WHAT IMPROVES WHEN A MICROSCOPY ENCODER IS UPDATED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Microscopy encoders support scientific measurements through classifiers, segmentation decoders and measurement rules. We study which update gains survive the choice to co-train a head, leave it unchanged, or fit it afresh to frozen features. In matched dense-encoder forks across ten seeds, absolute task weighting leads normalization by 4.92 percentage points of culture balanced accuracy after 512 updates through co-trained heads and 4.22 through unchanged source heads; fresh-head contrasts remain unresolved. The preference is therefore measurable as compatibility with an existing classifier. The co-trained–fresh interaction persists under fresh-optimizer and separate-clipping controls, but equal-budget per-arm learning-rate tuning removes the early absolute advantage. Later heads diverge in recipe ranking without meeting the prespecified criterion for a co-trained head hiding resolved accuracy loss. DINOv2-small does not reproduce the early preference at its default rates. SCOPE, a completed 4B-parameter microscopy model, supplies complementary positive evidence: an unchanged classifier improves from 0.6235 to 0.9798 on terminal features, and fresh fits also improve on the previously supervised culture task. On 23 held-out Allen plates, brightfield structure Dice is 0.312, exceeding a 37M-parameter scratch control by 0.056; most of this capability predates continuation. Spatial controls show how decoder capacity and feature access change the update contrast. On held-out larvae, improved counts coexist with reduced strict object overlap, while a validation-based retention gate has no consistent advantage on external reporter assays. Most follow-ups reuse inspected panels. The study makes encoder progress interpretable through the complete use: which readout changes, which labels and optimization it receives, and which biological measurement the resulting procedure supports.

1 INTRODUCTION

A laboratory uses an encoder through a readout: a classifier identifies a culture, a decoder delineates a structure, and a counting rule turns predicted objects into a measurement. Updating the encoder can improve any of these uses, but the evaluation must specify what happens to the readout. A *co-trained* head evaluates the encoder–head pair after both have learned. An *unchanged* head tests compatibility with an existing interface. A *fresh* head, fitted after the encoder is frozen, tests what that fitting procedure can recover. These are distinct deployment choices, not competing definitions of a correct score.

We examine these choices through matched interventions and a completed microscopy training lineage. Paired copies of a dense encoder or DINOv2-small start from identical states and continue under different task weightings. SCOPE, a 4B-parameter shared microscopy encoder, supplies the longitudinal study: culture classifiers, brightfield structure predictions, spatial readouts and object measurements. The forks identify recipe effects within their families; SCOPE establishes which capabilities change along one trained lineage.

Three findings connect the studies. First, the early dense-model preference is readout-specific. At the original rate and 512 updates, absolute weighting leads normalization by 4.9 percentage points through co-trained heads and 4.2 through unchanged heads, while fresh-head contrasts remain unresolved (Figure 1b). The co-trained–fresh interaction survives optimizer and clipping controls;

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.

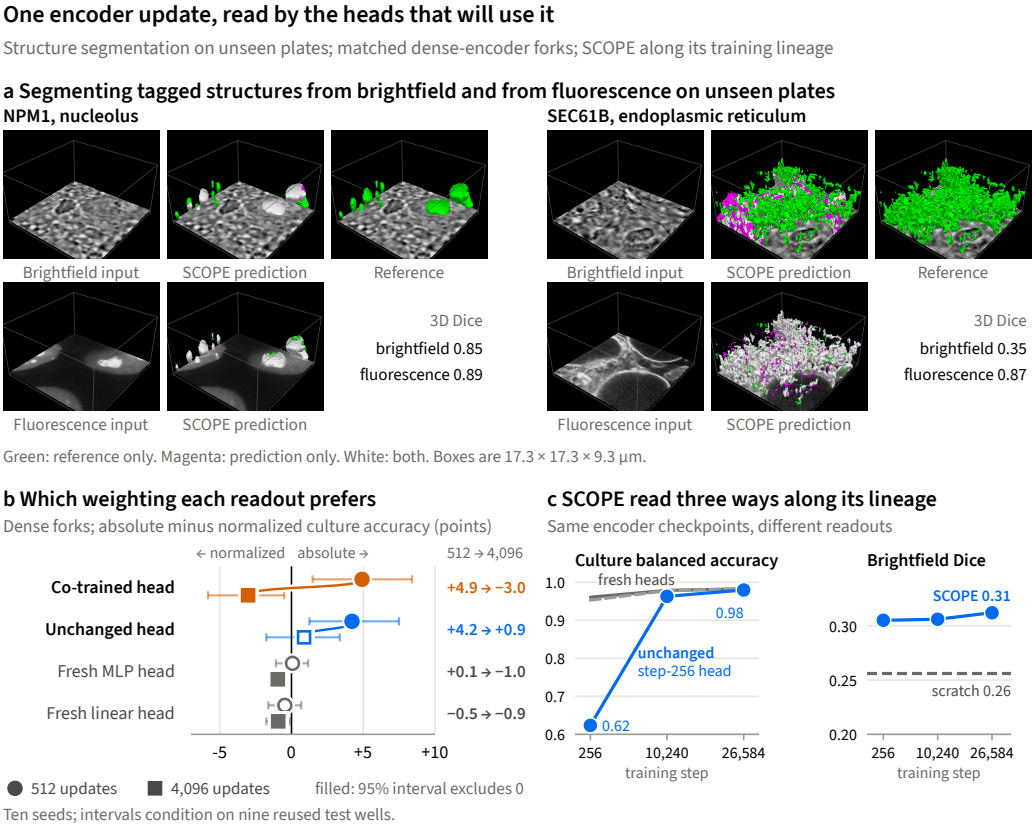

Figure 1: **One encoder update, read by the heads that will use it.** **(a)** Terminal SCOPE on two outcome-ranked, median-BF-Dice held-out crops. Floors show input slices; Dice uses full 3D crops. Fluorescence supplies the reporter used to derive references; brightfield does not. Selection and display-only tolerances: Appendix F.4. **(b)** Separate dense forks from absolute-trained sources at original rates. Circles/squares: 512/4,096 updates. Arrows connect means; bars are pointwise 95% bootstrap intervals over ten seeds, conditional on nine reused test wells. Filled marks indicate intervals excluding zero. **(c)** SCOPE's unchanged step-256 MLP and fresh linear/MLP heads (460 fitting and 128 validation images), evaluated on 1,460 images in nine reused wells. Fresh heads start higher and also improve. Right: brightfield macro Dice on 23 held-out plates versus the 37.2M-parameter scratch control; most of SCOPE's advantage predates continuation (Appendix F).

equal-budget per-arm rate tuning does not retain the early absolute preference. Later heads separate in ranking, yet no tested cell meets the prespecified criterion for hidden accuracy loss.

Second, useful SCOPE gains survive readout choice. Its unchanged incoming classifier gains approximately 36 culture-accuracy points on terminal features, and fresh heads also improve (Figure 1c). Held-out-plate brightfield segmentation exceeds a 37M-parameter scratch control, with most of that capability inherited before continuation. Spatial controls then show how decoder capacity and pyramid-feature access change the measured update advantage.

Third, the requested measurement determines the comparison. Lower count errors on held-out larvae accompany fewer strict-overlap matches; a variable-count nuclear panel also ranks pipelines differently by endpoint. Frozen reporter assays test a validation-based retention rule, whose advantage varies across families and assays. The next measurement requires its own evaluation rather than inheriting a preference from another score.

The underlying distinctions have precedents. Classifier rebalancing (Kang et al., 2020), losses (Kornblith et al., 2021), and readout protocols (Kornblith et al., 2019b; He et al., 2022) affect feature assessment; fixed-head performance and representation forgetting are distinct (Davari et al.,

2022), and fine-tuning can distort pretrained features (Kumar et al., 2022). Backward-compatible and regression-free updates address deployment constraints (Shen et al., 2020; Yan et al., 2021), while tuned scalarization changes multitask comparisons (Xin et al., 2022; Kurin et al., 2022). Microscopy benchmarks likewise expose protocol-dependent rankings (Teuber et al., 2026; Hayir et al., 2026; Treis et al., 2026). Our contribution is the connected experimental comparison: co-trained, unchanged and fresh heads on matched updates, followed by spatial and measurement controls in a shared microscopy system. It makes the update claim operational—which encoder training reaches, which readout puts it to use, and which biological measurement that procedure supports.

2 SCOPE AS A LONGITUDINAL CASE STUDY

SCOPE shares spatial computation across imaging channels and tasks with incomplete annotation. Its completed training lineage lets us examine how those shared features serve different readouts. Channel-flexible encoding, partial-label segmentation and label-free prediction have established precedents (Bao et al., 2023; Chen et al., 2023; Zhang et al., 2021; Ounkomol et al., 2018; Christiansen et al., 2018); our experiments follow the capabilities of the trained system rather than isolate its architectural components.

An explicit observation contract. Typed channels identify the measurements, a voxel-to-world affine places the view in space, and task queries specify the requested outputs. A shared convolutional stem embeds each channel with its type and fuses available channels by voxelwise softmax. Jointly reordering channels and types therefore leaves the fused result unchanged. Supervision distinguishes an unobserved target from an observed absence: a missing annotation contributes no loss, whereas a valid empty mask supplies negative supervision. Missing reporters are not zero-intensity targets, and training-only teacher features remain outside the student input.

Shared computation. A residual pyramid with total stride $(4, 8, 8)$ feeds a 28-block, width-1536 transformer. Local attention is supplemented by compressed spatial blocks (DeepSeek-AI, 2026), with rotary phases derived from physical coordinates (Su et al., 2021; Khan et al., 2026). Four residual streams use hyper-connections (Xie et al., 2025); 14 mixture-of-experts layers select four of 32 routed experts per token alongside a shared expert. A U-shaped decoder (Ronneberger et al., 2015) and task readouts produce segmentation, fluorescence prediction, instance cues and eight-class LIVECell culture classification (Edlund et al., 2021). Auxiliary readouts predict features and align visual descriptors to verified localization annotations. One pass means one student encoder evaluation per input view, excluding additional training-teacher evaluations.

Objective. Let S_e, I_e, C_e indicate eligible segmentation, instance and culture targets for episode e, and let $\mathcal{A}_e$ index eligible auxiliary targets. The continuation uses

$$L_e = 2S_e R_{\text{seg}} + 0.1S_e L_{\text{BCE}} + \sum_{k \in \mathcal{A}_e} w_k \ell_k + 0.25 I_e L_{\text{inst}} + 0.15 C_e L_{\text{class}} + 0.002 L_{\text{MoE}} + 0.01 L_{\text{indexer}}, \quad (1)$$

where R_{seg} averages observed-query risks, each combining $0.7L_{\text{softDice}} + 0.3L_{\text{Brier}}$; for empty references, the Dice term is replaced by mean foreground probability. Auxiliary coefficients are $(0.3, 0.1, 0.05, 0.1, 0.2)$ for reconstruction, self prediction, physical context, paired latent prediction and localization. These coefficients are *absolute*: task availability determines which losses are evaluated, without renormalizing their coefficients. Their comparative effect is tested in the smaller matched forks, not inferred from SCOPE's training outcome.

Training history and comparison points. The continuation completes 26,584 updates from previously trained weights. The first 128 freeze the encoder while training eligible decoder and readout parameters; subsequent updates train the permitted encoder, decoder and heads with an EMA teacher of momentum 0.996. Updates average seven or ten episodes per rank across three ranks. The catalog contains 319,612 observation contracts, mostly microscopy with some structural MRI (Van Essen et al., 2013); repeated views are not independent acquisitions. Appendices E.2 and E specify the inherited history and continuation. Checkpoints at steps 256, 10,240 and 26,584 are termed *incoming*, *adapted* and *terminal*. Incoming SCOPE is already pretrained; terminal comparisons were added post hoc.

3 COMPARING UPDATES UNDER DECLARED PROCEDURES

What is allowed to change? Let P be the observation pipeline and $D_h[E]$ the readout supplied or fitted for encoder checkpoint E by procedure h. The procedure specifies the head's origin or initialization, fitting labels, update budget and validation-selection rule, including its associated training history when the head is co-trained. Conditional on the resulting fitted readout, define

$$R_h(E) = \mathbb{E}_{(X,Y)\sim Q}[\ell(D_h[E](E(P(X))), Y)], \qquad \Delta_h(E_a, E_b) = R_h(E_b) - R_h(E_a), \tag{2}$$

for evaluation population Q and task loss ℓ. A positive contrast favors E_a under that procedure. Co-trained, unchanged and fresh heads test joint adaptation, compatibility and fresh decodability, respectively. *Repair* continues training a joint head on frozen features, allowing validation to retain zero updates. Fresh fitting counts as unseen-task transfer only when the task is new to the encoder. These finite procedures include their label and compute costs; they do not estimate Bayes risk or all information contained in the features.

Matched weighting interventions. A dense planar encoder and DINOv2-small (Oquab et al., 2024) have 360,697 and 22,209,705 trainable parameters including heads. Each carries a BBBC039 nuclear readout (Caicedo et al., 2019; Ljosa et al., 2012) and a LIVECell culture MLP. Their *source* models are trained separately from SCOPE for 3,600 (dense) or 900 (DINOv2) updates. Each paired *fork* restores the same weights, AdamW state (Loshchilov & Hutter, 2019), randomness and upcoming batches into its weighting branches, or *arms*. Each update draws two nuclear and two culture records. With canonical nuclear Dice/Brier/BCE risk L_n divided by 2.1 and culture cross-entropy L_c,

$$L_{\text{abs}} = \tfrac{1}{2}(2.1L_n + 0.15L_c), \qquad L_{\text{norm}} = \tfrac{1}{2}(L_n + L_c), \qquad L_{\text{ref}} = 1.125L_{\text{norm}}. \tag{3}$$

Both tasks occur in every update: the nuclear:culture coefficient ratio is 14:1 versus 1:1, and the reference-scaled arm restores total coefficient mass without restoring the ratio. AdamW uses $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, decay 10^{-4} and clipping at 5. Encoder rates are 3×10^{-4} for dense and 10^{-5} for DINOv2; heads use 3×10^{-4}. These interventions estimate the complete recipe effect, including its interaction with optimizer state and clipping.

Availability and optimization controls. Ten seeds cross both source weightings with both arm objectives over 4,096 updates. Fresh readouts are logistic regression and an original-architecture LayerNorm–MLP fitted for 2,048 updates with validation selection. Each *unchanged* head is its source's co-trained culture head applied to both arms without further training. Controls reset AdamW, vary shared or separate encoder/head rates, tune rates per arm down to $0.015\times$, warm heads on frozen features, or change clipping. Per-arm tuning uses each readout's own validation criterion, so different procedures can select different encoders. Model-state digests match the original three shared seeds. Appendices B.2, C.1 and C.6 retain the grids and replication checks; unequal source budgets and rates limit architecture-only interpretation.

Two follow-ups use the ten absolute-trained dense sources. One resets AdamW, warms up for 256 updates and decays rates to zero at 4,096, with $0.5\times$, $1\times$ and $2\times$ peaks (Appendix C.4). The other samples nuclei-only, culture-only or both with equal probability and matched subset schedules. Each present task draws records from its own dataset, rather than supplying a second annotation of one image. Arm A keeps absolute coefficients, E assigns each present task coefficient $1/2$, and N averages canonical losses over present tasks (Appendix D). Rules preceded new fitting, but these wells had already been examined. The attempted full-SCOPE weighting comparison did not finish and supplies no result here.

Evaluation populations. Culture uses 460/128/1,460 train/validation/test images; nine test wells cover eight classes, seven with one well each. BBBC039 uses 122/42/36 fields from 12/4/4 plates; BBBC050 uses 99/22/44 volumes from 9/2/4 embryos. These reused panels support exploratory comparisons. *Source-exposed* denotes populations present in the encoder lineage. The native-3D *C. elegans* study (Long et al., 2009; Hirsch & Kainmueller, 2020) fixes five pipelines, three head seeds and instance conversion using 18 training and three validation larvae before scoring seven test larvae; the terminal extension reuses them. BBBC013 and BBBC014 (Ravkin & Broad Institute Imaging Platform, n.d.a;n) each contribute one plate scored by 26 frozen pipelines without on-assay

fitting. Only DNA enters segmentation; the observed reporter is measured afterward by a common operator.

Allen imaging (Viana et al., 2023) supplies 256 label-independently sampled crops covering 24 structures on 23 test plates disjoint from selected training and validation plates. Audited lineage catalogs contain no matching test keys or plate IDs; pixel duplicates and unlisted ancestors remain outside that audit. Algorithmic Allen masks are reference segmentations, not independently adjudicated biological truth. Dice averages nonempty crops within structures, then structures equally; 22 empty-reference crops supply a separate foreground false-positive endpoint.

Uncertainty and selection. Fork intervals resample training seeds conditional on reused culture wells; other studies use their declared acquisition units. Allen resampling retains repeated plate draws and varies plate/structure composition, not new plates within a fixed structure. Head seeds are not biological replicates. Intervals are pointwise, with grid-wide sensitivities and post-hoc analyses identified separately. A contrast is *resolved* when its 95% interval excludes zero; an interval spanning zero establishes neither equality nor equivalence. Readout interactions are estimated from paired differences, not from a contrast between significance labels.

In the original three-seed forks, four validation-only rules choose among 13 fixed candidates: retain the source; maximize culture accuracy; maximize it without worsening point-estimate nuclear F1 or count MAE; or minimize a joint culture, object and count risk. The third is the *measurement-retention gate*; the fourth is *ordinary joint validation*. Tie-breaking is fixed, and later assay outcomes select neither the rules nor their candidate pipelines.

4 Results

4.1 Co-trained and unchanged heads share an early advantage

At the original dense learning rate and 512 updates, absolute weighting exceeds normalized weighting by 4.92 percentage points of culture balanced accuracy through the co-trained head (95% seed interval $[1.47, 8.39]$). Fresh linear and MLP fits give contrasts of -0.46 and $+0.07$ points, respectively, with both intervals spanning zero (Table 1). The directly paired joint-minus-MLP interaction is $+4.84$ $[1.13, 8.99]$. From normalized-trained sources, the joint contrast is also positive, $+3.23$ $[0.79, 5.67]$: the preference does not require continuing the source objective.

The unchanged-head control asks whether updating the head is necessary for this preference. The source head already registers it when held fixed: $+4.22$ $[1.25, 7.49]$, with an unchanged-minus-joint contrast of -0.70 $[-1.85, 0.53]$ (Figure 1b). A normalized-trained source's unchanged head likewise favors absolute weighting, $+3.01$ $[1.04, 4.94]$. Thus the early effect is measurable as compatibility with an existing classifier, as well as through joint adaptation. The fresh fits do not establish the same advantage; their unresolved contrasts do not imply equal information in the two representations (Appendix B.3).

Learning-rate controls identify the range of that result. The early joint contrast changes from -0.63 at shared $0.15\times$ to $+11.20$ at $2\times$; at $0.5\times$, joint and fresh-MLP readouts favor opposite recipes (Appendix Figure 3). Equal-budget per-arm validation tuning gives an unresolved joint contrast of -0.37 $[-1.32, 0.55]$. Extending the search to $0.015\times$ reduces lower-edge selection from nine to one of ten normalized-arm choices. The tuned joint contrast becomes -0.92 $[-1.83, -0.10]$, while both fresh contrasts remain unresolved (Appendix C.2). The original absolute-weighting advantage therefore belongs to the shared-rate regime, not to a generally superior tuned recipe.

The effect also survives changes to optimizer history and gradient handling. Fresh AdamW with 256 warmup updates and cosine decay gives a joint-minus-MLP interaction of $+3.50$ $[1.97, 5.40]$; 59% of this estimate comes from BT474's single test well, and per-arm tuning again leaves the interaction unresolved (Appendix C.4). Separately clipping encoder and head groups preserves the matched-fork interaction, $+6.46$ $[4.35, 9.03]$, positive in all ten seeds. Removing clipping gives $+3.49$ $[0.74, 6.28]$. Global norm coupling is therefore not necessary for this effect. Its full causal contribution remains open: two unstable unclipped seeds leave the variable-availability result unresolved (Appendix C.5).

Table 1: **Readout rankings depend on learning rate and horizon.** Absolute minus availability-normalized culture balanced accuracy (points), absolute-trained sources, inherited AdamW: ten-seed means and pointwise 95% seed-bootstrap intervals, conditional on nine reused test wells. Multipliers scale encoder and head rates in both arms; the tuned row selects each arm's multiplier on validation with the readout's own criterion. The late $2\times$ MLP favors absolute weighting; tuned procedures can choose different encoders for different readouts.

	512 updates			4,096 updates		
Condition	Joint	Fresh linear	Fresh MLP	Joint	Fresh linear	Fresh MLP
Shared 0.15×	-0.63 $[-1.27, -0.05]$	-0.97 $[-1.57, -0.48]$	-1.13 $[-1.88, -0.56]$	-1.05 $[-1.82, -0.25]$	-1.00 $[-1.63, -0.39]$	-0.67 $[-1.20, -0.14]$
Shared 0.5×	$+2.50$ $[+0.15, +5.01]$	-1.57 $[-2.45, -0.76]$	-0.97 $[-1.90, -0.14]$	-1.82 $[-3.02, -0.40]$	-1.17 $[-2.08, -0.33]$	-0.82 $[-1.49, -0.27]$
Shared 1×	$+4.92$ $[+1.47, +8.39]$	-0.46 $[-1.61, +0.68]$	$+0.07$ $[-1.08, +1.16]$	-3.02 $[-5.81, -0.51]$	-0.92 $[-1.75, -0.11]$	-0.95 $[-1.37, -0.53]$
Shared 2×	$+11.20$ $[+6.49, +16.33]$	$+5.78$ $[+3.38, +8.53]$	$+8.28$ $[+4.73, +12.22]$	$+1.63$ $[-1.01, +4.20]$	$+0.56$ $[-0.32, +1.68]$	$+0.95$ $[+0.07, +1.92]$
Tuned per arm	-0.37 $[-1.32, +0.55]$	-0.77 $[-2.30, +0.86]$	$+0.06$ $[-1.14, +1.27]$	-1.67 $[-3.20, -0.29]$	-0.67 $[-1.19, -0.12]$	-1.01 $[-1.61, -0.38]$
Tuned, wider grid	-0.92 $[-1.83, -0.10]$	-0.78 $[-2.35, +0.83]$	-0.33 $[-1.27, +0.50]$	-1.99 $[-3.53, -0.53]$	-0.67 $[-1.19, -0.12]$	-1.01 $[-1.61, -0.38]$

Variable task availability provides a further contrast. Two unnormalized recipes give an A-minus-E joint-minus-MLP interaction of $+6.38$ $[3.11, 9.18]$ points, so per-episode normalization is not required for a readout interaction. Ratio and scale change together in that comparison (Appendix D). DINOv2-small defines the family boundary: at its default rates, no readout resolves an early weighting preference from either source. Tripling its encoder rate instead moves all readouts toward absolute weighting by 7–20 points, while higher rates enter the reported culture-collapse regime (Appendix C.6).

4.2 Training time separates recipe ranking from retained accuracy

At 4,096 updates, the early joint-versus-fresh interaction is unresolved. Under the constant-rate recipe, joint and fresh heads favor normalized weighting at shared rates $0.15\times$–$1\times$ and under wider-grid per-arm tuning (tuned joint -1.99 $[-3.53, -0.53]$). The $2\times$ fresh MLP instead favors absolute weighting, $+0.95$ $[0.07, 1.92]$. Horizon and rate must therefore accompany a recipe ranking.

At shared $1\times$, the co-trained head favors normalized weighting, -3.02 $[-5.81, -0.51]$, while the unchanged-head contrast is unresolved, $+0.88$ $[-1.75, 3.36]$. Their paired difference is $+3.90$ $[1.03, 7.52]$. This distinguishes ranking from loss of an existing capability: none of the 26 prespecified cells meets the criterion for a co-trained head hiding an unchanged-head accuracy loss. Six meet the ranking-only criterion; 20 meet neither, under unadjusted intervals. In cells with resolved unchanged-head regression, co-trained performance also regresses. These findings do not prove the absence of smaller losses. The late joint preference is also sensitive to the Huh7 and SKOV3 wells (Appendices B.3 and C.1).

The second trained task adds a separate constraint. At $1\times$/512, the culture advantage accompanies -0.20 points of nuclear F1 and $+0.253$ objects of count error; at $2\times$, both tasks improve. A multitask preference therefore requires their joint comparison rather than culture accuracy alone (Appendix D.1).

4.3 SCOPE gains support culture decoding and brightfield segmentation

SCOPE supplies positive evidence across several readouts of one trained lineage. Its unchanged step-256 classifier improves from 0.6235 to 0.9798 balanced accuracy on the incoming versus terminal features of 1,460 held-out images (Table 2). Fresh linear fits improve from 0.9608 to 0.9826 with 460 fitting images and from 0.7853 to 0.9832 with 51. Fresh MLPs also improve. These are same-task representation gains: culture labels already supervised the encoder, and both fitting budgets use the 128 labeled validation images for selection. DINOv2 CLS remains stronger at the full fitting budget, 0.9927, under a different pretraining procedure.

Table 2: **Culture gains survive several readouts; object outcomes are task-specific.** The SCOPE update under fixed procedures. Each row reads the step-256, -10,240 and -26,584 encoders with one procedure. Culture: 1,460 reused test images in nine wells; culture labels also supervised the encoder. BBBC039: 36 fields from four plates. *C. elegans*: seven larvae, expanded instance-conversion grid. The 26,584 column is a later post-hoc extension; for *C. elegans* it reuses the locked larvae. Fresh MLP rows average three fitting seeds. Controls differ in pretraining, architecture and fitting budget and are not matched ablations.

Readout procedure	Endpoint	256	10,240	26,584	Strong control
Culture: inherited step-256 head	Bal. acc.	0.6235	0.9630	0.9798	–
Culture: fresh linear, 51 images	Bal. acc.	0.7853	0.9781	0.9832	–
Culture: fresh linear, 460 images	Bal. acc.	0.9608	0.9788	0.9826	DINOv2 CLS 0.9927
Culture: fresh MLP, 460 images	Bal. acc.	0.9530	0.9779	0.9832	–
BBBC039: strong object readout	F1@0.5	0.9472	0.9476	0.9467	Tuned Cellpose 0.9699
C. elegans: full pyramid	F1@0.5	0.8075	0.8096	0.8071	Scratch 3D U-Net 0.7885

Head identity matters to compatibility. A *source-fitted linear probe*, unlike the inherited classifier above, drops from 0.9608 to 0.9475 on adapted features. Training-only alignment of feature means and variances restores 0.9768 (Appendix F.2). This localizes a correctable interface mismatch for that probe, not a general regression of every source head.

On 23 held-out Allen plates, terminal brightfield macro Dice is 0.312 $[0.224, 0.407]$. A 37.2M-parameter scratch U-Net trained on the same Allen records for 60,000 updates per seed reaches 0.256 $[0.174, 0.347]$; the paired advantage is $+0.056$ $[0.037, 0.079]$. The original 16.5M control reaches 0.242. These are positive system-level comparisons under the reported budgets. Both scratch results average two seeds, and the larger control's selected checkpoints occur at its budget endpoint; convergence, capacity and lifetime compute are not matched to SCOPE.

The lineage comparison separates this capability from the continuation gain: terminal-minus-incoming Dice is $+0.0076$ $[0.0040, 0.0115]$. NPM1, FBL and LMNB1 exceed 0.73 Dice, whereas six structures remain below 0.10. The held-out-plate advantage is therefore largely inherited and uneven across structures (Appendix F.3). Outcome-ranked qualitative examples show the corresponding morphology in brightfield and fluorescence; fluorescence agreement uses references derived from that same imaging channel (Appendix F.4).

4.4 Spatial readouts expose the roles of capacity and feature access

On BBBC039, strong object readouts give F1 at IoU 0.5 of 0.9472, 0.9476 and 0.9467 on incoming, adapted and terminal features, with unresolved terminal contrasts. Tuned Cellpose (Pachitariu et al., 2025) reaches 0.9699 and count MAE 2.60 under its own fitting budget. The culture gains therefore coexist with a nuclear comparison that does not resolve an update benefit.

BBBC050 embryos (Tokuoka et al., 2020) show how the readout changes that comparison. A restricted token/half-skip head favors adapted features, Dice 0.7605 versus 0.7101; full-pyramid fusion yields 0.7873 versus 0.7942, with scratch at 0.8221. A controlled extension then fixes the encoder pair and compares the original restricted head, a larger restricted-input head, and full-pyramid fusion. At the prespecified 3,600-update allowance, their adapted-minus- incoming gains are $+0.0596$, $+0.0190$ and -0.0160. The directly paired full-pyramid-minus-larger-restricted interaction is -0.0350 $[-0.0460, -0.0224]$. Larger restricted capacity reduces the initial contrast; access to additional scales changes it further. The latter comparison retains adapter differences, three fitted seeds and four reused embryos (Appendix G.3).

Input processing provides a complementary control. On BBBC021, training-matched planar layout raises adapted-token compound-excluded phenotype accuracy from 0.5853 to 0.6823. With acquisition week additionally excluded, 0.4284 versus 0.4299 is unresolved. Pyramid effects vary by endpoint, and DINOv2 and CellProfiler remain ahead of the adapted tokens (Broad Institute Imaging Platform, n.d.; Caie et al., 2010; Ljosa et al., 2013) (Appendix I). Neither fresh culture accuracy nor one favorable spatial probe substitutes for these task-specific tests. SCOPE's devel-

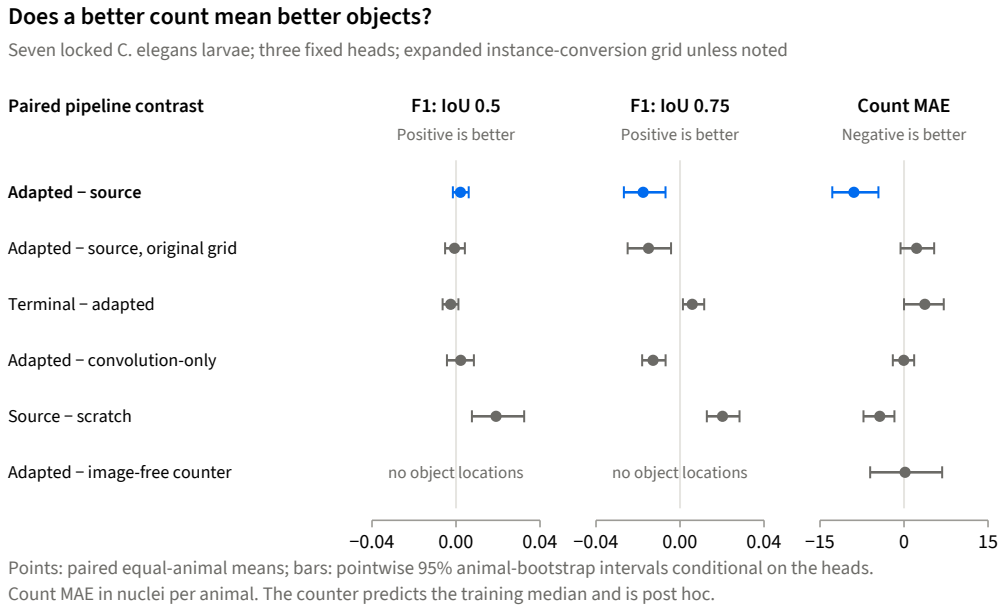

Figure 2: **An update's benefit depends on the requested measurement.** Paired pipeline contrasts on seven *C. elegans* larvae. Points are equal-animal means over three fitted heads; bars are pointwise 95% animal-bootstrap intervals conditional on those heads. Source, adapted and terminal denote steps 256, 10,240 and 26,584. All receive the expanded validation-selected instance conversion. The terminal extension reuses these animals post hoc. The image-free training-median comparator predicts counts but has no object-localization output.

opment guardrail likewise passes only five of fourteen validations, including the terminal one; the plant empty-reference failure at step 20,480 remains recorded (Appendix E).

4.5 Counts and strict object overlap track different changes

Holding a binary foreground fixed exposes what instance conversion adds. On three development larvae, changing only the conversion setting leaves Dice at 0.8761 while object F1 changes from 0.6646 to 0.7260 and count MAE from 110.3 to 35.3 nuclei. Every pipeline subsequently receives the same expanded validation-selected conversion allowance. On seven locked test larvae, the primary adapted-minus-incoming F1-at-IoU-0.5 contrast is $+0.0021$ $[-0.0015, 0.0060]$. Count MAE improves from 17.57 to 8.62, while strict F1 at IoU 0.75 falls from 0.4123 to 0.3947 (Figure 2). The count benefit is unresolved under the original conversion grid. A post-hoc training-median counter, 544 nuclei for every image, reaches MAE 8.43 on this nearly fixed-count population but supplies no object locations.

One-to-one matching clarifies the difference: signed count error is $\mathrm{FP} - \mathrm{FN}$, whereas unmatched objects total $\mathrm{FP} + \mathrm{FN}$. Errors can cancel in the former while accumulating in the latter. At IoU 0.75, unmatched objects increase from 631.38 to 648.24 per animal. A post-hoc fate analysis of one head seed records fewer merge-classified reference nuclei in all seven animals, more split-classified nuclei in all seven, and fewer strict matches in six. These observed changes accompany better counts; the precedence-based categories do not isolate a causal mechanism or directly measure boundary accuracy (Appendix H.4).

The variable-count control establishes the population boundary. On 36 BBBC039 fields from four plates, reference counts range from 53 to 202; a training-median constant has MAE 31.89. Classical watershed instead has 1.89 fewer nuclei of count error than the adapted SCOPE semantic readout, but 0.0690 lower F1 at IoU 0.75, with both directions retained on all four plates. This is a post-hoc pipeline comparison, not an encoder-update effect; four plates limit its intervals, and the within-

SCOPE contrast is unresolved (Appendix H.5). Image-specific counting is useful here, yet remains distinct from strict object matching.

The terminal larva extension gives primary F1 0.8071 versus adapted 0.8096 with an unresolved paired contrast, and count MAE 12.33. Frozen SCOPE pipelines exceed the tested scratch model's primary F1 by approximately 0.019. A convolution-only control, however, exceeds the full adapted pipeline at IoU 0.75. These results support particular pretrained pipelines, without isolating a transformer or matched-compute benefit (Appendices H.2 and F).

4.6 Reporter assays test the reach of a validation rule

Frozen reporter deployment tests whether a measurement-retention rule carries its validation preference into another assay. In the original dense study, the gate selects lower culture accuracy than ordinary joint validation, 0.6785 versus 0.7322, at similar object F1 and count error. Gate-minus-joint signed dose associations then range from -0.0045 to $+0.0124$ on BBBC013 and from -0.0052 to $+0.0003$ on BBBC014, changing sign across families and series. SCOPE adaptation changes association by at most 0.0045; classical segmentation leads it on three of four series (Appendix J).

Reporter shifts remove the strong positive associations, supporting a role for spatial correspondence rather than verifying cellular localization. Each assay contributes one plate; BBBC014 also confounds dose with column and was specified after inspection of BBBC013. The fixed comparisons thus place an empirical boundary on the proposed gate: preserving the selected development metrics did not produce a consistent external-assay advantage. The next measurement remains an outcome to evaluate, not a consequence of passing the preceding validation rule.

5 Discussion

What the controls distinguish. The early dense-model preference is visible through both co-trained and unchanged heads, while fresh fitting leaves the primary contrast unresolved. This establishes compatibility with a particular classifier, not an ordering of all culture information. Per-arm rate tuning removes the early absolute preference; later ranking divergence meets no criterion for hidden accuracy loss. SCOPE supplies the complementary positive result: unchanged and fresh classifiers improve, and held-out-plate brightfield capability survives a stronger scratch comparison. Spatial and object controls then show which readout and measurement choices change the interpretation. Together, the studies identify useful outcomes under explicit conditions rather than one latent quantity shared by every score.

A reusable evaluation procedure. An update comparison should declare four components: the *optimization* budget and rate selection; the *readouts* permitted to change and their fitting labels; the *measurement*, including conversion to objects or counts; and the *population*, its acquisition units and prior exposure. Co-trained, unchanged and fresh-head comparisons then answer explicit use cases. Every trained task and a relevant simple control remain in the evaluation. A constant counter tests dependence on image-specific count variation; object matching tests a different output property.

Scope. Training seeds are not new acquisitions. Seven culture classes have one test well; BT474 dominates the scheduled recipe's primary interaction, and Huh7/SKOV3 determine the late joint contrast's nominal resolution. Follow-ups reuse inspected panels. SCOPE is one lineage without a completed matched 4B fork; scratch lifetime compute is unmatched and convergence unestablished. The Allen development panel includes 42 training fields, whereas the plate-disjoint test is separate, with algorithmic references and one plate per structure. Four nuclear plates, seven larvae and single-plate reporter assays limit acquisition-level generalization. The tested retention gate has no consistent external advantage.

Conclusion. An encoder update earns a useful interpretation through the readout and measurement it supports. Pairing co-trained, unchanged and fresh heads, then tracing their outputs into spatial measurements, distinguishes existing-head compatibility, fresh-readout gains and endpoint-specific trade-offs. Reporting those outcomes with their optimization, labels and acquisition scope makes progress interpretable where a laboratory puts the model to use.

ETHICS STATEMENT

Training data include de-identified structural MRI from the Human Connectome Project (Van Essen et al., 2013), used under its data-use terms, together with microscopy collections used under their respective terms. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657), funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University. No new human-subject data were collected. The package includes microscopy illustrations but no raw human imaging data or subject identifiers, and no clinical claim is made.

AI USE STATEMENT

Generative AI tools assisted with research ideas and experimental design, method implementation, data preparation, analysis and interpretation, and the formulation and checking of mathematical claims and proofs. They also assisted with literature searches, software, figures, manuscript drafting and revision, and reference formatting. The reported evidence comes from the specified experiments and derivations. Numerical and implementation checks are described in the appendices and accompanying analysis code. The authors are responsible for the final text, claims and artifacts.

REPRODUCIBILITY STATEMENT

The matched-fork objectives, source states, fitting budgets and selection rules are specified in Section 3 and Appendices B–C.6; SCOPE's model and consolidation procedure are specified in Appendix E. Evaluation populations, readouts and measurement procedures are stated alongside their results. The accompanying source package includes corrected Allen resampling code, structure-level aggregates and tests. It also supplies a standalone public-data dense-control implementation, validated so far by source checks and isolated synthetic CPU execution, not a completed independent biological reproduction. It does not include training images or a complete reproduction of the inherited SCOPE training lineage. Appendix K states the release scope and remaining reproducibility limits.

REFERENCES

Allen Institute for Cell Science. 3D Cell Viewer. https://www.allencell.org/3d-cell-viewer.html, n.d. Provider resource record; accessed September 25, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.08243.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023. URL https://arxiv.org/abs/2309.16108.

Biohub. DynaCell: Registry of open data on AWS. https://registry.opendata.aws/dynacell/, 2026. Provider registry record; accessed September 25, 2026.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Broad Institute Imaging Platform. BBBC039v1: Nuclei of U2OS cells in a chemical screen. https://bbbc.broadinstitute.org/BBBC039, 2018. Manual nuclear reference collection; accessed September 20, 2026.

Broad Institute Imaging Platform. BBBC021v1: Human MCF7 cells, compound-profiling experiment. https://bbbc.broadinstitute.org/BBBC021, n.d. Version 1 dataset record.

Juan C. Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W. Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J. Theis, and Anne E. Carpenter. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A*, 95(9):952–965, 2019. doi: 10.1002/cyto.a.23863.

Peter D. Caie, Rebecca E. Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E. Roberts, and Neil O. Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular Cancer Therapeutics*, 9(6):1913–1926, 2010. doi: 10.1158/1535-7163.MCT-09-1148.

Srinivas Niranj Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114–1121, 2024. doi: 10.1038/s41592-024-02241-6.

Jianxu Chen, Liya Ding, Matheus P. Viana, et al. The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. *bioRxiv*, 2020. doi: 10.1101/491035. URL https://www.biorxiv.org/content/10.1101/491035v2. Version 2, December 13, 2020.

Zitong Chen, Chau Pham, Siqi Wang, Michael Doron, Nikita Moshkov, Bryan A. Plummer, and Juan C. Caicedo. CHAMMI: A benchmark for channel-adaptive models in microscopy imaging. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2310.19224.

Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, et al. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, 2022. doi: 10.1126/science.abi6983.

Eric M. Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K. Lee, Piyush Goyal, William Fedus, Ryan Poplin, Andre Esteva, Marc Berndl, Lee L. Rubin, Philip Nelson, and Steven Finkbeiner. In silico labeling: Predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803.e19, 2018. doi: 10.1016/j.cell.2018.03.040.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01621. URL https://arxiv.org/abs/2203.13381.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Christoffer Edlund, Timothy R. Jackson, Nabeel Khalid, et al. LIVECell—a large-scale dataset for label-free live cell segmentation. *Nature Methods*, 18:1038–1045, 2021. doi: 10.1038/s41592-021-01249-6.

Funahashi Laboratory and Broad Institute Imaging Platform. BBBC050v2: Nuclei of mouse embryonic cells. https://bbbc.broadinstitute.org/BBBC050, 2020. Manual nuclear reference collection accompanying Tokuoka et al., npj Systems Biology and Applications 6, 32; accessed September 20, 2026.

Emre Hayir, Lorin Crawford, and Alex X. Lu. MorphoHELM: A comprehensive benchmark for evaluating representations for microscopy-based morphology assays. *arXiv preprint arXiv:2605.15383*, 2026. URL https://arxiv.org/abs/2605.15383.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

Peter Hirsch and Dagmar Kainmueller. An auxiliary task for learning nuclei segmentation in 3D microscopy images. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pp. 304–321, 2020. URL https://proceedings.mlr.press/v121/hirsch20a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.09217.

Dorian Kauffmann, Guillaume Gay, Julio Mateos-Langerak, Oriane Pourcelot, Virginie Georget, Maëlle Carraz, et al. 2D multimodal image collection for fluorescence prediction from transmitted light microscopy. *Scientific Data*, 13:743, 2026. doi: 10.1038/s41597-026-07004-w.

Amil Khan, Anantajit Subrahmanya, and B. S. Manjunath. Geometry matters in packed 3D attention. In *Advances in Neural Information Processing Systems*, 2026. Accepted for publication.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019a. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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*, pp. 2661–2671, 2019b. doi: 10.1109/CVPR.2019.00277.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2010.16402.

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. URL https://arxiv.org/abs/2202.10054.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.04122.

Ziwen Liu, Eduardo Hirata-Miyasaki, Soorya Pradeep, Johanna V. Rahm, Christian Foley, Talon Chandler, et al. Robust virtual staining of landmark organelles with Cytoland. *Nature Machine Intelligence*, 7:901–915, 2025. doi: 10.1038/s42256-025-01046-2.

Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637, 2012. doi: 10.1038/nmeth.2083.

Vebjorn Ljosa, Peter D. Caie, Rob ter Horst, Katherine L. Sokolnicki, Emma L. Jenkins, Sandeep Daya, Mark E. Roberts, Thouis R. Jones, Shantanu Singh, Auguste Genovesio, Paul A. Clemons, Neil O. Carragher, and Anne E. Carpenter. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *Journal of Biomolecular Screening*, 18(10):1321–1329, 2013. doi: 10.1177/1087057113503553.

David J. Logan and Anne E. Carpenter. Screening cellular feature measurements for image-based assay development. *Journal of Biomolecular Screening*, 15(7):840–846, 2010. doi: 10.1177/1087057110370895. URL https://journals.sagepub.com/doi/10.1177/1087057110370895.

Fuhui Long, Hanchuan Peng, Xiao Liu, Stuart K. Kim, and Eugene Myers. A 3D digital atlas of C. elegans and its application to single-cell analyses. *Nature Methods*, 6(9):667–672, 2009. doi: 10.1038/nmeth.1366.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15:917–920, 2018. doi: 10.1038/s41592-018-0111-2.

Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-SAM: Superhuman generalization for cellular segmentation. *bioRxiv*, 2025. doi: 10.1101/2025.04.28.651001.

Ilya Ravkin and Broad Institute Imaging Platform. BBBC013v1: Human U2OS cells cytoplasm–nucleus translocation. https://bbbc.broadinstitute.org/BBBC013, n.d.a. Version 1 dataset record; accessed September 21, 2026.

Ilya Ravkin and Broad Institute Imaging Platform. BBBC014v1: Human MCF7 and A549 cells cytoplasm–nucleus translocation. https://bbbc.broadinstitute.org/BBBC014, n.d.b. Version 1 dataset record.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2003.11942.

David R. Stirling, Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. CellProfiler 4: Improvements in speed, utility and usability. *BMC Bioinformatics*, 22:433, 2021. doi: 10.1186/s12859-021-04344-9.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015. URL https://arxiv.org/abs/1511.05547.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15:29, 2015. doi: 10.1186/s12880-015-0068-x. URL https://link.springer.com/article/10.1186/s12880-015-0068-x.

Carolin Teuber, Anwai Archit, Tobias Boothe, Peter Ditte, Jochen Rink, and Constantin Pape. Evaluating vision foundation models for pixel and object classification in microscopy. *arXiv preprint arXiv:2603.19802*, 2026. URL https://arxiv.org/abs/2603.19802.

Derek Thirstrup et al. BBBC034v1: Induced pluripotent human stem cells. https://bbbc.broadinstitute.org/BBBC034, 2018. Allen Institute for Cell Science dataset; accessed September 25, 2026.

Yuta Tokuoka, Takahiro G. Yamada, Daisuke Mashiko, Zenki Ikeda, Noriko F. Hiroi, Tetsuya J. Kobayashi, Kazuo Yamagata, and Akira Funahashi. 3D convolutional neural networks-based segmentation to acquire quantitative criteria of the nucleus during mouse embryogenesis. *npj Systems Biology and Applications*, 6:32, 2020. doi: 10.1038/s41540-020-00152-8.

Tim Treis, Nikita Moshkov, Johan Fredin Haslum, Shantanu Singh, and Fabian J. Theis. Same encoder, different winner: A paired-view framework for Cell Painting encoder evaluation. *arXiv preprint arXiv:2609.12761*, 2026. URL https://arxiv.org/abs/2609.12761v1. Version 1, September 11, 2026.

Srinivas C. Turaga, Kevin L. Briggman, Moritz Helmstaedter, Winfried Denk, and H. Sebastian Seung. Maximin affinity learning of image segmentation. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper_files/paper/2009/hash/68d30a9594728bc39aa24be94b319d21-Abstract.html.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, Kamil Ugurbil, and WU-Minn HCP Consortium. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, 2013. doi: 10.1016/j.neuroimage.2013.05.041.

Matheus P. Viana, Jianxu Chen, Theo A. Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E. Arakaki, et al. Integrated intracellular organization and its variations in human iPS cells. *Nature*, 613: 345–354, 2023. doi: 10.1038/s41586-022-05563-7.

Bang Wong. Points of view: Color blindness. *Nature Methods*, 8(6):441, 2011. doi: 10.1038/nmeth.1618.

Maria Wyrzykowska, Gabriel della Maggiora, Nikita Deshpande, Ashkan Mokarian, and Artur Yakimovich. A benchmark for virus infection reporter virtual staining in fluorescence and brightfield microscopy. *Scientific Data*, 12:886, 2025. doi: 10.1038/s41597-025-05194-3.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mHC: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025. URL https://arxiv.org/abs/2512.24880v1. Version 1, December 31, 2025.

Derrick Xin, Behrooz Ghorbani, Ankush Garg, Orhan Firat, and Justin Gilmer. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.11379.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2011.09161.

Ji-Hu Zhang, Thomas D. Y. Chung, and Kevin R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of Biomolecular Screening*, 4(2):67–73, 1999. doi: 10.1177/108705719900400206.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2011.10217.

A Evidence Map and Common Definitions

This supplement follows the comparisons that support the main paper: matched updates, SCOPE's learned capabilities, spatial readout controls, and biological-image measurements. The matched forks identify effects of a specified weighting procedure within the smaller encoder families. SCOPE provides a completed training lineage; it is not the treatment arm of those forks. The two studies inform the same evaluation question without supplying one another's missing causal comparison.

Question in the main paper	Evidence here
Does the early preference require co-training the classifier?	B, B.3
How do source, rates, optimizer and availability constrain the result?	C, D
What was trained before and during SCOPE's continuation?	E, E.2
Which culture and brightfield gains survive their controls?	F, F.3
Can readout capacity or additional spatial features change a contrast?	G, G.3
What do counts, strict matching and qualitative fates measure?	H.2, H.5
Do the same conclusions survive input layout and reporter deployment?	I, J
Which analyses and training procedures can the release reproduce?	K

Head and checkpoint identities. *Co-trained* (or *joint*) means updated with the encoder; *unchanged* means applying specified earlier coefficients without another fit. An inherited co-trained classifier and a probe fitted to source features are both eligible unchanged heads, but are different objects. *Fresh* fits a head to frozen features; *repair* continues a saved head on frozen features and allows validation to keep zero updates. Fresh fitting measures same-task decodability when the encoder already saw those task labels, not transfer to a new task. For SCOPE, incoming/source, adapted/intermediate/later and terminal mean steps 256, 10,240 and 26,584. Incoming is already pretrained. For the forks, source denotes the per-seed checkpoint, optimizer and observation position before branching. The older three-seed contrast is retained as developmental evidence, not substituted for the ten-seed estimate.

Inference and chronology. A culture contrast in *percentage points* is 100 times a difference in balanced accuracy. Unless otherwise stated, fork intervals use 10,000 paired training-seed bootstrap draws conditional on the nine reused test wells. They do not add independent biological units. Other studies identify their plate, embryo or animal resampling unit locally. An interval spanning zero is unresolved, not evidence of equality. Interactions are formed from paired differences before resampling; separate significance labels are not compared. Protocols specified before a follow-up's fitting can still follow inspection of the same evaluation population. We state that distinction where it matters.

Two retention rules remain separate. SCOPE's operational guardrail uses nonzero tolerances on development references. The smaller forks' measurement selector uses zero-tolerance validation point estimates to filter candidates. Neither is a statistical noninferiority certificate. The source archive and numerical release retain the full outcome history; this supplement removes repeated displays, not unfavorable conditions.

B Matched Forks and the Unchanged-Head Comparison

B.1 What is matched, and what is measured

The dense planar model has 360,697 trainable parameters including heads; DINOv2-small has 22,209,705 plus provider pretraining. Both have a nuclear segmentation head and a culture classifier. Dense sources receive 3,600 updates and DINOv2 sources 900. The dense source budget was increased after a readiness failure, before its continuation outcomes. Family comparisons therefore combine architecture, pretraining, rates and source budget.

BBBC039 contributes 122/42/36 train/validation/test fields from 12/4/4 plates, with manual instance references. LIVECell contributes 460/128/1,460 images; its nine test wells cover eight cultures, seven with only one well. Inputs use the same observed pixels, grayscale-to-RGB conversion, normalization and symmetric padding without resizing. Target masks, class identifiers and plate identifiers do not enter the encoder. Each fixed-composition update draws two nuclear and two culture

records from permutation streams that exhaust a training pool before repeating. These are separate records, not two labels on one acquisition.

Within family and seed, the branches restore identical weights, AdamW moments and counters, randomness and the next observation suffix. Unless a control states otherwise, AdamW uses $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, decay 10^{-4} and global gradient clipping at five. Encoder rates are 3×10^{-4} for dense and 10^{-5} for DINOv2; heads use 3×10^{-4}. With observed nuclear risk

$$L_n = \{2(0.7L_{\text{Dice}} + 0.3L_{\text{Brier}}) + 0.1L_{\text{BCE}}\}/2.1$$

and culture cross-entropy L_c, the three losses are

$$L_A = (2.1L_n + 0.15L_c)/2, \quad L_N = (L_n + L_c)/2, \quad L_{\text{ref}} = 1.125L_N.$$

Absolute versus normalized weighting changes a 14:1 coefficient ratio to 1:1. Reference scaling restores total coefficient mass, not that ratio. At a common parameter state and common batches, the gradients are $1.05g_n + 0.075g_c$, $0.5g_n + 0.5g_c$, and 1.125 times the latter. These are generally different directions as well as magnitudes. The intervention identifies the complete recipe effect, including the optimizer trajectory it induces, not a scale-independent normalization mechanism. Masking unavailable labels is not a correction for nonrandom annotation availability.

Readout fitting. The unchanged head is the culture LayerNorm–MLP saved with the corresponding source. Fresh linear fitting uses training-only standardization and $C \in \{0.001, 0.01, 0.1, 1, 10\}$, chosen by validation balanced accuracy with smallest-C ties. Fresh MLPs reproduce the original LayerNorm–128-unit–GELU–eight-output architecture. They receive 2,048 CPU FP32 AdamW updates, batch 32, rate 3×10^{-4}, decay 10^{-4} and clipping at five. Validation every 128 updates, including zero, selects the earliest best checkpoint. Initializations and sample orders are paired within seed. This matches head form, not joint versus frozen optimization history. All fitting uses 460 training and 128 validation images; no test outcome selects the head.

B.2 Replication, source weighting and horizon

Seeds 187–196 cross absolute- and normalized-trained sources with both continuation objectives for 4,096 updates. Inherited-state endpoints are 512, 1,024, 2,048 and 4,096; a reset-AdamW follow-up on absolute-trained sources is read at 512 and 4,096. Reset clears moments and counters in both branches while preserving model and sample state. Horizons were not chosen by test outcomes. Rules preceded these new fits but followed the original three-seed study on the same wells.

The saved model digests reproduce the original three sources and six 512-update branches (9/9), and six original reset branches (6/6). The three original joint/linear/MLP differences remain $+9.81/-2.67/-1.52$ points; the seven added seeds average $+2.82/+0.48/+0.75$. Thus the earlier fresh-head reversal is not reproduced by ten seeds; it was not removed by changing the original runs. The ten-seed early co-trained contrasts favor absolute weighting from both source objectives, while both fresh-head contrasts are unresolved. At 4,096, the absolute-source joint contrast is $-3.02\,[-5.81, -0.51]$; both source-specific fresh-MLP intervals are negative. The normalized-source joint and the cross-source linear verdicts remain unresolved. The registered horizon labels use their original shrinkage/persistence rules, not a retrospectively chosen sign test.

Reset changes the early joint contrast from $+4.92$ to $+1.66$, but the paired inherited-minus-reset difference is $+3.26\,[-0.81, 7.40]$. A reset explanation is therefore not established. At 4,096, reset MLP fitting favors normalization, $-0.75\,[-1.46, -0.12]$, while its linear contrast remains unresolved. Table 3 retains the complete source/horizon/reset grid and validation-well rows.

B.3 Compatibility with an unchanged source classifier

No model is trained for this control. In all 480 replay units, co-trained probabilities reproduce bitwise, and the unchanged head equals its source head rather than the branch's updated classifier. Its protocol and classification rules preceded loading targets for this control; the wells were already used elsewhere. Predictions were saved before its scoring. Let R, J, M, L denote the absolute-minus-normalized contrasts through the unchanged, co-trained, fresh-MLP and fresh-linear heads. The interaction $R - J$ or $R - M$ is computed within each seed before bootstrap resampling. At 512 updates, $R = +4.22\,[1.25, 7.49]$ and $R - J = -0.70\,[-1.85, 0.53]$. The unchanged-minus-fresh

Condition	Readout	Updates	Difference [95% interval]; +/−
Inherited, absolute source	Joint head	512	$+4.92$ [$+1.47$, $+8.39$] 8/2
		1,024	$+1.58$ [-1.49, $+5.21$] 5/5
		2,048	-0.08 [-3.30, $+2.53$] 7/3
		4,096	-3.02 [-5.81, -0.51] 3/7
	Joint (validation)	512	$+3.52$ [$+1.48$, $+5.78$] 8/1
		1,024	$+3.20$ [-0.70, $+7.81$] 6/4
		2,048	-1.48 [-4.14, $+1.41$] 2/7
		4,096	-4.06 [-6.64, -1.56] 1/7
	Fresh linear	512	-0.46 [-1.61, $+0.68$] 4/6
		1,024	$+0.10$ [-0.77, $+0.86$] 6/4
		2,048	-0.77 [-1.51, -0.02] 3/7
		4,096	-0.92 [-1.75, -0.11] 4/6
	Fresh MLP	512	$+0.07$ [-1.08, $+1.16$] 5/5
		1,024	-0.13 [-1.17, $+0.87$] 5/5
		2,048	-1.52 [-2.62, -0.45] 2/8
		4,096	-0.95 [-1.37, -0.53] 1/9
Inherited, normalized source	Joint head	512	$+3.23$ [$+0.79$, $+5.67$] 8/2
		1,024	$+0.57$ [-1.81, $+2.71$] 7/3
		2,048	-2.10 [-5.01, $+0.34$] 2/8
		4,096	-2.05 [-4.90, $+1.06$] 3/7
	Joint (validation)	512	$+2.73$ [$+0.08$, $+5.31$] 7/2
		1,024	$+0.78$ [-1.17, $+2.81$] 5/3
		2,048	-4.69 [-8.05, -2.03] 1/9
		4,096	-2.66 [-6.33, $+1.09$] 2/8
	Fresh linear	512	-0.57 [-1.62, $+0.49$] 4/6
		1,024	-1.23 [-2.28, -0.34] 3/7
		2,048	-1.41 [-2.15, -0.71] 1/9
		4,096	-0.85 [-1.67, $+0.00$] 2/8
	Fresh MLP	512	-0.91 [-2.00, $+0.03$] 2/8
		1,024	-1.42 [-2.66, -0.41] 2/8
		2,048	-2.50 [-3.48, -1.54] 0/10
		4,096	-0.83 [-1.66, -0.00] 3/7
Reset, absolute source	Joint head	512	$+1.66$ [-0.03, $+3.56$] 7/3
		4,096	-1.18 [-5.04, $+2.24$] 4/6
	Joint (validation)	512	$+2.81$ [$+0.31$, $+5.78$] 7/2
		4,096	-3.52 [-7.66, $+0.23$] 3/6
	Fresh linear	512	-0.52 [-1.52, $+0.37$] 4/6
		4,096	-0.54 [-1.13, $+0.01$] 4/6
	Fresh MLP	512	-0.47 [-1.52, $+0.45$] 4/6
		4,096	-0.75 [-1.46, -0.12] 3/7

Table 3: Complete ten-seed horizon grid. Entries are absolute-minus-normalized culture balanced accuracy in points, pointwise 95% seed-bootstrap intervals, and positive/negative seed counts; exact ties are omitted. Validation uses eight wells disjoint from the nine reused test wells.

Table 4: **An unchanged earlier head agrees with the co-trained head early and parts from it late.** Absolute-trained sources, inherited AdamW, shared $1\times$ rate. R: unchanged source head; J: co-trained head; M: fresh MLP.

Updates	Retained R	Joint J	Fresh MLP M	$R-J$	$R-M$
512	$+4.22$ [$+1.25$, $+7.49$]	$+4.92$ [$+1.47$, $+8.39$]	$+0.07$ [-1.08, $+1.16$]	-0.70 [-1.85, $+0.53$]	$+4.14$ [$+0.79$, $+7.95$]
1,024	$+2.10$ [-1.43, $+6.38$]	$+1.58$ [-1.49, $+5.21$]	-0.13 [-1.17, $+0.87$]	$+0.52$ [-1.71, $+2.63$]	$+2.23$ [-1.67, $+6.76$]
2,048	$+1.83$ [-2.34, $+5.97$]	-0.08 [-3.30, $+2.53$]	-1.52 [-2.62, -0.45]	$+1.91$ [-0.39, $+4.34$]	$+3.35$ [-1.06, $+8.18$]
4,096	$+0.88$ [-1.75, $+3.36$]	-3.02 [-5.81, -0.51]	-0.95 [-1.37, -0.53]	$+3.90$ [$+1.03$, $+7.52$]	$+1.83$ [-0.99, $+4.56$]

Table 5: **In three seeds, a training-recipe ranking depended on the readout.** Culture balanced accuracy, averaged over all three seeds, on the same 1,460 images in nine test wells. Fresh linear and original-architecture MLP heads are selected on validation only. The MLP is a post-hoc full-grid control. These are descriptive same-task results, not unseen-task transfer or independent-well confidence estimates. Families have different source budgets.

Encoder	Continuation	Jointly fitted	Fresh linear	Fresh MLP
Dense	Source	0.6829	0.8020	0.8014
Dense	Absolute	0.7242	0.8006	0.8148
Dense	Normalized	0.6261	0.8273	0.8300
Dense	Reference-scaled	0.6979	0.8109	0.8211
DINOv2-small	Source	0.9357	0.9582	0.9579
DINOv2-small	Absolute	0.9760	0.9819	0.9800
DINOv2-small	Normalized	0.9741	0.9770	0.9772
DINOv2-small	Reference-scaled	0.9780	0.9786	0.9783

interaction is $+4.14\,[0.79, 7.95]$. Compatibility with the saved classifier is therefore measurable without further classifier training. The fresh fits do not establish the same recipe advantage; they do not prove equal information. Normalized-trained sources also give $R = +3.01$, $J = +3.23$, $M = -0.91$; the scheduled recipe gives $R = +2.24$, $J = +2.76$, $M = -0.74$. The early pattern does not require continuing the source objective.

At 4,096, the paired $R - J$ is $+3.90\,[1.03, 7.52]$, while R alone is unresolved. This is separation in recipe ranking, not necessarily a loss relative to the source. Under the unchanged head, the absolute and normalized arms' source-relative changes are respectively $+1.52\,[-1.17, 4.09]$ and $-2.70\,[-6.43, 0.62]$ at 512; at 4,096 they are $+2.26\,[0.59, 4.07]$ and $+1.38\,[-0.88, 3.97]$. The reported 26-cell control classifies six cells as ranking-only and 20 as neither; none meets its prespecified hidden-loss criterion. Four arm-cells have resolved unchanged-head regression, accompanied by co-trained regression. The per-cell classification thresholds and all 26 source-relative vectors are not included in the released figure data, so that report-level classification is not independently reconstructed here. The release does provide the primary contrast vectors.

These source-relative results keep the empirical interpretation direct: recipe ranking, accuracy relative to the source, and the accuracy gap between heads are different comparisons. The absence of a detected hidden loss is not a bound on smaller losses. One source head per source and one reused culture panel delimit the claim.

B.4 The original three-seed result

The complete pilot grid below records the contrast that motivated the larger study. It is retained for provenance and interpretation of the associated external selectors, not as an additional replication of the ten-seed outcome. The original-architecture refit was post-hoc.

C Optimization and Family Controls

The controls vary the procedure that produces the encoder, not the test population. Their purpose is to locate the early compatibility result: which settings retain it, which alter it, and which proposed explanations remain unresolved. Unless stated otherwise, they use the ten dense absolute-trained sources and the readouts in Appendix B.

C.1 Shared rates and equal-budget per-arm tuning

Both encoder and head rates are multiplied by $m \in \{0.15, 0.5, 1, 2\}$ for 4,096 updates; 512 is the primary endpoint. The $1\times$ observations are reused, not new confirmation. For each seed, arm, endpoint and readout, tuning selects its own validation balanced-accuracy maximum. Ties prefer the rate closest to one on a log scale, then the smaller rate. Each arm has the same candidate grid and fitting allowance, but different readout procedures can select different encoders. All 180 new and 40 reused linear fits complete. Seed-187 model digests reproduce the original endpoints.

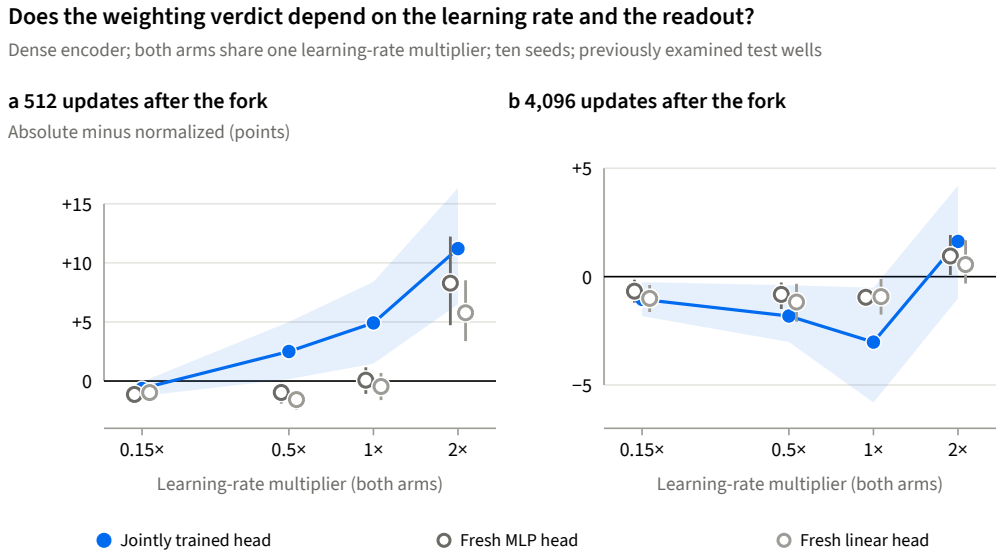

Figure 3: **The early preference depends on the optimization allowance.** Absolute-minus-normalized culture balanced accuracy under shared rates and readout-specific validation tuning. Means and pointwise 95% seed intervals condition on the nine reused test wells. Lower shared rates give concordant late directions; the $2\times$ condition prevents extending that result to the entire grid. The complete numerical grid is in main Table 1.

The main Table 1 retains the full shared-rate grid. At $0.15\times$, joint contrasts are $-0.63\,[-1.27, -0.05]$ and $-1.05\,[-1.82, -0.25]$ at 512 and 4,096. At $0.5\times$/512, joint/linear/MLP differences are $+2.50/-1.57/-0.97$ points. The paired joint-minus-linear interaction is $+4.07\,[1.72, 6.67]$ and joint-minus-MLP is $+3.47\,[0.87, 6.32]$. At $1\times$, these interactions are $+5.38\,[1.40, 9.45]$ and $+4.84\,[1.13, 8.99]$. Five of six early interactions at $0.5\times$–$2\times$ have pointwise intervals excluding zero; $2\times$ joint-minus-MLP does not. None at $0.15\times$/512 or any rate/4,096 excludes zero. Those unresolved intervals are not equivalence.

Original-grid per-arm tuning gives early joint difference $-0.37\,[-1.32, 0.55]$ and unresolved fresh-head differences. At 4,096, tuned joint/linear/MLP contrasts are $-1.67\,[-3.20, -0.29]$, $-0.67\,[-1.19, -0.12]$ and $-1.01\,[-1.61, -0.38]$. Shared rates $0.15\times$–$1\times$ likewise favor normalization late under all three readouts, but $2\times$ MLP favors absolute weighting, $+0.95\,[0.07, 1.92]$.

Multiplicity and population sensitivity. The grid uses pointwise intervals, not 30 simultaneous discoveries. A post-hoc exact paired sign-flip sensitivity assumes independent sign symmetry conditional on magnitudes; it is not justified by zero means alone. At $0.5\times$/512 the joint and MLP two-sided p values are 0.096 and 0.053. Ten signs give minimum attainable $2/2^{10}$, larger than Holm's initial $0.05/30$ threshold (Holm, 1979), so this sensitivity cannot reject the full family at that resolution. The estimates remain informative without a grid-wide discovery claim.

Dropping any one culture retains the early original joint preference ($+4.60$ to $+5.46$) and the late original MLP's negative interval. Dropping Huh7 or SKOV3 removes nominal resolution of the late joint and early normalized-source joint contrasts; without Huh7, late joint is $-2.27\,[-5.02, 0.06]$. These are fixed-panel sensitivities, not replication on new wells. Protocols and selection formulas preceded new fits, but no immutable pre-score rate-selection record was found: choices were persisted during analysis after scoring. The formulas use validation only. This supports the reported computation, not enforced blindness to a reused test.

C.2 Extending the tuning range

The original joint selector chose $0.15\times$, its lower boundary, for nine of ten normalized arms. A follow-up adds $0.05\times$ and $0.015\times$ to inherited training, and $0.05\times$ and $0.15\times$ to scheduled training. Both implementations reproduce the archived seed-187 $1\times$ state. All new forks and 160 readouts finish before scoring. One of ten normalized-arm choices remains at the extended lower boundary: censoring is reduced, not eliminated.

The extended inherited grid gives tuned joint $-0.92\,[-1.83, -0.10]$ at 512; tuned linear and MLP means are -0.78 and -0.33, both unresolved. At 4,096 all three favor normalization (joint $-1.99\,[-3.53, -0.53]$, linear -0.67, MLP -1.01). The new inherited shared rates give early joint contrasts -1.15 at $0.05\times$ and -1.02 at $0.015\times$. Every scheduled-grid tuned contrast remains unresolved at both endpoints; the early joint result is $-0.04\,[-1.71, 1.46]$. Thus the early shared-rate absolute preference is not retained under equal per-arm tuning within either tested grid. No globally optimal rate is claimed.

C.3 Head warmup, separate rates and bounded repair

Head warmup performs 2,048 absolute-weighted updates on a frozen encoder before forking. It preserves encoder weights and moments but advances the sample stream and head history; warm versus cold is not an initialization-only, observation-matched intervention. Its early joint contrast is $+4.03\,[0.50, 7.08]$, and cold-minus-warm change $+0.89\,[-4.70, 6.20]$ is unresolved. Validation head accuracy rises from 0.713 to 0.791 during warmup, then drops after branching. At 512, warm-minus-cold validation differences are unresolved for both arms. Warm-at-fork does not establish continuing head convergence.

Reading absolute- versus normalized-trained sources before an objective switch gives joint/linear/MLP contrasts $+4.36\,[-0.16, 9.99]$, $+0.99\,[-1.66, 3.68]$ and $+2.37\,[-0.26, 5.28]$, all unresolved. Raising only the head rate by $4\times$ or only the encoder rate by $2\times$ also leaves the predicted change in the early joint-minus-MLP discrepancy unresolved: $-2.72\,[-6.88, 1.81]$ and $-1.02\,[-5.50, 3.58]$. Faster heads retain a 14.6-point joint-to-MLP gap versus 11.9 originally; faster encoders instead favor absolute weighting through all readouts. These outcomes do not identify encoder/head lag as the mechanism.

The original three-seed repair experiment freezes each encoder and continues a saved joint head, or a paired random initialization, with fresh AdamW, 460 labels, batch 32, rate 3×10^{-4} and validation selection. Budgets are $0, 128, 256, 512, 1024, 2048$; selection includes zero updates. After 128 repair updates, normalized and absolute dense means are 0.8107 and 0.7928, with all three paired differences positive. DINOv2 does not reverse. This is a bounded post-hoc repair result in the three-seed subset whose early fresh-head reversal did not replicate; it is not promoted to the ten-seed headline. Moment-rescaling and reduced-rate pilot branches likewise remain developmental controls. Reset results are reported in Appendix B.2 rather than counted twice.

C.4 Fresh AdamW with warmup and cosine decay

Both arms retain the ten absolute-trained dense sources and sample states, but clear AdamW moments and counters. Encoder and head peak rates are $m\,3 \times 10^{-4}$, $m \in \{0.5, 1, 2\}$, multiplied at update u by

$$f(u) = \begin{cases} u/256, & 1 \le u \le 256, \\ \{1 + \cos[\pi(u - 256)/3840]\}/2, & 256 < u \le 4096. \end{cases}$$

All 60 forks and 120 readouts complete. Source model hashes, initial validation outputs and ordered record digests match the preceding study. Reset and schedule change jointly relative to inherited constant-rate training. The prespecified primary is the joint-minus-MLP weighting interaction at $1\times$/512. Per-arm tuning follows the same readout-specific validation rule, with equal three-rate budgets and fixed tie-breaking. The primary interaction is $+3.50\,[1.97, 5.40]$, positive in all ten seeds; joint-minus-linear is $+3.11\,[1.20, 5.11]$. Inherited moments and a constant rate are therefore not necessary for an early interaction in this setting. The paired change from inherited constant rates is unresolved, $-1.34\,[-6.39, 3.57]$, as is the change from constant-rate reset, $+1.37\,[-1.24, 4.60]$. Other shared rates and all tuned interactions are unresolved. At the late endpoint the interactions are near zero, with intervals still allowing nonzero effects; duration and schedule phase change together.

Peak	Updates	Joint	Linear	MLP	D_{MLP}
0.5×	512	$+0.72\ [-1.15, 2.88]$	$-1.01\ [-1.99, -0.12]$	$-0.61\ [-1.15, -0.18]$	$+1.33\ [-0.75, 3.80]$
1×	512	$+2.76\ [0.86, 4.91]$	$-0.34\ [-1.32, 0.55]$	$-0.74\ [-1.56, -0.03]$	$+3.50\ [1.97, 5.40]$
2×	512	$+5.00\ [-2.36, 12.95]$	$-0.76\ [-2.21, 0.69]$	$-0.85\ [-2.29, 0.90]$	$+5.85\ [-0.44, 12.23]$
Tuned	512	$+0.88\ [-1.10, 3.36]$	$-0.61\ [-1.98, 0.62]$	$-0.72\ [-1.59, 0.14]$	$+1.60\ [-0.51, 4.13]$
0.5×	4,096	$-0.48\ [-0.98, -0.01]$	$-0.62\ [-0.94, -0.23]$	$-0.58\ [-1.03, -0.14]$	$+0.10\ [-0.23, 0.39]$
1×	4,096	$-0.26\ [-0.83, 0.34]$	$-0.47\ [-0.86, +0.00]$	$-0.44\ [-0.89, -0.06]$	$+0.18\ [-0.17, 0.51]$
2×	4,096	$-0.57\ [-1.05, -0.12]$	$-0.39\ [-0.84, 0.11]$	$-0.56\ [-1.31, 0.15]$	$-0.01\ [-0.78, 0.55]$
Tuned	4,096	$-0.19\ [-0.90, 0.62]$	$-0.38\ [-0.87, 0.15]$	$-0.27\ [-0.80, 0.18]$	$+0.08\ [-0.80, 0.93]$

Table 6: **Shared-rate and validation-tuned culture contrasts under reset AdamW with warmup and cosine decay.** All values are percentage points, with pointwise 95% seed-bootstrap intervals. D_{MLP} is the paired joint-minus-MLP weighting interaction, not a difference of significance labels. Joint-minus-linear interactions are retained in the aggregate. Rounded-zero bounds are classified before rounding: the 1×/4,096 linear interval includes zero. No nuclear test readouts were collected for this recipe.

Table 7: **The matched-fork discrepancy persists under both clipping controls; the variable-availability outcome depends on the control.** Discrepancy D is the joint contrast minus the refit contrast at 512 updates, in points; Δ is its seed-paired change from the original global clip.

		Matched A−N		Variable E−A	
Clip	Refit	D	Δ	D	Δ
global	MLP	$+4.84$ $[+1.13, +8.99]$	–	-6.38 $[-9.18, -3.11]$	–
none	MLP	$+3.49$ $[+0.74, +6.28]$	-1.35 $[-6.73, +3.47]$	-4.11 $[-9.66, +2.53]$	$+2.27$ $[-3.70, +9.73]$
group	MLP	$+6.46$ $[+4.35, +9.03]$	$+1.62$ $[-2.37, +5.51]$	-4.75 $[-8.11, -1.00]$	$+1.63$ $[-1.00, +4.62]$
none	linear	$+4.65$ $[+1.79, +7.43]$	-0.73 $[-5.98, +4.34]$	-5.79 $[-11.43, +0.79]$	$+0.52$ $[-5.37, +7.86]$
group	linear	$+7.13$ $[+4.75, +9.81]$	$+1.75$ $[-1.74, +5.45]$	-5.73 $[-8.83, -2.09]$	$+0.59$ $[-1.74, +3.10]$

BT474's single test well contributes 59% of the primary point estimate. Removing it gives $+1.62\,[-1.00, 4.53]$; other single-class removals retain a positive interval. On separate validation wells, the early joint contrast is $+0.23\,[-3.67, 3.75]$. Neither sensitivity establishes a test–validation difference. No nuclear test outputs were collected for this recipe. The public aggregate preserves all seeds, both horizons, calibration and selection outcomes; no immutable pre-score rate-selection record was found.

C.5 Clipping the encoder and heads separately

The early normalized branches exceed the whole-gradient clip threshold on 54–59% of updates versus about 1% for absolute branches. Separate *group* clipping caps encoder and head groups independently at five; *none* removes clipping. The ten-seed matched A−N and variable E−A comparisons each run to 4,096. All 100 forks and 200 readouts finish; the original-clip implementation reproduces five archived seed-187 states. The matched joint-minus-MLP interaction remains positive without clipping, $+3.49\,[0.74, 6.28]$, and with separate clipping, $+6.46\,[4.35, 9.03]$, positive in all ten seeds in the latter. Global head–encoder norm coupling is not necessary for the observed matched discrepancy. Its full causal contribution is not identified: changes from global clipping remain unresolved. In variable availability, separate clipping retains the discrepancy; no clipping leaves both the interaction and its change unresolved. Eight seed contrasts are negative, while two unstable refits widen the interval. Excluding them would change the specified comparison.

Unclipped 1:1 branches reach gradient norms around 60–100. Variable normalized fresh-MLP accuracy at 512 falls from 81.1 to 70.9, while the absolute arm changes by less than 0.2 points. Thus clipping changes the learned features and refits as well as a head's immediate gradient. A clip factor

rescales gradients before AdamW; its moment history prevents interpreting that factor as a scalar on the final parameter step. The tests use one threshold and one dense family, with unadjusted intervals.

C.6 DINOv2 PROVIDES A FAMILY BOUNDARY

DINOv2 repeats both source weightings, seeds 187–196, the same records, readouts and horizons, at its own rates. Nine original model-state digests match; that does not assert equality of full optimizer/RNG payloads. Neither early joint contrast is resolved. At 4,096, absolute-source joint, linear and MLP intervals favor normalization, but normalized-source intervals remain unresolved. A cross-source decision that requires both source conditions therefore remains unresolved. At 512,

Table 8: **DINOv2-small at its own learning rates.** Absolute-minus-normalized culture balanced accuracy (points), ten-seed means and pointwise 95% seed-bootstrap intervals conditional on the nine reused test wells. The absolute-source late result is negative under every readout; the cross-source decision remains unresolved.

Source	Updates	Joint head	Fresh linear	Fresh MLP
Absolute	512	$+0.89\,[-1.21, +3.90]$	$-0.10\,[-1.11, +0.76]$	$-0.39\,[-1.53, +0.56]$
Absolute	4,096	$-0.41\,[-0.73, -0.09]$	$-0.38\,[-0.73, -0.03]$	$-0.35\,[-0.69, -0.03]$
Normalized	512	$-0.46\,[-2.19, +0.96]$	$-0.39\,[-1.42, +0.40]$	$-0.33\,[-1.31, +0.52]$
Normalized	4,096	$-0.06\,[-0.50, +0.35]$	$-0.05\,[-0.39, +0.32]$	$+0.04\,[-0.30, +0.38]$

DINO's joint-to-MLP gaps are 0.1–1.4 points (mean 0.6), versus 5.3–14.3 for dense; early clipping affects 1–3% of its updates. These are associations, not isolated architectural mechanisms. Raising only the DINO encoder rate by $10\times$ or $30\times$ makes joint culture validation reach chance within 128 updates while nuclear Dice stays near 0.95. After observing these trajectories, a validation-based collapse screen is specified before fresh refits; the later $2, 3, 5\times$ study uses it from the start. It flags a multiplier/horizon when at least six seeds have either arm's fresh-MLP validation accuracy below 0.5. The $5, 10, 30\times$ settings fail and are not treated as equivalence.

At $2\times$/512, all readout contrasts are unresolved. At $3\times$, all favor absolute weighting: joint $+14.84\,[2.09, 27.55]$, linear $+7.29\,[4.06, 10.47]$ and MLP $+20.19\,[10.75, 30.13]$. At both passing rates and horizons, paired joint-minus-refit intervals span zero. This is absence of a resolved interaction, not equivalent readouts. The failed high-rate manipulation and the same-direction $3\times$ result limit a universal head-lag explanation.

D VARIABLE AVAILABILITY AND THE JOINT-TASK COMPARISON

The fixed-composition forks compare task coefficients but never vary the number of tasks contributing to an update. This control varies that number and compares fixed equal coefficients with task-count normalization. It tests a constructed update-level availability regime, not two annotations of the same acquisition, and does not hold mean loss scale fixed.

Episodes and matched sources. Ten absolute-trained dense sources, seeds 187–196 at 3,600 updates, each produce three continuations. All three restore the same parameters, AdamW state, sampler state and randomness. A separate seeded generator chooses nuclei only, culture only or both, each with probability $1/3$; its 4,096-update schedule is identical across arms within a seed. Each present task draws two training records. An absent task draws none and executes no forward or loss. Thus "both" means two nuclear records and two separate culture records, not two labels on one image. The sampler traverses shuffled training pools and does not class-balance culture labels. Each update accumulates present-task gradients before one optimizer step. Thirty forks yield 60 encoder/horizon conditions at 512 and 4,096 updates; none failed.

Targets and weighting rules. The tasks remain nuclear foreground segmentation and eight-class culture classification, with no reconstruction target. Let L_n be the mean canonical nuclear risk, $\{2(0.7L_{\mathrm{Dice}} + 0.3L_{\mathrm{Brier}}) + 0.1L_{\mathrm{BCE}}\}/2.1$, and L_c mean culture cross-entropy. The nuclear loss excludes invalid padding before evaluating losses; an observed all-background reference is negative

supervision, not a missing task. Let I_n, I_c indicate task presence and $K = I_n + I_c \geq 1$. The three update losses are

$$\begin{aligned} L_A &= 1.05 I_n L_n + 0.075 I_c L_c, \\ L_E &= \tfrac{1}{2} I_n L_n + \tfrac{1}{2} I_c L_c, \\ L_N &= (I_n L_n + I_c L_c)/K. \end{aligned} \tag{D.1}$$

Here A retains the absolute coefficients, E uses equal fixed coefficients, and N normalizes equal base coefficients over present tasks. N is not the normalization of A's original unequal coefficients.

With an always-both schedule, A reduces to the original absolute update and E and N to the original normalized update. Before the new forks, seed 187's 512-update A and N continuations reproduced the corresponding earlier model states exactly. CPU checks also verify the E reduction and absent-task gradient semantics. Regenerating the variable schedules and restoring each source sampler reproduces the same ordered record-selection digests for all three arms at both horizons.

Normalization changes coefficient exposure. For probabilities (p_n, p_c, p_b) of the three subsets, E's expected nuclear and culture coefficients are $(p_n + p_b)/2$ and $(p_c + p_b)/2$; N's are $p_n + p_b/2$ and $p_c + p_b/2$. Under the registered uniform law,

$$\mathbb{E}[(\alpha_n, \alpha_c)] = \begin{cases} (0.7, 0.05), & A, \\ (1/3, 1/3), & E, \\ (1/2, 1/2), & N. \end{cases} \tag{D.2}$$

N therefore has 1.5 times E's expected total coefficient mass; A has 1.125 times E's. These are coefficients, not gradient-norm fractions. For realized subset counts (a, b, c), E's cumulative coefficients are $((a + c)/2, (b + c)/2)$ and N's are $(a + c/2, b + c/2)$, so finite schedules need not give equal aggregate task exposure. Across the ten full schedules, the counts are $(13394, 13709, 13857)$; mean per-update coefficients are $(0.332654, 0.336499)$ for E and $(0.496155, 0.503845)$ for N.

Consequently N minus E tests a normalization policy, including its mean scale change, not a scale-independent mechanism. E minus A changes both the task ratio and absolute magnitudes, not the ratio alone. Even identical task losses $L_n(\theta) = L_c(\theta) = \theta^2/2$ give expected gradients $2\theta/3$ under E and θ under N: different finite updates need not indicate task conflict or correction of annotation bias. Neither rule is a missing-not-at-random correction. The identity $(N - A) = (N - E) + (E - A)$ for endpoint contrasts does not identify a unique decomposition into ratio and normalization mechanisms.

Optimizer scope. Both learning rates are 3×10^{-4}, with AdamW weight decay 10^{-4}, inherited moments and clocks, BF16 forward autocast, and global gradient clipping at 5 after accumulation. Absent heads have no gradient, so their AdamW moments, clocks and weight decay do not advance. Observation and head-update exposure are matched across the new arms but lower than in an always-both continuation of equal step count. Averaged over seeds, clipped update fractions for A/E/N are $0.0066/0.4098/0.4871$ through 512 and $0.0052/0.2997/0.3502$ through 4,096. Clipping and moment changes are possible pathways of the weighting intervention, not pre-treatment confounders. These summaries do not isolate either pathway; Appendix C.5 tests clipping directly. AdamW does not guarantee invariance to current gradient rescaling with unchanged inherited moments and subset-dependent multipliers.

Readouts and uncertainty. Each snapshot receives its joint culture head, a fresh validation-selected linear classifier and a fresh LayerNorm–MLP. Fitting uses the same 460 training and 128 validation images; scoring uses 1,460 images from nine test wells. The pipeline requires completion records for all predictions before scoring; the panel had already been examined. Differences are formed within each training seed, then bootstrapped with 10,000 draws. Intervals are pointwise, unadjusted 95% intervals conditional on these wells, not independent biological replication or equivalence tests. The registered decision rule allows a detection in either fresh readout; it does not require both to agree. An independent CPU replay reproduces the culture contrasts and paired joint-minus-refit intervals from the saved seed rows.

Results and consequence. At 512, joint heads favor A over E and N, while all fresh-head intervals include zero (Table 9). The paired joint-minus-MLP difference for E−A is -6.38 points

Contrast	Readout	512 updates	4,096 updates
N−E	Joint	$+0.25$ $[-3.36, 3.54]$	$+0.91$ $[-0.96, 2.80]$
	Fresh linear	-0.26 $[-1.32, 0.74]$	-0.48 $[-1.08, -0.03]$
	Fresh MLP	-0.15 $[-0.84, 0.56]$	-0.03 $[-0.62, 0.59]$
E−A	Joint	-6.56 $[-9.39, -3.45]$	-0.13 $[-2.43, 2.10]$
	Fresh linear	-0.24 $[-1.45, 1.11]$	$+1.14$ $[0.35, 2.07]$
	Fresh MLP	-0.18 $[-0.85, 0.57]$	$+0.18$ $[-0.75, 0.89]$
N−A	Joint	-6.31 $[-9.95, -2.91]$	$+0.78$ $[-1.96, 3.13]$
	Fresh linear	-0.50 $[-1.63, 0.58]$	$+0.66$ $[0.09, 1.29]$
	Fresh MLP	-0.33 $[-0.93, 0.37]$	$+0.15$ $[-0.35, 0.66]$

Table 9: Paired culture balanced-accuracy contrasts in percentage points, with pointwise 95% seed intervals. Positive values favor the first named arm. The small late N−E linear contrast favors E; its MLP counterpart is unresolved, not equivalent.

Updates	LR factor	Culture BA (pp)	Nuclear F1 (pp)	Count MAE
512	0.15	-0.63	-0.13	-0.275
512	0.5	+2.50	-0.04	+0.447
512	1	+4.92	-0.20	+0.253
512	2	+11.20	+0.25	-0.025
4,096	0.15	-1.05	+0.06	-0.033
4,096	0.5	-1.82	+0.17	-0.064
4,096	1	-3.02	-0.05	+0.278
4,096	2	+1.63	+0.42	-0.178

Table 10: All paired-task means on the fixed nine culture wells and four nuclear plates. Differences are absolute minus normalized; accuracy and F1 are percentage points, count MAE objects per image. Positive accuracy/F1 and negative count-error differences favor absolute weighting. No confidence interval or new selection rule is implied.

$[-9.18, -3.11]$, so the early readout discrepancy is not inferred merely from different detection labels. At 4,096, fresh linear readouts favor E over N by 0.48 points and both E and N over A. Every fresh-MLP interval still includes zero. Detection only at the later linear endpoint is not itself evidence that the effect changed between horizons, nor does the linear/MLP detection difference establish a readout interaction.

The nuclear readout remains the joint decoder on 36 reused test fields, not a fresh nuclear fit. Its descriptive foreground-Dice contrasts include small adverse values for N−A: -0.062 points $[-0.135, -0.008]$ at 512 and -0.036 $[-0.059, -0.017]$ at 4,096. Late E−A is also negative, -0.057 $[-0.091, -0.027]$. Object-F1 and count-MAE intervals include zero for all three contrasts at both horizons; they do not certify retention or equivalence. Culture accuracy alone therefore does not identify an overall multitask winner.

This control establishes a variable-update-availability comparison and a small late linear-readout difference between normalization policies. It does not establish a pure ratio cause, a scale-independent benefit of normalization, a general feature-information ordering, or performance under co-observed targets on one acquisition. The original singleton-record derivation remains valid for the original fixed-composition forks; the present intervention has a different normalization unit and exposure law.

D.1 Both tasks belong in a recipe comparison

A post-hoc join pairs culture and nuclear outcomes from the original ten seeds, four shared rates and two horizons. It fits no model and selects no recipe. Nuclear F1 and count error use the existing co-trained decoder, not a fresh nuclear fit. At $1\times$/512, the $+4.92$-point culture contrast accompanies -0.20 nuclear-F1 points and $+0.253$ objects of count MAE. At $2\times$, both mean accuracy coordinates favor absolute weighting. The complete means below retain every rate/horizon; they are not Pareto-dominance tests or biological effect-size thresholds.

E SCOPE: ARCHITECTURE, TRAINING HISTORY AND OBJECTIVE

SCOPE is the longitudinal case study, not an architectural treatment in the small-model forks. The settings below specify the executed one-pass model. They distinguish stored compatibility modules, trainable parameters and operations that affect its predictions.

E.1 SELECTED SPATIAL PROCESSOR

Component	Consolidation setting
Typed pyramid	Widths $32, 64, 128, 256$; strides $(1, 2, 2)$, $(2, 2, 2)$, $(2, 2, 2)$; polyphase residuals enabled; channel-statistics fusion disabled.
Spatial bottleneck	28 blocks of width 1536; four residual streams.
Attention projections	24 query heads of width 64; query rank 768; one shared local K/V projection; separate compressed K/V and CSA index banks.
Attention banks	Local window $3 \times 7 \times 7$; CSA blocks $1 \times 2 \times 2$, top-32; HCA blocks $2 \times 8 \times 8$, all valid blocks.
Rotary coordinates	60 rotated dimensions, ten pairs per axis; fixed frequencies $\pi/2$ to 8π; fixed modality-specific reference lengths.
Output projection	Four groups of six heads, rank 384 per group; not rank-reducing relative to each 384-dimensional group input.
Feed-forward layers	14 MoE layers: 32 experts, top-4, routed width 1664, shared width 2304. The other 14 layers are dense, width 2304.
Workspace	Eight width-256 slots; updates after blocks $8, 11, 14, 17, 20, 23$; final read-only binding.
Decoder and mask features	Four pyramid skips; native mask-feature width 64; feature-axis normalization in the corrected query head, transition 1.0.
Geometry runtime	Explicit FP32 path, overriding the inherited legacy precision field.

All block indices are zero-based. Local-only attention occurs at $0, 1, 27$, CSA at even blocks 2–26, and HCA at odd blocks 3–25. The 13-CSA/12-HCA schedule is independent of the 14-MoE/14-dense schedule. The field `retained_blocks=24` controls activation retention, not depth. One pass is one student encoder evaluation for a view; teacher forwards during training are additional computation.

Spatial coordinates and compression. A token index j is anchored at native voxel $\mathbf{s} \odot j$, $\mathbf{s} = (4, 8, 8)$, and world position $r_j = (A\overline{\mathbf{s} \odot j})_{1:3}$. Crop, padding and flip transforms compose into the affine A. Grid indices determine local neighborhoods and block membership; physical coordinates determine RoPE. The encoder uses millimeters and dimensionless coordinates $\xi_j = L_m^{-1}(r_j - o)$, with fixed modality reference lengths and the view's valid-token mean o. Recentered rotary coordinates are not used to register separate views. A singleton depth axis is an input interface, not measured slice thickness.

Polyphase rearrangement preserves every input sample after zero padding, but its learned channel-reducing projection is not invertible in general. For a compressed block, featurewise softmax weights pool projected values before RMSNorm; its positional anchor is instead the mean of valid block coordinates. Thus a small block does not supply an exact pooling–rotation commutation guarantee. Keys are rotated and values are not. The separate CSA index bank scores all fine blocks and learns from detached dense-attention block mass; top-32 limits selected attention, not all candidate scoring. For an $8 \times 20 \times 20$ token grid the fine/coarse bank counts are 800/36. No linear end-to-end scaling claim follows from those counts.

Transport, experts and query readout. At each token, attention and feed-forward sublayers separately read four streams and update $H' = T^\top H + pb^\top$, with $T = (1 - \alpha)I + \alpha\,\mathrm{Sinkhorn}_{12}(L)$ and $\alpha = (1 - 0.1^{1/62})/2$. Reads sum to one; positive writes have mean one away from numerical floors. The identity component controls a fixed transport path, not the full input-dependent Jacobian. Finite Sinkhorn is approximate, and an outer workspace gate need not preserve an ideal candidate's contraction. These operator bounds are not used as evidence of the empirical readout effects.

For expert scores $s_e = \sqrt{\mathrm{softplus}(w_e^\top x)}$, selection uses the four largest $s_e + b_e$, while output weights normalize s_e over those four. The correction bias affects selection, not output weighting.

Contributions are restored to token-assignment order and summed in FP32; the expert calculations themselves are not all FP32. The change retains 32 stored experts and does not establish specialization or an isolated top-4 benefit.

Eight slots exchange information with tokens after the designated blocks; final binding reads the last tokens and surviving slots without another slot update. Query-to-slot weights mix normalized slot supports into a spatial prior. Content attention and a separate prior-weighted read refine the normalized label query in its tangent space before scoring the feature-normalized native mask field. The inherited unary correction and trained static residual readouts remain distinct from the encoder slots. Earlier action modules are frozen and no candidate search, regional read, policy rollout or GRPO loss is executed. Bias-free compatibility projections may evaluate $h - h_0 = 0$, yielding zero for finite operands rather than a learned iterative correction.

E.2 Inherited training history

Consolidation step 256 is not 256 updates from random initialization. The source-linked stages below are local or inherited counters with branches and resets; adding them would not reconstruct lifetime training. The oldest foundation exposure and intermediate action restarts remain incomplete. A final-versus-scratch comparison consequently tests complete fitted systems, not consolidation or scale alone. The relationship stage records 115,200 episode presentations over 12,800 updates with nine episodes per update; overlapping lane catalogs are not independent acquisitions. Its output initializes supervised post-training, whose selected step 4,480, rather than the later completed step, is inherited. The reported terminal endpoint is the completed top-4 schedule at 26,584. Original external studies keep their fixed 256/10,240 checkpoints; terminal culture, nuclear and larva evaluations are explicitly post-hoc extensions.

E.3 Observed training population and sampling

The retained 319,612 contracts comprise 318,998 ordinary and 614 new-domain records. Protected Allen, paired, biology and retention pools overlap this population; expanded views are not additional patients, wells or organisms.

Resource attribution. The named public resource families include Allen imaging (Allen Institute for Cell Science, n.d.; Viana et al., 2023), BBBC034 (Thirstrup et al., 2018), OpenCell (Cho et al., 2022), the JUMP pilot (Chandrasekaran et al., 2024), Light My Cells (Kauffmann et al., 2026), Cytoland (Liu et al., 2025), DynaCell (Biohub, 2026), and VIRVS (Wyrzykowska et al., 2025). These references credit the resource families, not the study-specific expansions or counts. The internal aliases and exact selected releases require the corresponding acquisition manifests; in particular, a family name alone does not resolve private or transformed source records.

The new-domain pool comprises 460 LIVECell, 83 annotated Arabidopsis and 71 unlabeled plant records. Protected Allen has 2,006 records, retention 1,750, biology 7,330 and paired 39,420. Every rank receives two protected Allen views, four ordinary and one retention record; every second update adds one record from each nonempty new, biology and paired pool. Thus

$$L_t = \frac{1}{3}\sum_{r=1}^{3}\frac{1}{m_t}\sum_{j=1}^{m_t} L_{rjt}, \qquad m_t \in \{7, 10\}. \tag{E.1}$$

Absolute within-episode coefficients do not remove the effects of episode averaging, pool overlap or frequency. The completion record's stream coverage is not exhaustive spatial coverage. Structural MRI supplies another scale/contrast regime, with no isolated efficacy or clinical claim. No CT or NPH observations were admitted; vocabulary entries alone are not exposure. Dataset research-use restrictions remain applicable.

E.4 Losses and target eligibility

The coefficients are those of main Equation (1); all task scales are one. Only eligible losses are evaluated. Segmentation validity intersects target and spatial masks. A wholly unannotated crop has no segmentation term, whereas an observed empty crop is supervised. For each observed query on

Stage	Selected incoming/outgoing state	Observations, objectives and trained scope
Multidomain foundation	Imported checkpoint at inherited step 75,520	Allen structures, paired and unpaired microscopy, owned volumetric images and structural MRI. The source configuration specifies segmentation, BF–FL feature prediction and EMA self-supervision with top-2 routing. Earlier initialization and cumulative exposure are not fully reconstructed.
Relationship pretraining	Foundation 75,520 → local 12,800	Protected Allen, paired images, replay and annotated-protein image lanes; segmentation, fluorescence reconstruction, self/context/paired features, localization and a sequence auxiliary. Full trainable visual model, three ranks, nine episodes per update.
Supervised post-training	Relationship 12,800 → selected 4,480	Structure and generic segmentation, reconstruction, localization and fixed-encoder replay. First 1,280 updates train heads; subsequent updates open the upper eight encoder blocks. The separate run reached 12,800, but the next stage takes step 4,480, not its terminal state.
Native continuation	Post-training 4,480 → local 10,996	The same task families; native stem, decoder, upper four blocks and selected heads train, while router and workspace remain frozen. One device; 582,123,311 parameters were reported trainable at admission.
Action-training ancestry	Native 10,996 → consolidation input at local step 9,543	Shared multitask targets plus region-reading and policy objectives. Contracts record the original joint-action predecessor at 7,153 and later executor migrations. This reconstruction does not close every intermediate restart; the final consolidation does not execute actions or GRPO.
Passive consolidation	Action input 9,543 → local 4,356	One-pass multitask supervision and new instance targets. The selected output initializes the top-4 branch; its 36,224-update plan was not completed.
Top-4 warmup and recovery	Reset branch counter $0 \to 128 \to 256$	The retained catalog of Table 12; new culture head, absolute task coefficients and four selected experts. Encoder warmup ends at 128. Recovery supplies the step-256 source used in the external comparisons.
Reported continuation	$256 \to 26,584$	Full eligible backbone, decoder and heads with the same optimizer, EMA and sampler state. The stopping policy changes; the strict retention thresholds do not. Step 10,240 is the original adapted checkpoint and 26,584 the post-hoc terminal extension.

Table 11: Source-linked training history recovered from configuration, handoff, migration and completion records. These are documented stages, not a controlled attribution of their benefits. A saved plan is not counted as completed training. The oldest foundation history and intermediate action restarts remain incomplete.

nonempty valid support V_q, write $p_v = \sigma(z_v)$ and $y_v \in \{0, 1\}$:

$$D_q = \begin{cases} 1 - \dfrac{2\sum_{V_q} p_v y_v + 10^{-6}}{\sum_{V_q} p_v + \sum_{V_q} y_v + 10^{-6}}, & \sum_{V_q} y_v > 0, \\ |V_q|^{-1} \sum_{V_q} p_v, & \sum_{V_q} y_v = 0, \end{cases} \tag{E.2}$$

$$R_q = 0.7 D_q + \frac{0.3}{|V_q|} \sum_{V_q} (p_v - y_v)^2, \tag{E.3}$$

$$B_q = |V_q|^{-1} \sum_{V_q} [\mathrm{softplus}(z_v) - y_v z_v]. \tag{E.4}$$

The squared probability term is the binary Brier score (Brier, 1950). The separate query means receive coefficients 2 and 0.1, not a coefficient of 0.2 on BCE. Its additional logit derivative is $0.1(p_v - y_v)/(Q_e|V_q|)$ for Q_e observed queries. It retains a corrective signal for a wrong finite logit

Family	Retained categories and counts	Total
Allen hiPSC	Atomic 8,128; paired 5,910; structure 2,006	16,044
Allen EMT	Atomic volumetric 1,576	1,576
BBBC034	Atomic 4; segmentation 1	5
Brain organoids	Atomic 1,275; segmentation 1,275	2,550
BF Cell Painting	Atomic 1,298; paired 558; registered target-student 461	2,317
Cardio	Atomic volumetric 22	22
CytoLand	Atomic 426; paired 284	710
DynaCell	Atomic 2,739; paired 1,826	4,565
FNet TOMM20	Atomic 160; paired 80	240
JUMP pilot	Atomic 14,448; paired 9,030	23,478
Light My Cells	Atomic 41,807; paired 3,827	45,634
OpenCell	Atomic planar 17,198; volumetric 5,054	22,252
QPI	Unlabeled atomic 1,097; segmentation 474	1,571
VIRVS	Atomic 3,498; paired 1,932	5,430
Owned BisQue volumes	Multichannel 49,788; single-channel views 114,688	164,476
HCP structural MRI	Paired 14,064; target-student views 14,064	28,128
LIVECell	Instance-annotated 460	460
Plant volumes	Arabidopsis instances 83; native unlabeled 71	154
All retained training contracts	35 categories; auxiliary pools overlap	319,612

Table 12: Post-selection training counts, not preselection catalog totals. "Atomic" denotes an input-view contract, not an independent biological unit; paired and target-student expansions can reuse its acquisition. The table includes unlabeled and partially annotated inputs without converting them into fully supervised examples.

without guaranteeing an upstream parameter gradient or calibrated probabilities. Focal, generalized Dice, deep-mask and semantic-boundary losses are not silently added.

Affinity-based segmentation has established precedent (Turaga et al., 2009); the following head uses its own observed-edge BCE, not the maximin objective. The instance head uses 64-channel mask features, a spatial convolution, 32-channel SiLU and five outputs: three oriented affinities, one boundary, and softplus density. Affinity endpoints require known, unambiguous ownership and at least one foreground endpoint. Unknown ownership is not a negative; foreground boundaries beside unknown ownership are excluded. Affinity/boundary BCE average the classes present. Each admitted centroid contributes a Gaussian of bandwidth two *voxels*, truncated at three bandwidths and renormalized on valid support. Under complete annotation,

$$L_{\text{density}} = \frac{\sum_{v \in V} [\rho_v - y_v \log \max(\rho_v, 10^{-8})]}{N + 1}, \qquad L_{\text{inst}} = L_{\text{aff}} + L_{\text{boundary}} + 0.1 L_{\text{density}}. \tag{E.5}$$

N counts admitted centroids inside the crop. The density integral and watershed component count are different outputs; count MAE is logged, not separately optimized. Incomplete annotations disable density/count supervision rather than provide zero targets.

The culture classifier concatenates valid-token means and $\sqrt{\text{variance} + 10^{-6}}$, then LayerNorm, a 256-unit layer, GELU and eight logits. Cross-entropy uses source-verified LIVECell cultures only; records lacking reference foreground receive no culture target. Other datasets are not negative examples of these classes.

Latent-target prediction follows an established learning paradigm (Assran et al., 2023); the target eligibility and physical correspondence below specify this implementation.

Auxiliary	Detached target, support and loss
Self prediction	Clean teacher tokens through a fixed width-256 projection; masked valid sites, or all valid sites if none are selected; mean cosine loss.
Context	Larger-view teacher features affine-resampled at student token centers; valid overlap, with no minimum-overlap cutoff on this path.
Paired latent	Paired-channel features at corresponding indices. Equal shape, matching affines and registered content are upstream requirements, not a registration operation performed by the loss.
Fluorescence	Measured normalized intensity, MSE plus 0.05 times a spacing-scaled difference penalty. Supplied-target views disable copying tasks.
Localization	KL between visual and annotation-defined neighborhoods on the fixed training candidate bank, using temperature 0.1. Entity identity remains private supervision, not visual input.

The BF+FL path checks affine equality; the BF-only paired-target path does not independently check it. Operator tests do not verify every registration. EMA momentum is 0.996 for student-trainable parameters and buffers are copied. Current consolidation does not supervise the inherited sequence head. Measured fluorescence, annotation neighborhoods, fixed-source features and moving-EMA features are distinct targets; none is an unobserved optical measurement.

E.5 Optimization and the two operational references

The first 128 branch updates freeze the encoder but train eligible mature and new heads. Peak rates are 5×10^{-6} for the foundation, 10^{-6} for routers and 2×10^{-5} for other heads; mature heads use $0.25 \times 5 \times 10^{-6}$ during warmup. Head ramp is 64 updates; backbone/router ramp is 256 after opening. The final 2,048 updates linearly reduce the multiplier toward 0.1. AdamW decay is 0.01; norm limits are five for the foundation and one for routers/heads. Routing bias updates every eight steps at rate 0.001, deadband 0.25 and cap 0.5. The step-256 continuation preserves optimizer, EMA, randomness and cursors but changes the stopping policy, not the strict thresholds.

Historical checks use cached earlier-passive predictions m_H; incoming checks use top-4 step 128 values m_I, not step 256. Allowed Dice drops are 0.03/0.005 and Brier increases 0.02/0.002 respectively. Both allow empty-target foreground-fraction increase 10^{-4}. Incoming culture may drop 0.05. Localization conditions are $b \leq 1.25 b_H + 0.05$ and $b \leq 1.10 b_I + 0.02$. Historical segmentation covers Allen BF/FL, organoid and QPI; incoming checks cover eligible fixed-panel metrics. Allen BF/FL Dice are mandatory.

The looser emergency rule uses the incoming panel: Dice drops of 0.03 for Allen BF and 0.05 otherwise, Brier/empty-FP increases of 0.03/0.005, culture drop 0.20, and localization increase $0.50|b_I| + 0.10$. Required missing/nonfinite metrics fail safety. A pass requires both strict references and safety; strict failures may permit continued training without becoming passes. Five of fourteen archived validations pass, including the terminal one. Step 20,480's plant empty-reference failure remains recorded. These are operational tolerances, not statistical equivalence margins.

E.6 Completed continuation and measured cost

All fourteen validations use the same ordered 1,481-observation panel, with metric-specific eligibility. The terminal summary below replaces repeated cumulative tables in this supplement; the accompanying earlier table sources retain the preceding checkpoints and adverse outcomes. Recorded updates 257–26,584 comprise 26,328 updates per rank with no missing or duplicated indices. Summed synchronized iteration timers give 1,230,484.0 device-seconds, or 341.8 device-hours across three GPUs. Timing excludes validation, checkpoint writing, startup, queueing and all ancestral training. The trainable count is 3,987,591,525 and peak recorded PyTorch allocation is 115.01 GiB per rank, not the configured memory cap or total device allocation. No lifetime-compute or FLOP estimate is inferred. The archived aggregate replay covers 11,536 domain/metric means and 378 exported rows; it does not re-execute the model or validate reference truth.

Measurement	Eligible n	512	10,240	22,528	24,576	26,584
Allen BF Dice	60	0.3284	0.3311	0.3303	0.3265	0.3293
Allen FL Dice	60	0.8153	0.8285	0.8338	0.8193	0.8299
LIVECell Dice	64	0.8869	0.9061	0.9069	0.9079	0.9102
Arabidopsis Dice	19	0.9612	0.9774	0.9832	0.9851	0.9803
Organoid Dice	51	0.9357	0.9356	0.9618	0.9605	0.9628
QPI Dice	31	0.8744	0.9011	0.8987	0.8966	0.9019
LIVECell count MAE	64	34.4456	15.0739	11.5279	13.4642	11.1626
Arabidopsis count MAE	20	5.2504	3.0553	1.8586	3.2577	2.7235
Culture balanced accuracy	64	0.7344	0.9375	0.9688	0.9688	0.9688
Fixed-source self loss	1481	0.3511	0.4878	0.5383	0.5439	0.5450
Fixed-source paired loss	395	0.4223	0.5289	0.5642	0.5644	0.5644
Fixed-source context loss	1046	0.3474	0.4341	0.4613	0.4676	0.4680
Measured fluorescence loss	255	0.1421	0.1196	0.1206	0.1165	0.1117
Verified-localization loss	124	1.2505	1.1792	1.0670	1.1026	1.0413
Current-EMA self loss	1481	0.0308	0.0326	0.0324	0.0316	0.0274
Current-EMA paired loss	395	0.2150	0.1787	0.1770	0.1698	0.1677
Current-EMA context loss	1046	0.0573	0.0534	0.0525	0.0518	0.0484
Operational gate	–	Fail	Pass	Pass	Pass	Pass

Table 13: Completed development trajectory. Five of fourteen archived checkpoints pass the operational criteria in Appendix E.5. Allen BF remains near 0.33 while culture and several fitted dense outputs improve; fixed-source feature losses increase. These are descriptive outcomes on one reused panel, not independent transfer tests. Terminal fresh-readout evaluations extend previously inspected panels post hoc; the Allen comparison's timing is specified separately in Appendix F.3.

F Culture and Brightfield Capabilities Along the SCOPE Lineage

F.1 Culture gains under fresh and unchanged heads

Completion at 26,584 fixes the terminal checkpoint by the schedule, not culture test accuracy. Its comparison with 256 and 10,240 extends previously inspected wells post hoc. The task already supplied encoder supervision. The experiment asks whether useful culture decoding improves under stated readout budgets, not whether an unseen task was learned from 51 labels.

Table 14: Fresh culture readouts at the completed production horizon. Balanced accuracy on the same 1,460 images in nine wells. Linear fitting is deterministic; MLP entries average three head seeds. Both 51/460-image fitting allowances use 128 validation labels. These counts describe readout fitting and validation, not the encoder's earlier culture supervision.

Readout	Fit images	Step 256	Step 10,240	Step 26,584
Linear	51	0.78526	0.97807	0.98325
Linear	460	0.96081	0.97878	0.98263
Original MLP	51	0.78863	0.97719	0.98368
Original MLP	460	0.95297	0.97785	0.98319

Fresh linear heads use training-only standardization and the original five-value C grid. MLPs use the original LayerNorm–256-unit–GELU–eight-logit architecture, 2,048 CPU FP32 AdamW updates, batch 32, rate 3×10^{-4}, decay 10^{-4} and clipping at five. Validation every 128 updates, including zero, retains the earliest best balanced accuracy. Initialization, sample order and fitting subset are paired across encoders within each seed. Both budgets also use 128 validation labels. All 24 learned readouts and both moment maps finish before new scoring on the 1,460 images in nine reused wells.

Terminal accuracy exceeds 10,240 under both readout families and in all three paired MLP seeds at both budgets. The full-budget MLP log loss instead changes from 0.08111 to 0.08203: the accuracy

gain is not universal across scores. DINOv2 CLS remains stronger at 460 fitting images (0.9927 versus terminal fresh-linear 0.9826), under different pretraining and processing. Seven culture classes have one test well; head seeds do not provide independent-well uncertainty.

Table 15: Actual inherited classifiers crossed with all three encoder checkpoints. Unlike a source-fitted fresh probe, head 256 is the classifier inherited from joint training. Test balanced accuracy; no fitting occurs in this cross.

Encoder	Head 256	Head 10,240	Head 26,584
256	0.62350	0.54351	0.54681
10,240	0.96298	0.97270	0.97469
26,584	0.97984	0.98192	0.98345

The actual unchanged inherited classifier improves from 0.62350 on incoming features to 0.96298 on adapted and 0.97984 on terminal features. A fresh source head is already much stronger than that inherited starting head. The cross therefore identifies useful compatibility changes, not a unique additive decomposition of classifier and encoder learning. Training-only moment alignment changes terminal/head-256 accuracy to 0.98019 while worsening log loss from 0.19539 to 0.45284; it is not an unqualified repair. Two training probes match live/cache inherited logits bitwise. Archived CPU checks replay 138 reductions and 35 readout executions, not every CUDA prediction; eleven FP32 sensitivity controls remain separately identified.

F.2 A SOURCE-FITTED LINEAR HEAD IS A DIFFERENT INTERFACE

The linear probe fitted on incoming frozen features drops from 0.9608 to 0.9475 on adapted features. A training-only coordinatewise map

$$T(z)_j = \mu_{0j} + \frac{\sigma_{0j}}{\sigma_{1j}}(z_j - \mu_{1j})$$

restores 0.9768; paired ridge alignment gives 0.9791. Scales below 10^{-8} are replaced by one. Both alignment methods use all 460 training images even for a head fitted on 51 labels. The weak-head ridge map lowers accuracy from 0.9448 to 0.8066, whereas moments give 0.9668. Global shuffles fail, but within-culture shuffles largely preserve the recovery. The latter use training labels and are diagnostic rather than label-free deployment. Thus acquisition-specific correspondence is not needed for this observed classification effect.

An elementary coordinate example explains the interpretive boundary. If $f_1 = Af_0 + b$ with fixed invertible A, then the affine readout $W_1 = W_0A^{-1}$, $c_1 = c_0 - W_0A^{-1}b$ exactly reproduces $W_0f_0 + c_0$. The converse gives the same population-risk infimum over all affine maps for a common loss. It does not preserve a fixed map, finite regularized fitting, or a nonlinear head's behavior. Nor does it establish that the learned features actually obey that model. Moment alignment has domain-adaptation precedents (Sun et al., 2015); representation-similarity statistics have their own transformation invariances (Kornblith et al., 2019a). Neither reference makes this fitted map an exact recovery of the learned features.

F.3 BRIGHTFIELD STRUCTURE PREDICTION ON HELD-OUT PLATES

The Allen imaging resource (Allen Institute for Cell Science, n.d.; Viana et al., 2023) and its fluorescence-segmentation toolkit (Chen et al., 2020) provide the resource and methodological context. The following split and sampled crops are specific to this study; a resource citation does not identify the exact mask-generator version used in each selected record. The test contains 256 acquisitions from 23 plates and 24 tagged structures; HIST1H2BJ and NUP153 share a plate. AAVS1 is excluded as in training. Selected training, validation and test plates/FOVs are disjoint. The inspected lineage manifests contain no matching test keys or plate IDs, but do not close opaque ancestral aliases or renamed pixel duplicates. A label-independent sampler chooses one $32 \times 160 \times 160$ crop per acquisition. Its validation inputs and records reproduce the earlier pipeline bitwise. Of the test crops, 234 have reference foreground and 22 enter a separate empty-reference foreground-fraction endpoint.

The passive wrapper receives its structure query at 256, 10,240 and 26,584. BF sees neither fluorescence nor references; FL sees the tagged channel and coarse-family channel identity. Allen reference masks derive from the fluorescence pipeline, so FL agreement is a privileged-input comparison, not an independent biological reference or a formal BF upper bound. The checkpoint set preceded prediction; the terminal endpoint was designated primary during inference, before reference scoring, not before any prediction existed. Threshold is 0.5. Nonempty crop Dice is averaged within structure and then over the 24 structures.

Corrected plate-composition resampling. For structure means d_s and sampled multiplicities $m_{p(s)}$ of their plates, each of 10,000 bootstrap draws uses

$$\widehat{d}^{*} = \frac{\sum_s m_{p(s)} d_s}{\sum_s m_{p(s)}}.$$

Each draw samples 23 plates with replacement and moves shared-plate structures together. Paired systems use identical draws. The original scorer discarded repeated-plate multiplicity; this correction preserves all point estimates and saved predictions but changes intervals. The result varies plate/structure composition conditional on fitted systems, not new plates within each fixed structure or scratch training seeds. These pointwise intervals are not simultaneous per-structure claims. Terminal BF macro Dice is $0.3124\,[0.2244, 0.4072]$. Against the 37.2M scratch control the difference is $+0.0559\,[0.0373, 0.0791]$, compared with $+0.070\,[0.049, 0.095]$ against the 16.5M control. Terminal-minus-incoming is $+0.0076\,[0.0040, 0.0115]$; adapted-minus-incoming is $+0.0008\,[-0.0027, 0.0037]$. Incoming already leads the larger scratch by $+0.0483\,[0.0311, 0.0692]$. Most of this BF capability predates the reported continuation.

System	Parameters	Updates/seed	Macro Dice [95% interval]
Original scratch	16.537M	30,000	$0.2423\,[0.1629, 0.3294]$
Larger scratch	37.202M	60,000	$0.2565\,[0.1740, 0.3472]$
SCOPE, step 26,584	$\sim$4B	Full lineage	$0.3124\,[0.2244, 0.4072]$

Table 16: The historical and enlarged controls on the same 234 nonempty test crops. Scratch entries average two fitted seeds' Dice scores before equal-structure aggregation; SCOPE is one trained lineage. Intervals use the corrected paired plate-cluster procedure. SCOPE's inherited training is not represented by its continuation step number, and total compute is unmatched.

The 16,536,664-parameter U-Net has one output per structure and is trained from scratch for two seeds, 30,000 updates each, batch eight. The same Allen records, data compiler, crop loader and segmentation loss supervise the tagged output. Validation every 2,500 steps selects 30,000 and 27,500. The 37,202,040-parameter extension uses widths 48/96/192/384/480 and two new seeds, each with 60,000 updates at batch eight. Data files, AdamW recipe, loss and validation population match the first control. Both larger seeds select 60,000, with validation macro Dice 0.2680 and 0.2647. Full completion at the selected endpoint is not a convergence demonstration. The scratch studies were specified after SCOPE outcomes but before their own fits; the extension is post-hoc. The export helper computes unused reference Dice, not used for input or checkpoint selection.

Scores average scratch seeds' Dice values, not their probabilities. The larger control closes 20.2% of the original point-estimate gap, while terminal SCOPE leads it on all 24 structure means. These are not 24 resolved independent contrasts. RAB5A remains difficult and favors the old scratch slightly: 0.01483 versus terminal 0.01363 and larger 0.01275. NPM1, FBL and LMNB1 achieve BF Dice 0.839, 0.764 and 0.735; six structures remain below 0.10. One plate per structure cannot separate plate and structure difficulty. Parameter count, convergence, ancestral exposure and lifetime compute are not matched; the positive finding concerns these bounded systems.

Cell- and nucleus-mask controls score 0.070 and 0.055 macro Dice. Selecting the better mask for each structure using test references gives 0.096: an oracle over two fixed masks, not a learnable-system bound. FL scores 0.819; terminal-minus-incoming FL is $+0.0191\,[0.0055, 0.0332]$. Terminal BF empty-reference foreground fraction is 1.2×10^{-4}, versus 0.9–1.6×10^{-3} for the original scratch seeds and $7.10 \times 10^{-4}/6.93 \times 10^{-4}$ for the larger seeds. These fractions concern empty algorithmic references, not a per-image false-positive probability or verified absence of biology. The

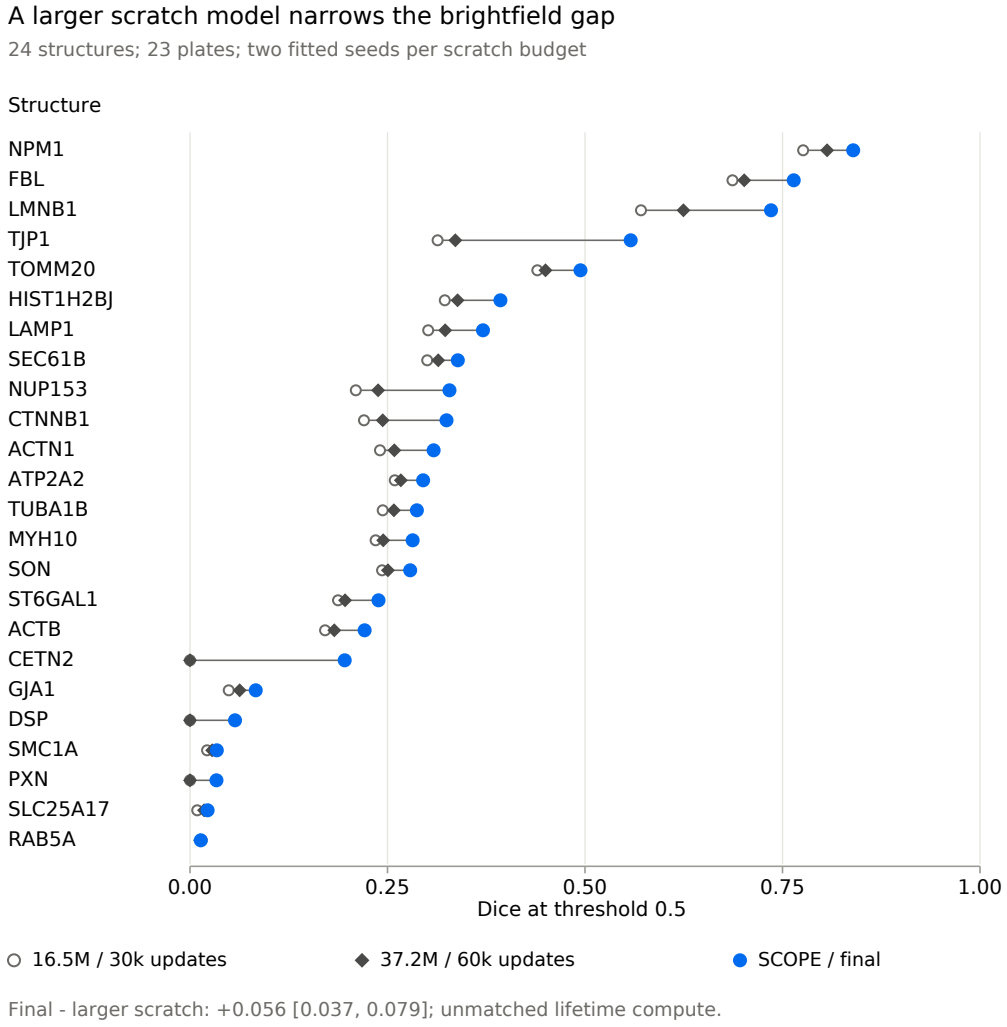

Figure 4: **A stronger scratch comparison narrows the BF gap while preserving structure heterogeneity.** Every structure and both scratch sizes are retained on the same test crops. Macro intervals use corrected plate-multiplicity resampling; per-structure values are point estimates. Larger capacity, doubled updates and different scratch seeds change together. References are algorithmic, with no within-structure plate replication.

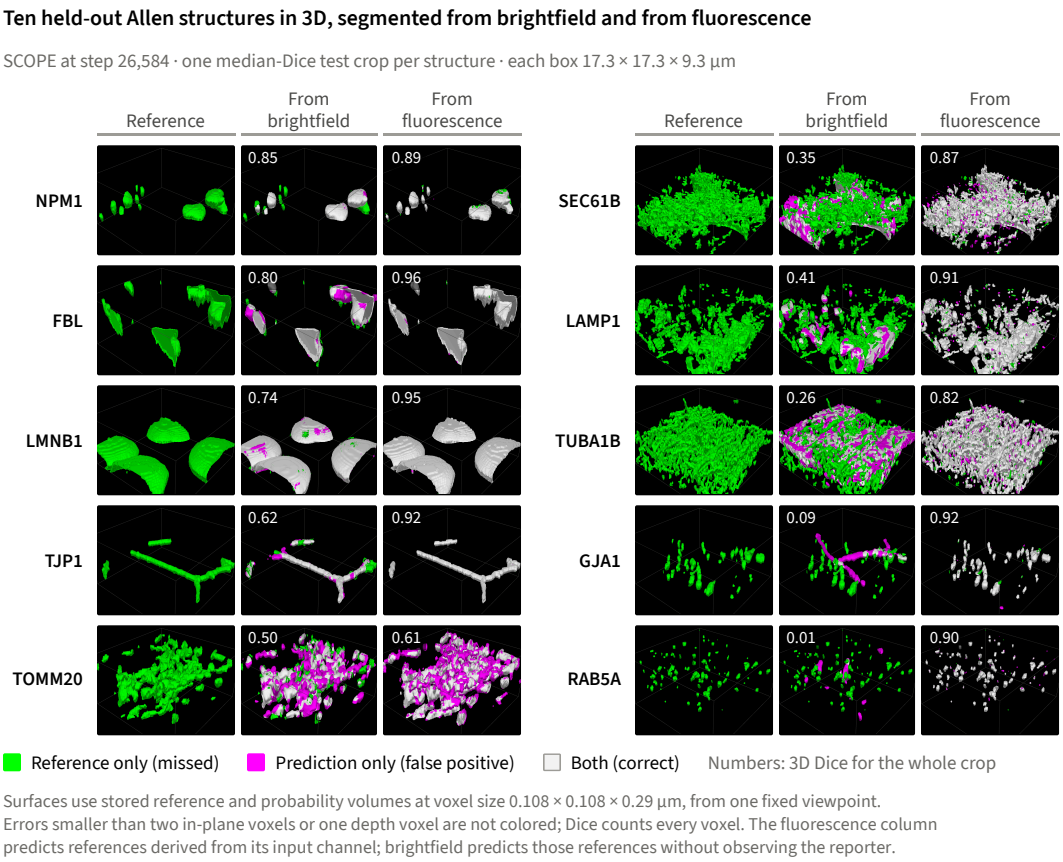

Figure 5: **The same held-out structures under BF and FL input.** Ten outcome-ranked median-BF examples at terminal step 26,584, rendered from stored volumes at voxel size $0.108 \times 0.108 \times 0.29\,\mu m$. Reference-only support is green, prediction-only magenta and shared white (Wong, 2011). Display color suppresses differences smaller than two in-plane or one depth voxel; numerical 3D Dice uses every voxel. FL predicts references derived from that same reporter, unlike BF. The display illustrates the scored outcomes, not an optical identifiability limit or independently adjudicated biological accuracy.

earlier 298-crop development panel includes 42 training FOVs from eight plates. Restricting it to 256 validation crops gives BF macro Dice 0.315 at both incoming/adapted checkpoints, versus test 0.305/0.306. That unpaired comparison does not demonstrate equivalence.

F.4 Qualitative support and example selection

The selected ten structures use the median-BF-Dice crop per structure; the slice or volumetric display is chosen using references. Selection was fixed before viewing images but is outcome-ranked, not representative random sampling. The source examples retain both stronger broad structures and weak punctate/network predictions. Model inference never receives references.

G Spatial Readouts: Capacity, Feature Access and Practical Baselines

The purpose of the stronger readouts is to test whether an apparent encoder gain persists when the source receives a more adequate way to express its features. All comparisons below keep their stated checkpoint pairs and reuse their original test groups; fresh head seeds do not create new acquisition populations.

G.1 BBBC039: OBJECT-AWARE DECODING AND TERMINAL SCOPE

BBBC039 (Broad Institute Imaging Platform, 2018; Caicedo et al., 2019) supplies the 200 manually annotated nuclear fields, which preserve touching-instance IDs and the full 520×696 native canvas. The primary split is 122/42/36 fields from 12/4/4 disjoint plates; the official 100/50/50 split shares plates and is a separate exploratory comparison. Input normalization and padding are target-independent. One observed plane is valid; depth padding is not acquired anatomy. Pixel pitch is unverified, so geometry is nominal. Searches over 238 current/historical-candidate manifests find no declared identity matches but do not close transformed aliases from BBBC022/BBBC038 ancestry.

Incoming/adapted fusion readouts use final tokens plus the half-resolution skip; skip-only and a raw-image U-Net are controls. The 79,265-parameter fusion and 118,625-parameter scratch models use three seeds, 900 updates, batch two, BCE plus half-weight foreground Dice, and validation every 150 updates at threshold 0.5. Samples, fitting initializations and order are paired across frozen checkpoints; no mask-guided crops or augmentation are used. Lattice-aware interpolation preserves the padding offset.

Object heads add normalized instance-distance and inner-boundary outputs. Loss weights are 0.5 foreground-distance MSE, 0.05 background-distance MSE, and 0.25 boundary BCE with positive weight four. Each selected head receives three geometric watershed settings; object heads additionally receive three learned distance/boundary settings, selected on validation. Both are reported. For the adapted object head, learned rather than geometric decoding changes object F1 from 0.8502 to 0.9476 with unchanged raw Dice; the paired seed/plate gain is $+0.0983\,[0.0886, 0.1113]$. Scratch also benefits. This tests decoding of trained outputs, not an encoder update.

Table 17: Terminal nuclear transfer with the strong object readout and validation-selected learned decoder. Equal-seed, equal-image means on 36 fields from four test plates. Dice uses the raw foreground field; F1 uses decoded instances. The paired plate-weighted intervals in the text have a different weighting and need not equal differences of these image-weighted means.

Encoder	Raw Dice	Object F1@0.5	Object F1@0.75	Count MAE
256	0.97214	0.94724	0.91565	7.89815
10,240	0.97223	0.94765	0.91772	7.83333
26,584	0.97229	0.94665	0.91513	7.87037
Scratch	0.97303	0.94306	0.90902	7.74074

The terminal extension adds 12 fits to the earlier 27 evaluations. Its plate-weighted F1 changes are $-0.00122\,[-0.00445, 0.00113]$ versus 10,240 and $-0.00110\,[-0.00306, 0.00086]$ versus source. Count and strict-IoU differences remain unresolved. Table means weight images; paired intervals weight plates and need not equal subtraction of table entries. The low-label semantic fit uses 13 images and still 42 validation masks. An initial terminal object-13 mismatch was retained as exploratory; three additional semantic-13 fits restore the matched comparison. No equivalence or new-population claim follows from the unresolved results.

G.2 BBBC050: OBSERVATION AND FITTING PROTOCOL

BBBC050v2 (Funahashi Laboratory & Broad Institute Imaging Platform, 2020; Tokuoka et al., 2020) supplies the 165 manually referenced volumes from 15 mouse embryos. Whole-embryo splitting uses 99/22/44 volumes from 9/2/4 fit/validation/test embryos; test microscopes and histone-fusion reporters differ from training. ZYX pitches are $(1.75, 0.8, 0.8)$ µm for training and $(2.0, 0.8, 0.8)$ for test. Generic fluorescence typing does not alias the reporter to a DNA dye. Targets and embryo/time identifiers are excluded from input. Counts use unique positive annotation IDs. Instance foreground is used consistently because separate semantic masks disagree on 30 training volumes. A 443-manifest audit finds no declared overlap, without resolving opaque ancestry. Router state is unchanged during frozen extraction.

The original restricted heads read final tokens and one half-resolution skip, predicting foreground, normalized per-instance physical distance and inner boundary. Three seeds use 1,800 AdamW up-

Features	Decoder	Raw Dice	$F1_{.5}$	$F1_{.75}$	Count MAE
Incoming: restricted	learned	0.7101	0.7182	0.0177	1.3712
Incoming: restricted	geometric	0.7101	0.6898	0.0149	1.8712
Later: restricted	learned	0.7605	0.7941	0.1363	1.3788
Later: restricted	geometric	0.7605	0.7590	0.1238	1.7273
Incoming: full pyramid	learned	0.7942	0.8381	0.2870	1.5455
Incoming: full pyramid	geometric	0.7942	0.8282	0.2815	1.8182
Later: full pyramid	learned	0.7873	0.8352	0.2621	1.7727
Later: full pyramid	geometric	0.7873	0.8133	0.2493	1.8258
Later: prefix only	learned	0.7943	0.8424	0.3186	1.6212
Later: prefix only	geometric	0.7943	0.8221	0.3142	1.7424
Scratch U-Net	learned	0.8221	0.8705	0.5218	1.3258
Scratch U-Net	geometric	0.8221	0.8502	0.5107	1.3788

Table 18: Full-pyramid controls and immutable earlier endpoints. Both selected decoders are retained. Raw Dice is shared across the two decoder rows.

dates at 3×10^{-4}, no augmentation, and validation Dice at 0.5 for checkpoint selection. Each receives learned prominence $0.1/0.2/0.3$ and geometric prominence $1/2/3$ µm grids with minimum volume 64 µm^3. Both decoding families are reported. The classical watershed's separate 27-setting search uses validation object F1, not the earlier count criterion.

Adding full-, quarter- and eighth-resolution skips creates a 341,635-parameter pyramid head; prefix-only removes final-token influence with 292,387 active parameters. Original half skips match bitwise on all 165 volumes per checkpoint, and 792 saved test masks replay. The richer incoming readout raises raw Dice from 0.7101 to 0.7942; adapted rises from 0.7605 to 0.7873. The resulting adapted-minus-incoming contrasts are $-0.0069\,[-0.0205, 0.0046]$ Dice and $-0.0029\,[-0.0248, 0.0222]$ F1. Both are unresolved, while scratch remains stronger. Additional feature access, adapters and capacity change together. For Δ_h = adapted-minus-incoming Dice under head h, the paired pyramid-minus-restricted interaction is -0.0573. Enumerating $3^3 4^4 = 6,912$ crossed seed/embryo bootstrap draws gives $[-0.0877, -0.0291]$, but a three-seed Student-t sensitivity interval conditional on these embryos is $[-0.1338, 0.0193]$. The sign resolution depends on the uncertainty procedure. Enumeration removes simulation error, not small-sample population uncertainty. This bundled comparison motivates, rather than replaces, the controlled capacity/access follow-up.

G.3 A CONTROLLED CAPACITY AND FEATURE-ACCESS COMPARISON

Readout A is the original restricted token/half-skip model (134,627 active parameters). B enlarges its capacity but obtains every additional scale by coordinate-aligned resampling of the same half skip (333,955 parameters). C uses the full encoder pyramid (341,635). B/C share the fusion trunk and its initial common tensors; their adapters differ. Thus capacity is closely matched, not identical function classes or inference cost.

All 18 fits retain the same two checkpoints, splits, three seeds and original foreground/distance/boundary losses. AdamW uses 3×10^{-4}, decay 10^{-4}, clipping at five and batch one. Validation every 300 updates chooses the earliest maximum Dice available within the prespecified 1,800- and 3,600-update allowances. Predictions precede scoring; no test outcome selects a budget or head. Define $\Gamma_{\text{cap}} = \Delta_B - \Delta_A$ and $\Gamma_{\text{access}} = \Delta_C - \Delta_B$. The prespecified primary is Γ_{access} at 3,600: -0.0350, crossed interval $[-0.0460, -0.0224]$ and seed-t sensitivity $[-0.0646, -0.0055]$. All twelve seed-by-embryo contrasts are negative. Capacity also reduces the contrast, $-0.0406\,[-0.0581, -0.0228]$, but its seed-$t$ interval $[-0.0838, 0.0027]$ spans zero. At 1,800, access is unresolved, $-0.0222\,[-0.0475, 0.0042]$.

This separates two interventions more closely than the original bundled comparison: larger restricted capacity reduces the contrast, and access to extra encoder scales changes it further under the longer allowance. Adapter and optimization differences remain part of that effect. Expanding validation selection to 3,600 improves none of the arms' mean test Dice; a larger candidate set does not certify convergence or deployment gain. These are raw-Dice outcomes. Object/count/volume results for the

Fitting allowance	Readout	Incoming Dice	Adapted Dice	Difference
1,800	A: restricted	0.7099	0.7603	$+0.0504$
	B: larger, restricted	0.7541	0.7675	$+0.0134$
	C: full pyramid	0.7943	0.7855	-0.0088
3,600	A: restricted	0.6880	0.7476	$+0.0596$
	B: larger, restricted	0.7495	0.7685	$+0.0190$
	C: full pyramid	0.7808	0.7648	-0.0160

Table 19: Fixed-checkpoint readout controls on the same four BBBC050 test embryos. Means weight embryos and fitting seeds equally. Each allowance uses its own validation-selected prefix of the same training trajectory. The 1,800-update rows are new paired fits, not replacements for the earlier historical experiment. Raw Dice is measured before instance conversion.

new A/B/C fits were not completed in the supplied artifact and are not inferred from the historical decoder study.

G.4 Label budget, output lattice and calibrated support

The restricted-head label-budget comparison uses one/three/nine fitting embryos (11/33/99 volumes) plus two validation embryos (22 volumes) at every budget. Nested subsets are paired within three subset/seed replicas, with 1,800 updates and the original losses and decoder grids. Adapted Dice is 0.7875/0.7735/0.7605 versus scratch 0.8276/0.8273/0.8221. F1 at 0.5 is 0.8102/0.8212/0.7941 versus 0.8776/0.8666/0.8705. Scratch also leads strict overlap at all budgets, while at three fitting embryos the adapted count error is lower (1.159 versus 1.674). Fitting-only labels exclude validation and prior encoder supervision; these results do not establish end-to-end label efficiency.

A target-only free-field fit on the head's exact output lattice achieves mean Dice 0.99737 and minimum 0.96378 over 44 volumes. This is a representability witness, not a learned prediction or optimal-Dice certificate. Validation-selected thresholds $0.1, \ldots, 0.9$ raise adapted Dice to 0.7705 versus scratch 0.8206; the approximately 34.5% foreground-volume deficit remains. Original FP16 probability rounding cannot explain it: threshold ambiguity bounds the uncalibrated volume error within $[-0.361130, -0.360887]$. These targeted controls limit resolution and threshold explanations, without diagnosing a unique source of the learned deficit.

G.5 Strong specialists and pooled-feature controls

A three-seed, 100-epoch fine-tuned Cellpose control matches adapted BBBC039 foreground Dice (0.97232 versus 0.97243) but has higher object F1 (0.96993 versus 0.94765) and lower count MAE (2.602 versus 7.833). Its optimization and pretraining differ; it is a practical comparator, not an architecture-only ablation. On BBBC050, expanding the frozen Cellpose 3D decoder search after inspecting the initial comparison raises F1 from 0.5243 to 0.8186 versus restricted adapted 0.7941; the paired interval spans zero. The earlier favorable SCOPE ranking is not robust to that decoder allowance.

Pooled-feature counts supply a separate negative control: BBBC050 adapted MAE improves over incoming (14.422 versus 20.479) but trails image moments (5.547) and validation-selected physical-distance maxima (0.841). The latter estimates counts, not accurate nuclear shapes. Altering the declared Z pitch with fixed pixels changes count predictions, a metadata sensitivity diagnostic rather than a legitimate correction. Conditional object-membership probes with exactly displacement-matched foreground endpoints yield adapted test AUROC 0.5181 versus raw-line 0.9414. This restricted PCA/readout failure does not establish information absence. These controls constrain a general morphological-transfer claim; their historical full displays remain in the source archive, not in a repeated sequence of figures here.

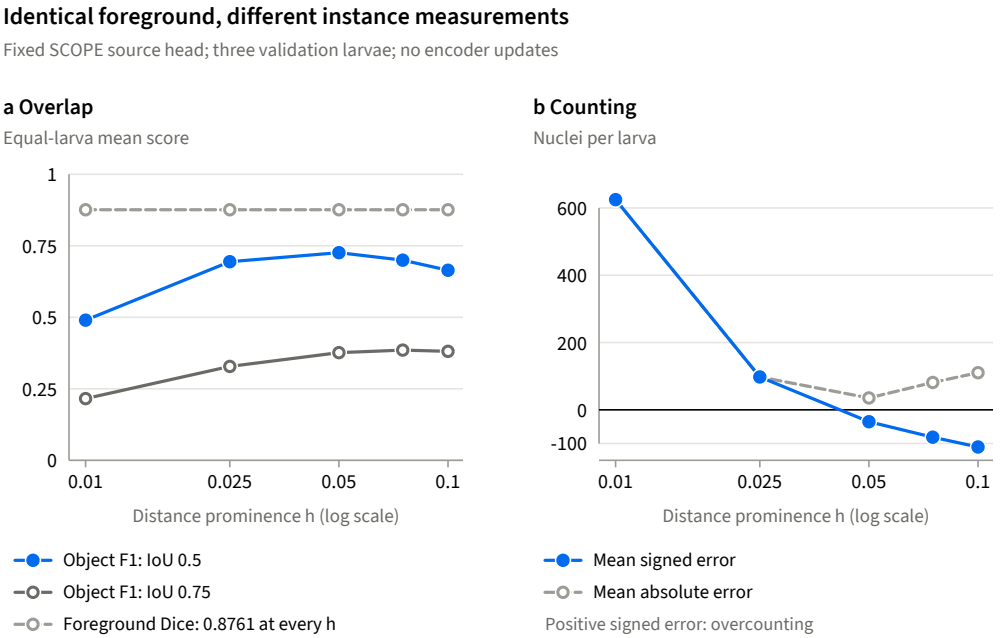

Figure 6: **The same foreground supports different object measurements.** One unchanged source head on three reused validation larvae, with only instance-separation prominence varied. Curves connect tested settings, not training time. The diagnostic is post-hoc; neither test animals nor encoder updates participate.

H Native Object Measurements and Counting

Segmentation metrics capture different output properties (Taha & Hanbury, 2015); nucleus-level error analysis distinguishes object errors from pixel agreement (Caicedo et al., 2019). A segmentation readout supplies both a foreground field and a partition of that field into objects. The following comparisons hold these components apart, then test the resulting pipelines on held-out animals. The primary adaptation contrast, stricter-overlap outcome, count controls and terminal extension remain separate: a gain in one does not stand in for another.

H.1 Fixed foreground, different instance partitions

On three development *C. elegans* larvae, one source full-pyramid head selected at update 1,600 produces fixed probability and normalized-distance fields (Long et al., 2009; Hirsch & Kainmueller, 2020). Only the watershed distance prominence changes, over $h \in \{0.01, 0.025, 0.05, 0.075, 0.10\}$; hole filling and the 128-voxel foreground-component cutoff remain fixed. Every decoded partition P_h has the same binary union $\{v : P_h(v) > 0\}$. Foreground Dice and topology of that union therefore remain unchanged, while object identities and counts can change. This is not a claim about the topology of individual reference objects or a new encoder result.

From $h = 0.10$ to 0.05, mean F1 at IoU 0.5 rises from 0.6646 to 0.7260 and count MAE falls from 110.3 to 35.3 nuclei per larva; foreground Dice remains 0.8761. Strict-IoU F1 changes from 0.3811 to 0.3764. At $h = 0.01$, mean count bias instead reaches $+624.7$. The original decoder optimum was at its search boundary. The appropriate follow-up is an equal expanded allowance for all models, not promotion of the best diagnostic setting.

H.2 EQUAL DEVELOPMENT ALLOWANCE AND A LOCKED ANIMAL TEST

Five conditions use source/adapted full-pyramid features, their respective convolution-only controls, or a scratch 3D U-Net. Each has three paired head seeds and 4,000 supervised updates on eighteen training larvae. Mean foreground Dice on three validation larvae selects the neural checkpoint; mean object F1 at IoU 0.5 then selects instance conversion. The original twelve-setting grid has three learned and three geometric prominences at two component-size cutoffs. The post-hoc expansion adds four lower learned prominences at both cutoffs to every unchanged head. All fifteen expanded optima use an interior learned prominence, 0.05 or 0.075. This finite procedure does not search all readouts or select the neural checkpoint by object F1.

The complete development grid matters: the first seed's original-grid F1 adaptation gain of 0.0713 becomes -0.0017 after averaging all three seeds. Expansion gives source/adapted validation F1 of 0.7598/0.7629, with only one seed favoring adaptation. These are selected validation outcomes, not test performance or evidence of equivalence.

Before test scoring, all fifteen heads and their original, expanded and geometric decoder choices are fixed. All 105 complete native prediction fields on seven test larvae are committed before references are opened; no test animal selects a threshold, checkpoint, seed or exclusion. Original reference IDs, including disconnected pieces, remain intact. The primary contrast is adapted-minus-source full-pyramid F1 at IoU 0.5 under the expanded learned decoder. The development amendment is not relabeled as preregistered.

This primary difference is $+0.00211$ with pointwise animal-bootstrap interval $[-0.00150, 0.00604]$; one of three head seeds favors adaptation. Count MAE falls from 17.57 to 8.62 nuclei per animal, a change of -8.95 $[-12.81, -4.57]$, with the same direction in all seed means. F1 at IoU 0.75 falls from 0.4123 to 0.3947, a change of -0.01756 $[-0.02674, -0.00690]$, negative on six of seven seed-averaged animals. Raw Dice changes by $+0.00019$ $[-0.00070, 0.00097]$. Figure 2 presents the main endpoint comparison.

Under the original decoder grid, source/adapted count MAE is 35.81/38.05; the difference $+2.24$ has interval $[-0.62, 5.38]$. Strict F1 still declines by 0.01501 $[0.00427, 0.02494]$. The expanded pipeline's count benefit is a held-out result, but not a decoder-independent representation advantage. The pretrained source exceeds scratch by 0.01906 F1 at IoU 0.5 $[0.00757, 0.03246]$ and 0.02023 at IoU 0.75 $[0.01277, 0.02834]$. This positive source benefit is distinct from the adaptation contrast. Adapted full-minus-prefix primary F1 is unresolved; strict F1 is lower by 0.01284 $[0.00687, 0.01803]$, and count MAE differs by -0.05 $[-2.00, 1.81]$. The prefix shares ancestral training and has fewer connected head parameters, so it is not an independently pretrained alternative.

Intervals resample the seven paired animals after averaging the three fixed heads. They condition on those heads and are pointwise, not simultaneous or future-seed guarantees. Acquisition-date hints overlap development; this is animal-held-out evaluation, not established batch- or new-laboratory independence. Declared-identity exposure checks leave opaque ancestry unresolved.

H.3 THE TERMINAL EXTENSION AND AN IMAGE-FREE COUNT PRIOR

The terminal extension adds six heads, three seeds for full and prefix features, under the same fitting and decoder procedures. Its 42 new native prediction fields precede scoring, but the seven animals have already been inspected: it is a post-hoc extension, not another locked population. Table 20 retains all three checkpoints, prefix and scratch controls, and the separately added constant counter.

Terminal-minus-adapted primary F1 is -0.00256 $[-0.00638, 0.00114]$; the three fitting-seed differences are $-0.01170, +0.00422, -0.00018$. Strict F1 improves by 0.00581 $[0.00138, 0.01152]$ over adapted features, while remaining 0.01176 below source. Count MAE is 12.33 versus 8.62. Terminal exceeds scratch on primary F1 by 0.01861 $[0.00717, 0.03135]$, but its full-minus-prefix contrast $+0.00153$ $[-0.00372, 0.00699]$ is unresolved. Under the original grid, its adapted-checkpoint primary contrast is likewise unresolved ($+0.00018$). These comparisons locate a useful pretrained pipeline without assigning a uniform gain to continued training.

The image-free control is the midpoint median of eighteen training reference counts: 544 nuclei. Introduced after the original locked outcomes, it gives 8.43 test MAE versus adapted full fusion's

Table 20: Completed native-3D comparison with the expanded validation-selected decoder. Equal-animal, equal-head means on seven reused animals. Full and prefix-only heads retain the same fitting procedure; the image-free control predicts only a count and has no object localization.

Encoder	Raw Dice	F1@0.5	F1@0.75	Count MAE
Full source	0.8743	0.8075	0.4123	17.57
Full intermediate	0.8744	0.8096	0.3947	8.62
Full terminal	0.8740	0.8071	0.4005	12.33
Prefix source	0.8753	0.8062	0.4099	9.71
Prefix intermediate	0.8750	0.8074	0.4075	8.67
Prefix terminal	0.8746	0.8055	0.4081	10.57
Scratch	0.8725	0.7885	0.3920	21.90
Train-only count constant	–	–	–	8.43

8.62, a paired difference of $+0.19$ $[-6.05, 6.81]$. It has no object locations and no defined object F1. This is a training-fitted count prior, not a test-fitted predictor or evidence that counts are generally recoverable without images.

For one-to-one matching, $P = \mathrm{TP} + \mathrm{FP}$ and $R = \mathrm{TP} + \mathrm{FN}$ imply

$$|P - R| = |\mathrm{FP} - \mathrm{FN}|, \qquad \mathrm{FP} + \mathrm{FN} = |P - R| + 2\min(\mathrm{FP}, \mathrm{FN}). \tag{H.1}$$

Unmatched objects can cancel in the count. At IoU 0.75, source/adapted mean unmatched objects increase from 631.38 to 648.24 ($+16.86$ $[4.57, 26.95]$) while count error falls. At IoU 0.5 the difference is -2.67 $[-6.71, 1.05]$. The identity is verified in the source analysis for all 315 pipeline–animal–decoder outcomes at both thresholds; it does not identify which objects physically merged, split or moved.

H.4 Reference-object fates clarify the spatial changes

The post-hoc fate display uses seed 187, source 256 and adapted 10,240, on all seven test animals. Each reference nucleus receives the first applicable category: strict match (IoU at least 0.75); otherwise a match at IoU at least 0.5 under the fixed cardinality-first matcher; merged when one prediction covers at least half this reference and at least half another; split when at least two predictions each place at least half their own volume inside this reference; missed when covered by less than 0.2; otherwise poorly localized. The precedence makes the categories exclusive but not a unique causal account of errors.

Mean merge-classified references fall from 75 to 38 in every animal, while split-classified references rise from 11 to 17 in every animal. Strict matches fall from 212 to 194 in six animals; loose matches rise from 200 to 241 in every animal. Fewer merge-classified references accompany the better counts alongside those other spatial changes. The one-head, outcome-based display does not estimate general merge/split rates or explain the ensemble count contrast causally.

H.5 Counting when counts vary

On the reused BBBC039 panel, 36 test fields from four plates have 53–202 reference nuclei, mean 132.92 and coefficient of variation 0.288. The training-median constant of 119.5 gives MAE 31.89, or 24.0% of the mean count. Each of 26 primary image-based counters does better on every plate: the MAE improvement ranges from 8.52 nuclei for image moments to 29.29 for fine-tuned Cellpose. The constant's competitiveness on larvae is thus population-specific. The comparison includes 13 primary instance pipelines, three ridge counters and ten aggregated dense conditions; twelve further instance pipelines are sensitivity controls. Head seeds are averaged within field, and SCOPE remains fixed at 256/10,240.

The localization distinction survives this greater count variation (Table 21). Classical watershed improves count MAE relative to the adapted SCOPE semantic readout but reduces strict F1 and leaves 18.83 more unmatched objects per field. Both directions hold on all four plates. The direction was known before this reanalysis; the grouped uncertainty strengthens the descriptive comparison, not a new causal claim. The within-SCOPE object-head count and F1 changes remain unresolved.

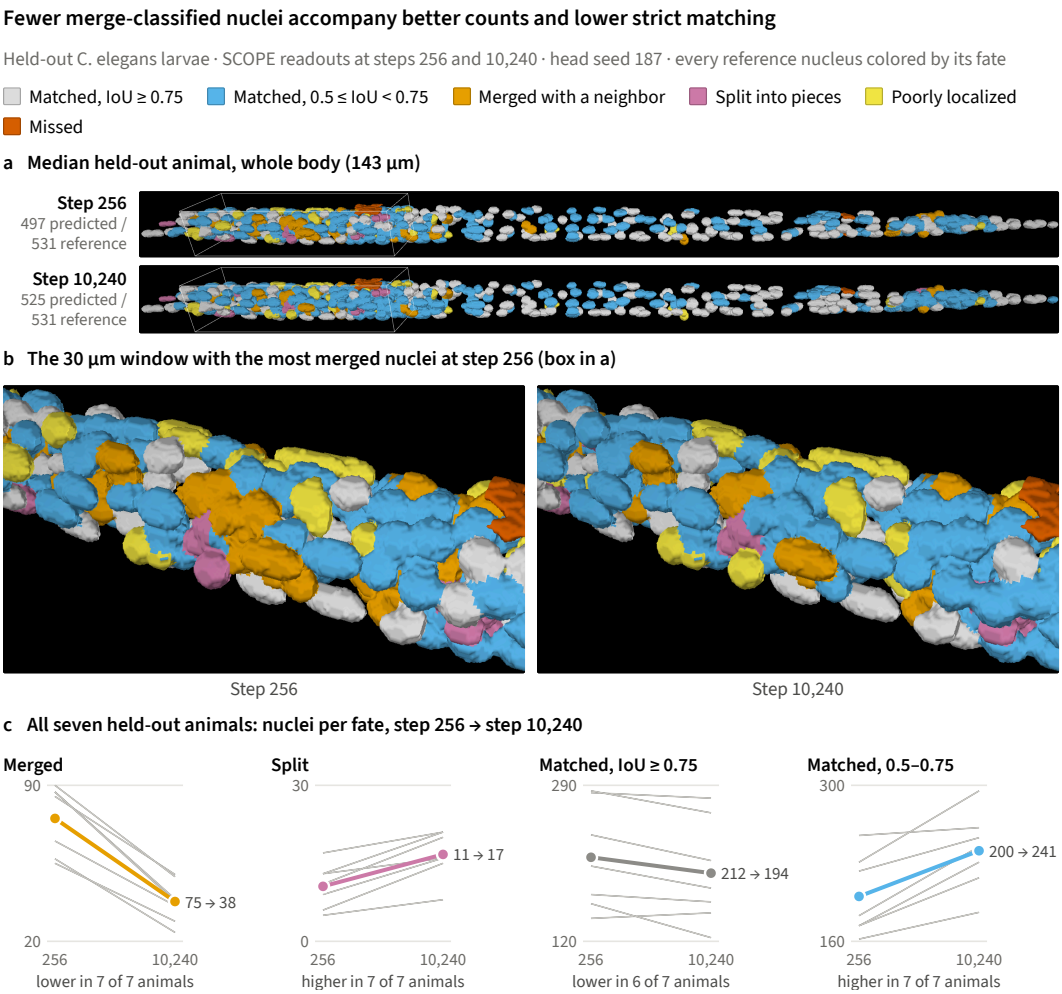

Figure 7: **Reference-object categories distinguish accompanying spatial changes.** Source and adapted seed-187 predictions use the original reference IDs. The displayed larva is selected by the stated outcome rank, and its 30-micrometer window is chosen for the source's merged category; it is not an outcome-independent typical case. The all-seven-animal panel supplies the population context. Category precedence is defined in the text; display choices do not change whole-volume scoring.

Contrast	Δ count MAE	Δ F1 at IoU 0.75
Classical minus SCOPE semantic, 10,240	-1.89 $[-3.09, -1.30]$	-0.0690 $[-0.0840, -0.0605]$
Learned minus geometric decoder, SCOPE object, 10,240	-1.25 $[-2.40, +0.22]$	$+0.1716$ $[+0.1552, +0.1881]$
SCOPE object, 10,240 minus 256	-0.06 $[-0.23, +0.03]$	$+0.0021$ $[-0.0013, +0.0054]$
Dense continuation, 4,096 minus 512	$+0.015$ $[-0.217, +0.206]$	$+0.0042$ $[+0.0019, +0.0074]$

Table 21: Focused contrasts on the same 36 fields. Lower count MAE and higher F1 are better. Brackets are descriptive 95% plate-cluster bootstrap intervals, conditional on the fitted pipelines; only four plates are available. The first row compares different pipelines, not encoder updates. The dense row averages forty related forks from ten seeds before resampling plates. MAE is nuclei per field; F1 differences are on the 0–1 scale.

Among the thirteen primary pipelines, Kendall τ_b between negative count MAE and F1 is 0.00 $[-0.08, 0.46]$ at IoU 0.5 and -0.10 $[-0.10, 0.36]$ at IoU 0.75. These intervals do not establish zero association; the concrete pipeline contrast, not a correlation null, supports measuring localization alongside counts.

The focused bootstrap resamples four plates 10,000 times and retains each plate's multiplicity m_g and number of fields n_g:

$$\widehat{\Delta}^* = \frac{\sum_g m_g \sum_{i=1}^{n_g} d_{gi}}{\sum_g m_g n_g}.$$

This field-weighted estimand differs from earlier equal-plate nuclear contrasts. Four plates yield only 35 distinct bootstrap multisets; intervals can under-cover, and four-of-four signs alone give two-sided $p = 0.125$ under sign exchangeability. The 78-pair exploratory scan is unadjusted. Anonymous plate sums in the release support focused replay, not recovery of private object masks or new biological replication.

I Observation Layout and Acquisition Context

A retained single plane can keep its pixels and world positions while padding changes convolutional phase, boundary context and tensor extent. The following controls vary the input layout, not the observed image or a single isolated architectural component. They use already inspected populations and retain both favorable and unresolved outcomes.

I.1 Phenotype prediction with compound and week exclusion

The image collection (Broad Institute Imaging Platform, n.d.; Caie et al., 2010) and published profile comparison (Ljosa et al., 2013) precede this study. Compound-excluded retrieval follows that profiling tradition; the fitting, aggregation and week-exclusion rules below define our comparison. CellProfiler 4 (Stirling et al., 2021) credits the software family, not the version that generated the historical profiles. The BBBC021 study contains 2,528 fields in 632 wells, including 302 treated wells, 38 compounds and twelve mechanisms. NSC excludes the query compound from fitting and the gallery; NSCB also excludes its acquisition week. Query-plate DMSO controls calibrate features without becoming a thirteenth class. Full-population scores retain unsupported targets; the supported NSCB population has matches for 274 wells, 33 compounds and ten mechanisms. Penalties are selected without query labels. Scores average compounds within mechanisms and then mechanisms equally.

The original depth-four, plane-zero and training depth-eight, plane-three layouts preserve views, pixels, pooling and normalization. Under matched balanced-ridge fitting, adapted final-token well-level NSC increases from 0.5853 to 0.6823, a change of 0.0970 $[0.0318, 0.1626]$. NSCB changes from 0.4284 to 0.4299: 0.0014 $[-0.0503, 0.0502]$. Intervals resample query compounds within mechanisms conditional on trained heads and overlapping fitting populations; they omit encoder-training variation and correction for the full exploratory family.

Layout	Readout	Raw Dice	Object F1	Count MAE
Original	Incoming fusion	0.97232	0.8451	7.611
Original	Later fusion	0.97234	0.8466	7.722
Original	Later skip	0.97189	0.8483	7.574
Training	Incoming fusion	0.97271	0.8433	8.065
Training	Later fusion	0.97282	0.8478	7.593
Training	Later skip	0.97182	0.8460	7.519

Table 22: Three head seeds on 36 reused test images, not twelve independent plates. Object F1 uses IoU 0.5; count MAE is decoded instance error. Paired plate-weighted contrasts are not differences between these rounded image-weighted entries.

At fixed training layout, all four final-token adapted-minus-incoming intervals include zero. Pyramid features have positive well-level lineage contrasts but not the corresponding compound-dose-week contrasts; adapted pyramid compound-dose-week NSC falls under the layout change. DINO CLS and CellProfiler exceed adapted tokens on all four tested endpoints, with different pretraining and cost. Layout sensitivity neither establishes a universal input correction nor makes every feature family uninformative.

I.2 Nuclear overlap and object readouts

For BBBC039, the original depth-four, plane-one layout becomes depth eight, plane three while measured pixels, normalization, XY padding and observed world positions remain fixed. Pitch is nominal, so this is not a physical-scale robustness experiment. Incoming fusion, adapted fusion and adapted skip-only features in both layouts receive three head seeds each: eighteen 900-update AdamW fits, with identical validation-only selection and area-32/prominence-2 watershed. Original layouts are refitted rather than replaced by favorable historical heads.

At the training layout, adapted-minus-incoming raw Dice is 0.000121 $[0.000019, 0.000246]$, approximately 0.012 percentage points. Object F1 changes by 0.00429 $[-0.00154, 0.01118]$ and count MAE by -0.351 $[-1.642, 0.698]$. The original-layout Dice contrast includes zero. The table averages images and seeds; the 35 paired contrasts instead average plates equally using crossed seed/plate bootstrap draws. The small resolved foreground change does not establish an improved object measurement. Source audits reproduce the original features, all eighteen selected-head probabilities and independent object matching; the shared head and watershed are not independently implemented predictors. Numerical agreement does not resolve opaque ancestral exposure or increase the four test plates.

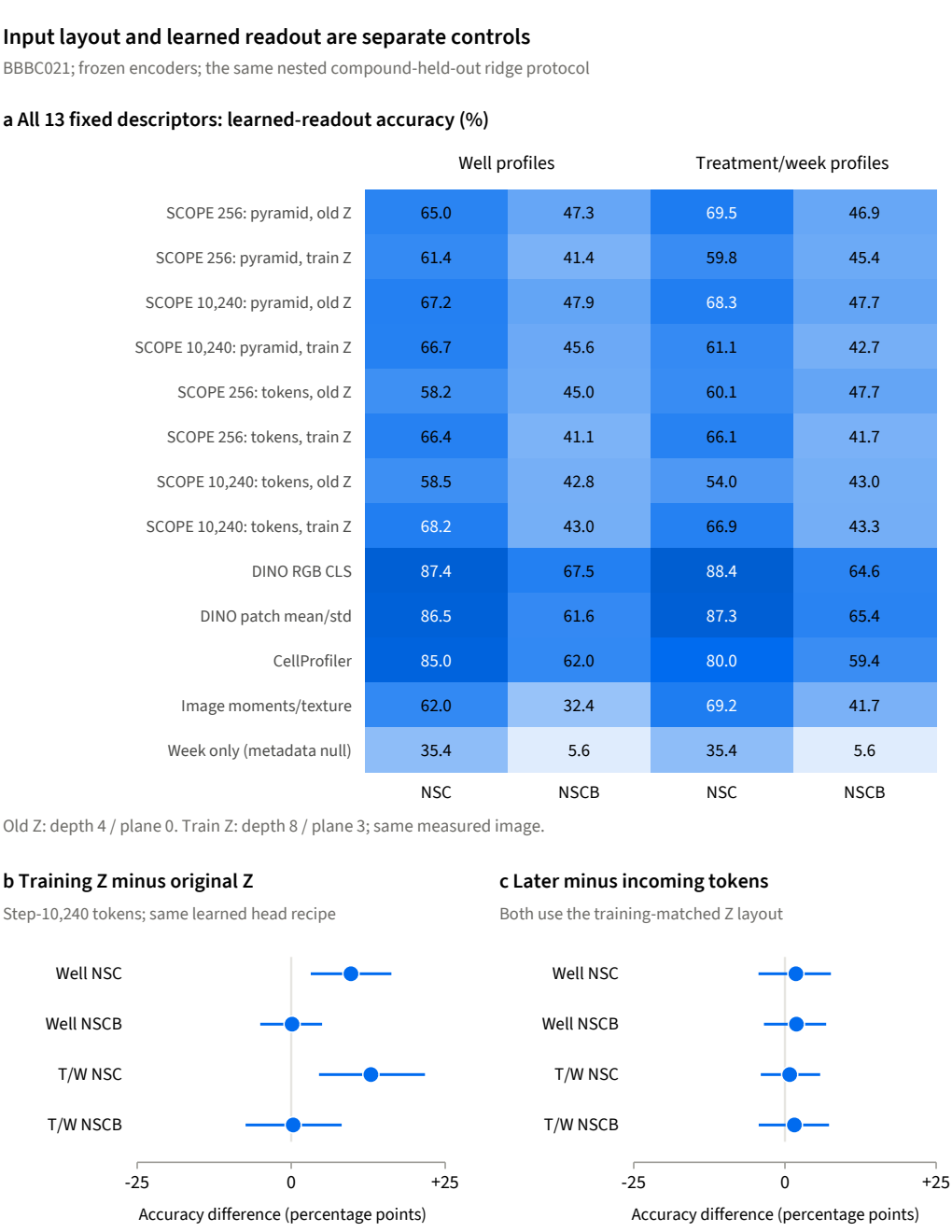

Figure 8: **Input layout changes some phenotype conclusions, not all of them.** The complete learned-readout grid retains the compound-excluded and compound-plus-week-excluded endpoints, stronger external features and adverse layout contrasts. This is a reused phenotype panel with conditional intervals, not a new acquisition replication.

J FROZEN SELECTION RULES ON EXTERNAL REPORTER ASSAYS

The readout controls diagnose why encoder comparisons can differ. A separate experiment tests a proposed practical response: whether a validation-based measurement gate improves the model selected for a new reporter assay. The unchanged rule and strong simple comparators are retained even where the gate does not gain an advantage.

J.1 CANDIDATE UPDATES AND SELECTION RULES

The original three-seed dense/DINOv2 study offers thirteen candidates per family/seed: the source and four snapshots from each of three 512-update continuations. Four rules use validation only: keep the source; maximize culture balanced accuracy; maximize culture accuracy subject to no point-estimate object-F1 or count-MAE regression; or minimize a joint risk. The joint score averages culture error, object-F1 error and count MAE divided by $\max(\text{source count MAE}, 1)$. Ties favor the source and then the declared arm/step order. The gate's zero-tolerance point estimates are not a multiple-comparison-adjusted noninferiority test. This selection experiment is distinct from the later ten-seed horizon and unchanged-head comparisons.

The complete initial test outcomes remain in Table 23. Dense gate selection is worse than joint validation on the three displayed means; DINOv2 trades culture accuracy for a small F1 change at equal mean count MAE. Equal-plate, seed-crossed gate-minus-culture-selection F1 intervals also include zero: dense -0.00204 $[-0.01010, 0.00642]$ and DINOv2 $+0.00113$ $[-0.00428, 0.00959]$. Their count changes are $+0.37589$ $[-0.66111, 1.73968]$ and -0.05089 $[-0.79792, 0.62639]$. These estimands differ from the table's equal-field means.

J.2 EXTERNAL OBSERVATIONS AND THE FIXED MEASUREMENT

BBBC013 contains one 96-well U2OS translocation plate with DNA and FKHR-EGFP images (Ravkin & Broad Institute Imaging Platform, n.d.a; Logan & Carpenter, 2010). Twenty-six unique pipelines are fixed before processing the assay: eight dense and eight DINO choices, three source and three adapted SCOPE heads, three scratch heads and one classical pipeline. Several selectors choose the same model. All existing heads and instance-conversion settings remain fixed; no assay image fits a head or threshold. Only DNA enters segmentation. Reporter intensity enters the measurement after all 2,496 pipeline–well prediction fields are sealed, and dose/well identities never enter the segmenter. Native tiling covers the full field. Source/manifold exposure checks leave opaque ancestral aliases and provider pretraining unresolved.

For predicted nucleus I_{wk}, define inner support N_{wk} by erosion of one 8-neighbor pixel. Its perinuclear support C_{wk} is background whose nearest nucleus is k and whose distance is in $(1, 10]$ native pixels. All predicted nuclei, including subsequently rejected objects, exclude ring pixels; the fixed distance-transform convention resolves ties. These rings are not verified cytoplasmic boundaries. Eligible objects avoid the image border and have at least eight pixels in both supports. The well measurement is

$$m_w = \frac{1}{|\mathcal{I}_w|} \sum_{k \in \mathcal{I}_w} \log_2 \frac{\overline{F}_{N_{wk}} + 0.5}{\overline{F}_{C_{wk}} + 0.5}. \tag{J.1}$$

The offset is in released 8-bit intensity units. Objects are equally weighted within well, then wells enter the analysis. A well without eligible objects is a failure, not an exclusion. Nominal pixel pitch does not calibrate the measurement radii in micrometers.

BBBC013 has two 36-well dose series (nine doses, four rows), eight shared positive and eight negative controls, and eight additional no-drug wells. No-drug wells are not empty-cell references. The primary endpoint is availability together with signed Spearman dose association; incomplete or constant series are undefined. The secondary shared-control statistic (Zhang et al., 1999) is $Z' = 1 - 3(s_+ + s_-)/|\mu_+ - \mu_-|$, using sample standard deviations; zero contrast is undefined. Shared controls do not create two replications.

BBBC014 is one further plate with a complete 48-well series for each of MCF7 and A549. The provider pipeline and four pixel-exact examples identify channel 2 as DNA/DAPI and channel 1 as NFkB/FITC. The same twenty-six pipelines, selections and native-pixel measurement rule are

Table 23: **A measurement gate is not automatically a better selector.** Validation-only choices among the same 13 checkpoints, evaluated on test observations. Each number averages all three seeds; nuclear F1 at IoU 0.5 and count MAE first average the 36 fields. These descriptive field-weighted means differ from the equal-plate paired intervals in the text. Kept counts source selections out of three, not rejected biological samples. The gate uses zero-tolerance validation point estimates, not a safety certificate.

Encoder	Selection	Culture BA	Object F1	Count MAE	Kept
Dense	Source only	0.6829	0.8443	7.59	3/3
Dense	Culture validation	0.7167	0.8442	7.33	1/3
Dense	Measurement gate	0.6785	0.8422	7.84	1/3
Dense	Joint validation	0.7322	0.8447	7.56	0/3
DINOv2-small	Source only	0.9357	0.7761	9.01	3/3
DINOv2-small	Culture validation	0.9661	0.7768	8.73	2/3
DINOv2-small	Measurement gate	0.9704	0.7775	8.58	2/3
DINOv2-small	Joint validation	0.9766	0.7752	8.58	0/3

	BBBC013		BBBC014	
Pipeline / contrast	Wortmannin	LY294002	MCF7	A549
SCOPE source	0.8775	0.8754	0.8478	0.9196
SCOPE adapted	0.8799	0.8799	0.8504	0.9189
Scratch	0.8734	0.8903	0.8597	0.9254
Classical segmentation	0.8668	0.9124	0.8830	0.9283
Whole-field channel correlation	0.8275	0.8606	0.8804	0.8992
Dense gate minus joint	+0.0124	-0.0035	-0.0052	-0.0022
DINOv2 gate minus joint	-0.0045	+0.0007	-0.0020	+0.0003

Table 24: **Frozen external measurements and selection comparisons.** Native signed dose associations; learned rows average three fixed head seeds. Classical and whole-field controls are single procedures. Every pipeline returns measurements on all 96 wells of each assay. Selector rows compare unchanged validation choices, not a rule fitted on these assays. One plate per assay and three fitted heads do not support an interval over new plates. The original assay diagrams retain every fixed-seed contrast.

unchanged. The extension was specified after BBBC013 outcomes but before its own outcomes; dose and column are confounded. Again all 2,496 predictions yield measurements. Neither assay supplies instance ground truth or absolute molecular-error labels.

J.3 Complete assay comparisons

Table 24 places the native pipeline outcomes beside the same selector contrasts on both assays. Source/adapted SCOPE comparisons remain attached to checkpoints 256/10,240, not the terminal checkpoint. The signed association is not a segmentation metric, and small changes are not evidence of equivalence.

The gate's advantage changes sign across families and endpoints. All three dense seeds favor it for BBBC013/Wortmannin, but that contrast does not represent the full panel. A half-field toroidal reporter shift removes the strong positive association while preserving predicted compartments. This shows dependence on spatial correspondence, not exact localization or an exchangeable biological null. Whole-field channel correlation retains substantial signal without segmenting cells. On BBBC013, its $Z' = 0.5978$ and the classical control's 0.7197 exceed source/adapted SCOPE's 0.3893/0.4406; scratch gives -0.0179. Raw reporter mean instead has negative associations $-0.8192/-0.7943$. These controls test what the measurement actually needs, rather than granting necessity to a large encoder.

The source audits reproduce all well reductions and independently inspect 104 predetermined pixel/geometry cases per assay, with the recorded 2×10^{-12} arithmetic tolerance. They validate execution, not biological reference truth. Two different assay plates extend the tested conditions

without supplying replicated plates of either assay. No selector was retuned in response to these outcomes. The completed comparison bounds the proposed gate's usefulness; it neither establishes a superior replacement nor makes the measured spatial signal invalid.

K Reproducibility, Access, and Interpretation

The supplement separates the implemented predictor, completed study outcomes and the computations included in this release. The public artifact supports substantial numerical reanalysis. It is not a complete retraining package for SCOPE or a release of every private neural prediction.

K.1 What defines a comparable run

Observation. Preserve input-channel semantics, permitted targets, native lattice, affine, valid support, masking and normalization. Layout and reference masks must not silently select an easier input. Nominal pitches in uncalibrated panels are not measured instrument calibration. Crop and frame counts remain separate from plate, well, embryo and animal replication.

State and updates. Weights, buffers, external runtime switches, trainable scope, sparse-attention phase and positional references jointly define the function. Resuming an update also requires optimizer moments and clocks, EMA, rank randomness, sampling cursors and schedule position. Source identities in Table 11 document known ancestry without completing missing early exposure. A zero gradient can advance optimizer state where an absent gradient does not; distributed reduction must preserve the declared episode/rank weighting.

Readout and decision. Specify co-trained, unchanged inherited, source-fitted or freshly fitted heads by origin, training labels, update count and validation rule. Include all validation annotations in label accounting. When each readout selects its own learning rate, it can select a different encoder as well as a different head. An expanded decoder grid is an increased deployment allowance even when no encoder weights change. Selecting it on development does not make its selected validation maximum a generalization estimate.

K.2 What the distributed computations reproduce

The package retains the uploaded executable scientific code, numerical archives and table sources unchanged. Pre-rendered figures allow compilation without model weights, GPU access, external downloads or shell escape. Figure regeneration may have separate optional dependencies; font binaries are not distributed in this revision.

For example, the following commands use the included aggregates. Output paths must be new because the replay scripts refuse silent overwriting:

```
python reanalysis/allen_resampling_v2.py \
  --input reanalysis/data/allen_large_20260924.json \
  --output /tmp/allen_large_replayed.json
python reanalysis/availability_replay.py \
  --input reanalysis/data/availability_20260924.json \
  --output /tmp/availability_replayed.json
python reanalysis/readout_interaction.py
python reanalysis/spatial_access_controls.py
python -m unittest discover -s reanalysis -p 'test_*.py' -v
```

The core does not implement interrupted-run resume; partial runs remain incomplete. Full GPU trajectory parity and independent full training are unverified for that extraction. Source and runtime versions, archive hashes and original analysis-specific seeds remain in the corresponding included files. The qualitative panels are outcome-ranked illustrations and retain their original source metadata; they do not add independent test cases. The exact category-level rules and complete per-cell source-relative vectors for the reported 26-cell hidden-loss classification are not fully distributed here. Its reported outcome must not be represented as an independently reconstructible null from only the displayed mean contrasts.

Included component	Scope of reproduction
`reproducibility/dense_core/` and companion utilities	Standalone scheduled-recipe dense core with fresh AdamW, warmup and cosine decay. It does not by itself restore every private inherited-optimizer source or the complete SCOPE training history.
`reanalysis/allen_resampling_v2.py`	Corrected plate-multiplicity bootstrap from the released Allen aggregates; not new predictions or biological-reference adjudication.
`reanalysis/availability_replay.py`	All 36 culture contrasts and paired readout interactions from the 60 saved seed/arm/horizon summaries.
`reanalysis/readout_interaction.py`	Historical restricted/full-pyramid interaction with crossed and three-seed Student-t sensitivity.
`reanalysis/spatial_access_controls.py`	Controlled A/B/C raw-Dice access and capacity interactions at both fitting allowances, including all 6,912 crossed seed/embryo resamples.
Anonymous variable-count aggregates	Plate sums for the focused field-weighted count/localization contrasts; no individual object masks or raw acquisitions.

Table 25: Reanalysis is matched to the released information. Archived source receipts describe additional checks; they are not new checks performed by compiling or reading this manuscript.

K.3 What remains outside this release

The full SCOPE weights, private feature caches, raw acquisition preparations, production CUDA kernels and every intermediate training state are not in this package. Public dataset access does not reconstruct unrecorded ancestral exposure or supply the private cached predictions. Research-use and imaging permissions remain applicable. Reproducing aggregate arithmetic, replaying neural predictions, rerunning training and testing new acquisitions are different levels of evidence.