Reassessing Spatial Context in Histopathology

Anonymous authors
Paper under double-blind review

Abstract

Spatial context may improve histopathology patch classification, but evaluation on patches from the same whole-slide image (WSI) can overstate generalization to unseen slides. In the SPIDER dataset records examined here, 2,436 WSI identifiers occur in both stored training and test partitions; these WSIs account for 84.59% of stored test targets. We construct a fixed WSI-disjoint split of 408,235 targets from 6,785 WSIs and compare full-view, dual-view, and kernel-regularized contextual transformers with two frozen feature encoders across three seeds. Relative to the full-view baseline, the dual-view system improves three-seed ensemble, four-organ macro-F1 by 2.03 percentage points for UNI2-h (95% CI $[1.10, 3.13]$, $p < 0.001$) and 0.86 points for ALICE (95% CI $[-0.08, 2.00]$, $p = 0.072$). This comparison also changes loss aggregation and normalization order; an exploratory within-architecture view-weight sweep shows no consistent advantage for center-view supervision. Adding a cross-slide conditional-moment penalty does not improve the primary metric for either encoder. The principal finding is a release-specific evaluation concern, while the contribution of center-view supervision to the observed recipe gain remains uncertain.

1 Introduction

A histopathology target patch is rarely an isolated object. Its surrounding epithelium, stroma, tissue boundary, and inflammatory background can disambiguate local morphology. Contextual attention heads exploit this fact by predicting the center of a spatial patch window from all nearby tokens (Vaswani et al., 2017; Dosovitskiy et al., 2021; Nechaev et al., 2025). The same spatial structure that makes context useful also makes evaluation fragile: neighboring targets from one WSI share stain, scanner, preparation, and tissue characteristics. If one WSI contributes patches to both training and test sets, a patch-level metric no longer measures generalization to unseen slides.

We study this problem on SPIDER, a multi-organ patch-classification resource with context windows (Nechaev et al., 2025). The SPIDER paper describes slide-level partitioning. The stored split assignments in the SPIDER records available for this study are not WSI-disjoint: 58,287 of 68,905 test targets come from WSIs also represented in training.

A second issue concerns what "context helps" should mean. Let the center view provide a coarse information set and the full window a refinement. For Bayes posteriors, the refined prediction may legitimately differ from the coarse prediction, while its conditional expectation given the coarse view must return the coarse posterior. This is a coherence condition, not pointwise invariance. It explains why forcing the two predictions to agree can suppress useful context. It also motivates a kernel conditional-moment penalty, adapted here to the clustered WSI sampling design.

We compare three systems on a fixed 70/15/15 WSI split: a full-view transformer, a shared transformer supervised on center and full views, and the shared model with a cross-slide moment penalty. Across two frozen feature representations and three training seeds, we evaluate 64,921 test targets and estimate uncertainty by resampling WSIs. The dual-view training recipe improves on the full-view baseline for UNI2-h, but the systems differ in normalization order and loss aggregation as well as view supervision. The moment penalty

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.

yields no improvement in the primary metric. Context corrects some center-view errors while introducing others.

Our contributions are:

1. a probabilistic account of contextual refinement that distinguishes conditional coherence from pointwise agreement;
2. a WSI-balanced cross-slide estimator with an exact finite-population sampling guarantee; and
3. an evaluation across two frozen representations that distinguishes the observed training-recipe effect from an unresolved mechanism and reports the negative regularization result.

The theoretical formulation draws on prior work on conditional coherence and kernel moments (Gögl et al., 2026; Muandet et al., 2020; Feng et al., 2019). Our contributions concern their application to the examined SPIDER records and the resulting evaluation.

2 Related Work

Spatial context in computational pathology. Patch encoders capture local morphology, whereas WSI models aggregate information over much larger tissue regions (Campanella et al., 2019; Lu et al., 2021). SPIDER occupies an intermediate scale by attaching a labeled target to a local spatial neighborhood (Nechaev et al., 2025). Context consistency has also been studied for semi-supervised histology segmentation under overlapping crops (Bashir et al., 2024). Our setting differs: labels are available for the target, context is observed rather than generated, and the central concern is WSI-clustered evaluation.

Frozen pathology representations. Foundation encoders provide transferable patch features for pathology tasks (Chen et al., 2024; Vorontsov et al., 2024). We evaluate frozen ALICE and UNI2-h patch representations (Li et al., 2026; Mahmood Lab, 2025). Their embeddings differ, but both represent the same SPIDER targets; they do not constitute independent cohorts.

Coherence under nested information. When one view refines another, Bayes predictions obey the tower property. Martingale-consistent learning makes this principle explicit and contrasts it with invariance objectives that suppress all prediction changes (Gögl et al., 2026). Our center/full construction uses the same distinction, but the regularizer operates on observed pathology neighborhoods and excludes products from the same WSI.

Kernel conditional moments and grouped learning. RKHS embeddings turn conditional moment restrictions into kernel-weighted unconditional moments (Muandet et al., 2020); related cross-product losses avoid direct double-sampling requirements in reinforcement learning (Feng et al., 2019). Kernel two-sample statistics likewise illustrate how U-statistics estimate population discrepancies (Gretton et al., 2012). Our estimator is not HSIC and does not test conditional independence. It estimates a finite-population update product between distinct WSIs under the hierarchical sampling design. The supervised objective also gives equal weight to WSIs and organs, related in spirit to group-aware optimization (Sagawa et al., 2020).

3 Problem Setup and Method

3.1 Nested views and the value of context

For target i, let $T_i = x_{i,12} \in \mathbb{R}^d$ be the frozen embedding of the center tile and let $N_i = (x_{i,0}, \ldots, x_{i,24})$ be the complete 5×5 window. The one-hot class vector is $Y_i \in \{0,1\}^{60}$ and the organ is $O_i \in \mathcal{O}$. Only labels valid for organ O_i participate in its softmax. Define the nested information sets

$$\mathcal{F}_1 = \sigma(T, O), \qquad \mathcal{F}_5 = \sigma(N, O), \qquad \mathcal{F}_1 \subseteq \mathcal{F}_5, \tag{1}$$

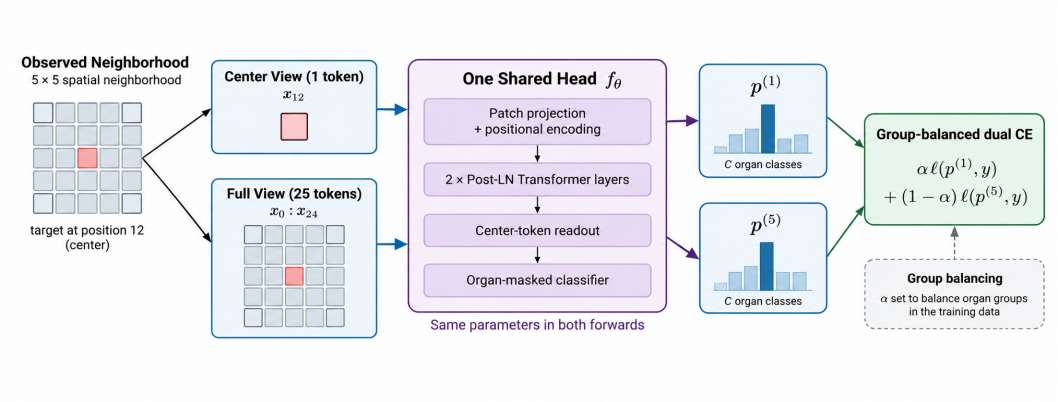

Figure 1: The shared transformer predicts from the center alone and from the full neighborhood, with common parameters and target-, WSI-, and organ-level loss averaging. The optional moment term pairs updates from distinct WSIs; inference uses the full neighborhood.

and Bayes posteriors $p_s^\star = \mathbb{E}[Y \mid \mathcal{F}_s]$ for $s \in \{1, 5\}$.

Proposition 1 (Coherent contextual update). Let $d^\star = p_5^\star - p_1^\star$. Then

$$\mathbb{E}[p_5^\star \mid \mathcal{F}_1] = p_1^\star, \qquad \mathbb{E}[d^\star \mid \mathcal{F}_1] = 0. \tag{2}$$

For multiclass Brier risk $\mathcal{R}(p) = \mathbb{E}\|Y - p\|_2^2$,

$$\mathcal{R}(p_1^\star) - \mathcal{R}(p_5^\star) = \mathbb{E}\|p_5^\star - p_1^\star\|_2^2 \geq 0. \tag{3}$$

The first identity is the tower property. For Equation 3, expand $Y - p_1^\star = (Y - p_5^\star) + (p_5^\star - p_1^\star)$. The cross term vanishes because $p_5^\star = \mathbb{E}[Y \mid \mathcal{F}_5]$ and $p_5^\star - p_1^\star$ is $\mathcal{F}_5$-measurable. Thus useful context generally requires a nonzero update. A pointwise penalty on $\|p^{(5)} - p^{(1)}\|^2$ can oppose the exact quantity by which refined Bayes prediction reduces Brier risk. Conditional mean-zero is the appropriate ideal, not equality on every target.

3.2 One shared contextual head

A shared head f_θ contains a linear projection to 256 dimensions, learned positions, two post-layer-normalized Transformer encoder layers with four heads and feed-forward width 1,024, a final layer normalization, and one 60-logit classifier. Both passes read the encoded token at position 12:

$$p_i^{(1)} = \mathrm{softmax}_{\mathcal{C}_{O_i}} f_\theta([T_i], [12]), \tag{4}$$

$$p_i^{(5)} = \mathrm{softmax}_{\mathcal{C}_{O_i}} f_\theta(N_i, [0{:}24]). \tag{5}$$

The two forwards share every parameter. Only the full-view pass is used at test time.

Training batches sample G_o WSIs in organ o and $n_{o,g}$ targets from WSI g. We average targets within WSI, WSIs within organ, then organs:

$$\mathcal{L}_{\mathrm{sup}} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \frac{1}{G_o} \sum_{g=1}^{G_o} \frac{1}{n_{o,g}} \sum_{i \in (o,g)} \left[\alpha \mathcal{L}_{\mathrm{CE}}(p_i^{(1)}, Y_i) + (1-\alpha) \mathcal{L}_{\mathrm{CE}}(p_i^{(5)}, Y_i) \right]. \tag{6}$$

Confirmatory dual-view runs use $\alpha = 0.5$. Figure 1 summarizes the computation.

3.3 A cross-slide conditional-moment proxy

Let $d_\theta(T, N, O) = p_\theta^{(5)} - p_\theta^{(1)}$. For organ o, let $\phi_o(T)$ be the feature map of an RBF kernel on L2-normalized center features. A practical proxy for Proposition 1 is

$$\mathcal{J}_o(\theta) = \|\mathbb{E}\left[d_\theta \, \phi_o(T) \mid O = o\right]\|^2_{\mathcal{H}_o \otimes \mathbb{R}^{60}}. \tag{7}$$

With an independent copy (T', d'_θ), Equation 7 expands to

$$\mathbb{E}\left[k_o(T, T')\langle d_\theta, d'_\theta\rangle \mid O = O' = o\right]. \tag{8}$$

The restriction tests mean updates against functions of the frozen center representation and is only a proxy for the ideal conditional expectation given $\mathcal{F}_1$. A small value does not establish calibrated posteriors or conditional independence.

We use a fixed three-scale kernel

$$k_o(t, t') = \frac{1}{3} \sum_{r\in\{1/4,1,4\}} \exp\left(-\frac{\|t - t'\|_2^2}{2r m_o}\right), \tag{9}$$

where m_o is a training-only within-organ median squared distance. To avoid treating patches within a WSI as independent, we form products only between targets from different WSIs:

$$\widehat{\mathcal{J}}_o = \frac{1}{G_o(G_o - 1)} \sum_{g=1}^{G_o} \sum_{\substack{h=1 \\ h\neq g}}^{G_o} \frac{1}{n_{o,g} n_{o,h}} \sum_{i\in(o,g)} \sum_{j\in(o,h)} k_o(T_i, T_j)\langle d_i, d_j\rangle. \tag{10}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \frac{\lambda}{|\mathcal{O}|} \sum_{o\in\mathcal{O}} \widehat{\mathcal{J}}_o. \tag{11}$$

We evaluate Equation 10 in FP32 and retain negative finite-batch values.

Proposition 2 (Sampling target). Suppose the sampler selects G_o WSI identifiers exchangeably without replacement and then samples targets uniformly within each selected WSI. Equation 10 is unbiased for the finite-population mean kernel-weighted update product over ordered pairs of distinct WSIs under that sampling design.

Conditioned on ordered WSI pair (g, h), the inner average is unbiased for its finite WSI-pair target mean. Exchangeability makes the average over the $G_o(G_o - 1)$ selected ordered pairs unbiased for the corresponding mean over all distinct WSI pairs. This sampling result concerns the distinct-WSI finite-population target. It does not make Equation 10 an unbiased estimator of Equation 7 under arbitrary target-level sampling.

4 Experimental Protocol

4.1 Data, provenance, and locked split

The SPIDER records analyzed here comprise 408,235 labeled target patches from 6,785 organ-prefixed WSI identifiers: 159,854 skin, 92,892 breast, 77,182 colorectal, and 78,307 thoracic targets. There are 60 pooled class names and 69 organ–class combinations. Available metadata include the WSI identifier, image name, spatial neighborhood, class label, and stored training or test assignment. Patient, institution, scanner, and stain-batch identifiers are unavailable, so WSI is the most stringent grouping unit that these records permit us to verify.

Figure 2 illustrates center tiles and labels across the four organs. These examples were selected to show appearance and label variation, independently of model predictions.

In the stored train/test assignment, 2,436 WSI identifiers occur in both subsets. They account for 58,287 test targets, or 84.59% of the stored test partition. We therefore construct a new stratified 70/15/15 partition of WSI identifiers using seed 42 and use it for all principal comparisons. Training, validation, and test comprise 4,749/1,018/1,018 WSIs and 280,030/63,284/64,921 targets, respectively. No WSI identifier, exact center row, or complete spatial window is shared across the new subsets. Figure 3 and Appendix A report counts by organ.

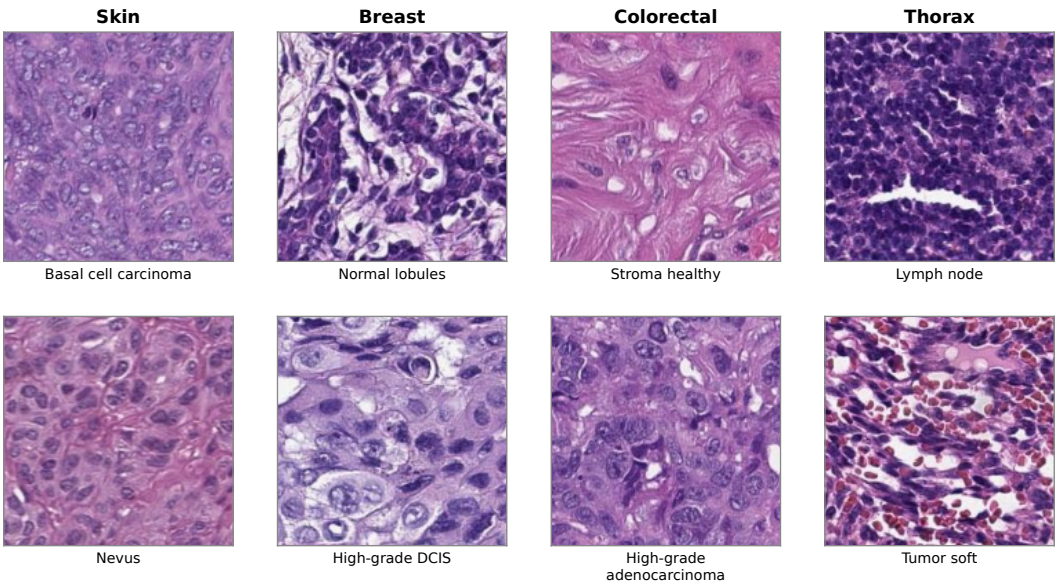

Figure 2: Examples of labeled center tiles from the four organs in the examined SPIDER records. The displayed labels are dataset annotations.

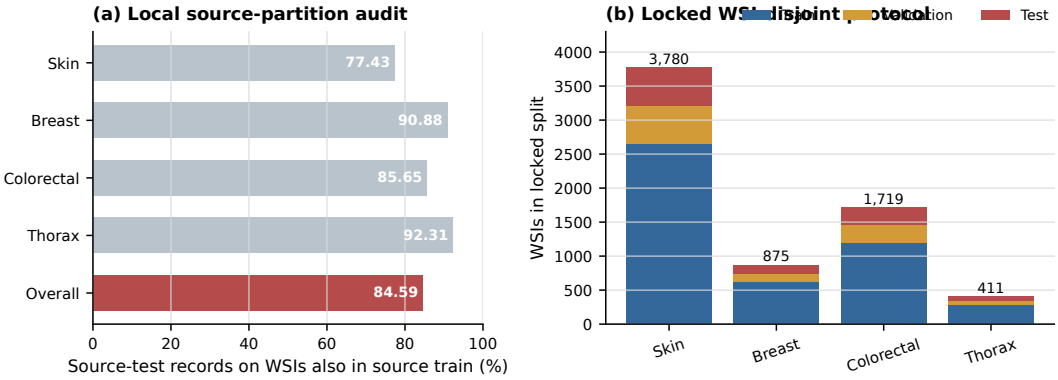

Figure 3: Partition audit for the SPIDER records examined here. (a) Percentage of targets in the stored test subset whose WSI identifier also occurs in training. (b) WSI composition of the new disjoint split. The two test subsets contain different targets.

4.2 Systems, training, and controls

We evaluate three systems. Full-view CE uses a two-layer pre-LN transformer trained on the 5×5 neighborhood with sample-averaged cross-entropy. Dual-view CE uses the post-LN shared head and the WSI- and organ-averaged loss in Equation 6, with $\alpha = 0.5$. Dual-view+Moment adds Equation 10, with $\lambda = 1$ for ALICE and $\lambda = 10$ for UNI2-h. All three systems have a 256-dimensional hidden representation, four attention heads, a feed-forward width of 1,024, no dropout, and the same batch construction and optimization budget. The first two systems also differ in normalization order and loss aggregation. Their comparison therefore estimates the performance difference between training recipes, not the isolated effect of center-view supervision.

A pointwise ALICE control adds $\lambda \|p^{(5)} - p^{(1)}\|_2^2$. An exploratory single-seed sweep varies $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ while keeping the post-LN shared architecture fixed. It assesses sensitivity to view weighting under a shorter training protocol.

Table 1: Test performance across three seeds (mean ± standard deviation). ECE, Brier score, and NLL are minimized. F1-4 is defined in Equation 12; F1-60 pools class names across organs.

Encoder	System	F1-4	F1-60	Acc. (%)	ECE (%)	Brier	NLL
ALICE	Full-view CE	0.8810 ± .0073	0.8597 ± .0080	87.16 ± .76	2.47 ± .91	0.1843 ± .0120	0.3716 ± .0244
ALICE	Dual-view CE	**0.8958 ± .0035**	**0.8754 ± .0041**	88.09 ± .41	1.68 ± .18	0.1668 ± .0057	0.3591 ± .0152
ALICE	Dual-view+Moment	0.8956 ± .0037	0.8750 ± .0044	88.09 ± .43	1.66 ± .19	0.1667 ± .0058	0.3580 ± .0145
UNI2-h	Full-view CE	0.8547 ± .0037	0.8267 ± .0030	84.34 ± .72	2.48 ± .60	0.2206 ± .0071	0.4682 ± .0097
UNI2-h	Dual-view CE	**0.8875 ± .0060**	**0.8672 ± .0075**	87.04 ± .74	1.37 ± .57	0.1814 ± .0121	0.3729 ± .0291
UNI2-h	Dual-view+Moment	0.8845 ± .0088	0.8641 ± .0096	87.07 ± .71	1.22 ± .33	0.1809 ± .0104	0.3730 ± .0317

Each confirmatory condition uses seeds 42, 43, and 44, 500 AdamW steps, learning rate 3×10^{-4}, weight decay 10^{-2}, and gradient clipping at 1. Regularization is zero through step 50, ramps linearly through step 100, then remains fixed. Each batch samples 64 WSIs from each of four organs and two targets per WSI (512 targets in total). We select the highest validation four-organ macro-F1 checkpoint and evaluate the locked test once.

4.3 Metrics and uncertainty

The primary metric averages class macro-F1 within organ and then averages organs:

$$M = \frac{1}{4} \sum_{o \in \mathcal{O}} \frac{1}{|\mathcal{C}_o|} \sum_{c \in \mathcal{C}_o} F1_{o,c}. \tag{12}$$

We also report pooled 60-class macro-F1, accuracy, 15-bin ECE (Guo et al., 2017), multiclass Brier score (Brier, 1950), and negative log-likelihood. For primary contrasts, we average probabilities across the three seeds, resample test WSIs with replacement within organ, and evaluate both models on the same 1,000 bootstrap draws. Percentile intervals quantify test-WSI sampling variation; they do not include training-seed uncertainty.

5 Results

5.1 Overlap in the stored partition

In the examined records, 2,436 WSI identifiers appear in both stored training and test subsets. A prior full-view evaluation on that stored partition yielded 91.80% accuracy with ALICE and 91.42% with UNI2-h. The corresponding mean accuracies across three seeds on the disjoint partition are 87.16% and 84.34%. These differences of 4.64 and 7.08 percentage points compare different training and test compositions; they do not estimate the amount of inflation caused by overlap alone. The overlap itself prevents the stored split from measuring generalization to unseen WSIs.

5.2 Comparison of full-view and dual-view systems

Table 1 reports mean ± standard deviation over training seeds. Dual-view CE exceeds Full-view CE in mean primary F1 by 1.48 points for ALICE and 3.28 points for UNI2-h. The probability-ensemble contrasts differ from the mean per-seed contrasts because F1 is computed after averaging seed-level probabilities. Figure 4 shows both summaries.

The paired WSI bootstrap gives Dual-view CE minus Full-view CE of +0.0086 for ALICE (95% CI $[-0.0008, 0.0200]$, $p = 0.072$) and +0.0203 for UNI2-h (95% CI $[0.0110, 0.0313]$, $p < 0.001$). These differences compare complete training recipes: center-view supervision, loss aggregation, and normalization order change together.

The within-architecture α sweep offers an exploratory check of the view-weighting effect. For ALICE, primary F1 changes from 0.8781 at $\alpha = 0$ to 0.8689 at $\alpha = 0.5$; for UNI2-h it changes from 0.8486 to 0.8347. The best nonzero setting is 0.8747 at $\alpha = 0.25$ for ALICE and 0.8541 at $\alpha = 0.75$ for UNI2-h. The sweep uses one seed and a shorter training budget, and its test-set comparisons are exploratory. It does not identify center supervision as the source of the recipe-level difference.

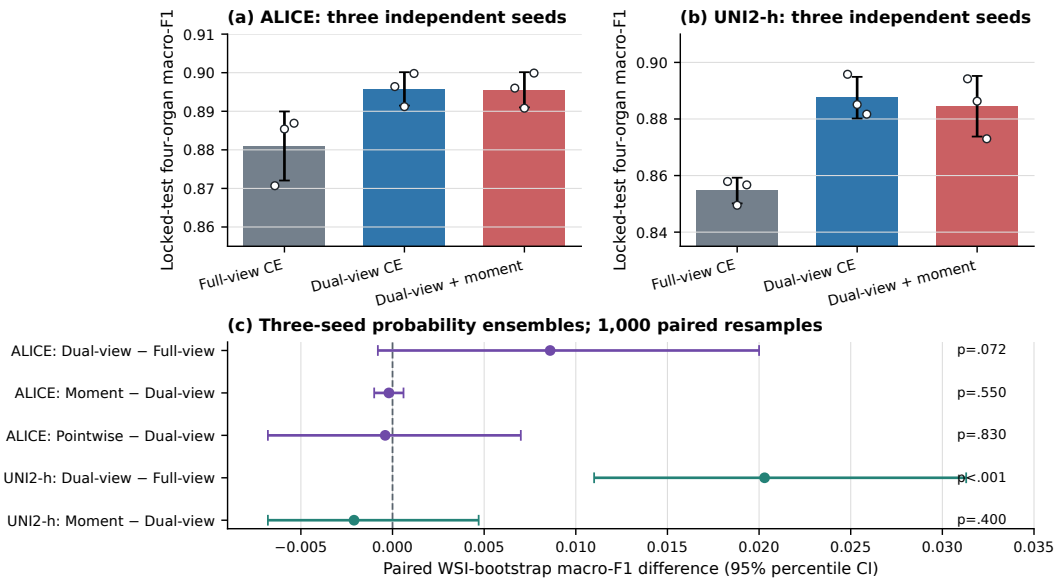

Figure 4: Test macro-F1 for individual seeds and their mean (top), with paired, WSI-bootstrap contrasts between three-seed probability ensembles (bottom). Intervals reflect variation from resampling test WSIs; they exclude variation in model training.

5.3 The cross-slide penalty does not improve macro-F1

The ensemble difference Dual-view+Moment minus Dual-view CE is -0.0002 for ALICE (95% CI $[-0.0010, 0.0006]$, $p = 0.55$) and -0.0021 for UNI2-h (95% CI $[-0.0068, 0.0047]$, $p = 0.40$). Neither comparison provides evidence of an improvement in the primary metric from the moment term. The ALICE pointwise-control contrast is also near zero: -0.0004 (95% CI $[-0.0068, 0.0070]$, $p = 0.83$). The pointwise penalty likewise shows no detectable ensemble macro-F1 benefit in this ALICE comparison.

Mean ECE decreases from Full-view CE to Dual-view CE for both encoders. Adding the moment term changes mean ECE from 1.68% to 1.66% for ALICE and from 1.37% to 1.22% for UNI2-h. We did not compute uncertainty intervals for calibration, so these small ECE differences are descriptive.

5.4 Context changes errors in both directions

On seed 42, every system has nonzero rescue and damage counts (Figure 5). For Dual-view CE, context rescues 1.31% and damages 0.79% of ALICE targets; the corresponding UNI2-h rates are 3.13% and 1.24%. For the moment system, damage is 0.85% and 1.52%. The identity Δaccuracy = rescue rate − damage rate holds up to rounding. The ALICE seed-42 comparison also shows lower F1 for the moment-regularized system than for the full-view baseline on several breast and skin labels (Appendix F). These analyses show that changes in prediction quality vary across targets and classes.

6 Discussion

Evaluation and model comparison. The observed overlap in the stored partition requires a WSI-disjoint split to assess performance on unseen slides. On that split, the dual-view, WSI- and organ-balanced, post-LN system outperforms the full-view, sample-averaged, pre-LN system for UNI2-h, with a positive paired WSI-bootstrap interval. The ALICE ensemble contrast is smaller and its interval includes zero.

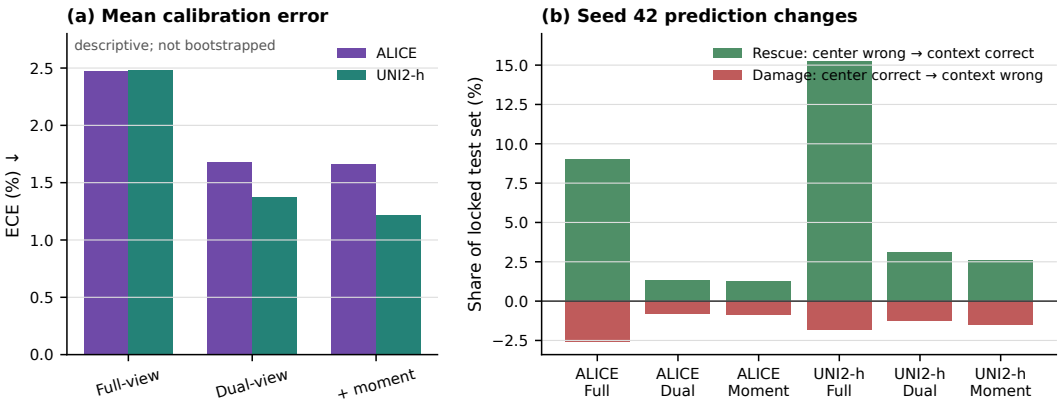

Figure 5: Descriptive behavior. (a) Mean 15-bin ECE across three seeds; no calibration bootstrap was run. (b) Seed-42 center-to-context changes on 64,921 locked-test targets. A rescue corrects a center error; damage overturns a correct center prediction.

Mechanistic interpretation. The principal comparison changes center-view supervision, loss aggregation, and normalization order simultaneously. The single-seed α sweep does not consistently favor equal weighting of the two views. Isolating center-view supervision will require matched architecture and aggregation, repeated seeds across the α grid, and evaluation on a test set reserved for that comparison or an appropriately nested selection protocol.

Conditional coherence. Proposition 1 shows that a refined Bayes posterior can change with context while remaining conditionally coherent with the center-view posterior. The absence of a macro-F1 gain from the moment penalty does not contradict that identity. It limits the empirical support for this particular surrogate, kernel, sampling design, and regularization strength.

Limitations. The study covers one dataset, two frozen encodings of the same records, and three confirmatory seeds. Independent replication on an external dataset was not performed. Patient, site, scanner, and stain-batch keys are absent, so WSI disjointness cannot establish patient- or institution-level generalization. The regularization weights were selected during earlier validation screening. Bootstrap intervals cover test-WSI sampling but not training randomness. ECE depends on binning and lacks paired uncertainty. The example images are illustrative. Patch-category classification does not establish diagnostic, prognostic, or treatment utility.

7 Conclusion

In the SPIDER records examined here, overlapping WSI identifiers in the stored partitions preclude an unseen-slide interpretation of the original test assignment. On a WSI-disjoint split, a dual-view, WSI-balanced training recipe improves on the full-view baseline for UNI2-h, although the comparison cannot isolate the effect of center supervision. A cross-slide conditional-moment penalty does not improve four-organ macro-F1 with either frozen representation. Together, these findings support slide-disjoint evaluation and more closely matched ablations before attributing contextual gains to a particular component.

Reproducibility Statement

Records retained for reproducibility include the WSI-disjoint partition, training-set kernel bandwidths, model checkpoints, test-set probabilities, partition-audit results, and figure-generation code. Appendix H summarizes the material associated with each analysis.

Ethics Statement

This retrospective computational study uses de-identified, publicly available pathology patch data from the SPIDER dataset (Nechaev et al., 2025). The models and contextual evaluation protocols explored herein are intended strictly for benchmarking representation generalization and slide-level partitioning integrity. They have not been clinically validated for diagnostic, prognostic, or treatment decisions.

LLM / AI Usage Statement

In accordance with the ICLR Author Guide, large language models (LLMs) were used solely as general-purpose assistive tools for plotting code, literature exploration, proofreading, and LaTeX formatting and citation audit. No scientific hypotheses, mathematical derivations, or empirical findings were generated by LLMs. All empirical experiments, mathematical derivations, and data analyses were executed, audited, and verified by the authors, who take full responsibility for all content.

References

Raja Muhammad Saad Bashir, Talha Qaiser, Shan E. Ahmed Raza, and Nasir M. Rajpoot. Consistency regularisation in varying contexts and feature perturbations for semi-supervised semantic segmentation of histology images. Medical Image Analysis, 91:102997, 2024. doi: 10.1016/j.media.2023.102997.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1):1–3, 1950.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Frederick Klauschen, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25(8):1301–1309, 2019.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3):850–862, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the Bellman equation. In Advances in Neural Information Processing Systems (NeurIPS), volume 32, 2019.

Moritz Gögl, Hanwen Xing, and Christopher Yau. Martingale-consistent self-supervised learning. arXiv preprint arXiv:2605.11846, 2026.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. Journal of Machine Learning Research, 13:723–773, 2012.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, pp. 1321–1330, 2017.

Jiawen Li, Tian Guan, Huijuan Shi, Xitong Ling, Mingxi Fu, Anjia Han, Chao He, and Yonghong He. Alice: Learning a general-purpose pathology foundation model from vision, vision-language, and slide-level experts. arXiv preprint arXiv:2607.09526, 2026.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 5(6):555–570, 2021.

Mahmood Lab. Uni2-h model card. https://huggingface.co/MahmoodLab/UNI2-h, 2025. Accessed September 2026.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), volume 124 of Proceedings of Machine Learning Research, pp. 41–50. PMLR, 2020.

Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Spider: A comprehensive multi-organ supervised pathology dataset and baseline models. arXiv preprint arXiv:2503.02876, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In International Conference on Learning Representations, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, 2017.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, et al. A foundation model for clinical-grade computational pathology and rare cancer detection. Nature Medicine, 30:2924–2935, 2024.