SOURCE GAINS, TARGET LOSSES: AUDITING GEOMETRY-DISTILLED VISUAL FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Geometric supervision promises visual representations that better capture scene structure, yet gains on source geometry tasks can hide a loss of boundary performance on other datasets. This creates a selection problem: an adapter can score better on the task it learns while becoming less useful for the task it is meant to support. We study this mismatch by adapting DINOv2 with depth and multi-view supervision from ScanNet++ and fitting new linear boundary probes on three labeled target datasets. With adapter capacity, training objective and budget held fixed, adaptation depth separates the outcomes: early-layer updates transfer better than late-layer updates in all three training runs, under five evaluation metrics and two dataset weightings. Source boundary F1 nevertheless favors a later layer group in every run, and the chosen adapters score below the unadapted backbone. We then test whether geometric learning can keep more of the backbone's boundary performance by penalizing changes to pretrained patch features. With a penalty weight chosen using target results, the mean target F1 loss falls from 0.015–0.020 in three unconstrained training runs to 0.0010 on average over three constrained runs, while source coplanarity gains remain. A preregistered test on COCO, a target set the project had never used, repeats the layer result: early windows beat both late windows and the unadapted backbone in all three runs, and source boundary F1 again picks a worse window. Adapting a backbone well therefore depends on where it learns geometry and on which pretrained features it keeps for transfer.

1 INTRODUCTION

Self-supervised vision transformers (Dosovitskiy et al., 2021) such as DINOv2 support dense prediction without task-specific labels (Oquab et al., 2023; Balestriero et al., 2023). Depth and multi-view supervision can tie these features more closely to scene structure (Hou et al., 2021; Hoyer et al., 2021). Object boundaries look like a natural beneficiary, because occlusions and depth changes often separate objects. The link is imperfect, however. A depth step can fall inside one object, and two adjacent objects can lie on one plane. Geometric training may therefore sharpen some boundary cues while weakening others that the pretrained features already carried.

This tension matters when a practitioner decides how to adapt a backbone. Before any target data are labeled, the only evidence at hand is performance on the geometric training tasks. A higher source score shows that an adapter learned geometry. The score says nothing about whether the adapter kept the features that a boundary predictor needs. We ask whether source scores still identify the adaptations that transfer best to boundary prediction.

To answer this question, we train low-rank adapters (LoRA) (Hu et al., 2021) on ScanNet++ RGB-D video (Yeshwanth et al., 2023) to predict four geometric relations while the DINOv2 weights stay frozen. We then freeze each adapted model and fit a new linear boundary probe on the labeled training split of SUN RGB-D, ScanNet and Hypersim. Linear probes are a standard test of what a representation makes accessible (Zhan et al., 2023). Because every model gets its own probe, each representation can learn its own mapping to the target labels, and a measured loss reflects the adapted features themselves. The unadapted backbone, probed in the same way, is the reference throughout.

Geometric learning and boundary transfer come apart in this test. Adapting all 24 blocks raises the source boundary and coplanarity scores but lowers mean target boundary F1 by 0.015 to 0.020 in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

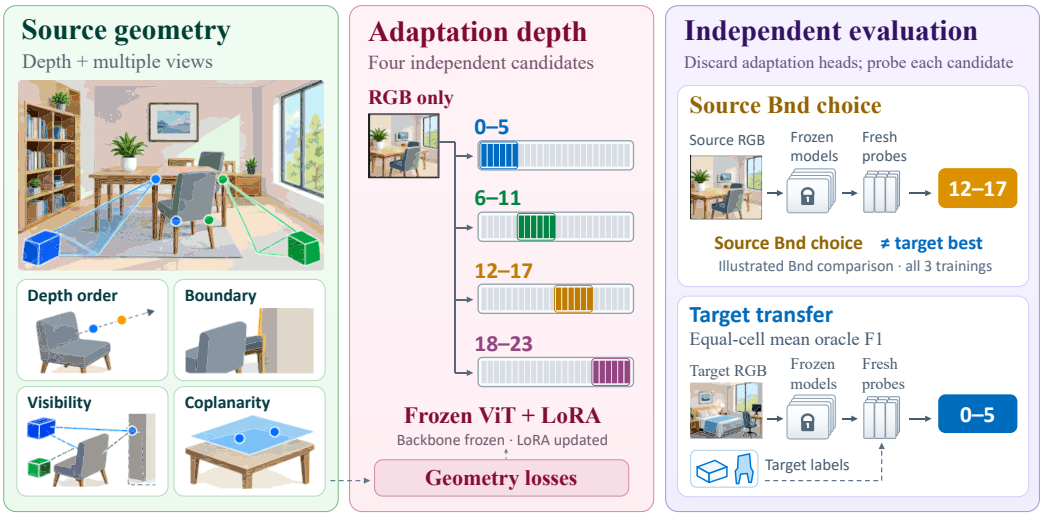

Figure 1: Study design. Left: depth and multiple views of ScanNet++ scenes supply four geometric relations between patches. Middle: each of four candidate adapters changes six consecutive blocks of a frozen DINOv2 ViT-L/14. Right: after training we discard the relation heads. A new source probe scores each candidate on ScanNet++ validation frames, and a new linear probe fitted on labeled target images measures boundary prediction. In all three training runs the source boundary score picks window 12–17, while window 0–5 has the highest mean target oracle F1. The scenes are schematic illustrations, not dataset images or model outputs.

three training runs. To locate the layers behind this loss, we compare four adapters of equal size, each confined to six consecutive blocks. We fixed the four windows and the two source selection rules before evaluating any window, and we repeated the whole train, score and select procedure three times. The two early windows beat the two late windows in every run, under all five target metrics and both dataset weightings. The source boundary score still picks window 12–17 each time. The source coplanarity score picks a near-best early window twice. In the third run it picks the deepest window, which falls 0.0203 below the unadapted backbone. A single run would have made coplanarity look like a reliable guide. A preregistered study on COCO, which the project had never used, repeats both findings on an untouched target set.

If the loss comes from changing features that the boundary probe relies on, limiting those changes should reduce it. We therefore add a penalty that keeps the adapted patch features close to the pretrained ones while the adapter learns the same geometry. Across three training runs the penalty leaves an average target loss of 0.0010, against 0.015 to 0.020 without it, and the coplanarity gain remains. In the run where we swept it, every weight from 0.3 to 3 removes the loss, and on COCO the weights chosen without COCO data keep boundary F1 at the backbone's level. Lowering the learning rate also shrinks the loss but gives up source boundary gain along the way. Two of the three preserved adapters beat the lowest tested learning rate on both source scores and target F1.

In practice, each candidate needs a comparison with the unadapted model as well as with the other candidates. A source-based choice needs repeating before anyone relies on it, and preserving pretrained features belongs in the training goal.

2 Related Work

Geometry in visual features. Geometry-aware pretraining can help dense tasks (Hou et al., 2021; Hoyer et al., 2021). Cross-view completion learns features for depth and flow by reconstructing a masked view from a second view of the same scene (Weinzaepfel et al., 2022). DUSt3R regresses 3D pointmaps from unposed image pairs (Wang et al., 2024), and dense depth predictors build directly on ViT features (Ranftl et al., 2021). Closest to our setting, FiT3D fine-tunes DINOv2 on ScanNet++ against multi-view-consistent rendered features, which improves linear-probe segmentation and depth (Yue et al., 2024). Probe3D tests frozen foundation-model features with probes for depth and surface normals and with zero-shot correspondence, and finds clear gaps in their 3D

awareness (El Banani et al., 2024). We share the ScanNet++ source and the probing approach, but we measure what geometric adaptation does to object boundaries on other datasets, and whether the source scores reveal it.

Probing and feature transfer. Linear classifier probes read out what intermediate layers encode (Alain & Bengio, 2017), and probes on DINO, DINOv2 and ViT layers measure what a representation makes accessible (Caron et al., 2021; Oquab et al., 2023; Raghu et al., 2021; Zhan et al., 2023). Transfer depends on the layer and the task (Yosinski et al., 2014; Azizpour et al., 2014; Zamir et al., 2018; Standley et al., 2020), training choices that raise ImageNet accuracy can make fixed features transfer worse (Kornblith et al., 2019), and fine-tuning can distort pretrained features (Kumar et al., 2022).

Which layers to adapt. Adapters (Houlsby et al., 2019), visual prompts (Jia et al., 2022), LoRA and its layer and rank variants (Hu et al., 2021; Zhang et al., 2023; Valipour et al., 2023) change a pretrained network at low cost. In surgical fine-tuning, the best layers to tune depend on the kind of distribution shift. The first block works best for image corruptions, and the last layer for shifts in the output (Lee et al., 2023). That work chooses layers with target data. We hold the geometric objective and the adapter size fixed, vary only the adapted layers, and ask whether source scores alone can make the choice.

Keeping pretrained behavior. Elastic weight consolidation penalizes changes to weights that mattered for an earlier task (Kirkpatrick et al., 2017), and L2-SP pulls fine-tuned weights toward their pretrained values (Li et al., 2018). Learning without Forgetting keeps a network's outputs on the old task while it learns a new one (Li & Hoiem, 2016), and WiSE-FT averages fine-tuned and pretrained weights to keep robustness under distribution shift (Wortsman et al., 2022). Our penalty acts on the final patch features instead, and we use it to test whether keeping them changes the balance between source gains and target losses.

Selecting models without target labels. Transferability estimators such as NCE, LEEP, LogME, H-score and PARC score a candidate with target labels or target features (Tran et al., 2019; Nguyen et al., 2020; You et al., 2021; Bao et al., 2019; Bolya et al., 2021), and Task2Vec embeds a task to pick a feature extractor (Achille et al., 2019). The ranking such metrics produce can flip under small changes to the evaluation setup (Agostinelli et al., 2022), a sensitivity we also find in source-based selection. Our rules use only source scores, which exist before any target labeling. Appendix D.4 reports LogME and H-score for the all-layer adapter.

Boundary labels. Boundary detectors usually learn from human contours (Arbeláez et al., 2011; Xie & Tu, 2015). Our target boundaries come from polygons, mesh reprojection or rendering (Song et al., 2015; Dai et al., 2017; Roberts et al., 2021), so we report each dataset separately.

3 Experimental Setup

3.1 Adapting the backbone to geometry

DINOv2 ViT-L/14 turns a 448×448 RGB image into 32×32 patch tokens. We keep its weights frozen and train rank-8 LoRA updates on the query, key and value projections of selected blocks. The update is scaled by $\alpha = 0.5$ directly, with no division by the rank. Training uses the ScanNet++ Stage-C split: 2,194 frames from 50 scenes, plus 224 validation frames from nine other scenes. Posed RGB-D frames supply four relations between patches: ordinal depth (Ord), depth and occlusion boundaries (Bnd), visibility across views (Vis) and coplanarity (Cop).

Each label is checked in a second view. Ord compares patch-center depths, Bnd marks depth edges where visibility flips, Vis counts the views in which a pixel stays visible, and Cop compares local plane fits (Appendix I.1). A two-layer head per relation reads the patch tokens. Ord and Cop use cross-entropy on 512 sampled patch pairs per frame, Bnd uses a focal loss ($\gamma = 2$, positive weight 10) and Vis a masked smooth-L_1 loss. The four losses are summed with equal weight.

Adapters train for five epochs with AdamW (learning rate 5×10^{-4}, weight decay 0.05, batch size eight, cosine schedule, two warmup epochs). We keep the checkpoint with the lowest epoch-mean training loss. The original all-layer checkpoint stores no optimization settings, so its settings come

from the training script. After training we discard the relation heads, which never serve as target predictors.

3.2 CANDIDATE ADAPTERS

ALL-4 adapts all 24 blocks. The main comparison uses four windows of six consecutive blocks each: 0–5, 6–11, 12–17 and 18–23. The windows share the rank, the adapted projections and the training recipe, so only their position differs. We recorded the windows, the two source selection rules and the target evaluation before evaluating the first window (Appendix F.1). The target datasets had entered earlier stages of the project. After the first result, we repeated the procedure with adapter seeds 43 and 44, giving three independent trainings per window. Window 18–23 is retrained in all three runs, and an older 18–23 checkpoint is reported only as a cross-check (Appendix F.1).

An earlier historical pool contains Early (blocks 0–7), 10–19 and ALL-4, each trained with adapter seeds 42–44. Earlier target experiments shaped this pool, so we use it as supporting evidence. The evidence thus has three levels. Window run 42 and the COCO study are preregistered. Runs 43 and 44, the penalty sweep and the matched ALL-4 runs are post hoc on the old target sets. The historical pool is supporting.

3.3 SOURCE SELECTION AND TARGET EVALUATION

Source scores. A new source probe reads each frozen candidate on the 224 ScanNet++ validation frames. The boundary rule ranks candidates by the pooled F1 of this probe at a fixed threshold (logit above zero), and the coplanarity rule by the pooled F1 of the two-class coplanarity probe. Each score averages three probe seeds and uses no target data. We report source gains separately, as mean per-image F1 differences against the unadapted backbone on the same frames. For each gain we count how many of 5,000 paired image resamples leave it non-positive. The frames come from only nine scenes, so this count understates scene-level uncertainty.

Target probes. An evaluation cell pairs one dataset or sensor with instance or semantic boundaries. The ten cells cover three SUN RGB-D sensors (kv1, kv2, xtion), Hypersim and ScanNet. A patch counts as a boundary patch when its fraction of boundary pixels exceeds a threshold fixed before any adapter was evaluated (Appendix A). For every model and cell, a new per-patch linear probe trains for 30 epochs with AdamW (learning rate 10^{-3}) and three probe seeds. Its positive-class weight follows the training prevalence of the cell (5.9 to 25.6). Under this recipe the unadapted backbone reaches a mean F1 of 0.4169. The historical pool and an earlier 14-cell set used a fixed positive weight of 16, under which the backbone reaches 0.4156. We never compare numbers across the two recipes. Paired comparisons use identical images, patches, labels and splits.

Target metrics. Oracle F1 is the pooled patch F1 at the best of 50 score quantiles, chosen for each model, probe seed and cell on the evaluation labels. It measures linear recoverability, the best a linear probe can do. Scene-held-out F1 is the deployable counterpart and the primary evidence for the preservation and COCO results. Four metrics complement oracle F1. Exact F1 searches every threshold. Scene-held-out F1 picks the threshold on one half of the scenes and scores the other half, over 20 random splits. Per-image AP averages precision over images with boundary pixels, and pooled AP computes it once per cell. We average cells in two ways: equally over the ten cells, or first within each dataset and then equally over SUN RGB-D, Hypersim and ScanNet.

Selection gap. Let $U_m(c)$ be the mean target score of candidate c under metric m. A source rule q picks $\hat{c}_q = \arg\max_c S_q(c)$, and its selection gap is $R_{q,m} = \max_c U_m(c) - U_m(\hat{c}_q)$ within the pool. We obtain 95% intervals by resampling test scenes within each dataset and sensor, applying the same resample to every candidate and probe seed, and recomputing the best candidate each time. The intervals hold adapters, probes and source choices fixed, so they describe variation over target scenes only.

Table 1: Windows picked by the source scores. For each training run and source rule, the picked window and its gap to the best window of that run. Intervals are 95% scene-resampling intervals for mean oracle F1 over the ten cells. Results for all five metrics are in Appendix F.

		ten-cell mean		datasets weighted equally	
run	rule	picked	gap [95%]	picked	gap
42	Bnd	12–17	0.0134 [0.0065, 0.0224]	12–17	0.0108
	Cop	6–11	0.0008 [0.0000, 0.0077]	6–11	0.0000
43	Bnd	12–17	0.0188 [0.0106, 0.0265]	12–17	0.0151
	Cop	6–11	0.0063 [0.0000, 0.0134]	6–11	0.0026
44	Bnd	12–17	0.0133 [0.0055, 0.0228]	12–17	0.0156
	Cop	**18–23**	**0.0249** [0.0156, 0.0353]	**18–23**	**0.0199**

4 RESULTS

4.1 GEOMETRIC LEARNING CAN REDUCE BOUNDARY TRANSFER

We first ask whether geometric learning helps boundary prediction when every layer adapts. The all-layer adapter improves both source tasks. New source probes gain 0.0322 boundary F1 and 0.0890 coplanarity F1 over the unadapted backbone, and neither gain turns non-positive in 5,000 paired image resamples (Appendix B). Target boundary prediction moves the other way. Mean oracle F1 over the ten cells falls by 0.0151, from 0.4169 to 0.4018, with a 95% interval of $[-0.0227, -0.0061]$ for this adapter. Two further independent ALL-4 trainings lose 0.0203 and 0.0197, and in held-out F1 the three lose 0.0173, 0.0212 and 0.0224, all with intervals below zero (Appendix M). F1 drops in nine cells, including all six SUN RGB-D cells, which lose 0.0155 to 0.0295. Only Hypersim instance boundaries improve, by 0.0048. An earlier 14-cell set agrees: 13 cells lose F1 (Table C5), and the mean change under scene-held-out thresholds is -0.0166 (Appendix H.3).

The loss does not depend on the oracle threshold: exact F1 falls by 0.0151 and pooled AP by 0.0210, while per-image AP moves less, by 0.0070 (Appendix G).

4.2 EARLY LAYERS TRANSFER BETTER, YET SOURCE SCORES FAVOR LATE LAYERS

The all-layer loss raises a narrower question: can some layers learn geometry without this cost? The four-window comparison answers it while holding adapter size and training fixed. In each of the three runs, both early windows beat both late windows in mean oracle F1 (Fig. 2b). Windows 0–5 and 6–11 gain 0.0027 to 0.0090 over the unadapted backbone, whereas windows 12–17 and 18–23 lose 0.0081 to 0.0238. The same ordering holds under all five metrics and both weightings. Within the early pair, 0–5 leads mean oracle F1 in every run, and 6–11 leads some of the other comparisons. Early windows also win within cells. In all 30 combinations of cell and run, an early window has the highest oracle F1, and both early windows beat both late windows in 27 of them (Fig. 3; margins in Appendix Q).

The source boundary score does not see this advantage. It picks window 12–17 in all three runs (Table 1). The choice misses 0.0134, 0.0188 and 0.0133 of mean target F1 relative to the best window, and all six intervals, three runs under two weightings, exclude zero. Across the five metrics, two weightings and three runs, all 30 gaps are positive, ranging from 0.0096 to 0.0232. The chosen window also falls below the unadapted backbone in all three runs, although two of these intervals include zero.

The coplanarity score does better, but only in two runs. It picks window 6–11 in the first two runs, with gaps of 0.0008 and 0.0063 whose intervals reach zero. In the third run it picks the deepest window, 18–23. That window trails the best by 0.0249 ($[0.0156, 0.0353]$) and scores 0.3966 against 0.4169 for the unadapted backbone, a loss of 0.0203 ($[-0.0288, -0.0098]$). The equal-cell gap stays above 0.022 under every metric, so in this run the rule picks a model worse than no adaptation.

The third coplanarity choice rests on small source differences: dropping one of the three source probes moves it to 0–5 (Appendix F.4).

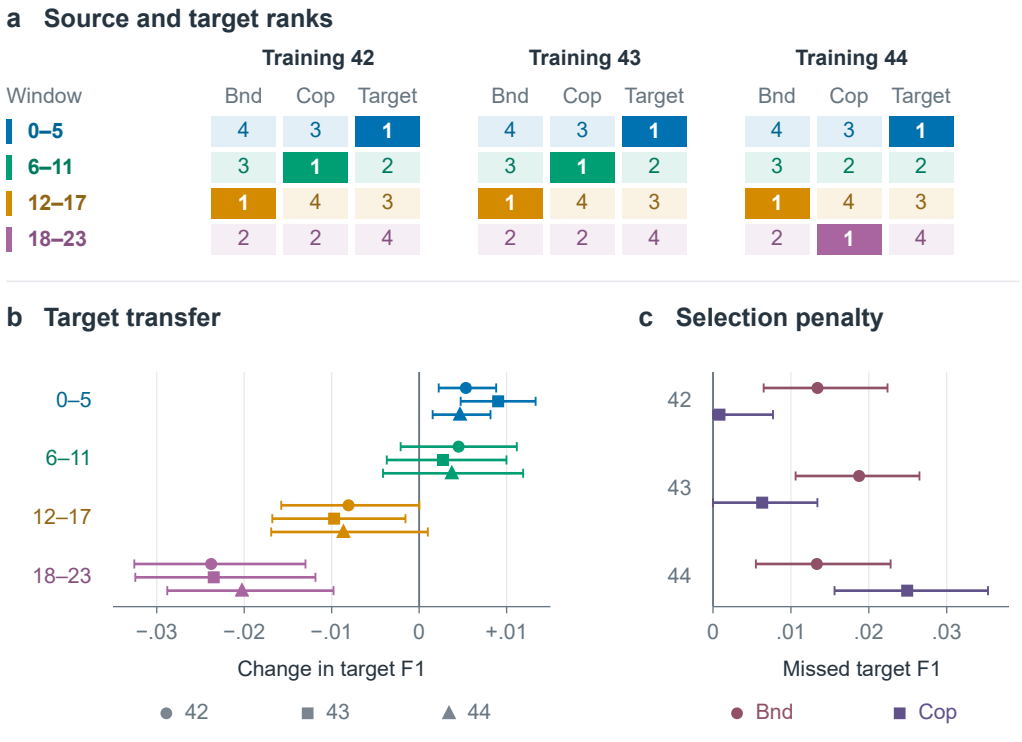

Figure 2: Source scores miss the windows that transfer best. (a) Ranks of the four windows by source boundary F1 (Bnd), source coplanarity F1 (Cop) and mean target oracle F1 in each training run; dark cells mark rank 1. (b) Mean change in target oracle F1 against the unadapted backbone for every window and run. (c) Gap between the window a source rule picks and the best window of that run. Every score averages three probes. Whiskers are 95% intervals from resampling target scenes, with the best window recomputed in each resample.

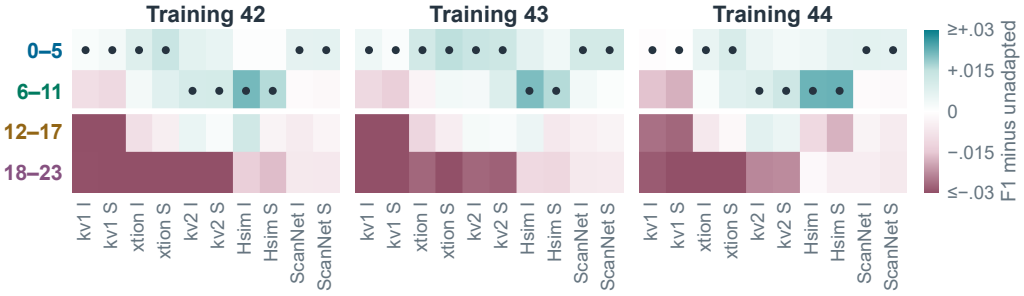

Figure 3: Early windows win cell by cell. Oracle F1 of each window minus the unadapted backbone in every target cell, for the three training runs; I and S denote instance and semantic boundaries. A dot marks the best window of each cell and run. Colors are clipped at ± 0.03; the largest loss is 0.044. Every entry averages three probe seeds.

The historical pool points the same way: Early (blocks 0–7) beats 10–19 and ALL-4 in all three runs despite the lowest source scores (Appendix H and Table K37).

4.3 Preserving pretrained features reduces the boundary loss

LoRA limits how many parameters change, but it does not limit how far the features move. We therefore add a penalty on the distance between the adapted and the pretrained patch features. Let F and F^0 be the adapted and the unadapted patch tokens after the final LayerNorm for a batch of

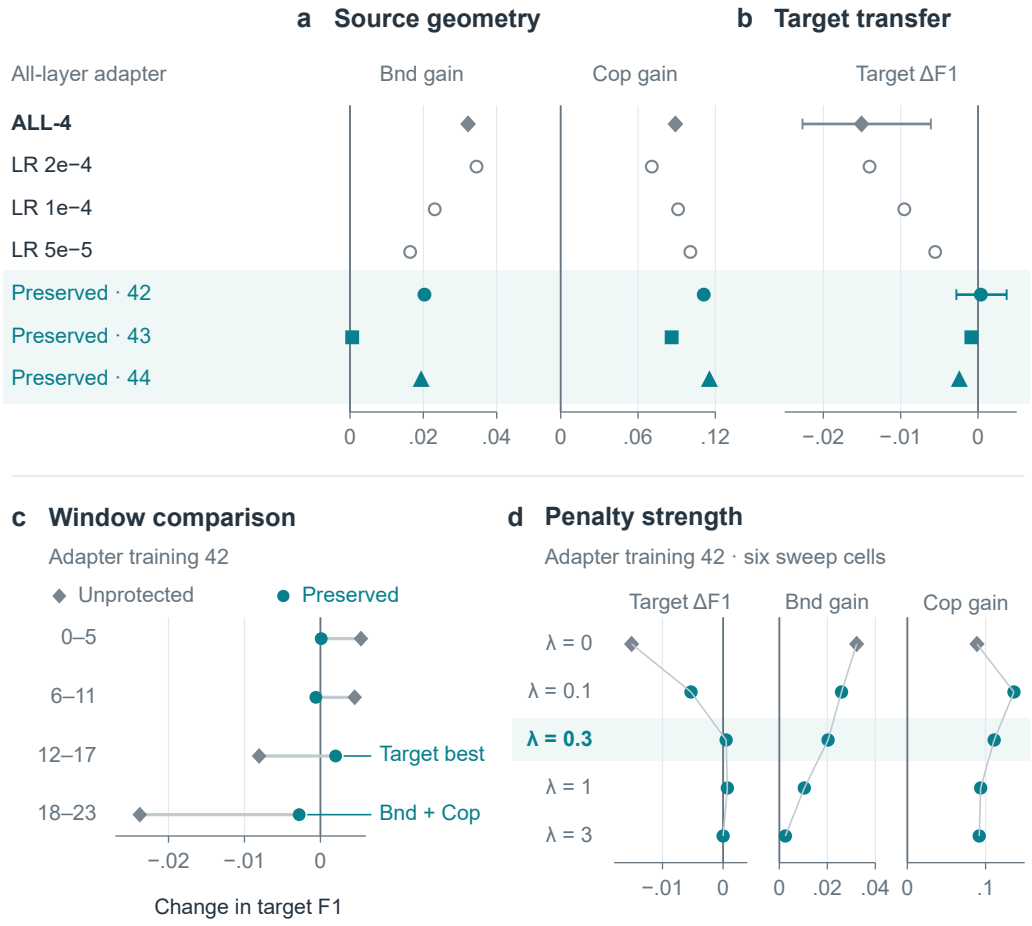

Figure 4: Feature preservation keeps geometric gains with a smaller boundary loss. All preserved models use $\lambda = 0.3$. (a) Source per-image F1 gains and (b) ten-cell change in target oracle F1 for the same all-layer adapters, one per row: the unprotected adapter (gray diamond), one training per lower learning rate (open circles) and three preserved trainings (teal). Whiskers are 95% scene-resampling intervals, available for the unprotected adapter and preserved run 42. (c) The four windows of run 42 with and without preservation; the labels mark the preserved pool's best window and the window both source rules pick. (d) Penalty sweep for adapter training 42: change in target F1 on the six sweep cells and source gains at each λ; the band marks the chosen weight. Every estimate averages three probes.

source images. The training loss becomes

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{geometry}} + \frac{\lambda}{BNd} \sum_{b,p,c} \left(F_{bpc} - F^0_{bpc}\right)^2,$$

where B is the batch size, N the number of patches and d the feature dimension. The frozen backbone computes F^0 without gradients. The penalty covers all patches, uses no target data and leaves adapter size unchanged (Appendix I.1). We chose $\lambda = 0.3$ from a sweep on six target instance cells with one adapter training, so this choice used target data. Section 4.4 tests weights chosen without it.

With the penalty, the three all-layer trainings change mean target F1 by $+0.0004$, -0.0009 and -0.0024 (mean -0.0010). The unprotected adapters of the same seeds change it by -0.0151, -0.0203 and -0.0197 (Fig. 4b). In held-out F1 each preserved adapter beats its same-seed unprotected adapter by 0.0174, 0.0197 and 0.0188, and in pooled AP by 0.0202, 0.0271 and 0.0203. All six intervals lie above zero, and they stay above zero with datasets weighted equally (Appendix M). The coplanarity gain survives in all three runs ($+0.1110$, $+0.0861$, $+0.1154$; no non-positive resample). The boundary gain survives in two runs ($+0.0203$ and $+0.0194$) and becomes uncertain in

Table 2: Feature preservation against lower learning rates. All-layer adapters with the same five-epoch budget. Source columns give per-image F1 gains; the target column gives the change in mean oracle F1 over the ten cells. Each row averages three probe seeds. Unprotected rows use adapter seed 42; preserved rows use learning rate 5×10^{-4}.

training setting	source Bnd gain	source Cop gain	target F1 change
learning rate 5×10^{-4} (original)	$+0.0322$	$+0.0890$	-0.0151
learning rate 2×10^{-4}	$+0.0345$	$+0.0707$	-0.0140
learning rate 1×10^{-4}	$+0.0231$	$+0.0911$	-0.0095
learning rate 5×10^{-5}	$+0.0164$	$+0.1006$	-0.0055
preservation, run 42	$+0.0203$	$+0.1110$	$+0.0004$
preservation, run 43	$+0.0006$	$+0.0861$	-0.0009
preservation, run 44	$+0.0194$	$+0.1154$	-0.0024

Table 3: COCO, a target set the project had never used. Held-out F1 differences per training run with 95% paired image-resampling intervals; instance and semantic cells weighted equally. The source boundary rule picks window 12–17 in every run.

comparison	run 42	run 43	run 44
early − late windows	$+.0081$ [.0063, .0098]	$+.0153$ [.0134, .0173]	$+.0115$ [.0095, .0134]
best window − 12–17	$+.0065$ [.0047, .0087]	$+.0161$ [.0139, .0184]	$+.0133$ [.0102, .0161]
ALL-4 − backbone	$-.0009$ [$-.0033$, .0012]	$-.0044$ [$-.0068$, $-.0021$]	$-.0040$ [$-.0066$, $-.0017$]
R2 (λ=1) − ALL-4	$+.0010$ [$-.0011$, .0035]	$+.0045$ [.0023, .0069]	$+.0043$ [.0021, .0069]
R1 (λ=0.1) − ALL-4	$+.0018$ [$-.0003$, .0038]	$+.0053$ [.0033, .0074]	$+.0051$ [.0033, .0072]

the third ($+0.0006$, non-positive in 24% of resamples). Over five metrics and two weightings, the three-run means lie between -0.0018 and $+0.0020$ (per cell: Table P46).

Preservation versus a smaller learning rate. A smaller learning rate also restrains adaptation, so we compare it with the penalty under the same five-epoch budget. Three unprotected ALL-4 adapters train at learning rates 2×10^{-4}, 10^{-4} and 5×10^{-5}, one training each (Table 2). Lower rates shrink the target loss, but the smallest rate still loses 0.0055. Preserved runs 42 and 44 beat this setting on source boundary gain, source coplanarity gain and target F1 at once. Run 43 keeps less boundary gain but loses less target F1 than every learning-rate setting. Per-image AP with equal dataset weights favors the smallest rate, and with one training per rate this control leaves the mechanism open (Appendix G.1).

Penalty strength. We then evaluated the other weights of the sweep on all ten cells (run 42). Against the backbone, held-out F1 changes by -0.0173, -0.0085, $+0.0001$, $+0.0008$ and -0.0002 at $\lambda = 0$, 0.1, 0.3, 1 and 3 (Appendix N). Any weight from 0.3 to 3 removes the loss, so the result does not hinge on 0.3, and $\lambda = 0.1$ halves it. The source boundary gain shrinks with the penalty, from $+0.0259$ to $+0.0025$, whereas the coplanarity gain stays above 0.09 for every $\lambda > 0$ (Fig. 4d).

Layer choice under preservation. Retraining the four windows with $\lambda = 0.3$, one training each, shrinks the spread of target means from 0.0291 to 0.0048, and both source rules then pick 18–23, the lowest window (Fig. 4c, Appendix F.6).

4.4 A PREREGISTERED TEST ON AN UNTOUCHED TARGET SET

The target datasets above shaped earlier stages of the project, and λ was chosen on them. We therefore ran one preregistered study on COCO-2017, which the project had never used (Appendix L). Before any COCO result, we fixed two rules that need no COCO data. R1 maximizes the sum of the two source gains and picks $\lambda = 0.1$. R2 maximizes held-out F1 on the old ten cells and picks $\lambda = 1$. Boundary labels follow the SUN RGB-D rules, and all 28 models (every weight with three runs) were probed on 3,018 training and 754 test images. Held-out F1 is the primary metric, with intervals from 1,000 paired image resamples.

The layer result transfers in full (Table 3). Early windows beat late windows by 0.008 to 0.015 in every run, which meets the preregistered criterion. Both early windows also beat the unadapted backbone, by 0.0075 to 0.0157, with all six intervals above zero. The source boundary rule again picks 12–17, which trails the best window in every run. ALL-4 again costs boundary F1, with intervals below zero in runs 43 and 44. Both weights chosen without COCO data recover that loss, with intervals above zero in both runs, and every preserved model stays within -0.0001 to $+0.0012$ of the backbone. Run 42's ALL-4 loses almost nothing on COCO, so its interval includes zero and the preregistered recovery test is not met in all three runs (Appendix L).

4.5 Checks on the source score and the target probe

Source scores without a fixed threshold. The boundary rule scores F1 at a fixed threshold, which might cause the miss. A post hoc rerun scored the source boundary probes of all twelve window adapters in four ways. Two are threshold-free (average precision and best F1 over all thresholds), one uses 50 quantiles, and one is the fixed threshold. Every scoring picks window 12–17 in all three runs, and resampling the nine validation scenes keeps that pick in 81% to 100% of draws (Appendix O). Removing the threshold does not remove the miss.

A more flexible target probe. A two-layer MLP probe on the 12 cells of the historical protocol narrows the all-layer gap in 11 cells, and the SUN kv1 cells stay negative (Appendix D.3). A probe trained on unadapted features and applied to adapted features lowers boundary-patch scores by 0.16 to 0.27 on the SUN sensors (Table K40). A more flexible probe recovers part of the loss, but neither check identifies what causes the rest.

5 Discussion and Conclusion

Geometric learning and boundary transfer respond differently to adaptation. When all layers adapt, the source tasks improve while target boundary F1 drops, even though every model receives a new target probe. The equal-capacity windows place this difference in depth. Early windows score above the unadapted backbone and late windows below it, in every run and under every metric we tested. The source boundary score favors the late layers each time. The layer ordering holds on COCO, a target set we fixed before any COCO result, where early adaptation also beats the unadapted backbone. There the best early window gains 0.008 to 0.016 over the backbone, so adapting the early layers to geometry helps boundary transfer outright.

Source-based selection can fail in two ways. A rule can miss a better candidate in the pool, or it can pick a model worse than no adaptation. Boundary scoring shows the first failure in all three runs, and coplanarity scoring shows the second in one run. Comparing only candidates would hide the second failure, and a single run would have hidden the instability of the coplanarity rule. Repeating the train, score and select procedure exposed both. Both echo earlier findings that the right layers depend on the shift (Lee et al., 2023) and that transferability rankings move with the setup (Agostinelli et al., 2022).

Feature preservation changes the trade-off between source and target. It keeps the coplanarity gain and part of the boundary gain while cutting the average target loss. Two preserved adapters also beat the lowest tested learning rate on both source scores and target F1, so restraining optimization and constraining the features act differently. The MLP probe adds a caution: a weaker linear readout does not mean the information has disappeared. A new linear probe still leaves the loss in place, and preservation reduces it. The remaining SUN kv1 losses mark the boundaries that are hardest to keep across sensors.

These conclusions cover boundary prediction from DINOv2 features adapted on ScanNet++, measured with probes trained on labeled target data. The old target sets had been used earlier in the project. COCO had not, but COCO is one additional dataset, used with landscape images only. We chose the paper's penalty weight with target feedback, and the COCO study shows that weights chosen without it also remove the loss. The preserved four-window comparison has one training per window. Same-capacity adaptation without geometric supervision and full target decoders remain untested.

REPRODUCIBILITY STATEMENT

Every number in the paper comes from saved predictions, probes or training logs. The appendix names the saved result files behind several tables, the identity checks that paired predictions had to pass, and both preregistration documents with their registered deviations (Appendices F.1 and L). The training objective, the feature-preservation term and the LoRA scaling are defined in Appendix I.1, and the data splits and label thresholds in Appendix A. Historical tables whose generating code did not survive a storage migration are marked as such.

AI USE STATEMENT

The authors conceived the research question and ideas, designed the experiments, decided every claim and led the writing of the paper. Generative AI tools served as research assistants. They helped implement and debug code, run and audit analyses, and draft and edit text under the authors' direction, and the authors reviewed and revised all such material. The authors drew Figure 1 by hand. The other figures and all tables are produced from saved predictions and logs, and every numerical claim traces to saved predictions, logs or an explicitly identified historical report. The authors take full responsibility for the submission.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6429–6438, 2019.

Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 303–321, 2022.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017.

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *arXiv preprint arXiv:1406.5774*, 2014.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari S. Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas J. Guibas. An information-theoretic approach to transferability in task transfer learning. In *IEEE International Conference on Image Processing (ICIP)*, pp. 2309–2313, 2019.

Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19301–19312, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799, 2019.

Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 709–727, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2019.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *International Conference on Learning Representations*, 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *Proceedings of the 14th European Conference on Computer Vision*, pp. 614–629, 2016.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2825–2834, 2018.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cédric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning (ICML)*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10912–10922, 2021.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, pp. 746–760. Springer, 2012.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 567–576, 2015.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9120–9132, 2020.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1395–1405, 2019.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 3274–3287, 2023.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3502–3516, 2022.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1395–1403, 2015.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning (ICML)*, 2021.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D feature representations by 3D-aware fine-tuning. In *Proceedings of the European Conference on Computer Vision*, pp. 57–74, 2024.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3712–3722, 2018.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. *arXiv preprint arXiv:2310.06836*, 2023.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

SUPPLEMENTARY MATERIAL

This supplement is a claim-audit companion to the main paper. The main text keeps the evidence needed to decide the three claims; details that would obscure the 9-page story are placed here. The table below shows how the supplement should be read: appendix results refine or bound the main claims, but they do not create broader claims than those stated in the paper.

Claim-to-evidence map.

claim	main-paper evidence	supplementary evidence and scope
All-layer geometric adaptation lowers target boundary F1	Section 4.1	Appendix B gives the source gains; Table G28 gives five metrics; Table C5 gives the 14-cell set; Appendix M gives three unprotected trainings.
Early windows transfer better, and source scores pick late windows	Section 4.2, Fig. 2, Table 1	Appendix F gives all metrics for twelve adapters and the cellwise record; Appendix F.4 the source-probe sensitivity; Appendix H the historical pool. Target datasets had entered earlier work.
Feature preservation keeps geometric gains with a smaller boundary loss	Section 4.3, Fig. 4, Table 2	Appendix G gives the seed replication, the per-cell table and the penalty sweep (Fig. G2); Appendix G.1 the learning-rate controls. The penalty weight was chosen with target feedback.

The remaining appendices follow this map. Appendix A gives datasets, boundary labels, preprocessing, and the post-audit construction of each cell. Appendices B to C provide source recoverability, geometry-alignment, and full trend tables. Appendix D expands the mechanism, transferability-estimator, calibration, loss-variant, non-boundary controls, and non-SUN multi-head mechanism evidence. Appendix E gives placement, bottleneck-family, cross-backbone, and depth-only extension results. Appendices I and J list the reproducibility package and the negative-results/self-correction log.

A DATASETS, BOUNDARY LABELS, AND PREPROCESSING

All probes operate on a 32×32 grid of patch tokens from a 448×448 letterboxed input. For each frame we build pixel-level boundary maps, mask invalid pixels, aggregate to patches, and threshold. We use identical code for all datasets, swapping only the per-dataset loader, so that the geometry alignment of Eq. (1) is measured under exactly the labels the probe consumes.

A.1 BOUNDARY LABEL DEFINITIONS

Let a frame provide a depth map D (metres), a semantic label map S, and an instance map I, each letterboxed (NEAREST for labels, BILINEAR for depth) with padded rows recorded in a validity mask. Pixel boundaries are:

- **Depth:** a pixel is a depth boundary if the local depth gradient magnitude exceeds τ_{depth} metres; invalid/zero-depth pixels are masked.
- **Semantic:** a pixel is a semantic boundary if any 4-neighbour has a different value in S.
- **Instance:** a pixel is an instance boundary if any 4-neighbour has a different value in I (a label change within the same semantic class still counts).

Pixel maps are aggregated to a patch by the mean indicator over the patch; a patch is positive if that fraction exceeds τ_{patch}, and valid if its valid-pixel fraction exceeds 0.5. We separate the depth-gradient threshold τ_{depth} (metres) from the patch fraction threshold τ_{patch} (unitless); conflating them was the cause of an early confounded pilot (Appendix J).

A.2 PER-DATASET SETTINGS

Table A1 summarizes the resources, splits, thresholds, and alignment span used by the main board. Real-sensor datasets use $\tau_{\text{patch}} = 0.10$ for instance/semantic; the rendered Hypersim uses $\tau_{\text{patch}} = 0.30$ to match real-sensor target positive rates (sharp synthetic edges otherwise inflate the positive rate). Depth-boundary $\tau_{\text{depth}} = 0.10\,\text{m}$ throughout. Target positive rate is a separate label statistic used for probe/loss audits; it is not the alignment IoU.

Table A1: Datasets, splits, and boundary-label settings. The final column reports the main-board target–depth IoU span over instance/semantic cells at the listed settings, not a positive rate. "annot." is the label-generation process, a second axis modulating transfer (Appendix B).

dataset	role	annot.	train/test	τ_{patch}	IoU span
NYUv2 (Silberman et al., 2012)	low-align real (anchor)	polygon	795/654	0.10/0.20	0.043–0.053
SUN RGB-D (Song et al., 2015)	4-sensor gradient	polygon	80/20 per coll.	0.10	0.035–0.114
Hypersim (Roberts et al., 2021)	high-align synthetic	render	2900/875	0.30	0.148–0.172
ScanNet-v2 (Dai et al., 2017)	high-IoU mesh-derived	mesh	1591/283	0.10	0.206–0.208
DIODE (Vasiljevic et al., 2019)	outdoor depth-only ext.	lidar	—	0.20	—

A.3 SUN RGB-D: FOUR-SENSOR STRATIFICATION (CRITICAL)

SUN RGB-D (Song et al., 2015) mixes four depth sensors (Structure Sensor / "kv1", Kinect v2 / "kv2", Asus Xtion / "xtion", Intel RealSense) across ten sub-collections. Because the sensors differ in resolution and noise, an unstratified split would let one sensor dominate the positive rate and IoU. We therefore (i) sample 80 scenes per sub-collection by `np.linspace` (kv1/kv2/xtion $= 160$ scenes each, realsense $= 261$, total 741), (ii) decode depth with the SUN RGB-D bit-shift convention $d \leftarrow ((d \gg 3) \mid (d \ll 13))/1000$ (metres), and (iii) take a scene-disjoint $80/20$ split *per sensor*. Treating each sensor as its own cell is what yields the within-dataset alignment gradient of Appendix B; per-sensor IoUs are kv1 0.035, realsense 0.072, xtion 0.074, kv2 0.114 (instance), confirming the gradient holds under one acquisition protocol.

A.4 SCANNET-V2 SUBSET AND THE MESH-DERIVED CAVEAT

We use a 100-scene / 1874-frame subset with three-way frame-id intersection of `depth`, `label-filt`, and `instance-filt`, plus the matching `color` frames (so the frozen-vs-adapter comparison is not contaminated by out-of-distribution gray-scale inputs; see Appendix J). Label/instance maps (968×1296) are NEAREST-resized onto the depth plane (480×640); depth is mm$/1000$. Scene-disjoint $80/20$ split ($1591/283$ frames). ScanNet's 2D labels are reprojected from a reconstructed 3D mesh (Dai et al., 2017); consequently label boundaries partly coincide with depth discontinuities *by construction*, which we quantify as constructive IoU inflation in Appendix B and which underlies the annotation-process axis of Appendix B.

A.5 HYPERSIM MANIFEST AND A CORRECTED PILOT

Hypersim (Roberts et al., 2021) provides exact per-pixel depth, semantic, and instance ground truth from physically based rendering. We use a stratified manifest of 2900 train / 875 test frames. Two granularities matter and are named separately throughout: a *sub-collection* is a Hypersim batch (`ai_VVV`) and a *scene* is one rendered environment within it (`ai_VVV_NNN`). The manifest draws from 31 train and 26 test sub-collections, avoiding the sampling collapse of an initial pilot that drew all training frames from two (Appendix J), and from 58 train and 35 test scenes. The split is scene-disjoint (no scene appears on both sides) but not sub-collection-disjoint: 18 sub-collections contribute different scenes to both sides. The 35 test scenes are the resampling units of the scene-paired bootstrap. Instance and semantic use $\tau_{\text{patch}} = 0.30$; the resulting per-patch test positive rates under the unadapted backbone are 0.0504 (instance) and 0.0383 (semantic). Depth uses a five-value τ_{depth} sweep $\{0.02, 0.05, 0.10, 0.20, 0.50\}$ with direction stable across all five (Appendix C).

Table B2: Source-domain recoverability (Δ vs. frozen, 3-seed paired image bootstrap, N=224, 5,000 resamples; "0/5000" means no resample left the gain non-positive). Bnd is mean per-image boundary F1 at logit>0; Cop is mean per-image argmax F1; Ord accuracy and Vis RMSE are near-neutral. MID-LoRA keeps the boundary gain and trades off Cop, consistent with late layers being where Cop is learned.

adapter	ΔBnd-F1	ΔCop-F1	ΔOrd-acc	ΔVis-RMSE
ALL-4	$+0.0322$ (0/5000)	$+0.0890$ (0/5000)	-0.012 (n.s.)	$+0.003$ (n.s.)
MID-LoRA	$+0.0344$ (0/5000)	-0.0169 (p=0.998)	-0.005 (n.s.)	$+0.003$ (n.s.)

Table B3: Dataset-level geometry alignment (test splits). IoU per Eq. (1); lift is base-rate-corrected. ScanNet's high IoU is partly constructive (mesh-derived labels borrow depth geometry), quantified in Appendix B.3.

cell	instance IoU	semantic IoU	lift (inst)
SUN-RGBD/kv1	0.035	0.035	—
NYU	0.053	0.043	1.78
SUN-RGBD/realsense	0.072	0.068	—
SUN-RGBD/xtion	0.074	0.073	—
SUN-RGBD/kv2	0.114	0.108	—
Hypersim	0.172	0.148	2.72
ScanNet-v2	0.208	0.206	—

B Source-Domain Recoverability and Geometry-Alignment Analysis

Alignment definition. For target boundary labels B_t and depth edges B_d on the jointly valid patches V of a cell,

$$\mathrm{IoU}(t,d) = \frac{\sum_{p\in V} B_t[p]B_d[p]}{\sum_{p\in V} \mathbf{1}[B_t[p]=1 \ \vee \ B_d[p]=1]}. \tag{1}$$

This is a property of the target labels and the depth maps, computed without any model prediction, so it is constant across adapters within a cell.

B.1 Source-domain recoverability table

Table B2 reports the source-domain linear-probe recoverability of the two adapters relative to frozen, the "recoverability signal" of Section 4.1. Both ALL-4 and MID-LoRA raise fixed-decision boundary F1 significantly; MID-LoRA preserves the boundary gain while trading off co-planarity, so the layer-selective variant is not a uniformly weaker adapter.

B.2 Dataset-level alignment and base-rate-corrected lift

Table B3 gives $\mathrm{IoU}(t, \text{depth})$ and the base-rate-corrected lift $P(\text{depth} \mid t)/P(\text{depth})$ on test splits. Hypersim and ScanNet sit at high IoU; the four real-sensor SUN RGB-D cells plus NYU span the low-to-mid range. The lift column shows the alignment ordering is not an artifact of differing depth-edge density.

B.3 Base-rate-matched sanity check

A natural objection is that Hypersim/ScanNet simply have denser depth boundaries, mechanically raising IoU. We control for this by sweeping the depth threshold to match depth positive rates across datasets and recomputing IoU and lift. Across a τ sweep that equalizes depth positive rate at $\approx 5\%/10\%/20\%$, NYU instance IoU stays in 0.031–0.078 (lift 1.05–1.46) while Hypersim stays in 0.114–0.165 (lift 2.59–2.85). The $\sim$ 2–3$\times$ alignment gap therefore survives base-rate matching and is intrinsic to the labels, not a depth-edge-density artifact. The full seven-row-per-dataset τ-matched table is provided with the released analysis JSON.

B.4 Rank-order stability of alignment under preprocessing

Because $\mathrm{IoU}(t, \mathrm{depth})$ is computed in the probe's patch space, a fair concern (raised in review) is whether the *ordering* of datasets by alignment — which is all the rank-based trend relies on — is stable under the patch size and threshold choices, rather than an artifact of a particular setting. Three pieces of evidence indicate it is. (i) *Depth threshold.* The base-rate-matched sweep above varies τ_{depth} over a range that changes the depth positive rate by $4\times$ ($\approx 5\%$ to 20%); across the entire sweep the NYU and Hypersim IoU ranges do not overlap (0.031–0.078 vs. 0.114–0.165), so the low-vs-high ordering is preserved at every depth threshold. (ii) *Patch threshold.* The instance/semantic patch threshold τ_{patch} is deliberately set per-pipeline to equalize positive rates (0.10 real, 0.30 rendered; Appendix A); the within-SUN-RGBD sensor ordering (kv1<realsense<xtion<kv2) is computed at a *single, shared* $\tau_{\text{patch}} = 0.10$, so that ordering — the cleanest part of the alignment trend — carries no cross-pipeline threshold confound at all. (iii) *Patch size.* Every measurement uses the fixed 32×32 DINOv2 patch grid, and we did not vary it. Whether a coarser grid preserves the dataset ordering is therefore untested here; the lift statistic normalizes by depth-boundary density, which removes one but not all of the grid's influence. We report the dataset *ranking* by alignment at this grid as the quantity the analysis uses, and the absolute IoU values as setting-dependent; the alignment trend (Section 4.1) and all Spearman correlations depend only on the ranking. A stricter evaluation-unit audit gives the same bounded conclusion: collapsing to four dataset means still leaves $\rho(\mathrm{IoU}, \Delta\mathrm{BF1}) = 0.60$, and keeping only one target type leaves instance-only $\rho = 0.79$ and semantic-only $\rho = 0.68$. The main weakening appears only in SUN-only or polygon-only subsets, which is why the main paper describes alignment as a benchmark-level descriptive axis rather than a portable target-only predictive law.

B.5 Approximating alignment from monocular predicted depth

The main experiments compute $\mathrm{IoU}(t, \mathrm{depth})$ from ground-truth RGB-D depth. As a deployment-realism check, we also ran a monocular proxy-depth audit: sensor depth was replaced by DA3/DPT-Large relative depth, and the predicted edge threshold was matched to the ground-truth depth-edge base rate in each cell. This audit is intentionally secondary because it mixes the alignment coordinate with monocular-depth error. On the full 14-cell board, mean-image GT-depth alignment correlates with ΔBF1 at $\rho_{\mathrm{s}} = 0.648$ ($p = 0.012$), while predicted-depth alignment gives $\rho_{\mathrm{s}} = 0.534$ ($p = 0.049$). These numbers should not be compared directly with the main global patch IoUs in Tables B3 and C4: this proxy-depth audit uses image-level IoUs and base-rate-matched predicted depth edges. Its role is only to show that a depth proxy can preserve part of the target-side ordering; it is not used in the paper's main evidence.

B.6 Image-level correlation

Within a dataset, do images whose target boundaries are more depth-aligned show larger gains? On Hypersim the per-image Spearman correlation between IoU and ΔpiAUPRC is positive and significant for both instance ($\rho = +0.106$, $p = 0.003$) and semantic ($\rho = +0.082$, $p = 0.025$). On NYU it is null ($\rho = -0.04$, $p \approx 0.3$), which we attribute to NYU's uniformly low IoU (mean 0.04) offering no within-dataset dynamic range to detect a trend; we therefore restrict the image-level claim to "within Hypersim, higher alignment is associated with larger gain," and rely on the cross-dataset board and within-SUN-RGBD sensor control for the main benchmark-level diagnostic claim. This is a deliberately conservative reading: the image-level effect is real but small and we do not over-state it.

C Complete Alignment-Trend Tables

C.1 Full 14-cell table, all four estimands

Table C4 lists the 14 cells. Table C5 adds the frozen baselines and the per-seed standard deviations underlying the paired bootstrap, so the reader can verify that the verdicts are not seed noise (per-seed std of Δ best-F1 is 0.001–0.004 across cells).

Table C4: Geometry-alignment board (14 cells, ascending IoU; ALL-4 adapter vs. frozen, 3-seed paired). Negative transfer is reproduced on *four* low-alignment real-sensor configurations (SUN RGB-D kv1, NYU, realsense, xtion), not a single dataset. Best-F1 transfer is generally less negative at higher alignment; the only positive best-F1 flip is the highest-alignment rendered cell. Mesh-derived ScanNet sits at the top of the IoU range yet is near-neutral, revealing a second, annotation-process axis (Appendix B). Δthr is the shift of the best-F1 threshold, a signature of the class-conditional margin alteration analyzed in Section 4.5.

dataset	target	IoU	annot.	ΔBF1	ΔpiAUPRC	ΔgAUPRC	Δthr
SUN/kv1	instance	0.035	polygon	−0.033	−0.014	−0.028	−0.238
SUN/kv1	semantic	0.035	polygon	−0.030	−0.015	−0.028	−0.218
NYU	semantic	0.043	polygon	−0.017	−0.009	−0.027	−0.058
NYU	instance	0.053	polygon	−0.015	−0.009	−0.024	−0.000
SUN/realsense	instance	0.072	polygon	−0.012	−0.005	−0.021	−0.055
SUN/realsense	semantic	0.072	polygon	−0.016	−0.010	−0.026	−0.104
SUN/xtion	instance	0.074	polygon	−0.027	−0.017	−0.024	−0.150
SUN/xtion	semantic	0.074	polygon	−0.022	−0.016	−0.021	+0.054
SUN/kv2	instance	0.114	polygon	−0.015	−0.016	−0.046	−0.165
SUN/kv2	semantic	0.114	polygon	−0.018	−0.015	−0.048	−0.301
Hypersim	semantic	0.148	render	−0.002	+0.012	−0.009	−0.330
Hypersim	instance	0.172	render	+0.005	+0.012	−0.003	−0.270
ScanNet-v2	semantic	0.206	mesh	−0.005	+0.003	−0.002	−0.076
ScanNet-v2	instance	0.208	mesh	−0.008	+0.001	−0.002	−0.055

Table C5: Full alignment-trend table with frozen baselines and 3-seed std. BF1_f is frozen best-F1; piAU_f is frozen per-image AUPRC. $\pm$ values on ΔBF1 are the sample standard deviation (ddof= 1) of the seed-paired Δ over probe seeds 42/43/44. Both BF1 and per-image AP baselines, deltas, and paired standard deviations are recomputed from the identity-bearing prediction files on September 11; zero-positive images are excluded from per-image AP.

cell	target	IoU	BF1_f	ΔBF1	piAU_f	ΔpiAUPRC
SUN/kv1	inst	0.035	0.348	−0.032±0.003	0.311	−0.014±0.001
SUN/kv1	sem	0.035	0.345	−0.030±0.001	0.309	−0.015±0.001
NYU	sem	0.043	0.348	−0.017±0.002	0.300	−0.009±0.000
NYU	inst	0.053	0.353	−0.015±0.001	0.300	−0.009±0.000
SUN/realsense	inst	0.072	0.279	−0.012±0.002	0.213	−0.005±0.003
SUN/realsense	sem	0.072	0.268	−0.016±0.001	0.215	−0.010±0.005
SUN/xtion	inst	0.074	0.303	−0.027±0.002	0.220	−0.017±0.000
SUN/xtion	sem	0.074	0.294	−0.022±0.003	0.215	−0.016±0.001
SUN/kv2	inst	0.114	0.408	−0.015±0.001	0.321	−0.016±0.001
SUN/kv2	sem	0.114	0.406	−0.018±0.002	0.318	−0.015±0.002
Hypersim	sem	0.148	0.485	−0.002±0.003	0.437	+0.012±0.002
Hypersim	inst	0.172	0.533	+0.005±0.003	0.492	+0.012±0.002
ScanNet	sem	0.206	0.509	−0.005±0.004	0.469	+0.003±0.001
ScanNet	inst	0.208	0.525	−0.008±0.001	0.482	+0.001±0.001

C.2 Spearman correlations

Over the 14 cells, the rank correlations are

$$\rho_s(\text{IoU}, \Delta\text{BF1}) = 0.74 \ (p = 2{\times}10^{-3}),$$
$$\rho_s(\text{IoU}, \Delta\text{gAUPRC}) = 0.61, \qquad \rho_s(\text{IoU}, \Delta\text{piAUPRC}) = 0.40 \ (p = 0.16).$$

On the 12-cell subset used for the estimand decomposition (Appendix D), $\rho_s(\text{IoU}, \Delta\text{BF1}) = 0.80$ ($p = 0.002$) and $\rho_s(\text{IoU}, \Delta\text{piAUPRC}) = 0.45$ ($p = 0.14$). We report all three estimands and both cell sets rather than selecting the most favorable; the best-F1 trend is significant and the AUPRC trend is same-signed but not, a discrepancy the mechanism explains (Section 4.5).

C.3 Dependence-aware alignment audit

The 14 cells are not independent samples: instance/semantic targets share images within each dataset or sensor. We therefore repeated the rank analysis with grouped units. Collapsing each target pair to

seven sensor-family units preserves the best-F1 trend ($\rho_s = 0.68$), and leaving out any dataset keeps the correlation positive (range 0.49–0.79). A group bootstrap over the seven sensor-family units gives median $\rho_s = 0.71$ with a wide 95% CI $[0.01, 0.96]$ and $P(\rho \leq 0) = 0.023$; bootstrapping only four dataset groups is too coarse and has a CI crossing zero. We therefore interpret alignment as a robust descriptive trend in this benchmark, not as an independent-cell causal law.

Table C6: Dependence-aware Spearman audit for the alignment board.

analysis unit	BF1 ρ_s	piAP ρ_s	gAP ρ_s
raw 14 cells	0.74	0.40	0.61
collapsed sensor-family (7 units)	0.68	0.43	0.64
collapsed dataset (4 units)	0.60	0.60	0.80
leave-one-dataset range, ΔBF1	0.49–0.79		
leave-one-sensor-family range, ΔBF1	0.58–0.82		
sensor-family group bootstrap, ΔBF1	median 0.71, 95% CI $[0.01, 0.96]$		

C.4 HYPERSIM DEPTH τ-SWEEP

On Hypersim depth, the ALL-4-vs-frozen per-image AUPRC Δ is positive and significant across all five depth thresholds: $\tau = 0.02$ ($+0.0086$), 0.05 ($+0.0095$), 0.10 ($+0.0078$), 0.20 ($+0.0071$), 0.50 ($+0.0141$), all $p < 0.005$. Direction is stable across the sweep, so the positive depth transfer on the highest-alignment dataset is not an artifact of a single threshold choice.

D MECHANISM: EXTENDED EVIDENCE

D.1 LAYER-WISE DRIFT AND REPRESENTATION SIMILARITY

Centered kernel alignment (CKA) between the frozen and adapted backbones shows that ALL-4 concentrates its representational change in the late blocks: the late-to-mid ratio of $(1 - \text{CKA})$ is $\approx$ 22–24$\times$, whereas MID-LORA confines change to its training blocks with a ratio of $\approx$ 6–8$\times$. The same qualitative signature appears on the ViT-B/14 backbone (Appendix E). A layer-wise linear probe (reading boundary structure from individual blocks) shows mid-block deltas near zero (preservation) and depth structure non-degraded across depth, with the instance/semantic degradation localized to the late blocks — consistent with the calibration account of altered feature–head compatibility; these diagnostics do not identify a unique geometric rotation mechanism.

D.2 FEATURE-PRESERVING ADAPTER BASELINE

As a lightweight feature-protection control, we retrained the same ALL-4 LoRA adapter with an additional token-level feature distillation loss: $\mathcal{L} = \mathcal{L}_{\text{Ord/Bnd/Vis/Cop}} + \lambda \| f_{\text{adapt}}(x) - f_{\text{frozen}}(x) \|_2^2$, with $\lambda = 1.0$. The adapter uses the same source data, LoRA rank, optimizer, and epoch budget as ALL-4; only the feature MSE term is added. We then ran the same unified base-rate-matched w^+ probe audit on the six representative cells used in the review control, and the matched ScanNet++ source Bnd/Cop probes in Table D8. The source-domain half is important: feature protection preserves

Table D7: Feature-preserving ALL-4 baseline on the six representative instance-boundary cells. Frozen and original ALL-4 values are from the matched unified-w^+ audit; feature-protected uses an ALL-4 LoRA adapter trained with $\lambda = 1.0$ feature MSE to the frozen backbone.

cell	IoU	frozen BF1	ALL-4 ΔBF1	protected ΔBF1
SUN kv1	0.035	0.349	-0.030	$+0.001$
NYU	0.053	0.353	-0.015	-0.000
SUN xtion	0.074	0.305	-0.028	-0.000
SUN kv2	0.114	0.409	-0.016	$+0.002$
Hypersim	0.172	0.533	$+0.005$	$+0.001$
ScanNet	0.208	0.528	-0.007	$+0.001$

Table D8: Superseded: source-domain side of the earlier feature-preserving control ($\lambda = 1.0$) on ScanNet++ Stage C validation, from a cached two-relation probe with different hyperparameters (3 seeds, global scalar F1). Not comparable with Table B2; the protocol-matched replacement is Table G30.

model	Bnd F1	ΔBnd	Cop F1	ΔCop
frozen	0.234 ± 0.015	0.000	0.207 ± 0.054	0.000
ALL-4	0.257 ± 0.015	$+0.023$	0.340 ± 0.058	$+0.133$
feature-protected	0.249 ± 0.014	$+0.015$	0.247 ± 0.075	$+0.040$

positive source recoverability, but not at the full ALL-4 magnitude. Because the cached feature-protected source probe reports global scalar F1, Table D8 should be read as a matched sign/trade-off control; the main source recoverability audit remains the per-image-bootstrap estimand in Table B2. The OOD half in Table D7 shows that explicit feature preservation nearly eliminates the negative BF1 transfer of ALL-4 on these cells. The source-side numbers below come from a cached two-relation probe with different hyperparameters and are superseded by the protocol-matched values in Tables G29 and G30; they are kept here as the record of what the earlier version reported.

Table D7 reports this six-cell target comparison; the ten-cell three-adapter comparison at $\lambda = 0.3$ is in Section 4.3.

D.3 LINEAR VS. MLP PROBE: PART OF THE GAP DEPENDS ON READOUT CAPACITY

A potential objection is that the late-layer "degradation" is an artifact of the linear probe class. Replacing the linear head with a 2-layer MLP (Linear-GELU-Linear, hidden 256) flips several source global-AUPRC deltas (e.g. MID-LORA at block 17: instance $-0.014 \rightarrow +0.009$, semantic $-0.017 \rightarrow +0.005$), so a nonlinear head recovers part of the readout gap. This does not identify an invertible feature rotation: such a transform should be absorbable by a retrained linear probe. Nonlinear distortion, conditioning, finite-data fitting, or optimization can also contribute. Per-seed intervals are tight enough that these flips are systematic, not noise (e.g. block-17 instance linear $\{-0.015, -0.014, -0.013\}$ vs. MLP $\{+0.015, +0.010, +0.002\}$). This is why we frame the result as a *readout-family mismatch* rather than an identified rotation or a loss of capability. We report the linear probe as the main OOD estimand because it is the standard recoverability readout. The OOD MLP experiment below then quantifies how much head flexibility can recover in the transfer setting.

OOD MLP head We reran the external OOD boundary probes with the same train/test splits and a 2-layer MLP head (Linear-GELU-Linear, hidden 256), keeping the frozen and ALL-4 backbones fixed. This is not the primary protocol — the paper studies linear recoverability — but it measures whether the negative transfer is merely a linear-head capacity artifact. The answer is no. MLP heads improve the ALL-4−frozen best-F1 gap in 11/12 cells, and often flip mid/high alignment cells positive, but the lowest-alignment cells remain negative.

dataset	target	IoU	linear gap	MLP gap	gain
SUN-kv1	inst.	.035	−.033	−.021	+.011
SUN-kv1	sem.	.035	−.030	−.006	+.023
SUN-real.	inst.	.072	−.012	−.002	+.010
SUN-real.	sem.	.072	−.016	−.020	−.005
SUN-xtion	inst.	.074	−.027	+.003	+.030
SUN-xtion	sem.	.074	−.022	+.004	+.026
SUN-kv2	inst.	.114	−.015	−.004	+.010
SUN-kv2	sem.	.114	−.018	−.005	+.012
Hypersim	inst.	.172	+.005	+.016	+.011
Hypersim	sem.	.148	−.002	+.003	+.005
ScanNet	inst.	.208	−.008	+.005	+.013
ScanNet	sem.	.206	−.005	+.007	+.012

Here "gap" is Δbest-F1 of ALL-4 minus frozen under the stated head class, and "gain" is MLP gap minus linear gap. The result is a useful boundary on the mechanism: head flexibility can compensate for part of the linear-readout mismatch, especially at moderate/high alignment, but it does not erase the target-side alignment dependence. We therefore keep the paper's main estimand as the standard linear probe and use the MLP experiment as a readout-flexibility control rather than as a replacement protocol.

D.4 Comparison with transferability estimators

To contextualize the target-side alignment diagnostic, we computed two representative feature–label transferability scores, LogME and H-score, across the same 14 OOD cells used in the alignment-trend table. We compute the change in each score from frozen to ALL-4 on the target train split and compare it to held-out best-F1 transfer. Alignment remains the strongest sign-associated coordinate in this board: $\rho_s(\mathrm{IoU}, \Delta\mathrm{BF1}) = +0.737$ ($p = 0.0026$), whereas ΔLogME and ΔH-score are anti-correlated with transfer in this setup (-0.565, $p = 0.035$; -0.706, $p = 0.0048$). The SUN instance subset shows the failure mode concretely: LogME increases after ALL-4 adaptation on all four sensors, and H-score increases on three of four, even though the actual best-F1 transfer is negative on all four:

sensor	ΔBF1	ΔLogME	ΔH-score
kv1	-0.033	$+0.034$	$+0.009$
realsense	-0.012	$+0.022$	-0.005
xtion	-0.027	$+0.027$	$+0.015$
kv2	-0.015	$+0.030$	$+0.031$

This negative comparison is consistent with the paper's central warning: feature–label readability on the target split can improve while transfer *relative to frozen* still worsens. Geometry alignment is not proposed as a replacement for LogME/H-score; rather, it captures a target-side sign-associated coordinate that these scores miss in this setting.

D.5 Class-conditional logit shift, separability, and counterfactual decomposition

Table K40 (main text) reports the class-conditional shift on the three-sensor gradient (kv1, xtion, kv2). The measurement holds the trained head fixed (seed-42; head and features are deterministic in eval mode and the per-patch logit sign is consistent across head seeds), runs inference over all valid test patches for each backbone, and computes paired t-tests separately over the positive and negative patch sets. **Class-conditionality.** On all three sensors the positive-class shift (-0.16 to -0.27) dwarfs the negative-class shift ($|\cdot| \leq 0.033$), so the positive–negative differential is carried almost entirely by the positive class — the signature of a margin compression rather than a uniform offset. The separability statistic confirms it: on xtion, d' (the standardized mean gap between positive and negative logits) falls 1.208 (frozen) $\rightarrow$ 1.085 (ALL-4) and is partially restored to 1.136 by MID-LoRA, mirroring the AUROC ordering in Appendix D.6. Paired-bootstrap 95% CIs make the same point more explicitly: for ALL-4, the class-conditional differential stays negative on all three sensors (kv1 $[-0.292, -0.200]$, xtion $[-0.234, -0.145]$, kv2 $[-0.214, -0.139]$), whereas MID-LoRA keeps negative intervals on kv1 and xtion but drives kv2 to a CI spanning zero ($[-0.051, +0.018]$).

Threshold-repair audit (Table K41). To test whether the loss is an easily removable calibration offset, we split the flat per-patch logits into a held-out half and an eval half (10 random splits). We then measure the best-F1 gap that remains after re-optimizing the decision threshold on the held-out half; a pure calibration offset would be largely removed by this step. We additionally fit a single positive temperature T on the held-out half as a weak sanity check; because this preserves score ordering, it is not independent evidence about oracle BF1. Across kv1/xtion/kv2, 94–105% of the total gap remains after threshold repair; this is an operational residual, not an identified additive component. The MID-LoRA differential restoration ($-0.245 \rightarrow -0.292$ kv1, $-0.192 \rightarrow -0.166$ xtion, $-0.178 \rightarrow -0.016$ kv2) tracks its best-F1 fail/partial/rescue pattern (Table E13) cell-for-cell. The population signature across all 14 cells is the $\Delta\mathrm{thr}$ column of Table C4: 13/14 cells show a negative best-F1-threshold shift consistent with the positive-class depression. The one exception (xtion semantic, $+0.054$) is small and within the noise of a single cell. We release raw per-patch logits for the xtion SUN audit and, for frozen vs. ALL-4, the full-board threshold-audit dumps of Appendix I.4, so the main per-cell threshold analysis can be regenerated directly.

NYU/Hypersim fixed-head audit. We also repeated the fixed-head score-space audit on the cross-dataset cells that motivated the geometry story. With retrained heads, NYU instance/semantic show negative best-F1 transfer under ALL-4 ($-0.015/-0.017$) and smaller losses under MID-LoRA ($-0.004/-0.005$); Hypersim instance is positive and Hypersim semantic is near neutral to mildly negative. Under fixed heads, all four cells show a large positive–negative margin drop, but the

Table D9: NYU/Hypersim multi-head fixed-head audit. We reuse three trained linear heads (seeds 42/43/44). Diagonal rows report retrained-head BF1 transfer; fixed frozen-head rows report the score-space margin shift under adapted features. Outside SUN, the margin consistently shrinks, but the classwise source of the shrinkage can differ from SUN.

cell	adapter	diag. ΔBF1	fixed-head Δmargin	Δpos	Δneg
NYU inst.	ALL-4	-0.015	-0.777	$+0.097$	$+0.874$
NYU sem.	ALL-4	-0.017	-0.806	$+0.090$	$+0.896$
Hypersim inst.	ALL-4	$+0.004$	-1.663	-0.042	$+1.621$
Hypersim sem.	ALL-4	-0.001	-1.647	$+0.037$	$+1.683$
NYU inst.	MID-LoRA	-0.004	-0.656	$+0.169$	$+0.824$
NYU sem.	MID-LoRA	-0.005	-0.688	$+0.102$	$+0.789$
Hypersim inst.	MID-LoRA	$+0.001$	-1.590	$+0.608$	$+2.198$
Hypersim sem.	MID-LoRA	-0.007	-1.589	$+0.608$	$+2.197$

Table D10: Recall-band Δ area (adapter $-$ frozen) and partial AUPRC. ΔBF1 is positive for kv2 MID-LoRA while full ΔAUPRC is ≈ 0: the F1-band win is cancelled by the low-recall-band loss.

cell	cfg	$[0, .25)$	$[.25, .5)$	$[.5, .75)$	$[.75, 1)$	full AUPRC	BF1
kv1	all4	-0.0098	-0.0072	-0.0083	-0.0041	-0.0291	-0.0298
kv1	midlora	-0.0157	-0.0121	-0.0109	-0.0055	-0.0439	-0.0473
kv2	all4	-0.0274	-0.0104	-0.0034	-0.0027	-0.0444	-0.0158
kv2	midlora	-0.0095	$+0.0022$	$+0.0050$	-0.0002	-0.0030	$+0.0172$

class contributing most to the drop differs from the SUN gradient: the negative-class score often rises strongly under the swapped adapted features. This is useful boundary evidence. It says the SUN positive-logit depression is the cleanest direct mechanism instance, and the broader benchmark supports the more general, operational statement used in the main text: adaptation creates a class-conditional margin alteration under the linear boundary readout, not always the identical class-specific displacement pattern.

D.6 Calibration metrics across sensors

Section 4.5 reports AUROC/ECE/Brier on xtion, computed from the dumped per-patch probabilities. The qualitative pattern — a small AUROC drop under adaptation (preserved ranking content), ALL-4 not improving calibration error while MID-LoRA improves it, and the shift localized to the positive class describes, in threshold-free metrics, the same class-conditional shift as the logit displacements of Table K40. We report calibration metrics on xtion because it is the cleanest released SUN mechanism cell; the same dump pipeline now also covers the full frozen/all-4 board for Appendix I.4. We caution that the absolute ECE is dominated by the strong class imbalance (positive rate $\approx 7.5\%$), which is why we also report confidence-ECE and Brier, and why the *differences* between configurations — not the absolute level — carry the calibration signal.

D.7 Full recall-band decomposition

Table D10 gives the band decomposition for both sensors and both ALL-4 and MID-LoRA, with partial-AUPRC truncations at recall ≤ 0.8 and ≤ 0.5. The pattern is consistent: adaptation concentrates its effect in the mid-recall band that best-F1 samples, while the extreme low- and high-recall bands (which dominate the AUPRC integral but are insensitive to a uniform logit shift) dilute the signal. This is the quantitative basis for preferring best-F1 as the mechanism-sensitive estimand while still reporting AUPRC.

D.8 Protocol sensitivity controls: focal and unweighted BCE

Table D11 lists, for every cell of the ten-cell board, the label source, patch threshold, frame counts, scene count, evaluation positive rate, the positive weight derived from training prevalence, and which candidates were evaluated under that recipe. The historical 14-cell board shares every setting except the positive weight, which was fixed at 16 on the real sensors; its cross-cell alignment analysis therefore stays descriptive. The main paper studies the standard linear boundary-probe pro-

Table D11: Protocol coverage of the ten-cell evaluation set. Every model on this set (unadapted backbone; ALL-4 with adapter seeds 42, 43, 44; the four preregistered windows; the feature-preserving adapters) was probed with one recipe: linear probe, 30 epochs, AdamW 10^{-3}, patch threshold τ_p, depth threshold $0.10\,\mathrm{m}$, positive weight w^+ from training prevalence. Probe seeds: three for every model; the unprotected ALL-4 adapters 43 and 44 were first probed with one seed each and later rerun with three (Appendix M). Saved per-patch scores, and therefore scene-paired intervals, exist for the unadapted backbone, the four windows, and ALL-4 and the $\lambda = 0.3$ adapter in all three runs. The historical 14-cell set differs in one setting, a fixed $w^+ = 16$ on the real sensors, and adds NYU and SUN realsense; positive rates quoted in the alignment appendix come from that historical run.

cell	label source	τ_p	train / eval frames	eval scenes	eval pos. rate	w^+
Hypersim instance	rendered instance ids	0.30	2900 / 875	35	0.050	21.1
Hypersim semantic	rendered semantic ids	0.30	2900 / 875	35	0.038	25.6
ScanNet instance	mesh-reprojected instances	0.10	1591 / 283	20	0.129	5.9
ScanNet semantic	mesh-reprojected semantics	0.10	1591 / 283	20	0.118	6.4
SUN kv1 instance	manual polygons	0.10	128 / 32	32	0.078	11.1
SUN kv1 semantic	manual polygons	0.10	128 / 32	32	0.078	11.4
SUN xtion instance	manual polygons	0.10	128 / 32	32	0.075	10.6
SUN xtion semantic	manual polygons	0.10	128 / 32	32	0.068	11.3
SUN kv2 instance	manual polygons	0.10	128 / 32	32	0.113	9.4
SUN kv2 semantic	manual polygons	0.10	128 / 32	32	0.110	10.3

tocol with class-balanced BCE. To test whether the low-vs-high alignment structure is an artifact of this particular loss weighting, we reran the same real-cell board (SUN four sensors plus ScanNet, frozen/ALL-4/MID-LoRA) under two alternative linear-head losses. **Focal loss.** Focal BCE improves best-F1 deltas over weighted BCE in only 1/20 adapter-cell comparisons, flips 0/20 negative cells to non-negative, and leaves 18 BCE-negative comparisons still negative. **Unweighted BCE.** Removing the positive-class weight attenuates several negatives but does not erase the regime structure: unweighted BCE improves best-F1 deltas in 16/20 weighted-vs-unweighted comparisons, flips only 1/20 negative cells to non-negative (SUN xtion semantic MID-LoRA), and leaves 17 weighted-BCE-negative comparisons still negative. Representative comparisons:

cell	cfg	weighted ΔBF1	focal ΔBF1	unweighted ΔBF1
SUN-kv1 inst.	ALL-4	-0.032	-0.030	-0.030
SUN-xtion inst.	ALL-4	-0.027	-0.030	-0.016
SUN-xtion sem.	MID-LoRA	-0.005	-0.014	$+0.000$
SUN-kv2 inst.	MID-LoRA	$+0.012$	$+0.007$	$+0.013$
ScanNet inst.	ALL-4	-0.008	-0.019	-0.005
ScanNet sem.	MID-LoRA	-0.004	-0.007	-0.007

These controls justify the main-text wording: the paper is robust to common linear-head loss variations, but it does not claim protocol invariance of every cell-level magnitude.

D.9 Surface-normal readout-family audit

Surface normals serve two roles. First, they rule out generic NYU feature degradation: the original adapters improve NYU normal error even when boundary transfer on NYU is negative. Second, the window sweep turns normals into a second readout-family selection audit. We train per-patch linear normal heads (30 epochs, cosine loss, 3 seeds) on ScanNet++ source normals and NYUv2 normals. Mean angular error is lower-is-better. Table D12 shows that the normal family has a milder source/OOD rank mismatch: source normals favor early 0–7, whereas NYU normals favor 5–14. This does not reproduce boundary-style strong negative transfer, but it shows that source-side selection safety is readout family dependent.

E Placement Interventions: Full Verdict Tables and Robustness

E.1 Complete mid-LoRA cross-band table

Figure E1 summarizes the best-F1 verdicts. Table E13 adds the per-image and global AUPRC deltas for all ten cells, so the band-dependence is visible across estimands. Note that the kv2 anti→pro flip is on best-F1 (and on global AUPRC) but *not* on per-image AUPRC (MID-LoRA piAUPRC

Table D12: Surface-normal selection audit (mean angular error, lower is better; 3 seeds). Window ranks are computed only within the four layer-window candidates. The second readout family shows a milder rank mismatch: source normals favor early 0–7, while NYU OOD normals favor 5–14.

candidate	source angle°	source win. rank	NYU angle°	NYU win. rank
frozen	44.08	—	26.25	—
ALL-4	—	—	25.80	—
MID-LoRA	44.39	—	—	—
early 0–7	43.96	1	25.69	2
5–14	44.70	3	25.41	1
10–19	44.59	2	25.71	3
late 18–23	44.93	4	26.06	4

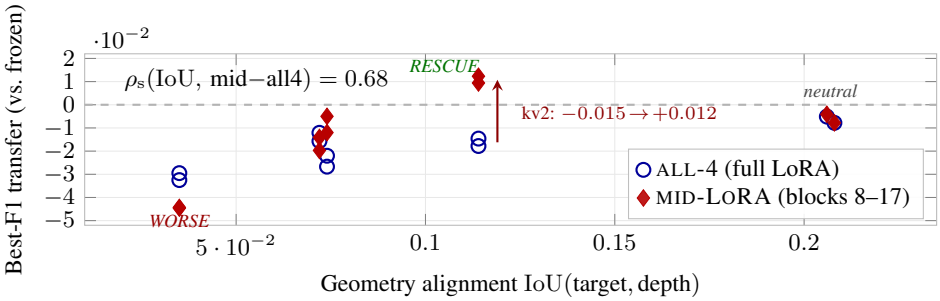

Figure E1: A placement stress test is alignment-conditioned. Under the same BF1 probe protocol, MID-LoRA flips transfer in the mid-alignment band (kv2, $-0.015 \rightarrow +0.012$), is unnecessary on near-neutral ScanNet, and is worse at extreme low alignment; its advantage over full adaptation grows with alignment ($\rho_s = 0.68$).

$\Delta = -0.015/-0.013$ remains slightly negative); this is exactly the estimand behavior the score-space account explains (Section 4.5) and we report it rather than claiming an "all-metric" flip.

E.2 Spearman correlations for the mitigation

Over the ten cells: $\rho_s(\text{IoU}, \Delta\text{BF1}_{\text{all4}}) = 0.73$, $\rho_s(\text{IoU}, \Delta\text{BF1}_{\text{mid}}) = 0.82$, $\rho_s(\text{IoU}, \texttt{mid}-\texttt{all4}) = 0.68$ on best-F1; $\rho_s(\text{IoU}, \Delta\text{gAUPRC}_{\text{mid}}) = 0.72$ and $\rho_s(\text{IoU}, \Delta\text{piAUPRC}_{\text{mid}}) = 0.53$. The MID-LoRA alignment trend is cleaner than ALL-4 on every estimand, and the advantage of layer-selectivity grows with alignment.

Table E13: Full MID-LoRA table (3-seed paired). The kv2 flip is positive on ΔBF1 and ΔgAUPRC but slightly negative on ΔpiAUPRC, consistent with the band decomposition (Appendix D.7).

		ΔBF1		ΔpiAP		ΔgAP	
cell	IoU	all4	mid	all4	mid	all4	mid
kv1 inst	0.035	−0.033	−0.045	−0.014	−0.020	−0.028	−0.039
kv1 sem	0.035	−0.030	−0.044	−0.015	−0.020	−0.028	−0.040
realsense inst	0.072	−0.012	−0.014	−0.005	−0.013	−0.021	−0.025
realsense sem	0.072	−0.016	−0.020	−0.010	−0.020	−0.026	−0.028
xtion inst	0.074	−0.027	−0.012	−0.017	−0.001	−0.024	−0.005
xtion sem	0.074	−0.022	−0.005	−0.016	+0.005	−0.021	+0.000
kv2 inst	0.114	−0.015	+0.012	−0.016	−0.015	−0.046	−0.007
kv2 sem	0.114	−0.018	+0.009	−0.015	−0.013	−0.048	−0.012
ScanNet inst	0.208	−0.008	−0.008	+0.001	−0.004	−0.002	−0.005
ScanNet sem	0.206	−0.005	−0.004	+0.003	−0.001	−0.002	−0.003

Table E14: Layer-window sensitivity (best-F1 transfer vs. frozen). Early 0–7 is best in every representative OOD cell; 5–14 is often second-best; late 18–23 is consistently risky.

cell	IoU	all	early 0–7	5–14	mid 8–17	late 18–23
kv1 inst	.035	−.033	+.002	−.032	−.045	−.033
kv1 sem	.035	−.030	+.003	−.033	−.044	−.033
xtion inst	.074	−.027	+.009	−.000	−.012	−.036
xtion sem	.074	−.022	+.012	+.007	−.005	−.039
kv2 inst	.114	−.015	+.015	+.009	+.012	−.032
kv2 sem	.114	−.018	+.014	+.007	+.009	−.030
Hypersim inst	.172	+.005	+.021	+.004	+.002	−.010
Hypersim sem	.148	−.002	+.017	+.003	−.007	−.013
ScanNet inst	.208	−.008	+.009	−.007	−.008	−.006
ScanNet sem	.206	−.005	+.009	−.005	−.004	−.004

Table E15: Full-board relation-subset and placement controls (best-F1 transfer vs. frozen). Early 0–7 is best in 12/12 cells.

cell	IoU	all	mid	early	late	Bnd+Cop	LoRA-norm
kv1 inst	.035	−.032	−.045	+.002	−.033	−.029	−.015
kv1 sem	.035	−.030	−.044	+.003	−.033	−.028	−.016
realsense inst	.072	−.012	−.014	+.006	−.011	−.018	−.010
realsense sem	.072	−.016	−.020	+.007	−.012	−.023	−.012
xtion inst	.074	−.027	−.012	+.009	−.036	−.005	−.018
xtion sem	.074	−.022	−.005	+.012	−.039	−.004	−.012
kv2 inst	.114	−.015	+.012	+.015	−.032	−.012	−.010
kv2 sem	.114	−.018	+.009	+.014	−.030	−.004	−.012
Hypersim inst	.172	+.005	+.002	+.021	−.010	+.003	+.010
Hypersim sem	.148	−.002	−.007	+.017	−.013	−.006	+.002
ScanNet inst	.208	−.008	−.008	+.009	−.006	−.005	−.006
ScanNet sem	.206	−.005	−.004	+.009	−.004	−.003	−.008

E.3 Layer-window sensitivity

Table E14 gives the complete window sweep and shows why the mitigation claim is no longer "mid-LoRA is optimal." The best OOD-preserving placement is early 0–7 in 10/10 representative cells; after the 6.29 completion with Bnd+Cop-only and LoRA-normal baselines this extends to 12/12 cells (Table E15). Mid 8–17 remains useful in some mid-alignment cells but is dominated by earlier placement.

E.4 Relation-subset controls on the full OOD board

Table E15 adds two controls requested by the review audit: Bnd+Cop-only and LoRA-normal. The result strengthens rather than weakens the placement conclusion. Relation subset identity changes magnitude, but no subset displaces early 0–7 on any of the 12 full-board cells. This is why the main text phrases placement as an OOD-preserving intervention under this protocol, not as a claim that a particular relation objective is universally optimal.

E.5 Rank robustness: early beats late at ranks 4 and 16

To test whether the layer effect is a rank-8 artifact, we train early 0–7 and late 18–23 adapters at ranks 4 and 16 on six representative OOD cells. Early beats late in 12/12 comparisons (Table E16); changing rank modulates magnitude but not the ordering.

E.6 Module placement: not Q/K/V-only, but not every module works

We next fix the block range to early 0–7 and vary the LoRA target module. The result is not a Q/K/V artifact: `attn_proj` matches Q/K/V on kv2 and ScanNet, and `mlp_fc2` is often strongest in the mid/high-alignment cells. The negative control is `mlp_fc1`, which is weak or negative despite improving source recoverability.

Table E16: Rank-placement robustness (best-F1 transfer vs. frozen). Positive "early–late" means early placement is better.

cell	e-r4	l-r4	diff	e-r16	l-r16	diff
kv1 inst	$+.001$	$-.038$	$+.039$	$-.005$	$-.018$	$+.013$
kv1 sem	$+.002$	$-.040$	$+.042$	$-.005$	$-.018$	$+.013$
kv2 inst	$+.016$	$-.029$	$+.046$	$+.016$	$-.030$	$+.046$
kv2 sem	$+.016$	$-.025$	$+.042$	$+.017$	$-.032$	$+.049$
ScanNet inst	$+.008$	$-.004$	$+.012$	$+.012$	$-.007$	$+.019$
ScanNet sem	$+.009$	$-.004$	$+.014$	$+.012$	$-.006$	$+.018$

Table E17: Early-module OOD sweep (best-F1 transfer vs. frozen). Non-QKV modules are non-negative in 10/18 comparisons; `attn_proj` and `mlp_fc2` preserve the early-placement effect in the cells where OOD reuse is plausible.

cell	`qkv`	`attn_proj`	`mlp_fc1`	`mlp_fc2`
kv1 inst	$+.002$	$-.007$	$-.007$	$-.005$
kv1 sem	$+.003$	$-.006$	$-.006$	$-.007$
kv2 inst	$+.015$	$+.015$	$+.009$	$+.016$
kv2 sem	$+.014$	$+.014$	$+.011$	$+.020$
ScanNet inst	$+.009$	$+.009$	$-.002$	$+.010$
ScanNet sem	$+.009$	$+.010$	$-.002$	$+.012$

E.7 Matched adapter-training-seed selection audit

The six-window OOD sweep in Table E14 motivates the claim-critical three-candidate pool: early 0–7, 10–19, and all-block. We evaluate that same pool using adapter-training seeds 42/43/44 and three fresh source-probe seeds per adapter, avoiding comparisons among source summaries from different recovered runs. Source Bnd uses pooled F1 at logit> 0 and source Cop uses argmax F1, matching the released Tier-1 implementation. The OOD utility is the equal-cell mean oracle BF1 over ten boundary cells after averaging three target-probe seeds per cell. Table E18 reports the complete matched audit. Early is target-best for all three adapter seeds, but Bnd and Cop never select it. The six target regrets are all positive (0.0185–0.0319; Table K37). This supports stability of the selection failure, not stability of a particular wrong ranking.

For the seed-42 six-window OOD board, a 10,000-replicate scene bootstrap that uses shared scene multiplicities for paired instance/semantic cells gives early $+0.0111$ BF1 (95% CI $[+0.0066, +0.0155]$). The Bnd-selected 10–19 regret is 0.0197 ($[0.0101, 0.0294]$), and the Cop-selected all-block regret is 0.0265 ($[0.0172, 0.0343]$); both are positive in all replicates. Dataset-balanced utility and every leave-one-dataset analysis also keep early as the target-best candidate. These intervals quantify target-scene sampling, whereas Table K37 separately reports adapter-training variation.

Table E18: Matched source scores and ten-cell target utility across adapter-training seeds. Source values and target cell values are each averaged over three probe seeds. Bold source entries are the score-selected candidates; bold target entries are the target-best candidates. Early is last on both source scores in every run. Regrets quoted in the text are computed from unrounded utilities; recomputing them from the four-decimal entries here can differ by 0.0001.

adapter seed	candidate	source Bnd F1	source Cop F1	target mean BF1
42	early	0.2460	0.2247	**0.4267**
42	10–19	**0.2707**	0.2660	0.4070
42	all-block	0.2604	**0.2896**	0.4002
43	early	0.2469	0.2222	**0.4242**
43	10–19	0.2667	0.2820	0.4084
43	all-block	**0.2692**	**0.3199**	0.3958
44	early	0.2432	0.2575	**0.4281**
44	10–19	0.2640	**0.2873**	0.4096
44	all-block	**0.2664**	0.2728	0.3963
frozen backbone ($\Delta U{=}0$)		—	—	0.4156

Table E19: Source–OOD trade-off across early modules. Source recoverability is measured on ScanNet++ Tier-1; OOD is mean best-F1 transfer over the six SUN/ScanNet cells in the module sweep.

early module	ΔBnd F1	ΔCop F1	mean OOD ΔBF1
`qkv`	+0.007	+0.036	+0.009
`attn_proj`	+0.005	+0.046	+0.006
`mlp_fc1`	+0.008	+0.019	+0.001
`mlp_fc2`	+0.010	+0.065	+0.008

Table E20: Source-based selection audit over early modules. Ranks are descending (1 is best). Source Bnd rank has Spearman $\rho = 0.00$ with mean-OOD rank; source Cop rank has $\rho = 0.40$. This small audit supports the paper's selection warning but does not claim a general checkpoint-selection law.

early module	ΔBnd F1	ΔCop F1	mean OOD ΔBF1	Bnd/Cop rank	OOD rank
`qkv`	+0.007	+0.036	+0.009	3/3	1
`attn_proj`	+0.005	+0.046	+0.006	4/2	3
`mlp_fc1`	+0.008	+0.019	+0.001	2/4	4
`mlp_fc2`	+0.010	+0.065	+0.008	1/1	2

E.8 Source recoverability for early modules

Table E19 summarizes the source–OOD trade-off. All early modules significantly improve source Bnd and Cop recoverability, but their OOD behavior differs. This is the most direct extension of the paper's main warning: source recoverability confirms that the adapter learned a source-readable signal, but it is insufficient evidence that the learned signal will transfer OOD.

E.9 Source-based selection audit for early modules

Table E19 can also be read as a small checkpoint-selection audit. Ranks in Table E20 are descending within the four early module candidates. Selecting the top module by either source Bnd or source Cop recoverability would choose `mlp_fc2`; selecting by mean OOD transfer would choose `qkv`. The regret is small in this four-candidate slice (0.001 mean BF1), so this is not a large-scale selection benchmark. Its value is diagnostic: source recoverability ranks do not reproduce the OOD-safety rank, and `mlp_fc1` is a concrete case where source Bnd improves more than `qkv`/`attn_proj` while mean OOD transfer nearly vanishes.

E.10 Bottleneck family-2 selection audit

We also audit a second adaptation family: a bottleneck adapter trained with the same relation-distillation objective. The completed board has five candidate windows (all blocks, early 0–7, 5–14, 10–19, late 18–23), source Tier-1 probes, 10 OOD boundary cells (SUN kv1/xtion/kv2, Hypersim, ScanNet; instance and semantic), and NYU normals. We use it as bounded adaptation-family breadth, not as a complete family-level benchmark over all PEFT parameterizations. Table E21 shows the selection failure: source Bnd and Cop both favor the 10–19 bottleneck, but mean OOD BF1 and NYU normals favor early 0–7.

Table E21: Bottleneck family-2 selection audit. Mean OOD is averaged over ten boundary cells; NYU normal is mean angular-error delta vs. frozen (lower is better). Source and OOD ranks again do not agree.

candidate	ΔBnd F1	ΔCop F1	mean OOD ΔBF1	NYU normal Δ°
all blocks	+0.035	+0.108	−0.013	−0.187
early 0–7	+0.026	+0.049	+0.003	−0.491
5–14	+0.024	+0.072	−0.006	−0.245
10–19	+0.043	+0.122	−0.010	−0.287
late 18–23	+0.017	+0.115	−0.015	+0.031

Table E22: ViT-B/14 external OOD replication (best-F1 transfer vs. frozen). The qualitative alignment-conditioned pattern persists under the smaller backbone.

cell	IoU	ALL-4	MID-LoRA
kv1 inst	.035	$-.038$	$-.028$
kv1 sem	.035	$-.037$	$-.025$
realsense inst	.072	$-.004$	$-.002$
realsense sem	.072	$-.011$	$-.009$
xtion inst	.074	$+.007$	$+.010$
xtion sem	.074	$+.012$	$+.015$
kv2 inst	.114	$+.007$	$+.006$
kv2 sem	.114	$+.006$	$+.004$
Hypersim inst	.172	$-.008$	$+.012$
Hypersim sem	.148	$-.004$	$+.003$
ScanNet inst	.208	$-.015$	$-.010$
ScanNet sem	.206	$-.013$	$-.010$

E.11 SPLIT-SEED SANITY

To confirm the kv1 WORSE verdict is not a split artifact, re-running kv1 with a different scene split gives MID-LoRA best-F1 $0.304/0.301$ (split 0) vs. $0.364/0.366$ (split 1) on instance/semantic. The absolute level drifts ≈ 0.06 (test $N = 32$ is small) but the paired Δ vs. frozen within each split has std 0.001–0.003, so the WORSE verdict is robust to the split.

E.12 VIT-B/14 CROSS-BACKBONE REPLICATION

Repeating the source recoverability, the mechanism diagnostics, and the layer-selective adapter on a DINOv2 ViT-B/14 backbone (mid-LoRA on blocks 4–8, proportional to ViT-L's 8–17) reproduces the broad qualitative profile: source boundary recoverability rises (best-F1 $+0.027$ to $+0.038$, non-positive in 0 of 5,000 paired resamples); co-planarity is not inflated under the mid variant; the CKA late-to-mid drift ratio is $\approx 22\times$ for full adaptation vs. ≈ 6–$8\times$ for mid; and the layer-wise late-block instance/semantic degradation matches the ViT-L pattern. We then ran the same external OOD board used for the ViT-L placement analysis (Table E22). The signs are not identical cell by cell, but the qualitative profile transfers: low-alignment SUN kv1 remains negative, the SUN mid-band becomes less negative or positive, Hypersim instance is rescued by the layer-selective variant, and ScanNet remains near-neutral to mildly negative. This is evidence that the profile is not confined to ViT-L, but it is not a cellwise sign-stability claim.

E.13 DIODE DEPTH-ONLY DOMAIN EXTENSION

As an out-of-domain (indoor+outdoor) check we run a depth-boundary-only probe on DIODE (Vasiljevic et al., 2019), which provides high-quality lidar depth but no instance/semantic ground truth (we therefore do *not* fabricate instance/semantic labels via segmentation models). DIODE serves only as a depth-boundary domain-extension point and is not part of the main alignment spectrum; we report its positive-rate audit with the released analysis and note that depth-boundary behavior is consistent with the high relevance of depth structure to the distilled adapter.

F PREREGISTERED FOUR-WINDOW POOL

F.1 WHAT WAS FROZEN, AND WHEN

The preregistration document was written on 2026-09-11, before any of the four windows had been scored on the source task or probed on a target cell (three of the four adapters had not yet been trained). Its SHA-256 is `8bc861793d3a83af21d940e3fbbff6a135807f86ce84d28b-ceb01b6033072ff1`; the file will be released with the code. Its clauses, in order:

1. **Candidate rule.** Divide the 24 transformer blocks of DINOv2 ViT-L/14 into four equal, non-overlapping, contiguous windows (0–5, 6–11, 12–17, 18–23) and inject one LoRA adapter per window with the ALL-4 recipe (rank 8, Q/K/V, equal relation weights, AdamW 5×10^{-4}, five epochs, adapter seed 42). The rule has no tunable parameter and consults no

target result. Window 18–23 coincides with an existing checkpoint and reuses it; the report must say so.

2. **Source rules.** The two recorded rules, unchanged: boundary, pooled F1 at the preset threshold of a fresh source probe; coplanarity, pooled F1 of the two-class source probe; each averaged over three probe seeds on the 224 ScanNet++ validation frames; no target data.
3. **Target quantities.** The ten-cell evaluation set; five metrics (oracle F1 over 50 quantiles, exact oracle F1, scene-held-out F1, per-image AP, pooled AP), all reported; equal-cell and dataset-balanced weighting, both reported; the gap between the selected and the best window within the pool.
4. **Decision.** Consistent with the paper if at least one source rule misses the target-best window with a positive gap under all five metrics; partly consistent if only one rule misses or the gap is positive under some metrics; refuted if both rules pick the target-best window, in which case the refutation is the first line of the report.
5. **Discipline.** No threshold, cell set, metric, weighting, or decision rule may change after results are seen; the worst cell and any interval touching zero are reported before means; the result may not be described as a prospective selection benchmark, because the target datasets had been used in earlier stages of the project; scene-paired intervals as in the historical pool; the permitted upper bound of the claim is that, under one pre-specified mechanical candidate rule, the two recorded source rules still fail to select the target-best candidate on this evaluation set.

The decision categories overlap. Read literally, a repetition in which one rule misses the target-best window with a positive gap under all five metrics satisfies both the "consistent" clause and the first branch of the "partly consistent" clause. The document is kept as written (its hash is frozen); Appendix F.2 therefore reports, for each rule, training and metric, whether the gap is positive, and does not assign a single category.

Deviations from the document, registered after the fact. (i) The document named the unadapted backbone's ten-cell mean as 0.4156, taken from the historical probe recipe; the matched recipe used for all prospective comparisons gives 0.4169, and both values appear in the paper with their recipes (Section 3.3). (ii) The document stated that the 18–23 checkpoint's optimization settings had been verified from its metadata; that checkpoint stores no optimization metadata (Appendix I.1), so the recipe of window 18–23 is attributed to the training script, not verified. (iii) The document specified one adapter training per window. We later ran the whole procedure twice more, with adapter seeds 43 and 44, re-fitting the source probes and re-applying both rules inside each repetition; the repetitions were added after the seed-42 results were known, and they change no threshold, cell set, metric, weighting or decision rule. (iv) The document reused the existing 18–23 checkpoint. Window 18–23 was instead retrained under the same recipe in all three trainings, and the reused checkpoint is kept as a cross-check. On the target the two seed-42 versions agree closely: 0.0002 apart in equal-cell oracle F1 and at most 0.0017 apart over the ten metric and weighting combinations. On the source they do not: the mean coplanarity score is 0.2809 for the reused checkpoint and 0.3415 for the retrained one (boundary 0.2473 against 0.2438). With the retrained checkpoint the coplanarity rule still selects 6–11 in the first training, but by 0.3465 to 0.3415 rather than to 0.2809; the source-score tables report the retrained adapters.

F.2 Candidates, source scores, and the outcome

The boundary rule selects window 12–17 in all three trainings, and its gap to the best window is positive under all five metrics and both weightings in each (Tables 1 and F24). The coplanarity rule selects 6–11 in the first two trainings, with equal-cell oracle-F1 gaps of 0.0008 and 0.0063 and a zero gap under dataset-balanced weighting in the first, and 18–23 in the third, with positive gaps under all five metrics and both weightings. The refutation condition, both rules selecting the target-best window, occurs in no training. Because the "consistent" and "partly consistent" clauses overlap, these per-rule, per-training results are the report; no single category is assigned.

Table F23: Source scores of the twelve preregistered adapters: mean pooled F1 over three source probe seeds on the 224 ScanNet++ validation frames. Bold marks each rule's pick inside its own adapter training. Structure is identical across all twelve (six blocks, Q/K/V, rank 8, epoch 4).

adapter training	window	checkpoint SHA-256 prefix	Bnd F1	Cop F1
42	0–5	`a9366fb3`	0.2296	0.2976
	6–11	`32b365b0`	0.2398	**0.3465**
	12–17	`13243942`	**0.2516**	0.2957
	18–23	`2f24c7ac`	0.2438	0.3415
43	0–5	`3d0c084f`	0.2296	0.2955
	6–11	`a58477ce`	0.2427	**0.3368**
	12–17	`3c2edf7c`	**0.2526**	0.2907
	18–23	`946a0418`	0.2482	0.3095
44	0–5	`4a1c0435`	0.2279	0.2963
	6–11	`e5a591de`	0.2444	0.3304
	12–17	`245db534`	**0.2577**	0.2909
	18–23	`c3e37992`	0.2462	**0.3504**

F.3 Evaluation-set identity and the bootstrap

The prospective probes were trained with the matched recipe on the same held-out images as the historical pool: per cell, the image count, the number of valid patches, the positive rate, and the mean valid patches per image agree exactly (for example SUN kv1 instance: 32 images, 22,731 patches, positive rate 0.0783, 710.3 patches per image; ScanNet instance: 283, 202,551, 0.1291, 715.7; Hypersim instance: 875, 560,911, 0.0504, 641.0). Per-patch scores of the 390 prospective probes (four windows in three adapter trainings and the unadapted backbone, ten cells, three probe seeds), plus 30 for the reused 18–23 checkpoint, were exported in the same file layout as the historical pool and scored by the same script.

The scene-paired bootstrap resamples scenes with replacement inside each of the five disjoint scene pools (Hypersim, ScanNet, SUN kv1, kv2, xtion) to the pool's own count, applies one resample to every window and probe seed and to the paired instance and semantic cells, recomputes oracle F1 on a fixed 50-quantile grid per file, and recomputes the best window in every replicate; 10,000 replicates, seed 914. Point estimates agree with the log-based cell means to within 5×10^{-4}. The SUN pools contain one image per scene, so their resampling is image-level. The intervals hold adapters, probes, and source choices fixed and describe target-scene variability only.

Per cell, an early window has the highest oracle F1 in all 30 combinations of cell and training run, and both early windows beat both late windows in 27. In the three exceptions, window 12–17 passes 0–5 but stays below 6–11: Hypersim instance boundaries in run 42 and both SUN kv2 cells in run 44. Under pooled AP, window 12–17 wins both SUN xtion cells in run 44.

Selection gap and unadapted reference. For adapter seed 42 under equal-cell oracle F1, the boundary-selected window trails the best window by 0.0134. This comprises a forgone gain of 0.0053 and a shortfall of 0.0081 below the unadapted backbone. The shortfall is the larger component, but its scene-paired interval includes zero. For seed 44, the coplanarity-selected window is 0.0203 below the backbone with interval $[-0.0288, -0.0098]$. Thus the candidate-relative gap and the difference from the backbone answer distinct comparison questions.

F.4 Sensitivity of source-based selection

In the third adapter training, the source coplanarity scores for window 0–5 are 0.1498, 0.3420, and 0.3973 across the three source probes. For window 18–23 they are 0.3208, 0.3710, and 0.3594. Dropping the first probe changes the selected window from 18–23 to 0–5; dropping either other probe leaves the choice at 18–23. Resampling the nine source validation scenes selects 18–23 in 89.1% of replicates, while the interval on its source-score advantage crosses zero. These checks hold the trained adapters fixed and assess variation in their source scores, separately from the target-scene intervals in the main text.

Table F24: Preregistered pool, all three adapter trainings: each window minus the unadapted backbone under five target metrics and two weightings, three probe seeds per entry. Bold marks the best window of that row, which is the reference for the gaps in Table 1.

		equal-cell				dataset-balanced			
metric	training	0–5	6–11	12–17	18–23	0–5	6–11	12–17	18–23
oracle F1	42	**+0.0053**	+0.0045	-0.0081	-0.0238	+0.0046	**+0.0063**	-0.0045	-0.0180
	43	**+0.0090**	+0.0027	-0.0097	-0.0235	**+0.0086**	+0.0060	-0.0064	-0.0168
	44	**+0.0047**	+0.0037	-0.0087	-0.0203	+0.0053	**+0.0066**	-0.0090	-0.0133
exact F1	42	**+0.0052**	+0.0043	-0.0083	-0.0240	+0.0044	**+0.0060**	-0.0046	-0.0182
	43	**+0.0091**	+0.0027	-0.0100	-0.0236	**+0.0087**	+0.0061	-0.0066	-0.0170
	44	**+0.0046**	+0.0037	-0.0087	-0.0205	+0.0052	**+0.0066**	-0.0090	-0.0135
held-out F1	42	**+0.0050**	+0.0042	-0.0090	-0.0264	+0.0040	**+0.0059**	-0.0052	-0.0199
	43	**+0.0088**	+0.0024	-0.0111	-0.0270	**+0.0083**	+0.0056	-0.0074	-0.0189
	44	+0.0034	**+0.0039**	-0.0094	-0.0229	+0.0042	**+0.0064**	-0.0100	-0.0151
per-image AP	42	**+0.0087**	+0.0032	-0.0039	-0.0149	**+0.0091**	+0.0075	-0.0005	-0.0085
	43	**+0.0135**	+0.0058	-0.0050	-0.0130	**+0.0147**	+0.0095	-0.0022	-0.0052
	44	**+0.0104**	+0.0052	-0.0067	-0.0122	**+0.0120**	+0.0112	-0.0035	-0.0055
pooled AP	42	**+0.0073**	+0.0009	-0.0082	-0.0336	+0.0058	**+0.0072**	-0.0046	-0.0257
	43	**+0.0112**	+0.0020	-0.0116	-0.0310	**+0.0108**	+0.0079	-0.0082	-0.0245
	44	**+0.0079**	+0.0051	-0.0131	-0.0280	+0.0092	**+0.0124**	-0.0108	-0.0197

F.5 Source boundary scores under three scoring rules, per probe seed

Table F25: Source boundary scores under three scoring rules. Window picked by each rule in the first run and its gap to the best window in mean oracle F1. The original run trained all four source probes; the rerun trained the boundary probe alone and used the older 18–23 checkpoint. The coplanarity choice of the first run is in Table 1.

source score	run	picked	gap
Bnd F1 at fixed threshold	original	12–17	0.0134
Bnd F1 at fixed threshold	boundary-only rerun	18–23	0.0293
Bnd best F1 over thresholds	boundary-only rerun	12–17	0.0134
Bnd average precision	boundary-only rerun	12–17	0.0134

The boundary-only rerun trained a fresh boundary probe on each window with the same protocol as the preregistered run but without the other three source probes, which changes the random stream at probe initialization. Table F26 gives the per-seed values behind Table F25: best-F1 and AP rank window 12–17 first in each of the three probe seeds, and the preset-threshold F1 ranks window 18–23 first in each.

Table F26: Boundary-only rerun, source validation frames, probe seeds 42 / 43 / 44.

window	best-F1	AP	F1 at preset threshold
0–5	0.2945 / 0.3081 / 0.2917	0.2495 / 0.2588 / 0.2424	0.2144 / 0.2340 / 0.2270
6–11	0.3253 / 0.3401 / 0.3220	0.2723 / 0.2789 / 0.2717	0.2175 / 0.2416 / 0.2367
12–17	**0.3335 / 0.3447 / 0.3460**	**0.2818 / 0.2891 / 0.2853**	0.2229 / 0.2427 / 0.2255
18–23	0.2880 / 0.2996 / 0.2946	0.2380 / 0.2444 / 0.2392	**0.2454 / 0.2571 / 0.2448**

F.6 Exploratory: the four windows retrained with feature preservation

All four windows were retrained with $\lambda = 0.3$ (one adapter each, the window 18–23 now trained fresh with the same script), scored on the source task with three probe seeds, and probed on the ten cells with three probe seeds. This experiment was not preregistered.

Table F27: Protected windows ($\lambda = 0.3$). Source F1 with three-seed standard deviation; target ten-cell mean oracle F1 (three probe seeds) and its difference from the unadapted backbone (0.4169).

window	source Bnd F1	source Cop F1	target mean	minus unadapted
0–5	0.2272 (0.0061)	0.2966 (0.1322)	0.4170	+0.0001
6–11	0.2334 (0.0064)	0.3003 (0.1305)	0.4162	−0.0006
12–17	0.2401 (0.0078)	0.3141 (0.1088)	**0.4189**	+0.0020
18–23	**0.2431** (0.0061)	**0.3207** (0.0780)	0.4141	−0.0028

Both source-score means select window 18–23; window 12–17 has the highest target mean in each of the three probe seeds, and the range of the four target means is 0.0048 against 0.0291 without the penalty. The source coplanarity scores have a three-seed standard deviation of 0.08–0.13, so the coplanarity pick is a ranking of means rather than a separated one.

G Feature preservation: adapter-seed replication and dose sweep

Table G28: Target changes under five metrics, ten cells, against the unadapted backbone. ALL-4 and the smallest learning rate (5×10^{-5}) each use adapter seed 42; preservation ($\lambda = 0.3$) averages three adapter trainings. All entries average three target-probe seeds per adapter.

	ten-cell mean			datasets weighted equally		
metric	ALL-4	preservation	low LR	ALL-4	preservation	low LR
oracle F1	−0.0151	−0.0010	−0.0055	−0.0094	−0.0001	−0.0013
exact F1	−0.0151	−0.0011	−0.0058	−0.0094	−0.0002	−0.0014
scene-held-out F1	−0.0173	−0.0017	−0.0078	−0.0108	−0.0005	−0.0027
per-image AP	−0.0070	+0.0001	−0.0025	−0.0009	+0.0020	+0.0030
pooled AP	−0.0210	−0.0018	−0.0076	−0.0132	+0.0004	−0.0011

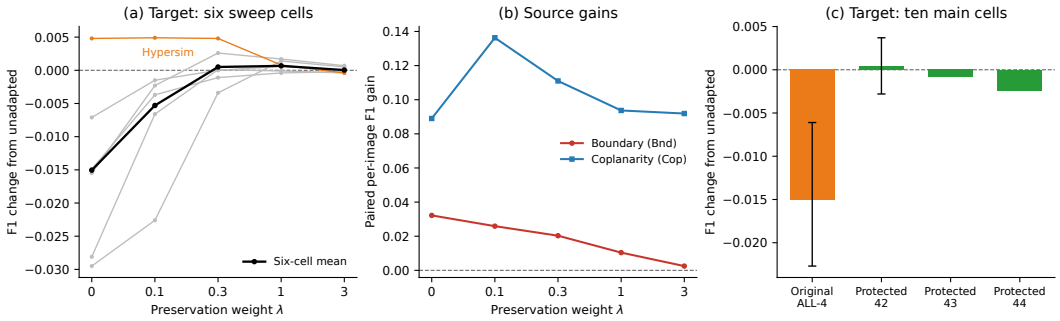

Figure G2: Feature preservation across penalty weights. (a) Mean oracle-F1 change against the unadapted backbone on the six sweep cells (black; adapter seed 42, three probe seeds per point), with individual cells in gray and Hypersim instance in orange. (b) Mean per-image source gains for boundary and coplanarity, three probe seeds per point. (c) Ten-cell mean oracle-F1 change for the single original ALL-4 adapter and three adapters trained with $\lambda = 0.3$. Every bar averages three probe seeds. Whiskers are the scene-paired 95% intervals available for the original adapter and preserved seed 42.

For $\lambda \in \{0.3, 1.0, 3.0\}$ the three probe seeds were run as three processes with a probe-seed offset while the deterministic test-pair sampling stayed anchored to base seed 42; the per-seed results were merged before the paired bootstrap. The per-seed runs reset the global random state at the start of each probe, so the split is equivalent to a single three-seed run up to CUDA nondeterminism; this equivalence was not checked bit for bit. An earlier version of this paper reported the $\lambda = 1.0$ and 3.0 coplanarity gains from probe seed 42 alone (+0.0099, non-positive in 8% of resamples, and −0.0031); that seed is the lowest of the three at every λ, and those single-seed values are superseded by the table.

Table G29: ALL-4 with $\lambda = 0.3$, three adapter seeds. Source values are three-probe-seed means with paired per-image gains against the unadapted backbone, with the share of 5,000 paired image resamples in which the gain is non-positive (column "gain ≤ 0"); target values are ten-cell means (three probe seeds) minus the unadapted backbone.

adapter seed	Bnd F1	Bnd gain	gain ≤ 0	Cop F1	Cop gain	target minus unadapted
42	0.2544	$+0.0203$	0/5000	0.3342	$+0.1110$	$+0.0004$
43	0.2332	$+0.0006$	24%	0.2996	$+0.0861$	-0.0009
44	0.2521	$+0.0194$	0/5000	0.3427	$+0.1154$	-0.0024

Table G30: Dose sweep, source side, all-block adapter seed 42. Paired per-image gains against the unadapted backbone, three probe seeds; per-seed coplanarity F1 in the last column shows the large probe-seed variance of that score (the unadapted backbone reaches 0.2072); "Cop gain ≤ 0" is the share of 5,000 paired resamples with a non-positive coplanarity gain.

λ	Bnd gain	Cop gain	Cop gain ≤ 0	Cop F1 per probe seed
0	$+0.0322$	$+0.0890$	0/5000	–
0.1	$+0.0259$	$+0.1363$	0/5000	–
0.3	$+0.0203$	$+0.1110$	0/5000	0.2704 / 0.3702 / 0.3619
1.0	$+0.0104$	$+0.0937$	0/5000	0.2148 / 0.3490 / 0.3642
3.0	$+0.0025$	$+0.0919$	0/5000	0.1992 / 0.3466 / 0.3743

G.1 Feature preservation and lower-learning-rate controls

The three lower-learning-rate settings use 2×10^{-4}, 10^{-4}, and 5×10^{-5}, with one adapter training per setting. Their source boundary gains are 0.0345, 0.0231, and 0.0164; the corresponding equal-cell target oracle-F1 changes are -0.0140, -0.0095, and -0.0055. A line fitted to these three settings gives a descriptive reference for their source–target relation. The protected adapters with seeds 42 and 44 lie 0.0081 and 0.0048 above this line. Seed 43 has a source boundary gain of 0.0006, outside the range covered by the lower-learning-rate settings, and is excluded from this interpolation. The fitted line is not an estimated population trade-off, and these three learning rates do not cover all forms of weaker adaptation. Under dataset-balanced per-image AP, the smallest learning rate improves over the unadapted backbone by 0.0030, compared with a three-training mean of 0.0020 with preservation.

For preservation at $\lambda = 0.3$, the third adapter training has the smallest recovery in eight of the ten metric–weighting combinations. Per-image AP under either weighting is the exception. Equal-cell pooled AP changes by -0.0009 to -0.0036 across the three trainings, all below the unadapted backbone. These per-training differences are included in the reported three-training means; the means lying within ± 0.002 do not imply that each training does. All metric-level results are retained in `preservation_three_seeds_v1/utilities.json` and `weak_update_analysis_v1/utilities.json` in the experiment records.

H Historical pool: intervals, exceptions, and the 14-cell held-out audit

The historical pool contains Early (blocks 0–7), 10–19, and ALL-4, each trained with adapter seeds 42, 43, and 44 under the historical probe recipe. Early has the lowest source boundary and coplanarity scores and the highest ten-cell target mean in each training. Its target means are 0.4267, 0.4242, and 0.4281, against 0.4156 for the unadapted backbone. Window 10–19 reaches at most 0.4096 and ALL-4 at most 0.4002. Both source rules therefore select an adapter below the backbone while another candidate improves on it. The selected-to-best gaps span 0.0185–0.0319 under oracle F1 and remain positive under exact F1, scene-held-out F1, per-image AP, pooled AP, and dataset-balanced weighting (Table K37). Leaving out one dataset at a time preserves Early as the winner (Table K38). Early beats both other candidates in all 150 per-cell comparisons, but falls below the backbone in nine SUN kv1 comparisons (Table H32).

H.1 SCENE-PAIRED INTERVALS OF THE SIX RECORDED SELECTION GAPS

The same resampling scheme as Appendix F (scenes resampled within each of the five pools, one resample applied to every candidate and probe seed, target-best recomputed per replicate, 50-quantile grid fixed, 10,000 replicates, seed 910) was applied to the historical pool with each adapter seed's recorded source choices held fixed. Every replicate had Early as the target-best candidate under equal-cell weighting; no replicate had a gap of zero or below in any of the twelve rows.

Table H31: Historical pool, oracle-F1 selection gap of the recorded source choice, scene-paired percentiles.

weighting	adapter seed, rule	selected	point	2.5%	97.5%
equal-cell	42, Bnd	10–19	0.0197	0.0101	0.0294
	42, Cop	ALL-4	0.0265	0.0172	0.0343
	43, Bnd and Cop	ALL-4	0.0284	0.0179	0.0375
	44, Bnd	ALL-4	0.0319	0.0215	0.0408
	44, Cop	10–19	0.0185	0.0104	0.0264
dataset-balanced	42, Bnd	10–19	0.0195	0.0089	0.0297
	42, Cop	ALL-4	0.0221	0.0133	0.0298
	43, Bnd and Cop	ALL-4	0.0254	0.0155	0.0342
	44, Bnd	ALL-4	0.0306	0.0206	0.0398
	44, Cop	10–19	0.0197	0.0115	0.0283

The three SUN pools contain one image per scene, so their resampling is image-level; the intervals cover target-scene variability given the saved adapters and probes and do not cover adapter-training or source-score variability.

H.2 WHERE THE BEST HISTORICAL CANDIDATE TRAILS THE UNADAPTED BACKBONE

Of the 150 cell-level comparisons (ten cells, five metrics, three adapter seeds, probe seeds averaged), Early trails the unadapted backbone in nine, all on SUN kv1 and all under threshold-based metrics; under per-image and pooled AP Early exceeds the backbone on kv1 by 0.0040 to 0.0237. Removing one dataset never changes the winner across the 105 aggregate slices (seven slices, five metrics, three adapter seeds); removing SUN, or keeping only Hypersim, promotes ALL-4 to runner-up in 6 of 15 metric–seed combinations.

Table H32: The nine cases in which Early scores below the unadapted backbone (Early minus backbone; historical probe recipe).

metric	adapter seed	cell	difference
oracle F1	43	SUN kv1 semantic	-0.0015
oracle F1	44	SUN kv1 instance	-0.0047
oracle F1	44	SUN kv1 semantic	-0.0051
exact F1	43	SUN kv1 semantic	-0.0017
exact F1	44	SUN kv1 instance	-0.0046
exact F1	44	SUN kv1 semantic	-0.0048
held-out F1	43	SUN kv1 semantic	-0.0022
held-out F1	44	SUN kv1 instance	-0.0045
held-out F1	44	SUN kv1 semantic	-0.0066

H.3 THE 14-CELL BOARD UNDER SCENE-GROUPED HELD-OUT THRESHOLDS

For the unadapted backbone and the all-block adapter on the historical 14-cell evaluation set, per-patch scores of the saved probes (three probe seeds) were re-scored with the decision threshold fitted on one half of the scenes and evaluated on the other, two-fold, over 20 repeated scene partitions shared by both models. The equal-cell mean change is -0.0166 (image-grouped partitions give -0.0165); 13 of 14 cells are negative, and 12 are negative for all three probe seeds. This audit supersedes the earlier random patch-half analysis of Appendix I.4, whose splits did not respect scene membership. Source

files: `runs_20260906/main_board_logits_with_ids_v1/grouped_scene_r20` and `mainboard_ap_20260911_v1`.

I REPRODUCIBILITY

I.1 SOURCE TRAINING OBJECTIVE

Let $F \in \mathbb{R}^{N\times d}$ be the adapted patch tokens of a frame ($N = 1024$, $d = 1024$). Two head types read them. A pairwise head maps a patch pair (p, q) through $[\,F_p,\ F_q,\ F_p - F_q,\ F_p \odot F_q\,] \in \mathbb{R}^{4d}$ and a two-layer GELU MLP with hidden width 384 to C logits; a per-patch head maps F_p through the same width to one logit. Ord uses a pairwise head with $C = 3$, Cop a pairwise head with $C = 2$, and Bnd and Vis per-patch heads.

The four relation losses are

$$\begin{aligned}
\mathcal{L}_{\mathrm{Ord}} &= \mathrm{CE}\big(z^{\mathrm{ord}},\, y^{\mathrm{ord}}\big), \quad y^{\mathrm{ord}} \in \{<, >, \approx\},\\
\mathcal{L}_{\mathrm{Bnd}} &= \operatorname*{mean}_{p\in V} \big(1-\pi_p\big)^{\gamma}\, \mathrm{BCE}_{w^+}\big(z_p^{\mathrm{bnd}},\, y_p^{\mathrm{bnd}}\big), \quad \gamma = 2,\ w^+ = 10,\\
\mathcal{L}_{\mathrm{Vis}} &= \operatorname*{mean}_{p\in V} \mathrm{smooth}L_1\big(\hat{v}_p,\, v_p\big), \quad v_p \in [0,1],\\
\mathcal{L}_{\mathrm{Cop}} &= \mathrm{CE}\big(z^{\mathrm{cop}},\, y^{\mathrm{cop}}\big), \quad y^{\mathrm{cop}} \in \{0,1\},
\end{aligned}$$

where V is the set of patches valid for that relation, π_p is the predicted probability of the observed class (so $(1-\pi_p)^{\gamma}$ is the focal factor), and BCE_{w^+} weights the positive term by w^+. Ord and Cop labels come from 512 patch pairs sampled per frame per relation; pair selection is reproducible when the deterministic-pairs option is set, which the evaluation loaders use. Vis and Bnd are dense over valid patches.

Feature-preservation term. Let $F \in \mathbb{R}^{B\times N\times d}$ be the student's final patch tokens for a batch, taken after the backbone's last LayerNorm (the `x_norm_patchtokens` output; the CLS token is excluded), and F^0 the same tokens from the unadapted backbone, computed under `no_grad` with the teacher in evaluation mode. The term is

$$\mathcal{L}_{\mathrm{feat}} = \frac{1}{BNd} \sum_{b,p,c} \big(F_{bpc} - F^0_{bpc}\big)^2,$$

a single layer, averaged over batch, tokens and channels, with no mask (unlike Bnd and Vis it includes invalid patches), no further normalization, and no detaching on the student side. The total loss is $\mathcal{L}_{\mathrm{Ord}} + \mathcal{L}_{\mathrm{Bnd}} + \mathcal{L}_{\mathrm{Vis}} + \mathcal{L}_{\mathrm{Cop}} + \lambda\,\mathcal{L}_{\mathrm{feat}}$ with $\lambda \in \{0, 0.1, 0.3, 1, 3\}$; because of the mean reduction, the scale of λ does not depend on B, N or d.

LoRA scaling. Each adapted projection computes $Wx + \alpha\, BAx$ with $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$, $r = 8$ and $\alpha = 0.5$ applied as a direct multiplier, not as α/r; B is zero-initialized. Read under the common α/r convention, the effective scale would be misstated by a factor of 8.

Label construction. Labels are computed from the posed RGB-D frames of ScanNet++ on the 32×32 patch grid and validated in partner views; each frame has ten partners drawn from three camera-baseline buckets in the ratio 4:4:2. A patch is valid when at least 10% of its pixels carry depth and at least 10% are foreground. For **Ord**, a pair (p, q) of valid patches is labeled $<$, $>$, or $\approx$ from the depths at the two patch centers with a deadband of 5% of the smaller depth; the pair is kept only if the same label holds after reprojecting both centers into at least one partner view whose depth agrees with the reprojected depth within 10%. For **Bnd**, a pixel is a boundary when its depth-gradient magnitude exceeds the 90th percentile of the valid gradients of that image and the cross-view visibility flips between the two sides of the gradient direction, which separates occlusion boundaries from surface folds; pixel labels are pooled to patches with the same validity mask. **Vis** is the fraction of the ten partner views in which the back-projected pixel is not occluded (again with the 10% depth-agreement test), a score in $[0, 1]$. For **Cop**, a local plane is fitted to the depth in a 7×7 window around each patch; a pair is coplanar when the two normals differ by less than 10° and the depth offset of one patch from the other's plane is below 5% of their mean depth, and a positive

Table I33: Adapter-training configuration. All variants use the ScanNet++ Stage-C train split, AdamW with weight decay 0.05, batch size 8 frames, 5 epochs, 2 warmup epochs, and equal relation weights unless stated otherwise. Relation heads are two-layer GELU MLPs with hidden dimension 384; Bnd/Vis are per-patch heads and Ord/Cop are pairwise heads.

variant	edited layers/modules	capacity	loss terms	LR
ALL-4	blocks 0–23, Q/K/V	LoRA r=8, α=0.5	Ord+Bnd+Vis+Cop	5×10^{-4}
MID-LORA	blocks 8–17, Q/K/V	LoRA r=8, α=0.5	Ord+Bnd+Vis+Cop	5×10^{-4}
placement sweep	selected block windows/modules	LoRA r=8 unless swept	Ord+Bnd+Vis+Cop	5×10^{-4}
protected	blocks 0–23, Q/K/V (and the four windows at λ=0.3)	LoRA r=8, α=0.5	four relations; feature MSE with $\lambda \in \{0.1, 0.3, 1.0, 3.0\}$	5×10^{-4}

pair is kept only if the same test passes in at least one partner view. These thresholds are the defaults of the label-generation script and were not tuned on any target result.

The total is $\mathcal{L} = \sum_r \omega_r \mathcal{L}_r$ with $\omega_{\text{Ord}} = \omega_{\text{Bnd}} = \omega_{\text{Vis}} = \omega_{\text{Cop}} = 1$ for every adapter in the main comparisons (the relation-subset controls of Appendix E.4 drop terms), and *no gradient normalization*: an earlier project plan described gradient normalization with a warmup, but the code that produced these checkpoints applies the plain weighted sum, and we report what was run. Class imbalance is handled only inside $\mathcal{L}_{\text{Bnd}}$ (focal factor plus $w^+ = 10$); Ord and Cop use unweighted cross-entropy and Vis an unweighted regression loss. The feature-preserving variant adds $\lambda \| F - F^{\text{frozen}} \|_2^2$ token-wise on top of this sum.

Optimization: AdamW, learning rate 5×10^{-4}, weight decay 0.05, cosine schedule with two warmup epochs, batch size eight frames, five epochs over the 2,194-frame Stage-C train split (1,370 steps). For every adapter trained after the checkpoint format was extended (the preregistered windows 0–5, 6–11, and 12–17, all feature-preserving adapters, and the ALL-4 adapter seeds 43 and 44), the relation weights, learning rate, epoch count, batch size, LoRA target modules, and target block list are stored in the checkpoint and were read back from the saved files rather than from a configuration template; the shipped YAML retains placeholder training fields and is not the source of these values. Two checkpoints predate this bookkeeping: the unprotected ALL-4 adapter (seed 42) and the older 18–23 checkpoint kept as a cross-check store only the LoRA tensors, the relation heads, the epoch index, and the final losses, and their training logs were lost in a storage migration. Their architecture (24 or 6 blocks, Q/K/V, rank 8) is verified from the tensors; their optimization settings are those of the training script, which are the values above, and the ALL-4 seeds 43 and 44 produced by the same script record exactly these values.

I.2 HYPERPARAMETERS

Backbone: DINOv2 ViT-L/14 (24 blocks, 1024-d), input 448×448, 32×32 patch grid. Main adapter: LoRA r=8, $\alpha = 0.5$ as a direct multiplier (Appendix I.1) on Q/K/V; ALL-4 on all blocks, MID-LORA on blocks 8–17. Placement sweeps additionally train early 0–7, 5–14, 10–19, and late 18–23 block windows; rank sweeps use $r \in \{4, 16\}$ for early vs. late; module sweeps use `qkv`, `attn_proj`, `mlp_fc1`, and `mlp_fc2` at early 0–7. Probes: per-patch linear head, AdamW with lr 10^{-3}, 30 epochs, 3 seeds $\{42, 43, 44\}$, deterministic test pairing, frame-uniform minibatches, and BCE averaged over all valid patches in the minibatch. The fixed boundary weights are $w^+ = 16/20/18$ for instance/semantic/depth in the NYU/SUN/ScanNet patch-audit protocol; Hypersim and the unified-w^+ control use $w^+ = (1 - p_{\text{train}})/p_{\text{train}}$, and the ScanNet++ source Bnd probe uses $w^+ = 10$. Source Bnd F1 uses logit> 0 and source Cop F1 uses argmax; target BF1 uses the threshold protocol below. Statistics: 5000-sample paired bootstrap over per-image metrics; we report mean$\pm$std and p-values.

I.3 BEST-F1 THRESHOLD PROTOCOL

The standard BF1 tables use an oracle operating-point protocol. For each seed/model/cell, we pool valid evaluation patches, sweep global thresholds over 50 empirical quantiles of the evaluation logits (0.5–0.999), record the threshold with the highest F1, and then average metrics across seeds. This is

Table I34: Full-board threshold audit. Recomputing the 14-cell board from per-patch logits shows that the low-alignment negative regime is not purely an evaluation-split oracle artifact. The held-out threshold protocol preserves the same 13/14 negative board and almost the same IoU ordering; a single fixed threshold weakens but does not erase the trend.

threshold protocol	$\rho_s(\mathrm{IoU}, \Delta\mathrm{BF1})$	p	non-negative cells
50-quantile oracle BF1	0.737	0.0026	1/14
exact sorted-score BF1	0.744	0.0023	1/14
held-out split BF1	0.726	0.0033	1/14
fixed logit>0 BF1	0.534	0.0491	6/14

Table I35: Per-cell threshold audit for ALL-4−frozen. Q50 is the standard 50-quantile oracle BF1 protocol; exact sweeps all sorted scores; held-out chooses the threshold on one random half and evaluates on the other (10 splits); $e - q$ is exact minus Q50.

cell	Q50 Δ	exact Δ	$e - q$	held-out f	held-out ALL-4	held-out Δ
SUN kv1 inst	−0.032	−0.033	−0.0004	0.344	0.311	−0.033
SUN kv1 sem	−0.030	−0.029	+0.0001	0.340	0.310	−0.030
NYU sem	−0.017	−0.017	−0.0000	0.347	0.329	−0.018
NYU inst	−0.015	−0.015	−0.0000	0.352	0.336	−0.015
SUN realsense inst	−0.012	−0.012	−0.0000	0.278	0.265	−0.013
SUN realsense sem	−0.016	−0.015	+0.0008	0.265	0.253	−0.012
SUN xtion inst	−0.027	−0.027	−0.0002	0.299	0.272	−0.026
SUN xtion sem	−0.022	−0.022	+0.0001	0.289	0.268	−0.020
SUN kv2 inst	−0.015	−0.015	−0.0002	0.405	0.390	−0.015
SUN kv2 sem	−0.018	−0.018	−0.0001	0.402	0.384	−0.018
Hypersim sem	−0.002	−0.001	+0.0007	0.484	0.483	−0.001
Hypersim inst	+0.005	+0.004	−0.0009	0.534	0.538	+0.004
ScanNet sem	−0.005	−0.005	−0.0000	0.509	0.505	−0.005
ScanNet inst	−0.008	−0.008	+0.0000	0.524	0.517	−0.008

why the main paper interprets BF1 as linear operating-point recoverability rather than deployment performance. The threshold selected for one seed is never reused for another seed or model, and logits are not averaged before threshold selection. The logit-dump mechanism audits use an exact sorted-score sweep, so the SUN fixed-head analysis is not sensitive to the 50-quantile approximation. By contrast, the threshold-repair test in Section 4.5 uses the SUN logit dumps to split patches into held-out/eval halves: the repair threshold or temperature is fit on the held-out half and evaluated on the paired eval half. Beyond that three-sensor SUN repair test, we now also present a full-board threshold audit from per-patch logits for frozen and ALL-4: the threshold is chosen on one half of each cell and evaluated on the other, with a fixed logit>0 check as a deployment-like contrast (Appendix I.4).

I.4 FULL-BOARD HELD-OUT-THRESHOLD AUDIT

To test whether the main best-F1 mismatch is merely an oracle-threshold artifact, we dumped per-patch logits for frozen and ALL-4 on the full 14-cell board and recomputed ΔBF1 under four protocols: (i) the standard 50-quantile evaluation-split oracle BF1 used in the main text, (ii) an exact sorted-score oracle sweep, (iii) a held-out half/eval-half BF1 where the threshold is fit on one half and evaluated on the other over 10 random splits, and (iv) a single deployment-like fixed threshold at logit>0. This audit does not replace the main estimand; it asks whether the low-alignment negative regime disappears once threshold choice is no longer oracle. All four rows are taken from a single later recomputation over the full-board logit dumps, so that the four protocols share one input; an earlier partial recomputation of the same held-out protocol on three of the four protocols gave $\rho_s = 0.702$ ($p = 0.005$), and we report the four-protocol run rather than mixing the two.

For auditability, Table I35 expands the full-board threshold check by cell. The exact sorted-score sweep never changes the sign of the standard 50-quantile BF1 delta, and held-out-threshold sign agreement with Q50 is 14/14. The held-out board preserves the same qualitative pattern: 13/14 cells remain negative, with only Hypersim instance positive.

The exact-vs-Q50 differences are all below 0.001 BF1 in magnitude, so the 50-quantile approximation does not create the sign pattern or the ordering used in the main alignment board.

The held-out column keeps the same qualitative reading as the main board: all low-alignment polygon-labeled real-sensor cells remain negative, Hypersim instance remains positive, and ScanNet stays near neutral. The fixed-threshold column is weaker, which is consistent with the paper's scope: best-F1 is still an operating-point estimand rather than a deployment metric. The important negative result is therefore narrower than "OOD utility in general" but broader than "oracle BF1 only on the evaluation split."

I.5 Checkpoints reused across all evaluations

Each adapter checkpoint is trained once on the ScanNet++ Stage-C indoor source split and then reused unchanged across its OOD evaluation cells; no adapter is retrained for a target dataset. The main diagnostic uses one ALL-4 checkpoint across NYU, Hypersim, SUN RGB-D (all sensors), ScanNet-v2, and DIODE. The placement/rank/module sweeps train additional source checkpoints only to test the intervention in Section 4.3; the probe heads are still the only target-specific trained components. The source split consists of 2194 training frames and 224 validation frames with posed RGB-D/DSLR imagery and per-scene camera intrinsics/extrinsics; source train/validation pairs are scene-disjoint. OOD evaluations are also scene-disjoint within their own datasets (SUN RGB-D and ScanNet-v2 use scene-level splits; Hypersim uses a scene-disjoint split; NYUv2 uses the standard 795/654 split), and no OOD evaluation frame is used to train an adapter. No adapter is retrained per target. The logit-dump experiment reuses the already-trained probe heads and only swaps the backbone, by design (Appendix D.5).

I.6 Compute

The historical main sweeps used one 80 GB GPU; the recovered September validation, source-selection, and adapter-seed reruns used one RTX 4090. Probe inference is $\sim$ 2.3 GB. Indicative wall-clock: the 16-process SUN RGB-D frozen/all-4 sweep $\approx$ 2h47m; the 4-process ScanNet sweep $\approx$ 7h19m; the 10-cell MID-LoRA cross-band chain $\approx$ 5h; the layer-window, rank, module, and source-recoverability enhancement queues run as serialized overnight jobs; the per-patch logit dump $\approx$ 3min (inference only); the NYU normal probe $\approx$ 4h (I/O bound).

I.7 Released artifacts

We will release: the boundary-probe and IoU-precheck code; the per-cell result files (per-seed best-F1, best-F1 threshold, per-image AUPRC arrays); the alignment-analysis JSONs (dataset-level IoU/lift, τ-matched sanity, image-level correlation); the recall-band decomposition; the SUN and full-board logit-dump summaries; the layer/rank/module/source-recoverability reports; and the dataset manifests (stratified scene lists) needed to reproduce every table.

J Negative Results, Abandoned Analyses, and Self-Correction Log

In the interest of an auditable record we document what did not work and what we corrected, since several of these directly shaped the claims we do and do not make.

Corrected Hypersim pilot (sampling + threshold confound) An initial Hypersim run reported negative depth-boundary transfer. We traced this to two confounds: (i) the training frames were drawn from only two sub-collections (sampling collapse), and (ii) the depth threshold was reused from NYU, pushing the synthetic positive rate to $\sim 29\%$ and changing the metric's meaning. The corrected v1 (stratified 31/26 sub-collection manifest, separated $\tau_{\text{depth}}/\tau_{\text{patch}}$, base-rate-matched positive rates) reverses the conclusion and is the only Hypersim result we report. The pilot is retained on disk for audit but never cited as a finding.

Replica was dropped for lack of native labels Our on-disk Replica build provides RGB+depth+poses and depth-derived boundary labels but no native instance/semantic ground truth. Rather than synthesize labels (which would couple them to depth and bias the IoU we study), we

dropped Replica from the alignment spectrum; SUN RGB-D's four sensors plus ScanNet/Hypersim already span the needed range.

No segmentation-model pseudo-labels on DIODE DIODE has no instance/semantic ground truth. We use it only for a depth-only domain-extension check and explicitly avoid fabricating instance labels via a segmentation model, which would introduce an uncontrolled annotation process.

ScanNet color frames were added to avoid an input confound The first ScanNet subset lacked RGB; using depth-as-grayscale would have fed DINOv2 out-of-distribution inputs and degraded *both* baselines. We added the matching color frames so the frozen-vs-adapter comparison is on equal footing.

Boundary of interpretation Per-image AUPRC is same-signed but not significant at $n = 14$, so it is a supporting threshold-integrated check rather than a confirmation of the BF1 trend. The adapter also does not improve all NYU tasks: boundary instance/semantic remain negative even as normals improve. Best-F1 is the primary estimand for the standard linear boundary-probe operating-point question studied here, not a universally preferred metric. MID-LoRA is one alignment-conditioned intervention, while the broader sweep shows early placement is the more robust OOD-preserving regime in this protocol. The module sweep is similarly bounded: `attn_proj` and `mlp_fc2` preserve the source–OOD trade-off better than `mlp_fc1`. Finally, we avoid a "U-curve" interpretation because the high-IoU plateau is near-neutral rather than a re-ascent, and the sensor bins are discrete points rather than a fitted curve.

K Additional Metric and Readout Audits

The September 11 candidate audit evaluates 300 saved prediction files using the protocols in Section 3.3. All comparisons check image, patch and label identity. Three adapter seeds and three probe seeds are kept distinct. The 20 repeated two-fold scene partitions are shared across models and paired target types. Held-out confusion counts are pooled across folds within each repeat; reported utility then averages repeats and probe seeds. Average precision uses complete tied-score groups; per-image averaging excludes zero-positive images with its denominator retained in the audit artifact. No model or head was retrained for this calculation.

The best candidate and positive regret survive all five target metrics and both equal-cell and equal-dataset weighting rules. The sole secondary-order change is seed-42 dataset-balanced piAP, where ALL-4 overtakes 10–19 while Early remains best. These are conditional ranking observations, not a claim of significance from 30 independent experiments. Historical source Bnd/Cop choices are unchanged. The audit script and input SHA-256 manifest will be released with the code. Candidate Q50 recomputation agrees with the recorded utilities; the maximum observed effect of switching strict $>$ to $\geq$ at the 50 thresholds is zero across these 300 files.

Table K36 gives the absolute target scores of the historical pool, so each gap in Table K37 can be read against the unadapted backbone; under held-out F1 the backbone scores 0.4102, Early 0.4190–0.4227, and neither other candidate exceeds 0.4035.

K.1 Complete metric-specific source-selection regrets

Table K39: Complete candidate ranking and regret audit. E, W, and A denote Early, 10–19, and ALL-4; EC/DW mean equal-cell/dataset-weighted.

weight	metric	seed	target order	Bnd regret	Cop regret
EC	Q50	42	$E > W > A$	0.01968	0.02647
EC	Q50	43	$E > W > A$	0.02838	0.02838
EC	Q50	44	$E > W > A$	0.03187	0.01853
EC	exact	42	$E > W > A$	0.01945	0.02642
EC	exact	43	$E > W > A$	0.02838	0.02838
EC	exact	44	$E > W > A$	0.03213	0.01853

weight	metric	seed	target order	Bnd	Cop
EC	held-out	42	$E > W > A$	0.01925	0.02764
EC	held-out	43	$E > W > A$	0.02937	0.02937
EC	held-out	44	$E > W > A$	0.03439	0.01919
EC	piAP	42	$E > W > A$	0.02148	0.02178
EC	piAP	43	$E > W > A$	0.02588	0.02588
EC	piAP	44	$E > W > A$	0.03377	0.02000
EC	gAP	42	$E > W > A$	0.02869	0.03472
EC	gAP	43	$E > W > A$	0.03832	0.03832
EC	gAP	44	$E > W > A$	0.03972	0.02504
DW	Q50	42	$E > W > A$	0.01950	0.02212
DW	Q50	43	$E > W > A$	0.02542	0.02542
DW	Q50	44	$E > W > A$	0.03062	0.01967
DW	exact	42	$E > W > A$	0.01926	0.02197
DW	exact	43	$E > W > A$	0.02551	0.02551
DW	exact	44	$E > W > A$	0.03088	0.01966
DW	held-out	42	$E > W > A$	0.01919	0.02254
DW	held-out	43	$E > W > A$	0.02576	0.02576
DW	held-out	44	$E > W > A$	0.03218	0.01982
DW	piAP	42	$E > A > W$	0.02006	0.01836
DW	piAP	43	$E > W > A$	0.02240	0.02240
DW	piAP	44	$E > W > A$	0.03140	0.02023
DW	gAP	42	$E > W > A$	0.03062	0.03116
DW	gAP	43	$E > W > A$	0.03487	0.03487
DW	gAP	44	$E > W > A$	0.04073	0.02689

K.2 Historical fixed-head and threshold-repair tables

The following tables concern the older SUN compatibility audit, not the new scene-held-out candidate protocol. The fixed-direction measurements do not identify optimal linear separability. The historical threshold-repair ratios are operational comparisons; the precise legacy generation script could not be recovered, so they are retained as historical supporting evidence rather than an independently reproduced causal decomposition.

L Preregistered COCO study

Preregistration. The study plan was written and hashed on 2026-09-25 at 17:18 CST, before any COCO image, label or result was produced (SHA-256 `e97df6f308018de3ea2feee7-26ea11546dc541e2d290f7f0a5b44c592de9d7c9`). It fixes the target set, the label rules, the 28 models, the probe recipe, the metrics, the two λ rules with their choices (R1: $\lambda = 0.1$; R2: $\lambda = 1$) and the criteria for supporting or refuting each outcome. No threshold, cell, split, metric or rule changed after the first COCO result.

Target set and labels. COCO-2017 `val2017` with panoptic annotations. We keep every image with width at least its height (the shared letterbox routine only supports landscape inputs), giving 3,772 images, split 80/20 at the image level with the SUN RGB-D procedure (3,018 training and 754 test images). Every panoptic segment, thing or stuff, is one instance region; category ids define the semantic cell. Segment ids reach 1.67×10^7, so each image is re-indexed densely before resizing, and instance boundaries compare full segment ids instead of the packed $\mathrm{label} \times 256 + \mathrm{instance}$ key used for the older datasets. A patch is positive when its boundary-pixel fraction exceeds 0.10, the value used for every real-image dataset. Positive rates are 0.153 (instance) and 0.148 (semantic).

Label audit. Train and test ids are disjoint and unique; the two cells share one validity mask; no patch is semantic-positive without being instance-positive. An independent reimplementation recomputed images, labels and masks for 40 random test images and matched the cache with zero differing entries. A visual overlay of boundary patches was inspected before any model was evaluated.

Table K36: Mean target scores in the historical candidate pool under three metrics. Equal-cell means over the ten cells, three probe seeds per cell, historical probe recipe. The unadapted backbone has no adapter seed. Bold marks the best adapted candidate within each adapter training.

adapter seed	candidate	held-out F1	per-image AP	pooled AP
—	unadapted	0.4102	0.3573	0.3463
42	Early	**0.4204**	**0.3725**	**0.3600**
42	10–19	0.4012	0.3511	0.3313
42	ALL-4	0.3928	0.3508	0.3252
43	Early	**0.4190**	**0.3729**	**0.3567**
43	10–19	0.4030	0.3487	0.3343
43	ALL-4	0.3896	0.3470	0.3184
44	Early	**0.4227**	**0.3751**	**0.3619**
44	10–19	0.4035	0.3551	0.3369
44	ALL-4	0.3883	0.3414	0.3222

Table K37: Selection cost in the historical pool. Range over three adapter trainings and two source rules of the target-score gap between the selected and the best candidate. Early is the best candidate in every row.

target metric	equal-cell	dataset-balanced
Oracle F1 (50 quantiles)	0.0185–0.0319	0.0195–0.0306
Exact oracle F1	0.0185–0.0321	0.0193–0.0309
Scene-held-out F1	0.0192–0.0344	0.0192–0.0322
Per-image AP	0.0200–0.0338	0.0184–0.0314
Pooled AP	0.0250–0.0397	0.0269–0.0407

Evaluation and reproducibility. The six new adapters ($\lambda \in \{0.1, 1, 3\}$, seeds 43 and 44) were trained with the command of the $\lambda = 0.3$ seed-43/44 adapters, changing only the weight and seed. All 28 models used the matched probe recipe with cached features. A second, independently scheduled run of the same evaluation completed 45 of the 56 jobs before it was stopped; its 135 logit files are bitwise identical to ours.

Analysis. Held-out F1 uses the 20 fixed two-fold splits of the main analysis with each image as its own group; thresholds are refit inside each of 1,000 paired image resamples shared by every model, cell and probe seed. Point metrics were checked file by file against the original point engine. While reviewing the analysis code, before its output was read, we found that our first implementation judged O4 on interval lower bounds, which is stricter than the preregistered point-estimate criterion; we restored the preregistered rule. O4 is supported under both.

M MATCHED UNPROTECTED ADAPTERS

Two further ALL-4 adapters (seeds 43 and 44, trained independently for the historical pool) were probed with the matched recipe on the ten cells, which gives an unprotected adapter for every preserved seed. Table M43 lists paired scene-resampling intervals.

Oracle F1 gives the same picture: the unprotected adapters lose 0.0151, 0.0203 and 0.0197, the preserved adapters change it by $+0.0004$, -0.0009 and -0.0024. Against the smallest learning rate, the preserved advantage depends on run and weighting: with equal cell weights the held-out F1 interval lies above zero only for run 42, and with datasets weighted equally all such intervals include zero.

Table K38: Historical pool with one dataset left out. Gap in oracle F1 between each non-Early candidate and Early (the best candidate in every slice), for adapter seeds 42/43/44, and how often each candidate is runner-up over the 15 metric–seed combinations of a slice.

cells used	10–19 behind Early	ALL-4 behind Early	runner-up
all ten	0.0197/0.0158/0.0185	0.0265/0.0284/0.0319	10–19 in 15/15
without Hypersim	0.0180/0.0181/0.0173	0.0286/0.0289/0.0311	10–19 in 15/15
without SUN	0.0193/0.0110/0.0211	0.0167/0.0217/0.0291	10–19 in 9/15, ALL-4 in 6/15
without ScanNet	0.0215/0.0160/0.0184	0.0292/0.0311/0.0340	10–19 in 15/15

Table K40: Class-conditional logit shift in the controlled SUN audit (same head, same patches, backbone swapped) across three sensors spanning the alignment gradient. Δ^+/Δ^- are the mean paired logit shift on positive/negative patches; the differential $\Delta^+ - \Delta^-$ measures the margin-alteration component. ALL-4 depresses positive logits while negative logits barely move, ruling out a uniform calibration offset in this clean SUN audit. MID-LORA restores the margin only in the aligned regime, matching its best-F1 fail/partial/rescue pattern. N^+ is the number of positive patches. The parenthesized IoU values are historical annotations whose exact subset provenance could not be recovered (xtion is 0.074 in Table B3). They do not enter the logit calculation; use the main-board values for alignment analysis.

sensor (IoU), N^+	config − frozen	Δ^+ logit	Δ^- logit	differential
kv1 (0.035), 1779	ALL-4	−0.271	−0.026	−0.245
	MID-LORA	−0.318	−0.026	−0.292
xtion (0.078), 1752	ALL-4	−0.159	+0.033	−0.192
	MID-LORA	−0.138	+0.028	−0.166
kv2 (0.114), 2483	ALL-4	−0.202	−0.023	−0.178
	MID-LORA	−0.016	+0.000	−0.016

N PENALTY WEIGHT ON ALL TEN CELLS

O SOURCE BOUNDARY SCORES ACROSS ALL THREE RUNS

A post hoc rerun rescored the saved source boundary probes of all twelve window adapters. Table O45 gives the share of 10,000 joint resamples of the nine validation scenes in which window 12–17 wins; it wins at the point estimate under every scoring and in every run. The fixed-threshold coplanarity score picks 6–11, 6–11 and 18–23 (shares 63%, 97%, 89%). This rerun uses a newer runtime; counts differ slightly from the original runtime, so we do not claim bitwise reproduction across runtimes.

P PER-CELL RECOVERY, FRESH HYPERSIM SCENES AND A LABEL-LOADER FIX

Fresh Hypersim scenes. We also evaluated the preserved and matched unprotected adapters, with frozen probes and thresholds, on 75 frames from three Hypersim sub-collections not in the project's

Table K41: Best-F1 drop is not removed by calibration repair. Operational counterfactual for the ALL-4−frozen best-F1 gap. The "residual" column is the gap remaining after re-optimizing the threshold over 10 held-out/eval splits per sensor. Values above 100% are possible because this is an operational ratio for nonlinear F1, not an additive variance decomposition. Re-optimizing the threshold removes $\leq 6\%$ of the gap. A single positive temperature is order-preserving and therefore cannot change oracle best-F1, so no temperature column is reported.

sensor	total ΔBF1 gap	residual after thr. repair	threshold-recovered
kv1	−0.028	−0.029 (105%)	+0.001 ($\sim$ 0%)
xtion	−0.029	−0.029 (100%)	+0.000 ($\sim$ 0%)
kv2	−0.015	−0.014 (94%)	−0.001 (6%)

Table L42: COCO: held-out F1 of each model minus the unadapted backbone (0.5298), and preserved models minus the same-seed ALL-4, with 95% paired image-resampling intervals.

model	run 42	run 43	run 44
window 0–5	+.0082 [.0069, .0092]	+.0157 [.0139, .0176]	+.0083 [.0071, .0096]
window 6–11	+.0075 [.0050, .0100]	+.0115 [.0088, .0141]	+.0107 [.0077, .0135]
window 12–17	+.0017 [−.0006, .0038]	−.0004 [−.0027, .0020]	−.0026 [−.0052, .0000]
window 18–23	−.0022 [−.0041, −.0001]	−.0030 [−.0050, −.0010]	−.0014 [−.0034, .0005]
ALL-4	−.0009 [−.0033, .0012]	−.0044 [−.0068, −.0021]	−.0040 [−.0066, −.0017]
$\lambda = 0.1$ minus ALL-4	+.0018 [−.0003, .0038]	+.0053 [.0033, .0074]	+.0051 [.0033, .0072]
$\lambda = 0.3$ minus ALL-4	+.0018 [−.0003, .0043]	+.0049 [.0029, .0073]	+.0052 [.0030, .0076]
$\lambda = 1$ minus ALL-4	+.0010 [−.0011, .0035]	+.0045 [.0023, .0069]	+.0043 [.0021, .0069]
$\lambda = 3$ minus ALL-4	+.0008 [−.0013, .0033]	+.0043 [.0021, .0067]	+.0042 [.0019, .0067]

Table M43: Ten cells, paired 95% scene-resampling intervals (1,000 draws for held-out F1 and pooled AP). Unprotected ALL-4 and preserved ($\lambda = 0.3$) adapters of the same seed. With datasets weighted equally, every preserved-minus-unprotected interval also lies above zero.

metric	run	unprotected minus backbone	preserved minus backbone	preserved minus unprotected
held-out F1	42	−.0173 [−.0266, −.0074]	+.0001 [−.0042, .0043]	+.0174 [.0075, .0269]
	43	−.0212 [−.0313, −.0109]	−.0015 [−.0050, .0031]	+.0197 [.0108, .0301]
	44	−.0224 [−.0322, −.0108]	−.0036 [−.0071, .0011]	+.0188 [.0083, .0294]
pooled AP	42	−.0210 [−.0313, −.0092]	−.0009 [−.0046, .0032]	+.0202 [.0096, .0291]
	43	−.0281 [−.0386, −.0148]	−.0010 [−.0043, .0028]	+.0271 [.0153, .0367]
	44	−.0240 [−.0350, −.0119]	−.0036 [−.0068, .0002]	+.0203 [.0082, .0306]

inventory. Averaged over seeds, preservation improves instance F1 by 0.8 points and pooled AP by 0.7 points over ALL-4, but seed 43 is negative for F1 and seeds 43 and 44 are negative for AP. With three scenes this check is descriptive only.

Label loader. The historical Hypersim loader cast instance ids to 8 bits and compared a packed $\text{semantic} \times 256 + \text{instance}$ key. Recomputing all 3,850 frames with full-precision tuple comparison leaves every patch label and mask of the 2,900 training and 875 test frames unchanged, so no reported result changes; one patch of one fresh frame changes. The selected SUN RGB-D entries show no collisions, and ScanNet's stored instance images are 8-bit, so their ids are within range by construction.

Q EARLY-OVER-LATE MARGINS ON THE OLD TARGETS

Table N44: Adapter seed 42, ten cells weighted equally, three probe seeds: change against the unadapted backbone. The weights were already inspected on six of these cells, so this is a sensitivity analysis, not a validation.

λ	oracle F1	exact F1	held-out F1	per-image AP	pooled AP
0	$-.0151$	$-.0151$	$-.0173$	$-.0070$	$-.0210$
0.1	$-.0068$	$-.0069$	$-.0085$	$-.0029$	$-.0091$
0.3	$+.0004$	$+.0002$	$+.0001$	$+.0001$	$-.0009$
1	$+.0008$	$+.0007$	$+.0008$	$+.0011$	$+.0009$
3	$+.0001$	$+.0000$	$-.0002$	$+.0004$	$+.0006$

Table O45: Selection share of window 12–17 under four source boundary scores.

source score	run 42	run 43	run 44
average precision	96.8%	93.8%	82.6%
best F1, all thresholds	100.0%	99.4%	96.1%
best F1, 50 quantiles	100.0%	99.3%	96.1%
F1 at fixed threshold	88.1%	81.0%	99.2%

Table P46: Where preservation recovers the loss. Oracle F1 minus the unadapted backbone in each of the ten cells, three probe seeds per entry, for the single unprotected all-layer adapter and three independently trained adapters with $\lambda = 0.3$; -0.0000 marks a value that is negative before rounding.

cell	ALL-4 ($\lambda = 0$)	$\lambda = 0.3$, seed 42	seed 43	seed 44
Hypersim instance	$+0.0048$	$+0.0048$	$+0.0049$	-0.0015
Hypersim semantic	-0.0021	$+0.0015$	$+0.0032$	-0.0018
ScanNet instance	-0.0071	-0.0000	-0.0001	$+0.0005$
ScanNet semantic	-0.0047	$+0.0001$	-0.0001	$+0.0007$
SUN kv1 instance	-0.0295	-0.0034	-0.0039	-0.0073
SUN kv1 semantic	-0.0289	-0.0053	-0.0057	-0.0085
SUN xtion instance	-0.0281	$+0.0001$	-0.0040	-0.0004
SUN xtion semantic	-0.0218	$+0.0003$	-0.0052	-0.0028
SUN kv2 instance	-0.0155	$+0.0026$	$+0.0020$	-0.0019
SUN kv2 semantic	-0.0178	$+0.0029$	$+0.0002$	-0.0013
ten-cell mean	-0.0151	$+0.0004$	-0.0009	-0.0024

Table Q47: Early windows beat late windows. For each training run, the margin is the lower early-window score minus the higher late-window score. Entries give its minimum over the three runs, after averaging three probe seeds. Left: five metrics, averaged over the ten cells (ten-cell) or over datasets weighted equally (datasets). Right: oracle F1 in individual instance (I) and semantic (S) cells, where negative margins mark local exceptions.

mean over cells			single cells, oracle F1		
metric	ten-cell	datasets	dataset or sensor	I	S
oracle F1	.0124	.0091	SUN kv1	$+.0101$	$+.0089$
exact F1	.0124	.0090	SUN xtion	$+.0078$	$+.0081$
scene-held-out F1	.0128	.0091	SUN kv2	$-.0023$	$-.0019$
per-image AP	.0071	.0080	Hypersim	$-.0112$	$+.0040$
pooled AP	.0091	.0104	ScanNet	$+.0021$	$+.0009$