BEYOND SOURCE-DOMAIN ACCURACY: CONTROLLED STUDY ON ADAPTATION GRANULARITY FOR IMAGING SONAR REPRESENTATION TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Transferring pre-trained vision foundation models to imaging sonar tasks has become a prevalent solution for underwater perception with limited data. Existing studies generally follow the empirical assumption that higher source-domain accuracy yields better cross-domain transfer performance, and thus focus primarily on improving source-domain capabilities. However, this common belief lacks systematic controlled validation, and how adaptation granularity governs sonar transfer performance remains unclear. To address this gap, this paper constructs a controlled experimental framework based on a 576M-parameter visual encoder. We systematically evaluate six feature adaptation schemes covering token-level scalar adjustment, channel-level transformation, and global affine mapping, from both source-domain classification accuracy and sonar cross-domain transfer perspectives. Experimental results show that fine-grained adaptation with stronger representation capacity improves source-domain accuracy yet degrades sonar transfer performance in the focal comparison. In contrast, coarse-grained scalar adaptation achieves a transfer accuracy of 96.5%, outperforming source-domain fine-tuning (95.1%) with stable and reproducible gains. A matrix of three backbones and three source domains, all evaluated on the same sonar target, further shows that the preferred granularity changes with the setting. Ablation studies verify that such benefits arise from content-dependent adaptation rather than simple feature aggregation with frozen encoders. Under identical training budgets, comprehensive comparisons with sonar masked autoencoders, LoRA, and drift-anchored fine-tuning further clarify the applicability of our findings. This paper demonstrates that adaptation granularity is a long-overlooked critical dimension in sonar cross-modal transfer. Relying solely on source-domain accuracy leads to misleading evaluation of adaptation strategies. Our controlled experiments clearly characterize the potential and limitations of vision-to-sonar transfer, providing practical guidance for underwater cross-modal perception adaptation design.

1 INTRODUCTION

The success of foundation-model adaptation is usually measured on the task used for fine-tuning. In many applications, however, the adapted encoder is subsequently reused on another dataset, sensor, or label space. This creates a second objective: learning a representation that remains useful beyond the source task. A modification that improves source accuracy may therefore be an unfavorable choice for representation transfer. Understanding which decisions cause these rankings to diverge matters when labels are scarce and repeated backbone training is expensive.

One such decision is the granularity of feature correction. A lightweight branch can reweight a token with a single scalar, assign separate gains to channel groups, or transform each channel independently. Finer correction expands the branch's function class and can better fit source-specific feature statistics; its effect on transfer is less clear. A scalar preserves each token's direction, while a channel-wise gain can rotate it. Yet when the backbone and correction are trained jointly, even a scalar branch can influence the features learned upstream. Output expressivity alone is therefore an incomplete description of the adaptation process.

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.

Existing research establishes that fine-tuning can distort pretrained features and compromise performance under distribution shift (Kumar et al., 2022; Wortsman et al., 2022). We study a more specific question: *does increasing the granularity of a jointly trained feature correction improve the representation available to a new target-task readout?* The distinction matters because a fresh target classifier can absorb some changes that would harm the original source classifier. We evaluate the adapted representation with a frozen encoder and a new linear probe, separating target representation quality from direct reuse of the source decision boundary.

If source accuracy were a reliable guide to transfer, the finest corrections should transfer best. Imaging sonar shows otherwise. We construct a bounded correction family—scalar, grouped, and channel-wise variants sharing input features, gain range, initialization, and training protocol, with a single variable controlling channel sharing. On a 576M Qwen visual backbone adapted to sonar, every finer or affine variant exceeds the scalar arm's source accuracy yet falls short of its target accuracy: scalar correction yields 96.54% target accuracy versus 95.45% for channel correction at the same training seed. The complete scalar pipeline also improves over plain source fine-tuning by 1.67 percentage points across three paired training seeds.

We then test how far this separation extends. A matrix of three backbones and three source domains shows that the preferred granularity changes with the setting: Qwen favors scalar on sonar and channel on EuroSAT, while DINOv2 exhibits the opposite mean ordering, and several cells differ by margins comparable to their seed variability. Finally, controlled regularization experiments ask whether the scalar benefit can be reproduced by restraining a channel branch; neither gain-magnitude penalties nor strong suppression of feature rotation succeeds. These failures localize rather than merely qualify the result: the benefit is not reconstructable from the branch's output behavior alone, pointing to the joint optimization of backbone and correction as the operative ingredient.

The paper makes three contributions:

- A bounded feature-correction framework that varies token-to-channel granularity within one parameterization and supports controlled comparisons during full-backbone adaptation.
- An empirical characterization of source–target ranking differences across three backbones and three source domains, including a three-seed whole-pipeline benefit on sonar and the limits of a general scalar preference.
- A focused analysis of correction geometry and output regularization showing that two natural constraints do not reproduce the scalar benefit, motivating closer study of joint backbone–correction optimization.

All matrix cells share one sonar target task. Three levels of evidence support these claims, and we grade them throughout: the three-seed paired pipeline benefit is the strongest result, the focal granularity ordering rests on single-seed controlled contrasts, and matrix cells vary in seed coverage. Our conclusions concern representation transfer into this task family.

2 Related Work

Fine-tuning and transfer. Fine-tuning can improve source performance while reducing robustness through feature distortion: linear probing before fine-tuning mitigates this (Kumar et al., 2022), weight interpolation improves the accuracy–robustness trade-off (Wortsman et al., 2022), and robust few-shot adaptation recipes continue this line (Wang et al., 2025). Other work identifies strong source–target accuracy correlations within particular evaluation families (Miller et al., 2021) and emphasizes the role of model selection in domain generalization (Gulrajani & Lopez-Paz, 2021). We examine a different transfer protocol: target labels train a fresh readout of the adapted representation. Our variable of interest is the correction branch's granularity within joint source adaptation.

Capacity and feature adaptation. Adapters, LoRA, and bias-only fine-tuning control adaptation by restricting trainable parameters (Houlsby et al., 2019; Hu et al., 2022; Zaken et al., 2022; Prottasha et al., 2025). Our family instead controls channel sharing in a feature correction while allowing the backbone to update. Its scalar endpoint acts as token reweighting, whereas channel and affine endpoints can change token directions. This distinction separates the output freedom of a small

branch from the capacity of the full model: a small correction head constrains the head's own operation, not the function class of the entire adapted encoder, whose backbone remains fully trainable. We compare these endpoints within the same family and use a wider scalar head to assess whether output granularity alone accounts for the observed result.

Visual representations for sonar. Vision transformers, the DINOv2 and DINOv3 self-supervised families, and masked autoencoders provide complementary starting points for representation learning (Dosovitskiy et al., 2021; Oquab et al., 2024; Siméoni et al., 2025; He et al., 2022). Recent sonar work spans self-supervision with backbone-size-dependent effects, physics-informed self-distillation, zero-shot-searched detection transformers, RGB–sonar tracking, adaptation of vision foundation models to acoustics, and linear probing for underwater acoustic recognition (Kwon et al., 2026; Hamoda et al., 2026; Gu et al., 2025; Li et al., 2025; Zhang et al., 2026; Hummel et al., 2026). We study a Qwen vision–language visual tower (Bai et al., 2025) alongside two other backbones, using public data including UATD and Marine Debris FLS (Xie et al., 2022; Valdenegro-Toro et al., 2025). The focus is how correction design changes transfer, rather than a new acoustic pretraining objective.

3 METHOD

3.1 SOURCE ADAPTATION AND TARGET REPRESENTATION TRANSFER

We formalize the two-stage protocol shared by all comparisons: joint source adaptation, then frozen-encoder target probing. Let $\mathcal{D}_s$ and $\mathcal{D}_t$ denote labeled source and target datasets, potentially with different label spaces. An encoder F_θ produces N tokens of width D. A correction C_ϕ transforms these tokens before mean pooling, yielding a representation $v_{\theta,\phi}(x)$. During source adaptation, the encoder, correction, and source classifier h_ψ are trained jointly:

$$\min_{\theta,\phi,\psi} \mathbb{E}_{(x,y)\sim\mathcal{D}_s} \left[\ell_{\mathrm{CE}}(h_\psi(v_{\theta,\phi}(x)), y)\right]. \tag{1}$$

For target evaluation, θ and ϕ are frozen and a new linear classifier is fitted using target training labels. Figure 1 summarizes the two stages. This protocol measures how well target information can be read from the source-adapted representation. It does not require the source and target to share classes.

3.2 BOUNDED CORRECTION WITH VARIABLE GRANULARITY

For token i, let $f_i^c \in \mathbb{R}^D$ be its feature after the common background-subtraction step, and let $p_i \in [-1, 1]^2$ be its normalized spatial coordinate. The correction uses a normalized token $u_i = \mathrm{LN}(f_i^c)$ and a content-and-coordinate prediction branch:

$$t_i = \mathrm{GELU}(W_h[f_i^c; p_i] + b_h), \qquad a_i = W_G t_i + c_G \in \mathbb{R}^G. \tag{2}$$

The hidden width is 64. Channels are partitioned into G groups, and the branch predicts one positive gain per group:

$$m_i = \exp(b \tanh(a_i)), \quad b = \log 8, \qquad \bar{f}_i = m_i \odot u_i, \qquad v_{\theta,\phi}(x) = \frac{1}{N}\sum_{i=1}^{N} \bar{f}_i. \tag{3}$$

Each gain is broadcast over its channel group. The bounded exponential permits both suppression and amplification, with $1/8 < m_i < 8$; the bound was fixed across all experiments and arms. Zero-initializing the output projection gives unit gains at initialization. The initial correction is therefore an identity on u_i; the preceding normalization and subtraction remain part of the shared pipeline.

The three principal variants are:

- **Scalar ($G = 1$):** one gain per token, shared across all feature channels. Each token retains its direction, and the gain controls its contribution to pooling.
- **Grouped ($G = 8$):** independent gains for eight channel blocks, allowing different feature subspaces to be rescaled separately.

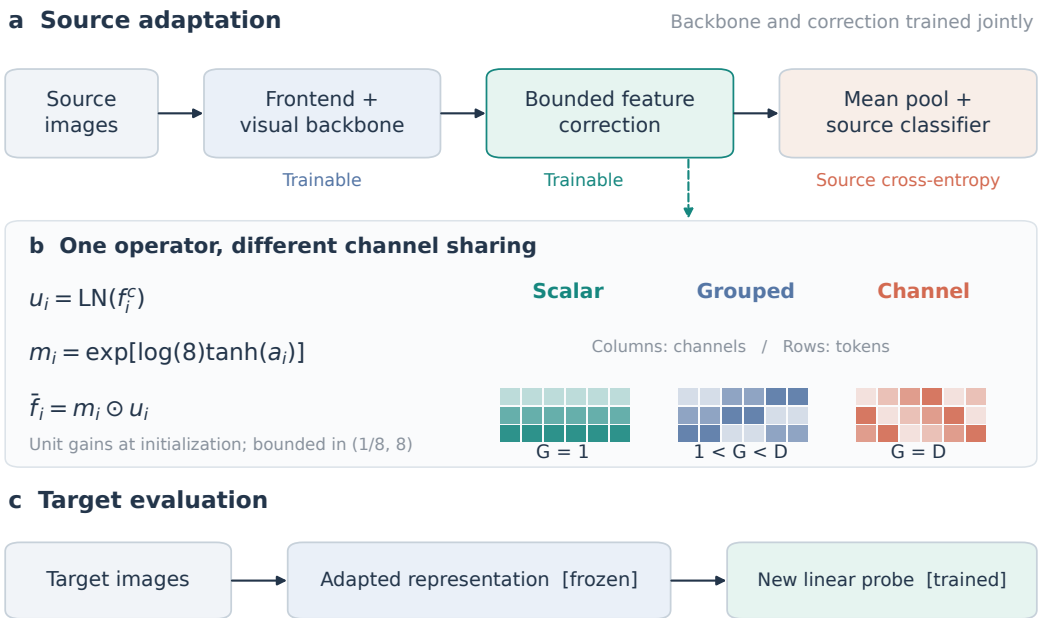

Figure 1: **Bounded feature correction.** (a) The backbone and correction are jointly adapted using source labels. (b) The same bounded gain operator ranges from token-scalar to channel-wise correction by changing which channels share a multiplier. Colored grids are schematic gain matrices. (c) The adapted representation is frozen for evaluation with a new target linear probe.

- **Channel** ($G = D$)**:** independent gains for every channel, allowing both the magnitude and direction of a token to change.

The focal sonar study also evaluates $G = 2, 32$ and an affine extension,

$$\bar{f}_i^{\mathrm{aff}} = m_i \odot u_i + \tanh(W_d t_i + d), \tag{4}$$

whose offset projection is zero-initialized. For compatible group partitions, finer correction classes contain coarser ones by tying the corresponding outputs. This is a statement about the correction at fixed input features; joint training also changes the backbone.

3.3 Controlling capacity and training the representation

All granularity arms within a setting share the frontend, token preprocessing, hidden width, gain bounds, initialization, and source-training schedule. On the Qwen tower, scalar, channel, and affine heads contain approximately 74K, 149K, and 224K parameters. Because changing G also changes parameter count and optimization, we additionally widen the scalar head to approximately 150K parameters. This control varies head capacity without untying channel gains.

In the Qwen sonar pipeline, the frontend combines a gated residual convolutional stem and a positional pathway derived from range and azimuth proxies. Together with the correction, the new components contain approximately 8.1M parameters relative to the 576.4M backbone. A classifier warm-up with the encoder frozen is followed by full joint fine-tuning. All learned corrections act on tokens immediately before final mean pooling; the correction changes what pooling aggregates, not how tokens are combined. At evaluation, the source classifier is discarded and frozen representations are assessed with the same linear-probe procedure across all arms. The experiment-specific backbone and dataset configurations are detailed in Section 4.1.

4 EXPERIMENTS

Our experiments address three questions: how correction granularity affects the focal sonar result, whether its ordering extends across backbones and source domains, and whether output constraints reproduce the scalar benefit.

4.1 EXPERIMENTAL SETUP

Backbones and datasets. We evaluate the 576.4M Qwen visual tower, self-supervised DINOv2-L (approximately 300M parameters), and supervised ViT-B/16 (approximately 86M parameters). Their feature widths are 1152, 1024, and 768. Source adaptation uses one of three datasets: a 28-class sonar corpus with 61,442 frames and sequence-level train/validation/test sizes of 48,356/6,466/6,620; EuroSAT (Helber et al., 2019), with ten classes and a fixed stratified split of 18,900/4,050/4,050 images; or DTD (Cimpoi et al., 2014), with 47 classes and official split 1 (1,880 images per partition).

All cells transfer to the Marine Debris FLS turntable task: 12 classes acquired with an ARIS Explorer 3000 sonar. The target probe split is 1,280/225/645 images; validation is a stratified 15% of the official training partition, selected with seed 0. Inputs use 448×448 resolution and ImageNet normalization; grayscale images are replicated across three channels. The target encoder is frozen, and target labels train only a new probe. Thus the matrix varies source domains while keeping the target task fixed.

Implementation. Source training uses 200 classifier warm-up steps at learning rate 10^{-3}, followed by three full-backbone epochs at 10^{-5} for the backbone and 10^{-4} for new modules and the classifier. We use AdamW, weight decay 0.01, a cosine schedule, 50 warm-up steps, gradient clipping at 1.0, and nominal batch size 16. Probe features are standardized using training statistics; probes use AdamW at 10^{-3} for 300 epochs, weight decay 0.01, a cosine schedule, and batch size 4,096. The best probe epoch is selected by validation accuracy. Hyperparameters are shared across correction arms rather than separately tuned for each baseline. Three source epochs also imply different update counts across datasets (9,066 sonar, 3,543 EuroSAT, 352 DTD; Appendix B), so equal epochs are not equal compute. Further details appear in Appendix B.

Metrics and seeds. We report source and target linear-probe (LP) accuracy. Each run averages probe seeds 42, 123, and 2024. Most correction arms also use these three *training* seeds. Qwen–sonar grouped and channel arms have only training seed 42; its scalar and plain-fine-tuning arms have all three. DINOv2–sonar scalar/channel matrix arms have seeds 123 and 2024, with earlier seed-42 runs listed separately. Plain-fine-tuning references elsewhere generally use one training seed. Figures distinguish paired training-seed values from probe variability; Appendix A provides all available runs. Probe-seed variation does not estimate training-seed uncertainty: we report paired contrasts where possible and do not treat repeated predictions of the same target images as independent observations for a pooled significance test.

4.2 MAIN RESULTS ON SONAR TRANSFER

A scalar bottleneck yields the best focal transfer. Figure 2a plots source accuracy directly against target accuracy. Scalar correction reaches 96.54% target LP, compared with 95.45% for channel correction at the same training seed. Across the extended ladder, grouped and affine corrections obtain 78.7–79.0% source LP versus 78.5% for scalar, but only 94.5–95.6% target LP versus 96.5%. The additional output freedom improves source fit in this comparison without improving the target representation. The ordering is not monotonic with group count.

The complete pipeline improves across training seeds. Relative to plain source fine-tuning, the scalar pipeline gains $+1.45$, $+0.98$, and $+2.59$ percentage points for training seeds 42, 123, and 2024 (Figure 2b). Mean target accuracy increases from 94.28% to 95.95%, a gain of 1.67 points. This is the replicated system-level result. The scalar–channel comparison has one paired training seed, so it supplies narrower evidence about granularity itself: the two comparisons answer different questions, one concerning the head's output freedom at a matched seed, the other the complete system against an established baseline.

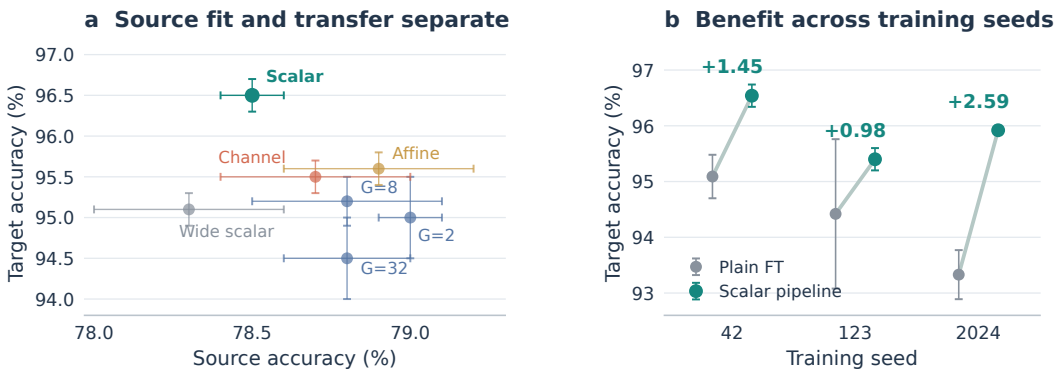

Figure 2: **The focal Qwen–sonar result.** (a) Source and target LP accuracy for the extended correction family at training seed 42; error bars are probe-seed standard deviations. Scalar has the highest target accuracy although finer variants have higher source means. (b) The complete scalar pipeline improves over plain fine-tuning in all three paired training seeds. This second comparison includes the frontend and does not isolate granularity.

Table 1: Qwen–sonar target LP accuracy. Each entry averages three probe seeds; $\pm$ is their standard deviation. The final row averages the three training-seed means. The comparison includes the complete frontend and scalar correction, so it does not isolate the gain head.

Training seed	Plain source FT	Scalar pipeline	Difference
42	95.09 ± 0.39	96.54 ± 0.20	$+1.45$
123	94.42 ± 1.34	95.40 ± 0.20	$+0.98$
2024	93.33 ± 0.44	95.92 ± 0.08	$+2.59$
Mean	94.28	95.95	$+1.67$

Comparison with alternative adaptation routes. At seed 42, plain fine-tuning reaches 95.09% target LP, rank-16 LoRA 94.16%, drift-anchored fine-tuning 93.90%, and bias-only fine-tuning 91.68%. The scalar pipeline reaches 96.54% under the same source schedule. These values compare common-protocol behavior, not separately optimized best performance for each baseline. A from-scratch ViT-S masked autoencoder trained on the same images reaches 76.95%; its different model size, initialization, and training duration make it a separate resource-specific reference (Appendix C).

4.3 Generalization Across Backbones and Source Domains

Mean preferences reverse across sources. Figure 3 summarizes all nine backbone–source settings. Qwen favors scalar on sonar ($+1.09$ points at the matched seed) and channel on EuroSAT (-0.43 points averaged over three pairs). DINOv2 shows the opposite source-domain pattern: channel leads on sonar (-0.78), while scalar leads on EuroSAT ($+0.78$). Thus the preferred granularity cannot be inferred from the backbone alone or transferred directly from the initial sonar comparison.

The paired values clarify the strength of these observations. Both DINOv2–sonar matrix seeds favor channel. DINOv2–EuroSAT gives scalar–channel differences of -0.31, $+1.29$, and $+1.35$, so its mean preference is sensitive to training randomness. Qwen–sonar has the largest positive mean but only one paired granularity seed. These distinctions motivate reporting the individual contrasts alongside the matrix rather than interpreting every cell mean as a resolved ordering.

Some settings offer little separation. On DTD, absolute scalar–channel mean differences are at most 0.33 points across backbones. ViT-B correction-arm means lie approximately between 92.25% and 92.90%, close to its frozen-feature accuracy of 92.71%. Several arms numerically exceed the frozen reference, but the net changes are small. Similar target accuracy does not imply that their representations are unchanged; it establishes only limited separation under this probe and protocol,

Table 2: Source (src) and target (ext) linear-probe accuracy (%). Means use the available training seeds, each averaged over three probe seeds. Bold marks the highest target mean among adapted arms and does not imply statistical significance. Δ_{s-c} is scalar minus channel target accuracy. †Qwen–sonar uses the matched seed-42 ladder; its scalar three-seed mean is 95.95. ‡DINOv2–sonar scalar/channel have two matrix training seeds and grouped-8 has three; earlier seed-42 references are listed separately in Appendix A.

		Frozen	Plain FT		Scalar		G=8		Channel		
Tower	Domain	ext	src	ext	src	ext	src	ext	src	ext	Δ_{s-c}
Qwen-VL 576M†	Sonar	91.8	77.7	95.1	78.5	**96.5**	78.8	95.2	78.7	95.5	+1.09
	EuroSAT	91.8	97.9	91.2	97.6	93.8	97.6	93.7	97.6	**94.2**	−0.43
	DTD	91.8	80.3	94.1	82.3	94.8	81.9	94.6	82.2	**94.9**	−0.11
DINOv2-L 300M‡	Sonar	–	–	–	80.8	93.3	80.5	93.1	80.5	**94.0**	−0.78
	EuroSAT	89.9	98.2	91.8	98.0	**92.9**	97.9	91.8	98.0	92.1	+0.78
	DTD	89.9	83.1	**90.7**	82.4	89.9	82.7	89.4	82.9	90.3	−0.33
ViT-B/16 86M	Sonar	92.7	77.8	91.7	78.3	92.4	78.1	**92.8**	78.3	92.5	−0.14
	EuroSAT	92.7	96.1	92.4	96.3	92.7	96.2	**92.9**	96.4	92.2	+0.48
	DTD	92.7	67.3	92.8	67.0	92.8	67.0	**92.9**	67.3	92.8	+0.00

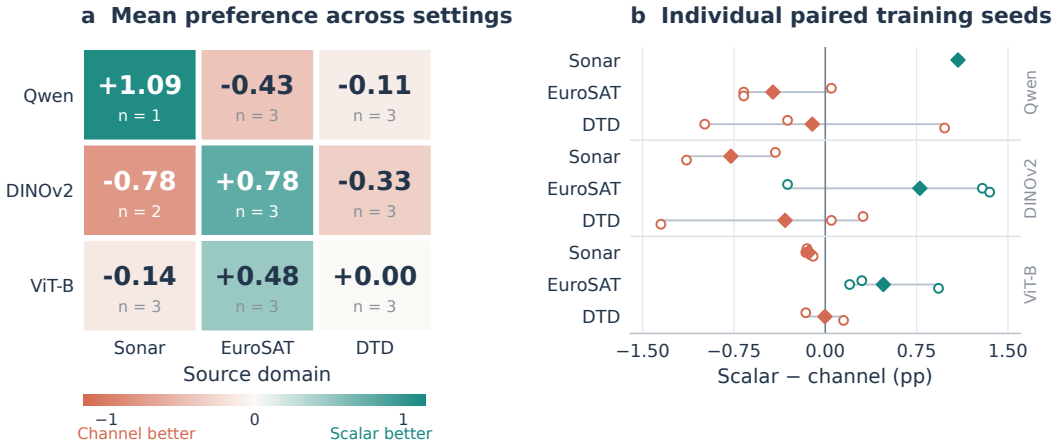

Figure 3: **The preferred granularity changes across settings.** (a) Scalar-minus-channel target accuracy (percentage points); n is the number of paired training seeds. Teal favors scalar and coral favors channel. (b) Individual paired contrasts (circles) and their mean (diamonds); horizontal lines span observed values, not confidence intervals. Values are calculated from the reported rounded seed means. Absolute source and target accuracies are in Table 2.

and our measurements do not include the feature-change intervention that would separate such cases. With 645 target test images, one image corresponds to approximately 0.155 percentage points for an individual run.

Source and target rankings can disagree. On Qwen–sonar, grouped-8 raises source LP from 78.50% to 78.76% while reducing target LP from 96.54% to 95.24%. On DINOv2–sonar, scalar has the higher source mean and channel the higher target mean. These are direct examples of a source–target ranking mismatch. We treat the source test scores as descriptive results; a prospective model-selection study would instead choose on source validation data and measure target regret on a separate test set.

Table 3: Retrained Qwen–sonar controls at training seed 42. Entries are percentages; ± is probe-seed standard deviation. Each arm is trained under the same source protocol.

Configuration	Source LP	Target LP
Frontend only	78.1 ± 0.2	95.4 ± 0.4
Frontend + background subtraction + LN	78.8 ± 0.2	95.6 ± 0.4
Coordinate-only gain	78.4 ± 0.2	94.6 ± 0.3
Feature-only gain	78.7 ± 0.2	95.8 ± 0.3
Unbounded exponential gain	78.6 ± 0.2	94.8 ± 0.2
Bounded scalar gain	78.5 ± 0.1	96.5 ± 0.2

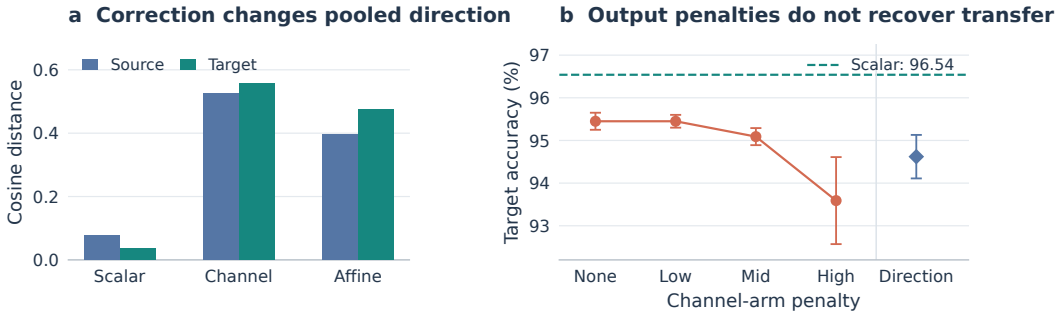

Figure 4: **Large directional changes accompany channel correction, but output penalties do not recover the scalar result.** (a) Mean cosine distance between pooled features before and after correction, measured on 128 source or target frames per domain. (b) Channel-arm target accuracy under gain-logit penalties and a direction penalty; error bars show three-probe-seed standard deviations. All arms use training seed 42. The dashed scalar reference is the same focal checkpoint, not a separate regularization experiment.

4.4 Ablation Study

Retrained controls separate the gain from its input pipeline and output bound. **Content, position, and boundedness.** Table 3 reports retrained seed-42 controls. Coordinate-only gain hurts transfer (94.6%), feature-only gain recovers part of the result (95.8%), and the bounded content-and-coordinate branch is best (96.5%); replacing its bounded gain by a clipped exponential lowers accuracy to 94.8%. These comparisons favor the full bounded design within the tested configuration.

Head capacity at fixed granularity. A wider scalar head with approximately 150K parameters, close to the channel head's size, reaches only 95.1% target LP: keeping scalar outputs does not by itself preserve the result, although the width control changes optimization as well as capacity.

Component effects across seeds. Adding the pipeline to the frontend-only arm changes target accuracy by $+1.14$ points at seed 42 but -1.09 at seed 123, whereas adding the scalar gain to the subtraction-plus-normalization control is positive in both available seeds ($+0.93$ and $+1.60$; Appendix C). The whole-pipeline improvement over plain fine-tuning should therefore not be read as an independent contribution from every added component.

4.5 Analysis of Correction Geometry and Regularization

This section asks what separates scalar from its finer variants, and whether simpler means can reproduce it.

Channel correction changes pooled direction more strongly. Figure 4a shows that channel and affine corrections rotate the pooled feature far more than scalar, whose observed rotation stays small despite unequal token weights; affine has more parameters yet less rotation than channel, so rotation is not monotonic in head size. Channel also has the smallest source-to-target gain-distribution KS statistic (0.061 versus 0.207 for scalar and 0.107 for affine), so marginal gain instability alone does

not explain the ordering, although effects involving the joint distribution of gains and features remain possible.

Penalizing output changes is insufficient. We retrain the channel arm with a gain-logit penalty $\lambda \operatorname{mean}(a_i^2)$, with coefficients calibrated to approximately 3%, 10%, and 30% of source cross-entropy: none improves on the unregularized arm, and the strongest penalty costs 1.86 points (Figure 4b). A pooled-feature direction penalty with coefficient 1.3 reaches 94.62%, 0.83 points below the unregularized arm.

The direction penalty is active: at step 300, its penalized distance is 0.027 versus approximately 0.17 in the unanchored calibration run, falling to 7.6×10^{-4} by step 9,000. Strong suppression of this quantity therefore does not reproduce scalar-level transfer. These experiments establish a failure of the tested remedies; because penalties also modify joint optimization and can remove useful transformations, they do not prove that magnitude or direction is causally irrelevant.

Post-hoc routes do not recover the benefit. When the source-fine-tuned backbone is frozen and only aggregation modules are trained, scalar correction reaches 94.88% and 94.68% for two module seeds, versus 95.09% for mean pooling. Grouped scaling reaches approximately 95.22% in the aggregate screening. None recovers the jointly trained scalar pipeline's 96.54% result. Separately, plain fine-tuning changes backbone features substantially despite per-block weight cosines of at least 0.99999 to initialization: mean token-feature cosine to the initialization teacher is 0.544 at block 16 and 0.880 at the final block (0.535/0.884 under a training-split repetition), while a drift-anchored arm retains more similar features (0.596/0.933) but transfers worse (Appendix D). Weight-space proximity is therefore an insufficient proxy for functional preservation, and closeness to initialization should not be read as evidence that no learning occurred. Together with the regularization failures above, these results locate the operative ingredient in the joint optimization of backbone and correction: what matters is not the correction's output format but that the backbone and this branch learn together. Isolating which component of joint training produces the benefit—for example, by removing or exchanging corrections across backbones—remains open.

4.6 Scope and Limitations

Where target development labels are available, the granularity choice should be validated on a held-out target split, with the target test set reserved for final evaluation: source accuracy alone is an unreliable selector in our experiments, and design decisions should rest on paired multi-seed contrasts. Within the tested family, bounded scalar correction is a strong first candidate for sonar-like transfer under scarce labels, combining the best focal-case transfer with a constrained, interpretable output space; weight- and feature-similarity diagnostics should complement, not replace, target probes. The scope of these conclusions is bounded: all matrix results concern one sonar target task with a relatively small test set, a second scene from the same family would not substitute for independent replication, backbones differ simultaneously in architecture, pretraining, scale, and frontend, shared hyperparameters may favor some configurations, training-seed coverage is unequal, most interventions use one training seed, and the same target benchmark informed exploratory analysis. The evidence supports the observed conditional rankings and the failure of two specific regularizers; it does not establish a universal granularity rule or a causal account of backbone optimization.

5 Conclusion

We studied adaptation granularity through a bounded feature-correction family trained jointly with visual backbones. In the focal sonar setting, scalar correction yields the strongest target representation, and the complete pipeline improves over plain source fine-tuning by 1.67 percentage points across three paired training seeds; the scalar–channel ordering, however, changes across backbones and source domains. Geometry measurements, regularization experiments, and post-hoc aggregation on a frozen backbone show that this benefit cannot be reproduced by constraining or re-creating the branch's outputs after the fact: it lies in the joint optimization of backbone and correction. These findings position correction granularity as a consequential adaptation choice whose value depends on the source, backbone, and target evaluation, best validated on held-out target data rather than inferred from source-task fit.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing parts of the code. We have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on the research methodology or experiments, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and the remaining required-disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work: the AI-assisted code was checked by two authors to ensure its correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study uses public image datasets and involves no human-subject experiment. Sonar perception can support environmental monitoring and underwater inspection and can also have dual-use applications. Our experiments concern research classification benchmarks and do not establish deployment reliability or safety.

REPRODUCIBILITY STATEMENT

Section 3 defines the correction family and adaptation procedure. Section 4.1 specifies datasets, optimization, and seed coverage. Appendix A reports the complete matrix and available training seeds; Appendices B–E provide additional training details, controls, intervention calibration, and data handling. Probe variability and training variability are distinguished throughout.

REFERENCES

S. Bai et al. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. arXiv:2010.11929.

XiaoTong Gu, Shengyu Tang, Yiming Cao, and Changdong Yu. Underwater object detection in sonar imagery with detection transformer and zero-shot neural architecture search. arXiv preprint arXiv:2505.06694, 2025.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

T. Hamoda, H. Rajani, and N. Gracias. BenthicDINO: Physics-informed self-distillation for view-invariant side-scan sonar representations. arXiv:2608.23215, 2026.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2790–2799, 2019.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

H. I. Hummel, S. Bhulai, R. D. van der Mei, and B. Ghani. Linear probing enables ship-radiated noise recognition with pretrained audio embeddings. *Ecological Informatics*, 95:103709, 2026. arXiv:2601.08358.

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*, 2022.

Taeyoun Kwon, Youngwon Choi, Hyeonyu Kim, Myeongkyun Cho, Junhyeok Choi, and Moon Hwan Kim. Mine-JEPA: In-domain self-supervised learning for mine-like object classification in side-scan sonar. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2026. MACVi Workshop, arXiv:2604.00383.

Y. Li, B. Wang, J. Sun, X. Wu, and Y. Li. Rgb-sonar tracking benchmark and spatial cross-attention transformer tracker. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3): 2260–2275, 2025. arXiv:2406.07189.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, 2021.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification, arXiv:2304.07193.

Nusrat Jahan Prottasha, Upama Roy Chowdhury, Shetu Mohanto, Tasfia Nuzhat, Abdullah As Sami, Md Shamol Ali, Md Shohanur Islam Sobuj, Hafijur Raman, Md Kowsher, and Ozlem Ozmen Garibay. PEFT A2Z: Parameter-efficient fine-tuning survey for large language and vision models. arXiv preprint arXiv:2504.14117, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Timothée Darcet, Théo Moutakanni, Piotr Bojanowski, et al. DINOv3. arXiv preprint arXiv:2508.10104, 2025.

Matias Valdenegro-Toro, Deepan Chakravarthi Padmanabhan, Deepak Singh, Bilal Wehbe, and Yvan Petillot. The marine debris forward-looking sonar datasets. arXiv:2503.22880, 2025.

Hanxin Wang, Tian Liu, and Shu Kong. Robust few-shot vision-language model adaptation. arXiv preprint arXiv:2506.04713, 2025.

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*, pp. 7959–7971, 2022.

K. Xie, J. Yang, and K. Qiu. A dataset with multibeam forward-looking sonar for underwater object detection. *Scientific Data*, 9(1):739, 2022. doi: 10.1038/s41597-022-01854-w.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3623–3627, 2022.

K. Zhang et al. Adapting vision foundation models to acoustics for pose-free 3D sonar reconstruction. arXiv:2609.06261, 2026.

A FULL RESULTS AND TRAINING-SEED COVERAGE

Table 4 lists the available matrix runs. Qwen–sonar uses the matched seed-42 ladder; its three-seed plain-fine-tuning comparison is reported in the main text (Table 1). Earlier DINOv2–sonar runs are separate from the matrix means. Original Qwen source LP values use one trained checkpoint per correction arm.

Table 4: Per-seed detail for the 3×3 matrix: external turntable LP and source-domain test LP accuracy (%), each entry the mean over three probe seeds. '–': not run. †Qwen×Sonar entries are the initial sonar runs (source LP of the gain arms uses one training checkpoint per arm). ‡E13 single-seed DINOv2×Sonar runs from the initial sonar experiments, shown for reference and not folded into the cell means of Table 2.

			External LP			Source LP		
Tower	Domain	Arm	s42	s123	s2024	s42	s123	s2024
Qwen-VL 576M	Sonar†	Frozen		91.83		–	–	–
		Plain FT	95.09	–	–	77.70	–	–
		Scalar	96.54	95.40	95.92		78.50	
		G=8	95.24	–	–		78.76	
		Channel	95.45	–	–		78.68	
Qwen-VL 576M	EuroSAT	Frozen		91.83			95.04	
		Plain FT	91.16	–	–	97.94	–	–
		Scalar	93.90	93.80	93.64	97.47	97.62	97.60
		G=8	93.69	93.59	93.75	97.62	97.65	97.67
		Channel	93.85	94.47	94.31	97.55	97.59	97.73
Qwen-VL 576M	DTD	Frozen		91.83			76.28	
		Plain FT	94.11	–	–	80.30	–	–
		Scalar	94.83	94.00	95.45	81.84	82.48	82.43
		G=8	94.62	94.99	94.26	82.00	82.20	81.61
		Channel	95.14	94.99	94.47	82.18	82.43	82.04
DINOv2-L 300M	Sonar	Scalar	–	93.23	93.28	–	80.55	81.06
		G=8	93.90	92.82	92.66	80.95	80.73	79.76
		Channel	–	93.64	94.42	–	80.05	81.05
DINOv2-L 300M	EuroSAT	Frozen		89.87			95.73	
		Plain FT	91.78	–	–	98.23	–	–
		Scalar	92.56	92.09	94.06	97.98	97.96	98.01
		G=8	92.97	91.47	90.96	98.02	97.80	97.97
		Channel	92.87	90.80	92.71	97.95	98.14	98.02
DINOv2-L 300M	DTD	Frozen		89.87			81.86	
		Plain FT	90.70	–	–	83.09	–	–
		Scalar	90.65	90.44	88.73	82.55	82.37	82.27
		G=8	89.46	88.78	89.97	82.95	82.73	82.50
		Channel	90.34	90.39	90.08	82.64	83.05	82.98
DINOv2-L 300M	Sonar‡	Frozen	89.87	–	–	76.60	–	–
		Plain FT	93.02	–	–	80.40	–	–
		Scalar	92.56	–	–	81.01	–	–
		Channel	93.43	–	–	80.44	–	–
ViT-B/16 86M	Sonar	Frozen		92.71			76.37	
		Plain FT	91.68	–	–	77.83	–	–
		Scalar	92.61	91.88	92.56	78.64	78.54	77.82
		G=8	92.97	92.61	92.82	78.13	78.16	78.04
		Channel	92.76	92.04	92.66	78.07	78.46	78.51
ViT-B/16 86M	EuroSAT	Frozen		92.71			95.73	
		Plain FT	92.40	–	–	96.06	–	–
		Scalar	92.45	92.76	92.97	96.48	96.26	96.30
		G=8	92.40	93.33	92.97	96.19	96.35	96.19
		Channel	92.15	92.56	92.04	96.40	96.42	96.33
ViT-B/16 86M	DTD	Frozen		92.71			66.24	
		Plain FT	92.76	–	–	67.29	–	–
		Scalar	92.71	92.81	93.02	66.81	67.29	66.93
		G=8	92.92	92.71	93.02	66.72	67.02	67.20
		Channel	92.87	92.81	92.87	67.18	67.32	67.46

Table 5: Extended Qwen–sonar ladder, training seed 42; uncertainty is over probe seeds. The wide scalar head varies parameter count at fixed output granularity.

Correction	Groups	Source LP (%)	Target LP (%)
Scalar	1	78.5 ± 0.1	96.5 ± 0.2
Grouped	2	79.0 ± 0.1	95.0 ± 0.5
Grouped	8	78.8 ± 0.3	95.2 ± 0.3
Grouped	32	78.8 ± 0.2	94.5 ± 0.5
Channel	1152	78.7 ± 0.3	95.5 ± 0.2
Channel + affine	1152	78.9 ± 0.3	95.6 ± 0.2
Wide scalar (150K parameters)	1	78.3 ± 0.3	95.1 ± 0.2

B Optimization and correction details

All matrix arms use the source training and probe hyperparameters in Section 4.1. At nominal batch size 16, three source epochs correspond to 9,066 sonar updates, 3,543 EuroSAT updates, and 352 DTD updates, excluding classifier warm-up. Equal epoch counts are not equal compute budgets across datasets or backbones. The Qwen sonar adaptation takes approximately 1.5–2 GPU-hours per arm in the original experimental setup; this is a reported runtime, not a hardware-normalized comparison.

The sonar frontend uses cascaded 3×3, depthwise 5×5, and 7×7 convolutional operations behind a zero-initialized residual gate. Its positional pathway adds Fourier features of available range and azimuth proxies to token embeddings. These are image-derived proxies rather than calibrated physical coordinates. Background subtraction and layer normalization are shared by the gain-family arms and retained in the no-gain control. The correction equations specify the varied component; the frontend-only comparison additionally changes this common preprocessing. All learned corrections are applied before final mean pooling.

C Additional controls and adaptation references

Rank-16 LoRA adapts query/value projections (approximately 2.99M trainable parameters). Drift anchoring adds a cosine penalty to features from a frozen initialization teacher at an intermediate and the final block, with coefficient 0.1. Bias-only fine-tuning updates backbone biases. These alternatives share the source protocol but were not individually tuned over learning rate, rank, or penalty strength. The in-domain masked-autoencoder reference trains a ViT-S/16 from scratch for 100 epochs with 75% masking on the same 48,356 sonar training images and the aligned 448×448, three-channel pipeline. It therefore controls the source data and input/probe protocol, not initialization, model scale, or compute.

Table 6: Source-protocol references on Qwen–sonar, except for the separately trained ViT-S MAE. One training seed; $\pm$ reflects three probe seeds.

Route	Source LP (%)	Target LP (%)
Plain source fine-tuning	77.7 ± 0.2	95.09 ± 0.39
Bias-only fine-tuning	74.4 ± 0.1	91.68 ± 0.72
Rank-16 LoRA	78.1 ± 0.1	94.16 ± 0.37
Drift-anchored fine-tuning	77.7 ± 0.1	93.90 ± 0.20
In-domain MAE (ViT-S)	63.5 ± 0.2	76.95 ± 0.26
Scalar pipeline	78.5 ± 0.1	96.54 ± 0.20

D Regularization calibration and representation diagnostics

The gain-logit penalty was calibrated after the zero-initialized ramp, over steps ≥ 51. The calibration window had cross-entropy 1.1860 and mean squared gain logits 0.1205. Coefficients

Table 7: Available multi-seed component controls on the focal sonar task. All entries are target LP means over three probe seeds. Missing entries were not run in the supplied study.

Pipeline	Seed 42	Seed 123	Seed 2024
Scalar	96.54	95.40	95.92
Frontend only	95.40	96.49	—
Frontend + subtraction + LN	95.61	93.80	—

0.295, 0.985, and 3.09 gave penalty-to-cross-entropy ratios of approximately 3.004%, 10.012%, and 30.037% in that window. These ratios describe calibration, not a fixed contribution throughout training. The direction-penalty coefficient was 1.3; only this coefficient was evaluated. Its manipulation check is measured during training and should not be conflated with the separate held-out pooled-feature distances.

The representation measurements use the seed-42 scalar/channel/affine checkpoint family and 128 frames per evaluation domain. Source/target cosine distances before and after correction are 0.0758/0.0370, 0.5258/0.5571, and 0.3950/0.4766 respectively. Source-to-target gain-distribution KS statistics are 0.207, 0.061, and 0.107. These small-sample diagnostics characterize the reported checkpoints; they do not establish population-level invariances.

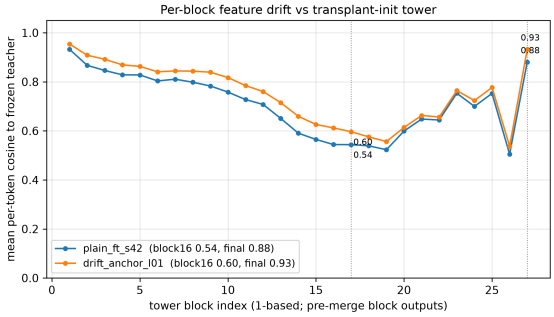

Figure 5: Feature cosine to a frozen initialization teacher across backbone blocks on 128 source frames. Drift anchoring retains more similar features but has lower target accuracy under the tested protocol. The curves concern plain and drift-anchored fine-tuning; they are not a scalar-versus-channel intervention.

In the drift profile, plain fine-tuning has block-16/final feature cosines 0.544/0.880, versus 0.596/0.933 for drift anchoring. A training-split repetition yields 0.535/0.884 for plain fine-tuning. The weight-space cosines remain at least 0.99999. Weight cosine is insensitive to some functionally important changes; its numerical proximity to one should not be interpreted as evidence that no learning occurred.

E Data handling and scope of the audit

The source sonar corpus comprises five public datasets (including UATD), sequentially split into 48,356/6,466/6,620 frames. EuroSAT uses a seed-0 stratified split, and DTD uses official split 1. Target train/validation partitions are derived from the official Marine Debris FLS training set, with the official test set retained. Probe feature standardization relies solely on training statistics.

An overlap audit of all 61,442 source sonar frames against 2,150 target images (via exact-image and perceptual-hash checks) finds no exact duplicates or identical perceptual hashes; near-hash matches arise from dark sonar backgrounds. This audit only covers duplicate images, not all acquisition or semantic dependencies. Shared ARIS-family sensors and synonymous labels (*tyre*/*tire*) across corpora limit domain independence interpretations. A second watertank scene from the same target dataset family is not used to claim independent target-task generality.