FOCAL: AGENT-GUIDED TISSUE MASKING BY VISUAL RECONCILIATION OF SEGMENTATION DISAGREEMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tissue detection in whole-slide images (WSIs) is a critical preprocessing step for computational pathology pipelines. Classical approaches rely on thresholding and morphological operations, while deep learning can improve accuracy but depends on diverse annotated training data and can still struggle to generalize across organs and stains, particularly on challenging slides. We propose FOCAL (Foundation model and Oracle Classification for Annotation-free tissue Localization), an agent-guided tissue detection pipeline that combines candidate masks with visual reasoning to resolve segmentation disagreements without additional task-specific annotation or detector fine-tuning. FOCAL separates consensus from disputed regions, uses a vision-language agent to adjudicate tissue versus background in the original image context, and reconstructs the final mask while retaining regional decision reasons and artifact labels. We evaluate on four datasets (TWI, HEST, ACROBAT and KPMP) totaling 6,203 WSIs across diverse tissue types, stains and scanning platforms. FOCAL achieves strong overall performance and is especially robust on challenging slides with artifacts, weak staining and fragmented tissue, etc. Beyond mask prediction, we organize artifact-labelled regions into a traceable resource for targeted retrieval of challenging cases and future artifact-detection, quality-control and tissue-segmentation robustness benchmarks after human verification and necessary revision. Code, generated tissue masks, documented artifact annotations, and curated challenge sets will be released upon paper acceptance.

1 INTRODUCTION

Tissue detection identifies the regions of a whole-slide image (WSI) that should be retained for analysis, providing the basis for patch extraction, feature representation and subsequent modeling in many computational pathology pipelines (Lu et al., 2021; Pocock et al., 2022; Dolezal et al., 2024). It therefore determines what downstream models can see: missed tissue is excluded before its potentially informative morphology can be evaluated, whereas falsely retained glass or artifacts consumes computation and can introduce misleading image content. Quality-control studies show that preparation and scanning artifacts can substantially degrade diagnostic model performance, making the consequences of incorrect region selection extend beyond the tissue mask itself (Schömig-Markiefka et al., 2021; Weng et al., 2024). Interactive software enables researchers to inspect and correct image regions (Bankhead et al., 2017), but reviewing every slide becomes increasingly burdensome as collections grow in size and diversity (Janowczyk et al., 2019; Jurgas et al., 2024). Reliable tissue detection across organs, stains and imaging conditions is consequently essential for preserving informative material without making manual correction a bottleneck.

Existing methods address this problem through increasingly expressive image representations. Thresholding and morphology offer efficient tissue selection from color or intensity cues (Otsu, 1979), while texture-based and hybrid methods such as EntropyMasker and Double-Pass incorporate complementary information (Song et al., 2023; Ceachi et al., 2025). Learned detectors such as DIAG and GrandQC-T use training data to capture richer tissue representations (Bandi et al., 2019; Weng et al., 2024), while AtlasPatch adapts a segmentation foundation model to tissue detection (Alagha et al., 2026). However, evaluations of general-purpose segmentation models on medical images reveal considerable variation across tasks and sensitivity to ambiguous prompts (Mazurowski et al., 2023). These advances improve accuracy, yet their decisions remain sensitive to the image evidence and, for learned models, the training distribution, with documented performance deterioration un-

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.

der acquisition shifts and artifacts (Stacke et al., 2021; Howard et al., 2021; Schömig-Markiefka et al., 2021). Weak staining, fragmented tissue, ambiguous boundaries and non-tissue structures that resemble tissue can therefore yield conflicting predictions even from otherwise effective detectors (Song et al., 2023; Weng et al., 2024). High average accuracy can coexist with severe failures on individual slides, and majority voting alone cannot establish whether a disputed region contains tissue or simply reflects a shared detector error. The remaining need is a mechanism that preserves existing detectors' strengths while examining the original image evidence to resolve disagreements and correct unreliable predictions.

Biomedical agents offer a way to organize this additional reasoning through tool selection, intermediate-result inspection and feedback-guided revision (Huang et al., 2025; Xiao et al., 2024; Alber et al., 2026). Within this paradigm, visual agents combine image interpretation with targeted exploration, connecting local observations to broader spatial context and using the resulting evidence to guide subsequent actions (Lu et al., 2024; Chen et al., 2025). This capability is particularly relevant to tissue detection, where a difficult decision may depend on local texture, neighboring tissue and the spatial context of the entire slide. Given conflicting masks, an agent can organize an inspect–judge–correct–review process that examines the original region, determines whether it should be retained, applies a mask operation and revisits the result in its wider context. Existing detectors supply candidate regions, allowing attention to concentrate on disagreements or suspicious outputs while coordinates, visual observations and decisions remain available for inspection. This combination of image evidence, selective tool use and feedback provides a concrete route toward more reliable tissue detection under challenging conditions.

We introduce FOCAL (Foundation model and Oracle Classification for Annotation-free tissue Localization), an agent-guided tissue detection framework that requires no additional task-specific annotation or detector fine-tuning. FOCAL integrates candidate masks on a shared grid, separates consensus from disputed regions and uses local visual adjudication to reconstruct a tissue mask with recorded regional reasons, while supporting subsequent whole-slide review and correction. Its central contribution is to turn detector outputs into inspectable and correctable tissue decisions, with particular emphasis on preserving reliable coverage when staining is weak, tissue is fragmented or artifacts resemble tissue. Across multiple datasets, organs and staining conditions, we assess overall performance, the lower performance tail and challenging visual examples, demonstrating strong aggregate performance alongside improved robustness on difficult slides. Beyond mask generation, FOCAL links artifact labels and regional decisions to slide identities, coordinates and image crops, creating a traceable resource for retrieving challenging cases and examining failure modes. Following human verification and necessary revision, these records can also support targeted benchmarks for artifact detection, quality control and tissue-segmentation robustness. We further contribute a tissue-detection benchmark on ACROBAT: manual tissue annotations for 140 IHC and 67 H&E breast slides, together with reference masks for all 4,212 slides imputed from them, which we will release.

2 METHOD

FOCAL combines candidate tissue masks with an image-guided agent to resolve regional disagreements and support whole-slide correction, without additional task-specific annotation or detector fine-tuning (Figure 1). The implementation and execution details are provided in Appendix C.

Spatial representation. For a WSI $I \in \mathbb{R}^{H\times W\times 3}$, we define a working image I' on a grid Ω by limiting its longest side to L while preserving its aspect ratio. Each detector F_j runs at its native inference scale, and Π_j aligns its binary prediction to this shared grid:

$$\alpha = \min\left(1, \frac{L}{\max(H,W)}\right), \quad I' = \mathrm{Resize}_\alpha(I), \quad M_j = \Pi_j[F_j(I)] \in \{0,1\}^\Omega. \tag{1}$$

Consensus and disagreement. For J candidate masks, the tissue vote fraction $q(x)$ separates pixels unanimously classified as tissue, U, from pixels with conflicting predictions, D:

$$q(x) = \frac{1}{J}\sum_{j=1}^{J} M_j(x), \quad U = \{x : q(x) = 1\}, \quad D = \{x : 0 < q(x) < 1\}. \tag{2}$$

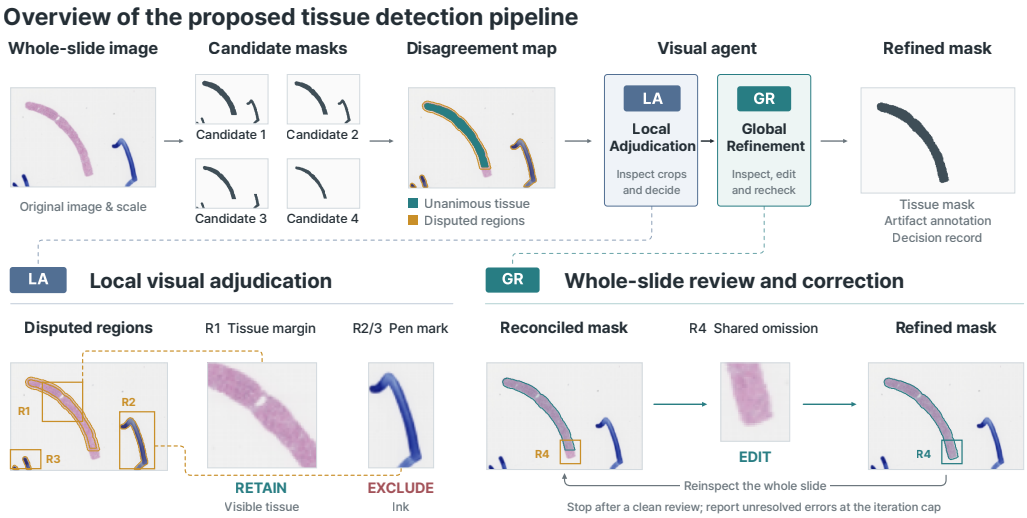

Figure 1: **FOCAL workflow schematic.** Overview of the agent-guided tissue detection pipeline. Candidate masks 1–4 represent generic detector outputs, used to identify disagreement regions for agent review. Local adjudication (LA) resolves candidate disagreements through visual tissue/background decisions. Global refinement (GR) denotes the optional whole-slide review-and-correction stage. The pipeline produces a final tissue mask with regional decision records and artifact descriptors.

Region construction. We group nearby disagreements into regions for review using morphological closing $\mathcal{C}_\kappa$ and connected-component extraction CC. Intersecting each component with D ensures that grouping does not add pixels to the disputed set:

$$\{C_k\}_{k=1}^{K} = \mathrm{CC}(\mathcal{C}_\kappa(D)), \qquad R_k = C_k \cap D, \qquad \bigcup_{k=1}^{K} R_k = D. \tag{3}$$

Region area and the review budget determine which regions receive visual review, indexed by $\mathcal{V}$, and which receive automatic assignment, indexed by $\mathcal{A}$.

Local visual adjudication. For each region selected for review, $\mathcal{P}$ provides the agent with thumbnail context, region overlays and selected image crops. Using this visual evidence and the available review history $\mathcal{H}_k$, the agent $\mathcal{A}_\phi$ returns a tissue/background decision z_k, an optional artifact label ℓ_k and a visual explanation r_k:

$$(z_k, \ell_k, r_k) = \mathcal{A}_\phi(\mathcal{P}(I', R_k), \mathcal{H}_k), \quad z_k \in \{0,1\}, \quad \ell_k \in \mathcal{L} \cup \{\perp\}, \quad k \in \mathcal{V}. \tag{4}$$

Here $\mathcal{L}$ is the artifact vocabulary and $\perp$ indicates that no artifact label is assigned. Each reviewed decision is saved with its region identifier and coordinates b_k for subsequent inspection:

$$\mathcal{D} = \{(\mathrm{id}_k, b_k, z_k, \ell_k, r_k) : k \in \mathcal{V}\}. \tag{5}$$

Mask reconstruction. Reviewed regions follow the agent's decision, while automatically assigned regions use a threshold on their mean candidate vote. The reconstructed mask retains unanimous tissue and adds disputed regions classified as tissue:

$$\bar{q}_k = \frac{1}{|R_k|} \sum_{x \in R_k} q(x), \qquad \tilde{z}_k = \begin{cases} z_k, & k \in \mathcal{V}, \\ \mathbf{1}[\bar{q}_k \geq 1/2], & k \in \mathcal{A}. \end{cases} \tag{6}$$

$$M^{(0)} = U \cup \bigcup_{k:\tilde{z}_k=1} R_k. \tag{7}$$

Whole-slide review and export. In the optional global refinement stage, the agent reviews the reconstructed mask overlaid on the whole-slide image and proposes tissue additions Δ_t^+ and removals Δ_t^-. Accepted edits update the mask, which is then mapped to the original image dimensions by nearest-neighbor resampling Π_0:

$$M^{(t+1)} = (M^{(t)} \setminus \Delta_t^-) \cup \Delta_t^+, \qquad \widehat{M} = \Pi_0[M^{(T)}], \quad T \geq 0. \tag{8}$$

Here $T = 0$ denotes export after local adjudication without global refinement.

3 Evaluation protocol

3.1 Datasets

Our datasets span multiple organs and diverse staining protocols, providing complementary settings for tissue-mask evaluation. **Tissues with Issues (TWI)** (Polejowska et al., 2026) contains 555 WSIs spanning H&E, IHC and special stains, with tissue annotations initialized by a model and refined by two professional annotators. **HEST** (Jaume et al., 2024) contributes 1,255 histology images paired with spatial transcriptomic profiles. Reference masks are rasterized from the released model-generated tissue contours. **KPMP** (Lake et al., 2023) comprises 181 H&E Visium images from the Kidney Precision Medicine Project, with spatial reference masks interpolated from Space Ranger tissue-spot labels within the capture region (10x Genomics, 2026). Cases with reference limitations identified by manual review are listed in Appendix Table 7. **ACROBAT** (Weitz et al., 2023) contains 4,212 breast WSIs, including 1,153 H&E and 3,059 IHC slides. In the absence of ground truth, we derive reference masks by supervised learning with a human in the loop: we manually annotated the tissue of 140 IHC and 67 H&E slides and used them to train DeepLabV3 tissue-segmentation models (Jaume et al., 2024), one for IHC and one for H&E, which impute the tissue masks of the remaining slides in the dataset. These manual annotations form a new benchmark that we will release. FOCAL-generated tissue masks for all 4,212 slides, together with those from the other datasets, will also be released as a high-quality community resource. Additional reference and cohort details are provided in Appendix A.

3.2 Baseline methods

We compare against seven baselines in two groups.

Non-supervised methods. These methods require no labeled training data. *Otsu* (Otsu, 1979) separates tissue from background using a global grayscale threshold that maximizes between-class variance. *CLAM* (Lu et al., 2021) applies a fixed threshold to the median-filtered HSV saturation channel, followed by morphological closing and contour filtering. *Double-Pass* (Ceachi et al., 2025) combines near-gray pixel filtering with two-cluster K-means on RGB values and morphological cleanup. *EntropyMasker* (Song et al., 2023) thresholds local grayscale Shannon entropy using a local minimum of its histogram and retains the largest connected component.

Pretrained tissue segmentation models. We apply released models without additional fine-tuning. *AtlasPatch* (Alagha et al., 2026) adapts SAM2 to multi-cohort WSI thumbnails by fine-tuning its normalization layers, approximately 0.076% of the tiny model's parameters. *GrandQC-T* (Weng et al., 2024) uses a UNet++ decoder with an EfficientNet-B0 encoder for tissue–background segmentation, trained on 208 H&E WSIs and applied at 10 μm/pixel. *DIAG* (Bandi et al., 2019) is a fully convolutional tissue detector developed across multiple centers and tissue–stain combinations for resolution-agnostic segmentation. Implementation settings are provided in Appendix C.

Training-source dependencies. AtlasPatch results are marked NA for TWI groups containing TCGA slides to account for overlap with its reported training sources. Appendix D documents the training-source dependencies of GrandQC-T and FOCAL's candidate models.

4 RESULTS

4.1 OVERALL PERFORMANCE

FOCAL achieves strongest overall performance across heterogeneous tissue types and staining protocols, attaining the highest reported mean Dice, IoU, and pixel accuracy on all four evaluated cohorts (Table 1). On the challenging, multi-stain TWI benchmark, FOCAL achieves mean Dice of 0.955, IoU of 0.926, and pixel accuracy of 0.974, substantially outperforming the strongest reported baseline, DIAG, by 5.8, 8.4, and 3.2 percentage points, respectively. FOCAL also delivers consistent gains on HEST and the KPMP cohort, reaching mean Dice of 0.965 and 0.891, respectively, and surpassing the strongest reported baselines across all three metrics. On all 4,212 ACROBAT slides, FOCAL reaches mean Dice of 0.974 on H&E and 0.943 on IHC, ahead of the strongest baseline, AtlasPatch (0.926 and 0.773; Section 4.2.3). Yet weak staining, artifacts and fragmented tissue can still defeat effective detectors (Song et al., 2023; Weng et al., 2024), and correcting such errors manually is costly (Janowczyk et al., 2019); we therefore examine lower-tail performance and visual examples next.

Table 1: Tissue detection on TWI, HEST, KPMP and ACROBAT.

	TWI			HEST			KPMP			ACROBAT					
										H&E (n = 1,153)			IHC (n = 3,059)		
Method	Dice	IoU	Accuracy	Dice	IoU	Accuracy	Dice	IoU	Accuracy	Dice	IoU	Accuracy	Dice	IoU	Accuracy
FOCAL	**0.955**	**0.926**	**0.974**	**0.965**	**0.935**	**0.982**	**0.891**	**0.807**	**0.957**	**0.974**	**0.951**	**0.976**	**0.943**	**0.907**	**0.966**
EntropyMasker	0.600	0.515	0.861	0.690	0.613	0.873	0.382	0.272	0.794	0.884	0.816	0.907	0.689	0.578	0.838
GrandQC-T	0.717	0.679	0.890	0.890	0.825	0.937	0.879	0.790	0.953	0.901	0.837	0.912	0.065	0.039	0.643
CLAM	0.569	0.510	0.843	0.521	0.459	0.758	0.512	0.440	0.865	0.678	0.557	0.756	0.244	0.172	0.659
AtlasPatch	NA†	NA†	NA†	0.952	0.921	0.975	0.872	0.800	0.932	0.926	0.877	0.936	0.773	0.675	0.871
Double-Pass	0.857	0.779	0.880	0.718	0.606	0.749	0.772	0.656	0.872	0.693	0.562	0.766	0.279	0.180	0.674
DIAG	0.897	0.842	0.942	0.906	0.848	0.940	0.868	0.783	0.948	0.896	0.824	0.908	0.574	0.429	0.785
Otsu	0.595	0.507	0.796	0.862	0.779	0.928	0.852	0.747	0.951	0.436	0.312	0.640	0.122	0.076	0.647

Unweighted slide means. Best displayed values are **bold**, second best underlined (ties at three decimals).
† AtlasPatch on overall TWI is withheld under the TCGA source-overlap reporting rule.
ACROBAT: all 1,153 H&E and 3,059 IHC slides, reported separately; EntropyMasker produced no mask for 3 H&E slides.

4.2 ROBUSTNESS ACROSS CHALLENGING TISSUE AND IMAGING CONDITIONS

4.2.1 TWI: ROBUSTNESS ON CHALLENGING SLIDES

FOCAL demonstrates strong robustness on the challenging, multi-stain TWI benchmark, with a particularly large advantage on IHC and fewer low-scoring slides across the full cohort (Figure 2a). Across all 555 slides, FOCAL achieves mean Dice of 0.955, compared with 0.897 for DIAG, the strongest displayed comparator by overall mean Dice. Its fifth-percentile Dice reaches 0.898 versus 0.506, and the number of slides below Dice 0.8 falls from 89 to 13 (Appendix Tables 4 and 5).

FOCAL's advantage becomes more pronounced on IHC, where it maintains mean Dice of 0.955 compared with 0.807 for DIAG. At the fifth percentile, FOCAL achieves Dice of 0.894, IoU of 0.809 and accuracy of 0.961. These exceed the strongest displayed baseline for each metric, which reaches 0.397, 0.247 and 0.696, respectively. Only 4 of 214 IHC slides fall below Dice 0.8 with FOCAL, compared with 73 for DIAG. On H&E, the average improvement is more modest, with mean Dice of 0.954 versus 0.950 for GrandQC-T, but the fifth-percentile Dice still improves from 0.801 to 0.894 (Appendix Tables 4 and 5).

FOCAL thus combines competitive H&E performance with substantially stronger resilience to the stain variability represented by IHC. TWI brings together diverse staining protocols and challenging imaging conditions, including pen marks, defocus and background artifacts. Its model-initialized annotations were refined by two annotators under a common annotation protocol (Polejowska et al., 2026). This human-refined reference makes TWI a valuable test of performance on difficult tissue–background distinctions, where FOCAL's lower-tail gains are especially clear. Figures 2b and 2c illustrate H&E and IHC examples in which FOCAL preserves fragmented and weakly stained tissue with high accuracy, while the other displayed methods produce substantially less accurate masks.

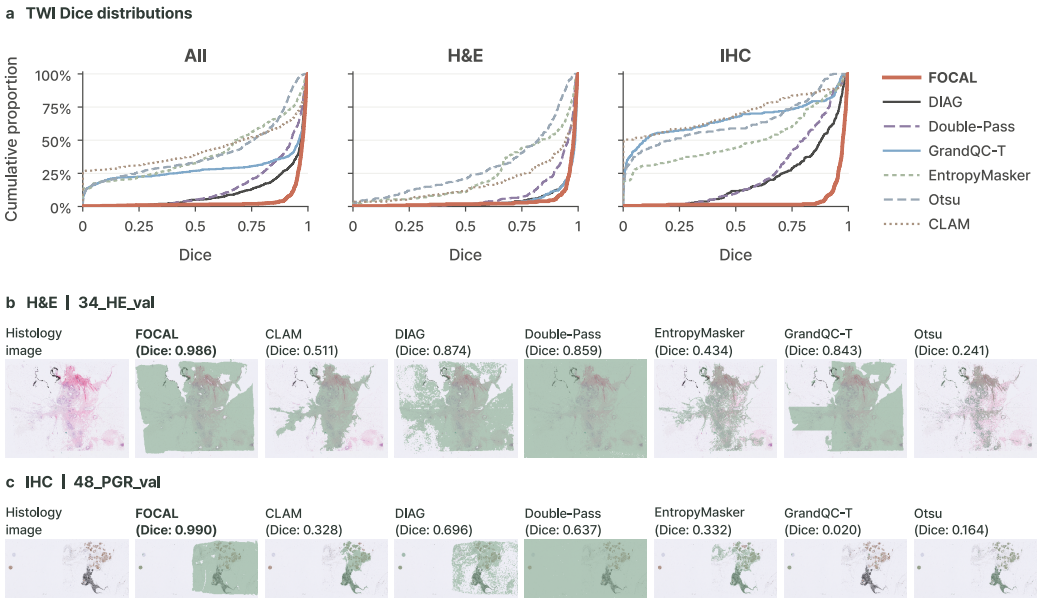

Figure 2: **FOCAL robustness across TWI stain groups and selected challenging slides.** (a) Dice ECDFs for All (555), H&E (276) and IHC (214). Lower curves indicate fewer low-scoring slides. (b,c) H&E `34_HE_val` and IHC `48_PGR_val`, shown in original color with identical sage-green prediction overlays.

FOCAL achieves Dice of 0.986 and 0.990, respectively, compared with only 0.874 and 0.696 for the strongest displayed baseline in each case.

4.2.2 HEST: ROBUSTNESS ON LARGE-SCALE H&E DATA

FOCAL retains strong low-tail performance when we extend the evaluation to 1,255 HEST histology images, broadening the assessment of robustness to a large, multi-organ H&E collection (Jaume et al., 2024). FOCAL achieves mean Dice of 0.965, compared with 0.952 for AtlasPatch, while maintaining a stronger lower tail across Dice, IoU and accuracy (Figure 3a–c).

FOCAL reaches fifth-percentile Dice, IoU and accuracy of 0.901, 0.820 and 0.946, compared with 0.849, 0.737 and 0.928 for AtlasPatch (Figure 3a). At the tenth percentile, FOCAL retains the lead on all three metrics, reaching 0.931, 0.871 and 0.966, respectively (Figure 3b and Appendix Table 6). FOCAL also reduces the number of slides below Dice 0.8 from 49 to 21 and those below Dice 0.5 from 22 to 2 relative to AtlasPatch (Figure 3c). Its advantage therefore extends beyond a modest increase in the average to substantially fewer large disagreements with the reference. FOCAL's consistent lower-tail gains across this larger collection complement the stain-specific robustness observed on TWI.

4.2.3 ACROBAT: LARGE-SCALE CROSS-STAIN ROBUSTNESS

ACROBAT contains 1,153 H&E and 3,059 IHC breast slides spanning ER, PGR, HER2 and KI67 (Weitz et al., 2023). Every slide is evaluated against the reference masks imputed by the DeepLabV3 models trained on our manual annotations (Section 3), and the two stain groups are reported separately (Table 1). On H&E, FOCAL attains mean Dice of 0.974, IoU of 0.951 and accuracy of 0.976, followed by AtlasPatch (Dice 0.926), GrandQC-T (0.901) and DIAG (0.896). On IHC the gap widens: FOCAL attains Dice of 0.943, IoU of 0.907 and accuracy of 0.966, while the strongest baseline, AtlasPatch, reaches Dice of 0.773, followed by EntropyMasker (0.689) and DIAG (0.574). GrandQC-T, trained on H&E, falls from Dice 0.901 on H&E to 0.065 on IHC, and the thresholding baselines also degrade sharply on IHC. FOCAL thus extends the IHC robustness observed on TWI to a collection of over three thousand IHC slides.

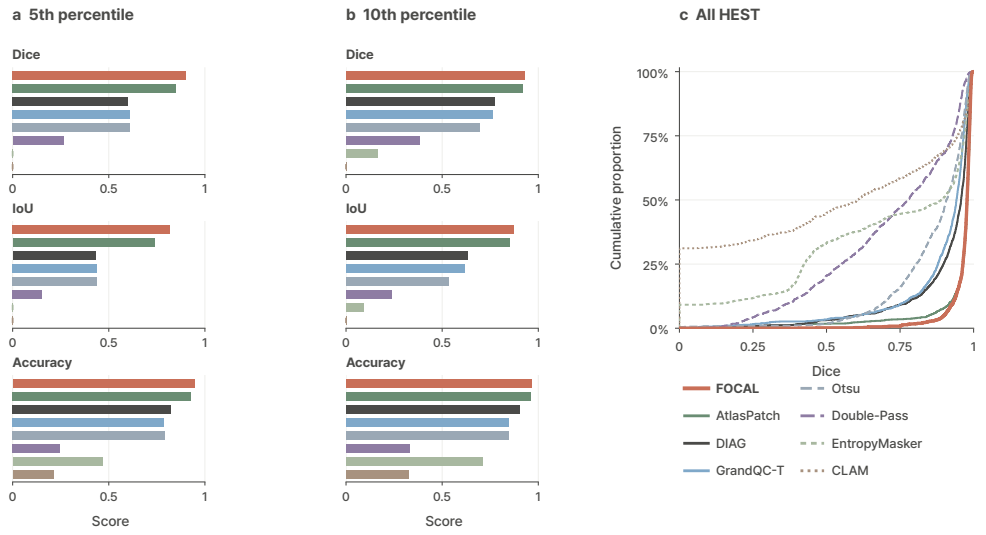

Figure 3: **HEST overall reference agreement and low-tail performance.** (a,b) Fifth and tenth percentiles of Dice, IoU and pixel accuracy, respectively. (c) Dice ECDFs over all 1,255 slides. Lower ECDF curves and higher percentile bars indicate stronger performance. Method colors are consistent across panels.

4.2.4 KPMP: ROBUSTNESS ON CHALLENGING KIDNEY SLIDES

KPMP presents well-documented challenges arising from limited biopsy material (El-Achkar et al., 2021), heterogeneous healthy and injured tissue architecture (Lake et al., 2023), and preparation and imaging risks such as tissue folds, overlap, inadequate image resolution and fiducial misalignment (Eadon et al., 2021), making reliable tissue coverage particularly important. Against this challenging background, FOCAL averages 0.891 Dice and retains 0.769 Dice when averaging the lowest-scoring 10% of slides, exceeding GrandQC-T's corresponding scores of 0.879 and 0.728 (Table 1). FOCAL achieves fifth-percentile Dice and IoU of 0.796 and 0.661, compared with 0.763 and 0.617 for GrandQC-T. Its tenth-percentile Dice and IoU remain higher at 0.817 and 0.691, compared with 0.802 and 0.669. FOCAL has no slides below Dice 0.5 and only 9 below 0.8. Its fifth-percentile accuracy reaches 0.914, while its tenth-percentile accuracy of 0.926 ranks second to AtlasPatch at 0.931 (Table 2).

Table 2: KPMP low-tail performance and empty predictions.

	5th percentile			10th percentile			Slide counts		
Method	Dice	IoU	Accuracy	Dice	IoU	Accuracy	$D < 0.5$	$D < 0.8$	Empty
FOCAL	**0.796**	**0.661**	**0.914**	**0.817**	**0.691**	<u>0.926</u>	**0**	**9**	**0**
GrandQC-T	<u>0.763</u>	<u>0.617</u>	<u>0.910</u>	<u>0.802</u>	<u>0.669</u>	0.918	<u>1</u>	<u>12</u>	**0**
AtlasPatch	0.477	0.313	0.609	0.752	0.602	**0.931**	7	17	**0**
DIAG	0.600	0.428	0.857	0.752	0.603	0.902	2	19	**0**
Otsu	0.743	0.591	**0.914**	0.772	0.629	0.923	**0**	20	**0**
Double-Pass	0.391	0.243	0.492	0.493	0.327	0.727	14	48	**0**
CLAM	0.000	0.000	0.327	0.000	0.000	0.779	50	78	42
EntropyMasker	0.000	0.000	0.178	0.000	0.000	0.655	73	122	<u>25</u>

Best displayed values are **bold**, and second-best values are <u>underlined</u>.

Manual inspection shows that both FOCAL and the spatial reference can preserve tissue while rejecting a cluttered, noisy background, as illustrated by `30-10868`, but also reveals substantial reference omissions where FOCAL retains visibly more complete tissue coverage (Figure 4). In `30-11120`, the reference misses a large, readily visible tissue piece that FOCAL captures, even

though its overall outline is comparatively simple. In `29-10423`, the reference provides incomplete coverage of the fragmented tissue, whereas FOCAL preserves most of the separated tissue components. In the darker, lower-contrast `31-10100`, the reference again omits a major tissue piece that FOCAL successfully retains. Manual review flagged substantial spatial-reference errors in 50 of the 181 slides (Appendix Table 7), which suggests that the tabulated agreement scores may understate FOCAL's ability to preserve real tissue. These examples therefore highlight both a limitation of the reference and practical strengths of FOCAL: by preserving visible tissue that would otherwise be excluded from downstream image analysis, FOCAL could reduce manual correction and support more reliable, streamlined spatial pathology workflows.

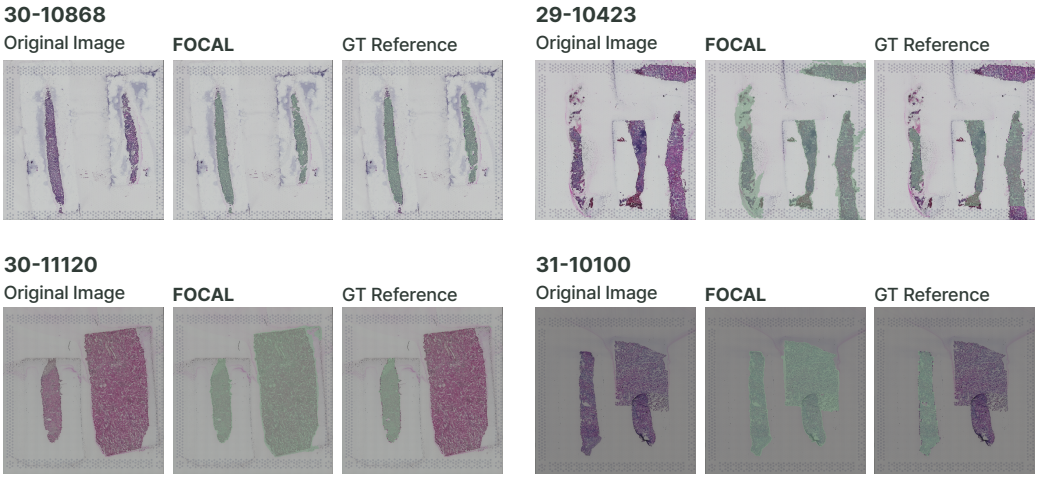

Figure 4: **FOCAL tissue coverage and KPMP ground truth (GT) spatial reference limitations.** Each triplet shows the original image, FOCAL and the spatial reference with identical sage-green mask overlays. `30-10868` provides an agreement example, while the other cases illustrate incomplete reference coverage of visible tissue.

4.3 Artifact annotations across datasets

Beyond evaluating tissue segmentation, FOCAL organizes its regional decisions and interference labels into a reusable resource spanning TWI, HEST, ACROBAT and KPMP, including H&E and IHC staining contexts (Table 3). By linking regional decisions to slide identifiers, coordinates and image crops, these records explain why tissue was retained or excluded and allow users to inspect the underlying visual evidence. Researchers can then retrieve targeted cases, such as IHC slides with pen marks or spatial transcriptomic slides with debris, turning scattered difficult regions into traceable examples for further study. After manual verification and necessary revision, these FOCAL-generated annotations can provide a starting point for independent benchmarks of artifact detection, pathology image quality control and tissue-segmentation robustness.

The selected ACROBAT examples connect these labels to recognizable image content and show FOCAL preserving the main visible tissue despite surrounding pale purplish pen marks that resemble stained tissue, debris and other challenging conditions (Appendix Figure 5), bringing interpretable regional evidence together with robust tissue coverage for subsequent review and reuse.

5 Discussion

FOCAL demonstrates the value of combining existing tissue detectors with visual reasoning to improve reliability where conventional mask generation is most vulnerable. Its strong overall performance is complemented by substantial gains in the lower performance tail, particularly on heterogeneous IHC slides, where weak staining, fragmented tissue and tissue-like artifacts challenge otherwise effective detectors. The human-refined TWI benchmark provides especially clear evidence of this advantage. By turning conflicting predictions into localized visual questions, FOCAL makes

Table 3: Recorded artifact annotations across TWI, HEST, ACROBAT and KPMP.

Artifact Labels	TWI	HEST	ACROBAT	KPMP
Debris / fibre	604 (252)	852 (354)	3,232 (1,615)	331 (79)
Pen / ink	145 (49)	1 (1)	1,117 (528)	4 (3)
Control marker	54 (19)	4,091 (267)	2,252 (676)	92 (14)
Coverslip / mountant	205 (63)	281 (115)	3,332 (1,438)	36 (16)
Scanner edge	49 (27)	278 (68)	1,074 (571)	14 (10)
Tissue fold	24 (13)	14 (7)	66 (45)	0 (0)
Out of focus	51 (33)	12 (8)	685 (434)	7 (7)
Bare glass	123 (56)	943 (325)	11,455 (972)	119 (30)
Boundary halo	1,622 (163)	3,637 (1,031)	178 (92)	18 (10)

Values denote artifact regions, with the number of contributing slides in parentheses.

tissue selection both more robust and easier to inspect, connecting the final mask to the image evidence behind individual decisions.

Manual review also reveals cases in which FOCAL preserves visible tissue more completely than the released reference, including large tissue fragments and low-contrast regions omitted from the reference mask (Figure 4). When such omissions fall within the evaluated region, correctly retained tissue is counted as a false positive, so reference-agreement scores can understate the quality of FOCAL's tissue coverage in these cases. These observations highlight the practical value of preserving informative tissue that would otherwise be discarded before downstream analysis.

Beyond segmentation, FOCAL records difficult regions with their slide identities, coordinates and visual explanations, supporting targeted case retrieval, failure analysis and, after expert verification, artifact-specific benchmarks. We will release the code, generated masks, regional annotations and curated challenge sets upon acceptance.

AI USE STATEMENT

FOCAL uses a large language model, Claude Opus 5 with low thinking effort, as its visual review agent (Section 2); this use is part of the method. Separately, the authors used LLMs to design or provide feedback on research methodology and experiments, to implement methods, to clean and reformat datasets, and to support qualitative and thematic data analysis. LLMs were not used to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for or write proofs, propose or refine hypotheses, assist with translation, or interpret results. The authors reviewed and verified all LLM-assisted work and take full responsibility for the contents of this paper.

REFERENCES

10x Genomics. Space ranger spatial outputs, 2026. URL https://www.10xgenomics.com/support/software/space-ranger/latest/analysis/outputs/spatial-outputs. Accessed September 22, 2026.

Ahmed Alagha, Christopher Leclerc, Yousef Kotp, Omar Metwally, Calvin Moras, Peter Rentopoulos, Ghodsiyeh Rostami, Bich Ngoc Nguyen, Jumanah Baig, Abdelhakim Khellaf, Vincent Quoc-Huy Trinh, Rabeb Mizouni, Hadi Otrok, Jamal Bentahar, and Mahdi S. Hosseini. AtlasPatch: An efficient and scalable tool for whole slide image preprocessing in computational pathology. *arXiv preprint arXiv:2602.03998v1*, 2026. doi: 10.48550/arXiv.2602.03998. URL https://arxiv.org/abs/2602.03998v1.

Samuel Alber, Bowen Chen, Eric Sun, Alina Isakova, Aaron J. Wilk, and James Zou. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. *Nature Methods*, 23:749–759, 2026. doi: 10.1038/s41592-026-03029-6. URL https://www.nature.com/articles/s41592-026-03029-6.

Peter Bandi, Maschenka Balkenhol, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. *PeerJ*, 7:e8242, 2019. doi: 10.7717/peerj.8242. PMID: 31871843.

Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7:16878, 2017. doi: 10.1038/s41598-017-17204-5. URL https://www.nature.com/articles/s41598-017-17204-5.

Bogdan Ceachi, Filip Muresan, Mihai Trascau, and Adina Magda Florea. Efficient tissue detection in whole-slide images using classical and hybrid methods: Benchmark on tcga cancer cohorts. *Cancers*, 17(17):2918, 2025. doi: 10.3390/cancers17172918. PMID: 40941015.

Chengkuan Chen, Luca L. Weishaupt, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Guillaume Jaume, Ming Y. Lu, and Faisal Mahmood. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964v1*, 2025. doi: 10.48550/arXiv.2506.20964. URL https://arxiv.org/abs/2506.20964v1.

James M. Dolezal, Sara Kochanny, Emma Dyer, Siddhi Ramesh, Andrew Srisuwananukorn, Matteo Sacco, Frederick M. Howard, Anran Li, Prajval Mohan, and Alexander T. Pearson. Slideflow: deep learning for digital histopathology with real-time whole-slide visualization. *BMC Bioinformatics*, 25:134, 2024. doi: 10.1186/s12859-024-05758-x. URL https://link.springer.com/article/10.1186/s12859-024-05758-x.

Michael Eadon, Ricardo Melo Ferreira, Ying-Hua Cheng, and Tarek Ashkar. Spatial transcriptomics protocol. protocols.io, 2021. URL https://doi.org/10.17504/protocols.io.bz5hp836. Version 2. Kidney Precision Medicine Project.

Tarek M. El-Achkar, Michael T. Eadon, Rajasree Menon, Blue B. Lake, Tara K. Sigdel, Theodore Alexandrov, Samir Parikh, Guanshi Zhang, Dejan Dobi, Kenneth W. Dunn, et al. A multimodal and integrated approach to interrogate human kidney biopsies with rigor and reproducibility: guidelines from the Kidney Precision Medicine Project. *Physiological Genomics*, 53(1):1–11, 2021. doi: 10.1152/physiolgenomics.00104.2020. URL https://doi.org/10.1152/physiolgenomics.00104.2020.

Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12:4423, 2021. doi: 10.1038/s41467-021-24698-1. URL https://www.nature.com/articles/s41467-021-24698-1.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical AI agent. *bioRxiv preprint*, 2025. doi: 10.1101/2025.05.30.656746. URL https://doi.org/10.1101/2025.05.30.656746.

Andrew Janowczyk, Ren Zuo, Hannah Gilmore, Michael Feldman, and Anant Madabhushi. HistoQC: An open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics*, 3:1–7, 2019. doi: 10.1200/CCI.18.00157. URL https://ascopubs.org/doi/10.1200/CCI.18.00157.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Perez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Datasets and Benchmarks Track, Spotlight.

Artur Jurgas, Marek Wodzinski, Marina D'Amato, Jeroen van der Laak, Manfredo Atzori, and Henning Müller. Improving quality control of whole slide images by explicit artifact augmentation. *Scientific Reports*, 14:17847, 2024. doi: 10.1038/s41598-024-68667-2. URL https://www.nature.com/articles/s41598-024-68667-2.

Blue B. Lake, Rajasree Menon, Seth Winfree, Qiwen Hu, Ricardo Melo Ferreira, Kian Kalhor, Daria Barwinska, Edgar A. Otto, Michael Ferkowicz, et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619:585–594, 2023. doi: 10.1038/s41586-023-05769-3. URL https://www.nature.com/articles/s41586-023-05769-3.

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. doi: 10.1038/s41551-020-00682-w. PMID: 33649564.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative AI copilot for human pathology. *Nature*, 634: 466–473, 2024. doi: 10.1038/s41586-024-07618-3. URL https://www.nature.com/articles/s41586-024-07618-3.

Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918, 2023. doi: 10.1016/j.media.2023.102918. URL https://doi.org/10.1016/j.media.2023.102918.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Johnathan Pocock, Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Srijay Deshpande, Giorgos Hadjigeorghiou, Adam Shephard, Raja Muhammad Saad Bashir, Mohsin Bilal, Wenqi Lu, David Epstein, Fayyaz Minhas, Nasir M. Rajpoot, and Shan E Ahmed Raza. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, 2:120, 2022. doi: 10.1038/s43856-022-00186-5. URL https://www.nature.com/articles/s43856-022-00186-5.

Agata Polejowska, Nefise Uysal, Siem de Jong, Fazael Ayatollahi, Roan Sherif, Muradije Demirel-Andishmand, Frederique Meeuwsen, Dominique van Midden, Linda Krijgsman, Geert Litjens, and Salma Dammak. Tissues with issues benchmark dataset, 2026. URL https://zenodo.org/records/21649203. Version 1.0.0.

Birgid Schömig-Markiefka, Alexey Pryalukhin, Wolfgang Hulla, Andrey Bychkov, Junya Fukuoka, Anant Madabhushi, Viktor Achter, Lech Nieroda, Reinhard Büttner, Alexander Quaas, and Yuri Tolkach. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Modern Pathology*, 34:2098–2108, 2021. doi: 10.1038/s41379-021-00859-x. URL https://www.nature.com/articles/s41379-021-00859-x.

Yipei Song, Francesco Cisternino, Joost M. Mekke, Gert J. de Borst, Dominique P. V. de Kleijn, Gerard Pasterkamp, Aryan Vink, Craig A. Glastonbury, Sander W. van der Laan, and Clint L. Miller. An automatic entropy method to efficiently mask histology whole-slide images. *Scientific Reports*, 13:4321, 2023. doi: 10.1038/s41598-023-29638-1. PMID: 36922520.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, 2021. doi: 10.1109/JBHI.2020.3032060. URL https://doi.org/10.1109/JBHI.2020.3032060.

Philippe Weitz, Masi Valkonen, Leslie Solorzano, Circe Carr, Kimmo Kartasalo, Constance Boissin, Sonja Koivukoski, Aino Kuusela, Dusan Rasic, Yanbo Feng, Sandra Sinius Pouplier, Abhinav Sharma, Kajsa Ledesma Eriksson, Leena Latonen, Anne-Vibeke Laenkholm, Johan Hartman, Pekka Ruusuvuori, and Mattias Rantalainen. A multi-stain breast cancer

histological whole-slide-image data set from routine diagnostics. *Scientific Data*, 10:562, 2023. doi: 10.1038/s41597-023-02422-6. URL https://www.nature.com/articles/s41597-023-02422-6.

Zhilong Weng, Alexander Seper, Alexey Pryalukhin, Fabian Mairinger, Claudia Wickenhauser, Marcus Bauer, Lennert Glamann, Hendrik Blaker, Thomas Lingscheidt, Wolfgang Hulla, Danny Jonigk, Simon Schallenberg, Andrey Bychkov, Junya Fukuoka, Martin Braun, Birgid Schomig-Markiefka, Sebastian Klein, Andreas Thiel, Katarzyna Bozek, George J. Netto, Alexander Quaas, Reinhard Buttner, and Yuri Tolkach. GrandQC: A comprehensive solution to quality control problem in digital pathology. *Nature Communications*, 15:10685, 2024. doi: 10.1038/s41467-024-54769-y. PMID: 39681557.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. CellAgent: An LLM-driven multi-agent framework for automated single-cell data analysis. *arXiv preprint arXiv:2407.09811v1*, 2024. doi: 10.48550/arXiv.2407.09811. URL https://arxiv.org/abs/2407.09811v1.

A DATASETS AND REFERENCE PREPARATION

Tissues with Issues (TWI). We obtained the 555 WSIs and accompanying tissue annotations from the official Zenodo release (Polejowska et al., 2026). The reference annotations were initialized by a model and refined by two annotators. We rasterized the released tissue contours onto the common evaluation grid, subtracting annotated holes and retaining all slides.

HEST. We use 1,255 histology images from HEST together with the accompanying tissue contours generated by the dataset's tissue-detection model (Jaume et al., 2024). For evaluation, these contours were mapped to the evaluation grid and rasterized into reference masks, preserving the tissue boundaries and interior holes specified in the released contours.

ACROBAT. From the official ACROBAT release (Weitz et al., 2023), which includes 4,212 breast WSIs comprising 1,153 H&E and 3,059 IHC images, we generated tissue masks for all slides. Reference masks were obtained by supervised learning with a human in the loop. We manually annotated the tissue of 140 IHC and 67 H&E slides, chosen to cover challenging cases, and used them to train DeepLabV3 tissue-segmentation models, one for IHC and one for H&E. On held-out annotated patients, the trained models reach mean IoU of 0.92 (H&E) and 0.82 (IHC) against the manual masks. The models then impute the tissue masks of all remaining slides. The manual annotations and the imputed reference masks will be released as a benchmark.

KPMP. We use 181 H&E Visium images from KPMP and the accompanying Space Ranger tissue-spot labels (Lake et al., 2023; 10x Genomics, 2026). Spatial reference masks were derived by mapping spots to image coordinates and assigning pixels within the capture region the tissue label of their nearest spot.

B COMPLETE NUMERICAL RESULTS

B.1 TWI OVERALL PERFORMANCE AND ROBUSTNESS

Tables 4 and 5 quantify the seven methods shown in Figure 2, using the same All, H&E and IHC groups. The first reports mean Dice, IoU and pixel accuracy. The second reports fifth and tenth percentiles of Dice, IoU and pixel accuracy, together with low-score counts and empty predictions.

Table 4: TWI overall and stain-stratified segmentation performance.

	All ($n = 555$)			H&E ($n = 276$)			IHC ($n = 214$)		
Method	Dice	IoU	Accuracy	Dice	IoU	Accuracy	Dice	IoU	Accuracy
FOCAL	**0.955**	**0.926**	**0.974**	**0.954**	**0.926**	**0.965**	**0.955**	**0.926**	**0.982**
DIAG	<u>0.897</u>	<u>0.842</u>	<u>0.942</u>	0.949	0.915	0.959	<u>0.807</u>	<u>0.715</u>	<u>0.904</u>
Double-Pass	0.857	0.779	0.880	0.910	0.852	0.911	0.762	0.647	0.811
GrandQC-T	0.717	0.679	0.890	<u>0.950</u>	<u>0.920</u>	<u>0.960</u>	0.339	0.288	0.769
EntropyMasker	0.600	0.515	0.861	0.759	0.667	0.865	0.498	0.418	0.839
Otsu	0.595	0.507	0.796	0.701	0.601	0.811	0.367	0.294	0.723
CLAM	0.569	0.510	0.843	0.821	0.757	0.881	0.297	0.244	0.765

Best values are **bold**, and second-best values are <u>underlined</u>.

Table 5: TWI low-tail performance and empty predictions.

	5th percentile			10th percentile			Slide counts		
Method	Dice	IoU	Accuracy	Dice	IoU	Accuracy	$D < 0.5$	$D < 0.8$	Empty
All ($n = 555$)									
FOCAL	**0.898**	**0.815**	**0.940**	**0.928**	**0.865**	**0.960**	**8**	**13**	1
DIAG	0.506	0.339	0.740	0.691	0.528	0.814	28	89	**0**
Double-Pass	0.504	0.337	0.474	0.630	0.460	0.630	27	142	**0**
GrandQC-T	0.000	0.000	0.514	0.008	0.004	0.643	149	177	4
EntropyMasker	0.000	0.000	0.585	0.000	0.000	0.689	181	334	65
Otsu	0.002	0.001	0.404	0.008	0.004	0.509	184	304	**0**
CLAM	0.000	0.000	0.495	0.000	0.000	0.584	217	306	128
H&E ($n = 276$)									
FOCAL	**0.894**	**0.808**	**0.898**	**0.929**	**0.867**	**0.945**	5	**9**	1
DIAG	0.778	0.636	0.815	0.878	0.783	0.895	**3**	16	**0**
Double-Pass	0.653	0.485	0.639	0.787	0.649	0.755	5	29	**0**
GrandQC-T	0.801	0.669	0.802	0.875	0.778	0.890	5	13	**0**
EntropyMasker	0.142	0.077	0.603	0.438	0.280	0.740	30	128	7
Otsu	0.084	0.044	0.449	0.221	0.124	0.586	58	131	**0**
CLAM	0.233	0.132	0.515	0.427	0.272	0.664	32	78	9
IHC ($n = 214$)									
FOCAL	**0.894**	**0.809**	**0.961**	**0.925**	**0.860**	**0.970**	**3**	**4**	**0**
DIAG	0.396	0.247	0.696	0.471	0.308	0.749	25	73	**0**
Double-Pass	0.397	0.247	0.401	0.503	0.336	0.471	21	109	**0**
GrandQC-T	0.000	0.000	0.441	0.000	0.000	0.494	144	163	4
EntropyMasker	0.000	0.000	0.558	0.000	0.000	0.631	95	145	31
Otsu	0.001	0.000	0.288	0.002	0.001	0.431	126	168	**0**
CLAM	0.000	0.000	0.452	0.000	0.000	0.516	148	181	87

Best values are **bold**, and second-best values are underlined.

B.2 HEST LOW-TAIL REFERENCE AGREEMENT

Overall HEST performance is already reported in Table 1. Table 6 reports fifth and tenth percentiles of Dice, IoU and pixel accuracy across all 1,255 HEST slides. Each metric is summarized independently, retaining empty predictions.

Table 6: HEST low-tail performance and empty predictions.

	5th percentile			10th percentile			Slide counts		
Method	Dice	IoU	Accuracy	Dice	IoU	Accuracy	$D < 0.5$	$D < 0.8$	Empty
FOCAL	**0.901**	**0.820**	**0.946**	**0.931**	**0.871**	**0.966**	**2**	**21**	**0**
AtlasPatch	0.849	0.737	0.928	0.919	0.851	0.961	22	49	**0**
DIAG	0.602	0.430	0.825	0.775	0.632	0.900	40	146	**0**
GrandQC-T	0.610	0.439	0.788	0.762	0.616	0.845	43	158	2
Otsu	0.611	0.439	0.791	0.695	0.533	0.844	22	303	**0**
Double-Pass	0.264	0.152	0.245	0.380	0.234	0.332	257	670	**0**
EntropyMasker	0.000	0.000	0.467	0.166	0.090	0.708	416	570	115
CLAM	0.000	0.000	0.212	0.000	0.000	0.325	564	770	391

Best values are **bold**, and second-best values are underlined.

B.3 KPMP

Overall KPMP performance and lower-tail performance are already reported in Tables 1 and 2, respectively. Manual inspection of the original images, FOCAL overlays and spatial references identified 50 slides with reference limitations. Table 7 lists these cases, complementing the examples of incomplete tissue coverage discussed in the main text.

Table 7: KPMP spatial-reference limitations identified by manual review.

Slide ID	Slide ID	Slide ID	Slide ID	Slide ID
27-10039	27-10041	27-10047	27-10067	27-10088
27-10119	27-10146	27-10155	27-10207	27-10218
27-10228	27-10243	28-12178	28-12265-XY02	28-12626
28-12629	28-12634	28-12635	29-10016	30-10631
30-10929	30-11072	30-11084	30-11095	30-11120
30-11168	31-10040	31-10091	31-10100	31-10169
31-10322	31-10379	32-10333	33-10005	33-10375
34-10331	34-10889	34-11025	34-11042	34-11049
933-10155	933-10372	933-10777	933-10868	934-10002
934-10003	935-10061	935-10099	935-10111	935-10337

C Implementation details

C.1 Image and mask alignment

We preserve the image aspect ratio and limit the longest working dimension to 2,000 pixels. Each detector operates at its supported physical scale, and binary predictions are aligned to the working grid using nearest-neighbor sampling. Prediction and reference masks are evaluated at identical dimensions, with foreground encoded as 255 and background as 0. Table 8 summarizes the baseline configurations.

Table 8: Baseline and FOCAL implementation settings.

Method	Evaluated implementation / scale
Otsu	Grayscale inverse Otsu threshold on the working thumbnail. No additional learned model.
CLAM	HSV saturation median filter 7. Threshold 8. Closing 4. Contour-area settings $a_t = 100$, $a_h = 16$. At most 8 holes. Reference patch size 512.
EntropyMasker	8.36 μm/pixel. Histogram/mask disk radii 5/3. 30 histogram bins. Accepted local minimum in $(1, 4)$. Largest-component rule.
Double-Pass	10 μm/pixel. Near-gray tolerance 15. Morphology kernel 5. Minimum object size 5,000. Two-cluster K-means with fourfold spatial downsampling.
DIAG	Released convolutional detector. 2 μm/pixel. Patch size 1,024. Confidence threshold 0.8. Region/hole diagonal rules 100/250 μm.
AtlasPatch	Released tissue detector and supported slide-loading interface. Candidate thumbnail at its prescribed 1.25$\times$ operating scale.
GrandQC-T	Released binary tissue detector. 10 μm/pixel.
FOCAL	Visual adjudication of candidate disagreement regions on the shared grid by Claude Opus 5 (low thinking effort). No additional detector fine-tuning.

C.2 Disagreement regions

An elliptical closing kernel of 21 working pixels groups nearby disagreements, after which each connected component is intersected with the original disagreement mask. Regions are considered in descending area order for visual review, with a minimum area of 2×10^{-5} of the image and a configurable review budget. Regions below the area threshold or beyond the budget receive the per-region mean-vote assignment defined in Equation 6. When a mixed region is split, its disjoint child regions replace the parent for subsequent decisions and reconstruction.

C.3 Visual review and mask export

The agent is Claude Opus 5 run with low thinking effort. It examines the original thumbnail, region overlays and selected contextual crops, requesting closer views when needed. It judges whether tissue is present and clearly viewable, distinguishes internal cell-free spaces from external background, and records a region identifier, tissue/background decision and short visual explanation. Recorded identifiers are checked against the region map before accepted regions are combined with

the consensus and automatically assigned tissue. Masks are exported to the original image dimensions by nearest-neighbor resampling, while quantitative evaluation uses the common working grid directly.

D TRAINING DATA SOURCES

Table 9 summarizes the reported training sources of the evaluated tissue detectors and their relationship to our evaluation datasets.

Table 9: Training sources and evaluation reporting.

Method	Training sources and reporting
Otsu	No labeled training set. The threshold is estimated from each image (Otsu, 1979).
CLAM	The evaluated HSV-based tissue detector requires no labeled training set. The downstream trained CLAM classifier is not used (Lu et al., 2021).
Double-Pass	No labeled training set. Color filtering and clustering operate on the input image (Ceachi et al., 2025).
EntropyMasker	No labeled training set. Tissue selection uses local entropy and image-specific histogram analysis (Song et al., 2023).
DIAG	Trained on 50 WSIs spanning ten tissue–stain categories from five medical centers (Bandi et al., 2019). The reviewed publication does not identify our evaluation collections as training sources.
AtlasPatch	SAM2 adaptation uses a curated corpus drawn from CHUM, CAMELYON17, PANDA and TCGA (Alagha et al., 2026). Its training data overlap with the TCGA subset of TWI; results for TWI groups containing these slides are therefore marked NA and omitted from the main TWI figure.
GrandQC-T	The tissue detector was trained on 208 H&E WSIs drawn from TCGA and institutional cohorts (Weng et al., 2024). No slide-level overlap was identified between its TCGA training data and the TCGA slides in TWI, so this method is retained in the TWI comparison.
FOCAL	Combines candidate masks with agent-guided visual review and correction, without additional task-specific annotation or detector fine-tuning. Training-source overlap and potential data leakage issues are carefully addressed during processing.

E ACROBAT ARTIFACT EXAMPLES

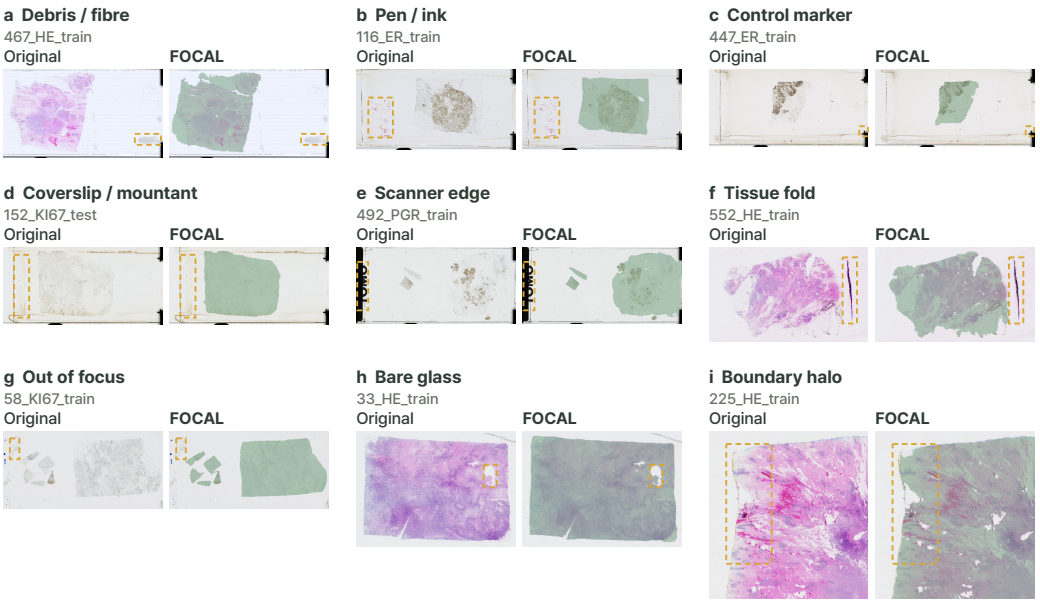

Figure 5: **ACROBAT examples of recorded artifact and background descriptors.** Each pair shows the original image and FOCAL with the same field of view. Yellow dashed boxes indicate the reviewed region.