for strict localization, but its AP50 changes must be considered separately; it is not automatically superior on both metrics. A frozen independent validation should retain this tradeoff rather than attribute all benefit to score weights. The current comparison still uses the historically explored test set.

7.6 What transfers, and what fails to transfer?

The same coordinate-only uniform representative increases AP75 by 0.653 and 0.566 points on community Small and Medium, respectively, relative to their unchanged agreement-selected anchors. Their AP50 changes are -0.013 and $+0.043$. These sources have single-view AP50 of 73.982 and 74.790 and precede the later training continuations. Together with the Nano controls, they support useful geometric agreement across these related model paths. The differing gains are not evidence of a monotonic scaling law.

Table 4 reports the practical combined pipeline alongside ordinary multi-view suppression. Relative to the same checkpoint’s single view, five-view uniform geometry adds 2.360 AP50 points on Small and 2.722 on Medium. Relative to ordinary five-view NMS, it adds 2.530 and 2.160 AP75 points. These comparisons establish useful accuracy improvements on the reproduced models; the restricted controls identify which parts arise from ordering and geometry. The five model passes impose additional cost and are not the latency-optimized single-view setting.

The contrast between simple consensus and a more elaborate repair helps define the current method. Using a well-supported object’s appearance to guide local translations improves Compact Nano by 0.217 AP50, but has slightly negative AP50 effects on rotation Nano, Small,

Table 3: Coordinate-only intervention. Scores, selected anchor identities, labels, order, and counts are fixed within each lineage. No suppression or truncation is rerun after replacement.

Geometry	Compact		Rotation lineage	
	AP50	AP75	AP50	AP75
Unchanged anchors	73.541	47.660	74.822	49.044
Weighted representative	73.597	48.575	74.848	49.501
Uniform representative	73.556	48.818	74.741	49.716
Random group member	72.708	46.678	73.767	47.484

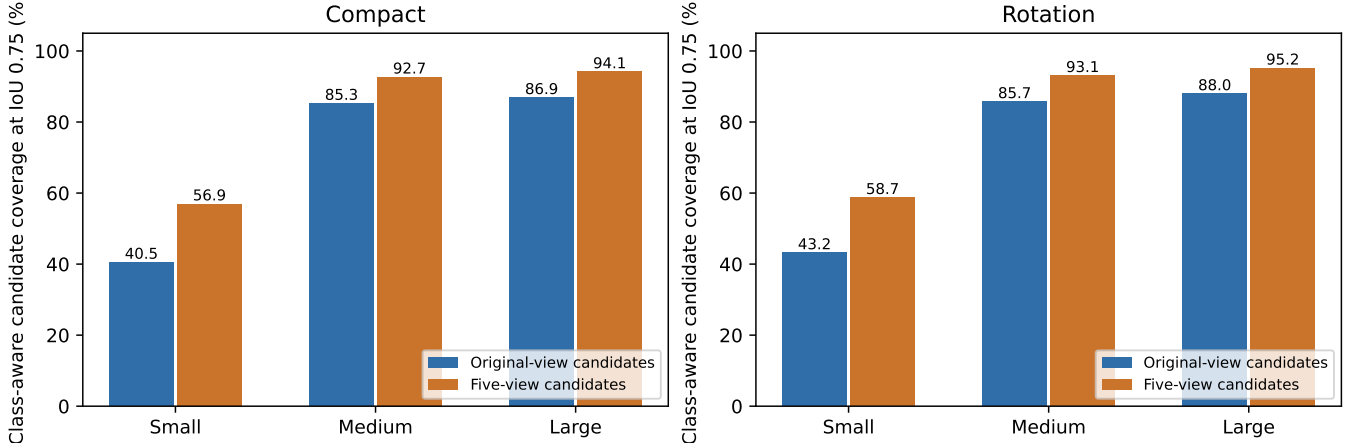


Figure 8: Class-aware candidate coverage on all valid test objects. Five views improve the upper bound most for small objects. The statistic uses cached top-300 predictions per view and allows candidate reuse; it is not one-to-one detection recall or coverage of every decoder query.

Table 4: Accuracy operating points on fixed community Small and Medium checkpoints. Five-view uniform geometry combines view expansion, agreement scoring, and coordinate selection. Its total gain must not be attributed to the final coordinate step alone. End-to-end latency of this complete pipeline remains to be measured.

Scale	Inference	AP50	AP75
Small	Single view	73.982	50.241
	Five-view NMS	76.068	50.786
	Five-view uniform	76.341	53.317
Medium	Single view	74.790	52.499
	Five-view NMS	77.137	53.270
	Five-view uniform	77.512	55.430

and Medium. It is therefore retained as a conditional result rather than a universal upgrade. In the explanatory A/B ship case, the local pool contains a proposal with IoU 0.756 for B, but the appearance rule selects one at approximately 0.471. The opportunity exists; the selection signal is insufficient. Appendix G reports full-set deltas, the visual case, and a diagnostic decomposition. This failure motivates better localization evidence while leaving the tested consensus gains intact.

7.7 A different prediction paradigm

Frozen rules transfer to a YOLO26-N oriented checkpoint evaluated on all 11,738 images. Fixed-box rescoring improves AP75 by 0.529 over ordinary five-view NMS. With agreement-selected identities and scores held fixed, uniform geometric selection adds 0.802 AP75, whereas random replacement loses 0.829. Thus, useful structure in the candidate pool is not restricted to the related DETR implementations. This is one additional trained checkpoint, not evidence of universal transfer.

A separate train-only quality readout uses frozen single-view features. Relative to its exactly matched native predictions, the visual readout adds 0.225 AP50 and 0.453 AP75. Metadata alone adds 0.059 and 0.275; class-preserving shuffled targets change the metrics by -0.008 and -0.038 . Classwise monotone scaling leaves both metrics exactly unchanged. The visual readout improves on the metadata control by 0.178 AP75, supporting useful additional feature information under this protocol. This small head is a learned addition, not a training-free method or a new detector architecture. It requires one detector pass plus feature pooling and scoring; complete deployment overhead is not yet measured.

Fresh native inference is identical across the five-view and learned-readout paths on every test image. It differs slightly from the historical prediction cache, whose AP50/AP75 are 69.5153/49.3750; all intervention deltas

Table 5: Full-set YOLO transfer and posttraining. Passes count detector forwards, not total cost. All rows share one detector checkpoint. Single-view learned read-outs preserve native boxes. Five-view geometric rows preserve agreement-selected identities and scores.

Method	Passes	AP50	AP75
Single view	1	69.517	49.361
Five-view NMS	5	72.496	50.012
Fixed-box rescoring	5	72.660	50.541
Agreement selection	5	72.706	50.709
Weighted geometry	5	72.776	51.129
Uniform geometry	5	72.927	51.511
Random geometry	5	72.474	49.880
Metadata quality	1	69.576	49.636
Visual quality	1	69.742	49.814
Shuffled quality targets	1	69.509	49.323

above use fresh matched predictions. With 999 paired image bootstrap replicates, the visual quality head's AP75 gain has a pointwise 95% percentile interval of [0.307, 0.585]; its gain over metadata alone has interval [0.079, 0.392]. The fixed-box multi-view gain and uniform coordinate-only gain have intervals [0.399, 0.681] and [0.543, 1.248]. Excluding the 23 exact duplicate test images retains a +0.451 AP75 quality-head gain. These conditional, exploratory intervals do not account for training seeds, geographic correlation, multiple testing, or historical test selection. The bootstrap uses the same independent polygon kernel as the point estimates; a faster kernel failed the unweighted parity gate and was not used.

7.8 Applicability revealed by controlled comparisons

Classifier-head test-time updates at two registered learning rates yield negligible full-set AP50 changes. Five stream orders of a FIFO cache yield approximately 0.006–0.022 AP50 gains but 0.118–0.152 AP75 losses. A refinement-stability score produces approximately +0.039 AP50 and −0.173 AP75. The tested graph-diffusion recipe also underperforms the simple agreement rule. These outcomes reject the tested recipes as compelling joint improvements; they do not reject all adaptation under domain shift.

Their role in the argument is specific. Agreement with earlier predictions or increased confidence is insufficient by itself to guarantee better geometry. Conversely, a full-set fixed-box scoring gain shows that geometric coverage is not the only limitation. These observations motivate a localized test of cross-view evidence without requiring that every unsuccessful trial be presented as a new scientific finding.

7.9 Sensitivity to exact duplicate images

Excluding the 23 byte-identical cross-split images leaves the representative's AP75 gain over agreement selection at approximately 0.630 for Compact and 0.365 for the rotation lineage. Standard and duplicate-excluded results are reported separately. This sensitivity reduces one specific concern about the measured delta, while leaving near duplicates, spatial correlation, and test-guided selection unresolved.

7.10 Execution cost is part of the method

Training-free acceleration also appears outside detection: AudioCache reuses layer computations in audio generation [24]. That work provides context for exploiting inference redundancy; our execution study measures CUDA Graph scheduling and checks output identity, with no layer-cache transfer experiment claimed.

On the current Nano checkpoint, CUDA Graph reproduces boxes and logits bitwise across all 11,738 test images. Paired shared-device measurements reduce median batch-one latency from 34.17 to 30.36 ms, approximately 11.15%. An earlier checkpoint measured 25.30 to 18.21 ms in a separate run; these timings are not pooled. The measured path begins with compressed image bytes already in host memory and includes decoding, preprocessing, host-to-device transfer, model execution, postprocessing, and an owned host output. It excludes disk, network, and serialization. The current-N timing uses the released primary state, but neither timing measures the five-view procedures; shared-device measurements are not an isolated deployment benchmark.

Dense-target training stress tests retain all ground truths with finite gradients under the tested batches. They demonstrate capacity and correctness, not faster convergence or universal throughput gains. Final deployment comparisons must include the cost of view generation and geometric postprocessing. We will not multiply a single-forward speedup by an unmeasured multi-view factor to invent an end-to-end operating point.

8 Discussion and Limits

The observations suggest an order for diagnosis: verify model state, test whether an optimization intervention still helps, then separate candidate coverage from ordering and geometric selection. This is a way to locate a remaining problem, not a mandatory sequence for every detector. State repair, a learning-rate intervention, and a frozen-box ordering change answer different causal questions even when all improve AP50. The useful mechanism claims are therefore tied to what each control holds constant.

The main RiO model family shares design and training ancestry. Cross-size comparisons additionally mix

query design, initialization, batch size, training stage, and evaluation horizons. They cannot identify capacity as the cause of a gain difference. The additional YOLO checkpoint broadens architecture coverage, but independent training seeds, other datasets, and a final accuracy–latency comparison remain necessary. Image-resampling intervals are conditional on selected models and data; they cannot undo historical test selection or account for training variability. Extra-Large and FAIR1M training remain incomplete at the evidence cut; neither is included in successful paper-target reproduction claims. The portable preview provides inference, evaluation, frozen multi-view improvement, training-image calibration, source-anchored adaptation and graph execution, with separately recorded validation scopes. Historical training records do not substitute for a certified portable end-to-end training recipe.

The current positive coordinate-only controls support simple consensus selection. Score weighting is not necessary for the observed gain, and local appearance is not a general improvement over uniform consensus. Better localization surrogates are a plausible research direction, but adding proposals or confidence alone is insufficient evidence for them. For Nano, the remaining primary-result paper gap is 0.0292 AP50 points; absent independent seed evidence, we attribute it neither to hardware noise nor to a proven capacity limit. The completed same-source learning-rate test did not improve Nano’s uncalibrated fixed endpoint, supporting closure of that tested recipe rather than further continuation solely to cross a benchmark threshold.

The extension to YOLO asks whether the same evidence can make a community model easier to deploy. Frozen-rule transfer tests dependence on the detector family; a train-only quality readout tests whether single-view features can repair ordering with unchanged geometry. Their positive complete-set point estimates support these limited claims, with positive image-paired intervals for the primary AP75 comparisons; deployment overhead remains unmeasured. Their controls and alternative interpretations appear in Appendix L. A diverse model library is useful insofar as it tests such questions under shared protocols.

Release artifacts and benchmark interpretation.

A public checkpoint, its configuration, and its reported score must be treated as a joint experimental artifact. Our bounded audit of one released RT-DETRv2-OBBLarge baseline did not recover its accompanying metric, despite matching the published file digest and cross-checking evaluation independently (Appendix J). This observation concerns that release configuration under our documented environment; it does not establish a cause or invalidate the full RiO-DETR results. We retain it as a reproducibility limitation, exclude it from successful reproduction claims, and use the paper’s reported metrics as reference targets for our independently

trained community models.

9 Conclusion

OpenRiO turns community reproduction of RiO-DETR into a basis for further development. Its selected Nano, Small, Medium and Large results approach or exceed the paper’s reference scores under explicit recipes and evaluation conventions, with independent metric checks. The reproduction process yields reusable lessons about checkpoint state and training, while controlled inference experiments show additional accuracy gains from ranking and geometric consensus. Separate execution measurements demonstrate a practical latency improvement. Mechanism analyses and negative controls explain the scope of these contributions; completing Extra-Large, additional datasets, independent training seeds and portable training will determine how broadly they transfer.

References

- [1] Atif Belal, Lilian Hollard, Marco Pedersoli, and Eric Granger. DESA-TTA: Dynamic EMA and Source Anchoring for Test-Time Adaptation. *arXiv preprint arXiv:2609.01795*, 2026. doi: 10.48550/arxiv.2609.01795. URL <https://arxiv.org/abs/2609.01795>.
- [2] Lukas Lao Beyer, Tianhong Li, Xinlei Chen, Sertac Karaman, and Kaiming He. Highly Compressed Tokenizer Can Generate Without Training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 4096–4114, 2025. URL <https://proceedings.mlr.press/v267/beyer25a.html>.
- [3] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A General Toolbox for Identifying Object Detection Errors. In *European Conference on Computer Vision*, pages 558–573, 2020. doi: 10.1007/978-3-030-58580-8_33. URL <https://arxiv.org/abs/2008.08115>.
- [4] clip5. TensorRT-DETR: inference deployment library. <https://github.com/clip5/TensorRT-DETR>, 2026. Accessed October 8, 2026.
- [5] RiO-DETR contributors. RT-DETRv2-OBBLDIOR-R release manifest and checkpoint metadata. https://github.com/RicePasteM/RiO-DETR/blob/master/model_zoo/rtdetr_v2_obb_diorr/manifest.json, 2026. Accessed October 8, 2026; local artifact snapshot retained.
- [6] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast

- and Memory-Efficient Exact Attention with IO-Awareness. *arXiv preprint arXiv:2205.14135*, 2022. doi: 10.48550/arxiv.2205.14135. URL <https://arxiv.org/abs/2205.14135>.
- [7] Zhicheng Du, Changyue Liu, Wenji Xi, Zhaotian Xie, Zhuo Deng, Ziheng Zhang, Yang Liu, and Lan Ma. Auditing Training-Free 3D Shape Retrieval with Diffused Geodesic Moments. *arXiv preprint arXiv:2605.29004*, 2026. doi: 10.48550/arxiv.2605.29004. URL <https://arxiv.org/abs/2605.29004>.
- [8] Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2959–2969, 2021. URL <https://proceedings.mlr.press/v139/elesedy21a.html>.
- [9] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet Pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4917–4926, 2019. doi: 10.1109/iccv.2019.00502. URL <https://arxiv.org/abs/1811.08883>.
- [10] Zhangchi Hu and collaborators. RiO-DETR: Official repository and release notice. <https://github.com/RicePasteM/RiO-DETR>, 2026. Accessed October 15, 2026.
- [11] Zhangchi Hu, Yifan Zhao, Yansong Peng, Wenzhang Sun, Jie Chen, Peixi Wu, Xinghao Wang, Xiangchen Yin, Dongsheng Jiang, Hebei Li, and Xiaoyan Sun. RiO-DETR: DETR for Real-time Oriented Object Detection. In *European Conference on Computer Vision*, pages 458–477, 2026. doi: 10.1007/978-3-032-37083-9_26. URL <https://arxiv.org/abs/2603.09411>.
- [12] Jiannan Huang, Aditya Kane, Fengzhe Zhou, Yunchao Wei, and Humphrey Shi. Revisiting Real-Time Detection Transformer with Efficient Encoder Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pages 6862–6871, 2026. doi: 10.1109/cvprf719.2026.00659. URL <https://doi.org/10.1109/cvprf71909.2026.00659>.
- [13] Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G. M. Araújo. CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022. URL <https://jmlr.org/papers/v23/21-1342.html>.
- [14] Zhengjian Kang, Jun Zhuang, Kangtong Mo, Qi Chen, Rui Liu, and Ye Zhang. PaQ-DETR: Learning Pattern and Quality-Aware Dynamic Queries for Object Detection. *arXiv preprint arXiv:2603.06917*, 2026. doi: 10.48550/arxiv.2603.06917. URL <https://arxiv.org/abs/2603.06917>.
- [15] Zijun Liao, Yian Zhao, Xin Shan, Yu Yan, Chang Liu, Lei Lu, Xiangyang Ji, and Jie Chen. RT-DETRv4: Painlessly Furthering Real-Time Object Detection with Vision Foundation Models. In *European Conference on Computer Vision*, pages 95–113, 2026. doi: 10.1007/978-3-032-37029-7_7. URL <https://arxiv.org/abs/2510.25257>.
- [16] LibreYOLO contributors. RT-DETRv2 OBB inference integration. <https://github.com/LibreYOLO/libreyolo/pull/733>, 2026. Merged August 8, 2026; inspected October 8, 2026.
- [17] Xu Lin, Jinlong Peng, Zhenye Gan, Jiawen Zhu, and Jun Liu. YOLO-Master: MOE-Accelerated with Specialized Transformers for Enhanced Real-time Detection. *arXiv preprint arXiv:2512.23273*, 2025. doi: 10.48550/arxiv.2512.23273. URL <https://arxiv.org/abs/2512.23273>.
- [18] Aiwen Liu, Chengguang Zhu, Gang Wang, Dandan Zhu, Haodong Lin, Yan Wang, Huiyu Zhou, and Zhengyi Pan. FSDC-DETR: A Frequency-Spatial Domain Collaborative DETR for Small Object Detection. *arXiv preprint arXiv:2607.05176*, 2026. doi: 10.48550/arxiv.2607.05176. URL <https://arxiv.org/abs/2607.05176>.
- [19] Kemal Oksuz, Selim Kuzucu, Tom Joy, and Puneet K. Dokania. MoCaE: Mixture of Calibrated Experts Significantly Improves Object Detection. *arXiv preprint arXiv:2309.14976*, 2023. doi: 10.48550/arxiv.2309.14976. URL <https://arxiv.org/abs/2309.14976>.
- [20] pablocpas. RiO-RTDETR Oriented: community implementation. <https://github.com/pablocpas/rt detr-obb>, 2026. Accessed October 8, 2026; source and configuration snapshot retained.
- [21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3962–3971, 2019. doi: 10.1109/cvpr.2019.00409.
- [22] Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement. *arXiv preprint arXiv:2410.13842*, 2024. doi: 10.48550/arxiv.2410.13842. URL <https://arxiv.org/abs/2410.13842>.

- [23] V. C. R. Sairam, Varun Gopal, Poornima Jain, Vineeth N. Balasubramanian, and Muhammad Haris Khan. Real-Time Source-Free Object Detection. In *European Conference on Computer Vision*, pages 305–323, 2026. doi: 10.1007/978-3-032-37020-4_17. URL <https://arxiv.org/abs/2606.31834>.
- [24] Qingyang Shi, Zhicheng Du, Jiasheng Lu, Yingshan Liang, Xinyu Zhang, Yiran Wang, Jing Peng, and Kehong Yuan. AudioCache: Accelerate Audio Generation With Training-Free Layer Caching. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025. doi: 10.1109/icassp49660.2025.10889043. URL <https://doi.org/10.1109/icassp49660.2025.10889043>.
- [25] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. doi: 10.1016/j.imavis.2021.104117. URL <https://arxiv.org/abs/1910.13302>.
- [26] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=uXl3bZLkr3c>.
- [27] Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet Strikes Back: An Improved Training Procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. doi: 10.48550/arxiv.2110.00476. URL <https://arxiv.org/abs/2110.00476>.
- [28] Dongyue Wu and Tao Ma. Mapping the Concept Landscape: Structural Perception of Global Distributions for Transparent Data Pruning. In *European Conference on Computer Vision*, pages 556–574, 2026. doi: 10.1007/978-3-032-37242-0_31. URL <https://arxiv.org/abs/2608.22858>.
- [29] Yifan Wu, Yiqi Wang, Xichen Ye, Wenjing Yan, Xiaoliang Li, Cheng Jin, Xiangyu Yue, and Weizhong Zhang. Distill on a Diet: Efficient Knowledge Distillation via Learnable Data Pruning. *arXiv preprint arXiv:2606.25488*, 2026. doi: 10.48550/arxiv.2606.25488. URL <https://arxiv.org/abs/2606.25488>.
- [30] Ryota Yagi. Extending Dataset Pruning to Object Detection: A Variance-based Approach. *arXiv preprint arXiv:2505.17245*, 2025. doi: 10.48550/arxiv.2505.17245. URL <https://arxiv.org/abs/2505.17245>.
- [31] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. VarifocalNet: An IoU-Aware Dense Object Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8510–8519, 2021. doi: 10.1109/cvpr46437.2021.00841. URL <https://doi.org/10.1109/cvpr46437.2021.00841>.
- [32] Yi Zhang, Di Xiong, Ge Gao, Xiangyue Zhang, Yihang Qiu, Yuxuan Zhou, and Shuo Chen. Dynamic Query as Budgeting for Tiny Object Detection. <https://neurips.cc/virtual/2026/poster/154187>, 2026. Official NeurIPS 2026 poster abstract; accessed October 13, 2026. Full implementation not inspected.
- [33] Qing Zhou, Junyu Gao, and Qi Wang. Scale Efficient Training for Large Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20458–20467, 2025. doi: 10.1109/cvpr52734.2025.01905. URL <https://arxiv.org/abs/2503.13385>.
- [34] Qing Zhou, Tao Yang, Bingxuan Zhao, Hongyuan Zhang, Junyu Gao, and Qi Wang. Inconsistency Biases in Dynamic Data Pruning. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=Zw1Uw7u6Su>.
- [35] Qing Zhou, Bingxuan Zhao, Tao Yang, Hongyuan Zhang, Junyu Gao, and Qi Wang. Batch Loss Score for Dynamic Data Pruning. *arXiv preprint arXiv:2604.04681*, 2026. doi: 10.48550/arxiv.2604.04681. URL <https://arxiv.org/abs/2604.04681>.
- [36] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. MMRotate: A Rotated Object Detection Benchmark using PyTorch. In *ACM International Conference on Multimedia*, pages 7331–7334, 2022. doi: 10.1145/3503161.3548541. URL <https://arxiv.org/abs/2204.13317>.
- [37] Hebao Zhu. Super Sparse DETR: YOLO-Competitive Convergence and Acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pages 6680–6687, 2026. doi: 10.1109/cvprf71909.2026.00641. URL <https://doi.org/10.1109/cvprf71909.2026.00641>.

A Registered Statistical Procedure

We draw 999 bootstrap samples of 11,738 image identifiers with replacement using seed 20261006. All methods use the same sampled image multiplicities. Within image, classwise greedy matches are computed from each

method’s fixed predictions at IoU 0.5 and 0.75. Multiplicity weights repeat the image’s detections and ground-truth count, while preserving the baseline global score order and tie convention. AP is recomputed for every sample; detections are not treated as independent sampling units. Explicit multiplicity expansion on synthetic examples checks the weighted AP implementation. The unweighted result must match the existing full-set metric before a method enters resampling.

Pointwise 2.5th and 97.5th percentiles describe paired differences. AP75 is the primary geometric endpoint and AP50 is reported alongside it. The analysis is exploratory, includes multiple comparisons without a simultaneous correction, and is conditional on frozen models. A fraction of positive bootstrap differences is not reported as a p -value. Scene-level identifiers are unavailable for a geographic cluster bootstrap; shared scenes and near duplicates may make image-level intervals optimistic.

B Registered Full-Set Diagnostic

The analysis compares the original-view cached predictions with their five-view union. Both class-agnostic and class-aware geometric coverage are measured at IoU 0.5 and 0.75. Object sizes use $\sqrt{wh} < 32$, $32 \leq \sqrt{wh} < 96$, and $\sqrt{wh} \geq 96$ original-image pixels. Image density is stratified by at most 10, 11–30, and more than 30 annotated objects. These are descriptive bins, not learned decision rules.

At confidence 0.3 and IoU 0.5, unmatched objects are assigned in order to absent cached geometry, absent compatible class, low compatible score, or matching competition. The partition uses only cached predictions; it cannot claim to explain all internal decoder queries. For fixed selected boxes, classwise score/support bins summarize best same-class IoU. Such associations can be confounded by object difficulty and do not replace an intervention.

C Statements Supporting the Inference Controls

These elementary statements delimit the interpretation of our methods; they are not presented as novel generalization results.

Proposition 1 (Coverage, matching capacity, and eleven-point AP). *Fix one class, IoU threshold, ground-truth set, and finite candidate pool with unchanged geometries and labels. For each image x , form a bipartite graph between candidates and valid ground truths, with an edge when overlap qualifies under the evaluator. Let m_x be its maximum matching size, c_x the number of ground truths having at least one edge, and K the per-image output cap. If $N > 0$ is the class’s total ground-*

truth count, then any selection and ordering of this pool satisfies

$$\text{TP} \leq M := \sum_x \min(K, m_x) \leq \sum_x c_x, \quad (5)$$

$$R_{\max} \leq M/N. \quad (6)$$

For VOC2007 eleven-point AP on a zero-to-one scale, define

$$U(M, N) = \begin{cases} 0, & M = 0, \\ \frac{1}{11} \sum_{j=0}^{10} \mathbf{1}\{jN \leq 10M\}, & M > 0. \end{cases} \quad (7)$$

Then $\text{AP}_{11} \leq U(M, N)$.

Proof. Each true positive uses one retained candidate and one previously unmatched ground truth, hence the true positives form a matching bounded by both m_x and K . Every matched ground truth is covered. Precision is at most one, and interpolated precision is zero at recall thresholds above the maximum achieved recall. Counting the attainable eleven recall thresholds gives the stated AP bound; with no true positives every precision is zero. $\square$

Candidate reuse can make coverage strictly larger than achievable recall: one candidate adjacent to two ground truths covers both, but matches at most one. Its maximum recall is 1/2, while a perfect first true positive gives eleven-point AP 6/11. Thus recall itself is not a valid numeric upper bound on eleven-point AP; the interpolation grid matters. Per-class caps also relax competition with other classes, so this bound may remain loose. It applies to selection and reordering from the same geometry pool, not to newly regressed or fused boxes. Maximum matching need not be attainable by the evaluator’s greedy assignment, so it is a diagnostic bound, not an oracle detector score. A dependency-free companion program checks the bound on all 682 binary graphs with one to three candidates and one to three ground truths. A corresponding full-dataset matching analysis is planned; the existing coverage plots remain explicitly candidate-reuse statistics.

An interaction must be measured rather than assumed away. Let A_{ab} be AP on one fixed set of output identities when score changes are enabled by $a \in \{0, 1\}$ and coordinate changes by $b \in \{0, 1\}$. With groups and the evaluation protocol held fixed, define

$$I = A_{11} - A_{10} - A_{01} + A_{00}. \quad (8)$$

Then $A_{11} - A_{00} = (A_{10} - A_{00}) + (A_{01} - A_{00}) + I$ exactly. AP’s ranking and matching operations give no general reason for $I = 0$. This is bookkeeping for a controlled factorial experiment, not a new causal-identification theorem. Our existing sequential controls do not provide

every cell of this same-identity factorial; their gains are not silently summed, and completing it is a planned analysis priority.

Proposition 2 (Order invariance on a fixed detection set). *For classwise AP with unchanged detections, matching convention, and tie order, a transformation that preserves the actual total score order within each class at every matching and ranking stage leaves AP unchanged. A strictly increasing transformation in exact arithmetic satisfies this condition; finite-precision implementations require a separate order check.*

Proof. The transformation preserves the ordered detection list within each class. Greedy matches, cumulative true and false positives, and the precision–recall curve are therefore identical. The same holds for eleven-point interpolation and the mean over classes. $\square$

Thus, the observed fixed-set gains cannot result merely from monotonically rescaling scores. Candidate-dependent support changes their relative order. This statement does not establish that support is the best available quality estimator.

Proposition 3 (Conditional surrogate guarantee). *Let A be a fixed, finite set of candidate boxes. Suppose a target distribution P over a single object’s geometry and the candidate distribution Q_G satisfy*

$$\sup_{b \in A} |R_P(b) - R_{Q_G}(b)| \leq \epsilon, \quad (9)$$

where $R_P(b) = \mathbb{E}_{z \sim P}[1 - \text{IoU}(b, z)]$. Then a minimizer $\hat{b}$ of R_{Q_G} over A satisfies

$$R_P(\hat{b}) - \min_{b \in A} R_P(b) \leq 2\epsilon. \quad (10)$$

Proof. For a target minimizer b^* , $R_P(\hat{b}) \leq R_{Q_G}(\hat{b}) + \epsilon \leq R_{Q_G}(b^*) + \epsilon \leq R_P(b^*) + 2\epsilon$. $\square$

This elementary bound locates the assumption that needs evidence: candidate agreement must approximate target geometric risk. It is not a new generalization theorem. We do not estimate ϵ or assume independent view errors. Systematic errors and groups that mix adjacent objects can invalidate the surrogate. AP additionally depends on class assignment, duplicates, ordering, and IoU thresholds, so the bound is not an AP-improvement guarantee.

Proposition 4 (Information available to a quality readout). *Let $Y \in \{0, 1\}$ indicate that a fixed predicted box has same-class ground-truth overlap at least τ . Let M contain its score, class, and geometric metadata, and let V denote frozen visual features. For unrestricted measurable probability predictors under binary cross-entropy loss ℓ , the difference in optimal population risks is*

$$\inf_g \mathbb{E} \ell(Y, g(M)) - \inf_h \mathbb{E} \ell(Y, h(M, V)) = I(Y; V | M). \quad (11)$$

Proof. Conditioned on any features X , expected binary cross-entropy is minimized by $\Pr(Y = 1 | X)$, with optimum $H(Y | X)$. Subtracting the optima for M and (M, V) gives the definition of conditional mutual information. Deterministic probabilities are interpreted by continuity. $\square$

This standard identity motivates a control, not a claim that our finite readout attains either optimum. A visual readout must outperform a metadata-only one to support useful additional predictive information. A class-preserving shuffled-target control probes whether apparent gains can instead arise from class priors or incidental reweighting. Fitted risks on finite, detector-exposed data are not estimates of conditional mutual information without further assumptions. Moreover, Y is geometric eligibility, not a true-positive label after duplicate competition; lower quality-prediction loss need not improve AP. Complete-set fixed-box AP is therefore a separate required endpoint.

D From Reproduction Failures to Testable Mechanisms

This October 12 update extends the OpenRiO experimental foundation with three distinct checks: exact computation under dense supervision, bounded parameter adaptation, and numerically faithful ranking. The underlying algebra and optimization principles are established ideas. Our contribution is their concrete use to diagnose, test, and improve this community implementation. These checks do not replace full detection evaluation or provide a theorem that benchmark AP must increase.

D.1 Dense supervision requires a structured computation

Let G be the largest ground-truth count in a microbatch, Q the matching query count, and $g = \max(1, \lfloor 100/G \rfloor)$ the denoising group count for $G > 0$. The implemented denoising sequence has

$$L = Q + 2Gg. \quad (12)$$

The configuration value 100 therefore does not cap the number of denoising queries. Even an efficient attention kernel can encounter a quadratic allocation if its caller first constructs an $L \times L$ attention mask. This was a material resource boundary in the multi-scale Extra-Large run. Its first epoch stopped at a per-process allocation limit; the exact failing batch was not retained, so a numerical match to a possible mask size is not treated as an identified failing input.

Proposition 5 (Exact allowed-set decomposition). *With zero attention dropout, a denoising query in group j can be evaluated against only the keys of group j and*

the matching queries; a matching query can be evaluated against only matching keys. Applying these computations in blocks gives the same real-arithmetic attention outputs and derivatives as the original mask.

Proof. For query i , let A_i be the allowed key indices. Masked softmax assigns zero probability to its complement and has denominator $\sum_{k \in A_i} \exp(q_i^\top k_k / \sqrt{d})$. Restricting the key and value arrays to A_i leaves both this denominator and the weighted value sum unchanged. Queries with the same allowed set can be grouped; tiling their rows does not change a row’s normalization. Differentiating these identical smooth expressions gives the same derivatives. Finite-precision kernel execution is checked separately. $\square$

The implementation replaces the dense mask with a small structural descriptor and evaluates query tiles using an existing efficient attention backend. This follows the established principle of memory-aware exact attention [6]; it is not a new linear-attention algorithm. Its attention arithmetic can remain quadratic in G . CPU and CUDA output, input-gradient, and parameter-gradient comparisons pass their declared tolerances. A complete Extra-Large forward, backward, and optimizer update with 12,000 synthetic targets retains all targets and uses 7.31 GiB peak reserved memory under the tested process budget. The revised run also passes a fresh actual-data probe before restarting. This establishes tested capacity and numerical equivalence, not a universal non-exhaustion guarantee or a matched throughput improvement.

An additional recipe issue is kept separate from this execution change: rolling a batch of one image for MixUp pairs the image with itself and duplicates its target list. The attention repair preserves that training semantics. Altering the augmentation requires its own matched experiment.

D.2 Source anchoring as a measured contraction

We transfer dynamic teacher averaging and source anchoring from DESA-TTA [1]. Let θ_0 be source parameters, u_t the actual Adam displacement, $z_t = \theta_t + u_t$, and a_t the anchoring coefficient, with $0 \leq a_t \leq 1$. The implemented update is

$$\theta_{t+1} = (1 - a_t)z_t + a_t\theta_0. \quad (13)$$

For a norm D with fixed source-dependent positive weights, homogeneity and the triangle inequality give

$$D(\theta_{t+1} - \theta_0) = (1 - a_t)D(z_t - \theta_0), \quad (14)$$

$$D(\theta_{t+1} - \theta_0) \leq (1 - a_t)\{D(\theta_t - \theta_0) + D(u_t)\}. \quad (15)$$

With constant $0 < a \leq 1$, $D(u_t) \leq U$, and initialization at the source, summing the geometric series yields

$D(\theta_t - \theta_0) \leq (1 - a)U[1 - (1 - a)^t]/a$. The actual controller can have $a_t = 0$, so this constant-coefficient bound cannot be asserted for its whole trajectory without additional conditions. Gradient clipping does not establish an Adam displacement bound. The new probes therefore record $D(u_t)$ and test the measured inequalities. Teacher EMA obeys the corresponding convexity bound when its mixing coefficient lies in $[0, 1]$ and both parameter states use the same norm and source reference.

Across 46 completed traces containing updates, the measured anchoring identity has maximum absolute residual 2.12×10^{-8} . The full-stream results indicate a stability–adaptation trade-off: for Nano on clean images, fixed EMA loses 6.066 AP50 points while dynamic EMA with anchoring loses 0.020. Under noise, unanchored dynamic adaptation can improve AP50 more than anchoring while harming AP75 or later recovery. All recorded full-stream rollback counts are zero; these results establish no rollback benefit.

To test the mechanism rather than merely correlate drift with AP, we register 18 full-stream interventions: Nano and Small, clean/noise/recovery streams, and classification-only/geometry-only/joint pseudo-label losses. Checkpoints, image order, matching rule, native loss weights, anchoring, and optimizer settings are held fixed. Shared normalization parameters can affect both scores and boxes, so these are loss interventions rather than isolated prediction heads. Future teacher targets may diverge between arms; realized update norms are recorded to expose this feedback and rival explanations. All 18 interventions have now completed full-stream evaluation. Source state, image order and split manifests match their frozen references. Under synthetic noise, joint adaptation changes AP50/AP75 by +0.6884/+0.1819 for Nano and +0.7428/+0.3802 for Small. Classification-only loss gives +0.6768/+0.1661 and +0.8248/+0.3631; geometry-only loss gives +0.1771/+0.1011 and +0.2661/+0.1495. These interventions suggest that the classification objective supplies much of this noise-stream benefit, without isolating a prediction head or proving a universal mechanism. On clean streams, all six model–loss combinations reduce both metrics. Recovery-stream outcomes are mixed: Nano joint adaptation gives +0.1585/+0.1539, while Small gives +0.2605/−0.0490. Thus source anchoring can support adaptation under a specified shift without constituting a universal clean-data improvement. These are exploratory single-detector-seed results on synthetic shifts; real-domain transfer, adaptation overhead and seed robustness remain open.

D.3 Ranking invariance has an implementation boundary

For a fixed detection set, classwise strictly increasing transformations preserve within-class score comparisons in exact arithmetic. With unchanged tie handling,

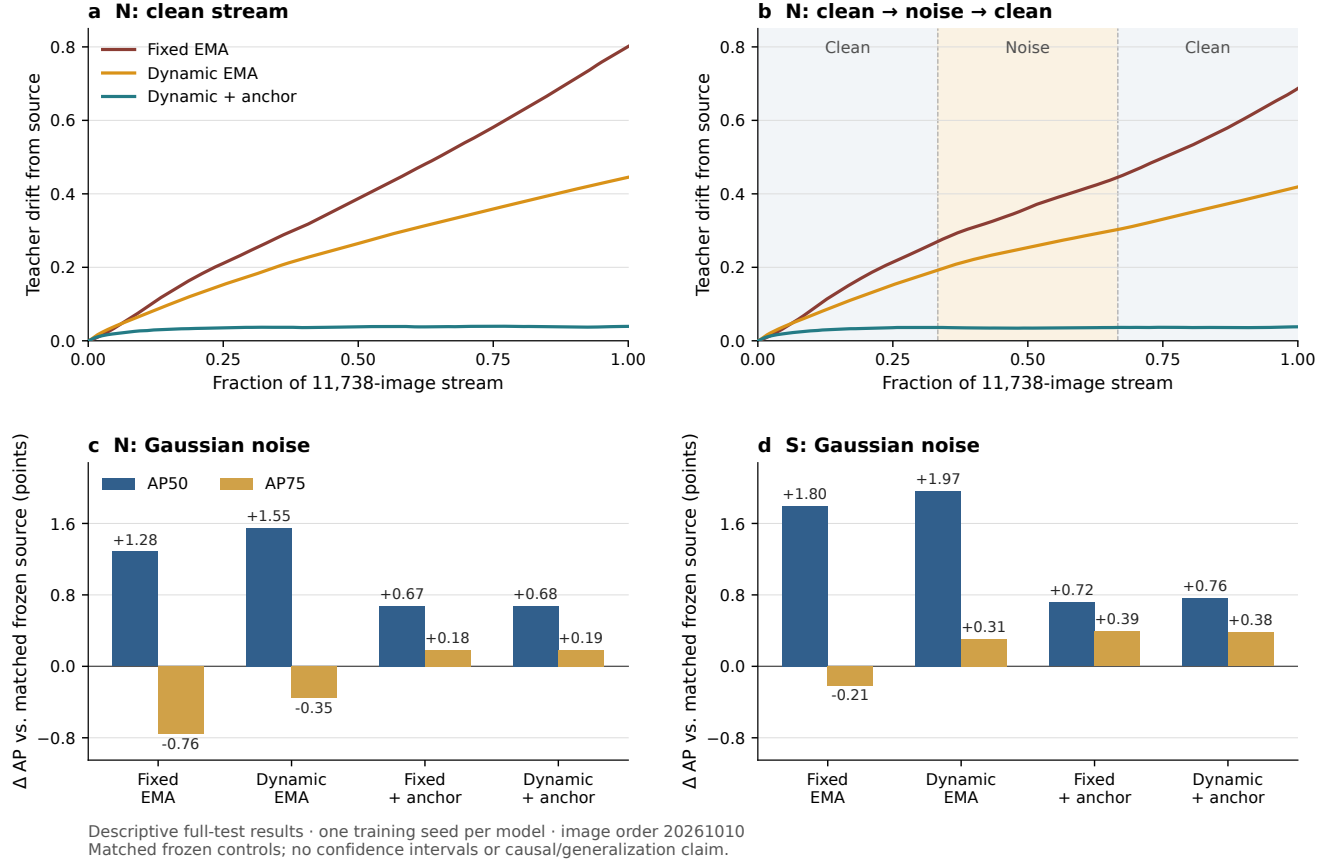


Figure 9: **Parameter stability and detection quality measure different effects.** Nano teacher drift is shown on clean and clean–noise–clean streams; paired noise AP50 and AP75 changes are shown for Nano and Small. Full-stream AP uses all 11,738 images and its matched frozen reference. Recovery phases use different images, so raw phase scores cannot establish recovery. These are descriptive results from one source training seed and one displayed order, with historical test exposure; no significance or causal claim follows from the plotted association.

greedy matching, suppression, and global AP ordering, they therefore preserve single-model AP. The statement requires every score-dependent stage to preserve those comparisons. Cross-class top- K filtering must remain fixed or be checked separately, since different classwise transforms can change it.

Our calibration audit found a concrete violation of that implementation condition: calibrated scores were stored in double precision, narrowed to single precision before per-image matching, but retained in double precision for global AP sorting. Newly created ties changed matched detection identities. On all 11,738 YOLO images, native AP50 was 69.51718 and the mixed-precision path gave 69.45010. Removing the narrowing restored native AP50 and AP75. The source candidates and fitted calibration were unchanged. The isolated repair preserves the original failed result and records score-order and matching differences; it does not redefine old benchmark tables silently. For multi-model unions, calibration can legitimately change cross-model ordering, so single-model invariance is a required control rather than a claim

that ensembling cannot improve AP.

We complete the control with eight frozen prediction variants on all 11,738 images, using double-precision scores and stable ordering for both matching and AP. The original polygon geometry is unchanged. Native and calibrated single-expert results agree exactly, including every class, for both Nano and YOLO. The direct union reaches 74.54608 AP50 and 48.82113 AP75; the calibrated union reaches 74.88735 and 49.43924, respectively: gains of 0.34127 and 0.61811 points under the same repaired protocol. This transfers the established calibrated-ensemble idea of MoCaE [19]; both detectors still execute. The calibrator was fitted on the 11,702 images remaining from the detector’s 11,725-image training-plus-validation pool after excluding the 23 documented exact test duplicates, not an independent calibration split. No refitting or threshold selection was performed for this repair. Historical test exposure remains, and these paired point estimates do not establish statistical significance or independent transfer. Original failed outputs are retained alongside the complete

repaired evaluation and input hashes.

Further theory should answer an observed question. Group averaging offers a principled diagnostic for correlated geometric errors [8]. For a finite or compact group with normalized group measure, an invariant input distribution, an equivariant target, and a unitary Hilbert-space output representation, group averaging is an orthogonal projection and removes the non-equivariant component of squared error. Detection matching and AP do not satisfy that formulation automatically. This motivates tests on fixed matched boxes, including symmetry-consistent inverse transforms and cross-view error correlation, before making any detection-performance claim. New statistical risk guarantees likewise require their own calibration and dependence assumptions; historically selected test streams do not supply them.

We now register a CPU study using the complete cached predictions of two Nano lineages. Native detection identities, classes, scores and counts are fixed. Reciprocal same-class matches at IoU at least 0.55 form complete quarter-turn groups; the horizontal flip is excluded from the group. On the same eligible anchors we compare native geometry, two-view and four-view moment means, confidence-weighted means, an IoU medoid and a seeded random member. A repeated-view identity control and quarter-turn covariance checks pass before full evaluation. Ground truth is used only for subsequent AP and fixed-assignment error diagnostics. The study tests whether correlated error limits geometric averaging and whether a reduced representation error translates to detection quality; the complete full-set comparisons are now available. Four-view uniform averaging changes AP50/AP75 by +0.0308/+0.6994 for the compact Nano path and +0.0302/+0.1801 for the rotated Nano path. Seeded random members instead reduce AP75 by 0.4290 and 0.4127 points. These are paired descriptive effects on historically exposed checkpoints, not guarantees from the projection theorem. This belongs to the optimization objective and does not enter the reproduction table.

A synthetic contract exposes a further representation boundary before any full-result selection: a square has an isotropic second moment, which loses the orientation of its boundary. Direct eigendecomposition can therefore change even a repeated identical square (IoU 0.76308 in the retained counterexample). A separate decoder control retains the native reference angle at numerical eigenvalue degeneracy. It passes polygon identity checks for square, nearly square and elongated boxes under quarter turns. Both the original and repaired full-data controls are retained. This fixes that specific identity failure; it does not make second moments injective on all rectangles or prove an AP benefit.

The square-safe decoder and original decoder give identical full-set AP in both tested paths. The synthetic identity repair is therefore a correctness contri-

bution without a demonstrated benchmark gain in these runs.

E Mechanism Evidence and Remaining Tests

The main argument separates observed effects from their possible explanations. Table 6 records the boundary of each claim. Experiments are grouped by the question they answer, not by the order in which they were run. The original matched-horizon evidence cut is October 8 at 19:35. The Nano completion and larger-scale status below are explicitly updated to October 9 at 23:44; earlier paired comparisons retain their frozen horizons.

E.1 Conditions under which optimization helps

Several historical alternatives—longer tails, learning-rate changes, effective batch changes, normalization interventions, and denoising changes—were evaluated from different parent checkpoints. They are useful recipe records but cannot be pooled as a balanced factorial experiment. In one earlier same-rate comparison, batch 8 and effective batch 32 reached 71.4841 and 71.4572 AP50. That result does not justify increasing batch solely because more device memory is available. Gradient accumulation also need not reproduce physical-batch normalization or ground-truth loss normalization. Likewise, three continuation seeds from one existing parent measure restart variability, not end-to-end training variance.

The latest Nano plateau consists of a fixed low-rate, rotation-augmented, frozen-normalization tail. Two consecutive 24-epoch windows met the controller’s operational criteria: improvement over the incumbent below 0.02 AP50 points, within-window range below 0.15, and absolute change over the final eight epochs below 0.1. These tolerances govern resource allocation; they are not estimated statistical noise levels. The final window’s best occurs at cumulative tail epoch 88, not at the endpoint. Figure 10 preserves the evaluated trajectory rather than presenting only the selected maximum.

E.2 Completed Nano comparison and calibration extension

The new pair starts from the same 71.8562 checkpoint and uses learning rates 2.5×10^{-7} and 10^{-6} . Both arms initialize a fresh optimizer; this is a controlled restart, not restoration of an identical previous optimizer state. Batch 8, accumulation 1, seed 1, quarter-turn augmentation, frozen normalization, and backbone learning-rate multiplier 0.1 are shared. Each arm has a finite 24-epoch budget with complete-set evaluation every four

Table 6: What the interventions establish, and what remains unresolved. All metric comparisons use the complete test set unless explicitly described as an illustrative case.

Question	Controlled observation	Interpretation boundary and next test
Checkpoint state	A two-buffer normalization repair recovers about 5.91 AP50 in one lineage; calibration fails elsewhere.	State mismatch can dominate a deficit. This does not identify every failed run’s cause. Audit state and compare raw, EMA, and calibration separately.
Remaining optimization	Higher-rate matched continuations outperform lower-rate Small and Medium controls at the observed horizons.	Evidence concerns those starting states and recipes. The completed Nano pair does not reproduce that uncalibrated endpoint advantage; calibration has a separate effect.
Batch and duration	Larger effective batch and repeated low-rate tails are not consistently better in historical Nano attempts.	Those attempts constrain recipes, not all large-batch training. Physical batch, accumulation, normalization, and update count must be distinguished.
Score ordering	Rescoring fixed identities improves AP75.	Additional selected geometry cannot explain that comparison. It does not show that support is a calibrated quality probability.
Geometric selection	Uniform and weighted representatives beat random coordinate replacement; uniform transfers to Small and Medium.	Structured candidate agreement is useful here. Score weighting is not necessary; shared view errors remain a failure mode.
Local appearance	Compact Nano improves, while three other paths do not; B has a good proposal but the selector misses it.	Added coverage and selection reliability are distinct. Report opportunity and regret distributions before designing a more complex selector.

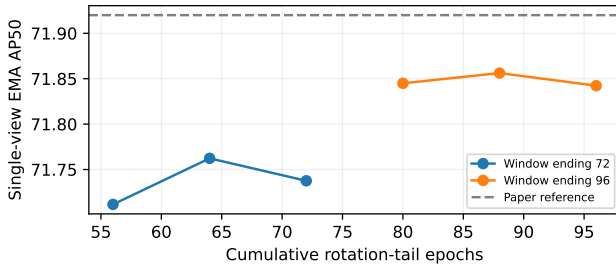


Figure 10: Nano’s late fixed-recipe trajectory. The paper reference supplies context, not a stopping threshold that must be crossed. A plateau here concerns this recipe and starting history.

epochs. EMA 0.999 is the primary endpoint; a simultaneous 0.9999 trace is secondary and follows the same parameter trajectory. Calibrated variants remain separate.

Both registered 24-epoch arms have completed; primary and calibrated outcomes are reported in Section 5. The launcher waits for one existing GPU process or fewer and at least 14 GB free, runs a batch probe, and limits the new worker’s allocation. It stops a new arm on two consecutive primary-EMA observations at least 0.25 AP50 below the starting checkpoint when the latter does not improve by more than 0.01. Failures block blind retries. The completed finite test triggered a calibration-subset robustness check rather than indefinite repetition to reach 71.92.

If the higher-rate arm improves over its matched con-

trol, Nano’s earlier plateau was recipe dependent at this starting state. If both flatten, this avenue has weaker justification; it does not prove a global optimum. If EMA settings differ on the same trajectory, the effect belongs to the reported model state as well as optimization history. Measuring realized layerwise updates and model-EMA separation could further test explanations based on frozen normalization, parameter motion, or averaging lag; these mechanisms are not established by the current AP curves.

E.3 Local repair contract and scope

The appearance probe builds a grayscale-and-gradient template on a canonical 32×16 patch. Teachers require support at least 0.8 and score at least 0.5; targets require support at most 0.65 and score at least 0.15. Up to three nearest same-class teachers are taken within 200 pixels, outside a size-dependent separation distance, with aspect ratios within a factor of two. We test 15 translations: offsets along the target’s local width of $\{-0.25, 0, 0.25\}$ and along its local height of $\{-0.5, -0.25, 0, 0.25, 0.5\}$. Angle and side lengths are preserved. Proposals overlapping another eligible same-class output above IoU 0.5 are rejected; the unchanged anchor remains valid. Template similarity allows a 180-degree patch reversal and averages over teachers. The objective subtracts 0.1 times squared normalized displacement and requires improvement of at least 0.05 over the anchor. The shuffled control permutes appearance similarities over the same proposals before applying the same penalty, validity mask, and gate; it is not uniform random replacement. No test annotations enter this rule.

The ship example is selected from image 21388 using the earlier explanatory A/B case. An offline oracle identifies proposal 6 with IoU 0.7564, while the appearance rule chooses proposal 11 at approximately 0.4706. This is deliberately a diagnostic upper opportunity within that local pool, not oracle AP or a reported inference method. Excluding exact cross-split duplicates leaves Compact’s AP50 gain over uniform at approximately 0.2192; this sensitivity does not make the historically inspected dataset independent.

The positive Compact result and negative rotation Nano result are especially important: nearly equal parameter counts do not imply equal residual errors or equal reliability of a visual proxy. The two paths also differ in queries and training ancestry. A stronger causal study would match architecture and training stage, stratify opportunity and regret by size, density, and view disagreement, and then freeze any revised rule before an independent evaluation. These are prospective tests, not completed mechanism evidence.

E.4 Incomplete scales

At the October 10 update, Large and Extra-Large have completed their 72-epoch base plans and additional continuations. Their selected uncalibrated EMA scores are 75.9915 and 73.9681 AP50, respectively, compared with paper references of 76.11 and 77.43. At matched 24-epoch horizons from each scale’s shared phase-best checkpoint, increasing the learning rate from 2.5×10^{-7} to 10^{-6} improves Large from 75.5449 to 75.8276 and Extra-Large from 73.6545 to 73.9093. Large retains rotation augmentation within its pair; Extra-Large retains its control augmentation within its pair. These results support further investigation of those continuations, not global convergence or a hardware explanation for the remaining gap. Extra-Large’s separate lower-rate rotation endpoint reaches 73.9681 EMA and 74.0703 raw; the latter is not substituted into the primary convention. The missing higher-rate rotation combination is registered from the same starting checkpoint. Both scales remain ongoing and excluded from completed reproduction claims. Reboot recovery restores the saved epoch-boundary optimizer, EMA and random state; intermediate evaluation cadence restarts, while the interrupted window’s total completed-epoch budget is preserved. The public-release discrepancy discussed separately is not used to explain these community trajectories.

F Inspecting Strong and Weak Predictions

We inspect the complete 11,738-image prediction caches of the frozen release candidates: primary Nano, Small, and Medium with AP50 71.8909, 75.0200, and 76.1754. This Medium checkpoint predates the latest ongoing

training result. Independent polygon matching reproduces their AP50 and AP75 within 0.0003 points. The candidate pool remains fixed at 300 predictions per image. Ground truth is used only for this offline analysis.

Coverage and ordering require different interventions. For each ground truth, candidate coverage asks whether any retained same-class prediction reaches the IoU threshold, irrespective of score and one-to-one matching. It is not detection recall or AP. Among 75,315 small objects with area below 32^2 pixels, coverage at IoU 0.5 is 74.73%, 83.17%, and 85.29% for Nano, Small, and Medium. At IoU 0.75 it is 43.21%, 56.48%, and 59.34%. Thus, many Nano small objects lack sufficiently precise geometry even before ranking is considered; changing scores alone cannot create these absent candidates. Architecture, training history and budget differ, so these cross-model observations do not isolate a causal effect of capacity.

For the 84,863 Nano predictions scored at least 0.7, 78,356 are independently matched true positives at IoU 0.5 and 60,975 at IoU 0.75. Among the remainder at IoU 0.5, the operational partition contains 840 duplicate matches, 3,274 partial same-class geometric matches, 237 cases overlapping another class, 2,101 with less than 0.1 overlap with any annotation, and 55 residual weak overlaps. The latter categories are geometric diagnostics; an unmatched prediction does not by itself establish an annotation error. These counts also show why high classification confidence should not be interpreted as a guarantee of precise localization.

A low score is not automatically a missed opportunity. Figure 11 shows cases selected by a fixed hash rule within predefined Nano categories, then inspects the same images across the released scales. In image 13539, a vehicle candidate has score 0.035 and IoU 0.753. Inspecting its competitors reveals another retained Nano candidate with score 0.860 and IoU 0.830: the low-scoring candidate is redundant, not evidence that the target needs confidence repair. In image 12928, a ship prediction scores 0.887 with IoU 0.356, while another retained candidate scores 0.620 with IoU 0.783. This is an actual confidence–geometry ordering mismatch. In image 22785, no retained Nano same-class candidate reaches IoU 0.5 for the highlighted small vehicle, so a score-only intervention cannot recover a qualifying box from that pool.

These examples distinguish ranking, redundancy and absent geometry. They motivate checking competing candidates before designing an apparent repair, and assessing each mechanism with full-set controls. They are illustrative selections rather than prevalence estimates or evidence that transferring another model’s boxes is an available inference method.

OpenRiO: distinguishing coverage, ranking and redundant candidates
 Green: annotation | Red: diagnostic candidate | Cyan: another retained Nano candidate

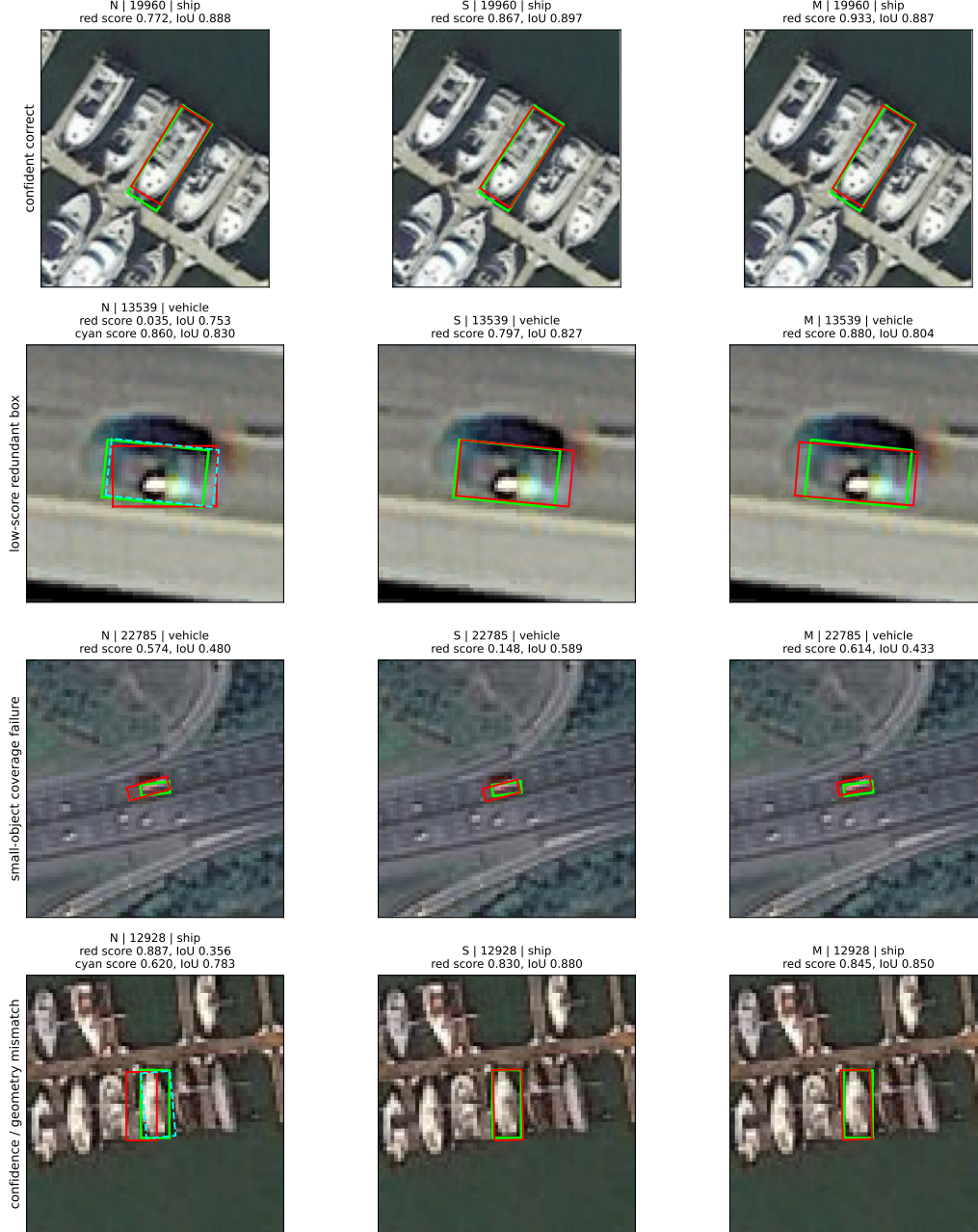


Figure 11: **Strong predictions and three distinct limitations.** Green denotes an annotation, red the displayed diagnostic candidate, and cyan another retained Nano candidate. The low-score row is explicitly a redundant candidate; the last row contains an ordering mismatch. Small and Medium candidates are selected by best same-class IoU for offline illustration of their available geometry, not by a deployable oracle. Other annotations and predictions are omitted for legibility. Selection rules, competing-candidate checks, and full-set statistics are retained with the figure.

G A Boundary Case: Appearance-Guided Repair

A more ambitious intervention uses a well-supported nearby object’s appearance to guide a local translation of a less-supported same-class prediction. It keeps the selected identities, labels, scores, and output counts fixed. This tests whether visual evidence from an easier detection can help repair a harder one, rather than merely improve its rank. Relative to uniform geometry, full-test AP50 changes are +0.217, −0.014, −0.076, and −0.015 for Compact Nano, rotation Nano, Small, and Medium. AP75 changes are +0.039, −0.004, −0.185, and −0.053. Shuffled local choices underperform in all four paths. Appearance therefore carries some useful information relative to this random control, but does not deliver a transferable improvement over the stronger uniform baseline. We retain the Compact result as a conditional signal, not a replacement default for all sizes.

A good proposal is not yet a successful repair.

The earlier ordering example contains a well-localized ship A and a poorly localized ship B. They are different objects; A supplies a possible appearance template, not B’s target coordinates. Local translation produces a B proposal with IoU 0.756 to its annotation, yet the appearance rule selects one at approximately 0.471, compared with 0.454 before repair (Figure 12). The added pool has useful geometric opportunity, but its selector fails to realize it. This case was selected for explanation and cannot establish failure frequencies.

For a fixed diagnostic annotation g , let a be the original anchor, A a proposal set containing a , and $\hat{b} \in A$ the selected proposal. Write $J(b) = \text{IoU}(b, g)$ and $J^* = \max_{b \in A} J(b)$. Then

$$\underbrace{J(\hat{b}) - J(a)}_{\text{realized local change}} = \underbrace{J^* - J(a)}_{\text{available repair}} - \underbrace{J^* - J(\hat{b})}_{\text{selection regret}}. \quad (16)$$

This identity separates two diagnostic quantities; it is not an AP decomposition or a novel bound. Ground truth is used only to measure them offline, never to select proposals at inference. For B, the approximately 0.302 available IoU improvement is offset by approximately 0.286 selection regret. Dense-object matching, category errors, duplicates, and score ordering still determine whether any such local change improves AP. The conditional surrogate statement in Appendix C explains why a better selection signal matters, without assuming that appearance similarity estimates geometric risk accurately.

H Explanatory Case Selection

Explanatory image cases. The ranking illustration uses test image 21388, category ship. The original scores of candidates A and B are 0.884 and 0.953, whereas their

support-adjusted scores are 0.843 and 0.775. Their respective annotation IoUs remain 0.881 and 0.454. The scan finds eligible reversals in 504 of 11,738 images under the registered illustration criteria; this count is not the number of all beneficial detections.

The negative geometry illustration uses image 19396 (storagetank): IoU changes from 0.842 to 0.516 with score fixed at 0.749. The same ground-truth identity is used before and after replacement.

The positive geometry illustration uses image 21906 (ship): IoU changes from 0.558 to 0.954 with score fixed at 0.379. The same ground-truth identity is used before and after replacement.

Ground truth is used to select and annotate these illustrations, never to produce their predictions. The appendix and machine-readable case records document selection rules, source hashes, coordinates, and crop bounds. The visual contrast motivates an interpretation; full-set controlled metrics determine whether that interpretation is useful.

I Additional Frozen-Rule Transfer

I.1 Frozen-rule transfer to public baselines

An additional registered experiment applies the same frozen rules to the authors’ public RT-DETR-OBB Nano and Small weights. The rules, group thresholds, score coefficients, and random seed are unchanged. These are public baseline models, not the unreleased full RiO-DETR implementation. The experiment tests model transfer on the same historically inspected DIOR-R dataset, not untouched-data generalization.

Public baseline N full-test AP50 / AP75: single view 62.748 / 36.990, ordinary five-view NMS 67.484 / 38.067, fixed-box scoring 68.448 / 39.708, agreement selection 68.575 / 40.229, weighted geometry 68.622 / 40.901, uniform geometry 68.676 / 41.098, random geometry 67.586 / 38.874. No method is selected or retuned from these outputs.

Public baseline S full-test AP50 / AP75: single view 73.739 / 48.725, ordinary five-view NMS 75.939 / 49.411, fixed-box scoring 76.111 / 50.438, agreement selection 76.173 / 50.905, weighted geometry 76.231 / 51.332, uniform geometry 76.231 / 51.644, random geometry 75.487 / 49.937. No method is selected or retuned from these outputs.

J A Bounded Audit of a Public Release

Question and scope. This audit, frozen on October 8, 2026, asks whether a specific released checkpoint re-

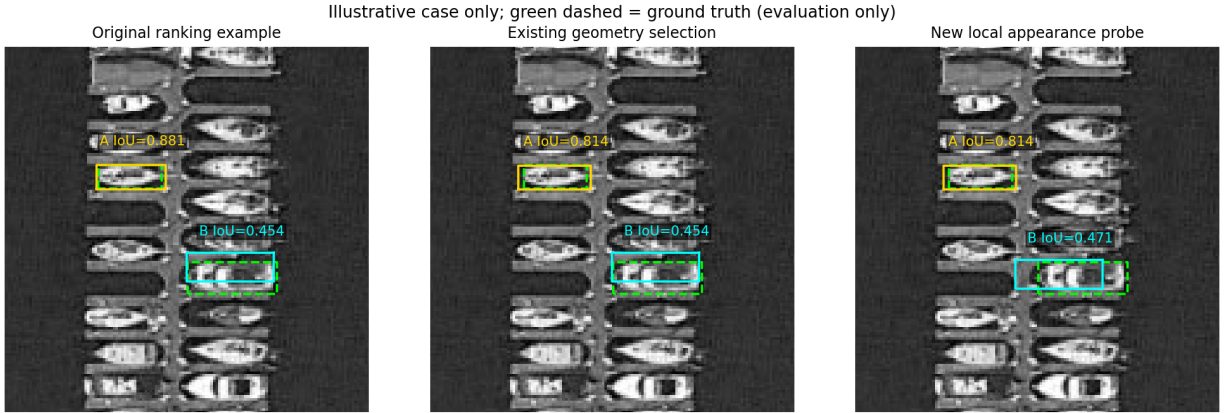


Figure 12: **Transferring evidence from A does not guarantee repair of B.** Left: the earlier ordering example. Middle: the existing uniform geometric selection. Right: the local appearance intervention changes B only slightly. Its proposal pool contains a much better-overlapping box according to an offline ground-truth oracle; that oracle is diagnostic, is not drawn here, and is not the deployed selection rule. The deliberately selected image explains the distinction between candidate opportunity and selection reliability; the full-test comparisons in the main text determine the method’s measured effect.

produces its accompanying DIOR-R metric under the documented public evaluation path. The artifact is the public RT-DETRv2-OBB Large baseline, not the full RiO-DETR implementation and not our independently trained community Large model. The release manifest identifies the exported state as EMA and reports 75.6878 AP50; its checkpoint epoch and best-metric epoch are 56 and 63, respectively [5]. The retained training log reports 75.2193 at epoch 56. Thus, the epoch mismatch alone does not explain the score we observe.

Artifact and evaluation controls. The downloaded checkpoint’s SHA-256 exactly matches the manifest. We resolve the original public configuration and strictly load the checkpoint. Local changes specify dataset paths, disable automatic backbone-pretraining downloads, and disable learnable affine blocks that are absent from the released state. A compatibility adapter maps two legacy image-transform entry points to the installed TorchVision interface without changing their transformation code. These changes are disclosed rather than described as an unmodified execution environment.

Evaluation uses all 11,738 test images, a single view, and the released fourth decoder output. State tensors match the checkpoint after loading, evaluation leaves buffers unchanged, and repeated inference on the first batch is identical. An independently written postprocessing calculation checks coordinate scaling, angle units, and query/class selection on that batch. A separate evaluator uses independently parsed annotations, polygon intersection, greedy matching, and VOC-style 11-point AP to rescore the complete prediction set. It returns 64.6978151 AP50, compared with 64.6978156 from the native evaluator. This agreement supports the reported local measurement; it does not establish identity with

Artifact or diagnostic condition	AP50
Release metadata	75.6878
Unchanged released EMA, native evaluation	64.6978
Same predictions, independent evaluation	64.6978
First decoder output	61.7616
Second decoder output	64.0138
Third decoder output	64.4473
Regenerated deterministic anchors	64.6988
Training-image BN recalibration	0.0002

Table 7: Bounded diagnostics of one public Large baseline. All locally measured values use the full test set. Alternative outputs, anchor regeneration, and recalibration are distinct diagnostic interventions; none replaces the unchanged released-EMA result.

every detail of the original training environment.

What the interventions rule out. Earlier decoder outputs do not recover the reported score, and regenerating deterministic anchors changes AP50 by less than 0.001 point (Table 7). The stored invalid-anchor mask agrees with native generation; finite anchor values show small numerical drift. Eight BN variance entries are approximately -1.4×10^{-12} , while their normalization epsilon is 10^{-5} , leaving positive normalization denominators. These observations are recorded without treating them as demonstrated explanations of the accuracy gap.

A separate calibration hypothesis estimates population moments from 256 fixed training images in two passes, updating all BN layers simultaneously after each pass. No test images or test annotations enter calibration. This particular procedure collapses accuracy and is rejected. It neither repairs the release nor establishes

that other calibration procedures would fail. Released weights and the community training runs remain unaffected.

Interpretation and disposition. The correspondence among this checkpoint, its evaluation configuration, and the accompanying metric remains unresolved. Our checks do not identify responsibility, prove an error in the paper, or establish a general defect in the public framework. The standard split also contains the previously documented cross-split duplicates; independent metric implementation does not remove that data limitation.

We archive the checkpoint digest, configuration, release metadata, relevant training-log entries, complete metric receipts, and calibration protocol. The diagnostic is closed without further experimental allocation. It remains an artifact-provenance case in this appendix, rather than a result used to strengthen our method claim. Community reconstruction proceeds from its own documented initialization and training, with the paper’s reported scores serving as performance references.

K Visual Comparisons with Explicit Sources and Controls

An artifact comparison is not a module ablation. Figure 13 places the public RT-DETRv2-OB B Large baseline and our independently trained Large model on identical DIOR-R crops. Their locally evaluated full-set AP50 values are 64.6978 and 76.1990. The former is the released baseline, not the unpublished complete RiO-DETR model. Its unresolved correspondence to the author’s reported metric is documented in Appendix J. Architecture, training history, and state differ; this comparison cannot assign the gap to a particular module or claim a controlled improvement over the author’s full method. Two examples improve localization and one worsens it.

A fixed-source comparison isolates coordinate selection. Figure 14 uses the earlier frozen Medium source released with the inference recipe, not the later 76.448 primary checkpoint. Its single-view AP50 is 74.7898. Agreement-selected five-view anchors yield 77.4687 AP50 and 54.8637 AP75; coordinate-only uniform consensus yields 77.5115 and 55.4299, whereas random members yield 77.0916 and 53.3546. The five-view cost is shared by all three panels. The first improvement includes additional image views and selection, while the latter contrast isolates geometric replacement at fixed scores and identities. In particular, the two increments must not be added to the latest primary Medium score.

Reproducibility and limits of visualization. The manifests record source-cache digests, image digests, pre-

diction identities, coordinates, scores, thresholds, and the deterministic case-selection rule. Within each chosen class, cases are nearest the median absolute IoU change among eligible upward or downward crossings, with image and object indices breaking ties. Preferred classes are ship, airplane, vehicle, bridge and baseball field. We do not label this a random or population-representative sample. The independent polygon implementation recomputes displayed IoUs; portable rendering needs only the manifests and the user’s dataset images. No attention weights or feature embeddings were measured in this experiment, so these figures make no claim about attention concentration, angular feature clustering, or representation disentanglement.

L Registered Extensions Beyond the Reproduced Family

This section records extension protocols deployed after the original training evidence cut. YOLO transfer and quality-readout evaluations have completed, with results in Section 7.7; the registered foundation-model guidance comparison has also completed. OpenRiO remains the primary reproduction, while the second architecture tests the reach of its optimization lessons.

Transfer before redesign. An existing YOLO26-N oriented detector, initialized from DOTA and fine-tuned on DIOR-R for 50 epochs, supplies a different prediction paradigm. Its legacy full-set predictions yield 69.5153 AP50 and 49.3750 AP75 under the independent evaluator. This training history differs from OpenRiO, so absolute scores do not constitute a matched-budget architecture ranking. A complete fresh five-view evaluation applies the previously frozen support and geometric rules, including fixed-set, uniform, and random controls, without fitting new coefficients to YOLO.

Can the single view supply its own quality evidence? Multi-view agreement makes an ordering defect actionable, but incurs additional detector passes. A simpler deployment hypothesis is that existing single-view features already contain some of the missing quality information. We freeze the detector and attach a small quality readout after native prediction selection. For each retained box, a symmetric rotated 3×3 grid samples the three feature maps entering the YOLO detection head. Average pooled visual features are combined with the score logit, log width and height, periodic angle coordinates, and class indicators. A 64-unit hidden layer predicts whether maximum same-class polygon IoU exceeds 0.75. Every coordinate, category, identity, and prediction count remains fixed.

Targets come only from the original training and validation image identifiers, excluding training images with audited byte-identical test counterparts. All ground



Figure 13: **The same scene can expose both progress and remaining errors.** The displayed box is the best-overlap same-class candidate with score at least 0.3 for the selected annotation; other predictions are hidden for readability. Dashed green denotes ground truth. Selection uses median absolute IoU changes among eligible within-class threshold crossings, including a counterexample, after scanning the full test set. Thus these are deliberately explanatory crops, not unbiased images or a precision comparison at a shared operating point. Image identifiers and exact boxes are retained in the figure manifest. DIOR/DIOR-R image credit: Gong Cheng and collaborators; image excerpts retain the dataset’s attribution terms.

One frozen Medium source: coordinates change; identities and scores stay fixed
Dashed green: GT. Geometry may improve or deteriorate.

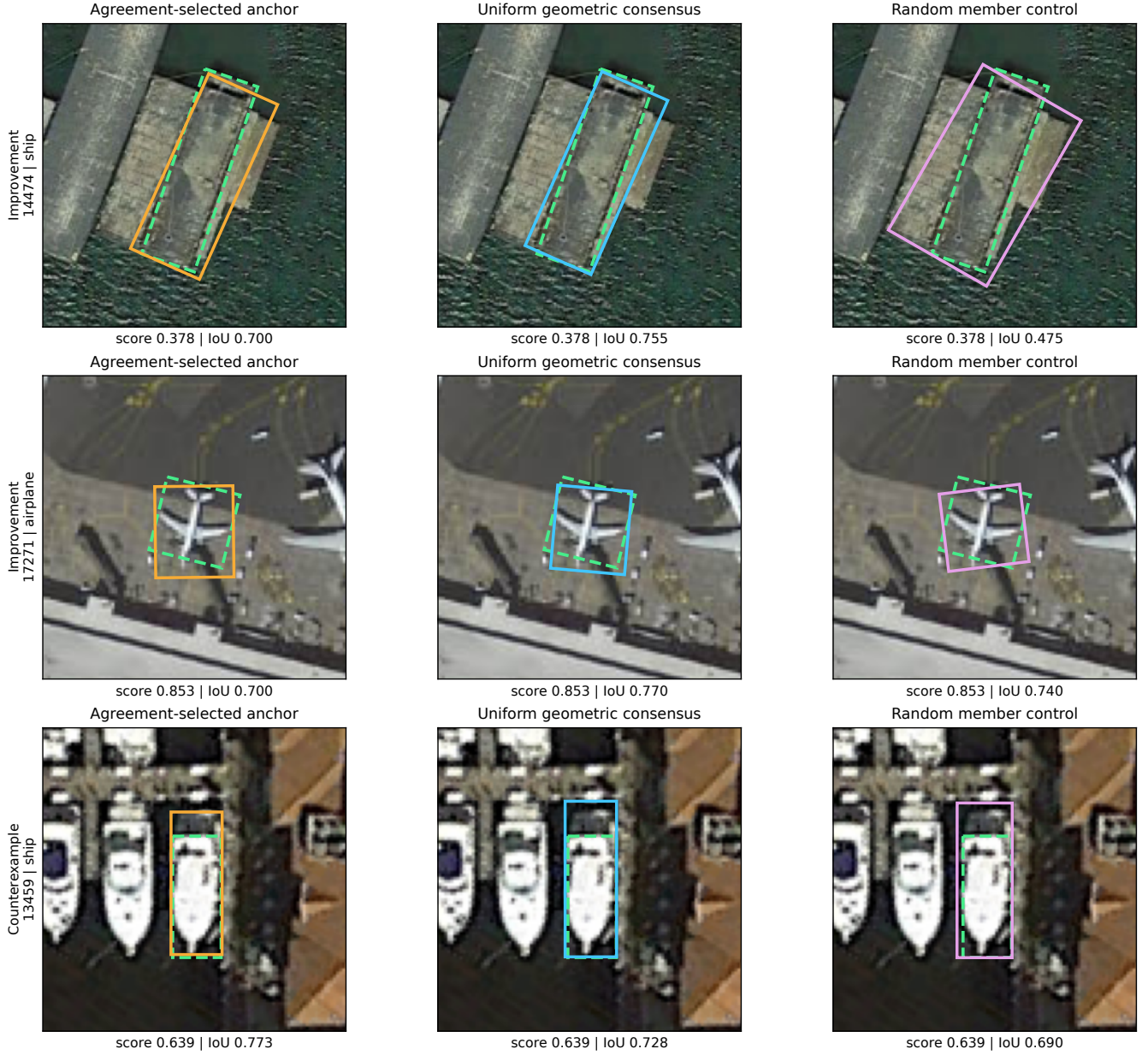


Figure 14: **Structured selection can repair a box and can also choose the wrong extent.** Columns share the same frozen model, image, anchor identity, class, and score; only coordinates change. Two median-change crossing examples and one reverse crossing illustrate the rule’s benefit and limit. The random control samples the same available groups, and ground truth is never a selection input to any inference method. Only one highlighted output is shown per panel. Source identities, exact coordinates and selection rules are recorded. DIOR/DIOR-R image credit: Gong Cheng and collaborators.

truths participate in target matching. Uniformly sampling at most 64 predictions per training image bounds feature-cache size; this does not truncate annotations or test outputs. Image-content hashes allocate a 20% training-side monitor for the readout. The frozen detector has previously seen these images, so this is not an independent detector validation set.

Three readouts receive the same registered eight-epoch budget: score and geometry metadata alone, metadata plus visual features, and visual features with class-preserving shuffled training targets. Training uses AdamW with learning rate 10^{-3} and batch size 2,048. The prediction rule is fixed in advance as $s' = s(0.75 + 0.25q)$. Native scores and a classwise monotone rescaling provide additional controls. Test annotations enter only the final evaluator, after readout fitting and prediction scoring. Both AP50 and AP75 on all 11,738 images are required; an improved auxiliary loss alone does not establish a better detector. Localization-aware scoring is established prior art [31]; the question here is whether a small frozen-detector readout repairs the diagnosed failure under these controls.

What each outcome would mean. If visual features outperform both metadata and shuffled targets in complete-set AP, a single-view quality correction becomes a candidate community artifact. If metadata suffices, the extra feature path is unnecessary. If quality loss improves but AP does not, quality estimation and detection ordering remain misaligned. If multi-view consensus helps while every single-view readout fails, the evidence favors acquiring complementary predictions over this attempted compression. None of these outcomes identifies an architectural cause merely from cross-model score differences.

Scope of the v4 extension. Local RT-DETRv4 Small through Extra-Large checkpoints are horizontal detectors. A compatible backbone is useful infrastructure but does not establish an oriented decoder, training loss, or benchmark result. Motivated by its training-only foundation-model guidance [15], our first oriented adaptation retains the verified Nano student and uses a locally available DINOv3-L teacher. It differs from the official teacher size and loss. Teacher inference uses BF16, a microbatch of one, and the same augmented image as the student, normalized and downsampled from 800 to 400 pixels. After removing the class and register tokens, its 25×25 patch grid aligns with the student’s stride-32 feature map.

Rather than learning a channel adapter, the loss matches off-diagonal pairwise cosine similarities between spatial positions. Relation transfer is established prior art [21]; this spatial implementation is a community hypothesis, not an exact reproduction of that method or of v4. A spatially shuffled teacher retains the same pre-trained weights but breaks positional correspondence.

Both interventions use loss weight five, learning rate 10^{-6} , batch eight, frozen normalization buffers, and primary EMA decay 0.999. The eight-epoch endpoint is compared with the first eight epochs of the already registered same-start, same-seed, constant-rate teacher-free continuation. No teacher or added parameters enter the exported detector. Real eight-batch probes verify finite student gradients, absent teacher gradients, and the memory budget before training. At the registered eight-epoch horizon, the teacher-free, aligned-teacher, and spatially shuffled conditions obtain 71.8700, 71.5507, and 71.5651 AP50. The aligned intervention is 0.3193 points below its matched control and 0.0144 below the shuffled condition. Correct spatial correspondence does not explain a useful gain in this recipe. We therefore do not extend this particular loss merely to spend more compute. This result bounds the tested teacher, layer, loss, and near-converged student state; it does not establish that foundation-model guidance or the official v4 method is ineffective. Foundation-model pretraining overlap is not independently audited.

M Extra-Large: Matching Structure Does Not Match Training

This supplementary audit was initiated on October 12 and updated on October 15, with the main reproduction table synchronized to the latest independently audited state. Its purpose is to identify plausible interventions, not to retrospectively explain a score by model capacity. The original repository revision and Extra-Large console log were rechecked against the current public release [10]; the log is unchanged from our October 4 snapshot. Our selected uncalibrated, single-view EMA reaches 75.1264 AP50 on all 11,738 DIOR-R test images, versus the paper’s 77.43. The original console log peaks at 77.6152; that distinct observation does not replace the paper benchmark.

A structural hypothesis that the evidence does not support. Extra-Large uses an HGNetv2-B5 backbone, 384-dimensional embeddings, three feature levels, and four decoder layers. Our reconstruction already implements these scale changes. Its 63,843,517 training parameters exactly match the original console log. The appendix hyperparameter table in RiO-DETR [11] describes DOTA and lists a feedforward width of 2048 for Extra-Large. The public DIOR configuration explicitly applies that width to the encoder, while the public decoder defaults to 1024. Our encoder and decoder follow this distinction. Widening all four decoder feedforward layers to 2048 would add 3,149,824 parameters and break the recorded count agreement. This is therefore an optimization hypothesis, not an established reproduction correction. Count agreement constrains possible implementations but does not establish equivalence of unpublished,

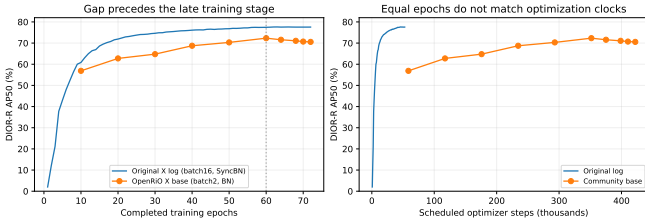


Figure 15: The Extra-Large gap appears before the final training stage. Curves show original-log and community-base EMA evaluations, aligned by completed epochs (left) and scheduled optimization steps (right). Scheduled steps include computation later removed from model ancestry by the author’s stage restoration; neither axis equalizes all training conditions. The two trajectories confound batch statistics, optimization, averaging, and reconstruction choices, so this comparison does not isolate a cause.

parameter-free geometric operations. A fresh CPU check strictly loaded the selected EMA and verified a finite full-resolution forward pass, including all four spatial query inputs; it is a structural check rather than an additional AP evaluation.

Training clocks provide a more concrete discrepancy. The original DIOR log records total batch16 across two ranks with synchronized batch normalization; our memory-constrained base run used physical batch2 and ordinary batch normalization. There are 732 versus 5862 scheduled optimizer updates per epoch. Consequently, an identical 2000-update warmup covers approximately 2.73 versus 0.34 epochs. With the same per-update EMA decay of 0.9999, the asymptotic half-life corresponds to approximately 110,898 versus 13,862 images. Warmup and checkpoint restoration modify the actual averaging history. The current continuation already normalizes EMA decay by image count and freezes normalization statistics, but it does not retrospectively change the base trajectory. The local cosine branch also differs from the original log’s learning-rate floor, and the author’s recorded restoration of the first-stage best state at epoch60 differs from our uninterrupted base phase followed by separate low-rate restarts.

At 20 completed epochs, the original log reports approximately 71.90 AP50 and our base run 62.75; by epoch60 our base reaches 72.32 before declining to 70.56 at epoch72. Subsequent community continuations recover and exceed that base result. At the earlier October 12 cut, two completed observation windows improved the best score by 0.331 and 0.153 points, respectively, so this route has not met its registered plateau criterion. Conversely, the tested two-layer normalization recalibration lowers the corresponding terminal EMA scores by 0.178 and 0.135 points. Together, these observations motivate matched-start tests of batch, averaging history,

and stage restoration. Gradient accumulation must be distinguished from a physical large batch: it does not reproduce synchronized normalization statistics, and averaging independently normalized microbatch losses need not reproduce the native full-batch objective. A physical batch16 reference requires a separate full-target memory admission test. The LR and stage-restoration contrasts are now registered and running or awaiting their shared branch point; the audit itself does not establish a causal explanation or a successful Extra-Large reproduction.

A fresh full-test evaluation of the updated selected EMA obtains 75.126405 AP50 and 52.643583 AP75, within 0.000556 AP50 of its training-side record. Annotation hashes were independently rechecked. This is a new inference audit of the same training seed, not an additional reproduction seed.

N Preserving the Objective Under a Memory Budget

The relevant question is not whether a larger backbone is inherently harder to reproduce, but which conditions remain invariant when its execution is changed. Our DIOR-R Medium and Large point estimates already exceed the paper references. Extra-Large retains the prescribed B5 backbone and logged parameter count, yet its early training trajectory differs. FAIR adds a second ambiguity: model size, training stage, and evaluation units change together. We therefore separate the units of *images*, *targets*, and *optimizer updates* before interpreting a performance gap.

An objective-level discrepancy. Let S_j be an unnormalized loss numerator and n_j its positive count in microbatch j . With a common model state and fixed predictions,

$$\frac{1}{K} \sum_{j=1}^K \frac{S_j}{\max(n_j, 1)} \neq \frac{\sum_{j=1}^K S_j}{\max(\sum_{j=1}^K n_j, 1)} \quad (17)$$

in general. Without the empty-target clamp, the left side assigns each microbatch a multiplier proportional to $1/n_j$ rather than a common target weight. The public criterion uses target counts for classification and a cross-layer matching-union count for ordinary box regression. Consequently, one scalar correction is insufficient. Our isolated worker accumulates two numerator-gradient sums and divides them by their respective group totals before gradient clipping and the AdamW update. Denoising terms retain their local replica-group average and use the combined target count. No image or ground-truth object is truncated.

This is an implementation of an explicitly stated objective, not a new loss or a general equivalence claim for physical batches. Global-batch denoising padding and random draws are not reproduced by local forward

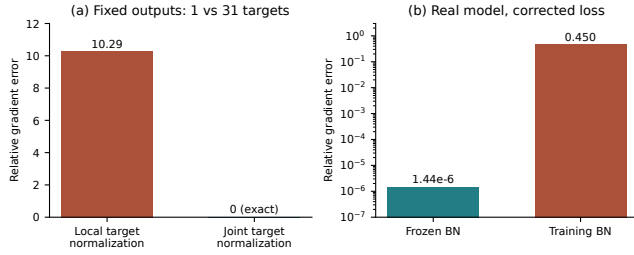


Figure 16: Two different obstacles to batch equivalence. Left: the actual focal loss on constructed outputs with 1 and 31 targets; independent microbatch normalization differs from a joint target denominator. Right: two real training images and the Medium model, with corrected loss normalization; the remaining discrepancy depends on whether BN statistics are shared. Denoising is disabled in the real-model diagnostic. These are gradient checks, not detector-accuracy experiments or population estimates.

passes. Batch normalization also remains local unless separately changed. We check the algebra using the actual periodic box, divergence, focal, auxiliary, encoder, and denoising losses, including empty-target batches. With fixed outputs and common denoising-group layout, eight cases agree with a joint-batch reference to below 10^{-5} relative gradient error. Zero-query denoising classification is explicitly evaluated as an empty sum; the native mean-times-count expression otherwise passes through 0/0. Other nonfinite branches are rejected before generic NaN replacement.

Separating normalization statistics from loss normalization. On two fixed, augmented FAIR training patches containing 2 and 19 objects, we compare the real Medium model’s joint physical batch of two with two single-image forwards. Denoising is disabled only for this diagnostic so that its padding does not confound the comparison. After the loss correction, frozen batch-normalization statistics give relative gradient error 1.44×10^{-6} . With training-mode statistics, the error is 0.450. These conditional checks establish distinct mechanisms; two images do not estimate their prevalence or predict an AP gain.

A protocol discrepancy precedes the FAIR accuracy comparison. The original Medium and Extra-Large logs both report 57,967 validation images and 503,948 ground truths. Our native parser instead counts 807,917 ground truths on 71,217 crops. After merging predictions, evaluation uses 201,311 ground truths in all 8,287 original validation scenes. Every original XML annotation was checked against its converted text file; the per-image classes, polygon coordinates, and difficulty flags agree. Removing difficulty-two crop annotations still leaves 791,934 objects and does not re-

solve the discrepancy. Nor do the inspected common scale/stride alternatives recover the logged image count. These checks localize an unresolved data/preprocessing boundary; they do not recover an unpublished source-image manifest.

Evaluation population	Images	Targets
Original FAIR logs	57,967	503,948
Community crops	71,217	807,917
Community scenes	8,287	201,311

The paper’s 43.6/47.4 references therefore remain targets rather than validated differences from our scene AP. The community split retains all validation scenes and excludes 54 byte-identical training duplicates. We do not select a crop or difficulty convention by its AP. The original complete-method FAIR logs disable AMP, as our runs do; the public baseline configuration’s AMP flag is a different reference.

Registered accuracy tests and their present boundary. Two Medium continuations start from the same epoch-fourteen EMA, with identical fresh AdamW state, data ordering, learning-rate schedule, local BN, and EMA decay. Only target normalization differs. They evaluate all crops and all merged scenes every two epochs, retain the full schedule through epoch36, continue a still-improving endpoint, and stop sustained large deterioration. Extra-Large transfer requires repeated matched-epoch gains and memory admission; this resource-allocation rule is not a significance test. Separate frozen-BN and physical-batch-four controls have passed dense and real-data admission. The physical-batch control is running in parallel; the frozen-BN continuation awaits GPU release. The physical-batch control also changes denoising grouping and may change worker-level augmentation grouping, so it is not treated as a pure BN intervention.

For DIOR Extra-Large, a state-preserving continuation from epoch40 uses the original log’s learning-rate floor. At epoch60 it forks a matched stage-restoration contrast, preserving the shared earlier trajectory. This intervention does not recover the author’s earlier physical-batch or SyncBN history. At this evidence cut the accuracy trials are running or conditionally scheduled: no AP benefit from the normalization correction, BN change, or stage restoration is claimed.

O From a Reproduced Detector to a Computational Budget

This section tests three computational budgets using matched controls. OpenRiO’s organizing contribution remains a usable, auditable community reproduction. Better execution, lower training exposure, and improved predictions are separate extensions of that artifact.

Historical reproduction boundary (October 14 snapshot). At the October 14 evidence cut, native full-set, single-view, uncalibrated EMA AP50 is 71.8908, 75.0200, 76.4483, 76.1990, and 75.0226 for Nano through Extra-Large. The corresponding paper references are 71.92, 74.44, 75.73, 76.11, and 77.43. Medium’s registered continuation has reached an empirical plateau, as has Large. Extra-Large continues under new recipe controls. Its older continuation resumed from its terminal optimizer, EMA and random states, preserving the prior plateau history. Fresh independent re-inference now verifies all five previously frozen endpoints. A later Extra-Large continuation reaches 75.0226 AP50 after four resumed epochs on the complete test set, a 0.0761-point increase. Its separate fresh evaluation gives 75.0230/52.9344 AP50/AP75 and passes independent CPU matching and AP verification. At the end of that 24-epoch window, unchanged EMA is 74.9571, raw weights 75.0016, and calibrated EMA 74.7775; none surpasses the selected unchanged-EMA endpoint. At that historical cut its paper gap was 2.4074 points; the latest audited 75.127 endpoint in Table 1 supersedes this status. These single-lineage observations establish neither seed robustness nor the cause of a remaining gap. Existing X schedule, normalization-clock and effective batch controls have priority over the new optimization pilots.

Can localization knowledge reduce required depth? Training-only foundation-model guidance did not improve the previously tested near-converged student over its matched control (Appendix L). A more direct question is whether localization already present in later decoder layers can improve earlier predictions. D-FINE introduces distribution refinement and localization self-distillation [22]. Our first test retains the existing point-regression detector, isolating conditional localization transfer before changing its representation. It is not an implementation of FDR or a reproduction of GO-LSD.

Let $C(b) \in \mathbb{R}^{4 \times 2}$ be cyclic corners of a predicted rectangle, and let $\mathcal{P}$ contain their cyclic and reversed permutations. For a training target g , define

$$d_g(b, t) = \frac{\min_{P \in \mathcal{P}} \|C(b) - PC(t)\|_F^2}{4 \max(w_g^2 + h_g^2, 10^{-6})}. \quad (18)$$

This is a normalized squared discrepancy, not a claimed metric. A change of equivalent angle or width–height encoding only permutes corners; closure of $\mathcal{P}$ leaves the minimum unchanged. Unlike covariance alone, the corner representation distinguishes a square from its 45-degree rotation. The minimum is piecewise differentiable, with ties handled by the implementation.

The final-layer Hungarian assignment fixes object identity. For each matched query, let b^ℓ be an earlier prediction and b^D the detached deepest prediction. The proposed training gate is

$$a_{\ell g} = \mathbf{1}\{d_g(b^D, g) + 10^{-4} < d_g(b^\ell, g)\}. \quad (19)$$

Transfer uses $a_{\ell g} d_g(b^\ell, b^D)/(1 + d_g(b^D, g))$, averaged over all eligible matched pairs. Teacher values and gate decisions are detached. An ungated control retains the same reliability weights and denominator. The original detection objective still supervises every ground truth; the gate removes only an auxiliary teacher target. No inference parameter is added. A locally better teacher under this proxy does not imply better IoU, ranking, or AP after optimization. The decisive outcome is the complete-set accuracy–latency frontier at two and three decoder layers, relative to both ordinary continuation and ungated transfer.

Can a detector need fewer image presentations?

SeTa motivates dynamic training-example budgets [33]; extending pruning to detection also requires object attribution and image-level aggregation [30]. We begin with controls that do not require an additional scoring network. Each epoch samples half the complete training images either uniformly or by training-label strata: class rarity, small-object presence, density, and elongation. Selected images retain every annotation. This policy is neither SeTa nor an assertion of unbiased detection gradients.

The cleaned pool contains 11,702 images, excluding 23 audited exact cross-split duplicates. The inherited detector’s earlier exposure is not erased by this exclusion. Four full-data epochs and eight half-data epochs each present 46,808 images, but have 23,404 and 23,408 updates respectively because of singleton final batches. Both counts are reported. Four-epoch half-data runs test a reduced exposure budget; the eight-epoch random control separates that budget from a longer epoch count. Selection overhead, actual images, updates, and wall time accompany full AP50/AP75. Success here would concern post-training only. A from-scratch convergence claim requires a new matched experiment from an earlier initialization. Le-DETR’s reduction of backbone pre-training cost [12] concerns a different resource and does not justify discarding the same fraction of DIOR training data.

Execution is another independent budget. The registered systems test uses the current Nano checkpoint to compare native execution, two-layer decoding, Conv-BN fusion, and CUDA Graph. Fusion changes floating-point arithmetic and therefore requires complete accuracy evaluation. Timing spans preloaded compressed image bytes through owned CPU detections, including pre-processing and applicable device transfers. Every backend receives warmup; paired input order is randomized. Shared-host or shared-device measurements are labelled explicitly. They do not establish an isolated service latency. Earlier graph speedups and their full-set bitwise parity remain tied to their original checkpoint.

Evidence required before expanding the architecture. Six fixed pilots compare ordinary continuation, ungated transfer, gated transfer, random half exposure, stratified half exposure, and equal cumulative random-half exposure. They use the same selected Nano EMA, seed, optimizer settings and normalization policy. Each endpoint evaluates all 11,738 test images; repeated use of this test-exposed lineage is disclosed. Actual-model CPU forward/backward, matching, sampling and fusion checks pass. Dense GPU memory probes pass with all 4,256 and 3,620 targets retained in the two stress fixtures (peak allocation 2.01 GiB approximately). The three full-exposure pilots have completed. Their three-layer AP50/AP75 is 71.8575/46.3024 for ordinary continuation, 71.8860/46.2818 for ungated transfer, and 71.8738/46.1982 for gated transfer. At two layers, the corresponding pairs are 70.7106/44.8626, 70.6922/44.8380, and 70.6845/44.8898. All three use 46,808 training-image exposures and complete test evaluation. Transfer does not establish a better shallow accuracy frontier over its matched control: the small gains are mixed across thresholds and depth, and only one seed has been tested. The random-half pilot has now completed 23,404 image presentations, half the full-exposure control, with three-layer AP50/AP75 of 71.8662/46.1264 and two-layer AP50/AP75 of 70.7565/44.8862. Its three-layer AP50 is within 0.01 points of ordinary continuation, but AP75 is 0.1761 lower. This is a useful exposure-budget observation, not evidence that training from scratch requires half the data. The coverage-half endpoint is 71.8723/45.9017: its AP50 remains close to the full-exposure control, but AP75 is 0.4007 points lower. The equal-exposure random-half control reaches 71.7891/46.0359 after eight epochs and 46,808 image presentations, below the full-exposure control by 0.0685/0.2665 points. All six registered training pilots are now complete. These controls do not establish an advantage of label-stratified selection over random selection or a free reduction in localization training cost. Shared-device training times do not establish a convergence-speed advantage.

The full GPU evaluation measures AP50/AP75 of 71.8908/46.0546 with three decoder layers and 70.7973/44.8193 with two. The latter therefore has to recover 1.0936 AP50 and 1.2353 AP75 to match the current detector. On a shared device, median end-to-end latency changes from 27.08 to 25.52 ms. Graph execution of the three-layer detector measures 21.49 ms on the same timing sample; this sample has only 32 distinct images, with identical raw outputs between graph and eager execution. A subsequent complete test-set check verifies bitwise identity of boxes and logits on all 11,738 images at 300 queries, preserving AP50/AP75 exactly. In that later paired timing run, median latency changes from 34.17 to 30.36 ms, an 11.15% reduction under shared-device load. These timings are contextual observations, not an isolated release benchmark; they must not be com-

bined with speedups from the earlier checkpoint. CPU evaluation on all 11,738 images gives 71.8904/46.0551. Conv-BN fusion does not improve median latency on either backend; the GPU AP75 shift of -0.0137 exceeds the registered 0.01 tolerance. We consequently do not adopt fusion as a default or combine its timing with graph speedups.

Separating metric correctness from checkpoint-state choice. An independently implemented OpenCV polygon intersection, greedy matching, and VOC2007 AP calculation verifies the current Nano predictions on all 11,738 images. Three-layer AP50/AP75 is 71.890834/46.054608; two-layer AP50/AP75 is 70.797279/44.819264. Each of the four differences from the native evaluator is below 10^{-6} percentage points. The complete test XML digest also matches the independently parsed annotations. This supports correctness of the implemented metric, without establishing identity to every paper protocol detail or removing historical duplicate exposure and adaptive test selection.

For each of the five frozen scale endpoints, a separate registered audit compares raw weights, unchanged EMA, EMA with raw normalization buffers, and two training-only calibration variants. Each calibration set contains 256 distinct images; the sets are disjoint and exclude the known exact cross-split duplicates. Calibration changes population moments at two specified backbone normalization sites while preserving learned parameters. All 25 scale-state combinations have now completed full inference and passed independent polygon matching and AP verification. The primary result remains unchanged EMA. Sensitivity to the calibration subset is not training-seed variance, and a better diagnostic variant will not silently replace the reproduction protocol. Fresh EMA AP50 differs from its recorded source by less than 0.0005 points at every scale. Both calibration variants reduce AP50 at every scale: their changes range from -0.049 to -0.064 for Nano, -0.303 to -0.378 for Small, -0.231 to -0.324 for Medium, -0.110 to -0.132 for Large, and -0.210 to -0.213 for Extra-Large. Replacing EMA normalization buffers with raw buffers does not reveal a useful recovery either. This endpoint evidence supports retaining unchanged EMA; the historical normalization repair is conditional on checkpoint history, not a universal post-training improvement.

The fixed-weight query study registers budgets of 300, 200, and 100 actual decoder queries, with full-set eager-graph output checks and paired byte-to-CPU timing. The current Nano uses learned query content; reducing the encoder selection count alone is therefore insufficient. The probe retains the first K learned embeddings together with K selected references, and verifies exact restoration at $K = 300$. It leaves the final top-300 class-query detection budget unchanged. This is a test-time budget probe rather than a reproduction of a learned query-budget method. Complete coverage diagnostics

separately measure whether correct-class candidates exist before attributing losses to ranking. Their ground-truth oracle quantities are diagnostic upper bounds, not deployable selection rules.

The complete diagnostic subsequently covers 124,443 ground-truth objects, including 75,315 with area below 32^2 pixels. For these small objects, the three-layer detector’s retained top-300 candidates attain 75.00% class-agnostic coverage at IoU 0.5, compared with 74.73% correct-class coverage. Correct-class coverage at IoU 0.75 is only 43.21%; the corresponding all-object values are 83.64% and 60.16%. This small class-agnostic gap motivates examining missing candidates and geometric precision before attributing small-object failures mainly to class assignment. Coverage is not AP, and the diagnostic does not observe candidates discarded before the top-300 output. It does not isolate a backbone cause or rule out ranking gains among existing boxes. The paired two-layer predictions and completed localization-transfer controls provide a more specific test of which localization quality can be retained with less decoder computation; the tested auxiliary transfer did not improve the shallow frontier over ordinary continuation.

If conditional transfer improves the shallow frontier over both controls, periodic residual distributions become a motivated next representation to study. Width–height equivalence, square orientation, confidence calibration, and elongated-object errors must remain explicit. A stronger architecture claim would require matched point/distribution baselines, seeds, another oriented dataset and complete accounting of training cost. If simple budget controls fail, that result instead limits which computation can be removed from the present reproduction without harming its utility.

Related directions and the next controlled comparisons. The October 13 literature audit separates computational budgets rather than equating all forms of efficiency. BLS estimates sample importance from batch losses [35], while RePB addresses score-context drift and temporal gradient bias [34]. MCL emphasizes explicit rare-concept coverage [28]; IF-Beta studies data selection for knowledge distillation [29]. These motivate testing score cost, temporal comparability, and object coverage independently. Their guarantees do not transfer automatically to a detection loss normalized by variable object counts. In particular, preserving expected gradients under sampling would not establish an identical optimization trajectory or an AP guarantee.

PaQ-DETR studies query utilization and quality-aware supervision, whereas Super Sparse DETR connects learned selection to exported structured computation [14, 37]. The official NeurIPS 2026 abstract for BUTR explicitly separates query count from query choice [32]. This motivates a conservative fixed-weight query-budget probe before changing assignments. We have inspected its official abstract, not a complete im-

plementation. Dynamic query lengths may conflict with graph capture: a small bank of fixed budget graphs must account for selection synchronization and cached memory. Thus query sparsity and graph acceleration require a joint measured frontier, not multiplication of two independently reported speedups. Our completed 200-query probe also preserves eager–graph outputs bitwise on the full test set, but AP50/AP75 drops to 70.8457/45.4156, losing 1.0451/0.6390 points relative to 300 queries. Its paired eager median is 30.71 ms versus 29.80 ms for native 300-query eager execution; shared-device variability prevents attributing this small timing difference to query count alone. The result does not establish an additional query-reduction speed advantage. A 100-query probe completes the registered comparison with AP50/AP75 of 68.8122/43.3421, losing 3.0786/2.7125 points relative to 300 queries. Eager and graph outputs remain bitwise identical on all 11,738 images for this budget as well. The observed accuracy cost supports retaining 300 queries as the default while optimizing their execution; it does not rule out a trained selection policy.

For missed candidates, FSDC-DETR motivates preservation of fine spatial detail [18]; this differs from reranking boxes that already cover the target. For explicit domain shifts, RT-SFOD motivates selective pseudo-label fusion and detection-aware feature diversity [23]. Neither entails a gain on clean DIOR-R. These studies guide future controls; they are not completed OpenRiO extensions. The community artifact and its verified reproduction remain the organizing contribution, with new claims limited to independently measured improvements over that artifact.

P Conditional Capacity After Reproduction

A reproduced detector makes it possible to ask whether conditional capacity adds useful information after ordinary training has largely saturated. YOLO-Master motivates conditional expert computation in real-time detection [17]. Its design is not transplanted unchanged: our bounded pilot operates on decoder queries and retains the entire pretrained detector as a frozen shared path.

For the input z to the final decoder feedforward output projection, we use

$$y = Wz + b + \sum_{i=1}^E p_i(z) U_i \text{SiLU}(V_i z), \quad E = 4. \quad (20)$$

Each residual expert has rank eight. Setting every U_i to zero recovers the original detector exactly at initialization. Only the residual experts and router are trained; normalization buffers and pretrained parameters remain fixed. Moving averages apply only to the new parameters, avoiding roundoff drift from repeatedly averaging the frozen source. Router weight decay is controlled sep-

arately, motivated by the diagnostic option added upstream, rather than by an established causal claim about router collapse.

Three matched pilots distinguish added capacity from conditional utility: a single residual bottleneck of width 32, four uniformly mixed rank-eight experts, and the content-dependent mixture. The same trained dynamic model is then evaluated with training-image mean probabilities. This intervention holds expert representations fixed while removing input dependence. If $e_i(z) = U_i \text{SiLU}(V_i z)$ and $d_i = p_i(z) - \bar{p}_i$, the squared local feature change is exactly $d^\top G(z) d$, where $G_{ij}(z) = e_i(z)^\top e_j(z)$. Probability variation alone therefore cannot establish the utility of routing; expert-output geometry matters as well. This feature identity provides neither an AP guarantee nor a causal explanation for dataset-level gains.

Constant mixing admits a useful deployment control: stack the input projections and concatenate the probability-weighted output projections to form one width-32 bottleneck. Mathematical equivalence still requires numerical and complete-dataset acceptance. A separate Top-2 inference probe dispatches only selected expert-query pairs. Kernel launch, indexing, and transfer costs remain included in paired end-to-end timing; sparse dispatch is not itself evidence of lower latency. The shared detector is always evaluated, so an added residual module does not imply a free speedup over the original model.

Nano and Small received two-epoch matched pilots using all 11,702 clean training images per epoch and the same seed. Every endpoint and registered inference variant uses the complete 11,738-image test set. Mean routes are fitted on 256 training images, without test-label fitting. Medium and Large are registered extensions conditional on a positive matched routing signal. That condition allocates computation; it does not establish significance. Historical test-guided source selection remains disclosed. Dense probes at all four scales retain 4,256 and 3,620 objects in two stress images, preserve the source state exactly, and pass backward and finite-output checks.

Routing activity does not establish routing utility. All six Nano/Small training pilots and their inference interventions have completed. Table 8 reports full-test endpoints without selecting a replacement reproduction checkpoint. Dynamic routing is below the best matched control by 0.0081 AP50 for Nano and 0.0101 for Small. Replacing the same dynamic model’s routes with their training-image mean changes AP50 by less than 0.01 at either scale. The predeclared expansion condition is therefore not met, and Medium/Large extensions are skipped. Independent CPU polygon matching and AP evaluation passes for all twelve prediction sets; the maximum absolute disagreement across their 24 AP50/AP75 values is 0.001464 points. This checks the implemented

Table 8: Frozen-source expert pilots on the complete DIOR-R test set. Each trained arm sees 23,404 training images. Fixed mean, fused mean and Top-2 reuse the dynamic arm’s weights. These are single-seed, two-epoch post-training observations, not converged architecture comparisons.

	Nano		Small	
	AP50	AP75	AP50	AP75
Frozen source	71.8908	46.0546	75.0202	51.5465
Single bottleneck	71.8949	46.1661	75.0177	51.5151
Uniform experts	71.8841	46.0774	75.0216	51.5030
Dynamic experts	71.8868	46.0758	75.0116	51.5091
Fixed mean	71.8785	46.0711	75.0190	51.5137
Fused fixed mean	71.8785	46.0711	75.0190	51.5137
Top-2 dispatch	71.8797	46.0876	75.0064	51.5123

metric, not seed robustness or complete equivalence to every paper protocol detail.

On the training calibration images, expert probability standard deviations across tokens range from 0.050 to 0.082 for Nano and from 0.041 to 0.074 for Small. A nearly uniform mean allocation therefore does not establish a constant or collapsed router. The fixed-route intervention instead shows that this measured input dependence has little effect on full-set AP under the tested budget. Expert similarity, small residual magnitudes and downstream insensitivity remain possible explanations; the present results do not distinguish them. Nano’s single bottleneck improves AP75 by 0.1115 over the frozen source, but this is neither a routing gain nor multi-seed evidence of a robust improvement.

Fixed-route fusion preserves the reported AP50 and AP75 exactly at both scales. In paired shared-device timing, Top-2 median latency is 29.50 versus 28.77 ms for dynamic routing on Nano, and 32.84 versus 31.21 ms on Small. These small, shared-device timing samples do not establish a release speedup; they demonstrate why actual sparse dispatch still needs an end-to-end measurement. This bounded result limits the tested frozen final-layer adapter, not conditional computation throughout the detector.

Q What Unsuccessful Interventions Teach Us

We organize exploratory outcomes by the explanations they constrain, rather than by the order in which experiments were run. A completed comparison establishes a boundary for that intervention; a resource-admission failure establishes no scientific result. Differences from the original paper’s training, architecture, and evaluation are retained with the corresponding experiment instead of being absorbed into the original method’s credit.

More training or larger effective batches are not universal repairs. Matched learning-rate continuations help Small and Medium but fail to give the same benefit at the selected Nano endpoint. For Extra-Large, one completed 24-epoch effective-batch-sixteen continuation reaches a best AP50 of 74.9831, versus 74.9785 for the matched batch-two control; their final values are 74.9039 and 74.8716. The small, single-seed differences do not establish a practical batch advantage. Accumulation changes optimizer-update frequency and does not reproduce physical-batch normalization statistics. The registered normalization-clock and further continuation controls remain in progress. This result does not explain the remaining gap to the paper.

A successful state correction need not transfer.

An early two-buffer repair recovers a large AP deficit, whereas calibration of other checkpoints reduces accuracy. At a separate Nano endpoint, training-image calibration improves the same parent across three subsets, but the magnitudes differ. The defensible lesson is to specify and test state changes on their own parents. A calibration result cannot silently replace an uncalibrated EMA benchmark entry.

A plausible quality proxy need not identify a better box.

Fixed-box cross-view rescoring and uniform geometric representatives improve the registered full-set comparisons. Refinement stability and the tested local appearance repair do not provide a general improvement. In the selected local-repair example, an available candidate has IoU 0.756 while the chosen candidate has IoU 0.471. This demonstrates a selector failure in that example, not the prevalence of this mechanism. The later train-only visual readout outperforms metadata and shuffled-target controls on one YOLO checkpoint; it uses different evidence and supervision and is reported separately.

Internal activity and stability are insufficient endpoints.

In the short Nano/Small expert pilots, routes vary with inputs but replacing them by their mean changes AP50 by less than 0.01. This motivates testing output sensitivity and stronger static controls; it neither proves expert collapse nor rejects MoE in general. Source-anchored adaptation improves specified synthetic-noise streams but harms the clean streams in the completed loss study. A bound on parameter drift is compatible with this outcome because it does not imply a bound with a favorable sign on detection AP. No rollback benefit is established when the evaluated traces contain no rollback events.

Cheaper computation must retain an explicit quality comparison. Reduced-query inference loses

accuracy, and the completed image-budget and shallow-decoder pilots do not establish a better accuracy-cost point over their controls. CUDA Graph has a different claim: measured latency reduction with complete-set prediction equivalence. Dense-target backward probes support retention of all annotations, while convergence speed remains a separate unproven claim. Each outcome is useful for a community baseline without being promoted to an architectural discovery.

R Further Analysis with a Clear Falsification Test

The following analyses are planned, not completed empirical results. They extend the existing evidence before introducing additional detector modules.

How much apparent coverage is lost to competition?

Compute the matching bound in Appendix C on the complete frozen prediction sets, stratified by object size and scene density. Compare candidate-reuse coverage, one-to-one matching capacity, and actual matches. If the first two nearly coincide, candidate competition is not a major explanation for that stratum. If they differ, a score-only policy cannot close the coverage statistic's apparent gap. Annotation-assisted bounds must not be evaluated as deployable detectors.

Do ranking and geometric selection solve complementary errors?

Complete a two-by-two comparison on one frozen set of identities: original or support-based scores crossed with original or consensus coordinates. Groups must be built once; counts, classes, and suppression remain fixed. Report each cell, the interaction, and the same image-paired uncertainty. A large interaction would rule out adding isolated gains. Transfer this unchanged rule to a newly completed dataset before further tuning there.

When does a small state change become a large metric change?

For paired calibration and adaptation outputs, measure score-order changes, matching changes, and upward/downward IoU-threshold crossings. A small parameter displacement can cross a discrete decision boundary. These diagnostics test where metric sensitivity occurs; they do not turn a parameter-distance bound into an AP guarantee. Fresh detector seeds and real domain shifts are needed to test the generality of the observed patterns.

S Packaging the Positive Interventions

The community preview separates reference reproduction states from the states needed to reproduce an op-

timization comparison. It contains ten complete detector states: the four primary models, the calibrated Nano, four historical sources for the multi-view experiments, and the calibration parent. Two additional normalization-parameter snapshots record the endpoints of the positive synthetic-noise adaptation runs. They are tied to exact source digests. Shipping a historical source is necessary for reproducing its paired improvement; it does not replace a stronger current single-view reference.

A deployable improvement includes its evaluation procedure. The release provides multi-view prediction and complete-set evaluation, training-image calibration, online adaptation with separate offline scoring, and paired eager/graph benchmarking. Fixed rules and their matched controls are explicit. Parameter-free inference uses an unchanged detector together with a different inference procedure, so it does not require a fictitious new trained weight. Online adaptation is a trajectory: each prediction precedes that image’s update. Its complete-stream AP therefore cannot be assigned to the terminal parameter snapshot as a static model score.

Validation has multiple scopes. Tensor identity checks bind exported sources to their research states. Fresh, spread-index five-view inference checks test re-located execution; complete cached-output comparisons test the postprocessing rules. The calibration tool reproduces its released target state bitwise. A separately ported CPU scorer reproduces both original complete adaptation-stream metrics exactly. Short adaptation checks exercise real gradients, frozen buffers, and a dense pseudo-target backward pass. Later predictions in a fresh adaptation run show numerical and ordering differences after updates; we do not claim a bitwise-identical cross-run trajectory. A fresh complete portable-stream re-evaluation has completed all 11,738 images for each source. Relative to retained matched frozen-stream controls, Nano gains 0.726654 AP50 and 0.203875 AP75; Small gains 0.774967 and 0.392308. The original frozen controls were not rerun in this addendum. These are re-executions of the same trained sources, not independent detector seeds, and do not replace the original comparison table. Shared-device timing checks validate the executable benchmark, not a universal latency guarantee.

Selecting deployment recipes does not erase exploration. The runnable recipe set prioritizes interventions with measured benefit and retains their indispensable baseline controls. Unsuccessful exploratory method implementations and weights are excluded from the deployment set; their hypotheses, outcomes, and explanatory limits remain in Appendix Q and the supporting evidence. This packaging decision is not an estimate of the success rate of the research search and does not remove adaptive test-selection limitations.

T Reproduction Closeout and Release

The reproduction phase closes after registered feasible routes have been reviewed with explicit stopping evidence under the available public information and hardware budget. This is not a claim to exhaust every conceivable hyperparameter or prove a global optimum. A new material implementation discrepancy or a new public release can reopen a route. Independent end-to-end seeds, portable installation, weight provenance and licensing, complete curves, and protocol documentation are separate public-release requirements. The present manuscript is an internal evidence snapshot and has not been submitted or publicly released.