Vision-Language Models as Zero-Annotation Oracles in Histopathology

Vishal Jain[1 [0009−0001−3447−9211]✉], Giorgio Buzzanca[2 [0009−0009−0340−8822]], Sarah Cechnicka[1 [0009−0008−3449−9379]], Maarten Naesens[3 [0000−0002−5625−0792]], Priyanka Koshy[4 [0000−0002−2313−5122]], Tri Nguyen[5 [0000−0001−6475−0706]], Jesper Kers[2 [0000−0002−2418−5279]], Candice Roufosse[1 [0000−0002−6490−4290]], and Bernhard Kainz[1,6 [0000−0002−7813−5023]]

[1] Imperial College London, London, UK
v.jain24@imperial.ac.uk
[2] Leiden University Medical Center, Leiden, The Netherlands
[3] KU Leuven, Leuven, Belgium
[4] University Hospitals Leuven, Leuven, Belgium
[5] University Medical Center Utrecht, Utrecht, The Netherlands
[6] Dept. AIBE, Friedrich-Alexander University Erlangen-Nürnberg, Germany

Abstract. Foreground segmentation is the critical first step of every computational pathology pipeline, yet existing methods rely on hand-tuned heuristics or supervised models that overfit to narrow stain and scanner distributions, failing silently on specialised stains. We propose a coarse-to-fine approach that recasts foreground segmentation as a visual perception task and leverages general-purpose vision-language models (VLMs) as zero-annotation oracles. Our key insight is that tissue-versus-background discrimination is a natural-image recognition problem, which transfers to histopathology, so VLMs trained on internet-scale corpora generalise where domain-specific models have limitations. We introduce Leica-75, a benchmark of 75 renal transplant whole-slide images spanning three stain families. On Leica-75, our method achieves the highest segmentation quality on out-of-distribution stains (Dice $0.858_{\pm 0.027}$ on Jones, $0.853_{\pm 0.041}$ on EVG) with 3.5× lower cross-stain variance than the strongest baseline and 7× better than the best pathology-specific model, while remaining competitive on in-distribution H&E. Few-shot prompting with automatically curated exemplars (*Auto-context*) rescues hard cases on Stress-32 (n=32), a curated stress-test subset (Dice 0.532 → 0.824 for the 2B model). VLM-based annotation review matches human expert consensus (κ=0.989 for blur detection; mean precision/recall grading accuracy 0.708 vs. human 0.646 for segmentation mask review). The resulting pseudo-labels are used to distil lightweight student models that are as performant as the teacher model while running for a fraction of the cost. Code is available at https://github.com/VishalJ99/vlm-wsi-auto-context; Leica-75 and Stress-32 will be released as part of a larger data release from our centre on Zenodo.

Keywords: Digital pathology · Foreground segmentation · Vision-language models · Zero-shot learning · Quality control

1 Introduction

Every computational pathology pipeline begins by separating tissue from background in a whole-slide image (WSI). At gigapixel scale, downstream methods typically operate on fixed-size patches and aggregate representations via multiple-instance learning (MIL) [16,4,9], creating an extreme class-imbalance setting in which diagnostically relevant regions may occupy only a small fraction of the slide [3]. Including background or artefactual patches among candidate regions can introduce non-diagnostic cues that encourage spurious correlations and degrade robustness [8,28,14]. Reliable foreground segmentation is therefore a foundational prerequisite whose errors propagate to all subsequent analysis.

In practice, tissue preparation and digitisation introduce substantial variation across institutions, scanners, and chemical protocols [8,10,13]. Classical methods based on fixed thresholds or unsupervised clustering (Otsu [20], K-means [17]) are brittle under distribution shift. Rule-based pipelines such as HistoQC [10] require manual recalibration per site, while supervised models such as HEST [11] and GrandQC [26] can overfit to their training distribution. A major contributor is limited stain diversity in widely used pathology tooling and current foundation models [6,24,15,19,7], which are typically trained on H&E-dominated datasets. This induces stain bias that is consequential in clinical workflows employing specialised stains such as Jones methenamine silver, Elastica van Gieson (EVG), and periodic acid-Schiff (PAS) [8,13].

For the purpose of WSI preprocessing, tissue-versus-background discrimination depends primarily on generic visual cues (colour, texture, and structural regularity) rather than histopathological semantics. This suggests that general-purpose vision-language models (VLMs) [1], trained on internet-scale corpora, can provide a strong prior for stain and site-agnostic foreground detection. If such a prior can be leveraged reliably, it offers a path to cross-site deployment without the per-institution tuning that H&E-centric pipelines often require. However, naïvely applying VLM inference at WSI scale is prohibitively expensive. We therefore introduce a coarse-to-fine method that uses VLMs as zero-annotation oracles while aggressively restricting the number of expensive calls, and we use the resulting pseudo-labels to distil lightweight student models for fast, low-cost deployment.

Therefore, our **contributions** are: **(1)** A coarse-to-fine method that leverages general-purpose VLMs as zero-annotation oracles for WSI foreground segmentation, and a distillation pathway that converts VLM pseudo-labels into lightweight student models suitable for deployment. **(2)** The first systematic evaluation of general-purpose VLMs as zero-annotation oracles for stain-agnostic foreground segmentation across three stain families, showing improved robustness on out-of-distribution stains relative to both supervised pathology tools and a strong general-purpose segmenter. **(3)** Evidence that VLMs can provide practical quality control at scale on two tasks: (i) near-perfect binary out-of-focus patch triage, and (ii) precision/recall mask review on a synthetic corruption benchmark, where VLM reviewers match or exceed the strongest human rater.

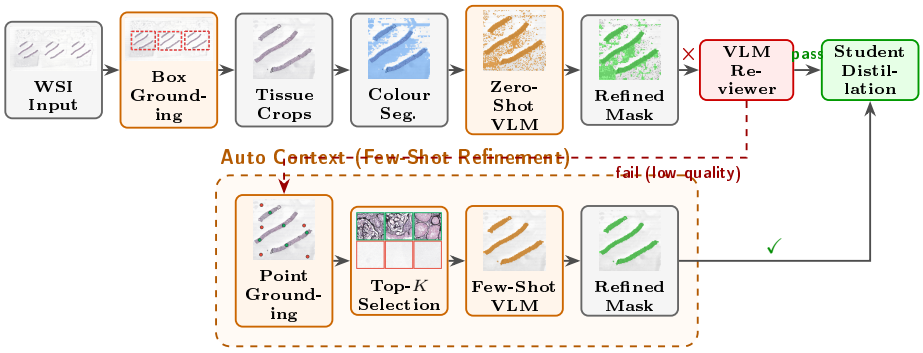

Fig. 1. The primary fast path (top) uses VLM-guided region proposals and zero-shot patch classification to generate initial tissue masks. A VLM reviewer (red) automatically audits quality; failed masks are routed to an *Auto Context* loop (bottom) where slide-specific exemplars are mined to rescue performance via few-shot refinement. Validated pseudo-labels are ultimately distilled into a lightweight student model for deployment.

Related work. Foreground segmentation has progressed from classical thresholding and clustering [20,17] and rule-based quality control (HistoQC [10]) to supervised pipelines embedded in computational pathology toolkits such as TIAToolbox [21] and TRIDENT [29], which incorporate GrandQC [26] and HEST [11]. However, these models are predominantly trained on H&E and inherit stain-specific biases that can degrade performance on specialised stains [13,8,28]. VLMs remain largely unexplored for histopathology preprocessing; the closest work is CORE [18], which pairs Florence-2 [27] with SAM [12] for tissue mask extraction but reverts to a supervised U-Net when staining quality degrades. We adopt the concept-promptable SAM 3 [5] as a strong general-purpose baseline. However, we treat segmentation as a general visual perception task solvable via text prompting of general-purpose VLMs [25,2], extend this to automated exemplar selection for few-shot refinement, and use VLM-generated pseudo-labels to distil lightweight students. We emphasize patch-level tissue labels, aligning the segmentation output with the primary consumer in modern pipelines: MIL-based WSI aggregation [9,4,22], which typically requires patch-level filtering rather than pixel-accurate boundaries.

2 Method

VLM-guided tissue core localisation. Given a WSI $\mathbf{W}$ at native resolution $H_0 \times W_0$, we extract a thumbnail $\mathbf{T} \in \mathbb{R}^{h \times w \times 3}$ at a fixed downsampling factor d (typically d=64, yielding h=H_0/d, w=W_0/d). Rather than relying on handcrafted colour thresholds, we query a VLM with $\mathbf{T}$ and a structured prompt requesting bounding boxes around tissue cores visible in the thumbnail. The VLM returns a set of axis-aligned bounding boxes $\mathcal{B} = \{b_i\}_{i=1}^{B}$, where each

$b_i = (x_i^{\min}, y_i^{\min}, x_i^{\max}, y_i^{\max})$ specifies the top-left and bottom-right corners of a tissue core in thumbnail coordinates. These bounding boxes define the regions of interest for the subsequent colour-based segmentation.

Two-pass colour segmentation. We refine the thumbnail boxes from Stage 1 using K-means in CIELAB space.

Global pass. We cluster all thumbnail pixels into K=2 clusters with assignments $z_{ij} \in \{1, 2\}$ and centroids $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$. We identify the global background centroid as the majority cluster, $k_{\text{bg}} = \arg\max_k \left|\{(i,j) : z_{ij} = k\}\right|, \qquad \boldsymbol{\mu}_{\text{bg}} = \boldsymbol{\mu}_{k_{\text{bg}}}$.

Local pass. For each VLM box $b_i \in \mathcal{B}$ we cluster pixels in the crop $\mathbf{T}_i$ into K=3 clusters with centroids $\{\boldsymbol{\mu}_k^{(i)}\}_{k=1}^3$. The crop background cluster is chosen by nearest-centroid matching,

$$k_{\text{bg}}^{(i)} = \arg\min_k \left\| \boldsymbol{\mu}_k^{(i)} - \boldsymbol{\mu}_{\text{bg}} \right\|_2, \tag{1}$$

and the binary crop mask is $\mathbf{M}^{(i)} = \mathbf{1}[z^{(i)} \neq k_{\text{bg}}^{(i)}]$. Using K=3 locally reduces failures when a dominant artefact would otherwise absorb a cluster under K=2. We combine crop masks via $\mathbf{M}^{(1)}(u,v) = \max_{i \in \{1,\ldots,B\}} \mathbf{M}^{(i)}(u,v)$, and upsample $\mathbf{M}^{(1)}$ to level-0 coordinates to restrict subsequent VLM evaluation.

Zero-shot patch classification. We partition the tissue regions identified by $\mathbf{M}^{(1)}$ into a regular grid of non-overlapping patches $\{x_n\}_{n=1}^N$ of size 512×512 pixels at level-0. Each patch is presented to the VLM with a structured prompt requesting two labels: (i) a binary tissue classification $\hat{y}_n^{(0)} \in \{0, 1\}$, and (ii) a focus-quality assessment $q_n \in \{\textit{sharp}, \textit{mild blur}, \textit{out of focus}\}$. Both labels come from a single VLM call, so the blur map is obtained at no additional cost. Aggregating $\{\hat{y}_n^{(0)}\}$ and $\{q_n\}$ yields a tissue map and a focus-quality map respectively.

VLM-based segmentation mask review. Once the tissue mask has been produced, we employ a frontier VLM (Gemini 3 series) as an automated annotation reviewer. The reviewer receives two inputs: (i) the tissue-core crop from the WSI thumbnail, and (ii) the same crop with the predicted mask overlaid. It is prompted to assess the overall quality of the segmentation and returns discrete quality grades for both precision and recall, $g_p, g_r \in \{1, 2, 3, 4\}$, ordered from 1 (best) to 4 (worst). We define a strict gate τ_g=1, so grades 2–4 would trigger the few-shot refinement stage described in Stages (iv–v). Reported reviewer and segmentation results evaluate these components separately, rather than the gated loop end-to-end.

High-magnification exemplar selection (*Auto-context*). Should refinement be triggered (e.g. by reviewer grades exceeding τ_g), we query the VLM to perform point grounding on the tissue-core crop, asking it to identify candidate tissue and background locations. This yields spatially informed candidate pools $\mathcal{G}^+$ (tissue) and $\mathcal{G}^-$ (background) that are better balanced than naïve random sampling, which would be dominated by background. Because point grounding is imprecise, we present the corresponding high-magnification patches to the VLM in a comparative re-ranking prompt and select the top-K tissue exemplars $\mathcal{E}^+$ and top-K background exemplars $\mathcal{E}^-$, forming the in-context exemplar set $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$.

Few-shot patch classification. The exemplar set $\mathcal{E}$ is prepended to the VLM prompt as labelled demonstrations, and all patches are re-classified. The updated label for patch n is $\hat{y}_n = f_{\text{VLM}}(x_n \mid \mathcal{E})$, where f_{VLM} denotes the VLM response under the few-shot prompt. This stage targets whole slide images where zero-shot performance is limited by background staining artefacts whose high-magnification appearance mimics cellular structures.
Post-processing. We apply lightweight morphological cleanup on this grid using 4-neighbour connectivity: (i) remove foreground connected components with area < 3 patches, and (ii) fill background holes with area ≤ 10 patches.
Knowledge distillation. We perform *hard-label* distillation from VLM pseudo-labels at patch level. A lightweight student f_θ (MobileNetV3-Large-100, 4.2 M params, 2-class head) is trained with standard cross-entropy on canonical binary targets $\hat{y}_n \in \{0, 1\}$; no temperature scaling or teacher-logit matching is used.

3 Evaluation

We evaluate on Leica-75: 75 renal transplant WSIs (25 per stain: EVG, H&E, Jones) from 68 patients, scanned on a Leica system. Seven patients are represented by two WSIs; in four cases, the two slides use different stains. Slides are partitioned into non-overlapping 512×512 level-0 patches. Ground-truth masks are manually curated at patch level (one binary tissue/background label per patch), rather than at pixel level, across 220 annotated tissue cores; a patch is foreground when at least 50% of its area within the annotated core is painted as tissue. We additionally define a curated stress-test cohort, Stress-32 (n=32 WSIs from 25 patients; 23 Jones, 9 EVG), comprising slides where zero-shot predictions confuse background artefacts with tissue under high magnification. Both benchmarks (foreground masks and case lists) will be publicly released on Zenodo as part of a larger institutional data release (summer 2026). For blur assessment, we sample 200 patches from 186 WSIs, stratified by Qwen3 VL 8B predicted focus labels (sharp / mild blur / out of focus) into approximately equal thirds, and collect independent human judgements for each patch. For evaluating VLMs as segmentation-mask reviewers, we discretise precision and recall into four grades (excellent $> 95\%$, good 80–95%, partial 50–80%, poor $<50\%$) and construct a balanced benchmark by manually corrupting expert-verified masks to obtain uniform coverage over the resulting 4×4×3 grid (precision × recall × stain), yielding 96 masks with known ground-truth grades. We distil two student models using Leica-75 or Stress-32 following a 70/10/20 WSI-level train/val/test split, sampling 5,000 foreground and background patches from the training split with standard data augmentations (colour jitter, random rotations, flips). The Leica-75 test split has no patient overlap with either train or validation, although one patient spans the train and validation splits. For Stress-32 evaluation, we exclude two test WSIs from patients also present in training, leaving five test WSIs without patient overlap. Leica-75 and Stress-32 share three patients, including one exact WSI. We compare six baselines (Otsu [20], K-means [17], HistoQC [10], TRIDENT-GrandQC [26,29], TRIDENT-HEST [11,29], and SAM3 [5] prompted

Table 1. Dice (mean$_{\pm\text{std}}$) on 75 Leica WSIs; 2B/8B denote Qwen3-VL oracle variants. Best per stain **bold**. SAM3 is the strongest non-VLM baseline on EVG/Jones (GrandQC on H&E); per-stain significance reported in the text.

	Baselines						*Ours (VLM oracle)*			
Stain	Otsu	K-means	HistoQC	GrandQC	HEST	SAM3	2B-ZS	8B-ZS	2B-FS	8B-FS
EVG	$0.638_{\pm.284}$	$0.667_{\pm.270}$	$0.548_{\pm.276}$	$0.769_{\pm.148}$	$0.793_{\pm.049}$	$0.801_{\pm.119}$	$0.835_{\pm.123}$	$\mathbf{0.853}_{\pm.041}$	$0.844_{\pm.039}$	$0.847_{\pm.035}$
H&E	$0.679_{\pm.204}$	$0.667_{\pm.224}$	$0.224_{\pm.235}$	$\mathbf{0.870}_{\pm.047}$	$0.807_{\pm.075}$	$0.849_{\pm.083}$	$0.852_{\pm.034}$	$0.843_{\pm.045}$	$0.829_{\pm.047}$	$0.828_{\pm.047}$
Jones	$0.644_{\pm.239}$	$0.637_{\pm.244}$	$0.281_{\pm.298}$	$0.775_{\pm.194}$	$0.743_{\pm.109}$	$0.809_{\pm.189}$	$0.854_{\pm.072}$	$\mathbf{0.858}_{\pm.027}$	$0.843_{\pm.031}$	$0.840_{\pm.031}$

with our Stage-1 boxes) against four VLM oracle configurations (Qwen3 VL 2B/8B, zero-shot and few-shot [1]). For few-shot Auto-context, we set K=1. Statistical comparisons use paired Wilcoxon signed-rank tests and, for per-slide variance, a paired variance-ratio permutation test; we treat $p<0.05$ as significant and report exact, uncorrected p-values. Finally, we evaluate 264 WSIs from an external kidney-pathology centre in a different country.

Foreground segmentation. Tab. 1 reports stain-stratified Dice. SAM3, a strong general-purpose segmenter, surpasses the supervised pathology baselines overall (mean Dice 0.820 vs. GrandQC 0.805, HEST 0.781) and is thus a credible modern comparator; nonetheless VLM methods achieve the highest mean Dice on both specialized stains (Qwen 8B-ZS $0.853_{\pm0.041}$ EVG, $0.858_{\pm0.027}$ Jones vs. SAM3 $0.801_{\pm0.119}$, $0.809_{\pm0.189}$). The VLM significantly exceeds SAM3 on EVG (8B-ZS p=0.012, 2B-ZS p=0.004, paired Wilcoxon) but not on Jones (p=0.96), where the higher mean is robustness-driven: SAM3 collapses on a minority of out-of-distribution slides (worst-case Δ up to +0.85 Dice) rather than losing uniformly. This robustness is itself significant; Qwen 8B-ZS shows far lower per-slide Dice variance than SAM3 (SD 0.038 vs. 0.137; variance-ratio permutation $p<0.001$, Brown–Forsythe p=0.002). On in-distribution H&E, GrandQC achieves $0.870_{\pm0.047}$ owing to its supervised training advantage; VLM methods remain competitive at $0.852_{\pm0.034}$ without any in-context examples. HistoQC achieves moderate precision but catastrophically low recall (0.150 on H&E, 0.209 on Jones), discarding the majority of viable tissue. Aggregating across stains, our 8B-ZS oracle achieves the highest mean Dice ($\bar{D}$=0.851) and the lowest cross-stain standard deviation of any method (σ_s=0.006): 3.5× lower than the strongest general-purpose baseline (SAM3, σ_s=0.021) and 7× lower than the best pathology-specific model (GrandQC, σ_s=0.046, $D_{\min}$=0.769), confirming more consistent predictions across stain families. To verify that this is not an artefact of the stages preceding the VLM, we ran a fairness ablation giving the strongest baselines (GrandQC, SAM3) identical bounding-box and colour-mask preprocessing: Qwen-8B zero-shot still yields higher recall on all three stains (vs. GrandQC ΔR=+0.115/+0.040/+0.133; vs. SAM3 +0.070/+0.028/+0.068, EVG/H&E/Jones), confirming the gain originates from the VLM decision layer itself. Under-segmentation is the more dangerous failure mode because it irrecoverably excludes diagnostically relevant patches from all downstream analysis.

Component ablation. Tab. 2 quantifies the contribution of each pipeline stage on Stress-32. Naïve WSI-level zero-shot classification (row 1) achieves $0.107_{\pm.060}$

Table 2. Pipeline ablation on Stress-32 (n=32) using Qwen3-VL 2B/8B. Each row adds one component.

#	Configuration	Dice	Prec	Rec	Acc	ΔD
1	2B WSI-level ZS	$.107_{\pm.060}$	$.058_{\pm.034}$	$.976_{\pm.019}$	$.735_{\pm.073}$	–
2	+VLM bbox filter	$.426_{\pm.107}$	$.279_{\pm.092}$	$.976_{\pm.019}$	$.958_{\pm.028}$	+.319
3	+Colour filter	$.520_{\pm.159}$	$.378_{\pm.176}$	$.952_{\pm.034}$	$.971_{\pm.020}$	+.094
4	+Auto-context 2B	$.812_{\pm.093}$	$.811_{\pm.133}$	$.832_{\pm.093}$	$.994_{\pm.005}$	+.292
5	+Auto-context 8B	$.836_{\pm.063}$	$.829_{\pm.073}$	$.846_{\pm.076}$	$.995_{\pm.004}$	+.024
6	+Post-processing	$\mathbf{.848_{\pm.061}}$	$.850_{\pm.067}$	$.850_{\pm.074}$	$.995_{\pm.004}$	+.013

Table 3. Zero-shot versus few-shot comparison on Stress-32 (n=32) using Qwen3-VL 2B/8B. $^{***}p<0.001$ (paired Wilcoxon).

	Absolute				*Δ(FS−ZS)*	
	2B-ZS	2B-FS	8B-ZS	8B-FS	Δ2B	Δ8B
Dice	$.532_{\pm.163}$	$.824_{\pm.091}$	$.790_{\pm.096}$	$\mathbf{.848_{\pm.061}}$	$+.292^{***}$	$+.058^{***}$
Prec	$.396_{\pm.181}$	$.831_{\pm.125}$	$.731_{\pm.132}$	$\mathbf{.850_{\pm.067}}$	$+.434^{***}$	$+.119^{***}$
Rec	$\mathbf{.932_{\pm.054}}$	$.834_{\pm.095}$	$.880_{\pm.069}$	$.850_{\pm.074}$	$-.098^{***}$	$-.030^{***}$
Acc	$.973_{\pm.021}$	$.994_{\pm.005}$	$.992_{\pm.006}$	$\mathbf{.995_{\pm.004}}$	$+.022^{***}$	$+.003^{***}$

Dice due to extreme base-rate imbalance: tissue typically occupies only a small fraction of the slide, so even a low false-positive rate yields many spurious foreground patches and collapses precision. VLM bounding-box filtering (row 2) provides the first major gain ($\Delta D = +0.319$) by restricting processing to VLM-localised tissue regions. Two-pass colour refinement adds a further +0.094 (row 3). The largest precision gain comes from *Auto-context* exemplar mining (K=1) and few-shot prompting (row 4), which increases Dice by +0.292 and improves precision from 0.378 to 0.811 while maintaining high recall (0.832). Scaling from 2B to 8B yields an additional +0.024 (row 5), and lightweight morphological post-processing contributes a final +0.013 to reach $0.848_{\pm.061}$ Dice (row 6).

Few-shot stress test. On the Stress-32 set (Tab. 3), Qwen 2B-ZS achieves $0.532_{\pm 0.163}$ Dice (precision 0.396) due to deceptive background textures, while Qwen 8B-ZS reaches $0.790_{\pm 0.096}$. Few-shot prompting with automatically curated exemplars yields striking improvements: Qwen 2B rises to $0.824_{\pm 0.091}$ (Δ=+0.292, $p<0.001$) and Qwen 8B to $0.848_{\pm 0.061}$ ($\Delta = +0.058$, $p<0.001$). The mechanism is precision-driven (+0.434 for 2B): the model learns to reject deceptive background patches by leveraging slide-specific exemplars, with individual case gains up to +0.59 Dice.

Knowledge distillation. Tab. 4 compares each student with the corresponding Qwen-8B pseudo-label generator on the same held-out cases. The lightweight MobileNetV3 student (4.2 M parameters) approaches teacher performance while running ∼100× faster (586 vs. 4–6 patches/sec). On Leica-75 the student achieves $F1$=0.884 (teacher 0.878); on the Stress-32 test set it achieves $F1$=0.895 (teacher 0.910). Cross-set transfer remains asymmetric: the Stress-32 student transfers well to Leica-75 ($F1$=0.879), whereas the Leica-75 student degrades on Stress-32 ($F1$=0.750).

Table 4. Knowledge distillation on held-out test WSIs (Leica-75: n=15; Stress-32: n=5). Metrics are pooled within annotated tissue-core regions.

Eval	Model	Params	$F1\uparrow$	IoU↑	$P\uparrow$	$R\uparrow$	p/s†
Leica-75	Qwen-8B ZS teach.	∼8B	.878	.783	.940	.824	6
	Student (Leica-75)	4.2M	**.884**	**.792**	.931	.841	586
Stress-32	Qwen-8B FS teach.	∼8B	**.910**	**.835**	.918	.902	4
	Student (Stress-32)	4.2M	.895	.810	.858	.935	586
Student (Stress-32)→Leica-75 test		4.2M	.879	.784	.924	.838	586
Student (Leica-75)→Stress-32 test		4.2M	.750	.600	.625	.937	586

†Patches/sec, single GPU.

Table 5. Blur detection (n=200): Qwen 8B vs. human consensus.

Metric	Est.	95 % CI
Hum. α (3-cl.)	$.927_{\pm.016}$	[.894, .955]
VLM κ_w (3-cl.)	$.708_{\pm.036}$	[.632, .773]
VLM κ (bin.)	$.989_{\pm.009}$	[.963, 1.00]
VLM sens. (bin.)	$1.00_{\pm.000}$	[1.00, 1.00]
VLM spec. (bin.)	$.993_{\pm.006}$	[.977, 1.00]

Table 6. Mask-quality reviewer alignment on corrupted masks (n=96). Best VLM **bold**; best human underlined; ‡VLM exceeds best human. Bottom rows: group means. QWK denotes quadratic-weighted Cohen's κ; MAE denotes mean absolute error.

	Accuracy (grade-level)			*QWK (ordinal)*		*MAE (%)*↓	
Rater	Prec	Rec	Joint	Prec	Rec	Prec	Rec
Hum. (Gi)	$.760_{\pm.089}$	$.750_{\pm.089}$	$.604_{\pm.094}$	$.886_{\pm.051}$	$.896_{\pm.040}$	$4.35_{\pm0.96}$	$4.35_{\pm1.03}$
Hum. (C)	$.729_{\pm.089}$	$.854_{\pm.073}$	$.646_{\pm.099}$	$.887_{\pm.045}$	$.941_{\pm.034}$	$6.30_{\pm0.96}$	$3.97_{\pm0.92}$
Hum. (V)	$.688_{\pm.099}$	$.844_{\pm.073}$	$.594_{\pm.099}$	$.862_{\pm.051}$	$.935_{\pm.032}$	$6.09_{\pm1.43}$	$3.26_{\pm0.65}$
Gem. 3.1 Pro	$.740_{\pm.089}$	$\mathbf{.885_{\pm.062}}^{\ddagger}$	$.635_{\pm.094}$	$.892^{\ddagger}_{\pm.045}$	$\mathbf{.953_{\pm.030}}^{\ddagger}$	$7.55_{\pm1.39}$	$4.58_{\pm0.85}$
Gem. 3 Flash	$.781^{\ddagger}_{\pm.083}$	$.875^{\ddagger}_{\pm.068}$	$.667^{\ddagger}_{\pm.089}$	$\mathbf{.896_{\pm.053}}^{\ddagger}$	$.951^{\ddagger}_{\pm.031}$	$\mathbf{6.09_{\pm1.11}}$	$\mathbf{3.99_{\pm0.62}}$
Gem. 3 Pro	$\mathbf{.792_{\pm.083}}^{\ddagger}$	$.875^{\ddagger}_{\pm.068}$	$\mathbf{.708_{\pm.094}}^{\ddagger}$	$.894^{\ddagger}_{\pm.054}$	$.915_{\pm.080}$	$7.49_{\pm1.28}$	$4.82_{\pm1.25}$
Human mean	*.726*	*.816*	–	*.878*	*.924*	*5.58*	*3.86*
VLM mean	*.771*	*.878*	–	*.894*	*.939*	*7.04*	*4.46*

Blur detection. For focus quality assessment (Tab. 5), Qwen 8B achieves near-perfect binary agreement with human consensus (out-of-focus vs. in-focus, with *sharp/mild blur* collapsed; κ=0.989, sensitivity 1.000, specificity 0.993) and strong 3-class ordinal agreement (κ_w=0.708) with zero extreme misclassifications. This confirms that VLMs can serve as reliable automated quality gates for out-of-focus patches at no additional inference cost, since blur labels are extracted jointly with tissue classifications. For VLM-free deployment, the gate also distils cheaply: a logistic-regression probe on frozen DINOv3-small [23] features reproduces the binary Qwen 8B blur labels at $0.995_{\pm.010}$ accuracy under case-grouped 5-fold cross-validation ($F1$=0.992; one miss in 200), confirming the blur boundary is near-linearly separable in self-supervised feature space.

VLM reviewer alignment. For mask-quality grading (Tab. 6), we evaluate three Gemini reviewers (Gemini 3 Pro, 3 Flash, and 3.1 Pro); Gemini 3 Pro achieves the highest joint exact-match accuracy ($0.708_{\pm0.094}$), exceeding the best human rater ($0.646_{\pm0.099}$) by +0.062. Group means show VLMs outperform humans on grade-level accuracy (0.771/0.878 vs. 0.726/0.816 for precision/recall) and ordinal agreement (QWK 0.894/0.939 vs. 0.878/0.924), while humans retain a calibration advantage (MAE 5.58 vs. 7.04). Inter-rater QWK averages 0.904 for human-human, 0.884 for VLM-VLM, and 0.857 for human-VLM pairs. On a

complementary subjective review task, binarising grades as excellent vs. needs review (grades 2-4), VLM-human shows κ=0.717, confirming VLMs can reliably triage masks for human escalation even when given no explicit scoring criteria.
Multi-centre validation. Our implementation was run at an external centre on a separate renal cohort of 264 WSIs (58 H&E, 128 Jones, 78 PAS) from six hospitals, yielding 353 valid human-annotated regions of interest (ROIs), of which 340 contained reference foreground and contributed to recall. Qwen-8B zero-shot achieves an overall Dice of $0.784_{\pm 0.199}$ with high recall ($0.964_{\pm 0.096}$) and moderate precision ($0.701_{\pm 0.205}$). Per-hospital Dice ranges from 0.715 to 0.840. The lower precision relative to the Leica-75 benchmark is expected, as this external cohort was curated as a deliberately challenging evaluation set. The consistently high recall (≥ 0.94 across hospitals) confirms that our method reliably detects tissue across institutions without site-specific tuning.

4 Discussion

General-purpose VLMs match or surpass dedicated pathology tools and a strong general segmenter on out-of-distribution stains with markedly lower cross-stain variance, and distilled students preserve this at $\sim$100$\times$ throughput. Coverage spans three stain families but a single organ (kidney) with relatively favorable tissue contrast, so transfer to sparse or low-contrast specimens remains untested. The Stage-1 bounding-box filter caps achievable recall, and slides where it fails entirely never reach the segmentation layer. The robustness advantage over SAM3 on Jones is driven by tail behavior on a minority of slides. Broader organ and scanner coverage, and evaluation of the VLM reviewer on naturalistically generated masks, are the main directions for future work.

5 Conclusion

By reframing foreground segmentation as a perception task, we show that general-purpose VLMs outperform both dedicated pathology tools and a strong general-purpose segmenter (SAM3) on out-of-distribution stains (Dice 0.858 Jones, 0.853 EVG) with 3.5–7$\times$ lower cross-stain variance, while Auto-context rescues Stress-32 hard cases for Qwen 2B (+0.292 mean Dice) and VLM quality gates match humans (blur κ=0.989; mask-review accuracy 0.708 vs. 0.646). Distillation into lightweight students decouples deployment from frontier-model cost.

Acknowledgments. S. Cechnicka is supported by the UKRI Centre for Doctoral Training AI4Health (EP/S023283/1) and a UKRI Round 6 IAA award. We acknowledge HPC resources from NHR@FAU (projects b143dc and b180dc), funded by federal and Bavarian state authorities. We thank Gerhard Wellein and his team for their HPC support. NHR@FAU hardware is partially funded by DFG project 440719683. Additional support was received from ERC projects MIA-NORMAL (101083647) and CHARMS (101246053), and DFG projects 513220538 and 512819079. Dr. Roufosse is supported by the National Institute for Health Research (NIHR) Biomedical Research

Centre based at Imperial College Healthcare NHS Trust and Imperial College London (ICL). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. Dr. Roufosse's research activity is made possible with generous support from Sidharth and Indira Burman. Human samples used in this research project were obtained from the Imperial College Healthcare Tissue & Biobank (ICHTB). ICHTB is supported by the NIHR Biomedical Research Centre based at Imperial College Healthcare NHS Trust and ICL. ICHTB is approved by Wales REC3 to release human material for research (22/WA/2836).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv:2511.21631 (2025), https://arxiv.org/abs/2511.21631
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., et al.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
3. Bejnordi, B.E., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A., the CAMELYON16 Consortium: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017). https://doi.org/10.1001/jama.2017.14585
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busber, K.J., Madabhushi, A., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (2019). https://doi.org/10.1038/s41591-019-0508-1
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., et al.: SAM 3: Segment anything with concepts. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=r35clVtGzw
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nat. Med. **30**, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3
7. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Camara, A., Mac Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. MedRxiv (2023). https://doi.org/10.1101/2023.07.21.23292757, https://www.medrxiv.org/content/10.1101/2023.07.21.23292757v3
8. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat. Commun. **12**(1), 4423 (2021). https://doi.org/10.1038/s41467-021-24698-1
9. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: ICML'18. pp. 2127–2136. PMLR (2018)
10. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M.D., Madabhushi, A.: HistoQC: An open-source quality control tool for digital pathology slides. JCO Clin. Cancer Inform. **3**, 1–7 (2019). https://doi.org/10.1200/CCI.18.00157

11. Jaume, G., Doucet, P., Song, A.H., Lu, M.Y., Almagro-Perez, C., Wagner, S.J., Vaidya, A.J., Chen, R.J., Williamson, D.F., Kim, A., et al.: HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track (2024). https://doi.org/10.52202/079017-1704, https://arxiv.org/abs/2406.16192
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023). https://doi.org/10.1109/ICCV51070.2023.00371, https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html
13. Kömen, J., de Jong, E.D., Hense, J., Marienwald, H., Dippel, J., Naumann, P., Marcus, E., Ruff, L., Alber, M., Teuwen, J., Klauschen, F., Müller, K.R.: Towards robust foundation models for digital pathology. Nature Communications **17**(1), 5218 (2026). https://doi.org/10.1038/s41467-026-73923-2
14. Lin, W., Liu, S., Zhu, R., Lin, Y., Wang, B., Wang, L.: Beyond diagnostic performance: Revealing and quantifying ethical risks in pathology foundation models (2025), https://arxiv.org/abs/2502.16889
15. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nat. Med. **30**, 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4
16. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**(6), 555–570 (2021). https://doi.org/10.1038/s41551-020-00682-w
17. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. pp. 281–297. University of California Press, Berkeley, CA (1967), https://digicoll.lib.berkeley.edu/record/113015
18. Nasir, E.S., Elhaminia, B., Eastwood, M., King, C., Cain, O., Harper, L., Moss, P., Chanouzas, D., Snead, D., Rajpoot, N., Shephard, A., Raza, S.E.A.: CORE – a cell-level coarse-to-fine image registration engine for multi-stain image alignment. arXiv:2511.03826 (2025), https://arxiv.org/abs/2511.03826
19. Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: A family of foundational vision transformers for pathology. arXiv:2406.05074 (2024). https://doi.org/10.48550/arXiv.2406.05074, https://arxiv.org/abs/2406.05074
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
21. Pocock, J., Graham, S., Vu, Q.D., Jahanifar, M., Deshpande, S., Hadjigeorghiou, G., Shephard, A., Bashir, R.M.S., Bilal, M., Lu, W., et al.: TIAToolbox as an end-to-end library for advanced tissue image analytics. Commun. Med. **2**(1), 120 (2022). https://doi.org/10.1038/s43856-022-00186-5
22. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: NeurIPS'21. pp. 2136–2147 (2021)
23. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L.,

Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=2NlGyqNjns, published 15 May 2026
24. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nat. Med. **30**(10), 2924–2935 (2024). https://doi.org/10.1038/s41591-024-03141-0
25. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191 (2024), https://arxiv.org/abs/2409.12191
26. Weng, Z., Seper, A., Pryalukhin, A., Mairinger, F., Wickenhauser, C., Bauer, M., Glamann, L., Bläker, H., Lingscheidt, T., Hulla, W., et al.: GrandQC: A comprehensive solution to quality control problem in digital pathology. Nat. Commun. **15**(1), 10685 (2024). https://doi.org/10.1038/s41467-024-54769-y
27. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4818–4829 (June 2024). https://doi.org/10.1109/CVPR52733.2024.00461, https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Florence-2_Advancing_a_Unified_Representation_for_a_Variety_of_Vision_CVPR_2024_paper.html
28. Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Porter, C.W., Coffee, M., Zeng, Z., Li, S., Wang, Z., Huang, T.H.K., Rehg, J.M., Kautz, H., Zhang, A.: The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=kIuqPmS1b1, published 21 February 2026
29. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating data processing and benchmarking of AI models for pathology. arXiv:2502.06750 (2025), https://arxiv.org/abs/2502.06750