
One-Slide Calibration of Pathology Foundation Models

Ming Ren Hou
App-In Club
ming.hou@appinclub.org

Tianyi Huang
Ryquo
tianyi@ryquo.com

Abstract

Scanner variation changes how pathology foundation models represent the same tissue. We introduce SLIDERULER, which uses regions within a slide as internal controls to estimate and correct acquisition-induced shifts in other regions. A transfer map learned from paired rescans enables calibration from a single scan at inference while keeping the foundation model fixed. Across two encoders and five SCORPION scanners, learned transfer reduces mean target-to-source embedding distance by 16.3–38.5% relative to raw embeddings. Comparisons with unrelated same-scanner controls reveal a positive same-slide contribution across all four evaluation settings, including scanner holdout. A source-anchored variant reduces source-feature displacement by 47.7–83.6% relative to learned transfer while retaining most of its alignment gain. By drawing calibration information from the slide itself, SLIDERULER offers a path toward more consistent use of frozen pathology models across imaging systems.

1 Introduction

A pathology foundation model sees both tissue and the imaging system that digitized it. Optical response, focus, compression, and color processing can change a representation even when the glass slide is fixed [4, 6]. Scanner robustness is therefore partly a measurement problem: how can a model distinguish acquisition variation from the biological information that its downstream tasks require? This distinction matters for deployment across hospitals and devices, where retraining a foundation model or assembling a target-domain calibration cohort may be impractical.

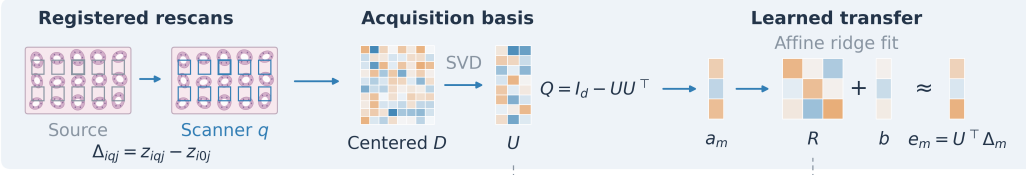
Paired rescans provide a direct way to measure acquisition variation. Comparing the same tissue across devices reduces the biological confounding present in comparisons between unrelated cohorts. Recent methods use this evidence to learn global transformations of frozen embeddings [3, 12]. An additional question remains: *does the slide being processed contain information about its own correction beyond the correction shared across slides?* Answering it requires a comparison that preserves the scanner while changing the identity of the control tissue.

SLIDERULER estimates acquisition shifts from other regions of the same slide. Paired training scans define an acquisition subspace and supervise a map from control displacement to query-region displacement. At inference, the model uses source prototypes and controls from the current scan, without requiring a paired rescan, scanner identity, or target-cohort statistics.

We study two estimators. Learned transfer combines direct control subtraction with an affine ridge map. A source-anchored variant combines a control-modulated global shift with a regularized residual, using source-feature displacement to guide model selection. Prostate biopsies provide a concrete setting for tissue-defined controls: benign glands and stroma can inform the correction of a candidate tumor region.

Our contributions connect three aspects of one-slide calibration. First, same-scanner control replacement measures the additional predictive information supplied by the current slide. Second, a source-anchored residual estimator provides an explicit tradeoff between target alignment and source-

TRAINING



INFERENCE

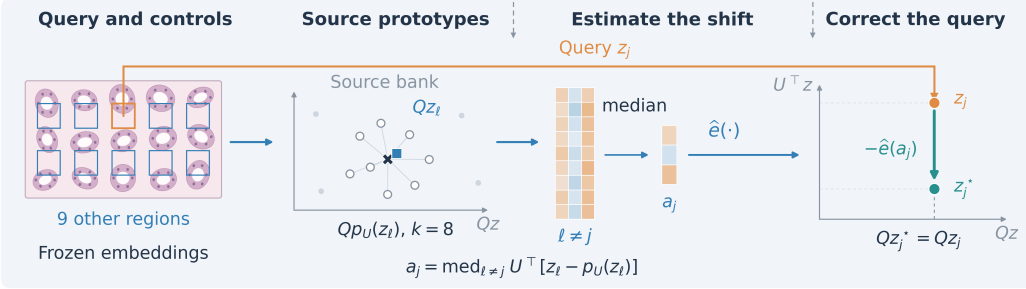


Figure 1: **One-slide calibration with learned transfer.** Top: paired rescans define the acquisition basis U from residuals centered across scanners within each registered region and train the transfer map. Bottom: a single scan supplies control displacements relative to source prototypes; learned transfer corrects the query while preserving Qz_j . Controls exclude the query; training prototype matches exclude its entire slide. The nine-control SCORPION example is schematic.

feature displacement. Third, patient-held-out tissue selection and a fixed-cohort prostate-biopsy study evaluate the availability of tissue-defined controls. These analyses connect a slide’s calibration signal to the correction it induces and the tissue needed to estimate it.

2 Related work

Acquisition harmonization. Color transfer and stain normalization address appearance variation [11, 14]; ComBat and CORAL adjust feature distributions using batch structure or domain statistics [7, 17]. Paired-data approaches such as FEATMAP and post-hoc acquisition shrinkage learn corrections directly in foundation-model embedding spaces [3, 12]. PathoROB evaluates biological representation structure alongside downstream robustness, motivating evaluation of both alignment and information preservation [9]. Our focus is the incremental information available within one test slide. We compare against explicit pooled-mean, scanner-mean, affine-ridge, and subspace-shrinkage implementations.

Control tissue and scanner benchmarks. MuSCID uses off-target organs collected at a deployment site for cross-organ calibration [20]; non-tumor regions have also been used to define pseudo-domains for pathology domain generalization [16]. SLIDERULER uses within-slide controls to predict a region-specific acquisition shift. SCORPION supplies 480 aligned regions scanned on five devices [15]. PAR contains three scans of each of 339 prostate-biopsy slides from 185 patients [1]. Together, paired-region and whole-slide data support separate measurements of acquisition transfer and control availability.

3 One-slide acquisition calibration

3.1 Paired training and within-slide controls

Let $z_{iqj} \in \mathbb{R}^d$ be a frozen encoder’s embedding for region j of slide i on scanner q , standardized with source-training statistics. Scanner 0 is the designated source. Registered training regions define shifts $\Delta_{iqj} = z_{iqj} - z_{i0j}$. All scans of a physical slide remain in the same fold. Training permits paired rescans; inference receives only the query scan.

Let $\mathcal{Q}_{\text{tr}}$ contain the permitted training scanners, including source scanner 0. For learned transfer, center each registered region across these scanners and stack the residuals:

$$d_{iqj} = z_{iqj} - \frac{1}{|\mathcal{Q}_{\text{tr}}|} \sum_{q' \in \mathcal{Q}_{\text{tr}}} z_{iq'j}, \quad U = [u_1, \dots, u_r], \quad Q = I_d - UU^\top. \quad (1)$$

The rows of D are $d_{iqj}^\top$, and $u_1, \dots, u_r$ are its leading right singular vectors. Thus $U \in \mathbb{R}^{d \times r}$ has orthonormal columns. We retain the smallest rank reaching 90% of residual energy, or the rank cap of 16 when that threshold is not reached. The projector Q removes these acquisition coordinates for prototype matching.

Let $p_U(z)$ be the mean of the $k = 8$ nearest source-bank embeddings to z under distance $\|Q(z - p)\|_2$. The control vector for query region j in scan X is

$$a_j(X; U) = \text{med}_{\ell \neq j} U^\top [z_\ell - p_U(z_\ell)]. \quad (2)$$

The median is coordinate-wise, and $a_j \in \mathbb{R}^r$. Training matches exclude every prototype from the query slide; held-out slides use the training bank. On SCORPION, each query has nine other regions as controls. A tissue-defined extension replaces these regions with selected benign-gland and stromal fields, with separate prototype banks and transfer maps (Appendix E).

3.2 Learned transfer

Write m for a training slide/scanner/query example, a_m for its control vector, and $e_m = U^\top \Delta_m$. Fit an affine ridge map with $R \in \mathbb{R}^{r \times r}$ and $b \in \mathbb{R}^r$, then blend its prediction with direct subtraction:

$$(R, b) = \arg \min_{R, b} \sum_m \|e_m - Ra_m - b\|_2^2 + \lambda \|R\|_F^2, \quad (3)$$

$$\widehat{e}(a) = (1 - \alpha)a + \alpha(Ra + b), \quad z_j^* = z_j - U\widehat{e}(a_j).$$

Source examples have $\Delta_m = 0$. The original region-transfer estimator uses $\lambda = 1$. Five inner physical-slide folds rebuild standardization, basis, prototypes, and regression. The shared blend $\alpha \in \{0, .01, \dots, 1\}$ minimizes a worst-scanner score based on slide-mean projected residual distance (Appendix A). The direct-control baseline uses $\alpha = 0$.

3.3 Source-anchored global and residual correction

The second estimator balances target alignment against source-feature displacement. Its basis U_A is fitted to group-weighted, *source-referenced* shifts Δ , and $g = \mathbb{E}_w[\Delta] \in \mathbb{R}^d$ is their pooled mean over non-source scanners. Each training group receives equal total mass. Define the centered control vector and global activation by

$$\widetilde{a}_j = a_j(X; U_A) - \mu_0, \quad v = U_A^\top g, \quad t(a) = \text{clip}\left(\frac{a^\top v}{v^\top v}, 0, 2\right), \quad (4)$$

where μ_0 is the group-weighted mean of source control vectors computed with own-group prototypes excluded. Set $t = 0$ if $v^\top v \leq 10^{-8}$. The global prediction $t(a)g$ retains the full embedding dimension.

A zero-intercept ridge map predicts the remaining acquisition coordinates:

$$B_\lambda = \arg \min_B \sum_m w_m \|U_A^\top [\Delta_m - t(\widetilde{a}_m)g] - B\widetilde{a}_m\|_2^2 + \lambda \|B\|_F^2, \quad (5)$$

$$r(a) = \beta B_\lambda a \min\left\{1, \frac{\|a\|_2}{\max(\|\beta B_\lambda a\|_2, 10^{-8})}\right\}, \quad \widehat{\Delta}_A(a) = t(a)g + U_A r(a). \quad (6)$$

The fitted correction is $z_j^* = z_j - \widehat{\Delta}_A(\widetilde{a}_j)$. Source examples again have zero target shift. The residual cap bounds its acquisition-coordinate norm by the control norm. At the learned anchor $\widetilde{a} = 0$, the correction is exactly zero.

Four penalties $\lambda \in \{.1, 1, 10, 100\}$ and three residual weights $\beta \in \{.25, .5, 1\}$, plus global-only and identity, give 14 candidates. Five inner grouped folds select

$$M^* = \arg \min_{M \in \mathcal{M}} \left[\max_{q \in \mathcal{Q}_{\text{tr}} \setminus \{0\}} \|\Delta - \widehat{\Delta}_M\|_{2_q} + \|\widehat{\Delta}_M\|_{2_0} \right]. \quad (7)$$

Each mean averages regions within a slide, slides within a group, and then groups equally. Every learned component, including μ_0 and g , is rebuilt inside each fold. The selected candidate is refitted on the outer training set. The global-only candidate retains control-dependent activation.

3.4 Tissue-defined controls

For prostate biopsies, we train a three-class linear selector on frozen Prost40M embeddings from an external annotated dataset [2]. Its classes are benign gland, residual stroma, and tumor. A field is retained when its largest predicted class probability is at least 0.8; otherwise it is unassigned. The cutoff is fixed before evaluation and applies to uncalibrated probability estimates. Five outer patient-grouped folds evaluate the selector; three inner folds select logistic-regression regularization by class- and patient-balanced out-of-fold log loss. Both levels use stratified group splits. The final selector is fitted on all external patients and then frozen for PAR, where fields in each scanner image are selected independently from their uncorrected embeddings.

4 Experimental design

Paired scanner evaluation. SCORPION contains 48 physical slides, ten registered regions per slide, and five scanners [15]. We extract 384-dimensional embeddings with frozen Prost40M and DistillPath-IS16-HOpt0 [5, 8, 13, 18]. Leica AT2 is the source. Six outer splits use 32 training, eight reserved validation, and eight test slides. The calibrated protocol permits training pairs from all scanners; scanner holdout excludes the evaluated device from all fitting and selection. Each encoder/protocol condition contains 1,920 held-out target-region comparisons.

Comparisons and endpoints. We evaluate raw embeddings, pooled-mean and scanner-mean correction, affine ridge, subspace shrinkage, direct controls, learned transfer, and the anchored variant. All SCORPION alignment comparisons apply the estimators without the optional strict gate. We measure mean target-to-source Euclidean distance in source-standardized coordinates, source-feature displacement $\|z^* - z\|_2$ on source scans, and top-1 same-region retrieval accuracy among the 80 source regions in an outer test fold. Material distance worsening means an increase exceeding 2% of the median raw non-source distance on the eight reserved validation slides. The anchored variant does not use those slides for fitting or candidate selection.

Isolating slide identity. Thirty-two seeded derangements replace only the control regions with those of another slide on the same scanner and in the same test fold. The query embedding remains fixed. We average the paired effects across permutations before computing pointwise 95% intervals from 5,000 physical-slide bootstrap resamples. The comparison measures the benefit of same-slide controls while holding the query, scanner, and learned estimator fixed. The donor pool and seeded assignments remain fixed during resampling.

External selection and PAR sampling. The external selector study uses 8,325 fields from 37 slides and 32 patients. Evaluation gives each patient total weight one; 95% intervals use 2,000 patient bootstrap resamples of saved out-of-fold predictions. For PAR, a fixed patient ordering selects 20 patients and all their biopsies, yielding 35 slides and 105 scans. Reference and dense policies request 512 and 2,048 fields per scan, respectively, sampled independently on each scanner image at a nominal $112\ \mu\text{m}$ field size, with up to 128 and 512 registered fields. The policies use non-nested sampled locations and otherwise fixed extraction, selection, and registration settings. Control-count eligibility requires at least 16 benign-gland and 16 stromal fields per scan. All assigned scans remain in the denominators.

Controlled compartment experiment. A 64-run synthetic suite combines eight regimes and eight seeds, with known acquisition shifts and separate benign, stromal, and lesion compartments. It evaluates the two-compartment transfer and selective-correction implementation. All seeds contribute to the reported means; a failed outer screen returns the uncorrected embeddings. The suite tests acquisition-sensitive and already robust ordinal heads, tissue-response differences, scanner extrapolation, and calibration-set size.

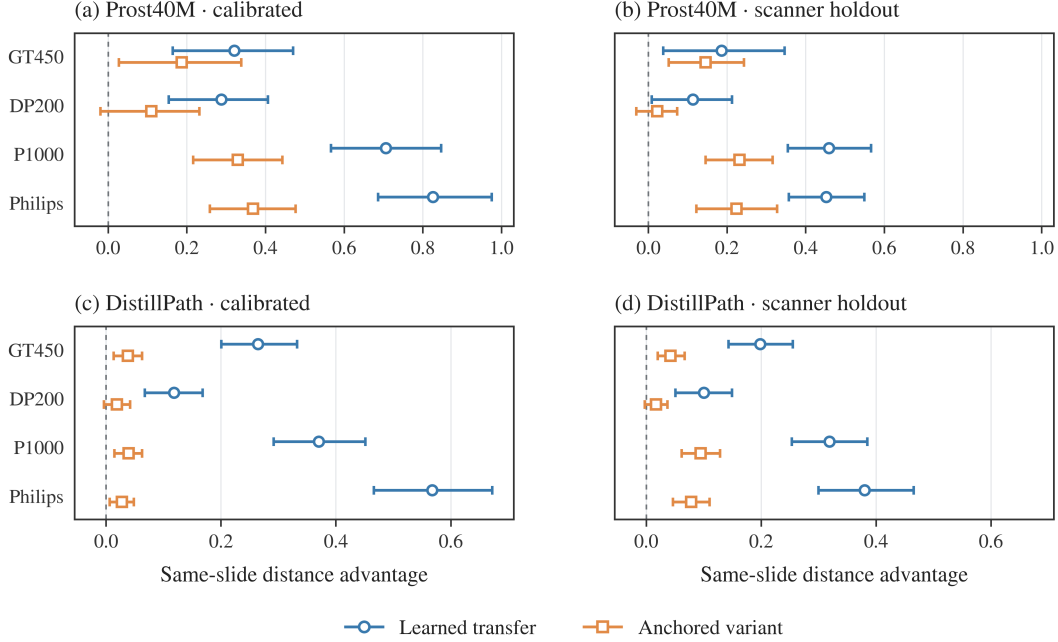


Figure 2: **Slide-specific acquisition information across scanners.** Same-slide distance advantage over unrelated same-scanner controls, disaggregated by target device. Positive values favor same-slide controls. Points and bars show mean differences and pointwise 95% physical-slide bootstrap intervals over 48 slides, after averaging 32 within-fold donor permutations. The query embedding is unchanged. Limits are shared within encoder rows; distances use each encoder’s source-training scale.

5 Results

5.1 Same-slide controls improve acquisition correction

Learned transfer reduces paired distance relative to raw embeddings in every encoder/protocol setting, by 16.3–38.5% (Table 1). Same-slide controls improve on unrelated same-scanner controls by 0.535 and 0.303 for Prost40M, and by 0.330 and 0.249 for DistillPath, in calibrated and held-out settings respectively. All four pooled intervals exclude zero, as do all 16 learned-transfer scanner-specific intervals; the pooled effect is positive under each of the 32 permutations. Figure 2 disaggregates the effect by device. These results show that same-slide controls improve prediction of paired acquisition shifts beyond unrelated controls from the same device.

Table 1: **Paired embedding alignment on SCORPION.** Mean target-to-source distance; lower is better. All corrections are ungated. Bold denotes the lowest point estimate in each row. Scanner mean falls back to pooled mean for a held-out scanner. Distances use each encoder’s source-training scale.

Encoder	Protocol	Raw	Scanner mean	Affine ridge	Shrinkage	Learned	Anchored
Prost40M	Calibrated	18.654	11.328	9.640	13.181	11.470	12.030
	Holdout	18.654	15.940	13.294	14.801	13.871	14.196
DistillPath	Calibrated	17.551	13.805	14.447	13.808	13.843	14.591
	Holdout	17.551	15.840	15.420	14.690	14.682	14.973

Global baselines remain competitive in mean distance: affine ridge attains the lowest Prost40M values, while scanner mean and subspace shrinkage perform well for DistillPath. Retrieval and distance-worsening rates provide complementary comparisons. Under Prost40M scanner holdout, learned transfer achieves 99.74% retrieval accuracy versus 98.18% for affine ridge, a paired advantage of 1.56 percentage points (95% CI 0.83–2.34), with material distance worsening of 3.07% versus 19.58%. Under DistillPath holdout, learned transfer and shrinkage have similar mean distance, with worsening rates of 2.40% and 6.77%, respectively.

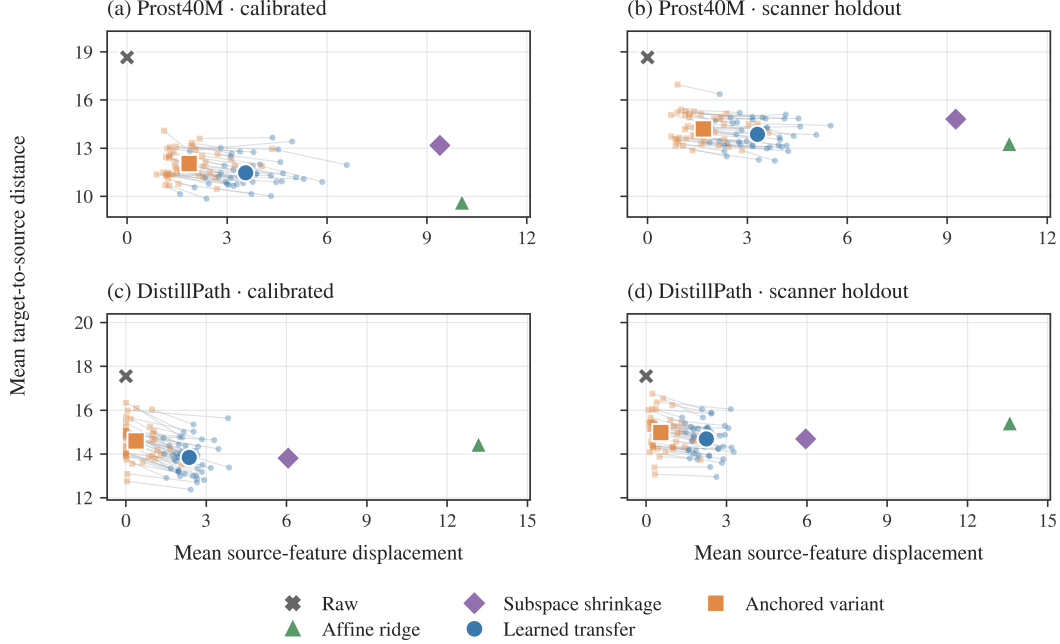


Figure 3: **Alignment and source displacement form distinct evaluation axes.** Large markers are method means; faint blue/orange points show learned/anchored results for each of 48 physical slides, joined within slide. Lower values on both axes are preferable. Target values average ten regions and four target scanners; source values average ten regions and one fitted model per slide when calibrated, or four models under scanner holdout. Axis limits are shared within encoder rows; distances use source-standardized coordinates.

5.2 Source anchoring changes the operating point

The anchored estimator reduces mean source-feature displacement by 47.7–83.6% across the four settings. Figure 3 shows the joint alignment–displacement comparison and its physical-slide variation; source displacement decreases on all 48 slides in every setting. This reduction accompanies an increase of 0.291–0.749 in mean target distance relative to learned transfer; all four paired intervals exclude zero. Pooled material distance-worsening rates also decrease in every setting (Appendix C). Anchoring retains 79.8–93.2% of learned transfer’s distance reduction over raw embeddings, providing a way to reduce source displacement while preserving most of the target alignment gain.

Training-only model selection retains a residual component in 42 of 60 unique fits and selects the control-modulated global correction in the remaining 18. All Prost40M fits retain a residual, whereas all six calibrated DistillPath fits select the global correction. This variation shows how model selection adjusts correction complexity across encoders. The anchored estimator retains positive pooled same-slide advantages of 0.031–0.249 (Appendix B).

5.3 Denser sampling increases tissue-control availability

Table 2: **External patient-held-out tissue selection.** Precision and recall are percentages with 95% patient bootstrap intervals. Counts denote reference-class fields and patients containing that class. Residual stroma is the dataset’s postprocessed remaining-tissue class.

Reference class	Fields	Patients	Precision	Recall
Benign gland	337	21	89.37 [80.05, 95.72]	89.51 [82.19, 94.40]
Residual stroma	7,028	32	99.98 [99.93, 100.00]	98.16 [96.78, 99.18]
Tumor	960	20	96.32 [91.85, 98.69]	81.40 [71.34, 89.38]

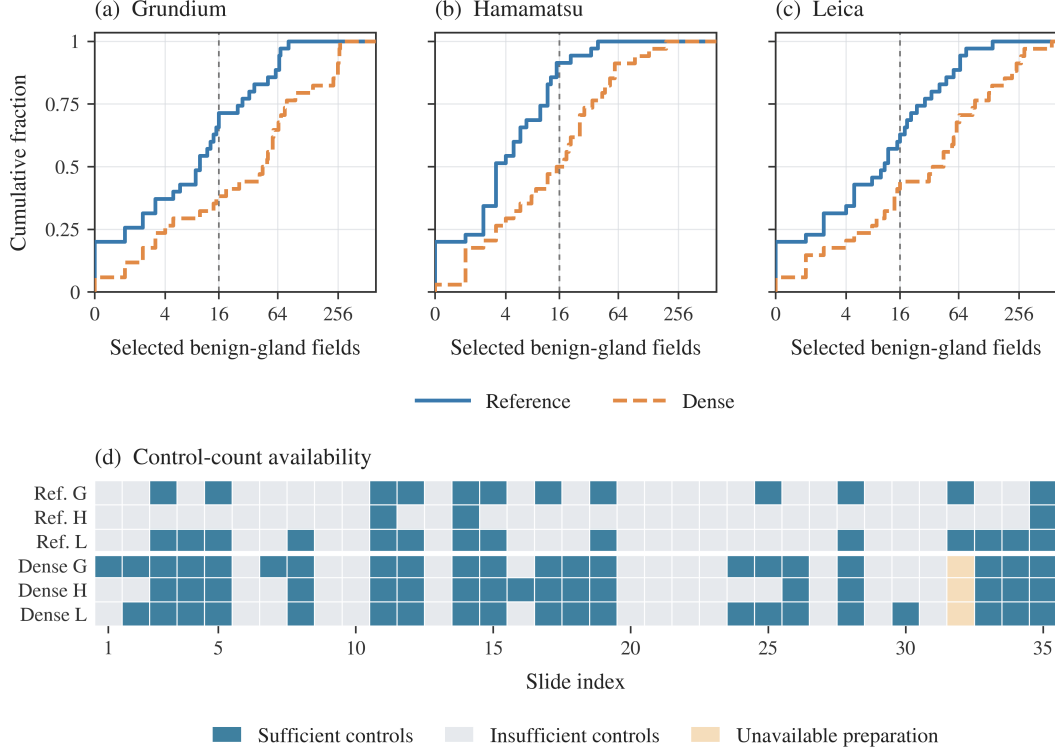


Figure 4: **Control availability across sampling policies.** (a–c) Empirical cumulative distributions of benign-gland counts for 35 reference and 34 dense scans per scanner. The horizontal coordinate is $\log_2(n + 1)$, with ticks labeled by count n ; dashed lines mark 16 fields. (d) Availability requires at least 16 benign-gland and 16 stromal fields per scan. Columns represent all 35 slides, with unavailable preparations marked. G/H/L denote Grundium/Hamamatsu/Leica. Sampling locations are non-nested.

The external selector achieves patient-weighted precision of 89.37% for benign glands, 99.98% for residual stroma, and 96.32% for tumor (Table 2). Coverage is 96.75%. Four additional patient partitions give benign precision of 90.81–93.05% at the same cutoff.

On PAR, dense sampling doubles the number of scans meeting both control-count requirements from 29/105 to 58/105 (Table 3). Among the 102 scans available under both policies, 27 pass under both, 31 pass only under dense sampling, and none pass only under reference sampling. Hamamatsu shows the largest absolute increase, from 3 to 17 eligible scans out of 35. Every available dense scan meets the stromal count threshold, identifying benign-gland sampling as the remaining constraint on control availability. The full distributions and paired scanner outcomes appear in Figure 4. At slide level, all three scanners meet the control requirements on 3/35 reference slides and 16/35 dense slides.

Table 3: **PAR control-count availability.** Scans with at least 16 selected benign-gland and 16 stromal fields. Each scanner has 35 assigned scans. The three unavailable dense scans remain in the denominator.

Policy	Grundium	Hamamatsu	Leica	Overall
Reference (512 fields)	12/35	3/35	14/35	29/105 (27.6%)
Dense (2,048 fields)	21/35	17/35	20/35	58/105 (55.2%)

5.4 Controlled tests explain when transfer helps

In the acquisition-sensitive synthetic regime, the two-compartment method reduces shift root-mean-square error (RMSE) from 4.014 to 1.894, with paired reduction 2.120 (95% CI 1.468–2.512), and increases worst-target quadratically weighted kappa (QWK) from 0.828 to 0.949. Grade flips fall

from 56.7% to 15.7% across all eight seeds. Source embeddings remain unchanged in this regime. The corresponding kappa gain is smaller for an already robust head, 0.009 versus 0.121. When lesion and control responses differ strongly, learned ungated transfer improves RMSE from 3.766 for direct subtraction to 1.662. These comparisons link the benefit of learned transfer to tissue-response differences and the downstream head’s sensitivity to acquisition shifts.

6 Scope and limitations

Our evidence concerns embedding calibration and tissue-control availability. SCORPION provides aligned regions and physical-slide groups, but no patient identifiers, compartment annotations, or clinical grades. Its bootstrap intervals condition on the fitted models and development choices; public metadata also do not permit a complete audit of pretraining overlap. Acquisition coordinates may contain biological variation, and reduced source-feature displacement does not establish preservation of downstream predictions. The optional strict gate accepts no SCORPION corrections at this sample size; the reported real-scanner results therefore concern the ungated estimators.

PAR control counts are an intermediate endpoint. Registration and matched-tumor requirements leave at most six eligible training patients, below the fixed ten-patient minimum, so no end-to-end PAR calibration or grading comparison was run. Control purity on PAR requires expert validation, and denser sampling yields spatially clustered fields. Synthetic stress tests show that agreement between control compartments does not protect against shared control drift or altered lesion responses. These findings identify the next validation priorities: expert-verified controls, patient-separated downstream evaluation, and assessment of correction decisions under broader acquisition changes. The method is intended for research use; its statistical screens do not provide clinical risk guarantees.

7 Conclusion

A slide can supply information about how its own acquisition has changed its representation. SLIDERULER uses this information to predict corrections from within-slide controls, while source anchoring makes the balance between target alignment and source-feature displacement explicit. The tissue-selection and sampling results connect this principle to control availability in prostate biopsies. Together, the findings support a complementary route to scanner robustness: adapting frozen pathology representations using information already present in each slide.

References

- [1] Peshawa J. Muhammad Ali, Navin Vincent, Saman S. Abdulla, Han N. Mohammed Fadhl, Anders Blilie, Kelvin Szolnoky, Julia Anna Mielcarz, Xiaoyi Ji, Kimmo Kartasalo, Abdulbasit K. Al-Talabani, and Nita Mulliqi. The par dataset: Prostate biopsy whole slide images from an underrepresented middle eastern population, 2026. URL <https://arxiv.org/abs/2512.03854>.
- [2] Adriana Katherine Calapaquí Terán, Abel A. González Bernad, Lucía Sánchez Magdaleno, Sara Marcos González, Roberto Carlos Delgado Bolton, Jaled Moustafá Calvo, Miriam Cobo Cano, Lara Lloret Iglesias, and José Javier Gómez Román. Annotated whole-slide images of core needle prostate biopsies. Zenodo, 2026. URL <https://doi.org/10.5281/zenodo.21780258>.
- [3] Leonhard Donle, Michael Phillips, Farieda Gaber, Siddhi Ramesh, Matteo Sacco, Sampsa Hautaniemi, Anni Virtanen, Keno Bressem, Lisa Adams, Kelsey Goon, Elena Nevins, Ryan A. Robinett, Sara Kochanny, Sasha Hassan, James Dolezal, Alexander T. Pearson, and Ernst Lengyel. Featmap: Targeted correction of acquisition signatures harmonizes medical foundation model embeddings and enables robust task generalization. *bioRxiv*, 2026. doi: 10.64898/2026.07.02.736184. URL <https://www.biorxiv.org/content/early/2026/07/08/2026.07.02.736184>.
- [4] Savannah R. Duenweg, Samuel A. Bobholz, Allison K. Lowman, Margaret A. Stebbins, Aleksandra Winiarz, Biprojit Nath, Fitzgerald Kyereme, Kenneth A. Iczkowski, and Peter S. LaViolette. Whole slide imaging (wsi) scanner differences influence optical and

- computed properties of digitized prostate cancer histology. *Journal of Pathology Informatics*, 14:100321, 2023. ISSN 2153-3539. doi: 10.1016/j.jpi.2023.100321. URL <https://doi.org/10.1016/j.jpi.2023.100321>.
- [5] Clément Grisi, Khrystyna Faryna, Nefise Uysal, Vittorio Agosti, Enrico Munari, Solène-Florence Kammerer-Jacquet, Paulo Guilherme de Oliveira Salles, Yuri Tolkach, Reinhard Büttner, Sofiya Semko, Maksym Pikul, Axel Heidenreich, Jeroen van der Laak, and Geert Litjens. Deep learning from routine histology improves risk stratification for biochemical recurrence in prostate cancer, 2026. URL <https://arxiv.org/abs/2603.14187>.
 - [6] Xiaoyi Ji, Richard Salmon, Nita Mulliqi, Umair Khan, Yinxi Wang, Anders Blilie, Henrik Olsson, Bodil Ginnerup Pedersen, Karina Dalsgaard Sørensen, Benedicte Parm Ulhøi, Svein R. Kjosavik, Emilius A.M. Janssen, Mattias Rantalainen, Lars Egevad, Pekka Ruusuvaori, Martin Eklund, and Kimmo Kartasalo. Physical color calibration of digital pathology scanners for robust artificial intelligence–assisted cancer diagnosis. *Modern Pathology*, 38(5):100715, 2025. ISSN 0893-3952. doi: 10.1016/j.modpat.2025.100715. URL <https://doi.org/10.1016/j.modpat.2025.100715>.
 - [7] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 01 2007. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj037. URL <https://doi.org/10.1093/biostatistics/kxj037>.
 - [8] Ramon Kaspar, Andrey Ignatov, and Valentina Boeva. Distillpath: An efficient 22m distilled pathology encoder approaching large foundation model performance, 2026. URL <https://arxiv.org/abs/2608.17872>.
 - [9] Jonah Kömen, Edwin D. de Jong, Julius Hense, Hannah Marienwald, Jonas Dippel, Philip Naumann, Eric Marcus, Lukas Ruff, Maximilian Alber, Jonas Teuwen, Frederick Klauschen, and Klaus-Robert Müller. Towards robust foundation models for digital pathology, 2025. URL <https://arxiv.org/abs/2507.17845>.
 - [10] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. ISSN 0047-259X. doi: 10.1016/S0047-259X(03)00096-4. URL [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4).
 - [11] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, 2009. doi: 10.1109/ISBI.2009.5193250. URL <https://doi.org/10.1109/ISBI.2009.5193250>.
 - [12] Valentin Oreiller, Darya Ardan, and Henning Müller. Post-hoc shrinkage of acquisition-induced variability in histopathology vision foundation model embeddings. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2026. URL <https://openreview.net/forum?id=x369KEBa23>.
 - [13] RamonK. Distillpath-is16-hopt0. Hugging Face, 2026. URL <https://huggingface.co/RamonK/DistillPath-IS16-H0pt0>.
 - [14] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. doi: 10.1109/38.946629. URL <https://doi.org/10.1109/38.946629>.
 - [15] Jeongun Ryu, Heon Song, Seungeun Lee, Soo Ick Cho, Jiwon Shin, Kyunghyun Paeng, and Sérgio Pereira. Scorpion: Addressing scanner-induced variability in histopathology, 2025. URL <https://arxiv.org/abs/2507.20907>.
 - [16] Yuki Shigeyasu, Shota Harada, Akihiko Yoshizawa, Kazuhiro Terada, Naoki Nakazima, Mariyo Kurata, Hiroyuki Abe, Tetsuo Ushiku, and Ryoma Bise. Domain generalization of pathological image segmentation by patch-level and wsi-level contrastive learning, 2025. URL <https://arxiv.org/abs/2508.07539>.

- [17] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation, 2015. URL <https://arxiv.org/abs/1511.05547>.
- [18] waticlems. Prost40m. Hugging Face, 2026. URL <https://huggingface.co/waticlems/Prost40M>.
- [19] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL <https://doi.org/10.1080/01621459.1927.10502953>.
- [20] Yufei Zhou, Can Koyuncu, Cheng Lu, Rainer Grobholz, Ian Katz, Anant Madabhushi, and Andrew Janowczyk. Multi-site cross-organ calibrated deep learning (musclD): Automated diagnosis of non-melanoma skin cancer. *Medical Image Analysis*, 84:102702, 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102702. URL <https://doi.org/10.1016/j.media.2022.102702>.

A Estimator and model-selection details

Original region-transfer estimator. The affine ridge in Equation 3 includes an unpenalized intercept. Its source examples use zero paired shift. Inner-fold transfer errors are projected Euclidean residuals $\|U^\top \Delta - \hat{e}\|_2$, averaged over regions within slide. For each permitted non-source scanner q , write the slide-mean error as $\bar{E}_q$ with sample standard deviation s_q over n_q slides. The training selection score is

$$\max_q (\bar{E}_q + 1.6448536 s_q / \sqrt{n_q}),$$

with ties resolved by mean scanner error, worst-scanner material-worsening fraction, and smaller blend weight. Standardization and all fitted geometric components are rebuilt in each inner fold.

Anchored estimator. Let G be the number of independent training groups, D the number of permitted domains including source, and J the number of regions per slide. Residual-regression weights give each group/domain equal total mass and sum to GDJ . This fixes the normalization of the ridge penalties. For groups with several biopsies, each biopsy receives equal weight within its group; prototype matching retains all source regions. The source mean and scale, source-referenced basis, prototype bank, mean source-control anchor, pooled shift, and residual map are refitted in every inner split. Exact ties follow a fixed candidate ordering: identity, anchored global-only, then residual candidates in penalty-major order. The global shift is full-dimensional; only the residual is constrained to the acquisition subspace. The basis and source centering therefore differ from the original estimator.

Global comparators. Pooled mean subtracts the mean paired shift over permitted non-source training scanners. Scanner mean uses a target scanner’s own training pairs when available and otherwise returns pooled mean. Affine ridge learns a regularized map from a query embedding to its source counterpart. Subspace shrinkage attenuates coordinates in a paired acquisition basis. Their hyperparameters are selected within training-only grouped folds. Paired source test regions provide evaluation references; all corrections use training-derived parameters and the query scan.

B Complete scanner comparisons

The same-slide increment accounts for 6.3–8.9% of learned transfer’s total distance reduction over raw embeddings. Unrelated same-scanner controls retain the remaining gain, separating the shared acquisition correction from the contribution of the current slide.

Table 4: Additional baseline distances, completing Table 1.

Encoder	Protocol	Pooled mean	Direct controls
Prost40M	Calibrated	14.131	12.215
	Holdout	15.940	13.837
DistillPath	Calibrated	14.985	14.051
	Holdout	15.840	14.770

Table 5: Learned transfer compared with global alternatives. Positive paired distance reductions favor learned transfer. Intervals use 5,000 paired physical-slide bootstrap resamples.

Encoder	Protocol	Versus affine ridge	Versus scanner mean	Versus shrinkage
Prost40M	Calibrated	−1.830 [−2.007, −1.649]	−.142 [−.251, −.035]	1.711 [1.565, 1.865]
	Holdout	−.576 [−.739, −.411]	2.069 [1.979, 2.158]	.930 [.792, 1.082]
DistillPath	Calibrated	.604 [.411, .798]	−.038 [−.091, .018]	−.035 [−.121, .052]
	Holdout	.738 [.553, .922]	1.158 [1.101, 1.219]	.008 [−.064, .079]

Table 6: Same-slide distance advantage over 32 wrong-slide assignments. Permutations are averaged before physical-slide resampling; the gain range describes variation across the 32 assignments.

Encoder	Protocol	Learned, 95% CI	Anchored, 95% CI	Learned gain range
Prost40M	Calibrated	.535 [.427,.641]	.249 [.169,.325]	.472-.591
	Holdout	.303 [.238,.369]	.156 [.099,.213]	.261-.348
DistillPath	Calibrated	.330 [.270,.392]	.031 [.015,.047]	.298-.367
	Holdout	.249 [.201,.299]	.058 [.041,.075]	.220-.277

Table 7: Top-1 retrieval accuracy (%) and material distance worsening (%). The worsening threshold is defined in Section 4.

Encoder	Protocol	Top-1 retrieval accuracy			Distance worsening	
		Raw	Learned	Anchored	Learned	Anchored
Prost40M	Calibrated	93.91	99.74	99.84	1.77	1.15
	Holdout	93.91	99.74	99.58	3.07	2.45
DistillPath	Calibrated	99.79	100.00	99.95	1.77	1.35
	Holdout	99.79	99.95	99.90	2.40	1.35

C Source-anchored refinement

Table 8: Paired anchored-minus-learned changes. Positive target-distance change means less alignment; negative worsening-rate change is favorable. Rate changes are percentage points.

Encoder	Protocol	Target distance, 95% CI	Retrieval change, 95% CI	Worsening change, 95% CI
Prost40M	Calibrated	.560 [.466,.652]	.104 [.000,.260]	-.625 [-1.094, -.260]
	Holdout	.326 [.264,.389]	-.156 [-.521, .156]	-.625 [-1.302, .000]
DistillPath	Calibrated	.749 [.677,.822]	-.052 [-.156, .000]	-.417 [-.833, -.052]
	Holdout	.291 [.241,.344]	-.052 [-.156, .000]	-1.042 [-1.562, -.573]

Every selected residual uses ridge penalty 100 and weight 0.25 or 0.5. The source-displacement means and physical-slide distributions in Figure 3 use the same evaluation units.

D Tissue selection and PAR preparation

External annotations and fitting. The external release contains 37 slides from 32 patients [2]. Benign labels cover whole glands, including lumen; the residual-stroma class is assigned by postprocessing remaining tissue. The selector uses 224-pixel inputs at $0.5 \mu\text{m}/\text{pixel}$ and adapts the native crop to each image’s recorded resolution. Training balances classes and patients within class; evaluation weights each patient’s complete field set to total one. Inner folds choose the inverse regularization strength $C \in \{.01, .1, 1, 10\}$ by minimizing out-of-fold log loss with equal class weight and equal patient weight within class. Scaling and all fitting are restricted to training patients.

All five primary outer folds select $C = 1$. Among 327 out-of-fold benign selections, 301 have benign reference labels, 24 tumor labels, and two residual-stroma labels. Patient-weighted accuracy, counting unassigned fields as errors, is 95.96%. Four additional randomized patient partitions, specified after the primary results, assess sensitivity to patient allocation at the same cutoff. Their benign precision is 90.81–93.05% and recall is 84.62–88.37%. PAR uses the selector fitted under the primary protocol.

Cohort and acquisition. The PAR cohort uses a deterministic patient ordering, excludes the two patients used to develop tissue-compartment extraction, and retains all biopsies and all three scanners for the selected 20 patients. Cohort selection uses no grading outcomes. Reference and dense sampling share the encoder, selector, cutoff, physical scale, and seed, with non-nested sampled locations. Thumbnail tissue fraction must be at least 0.55. Reference preparations are available for all 35 slides and dense preparations for 34; the unavailable dense preparation remains in all assigned-cohort denominators.

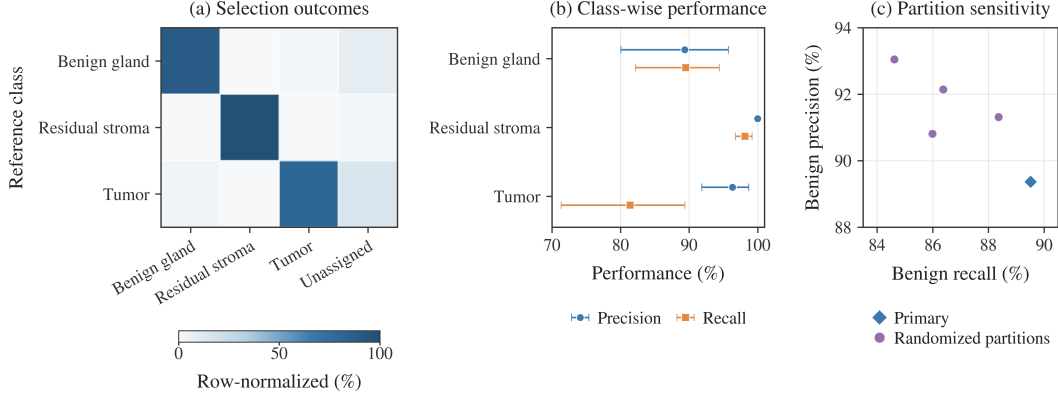


Figure 5: **External tissue-selector validation.** Each patient’s complete field set has total weight one. (a) Row-normalized primary out-of-fold selection outcomes, including unassigned fields at the fixed 0.8 cutoff. (b) Precision and recall with 95% intervals from 2,000 patient bootstrap resamples; recall includes unassigned reference fields. (c) Benign precision and recall for the primary partition and all four post-result randomized partitions of the same 32 patients. The four additional partitions assess sensitivity to patient allocation.

Table 9: Source-feature displacement and model selection. Counts refer to six unique fits per calibrated setting and 24 per holdout setting. Source means average 48 physical slides, each evaluated under one calibrated model or four holdout models.

Encoder	Protocol	Source displacement		Selected models		
		Learned	Anchored	Global-only	Residual	Identity
Prost40M	Calibrated	3.562	1.861	0	6	0
	Holdout	3.306	1.683	0	24	0
DistillPath	Calibrated	2.365	.388	6	0	0
	Holdout	2.248	.555	12	12	0

Registration and joint eligibility. Registration uses 448 μm context, 12 local probes, normalized cross-correlation at least 0.45, maximum local shift 64 μm , common support at least 0.5, and cycle-consistency median/95th-percentile limits of 96/192 μm . Automated registration quality control approves 22 reference trios and 21 dense trios, with 13 rejected under each policy and one dense preparation unavailable. The 34 matched-policy slides have identical decisions. Among approved dense trios, four Grundium–Hamamatsu and eight each Grundium–Leica and Hamamatsu–Leica pairs lack four jointly selected tumor fields. These restrictions can overlap with missing controls.

Table 10: Direction-specific calibration eligibility. Each fold assigns 16 training and four held-out patients. Eligibility requires both scans’ controls, approved registration, and four matched tumor fields. G: Grundium; H: Hamamatsu; L: Leica.

Policy	Direction	Slides /35	Patients /20	Training patients, folds 0–4
Reference	G→H	0	0	0, 0, 0, 0, 0
	G→L	2	2	2, 2, 0, 2, 2
	H→G	0	0	0, 0, 0, 0, 0
	H→L	0	0	0, 0, 0, 0, 0
	L→G	2	2	2, 2, 0, 2, 2
	L→H	0	0	0, 0, 0, 0, 0
Dense	G→H	5	4	3, 3, 2, 4, 4
	G→L	5	5	4, 4, 3, 4, 5
	H→G	5	4	3, 3, 2, 4, 4
	H→L	3	3	2, 2, 2, 3, 3
	L→G	5	5	4, 4, 3, 4, 5
	L→H	3	3	2, 2, 2, 3, 3

Pooling at least one eligible target direction gives reference patient totals G/H/L of 2/0/2 and dense totals of 6/4/5. Dense pooled training counts in folds 0–4 are G: 5,5,3,5,6; H: 3,3,2,4,4; L: 4,4,3,4,5. Multiple biopsies from one patient contribute one patient to these counts.

Table 11: Spatial clustering of selected controls. A nearby center is another selected center of the same class within 112 μm . The fraction measures spatial clustering of sampled fields.

Policy	Class	Selected fields	With nearby center	Fraction
Reference	Benign gland	1,667	459	27.5%
	Residual stroma	25,729	12,213	47.5%
Dense	Benign gland	6,251	4,016	64.2%
	Residual stroma	100,660	86,688	86.1%

E Two-compartment selective estimator

The synthetic study uses a compartment-specific instantiation. For controls $c \in \{b, s\}$, compute a_c by taking the coordinate-wise median over selected control tiles using a compartment-specific source bank. Subtract the coordinate-wise median source-control vector μ_c to obtain $d_c = a_c - \mu_c$. The lesion target e_m is the coordinate-wise median of projected registered lesion shifts. Two zero-intercept ridge maps predict e_m :

$$R_c = \arg \min_R \sum_m \|e_m - R d_{cm}\|_2^2 + \lambda \|R\|_F^2, \quad \hat{e} = \frac{1}{2} \sum_{c \in \{b, s\}} [(1 - \alpha) d_c + \alpha R_c d_c].$$

The basis is constructed by centering each registered region across the permitted scanners, as in Equation 1. This experiment uses ridge penalty $\lambda = 0.5$ and 21 blend weights $\alpha \in \{0, .05, \dots, 1\}$. Patient-grouped out-of-fold selection minimizes non-source mean residual norm, then any-worsening rate, then blend weight. The two compartments, source-median centering, and zero-intercept maps define the instantiation used in these synthetic experiments.

Write $\hat{e}_c = (1 - \alpha) d_c + \alpha R_c d_c$ for each compartment prediction. Selective correction requires sufficient controls; bounded dispersion and prototype-matching radius; supported joint control coordinates under Ledoit–Wolf covariance shrinkage [10]; and bounded disagreement

$$\rho = \frac{\|\hat{e}_b - \hat{e}_s\|_2}{\|\hat{e}_b\|_2 + \|\hat{e}_s\|_2 + \epsilon_\rho}.$$

The denominator stabilizer is 0.1 times the median non-source target-shift norm, with a numerical lower bound. Prediction norms must exceed the maximum source prediction norm, computed out of fold during threshold selection and recomputed after refitting the transfer maps. Candidate thresholds maximize eligible non-source coverage subject to grouped gain and event-rate screens. An outer patient cross-fit repeats the complete procedure from supplied embeddings. If it fails, the fitted model returns identity for every slide; otherwise rejected slides return identity individually. Accepted candidate lesion tiles receive $z_j^* = z_j - U \hat{e}$.

For gate selection, material worsening is an increase in projected residual norm exceeding 2% of the fitting set’s median non-source projected target norm. For binary group indicators H_i , let n be the number of evaluated groups and $\hat{p} = n^{-1} \sum_i H_i$. The one-sided 95% Wilson upper bound is [19]

$$U_W(\hat{p}, n) = \frac{\hat{p} + z^2/(2n) + z \sqrt{\hat{p}(1 - \hat{p})/n + z^2/(4n^2)}}{1 + z^2/n}, \quad z = 1.6448536. \quad (8)$$

The worsening screen uses patients with at least one accepted target scan; the source-correction screen uses all control-count-eligible source patients. The screens require the Wilson upper bounds for patient-any material worsening and patient-any source correction to be at most 5% and 2%, respectively, alongside nonnegative grouped lower alignment gain. Pooled and supplied domain-specific constraints are checked. At zero events, $U_W = z^2/(n + z^2)$ requires at least 52 groups for 5% and 133 for 2%. The separate SCORPION gate evaluates eight reserved validation slides using one region-control summary and a training-source prediction threshold, without the two-compartment disagreement test.

F Controlled synthetic results

The generator uses 48-dimensional embeddings, a rank-six acquisition process, three tissue compartments, and ordinal labels 0–5. Each scenario has eight seeds and 300 patients with a patient-disjoint 70/30 train/test split and 24 tiles per compartment; the 60- and 170-patient scenarios restrict paired calibration while source scaling, prototype banks, and ordinal-head fitting retain all 210 source-training patients. Let δ_m be the coordinate-wise median full-dimensional lesion target-minus-source difference for non-source scan m . Shift RMSE is $\sqrt{M^{-1} \sum_{m=1}^M \|\delta_m - \hat{\delta}_m\|_2^2}$, where $\hat{\delta}_m$ is the applied correction. Worst-target QWK is the minimum over non-source scanners, computed over all simulated labels 0–5. Grade flips average disagreement over all unordered scanner pairs and all test slides, including the source scanner. Paired inference uses 10,000 seed-level percentile bootstrap replicates and exact two-sided sign-flip tests, with Holm adjustment over three comparisons. In the acquisition-sensitive regime, RMSE reduction and QWK gain each have adjusted $p = 0.03125$. Source QWK is unchanged at 0.9976.

Table 12: Complete synthetic regimes. Means include all eight seeds; rejection by the outer screen returns the raw embeddings. Cov.: mean correction coverage. Audit: seeds passing the outer screen.

Regime	Audit	Shift RMSE		Worst QWK		Grade flip		Cov.
		Raw	Corrected	Raw	Corrected	Raw	Corrected	
Acquisition-sensitive	7/8	4.014	1.894	.828	.949	.567	.157	.804
Already robust head	7/8	4.031	1.893	.983	.993	.098	.038	.807
Unseen scanner	7/8	4.449	2.513	.790	.932	.617	.185	.753
Unseen, $1.5\times$ range	7/8	6.673	5.925	.643	.729	.714	.479	.365
Matched tissue response	7/8	3.504	1.630	.865	.961	.555	.150	.813
Strong response gap	6/8	5.371	2.985	.739	.908	.624	.275	.686
60 calibration patients	0/8	4.014	4.014	.828	.828	.567	.567	.000
170 calibration patients	5/8	4.014	2.535	.828	.911	.567	.289	.565

With matched tissue response, direct and learned ungated transfer have similar RMSE, 0.964 and 0.962. With a strong response gap, their RMSEs are 3.766 and 1.662. In the acquisition-sensitive regime, learned transfer improves on direct subtraction by 0.794 (95% CI 0.627–1.002; adjusted $p = 0.0234$). This isolates the benefit of learning tissue-dependent transfer in the controlled generator.

Table 13: Synthetic stress tests in the acquisition-sensitive regime. Material worsening exceeds 2% of the corresponding clean test regime’s median raw non-source shift norm. Accepted worsening is conditional on acceptance and undefined at zero coverage. Means of conditional rates omit seeds with undefined rates.

Stress	Ungated worsening	Coverage	Accepted worsening
One-control conflict	.738	.000	–
Concordant control drift	.964	.037	.990
Morphology out of support	.003	.000	–
Novel acquisition direction	.001	.000	–
Insufficient controls	.000	.000	–
25% selector contamination	.001	.415	.000
Wrong-slide controls	.697	.804	.692
Lesion-response mismatch	.266	.804	.272