MOLECULARLY GUIDED TISSUE SEGMENTATION AND RECOGNITION IN SPATIAL TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics reveals molecular organization within tissue, but translating these patterns into tissue boundaries and class labels remains challenging. We introduce Molecularly Oriented Spatial Segmentation and Tagging (MOSST), a framework that uses molecular evidence to guide both region proposal and tissue recognition. Spatial gene-expression patterns provide box-and-point prompts for SAM3 on H&E and FICTURE maps, supplemented by H&E proposals where molecular signal is weak. Candidate masks are selected without tissue-region annotations. GPT-6 then classifies each candidate using H&E and a Cell Map displaying annotated cell types within and around its boundary. We evaluate MOSST on ten lung tissue microarrays with 75 confidence-annotated ground-truth regions. Evaluation of all selected candidates yields 37/75 regions both localized and correctly classified under one-to-one matching at intersection-over-union ≥ 0.10. In a separate recognition experiment, candidates are paired to ground-truth regions by geometric overlap and held fixed. Adding Cell Map to H&E increases correct classifications from 33/63 to 42/63 and also outperforms the tested cell-composition text representation. These findings support combining molecularly guided region proposals with candidate-aligned cell-type maps for tissue recognition. Precise boundary delineation and generalization beyond the evaluated cohort require further validation.

1 INTRODUCTION

Spatial transcriptomics measures gene expression within tissue architecture (Ståhl et al., 2016). Linking these measurements to structures such as airways, vessels, and lymphoid aggregates requires locating each region, delineating its boundary, and assigning a tissue class. Spatial domains group locations by molecular similarity. Our target is a set of tissue-region masks with class labels that can be inspected on H&E.

Spatial-domain methods characterize molecular organization (Zhao et al., 2021; Hu et al., 2021; Dong & Zhang, 2022; Long et al., 2023; Si et al., 2024). LLMiniST uses spatial neighborhoods to construct ordered cell-type and gene lists for niche identification (Wei et al., 2025). These summaries encode neighborhood composition but do not show the within-neighborhood arrangement of cells against a tissue boundary. Promptable segmentation and image–text recognition offer tools for extracting and naming regions (Kirillov et al., 2023; Carion et al., 2026; Radford et al., 2021; Lu et al., 2024; Aleem et al., 2024). The question is how to connect molecular locations and cellular arrangement to tissue masks and their classes.

Expression patterns and tissue structures need not share boundaries. Molecular maps can locate candidate structures, but weak signal can leave some unrepresented. For recognition, a Cell Map displays annotated cell types relative to the candidate contour and surrounding H&E tissue. We test whether this map and its legend improve tissue naming.

Molecularly Oriented Spatial Segmentation and Tagging (MOSST) converts GeneMap regions into box-and-point prompts for SAM3 on H&E and FICTURE (Figure 1). An independent H&E branch covers weak molecular signal. Selection uses no GT information. GPT-6 classifies each retained mask from H&E and Cell Map, with FICTURE and factor-gene lists as an alternative input. Neither model requires task-specific fine-tuning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

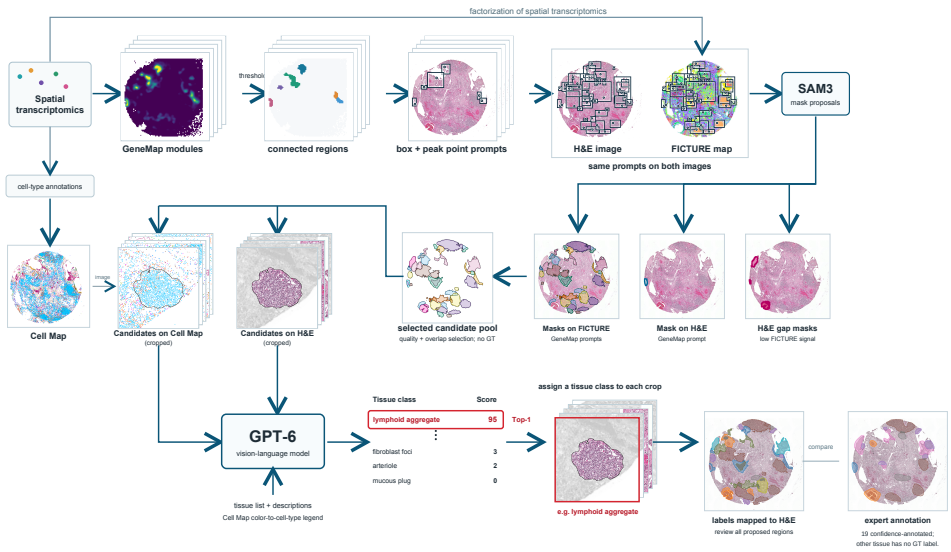

Figure 1: **Overall framework.** GeneMap scores are thresholded into connected regions that define box-and-point prompts. SAM3 produces masks from the same prompts on H&E and FICTURE, while an independent H&E branch covers locations with weak FICTURE signal (top). Candidate selection uses no GT information (middle). Each retained mask defines aligned H&E and Cell Map crops for GPT-6. Predicted classes are mapped back to the H&E core beside confidence-annotated GT regions (bottom). Images are from TMA39. The Cell Map color key appears in Figure 4.

MOSST connects molecular evidence to tissue regions in two ways: GeneMap locations guide mask proposals, and candidate-aligned Cell Maps provide spatial cell-type evidence for recognition. Across ten lung TMAs, MOSST gives the most correctly classified GT matches among the tested configurations, whose inputs and mask budgets differ. With candidates and GPT-6 fixed, adding Cell Map improves classification over H&E alone. This controlled result supports the map and its legend as recognition inputs, without isolating spatial arrangement from cell-type identity.

2 Related Work

Spatial-domain methods cluster expression profiles, sometimes incorporating histology, as in SpaGCN (Zhao et al., 2021; Hu et al., 2021; Dong & Zhang, 2022; Long et al., 2023). FICTURE estimates spatial expression factors (Si et al., 2024), while SEAL uses spatial transcriptomics to learn pathology representations (Hemker et al., 2026). LLMiniST predicts cell or spot niches from spatial-neighborhood cell-type and gene lists (Wei et al., 2025). MOSST instead generates region masks and supplies candidate-aligned H&E and Cell Map images, preserving cell positions relative to each contour. This is a difference in representation and output, not a measured advantage over LLMiniST. SaLIP already combines SAM and CLIP for medical segmentation (Aleem et al., 2024). Our contribution is the use of spatial molecular evidence within a segmentation–recognition workflow, rather than the cascade itself. We evaluate this workflow against automatic and language-guided systems and separately test molecular inputs on fixed candidates.

3 Method

3.1 Inputs, Output, and Four Label Systems

For each tissue microarray (TMA), the inputs are spatial transcript counts and coordinates, an H&E image H, a FICTURE factor map F, and a Cell Map L of annotated cell centers. Saved TMA-specific transforms place these measurements in the image coordinate system. The output is a set of binary candidate masks C_i with assigned tissue classes $\hat{y}_i$. GeneMap modules group genes

for proposing regions; FICTURE factors are expression programs; Cell Map labels are biologist-provided cell-type annotations used without relabeling; and the final output uses twelve tissue classes. These four vocabularies have distinct meanings. Neither a GeneMap module nor a FICTURE factor is a tissue diagnosis.

3.2 MOLECULARLY GUIDED CANDIDATE PROPOSALS

Transcript counts to GeneMaps. For gene g and image-aligned grid bin u, we sum the counts of Xenium transcripts mapped into that bin to obtain N_{gu}. A bin is eight low-resolution H&E pixels per side. Only bins with at least 20 total transcript counts enter the tissue support $\mathcal{T}$. Eligible genes have at least 500 counts across at least 25 occupied bins. Their count maps undergo $\log(1 + N_{gu})$, Gaussian smoothing with $\sigma = 1$ bin, and within-TMA standardization over $\mathcal{T}$. We retain the 120 genes with the strongest positive agreement between neighboring tissue-bin values, excluding nearly constant maps.

Within each TMA, Pearson correlations compare these standardized maps. Average-linkage clustering uses distance $1 - \text{correlation}$ and tests $K = 4, \ldots, 10$. We select the K maximizing standard silhouette plus mean adjusted Rand agreement with 100 reclusterings of 80% resampled tissue bins. Ties favor fewer singleton modules and then smaller K. Gene selection and module membership are recomputed for each TMA without annotations. A module's untrimmed GeneMap score is the mean standardized expression of its member genes:

$$E_m(u) = |G_m|^{-1} \sum_{g \in G_m} Z_g(u), \qquad u \in \mathcal{T}. \tag{1}$$

GeneMaps to connected regions and spatial prompts. Each E_m is thresholded independently over tissue bins. The original pass takes the strongest 1%, 2.5%, and 5% as core, middle, and outer support. An expanded pass takes 2.5%, 5%, and 10%. Each binary layer undergoes one 3×3 closing and hole filling, and is clipped to tissue. We enforce nested layers. Eight-neighbor core components smaller than two bins are discarded. If several cores share one outer-support component, each support bin is assigned to its nearest core. Regions in the same module are then merged only when they share at least three edge-neighbor bins. For each resulting region $R_{m,r}$, grid-bin edges give a tight image-coordinate box. Its highest-scoring bin supplies a positive point:

$$b_{m,r} = \text{bbox}(R_{m,r}), \qquad p_{m,r} = \text{center}\left(\arg\max_{u \in R_{m,r}} E_m(u) \right). \tag{2}$$

The executed configuration adds no extra GeneMap-bin margin. The connected region, box, and point are prompts, not segmentation masks.

Three sources of masks. Each box and point prompts SAM3 on the H&E image and separately on the FICTURE map:

$$\mathcal{P} = \bigcup_{m,r} [\text{SAM3}(H; b_{m,r}, p_{m,r}) \cup \text{SAM3}(F; b_{m,r}, p_{m,r})] \cup \mathcal{P}_H. \tag{3}$$

Using the same location on both images permits morphological and molecular appearance to yield different boundaries. The additional H&E branch $\mathcal{P}_H$ addresses locations where FICTURE has little signal. A 64-pixel block qualifies if H&E tissue occupies at least 35% and FICTURE color signal at most 8%. After 3×3 closing, connected areas of at least six blocks produce a box expanded by 10% and a positive point within H&E tissue. SAM3 segments the corresponding H&E crop. Its mask is returned to whole-core coordinates. This branch uses neither a GeneMap prompt nor GT regions. Figure 2 shows the three proposal branches. Appendix B.1 details the implementation.

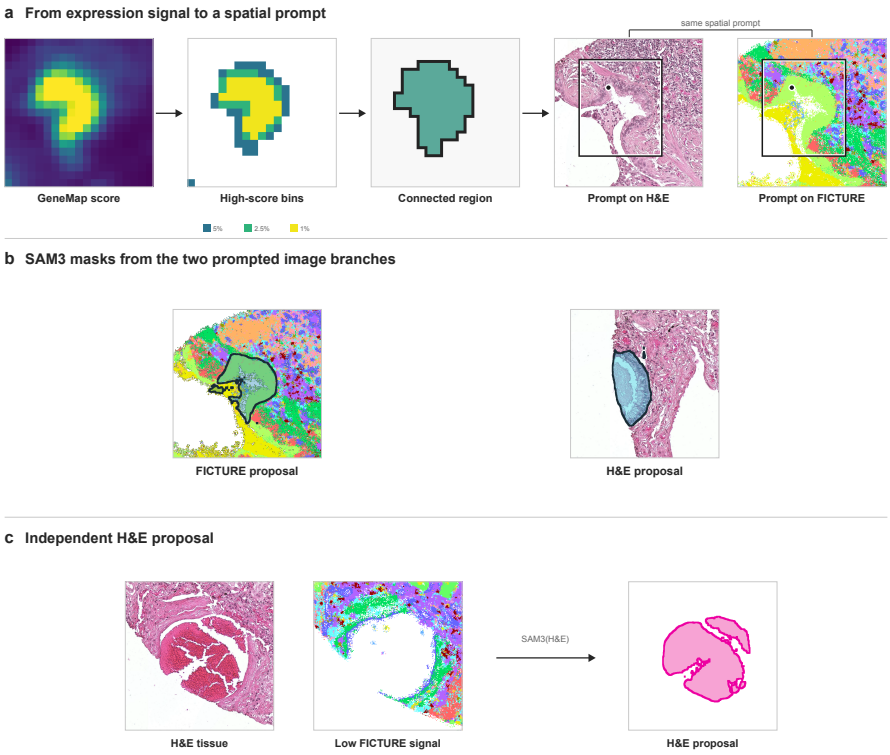

Figure 2: **Candidate generation.** (a) Thresholded GeneMap bins form connected regions, each providing a box and peak point for SAM3. (b) Example masks from the FICTURE and H&E branches, generated from different prompts. (c) The independent H&E branch covers tissue with weak FICTURE signal.

3.3 Candidate selection without GT information

The branches can produce overlapping masks. For n input masks, let $r_i^s, r_i^t \in [1/n, 1]$ be normalized SAM3-score and prompt-stability ranks. The executed quality score is $q_i = \mathbf{1}[\text{non-FICTURE source}] + (r_i^s + r_i^t)/4$. On a 512×512 working grid, if U is the already selected area, we choose

$$\nu_i(U) = \frac{|C_i \setminus U|}{|C_i|}, \qquad i^* = \underset{i:\nu_i(U)\geq .25}{\arg\max}\ q_i[.5 + .5\nu_i(U)]. \tag{4}$$

A selected mask removes remaining masks with full-resolution IoU at least .50. The development configuration caps selection at 40 per TMA. Neither GT class labels nor overlap with GT regions enter this greedy choice. The candidate-selection parameters, including the per-TMA limit, quality-score weights, and overlap threshold, were set during development on Visium HD and held fixed for the Xenium evaluation. The rule is not a global optimum. Figure 3 illustrates the selection sequence and a representative retained pool.

3.4 Recognition views and tissue-class assignment

Each selected mask fixes a close target crop and a wider context crop. Identical bounds and target contours are used across supplied image types. With $\phi_v(I, C_i)$ denoting an outlined crop,

$$\mathcal{V}_i(\mathcal{S}) = \{\phi_v(I, C_i) : I \in \mathcal{S},\ v \in \{\text{close}, \text{wider}\}\}. \tag{5}$$

The main workflow supplies H&E and Cell Map. The Cell Map plots saved cell centers with a fixed 24-color type key but does not define the mask. When FICTURE is supplied, the request also lists the top 20 loading genes per factor, without assigning cell-type names to factors. The recognizer

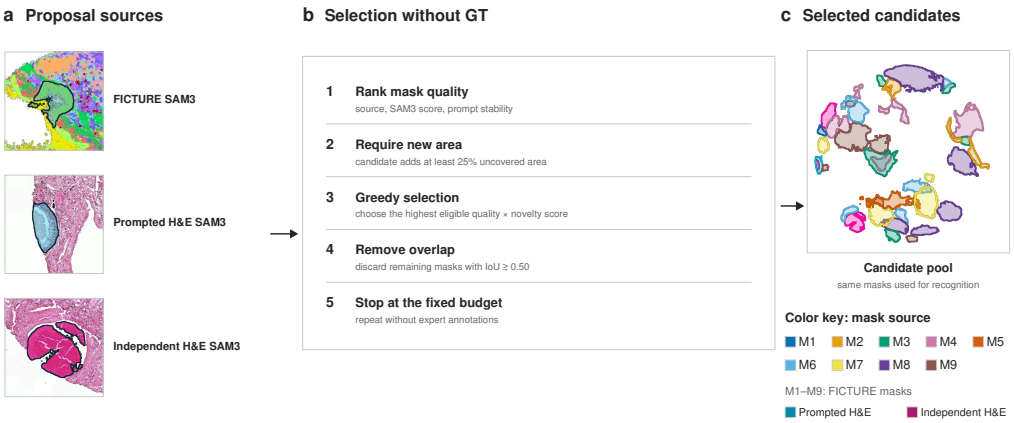

Figure 3: **Candidate selection.** (a) Masks from three proposal branches. (b) Greedy selection based on mask quality, additional coverage, overlap, and a per-TMA budget, without GT information. (c) Selected masks. Colors distinguish GeneMap modules for FICTURE masks and the two H&E sources.

receives tissue-class descriptions and a fixed response schema, and returns a support score s_{ik} for each offered class, a chosen label, and short evidence. The reported Top-1 class is

$$\hat{y}_i = \arg\max_{k \in \mathcal{Y}_i} s_{ik}. \tag{6}$$

Scores are model support values, not calibrated probabilities. Predicted masks are mapped back to H&E with their labels. Different-class masks may overlap or nest, as with an airway and an internal mucous plug (Appendix Figure 15). Figure 4 shows how the matched image views and their text keys enter GPT-6. Section 4 specifies the alternative recognition inputs and how they are compared. Appendix B.2 records crop and prompt details.

4 Experimental Design

4.1 Cohort and Annotation Scope

The workflow and candidate-selection parameters were established using an earlier Visium HD dataset and held fixed for evaluation on ten Xenium lung TMAs containing 75 regions selected and drawn by biologists. TMA39 is part of this evaluation cohort and is used to illustrate the workflow. Biologists also provided the cell-type annotations used to construct Cell Maps. Tissue-region annotations serve as ground truth (GT) for evaluation. The annotations do not form a dense map of all structures on a core, and their boundaries were not independently adjudicated for pixel-level precision. Annotators reviewed each core and labeled every structure they were confident identifying, so the 75 GT regions are an exhaustive record of what was confidently identifiable on each core, not a sampled subset of a larger set. Tissue outside these regions reflects insufficient annotator confidence, not an annotation budget. We therefore do not treat unannotated tissue as background or score every unmatched candidate as a false positive. These annotation limits apply to every result below.

4.2 Separating Candidate Availability, Boundary Agreement, and Recognition

An overall correct-label count alone cannot distinguish a missed location from a candidate with an inaccurate boundary or a candidate given the wrong tissue name. We therefore report these questions separately. First, *accessibility* asks whether at least one selected candidate C reaches a GT region G under

$$m(C, G) = \mathbf{1}\left[\frac{|C \cap G|}{\max(|C|, |G|)} \geq .10 \ \vee \ |C \cap G| = \min(|C|, |G|)\right]. \tag{7}$$

Containment can pass even with a large boundary mismatch. We use this permissive rule to measure candidate availability for confidence-annotated GT regions, not segmentation accuracy. We also

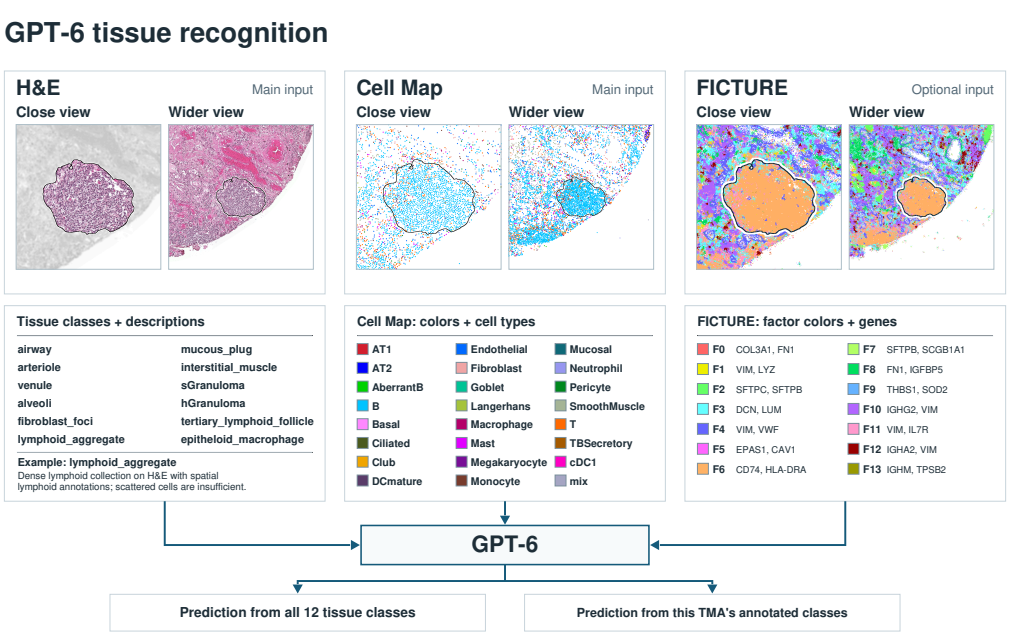

Figure 4: **GPT-6 recognition inputs.** Close and wider views share the candidate contour. H&E and Cell Map form B, with FICTURE optional. Requests include cell-type keys, 20 genes per FICTURE factor, and all offered class descriptions. The diagram abbreviates these texts. Full-12 is the main setting. The GT-derived TMA list is a separate oracle diagnostic. White FICTURE areas have no assigned factor.

evaluate one-to-one matches at IoU .10 and .50, separately maximizing geometric matches and matches with the correct class (Appendix A.2). Hand-drawn GT may include surrounding tissue, so IoU measures agreement with the recorded outline rather than tissue identity alone (Appendix Figure 15). We retain both IoU thresholds alongside recognition results.

For the fixed-candidate recognition experiment, each accessible GT region is assigned one eligible candidate by the largest intersection divided by the larger mask area. Candidate ID breaks ties. The predicted label is never used for that choice. Several GT regions can share one candidate. In total, 63 accessible GT regions map to 56 distinct candidates. Recognition is reported as correct/63, with correct/75 also shown to keep all 75 GT regions in view. These scores use GT geometry to choose representative candidates. They do not evaluate every generated mask.

4.3 Matched recognition-input conditions

The five original GPT-6 input conditions share 56 representative candidates, crop bounds, twelve class descriptions, response schema, and model settings. A subsequent composition-text control (F) replaces Cell Map with regional type frequencies (Appendix E.4). Their inputs are A: H&E; B: H&E and Cell Map; C: H&E and FICTURE; D: all three images; and E: Cell Map alone. Cell Map includes its color-to-cell-type key, while FICTURE includes factor colors and the top 20 genes for each factor. B is the primary recognition condition. A versus B is the focal input comparison. All statistical analyses are exploratory, not preregistered. Other input comparisons are descriptive. Table 3 reports the matched comparison.

4.4 Ranking metrics and class-list control

The primary *Full-12* setting offers the same twelve tissue classes for every TMA. A separate *TMA-restricted* control offers only the classes that occur in that TMA's GT regions. Restricted thus uses GT-derived class availability and is an oracle diagnostic, not an independent-use setting. It also changes which regions can serve as negatives for a class, so its AP cannot be interpreted as a matched-input gain over Full-12.

For each tissue class we calculate one-versus-rest AUROC from its support scores and average precision (AP), the stepwise precision–recall summary reported here as AUPRC. We calculate class metrics first and average the calculable classes equally, avoiding domination by frequent classes. A class with no positive or no negative GT region is NA, not zero. Metrics pool 63 GT assignments, not 63 independent predictions or ten TMA-level scores. Shared candidates repeat a score vector. Candidate bootstrapping preserves these groups but not within-TMA dependence, addressed separately by TMA resampling. Per-class positive counts range from one to nine (median five). AP varies with positive-class prevalence (Saito & Rehmsmeier, 2015). For the paired B–A Top-1 difference, we resample ten whole TMAs 10,000 times and recompute correct-label proportions, preserving shared candidates (Appendix D.4). AUROC and AP assess across-region score ranking, whereas Top-1 assesses the highest-scoring class within each region. The 0–100 support scores are model-reported values, not API log probabilities or calibrated probabilities.

4.5 Comparator families and denominators

The comparisons address three tasks. Eight image–text encoders test recognition on the 63 GT assignments. Their architecture and H&E rendering differ from GPT-6, so this comparison does not isolate the Cell Map effect. Proposal baselines test placement and boundary overlap against 75 GT regions, with mask budgets reported. Complete systems test localization with correct classification against all 75 GT regions, without GT-based candidate selection. GPT-6 B classifies all 340 MOSST masks under Full-12. Gemma supplies a separate H&E-only complete-pipeline control. These tasks have different denominators and do not form one leaderboard.

5 Results

The complete-system comparison tests automatic region localization and naming. Fixed-candidate experiments then distinguish recognizer choice from input choice. The clearest evidence for the contribution of Cell Map comes from A versus B, which keeps candidates, GPT-6, and the class list fixed.

5.1 Complete-system comparison

MOSST with GPT-6 classified all 340 selected masks from H&E and Cell Map under Full-12, without GT-based candidate pairing. It localized and correctly classified 37/75 GT regions at IoU $\geq$ .10 and 13/75 at IoU $\geq$ .50, the largest counts among the tested systems (Table 1). These comparisons describe the executed configurations, whose inputs and output budgets differ. The H&E-only Gemma control gave 28/75 correct outputs for MOSST and 6/75 for SAM2 automatic at IoU .10. This comparison holds recognition fixed but uses unequal mask counts. Comparing GPT-6 B with Gemma changes both model and input and cannot isolate the Cell Map contribution. The

Table 1: **Complete outputs against 75 GT regions.** Inputs refer to recognition, or joint segmentation and recognition for direct models. Both MOSST rows use molecularly guided proposals. *Masks* gives the number of outputs. Loc. and Correct use separate one-to-one matchings, with Correct requiring class agreement. Appendix Table 8 lists further configurations.

Pipeline	Recognition input	Masks	Loc. .10	Correct .10	Correct .50
SAM2 automatic + Gemma	H&E	294	19	6	3
SAM2 automatic + CONCH	H&E	294	19	7	1
Grounded MedSAM	H&E	120	5	5	1
LISA	H&E	11,703	11	5	1
GLaMM	H&E	134	5	5	1
MedPLIB	H&E	1,530	23	8	0
MedSAM3 LoRA	H&E	161	20	7	4
MOSST + Gemma	H&E	340	54	28	11
MOSST + GPT-6 (B)	H&E + Cell Map	340	**54**	**37**	**13**

candidate pool reaches 63/75 GT regions under the permissive accessibility rule, or 57/75 without

its complete-containment clause. Boundary agreement is stricter: 54/75 regions match at IoU .10 and 14/75 at IoU .50. MOSST does not lead every proposal-only metric. Appendix Table 9 reports the full comparison. Figure 5 maps all 40 TMA39 predictions back to H&E. Predictions outside the confidence-annotated GT regions remain unverified. Appendix Figure 14 distinguishes missing candidates from incorrect tissue labels.

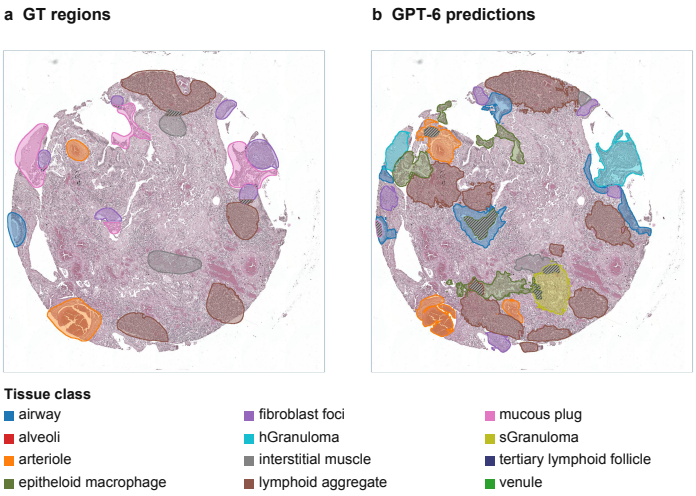

Figure 5: **Tissue labels mapped to TMA39.** Left, 19 GT regions. Right, 40 candidate masks classified by GPT-6 using H&E and Cell Map. Colors indicate tissue classes, and hatching marks overlapping predictions. Tissue outside the GT regions is unannotated.

5.2 Recognizer comparison on fixed candidates

On the same 56 candidates paired to 63 GT regions, H&E-only GPT-6 correctly classifies 33/63, versus 8–19/63 for eight image–text encoders (Table 2), and gives the largest AUROC and AP point estimates. Encoders use clean crops, whereas GPT-6 receives outlined views and a different scoring prompt. This compares recognition configurations, not architecture alone or autonomous localization. The input experiment below holds GPT-6 fixed.

Table 2: **H&E recognition on fixed candidates.** All rows use the same candidates, 63 GT assignments, and twelve tissue classes. Encoders average close and wider cosine scores. GPT-6 scores outlined views, so rendering and scoring also differ. TITAN uses its ROI adapter. Bold marks the largest point estimate.

Recognizer	H&E views	Top-1	AUROC	AP
PLIP	Clean	9/63	.5640	.1793
BiomedCLIP	Clean	12/63	.6951	.3384
QuiltNet	Clean	8/63	.6524	.2308
PathGen-CLIP-L	Clean	19/63	.7556	.3410
KEEP	Clean	14/63	.7767	.3249
CONCH	Clean	15/63	.7821	.4116
MUSK	Clean	19/63	.7468	.2972
TITAN (ROI)	Clean	9/63	.5866	.2184
GPT-6 (A)	Outlined	**33/63**	**.8650**	**.5231**

5.3 MOLECULAR INPUTS FOR TISSUE RECOGNITION

All conditions share 56 candidates paired by GT geometry to 63 regions, GPT-6 settings, and the twelve-class list. Adding Cell Map to H&E raises correct Top-1 labels from 33/63 (A) to 42/63 (B; Table 3). Ten GT assignments change from wrong to correct and one changes in reverse. Five TMAs improve, one declines, and four are unchanged. The gain is 14.3 percentage points, with an exploratory 95% TMA-bootstrap interval of 2.4–21.9 points, conditional on the recorded responses (Appendix D.4). Excluding six containment-only matches retains the nine-label gain (A: 28/57, B: 37/57).

Table 3: **Matched GPT-6 inputs for 63 accessible GT regions.** AUROC and AP are class-macro scores with exploratory 95% candidate-bootstrap intervals. FICTURE includes factor gene lists. B is the primary input. Bold marks the largest point estimate.

Input	Top-1	AUROC [95% CI]	AP [95% CI]
A: H&E	33/63	.8650 [.797, .908]	.5231 [.462, .663]
B: H&E + Cell Map	**42/63**	.9327 [.890, .961]	.6915 [.615, .792]
C: H&E + FICTURE	37/63	.9263 [.886, .954]	.6836 [.596, .780]
D: H&E + both maps	37/63	**.9362** [.899, .962]	**.7234** [.627, .828]
E: Cell Map only	27/63	.8794 [.815, .921]	.5972 [.501, .711]
F: H&E + composition text	35/63	.9190 [.873, .949]	.6300 [.551, .757]

H&E with FICTURE and factor-gene lists (C) reaches 37/63 without cell-type annotations. Cell Map alone (E) reaches 27/63. Adding FICTURE to B gives 37/63 (D), below B despite having the highest AUROC and AP point estimates. Paired D-minus-B ranking intervals include zero (Appendix Table 14), so FICTURE's added value alongside Cell Map remains uncertain. The GT-derived TMA-restricted class list reaches 50/63 but is an oracle diagnostic. A region-adapted LLMiniST-style composition-text control gives 35/63 with H&E, candidates, and GPT-6 fixed (Appendix E.4). B exceeds this text control by 11.1 percentage points (95% TMA-bootstrap interval 3.2–25.0). The AUROC and AP difference intervals include zero.

6 DISCUSSION AND LIMITATIONS

The controlled A–B comparison supports Cell Map and its legend as additions to H&E for tissue recognition in this cohort. The composition-text control also favors the tested Cell Map representation for Top-1 recognition. Differences in encoding and explicit counts mean that it does not isolate spatial arrangement alone. Complete-output results establish performance of the executed configurations, whose proposals, recognizers, and budgets differ. They do not isolate the molecular proposal contribution.

The cohort has ten lung TMAs, rare classes, and no external test set. One recorded response per candidate and condition does not establish stability across API calls. B correctly classifies 0/3 venules and 0/5 mucous plugs, versus 8/8 alveolar and 7/8 airway regions (Appendix Figure 16). Generalization and unannotated predictions require further expert assessment.

Annotation precision. Manual GT boundaries introduce uncertainty. The annotated airway in Appendix Figure 15 extends beyond the predicted contour. This disagreement alone does not identify the more accurate boundary. Low IoU may reflect annotation imprecision, segmentation error, or both, requiring independent boundary review.

7 CONCLUSION

MOSST links spatial molecular evidence to candidate regions and tissue classes. GPT-6 labels the complete candidate set without GT-based selection, and the matched-input experiment supports Cell Map for tissue recognition. Boundary agreement, unannotated predictions, and transfer beyond this lung cohort require further evaluation.

AI USE STATEMENT

AI tools assisted manuscript editing, code, and experiment preparation. The authors checked the analyses and claims and take responsibility for the paper.

REPRODUCIBILITY STATEMENT

The anonymous code supplement provides core functions, recorded prompts, derived scores, and overlap records for offline reproduction of recognition and complete-output metrics. We retain image hashes, model identifiers, and raw responses. The closed API may change, and fresh calls need not reproduce recorded responses. Raw responses, tissue images, and transcript data are not included.

REFERENCES

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and Sabine Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/tpami.2012.120.

Sidra Aleem, Fangyijie Wang, Mayug Maniparambil, Eric Arazo, Julia Dietlmeier, Kathleen Curran, Noel E. O'Connor, and Suzanne Little. Test-Time Adaptation with SaLIP: A Cascade of SAM and CLIP for Zero-shot Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5184–5193, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/DEF-AI-MIA/html/Aleem_Test-Time_Adaptation_with_SaLIP_A_Cascade_of_SAM_and_CLIP_CVPRW_2024_paper.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e0982cbc81401df3430ee1ff780dc7a2-Paper-Conference.pdf.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1), 2022. doi: 10.1038/s41467-022-29439-6.

Konstantin Hemker, Andrew H. Song, Cristina Almagro-Pérez, Guillaume Jaume, Sophia J. Wagner, Anurag Vaidya, Nikola Simidjievski, Mateja Jamnik, and Faisal Mahmood. Towards Spatial Transcriptomics-driven Pathology Foundation Models. arXiv preprint arXiv:2602.14177, 2026. URL https://arxiv.org/abs/2602.14177.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021. doi: 10.1038/s41592-021-01255-8.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Anglin Liu, Xu R. Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. MedSAM3: Delving into Segment Anything with Medical Concepts. *arXiv preprint arXiv:2511.19046*, 2025. URL https://arxiv.org/abs/2511.19046.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Computer Vision – ECCV 2024*, pp. 38–55. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72970-6_3.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1), 2023. doi: 10.1038/s41467-023-36796-3.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), 2024. doi: 10.1038/s41467-024-44824-z.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Yichen Si, ChangHee Lee, Yongha Hwang, Jeong H. Yun, Weiqiu Cheng, Chun-Seok Cho, Miguel Quiros, Asma Nusrat, Weizhou Zhang, Goo Jun, Sebastian Zöllner, Jun Hee Lee, and Hyun Min Kang. Ficture: scalable segmentation-free analysis of submicron-resolution spatial transcriptomics. *Nature Methods*, 21(10):1843–1854, 2024. doi: 10.1038/s41592-024-02415-2.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016. doi: 10.1126/science.aaf2403.

Huanhuan Wei, Xiao Luo, Hongyi Yu, Jinping Liang, Luning Yang, Lixing Lin, Alexandra Popa, and Xiting Yan. Identifying cellular niches in spatial transcriptomics: An investigation into the capabilities of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9275–9289, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.455. URL https://aclanthology.org/2025.acl-long.455/.

Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384, 2021. doi: 10.1038/s41587-021-00935-2.

A EVALUATION COHORT AND MEASURES

This section defines the evaluation units. It explains which GT regions have a candidate, how one candidate can be paired with several GT regions, and why localization and recognition use different denominators.

A.1 COHORT, GT REGIONS, AND FIXED EVALUATION UNITS

Table 4 lists the ten-TMA evaluation set, candidate counts, and recognition counts. The workflow and candidate-selection parameters were established on an earlier Visium HD dataset and held fixed for this Xenium evaluation. Each GT region is defined by one expert annotation. Each candidate is one proposed mask. A GT region is accessible only if a candidate meets the geometric rule below. Several GT regions can share one candidate, so the counts of GT regions, candidate masks, and model requests differ.

Table 4: Cohort accounting. Correct reports GPT-6 B under Full-12. The total class count is the union across TMAs. Unique inputs counts representative masks, each with H&E and Cell Map views.

TMA	Annotated classes	GT regions	Candidates	Accessible	Unique inputs	Correct
TMA07	4	8	40	4	4	3/4
TMA24	1	2	31	2	2	1/2
TMA29	4	7	25	7	6	3/7
TMA30	4	14	40	11	10	8/11
TMA31	3	3	40	3	3	2/3
TMA34	1	1	20	1	1	1/1
TMA36	4	8	24	7	5	4/7
TMA39	6	19	40	18	17	11/18
TMA41	4	8	40	7	5	6/7
TMA42	3	5	40	3	3	3/3
Total	12	75	340	63	56	42/63

Annotation scope. The cohort contains 75 GT regions. Because the annotation is confidence-based, unmarked tissue is not a confirmed negative and an unmatched prediction is not automatically a false positive.

Where GT information is used. Candidate masks and crop geometry are fixed before comparing recognizers. We use GT-region geometry to pair each GT region with a candidate for evaluation, but not to set that candidate's crop bounds. The resulting comparison tests recognition *given an available candidate*; it does not assess every output mask.

A.2 EVALUATION DEFINITIONS AND EDGE CASES

A.2.1 GEOMETRIC ELIGIBILITY AND REPRESENTATIVE SELECTION

A candidate and GT region can be paired if their intersection is at least 10% of the larger area, or if the smaller shape lies completely inside the larger one. We test complete containment by exact foreground-pixel equality, not an approximate 90% rule. If several candidates qualify for one GT region, we choose the one with the greatest intersection divided by the larger area. Candidate ID breaks ties. All recognizers use this same pairing. It is not one-to-one: one candidate can represent several GT regions.

Candidate accessibility is $63/75$, regardless of predicted labels. Recognition accuracy among accessible regions uses 63 as its denominator. A second summary retains all 75 GT regions, including the 12 without a candidate. Thus condition B's 42 correct labels appear as $42/63$ and $42/75$. The $42/75$ value still pairs candidates using GT geometry. It is not an evaluation of all 340 masks. Reusing one candidate prediction for several GT regions does not create additional model calls.

Assignment conventions in proposal and pipeline comparisons. For proposal benchmarking, masks are sorted by model score and then ID. The fixed pool retains its recorded order. At each IoU threshold, a mask is paired with the highest-IoU GT region not yet used. Mask average recall (AR) averages recall at IoU .50, .55, and so on through .95. The saved evaluator also checks it against the reference COCO implementation. Accessibility instead asks whether *any* candidate reaches each GT region and permits reuse. Mean best IoU takes each GT region's best overlap. These measures differ both in overlap threshold and in whether a candidate may be reused.

For complete-pipeline comparisons, localization and correctly classified counts are computed in separate one-to-one matchings. A localization edge requires the stated IoU threshold. A correctly classified edge additionally requires agreement between the predicted and GT tissue classes. Each matching maximizes the number of eligible pairs, using IoU to break ties. The two counts can therefore use different candidate–GT assignments. This differs from fixed-candidate recognition, where GT geometry selects one representative before its predicted class is evaluated. Counts are recall over the 75 confidence-annotated GT regions, not precision over unannotated tissue. One-to-one assignment limits reuse of instances, not overlap between their pixels. Distinct nested predictions can each match a separately annotated GT region when the geometric and class criteria are met. Missing annotations do not establish that an additional predicted structure is absent. The diagnostic table's near-duplicate count marks a mask when its IoU with an earlier kept mask is at least .90. This is separate from the .50 threshold used while selecting our candidate pool.

A.2.2 Class-wise ranking metrics

For class k, an eligible GT region is positive if its annotated class is k, and negative otherwise. Given positive scores s_p and negative scores s_n, the executed AUROC calculation is

$$\mathrm{AUROC}_k = \frac{1}{N_k^+ N_k^-} \sum_{p,n} \left[\mathbf{1}(s_p > s_n) + \tfrac{1}{2}\mathbf{1}(s_p = s_n) \right]. \tag{8}$$

Average precision is computed at distinct score thresholds, grouping tied scores together:

$$\mathrm{AP}_k = \sum_j (R_j - R_{j-1}) P_j. \tag{9}$$

We label this stepwise PR summary as AUPRC (AP). It is not trapezoidal integration of a PR curve. Macro AUROC and macro AP average equally across calculable classes. Total metrics pool GT regions across TMAs before calculating each class metric. They are not averages of per-TMA macro scores.

A.2.3 Why a metric can be NA

The saved evaluator requires at least one positive and one negative for both reported ranking metrics. Therefore a TMA containing only one class, such as the two accessible hGranuloma regions in TMA24, receives NA rather than zero. AP can be defined for some degenerate label sets under other conventions, but this report intentionally uses the same eligibility rule for both metrics. Restricted-arm ranking excludes rows where the class was not offered, which changes the negative set and can reduce the number of calculable classes from twelve to eleven. Missing scores for unoffered classes are not replaced with zeros.

Dependence and uncertainty. These summaries are descriptive. Reused predictions and observations from the same TMA are dependent. Appendix D.4 reports candidate-cluster bootstrap intervals and a ten-TMA resampling sensitivity check. Neither creates an independent external test cohort.

B Method and recognition-input details

This section follows the workflow from molecular maps to candidate masks and recognition requests. It records implementation details and provides enlarged examples.

B.1 PROPOSAL CONSTRUCTION AND SELECTION

From nine GeneMaps to spatial prompts. TMA39 contains nine expression-module score maps (Figure 7). Each map can produce several disconnected high-score regions, rather than a single mask. Within the tissue support, the original threshold triplet retains the top 1%, 2.5%, and 5% of scores. The expanded triplet uses 2.5%, 5%, and 10%. The highest threshold defines a core and the broadest defines its support. Binary closing and hole filling remove small gaps, and cores smaller than two grid bins are discarded. When several cores share connected support, each support bin is assigned to its nearest core. Regions sharing at least three edge-adjacent bins are merged. Each resulting connected region yields a tight box and its highest-score point in the common coordinate frame of the H&E image and FICTURE map. The region is a prompt source. Only the subsequent SAM3 call produces a candidate mask. The nine GeneMaps are distinct from the fourteen FICTURE factors and the 24 annotated cell types.

Independent H&E supplements. The third branch detects H&E tissue in locations with little FICTURE signal without using candidate masks or GT regions. On the saved common image grid, a 64-pixel block qualifies when at least 35% of its pixels meet the H&E tissue criterion and at most 8% carry FICTURE RGB signal (channel maximum at least 8). Binary closing with a 3×3 block structure precedes connected-component labeling. Components of at least six blocks receive a tight box expanded by 10% on each side. A positive point is chosen inside H&E tissue, preferring an interior, centrally located stained pixel without FICTURE signal. SAM3 runs on the corresponding H&E crop with the box and point, then the resulting binary mask is returned to whole-core coordinates and clipped to the core. Thus an H&E supplement is a mask from an H&E crop, not another H&E image. TMA39 yielded three selected gap prompts, R1, R2, and R4, and three retained masks. A component below six blocks (R3) was not prompted. Figure 6 enlarges R4. The detections are not tissue diagnoses.

SAM3 inference. We use the SAM3-base checkpoint with SHA-256 `9999e2341ceef5e136daa386eecb55cb414446a00ac2b55eb2dfd2f7c3cf8c9e`. The interactive predictor receives native-pixel boxes and one positive point, with internal coordinate normalization and `multimask_output=False`. Each call returns one binary mask at the input resolution and its predicted-IoU score. The decoder's internal dynamic-multimask stability fallback remains enabled, but only its single returned mask enters the pool. Loading receipts record 1,465 mapped state keys, including 309 interactive-predictor state entries, with no missing or unexpected keys. Inference uses evaluation mode and bfloat16 on an H200 GPU.

Prompt stability. For each candidate, we rerun its source image with the box scaled about its center by 0.9 and 1.1, keeping the positive point inside. A third variant shifts the point by 5% of the shorter box side, rounded and clipped to 1–32 pixels. We try right, left, down, then up and take the first point inside both the original mask and box. If none qualifies, only the two box variants are used. Perturbed masks receive the same foreground clipping as the original. Stability is their mean IoU with the original mask. It uses no GT annotation.

Selection under a spatial budget. Original and expanded proposal pools are combined with the independent histology branch. The selector assigns average ranks to tied SAM3 scores and stability values, then divides each rank by the pool size. Higher values receive higher ranks. The non-FICTURE offset in Eq. 4 gives eligible H&E masks priority before FICTURE masks, reserving their morphological alternatives in the finite pool. This is a fixed development heuristic, not a learned or optimal weighting. Equal selection scores retain the earlier candidate in the stored pool order. It requires at least 25% new area on a 512-pixel working grid and removes remaining masks with pairwise IoU at least 0.50 after each selection. Per-TMA candidate limits follow the fixed development configuration and never exceed 40. GT overlap is used only after selection for evaluation. The selected TMA39 pool contains 40 masks: 21 from the original pool and 19 from the expanded pool. Of the selected masks, 36 are SAM3 masks on FICTURE from GeneMap prompts, one is a SAM3 mask on H&E from a GeneMap prompt, and three are H&E supplement masks from the low-FICTURE-signal branch. These are counts *after* selection. A "selected FICTURE candidate" means one of the 36 selected masks, not a FICTURE factor map or a raw SAM3 call. The full cohort contains 340 selected candidates. Figure 7 traces one module from its map to the retained masks.

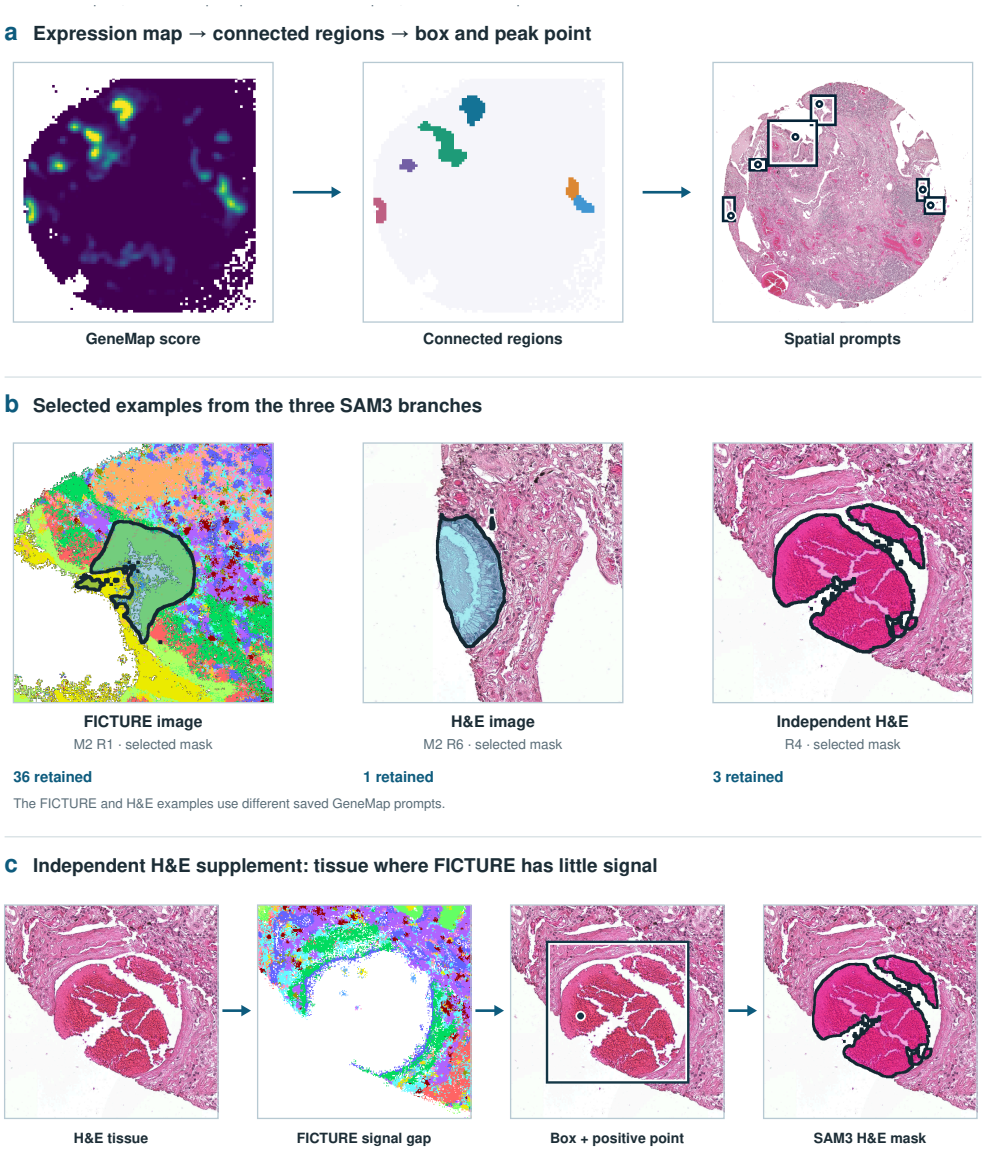

Figure 6: **Three proposal branches.** (a) GeneMap regions define box-and-point prompts. (b) SAM3 generates masks from FICTURE and H&E. The examples use different GeneMap regions. (c) An independent H&E proposal at a location with weak FICTURE signal. For TMA39, selection retains 36 FICTURE masks, one GeneMap-prompted H&E mask, and three independent H&E masks.

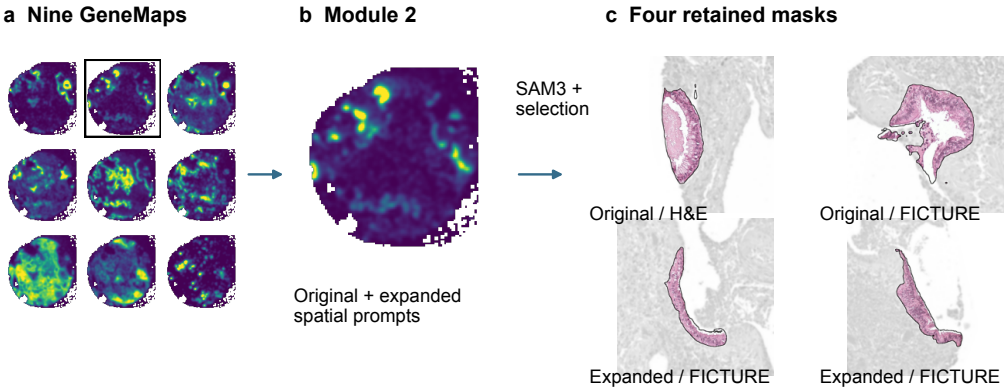

Figure 7: **Masks derived from one GeneMap.** Four selected TMA39 masks originate from module 2, using original or expanded prompts on FICTURE and H&E. Outlines mark candidate boundaries. Colors distinguish masks rather than tissue classes.



Figure 8: **Selected TMA39 candidates.** The pool contains 36 FICTURE masks, one GeneMap-prompted H&E mask, and three independent H&E masks. The color key distinguishes GeneMap modules and H&E proposal sources. Tissue classes are assigned later.

B.2 COORDINATE ALIGNMENT AND VISUAL INPUT CONSTRUCTION

FICTURE and Cell Map provenance. The FICTURE map uses the saved 14-factor, seed-1 model. At each resolved location the displayed color represents its assigned top factor. The exact factor-ID order uses the first fourteen colors of the official DEFAULT palette. Black in the inference image indicates no assigned factor. The no-signal region is shown white only in the paper figure. Factor-gene lists supplied to GPT-6 contain the twenty highest-loading genes per factor and do not assign cell-type names. The Cell Map uses biologist-provided `celltypeSimple` annotations without relabeling cells. These annotations describe individual cell identities, whereas the biologist-annotated tissue-region GT records regional boundaries and tissue classes. For TMA39, the 35,987 centers and labels were extracted from `nucXensFiltered.Sept5.Seurat.Robj`, using `coordinateX` and `coordinateY`. This experiment evaluates the supplied annotation representation, not the accuracy of the upstream cell-type annotations. After the stored transform each center is plotted with one of the 24 fixed label colors in the Cell Map legend of Figure 9. These center dots do not describe cell boundaries. Figure 9 shows both source paths and their distinct legends.

B.2.1 COORDINATE TRANSFORMATION

Let $x_i \in \mathbb{R}^2$ denote a source cell coordinate. The executed renderer applies the stored affine transform and origin correction, then scales to the common map size:

$$x_i^{\text{image}} = \big(A_{2\times 2} x_i + b - o\big) \odot (w/w_0, h/h_0). \tag{10}$$

The transform is specific to a TMA. The implementation checks that the H&E dimensions agree with the map dimensions and that each selected mask lies inside the stored core ROI. Example images in the manuscript use these accepted transforms. Quantitative coordinate-mapping error was not measured.

B.2.2 CANDIDATE-CENTERED CROPS

For a mask bounding box with width u and height v, the square crop side is

$$s = \min\{\max(s_{\min}, \text{round}(\alpha \max(u, v))), w, h\}. \tag{11}$$

The crop is centered at the bounding-box center and shifted, if necessary, to remain inside the image. Close and wider crops use $(\alpha, s_{\min}) = (1.6, 600)$ and $(3.0, 1100)$, respectively. These are image-pixel settings, not physical micrometer scales. The rendered maximum side lengths are 900 and 1024 pixels. Corresponding crops use identical bounds in all supplied image types. The saved request files call the wider crop the "medium" view. The image and crop bounds are unchanged.

B.2.3 CELL AND TARGET RENDERING

Annotated cell centers are rendered as colored circles with radius 3 pixels in each rendered crop. The 24-color mapping is fixed (Figure 9). Cell counts are not included in the primary prompt. A sparse map does not establish tissue acellularity. The renderer adds an external white halo and black contour using 13-pixel and 7-pixel maximum filters on the resized binary mask. The close H&E view uses the fixed target-emphasis helper, while the wider view retains surrounding histology.

Each submitted image contains the close panel at left and wider panel at right, separated by 20 pixels, with a 44-pixel title strip. The A–E requests present H&E first when provided, followed by the available molecular maps. The per-candidate manifest records the crop bounds, mask hash, and image hashes. The source-image colors are preserved in the qualitative figures.

B.3 TISSUE VOCABULARY AND INFERENCE CONTRACT

B.3.1 TWELVE TISSUE CLASSES

The following support criteria and limitations are transcribed from the executed full-vocabulary prompt. Dataset names such as `sGranuloma`, `hGranuloma`, and `epitheloid_macrophage` are preserved for consistency. They should not be expanded into a clinical diagnosis beyond the supplied evidence.

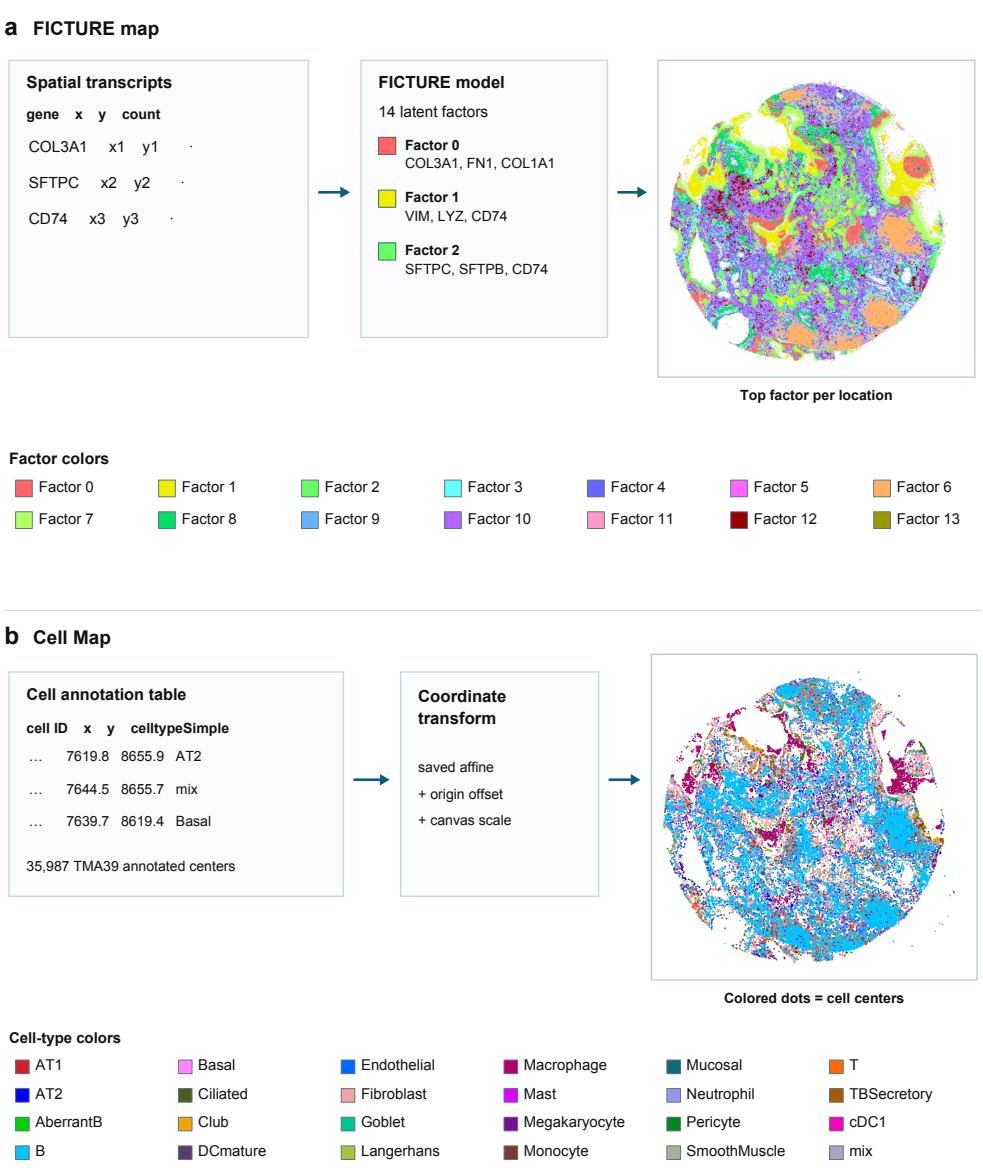

Figure 9: **FICTURE and Cell Map construction.** (a) Spatial transcripts are assigned to FICTURE factors, shown with the 14-factor color key and example genes. Transcript rows are schematic. (b) Annotated cell centers are plotted with the 24-cell-type color key. FICTURE factor colors do not encode cell types. White areas replace unassigned background for display only.

airway *Supporting evidence.* Airway epithelial lining and its associated wall, with a luminal relationship visible in the target or directly connected context. *Limits.* Nearby airway epithelium does not make luminal contents an airway wall. A partial wall can be supported without a complete circular lumen.

arteriole *Supporting evidence.* A vascular structure with an endothelial-luminal relationship and an organized circumferential smooth-muscle wall visible in the target or directly connected context. *Limits.* Blood or endothelial annotations alone do not establish an arterial subtype. If the mask contains only luminal material, describe it separately and report vessel-subtype limitations.

venule *Supporting evidence.* A vascular structure with a discernible thin wall relative to its lumen and limited organized circumferential smooth muscle. *Limits.* Do not choose venule simply because an arterial wall is not visible. An excluded, cropped, collapsed or unresolved wall is missing evidence, not positive venous evidence.

alveoli *Supporting evidence.* Alveolar spaces bounded by thin septa, supported by alveolar epithelial and capillary organization where visible. *Limits.* An empty space or AT1/AT2 annotations alone does not establish an alveolar structure.

fibroblast_foci *Supporting evidence.* For this dataset, a localized fibroblast-rich or collagen-rich tissue focus supported by its cellular or matrix appearance. *Limits.* Fibroblast annotations alone do not distinguish a focus from ordinary stroma or fibrosis in another structure. Do not infer a clinical UIP/IPF diagnosis from this dataset label.

lymphoid_aggregate *Supporting evidence.* A localized dense lymphoid collection supported by H&E cellularity and the distribution of lymphoid annotations. *Limits.* Scattered lymphocytes do not establish an aggregate. Distinguish a general aggregate from a demonstrably organized follicle.

mucous_plug *Supporting evidence.* Mucus-like material occupying an airway lumen, with composition and a convincing airway-luminal relationship considered together. Inflammatory cells and debris may be present. *Limits.* Sparse Cell Map dots, pink staining, fissures or missing expression signal alone do not establish mucus. Consider blood/fibrin-like material, collagen and necrotic debris; if composition or airway continuity is unresolved, report that.

interstitial_muscle *Supporting evidence.* Aligned viable smooth-muscle bundles or fascicles in tissue, supported by morphology and smooth-muscle annotations. *Limits.* Smooth-muscle annotations alone do not separate interstitial bundles from a vessel or airway wall. Acellular eosinophilic material is not sufficient evidence of viable muscle.

sGranuloma *Supporting evidence.* The dataset cellular-granuloma category: a compact or spatially coherent aggregate of epithelioid histiocytes, with or without multinucleated giant cells and other inflammatory cells. *Limits.* A complete capsule, central necrosis or lymphoid rim is not required. Macrophage enrichment alone is insufficient; missing architecture at a cropped boundary should be reported as uncertain rather than treated as absent.

hGranuloma *Supporting evidence.* The dataset hyalinized-granuloma category: a hyalinized or fibrotic nodular lesion for which residual granulomatous features or visible lesion organization support that interpretation. *Limits.* Dense, whorled or hyalinized collagen alone is not specific enough to settle this dataset label. Distinguish a nonspecific collagen-rich focus and luminal material. Do not equate the label automatically with the separate named disease pulmonary hyalinizing granuloma.

tertiary_lymphoid_follicle *Supporting evidence.* A lymphoid collection with visible follicular organization, supported by spatially organized B-cell and associated T-cell distributions; a germinal-center-like pattern can provide additional support when resolvable. *Limits.* B-cell abundance alone is insufficient. An unresolved follicular center does not prove its absence. Do not invent immunostains, follicular dendritic-cell networks or other features not supplied.

epitheloid_macrophage *Supporting evidence.* The dataset macrophage-rich-region category: histiocyte/epithelioid morphology supported by a localized macrophage-rich cellular distribution. *Limits.* Macrophage annotations alone do not establish epithelioid morphology. This category overlaps biologically with granulomas; compare spatial organization and report ambiguity when the two cannot be separated.

B.3.2 Model and output contract

The saved main experiment identifies the model as `openai/gpt-6-astra-pro`, with reasoning effort `medium`, high image detail, a maximum output length of 8192 tokens, and JSON-object output. The prompt requests one independent 0–100 support score for every offered class, plus target content, modality-specific observations, cross-modal agreement, conflicts, missing evidence, and a comparison between the leading two classes. These are support ratings, not calibrated probabilities. The final predicted label is checked against the largest offered-class score. The saved requests do not explicitly set temperature, top-p, or a sampling seed. Each input condition uses one recorded response per unique candidate, so the comparison does not estimate variability across repeated API calls.

The primary arm offers all twelve classes for every TMA. The restricted arm uses the classes present in that TMA's GT regions and therefore receives GT-derived vocabulary information. This arm uses a separate VLM prompt and inference. It is not implemented by masking the main VLM output. For encoder baselines, the restricted diagnostic masks the saved twelve-class scores without re-encoding the images. The two mechanisms must remain explicit in interpretation.

B.4 Recognition prompt and input differences

Each condition is a separate GPT-6 request for the same selected candidate. The fixed tissue definitions, target outline, crop bounds, scoring rule, and JSON response contract are shared. Table 5 lists every input change. A missing modality is marked unavailable, never treated as negative evidence. The complete example prompts, including all twelve definitions and the JSON schema, are preserved as `prompt_A_HE_only.txt`, `prompt_B_HE_CellMap24.txt`, `prompt_C_HE_FICTURE.txt`, `prompt_D_HE_CellMap24_FICTURE.txt`, and `prompt_E_CellMap_only.txt` in the source package. The Cell Map-only E request uses the same definitions and decision rules with H&E absent.

Table 5: **Recognition input contract.** Each image file contains close and wider views of the same outlined candidate. "Cell legend" maps 24 display colors to annotated cell types. "Factor genes" lists the top 20 loading genes for each displayed FICTURE factor; it supplies no cell-type names.

Condition	Image files	Additional text
A	H&E	Shared tissue definitions
B (main)	H&E, Cell Map	Shared definitions, cell legend
C	H&E, FICTURE	Shared definitions, factor IDs and genes
D	H&E, Cell Map, FICTURE	Shared definitions, cell legend, factor genes
E	Cell Map	Shared definitions, cell legend

B.4.1 Complete FICTURE factor gene key

The diagram in Figure 4 prints two genes per factor so that the colors and factor identities remain legible. The TMA39 example's executed GPT-6 requests included the following complete factor-to-gene list, copied from their saved prompt text.

Table 6: Complete FICTURE factor colors and top-20 gene lists for the TMA39 example supplied to GPT-6 in its archived C and D requests. Factors are expression components, not cell-type labels.

Factor	sRGB color	Top 20 genes, in prompt order
0	#ff6565	COL3A1, FN1, COL1A1, COL1A2, SPARC, LUM, VIM, POSTN, DCN, TIMP1, COL5A1, FSTL1, C3, SERPINF1, TGFBI, CTSK, THBS2, SPP1, MMP14, VCAN
1	#edee00	VIM, LYZ, CD74, HLA-DRA, FN1, SPP1, CD68, CHIT1, TGFBI, FCER1G, MRC1, MARCO, MS4A7, HLA-DRB1, CD44, C1QC, MMP9, LIPA, CTSK, LGMN
2	#65ff65	SFTPC, SFTPB, CD74, SFTPD, NAPSA, RNASE1, HOPX, PGC, LRRK2, C3, DMBT1, HLA-DRA, LAMP3, SOD2, SCGB3A2, CD44, AREG, MGST1, ATF4, EPAS1
3	#65feff	DCN, LUM, C3, VIM, FN1, PTGDS, C7, COL1A2, MGP, SPARCL1, SERPINF1, FSTL1, SFRP2, IGFBP7, SOD2, INMT, COL3A1, CCN2, CFH, SPARC
4	#6565ff	VIM, VWF, EPAS1, SPARCL1, IGFBP7, PECAM1, COL4A1, SPARC, POSTN, CCN2, CD74, HSPG2, SPRY1, ADAMTS9, CAV1, MMRN1, SOD2, PLVAP, TGFBR2, MGP
5	#ff65fe	EPAS1, CAV1, PECAM1, SPARC, AGER, VIM, FN1, FCN3, HSPA1A, BMPR2, BTG2, CD74, HOPX, ITGB1, DCN, TGFBR2, RNASE1, COL4A1, INMT, ATF4
6	#ffb265	CD74, HLA-DRA, HLA-DRB1, VIM, HLA-DQA1, CXCR4, CXCL13, MS4A1, CD44, HLA-DQB1, PTPRC, LYZ, CD79A, ACTG2, SOD2, C3, MARCKS, HLA-DRB5, LTF, NFKBIA
7	#b2ff65	SFTPB, SCGB1A1, SCGB3A2, CD74, RNASE1, CTSE, EPCAM, C3, KRT8, BPIFB1, MUC5B, HLA-DRA, CDH1, CCND2, KRT18, HOPX, CCN2, DMBT1, MGST1, AGR3
8	#00df5f	FN1, IGFBP5, MYH11, CCN2, ACTA2, MYLK, VIM, DES, SPARCL1, ELN, MYL9, IGFBP7, PALLD, ACTG2, SPARC, ITGB1, FSTL1, MGP, LTBP1, THBS1
9	#65b2ff	THBS1, SOD2, FOS, S100A9, CCN2, LYZ, TIMP1, VIM, AREG, BTG2, VEGFA, IGHD, LITAF, CCL2, FOSL2, MYC, CXCL8, IL6, HIF1A, NFKBIA
10	#b265ff	IGHG2, VIM, BTG2, ATF4, PECAM1, MZB1, CD79A, CD74, SSR3, DCN, IGFBP5, SEC11C, C3, CXCR4, COL3A1, SFRP2, ITGA6, IGHM, MERTK, CD44
11	#ff99cc	VIM, IL7R, CXCR4, PTPRC, CD8A, CCL5, CD74, CD44, CD2, CD3E, CCL4, DCN, FN1, SPARC, GZMA, HLA-DRA, COL3A1, GZMK, SPOCK2, CD3G
12	#990000	IGHA2, VIM, BTG2, ATF4, MZB1, CD74, PECAM1, SSR3, SEC11C, CD79A, MANF, COL3A1, IGHD, IGHM, DCN, ITGA6, AREG, SFRP2, CD44, CXCR4
13	#999900	IGHM, TPSB2, CPA3, TPSAB1, VIM, KIT, AREG, CD44, NFKBIA, CD74, HPGD, ID2, PLIN2, PTGS1, FCER1G, ATF4, FOSL2, BTG2, VEGFA, CSF1

The common instruction first asks the model to describe tissue or material *inside* the outline, use the wider view only for connected structures, compare the leading class with its closest alternative, and state missing or conflicting evidence. It then requests an independent 0–100 support score for every offered class. Scores need not sum to 100 and are not calibrated probabilities. The predicted class must have the highest score, with a fixed class-order tie break. This short description is for reading. The saved text files, rather than this paraphrase, define the executed requests. The TMA-restricted B control reuses B images but replaces the full twelve-class list with classes present in that TMA's GT regions. Because this list uses GT information, it is a diagnostic control rather than a deployable variant of the main workflow.

C BASELINE IMPLEMENTATION

These details specify the tested recognizers and medical segmentation systems. Input differences remain explicit because the comparisons do not isolate architecture alone.

C.1 BASELINE IMPLEMENTATION AND TEXT INPUTS

Table 7 identifies the eight encoder revisions recorded in the model registry. All eight evaluated encoders process the same clean H&E close and wider views, using the fixed geometric bounds. They do not receive Cell Map, a numerical cell-composition vector, or the target-outline rendering used by the main generative VLM. For normalized image features $z_{j,v}$ and text features t_k, the class score is

$$s_{jk} = \tfrac{1}{2}\left(z_{j,\text{close}}^{\top} t_k + z_{j,\text{wider}}^{\top} t_k\right). \tag{12}$$

No softmax is applied before ranking examples for AUROC or AP. In a Top-1 tie, encoders select the first class in the fixed class order. The reported primary score is the mean of the close and wider cosine similarities. No best-view selection is applied.

Table 7: Checkpoint revisions for the eight encoder runs. Full revision identifiers are retained in the model registry. TITAN is evaluated through its ROI adapter.

Encoder	Checkpoint revision (prefix)	Valid candidates
PLIP	`67ade53ddd32`	56
BiomedCLIP	`9f341de24bfb`	56
QuiltNet	`8ce77289ce35`	56
PathGen-CLIP-L	`feb8cbf4ff55`	56
KEEP	`28a25d95cc6b`	56
CONCH	`f9ca9f877171`	56
MUSK	`de1ffed28608`	56
TITAN (ROI)	`dac6773d9961`	56

PLIP uses its released Hugging Face processor and CLIP encoder. BiomedCLIP and QuiltNet use released OpenCLIP configurations and weights. PathGen-CLIP-L is loaded as ViT-L/14 with QuickGELU. Image size is recovered from its visual positional embeddings and the state dictionary is loaded strictly. KEEP uses bicubic resize to 224 pixels, center crop, and ImageNet mean/standard-deviation normalization. Text lengths are checked against each encoder's context limit. The recorded common runtime includes PyTorch 2.9.1, TorchVision 0.24.1+cu128, Transformers 4.49.0, and timm 1.0.15. These implementation details describe the saved runs, not newly executed comparisons.

C.1.1 ENCODER TEXT CONSTRUCTION

Each encoder class description used the template "An H&E histopathology image of [readable tissue name]. [supporting-evidence sentence]." The class-specific text follows the supporting-evidence language for the matching tissue class above. The readable names for the dataset-specific labels were *fibroblast-rich tissue focus*, *cellular granuloma*, *hyalinized granuloma*, and *macrophage-rich region*. Other labels use their displayed tissue names. These descriptions were fixed across models and TMAs. The exact twelve strings are supplied in `encoder_class_descriptions.txt` with the manuscript source.

Local Gemma configuration. The saved runners use Gemma-4-31B-Instruct with bfloat16 inference and sampling disabled. The complete-pipeline H&E control allows at most 2,048 new tokens per request. The fixed-candidate input-presentation controls allow 8,192. These are separate protocols.

C.2 MEDICAL SEGMENTATION BASELINES

These comparisons ask whether a medical segmenter can directly recover named tissue regions from H&E and class text. Every method processes all ten TMAs using the same twelve tissue phrases, without GT boxes, points, or masks. All returned regions are evaluated. Their natural output counts are not artificially equalized. Matching follows the maximum-cardinality one-to-one protocol of Table 8, with all 75 GT regions retained in the denominator.

Models and prompts. Grounded MedSAM combines Grounding DINO (Liu et al., 2024) with the released MedSAM ViT-B checkpoint (Ma et al., 2024). The detector supplies boxes and their text-query labels. MedSAM supplies masks, not class predictions. Detection and text thresholds are .30 and .25, and the mask threshold is .50. MedSAM3 (Liu et al., 2025) uses the released v1 LoRA adapter with its SAM3 base and official inference implementation, at resolution 1008, confidence threshold .50, and box-NMS threshold .50. The historical SAM3 checkpoint row instead uses the pre-existing project checkpoint, whose training provenance is not established by this experiment. It is a checkpoint-specific diagnostic, not a second reproduction of the published MedSAM3 model.

The shared text queries are airway, arteriole, venule, alveoli, fibroblast focus, lymphoid aggregate, mucous plug, smooth muscle, cellular granuloma, hyalinized granuloma, tertiary lymphoid follicle, and macrophage-rich region. Each maps deterministically to its corresponding study class. No molecular map or cohort-specific fitting is added to these runs.

Direct language-guided baselines. LISA uses `xinlai/LISA-13B-llama2-v1` (revision `b89000be11ad`), GLaMM uses `MBZUAI/GLaMM-FullScope` (`95a4ab0c9ab7`), and MedPLIB uses `Huangxs/MedPLIB-7b-2e` (`4720f90b5b38`). Each receives the whole H&E image and twelve separate class queries requesting all regions of that class. Inference uses bfloat16, sampling disabled, and at most 512 new tokens per query. The segmentation input has longest side 1024 for LISA/GLaMM and 256 for MedPLIB. Native mask logits are thresholded at zero and split into eight-connected components without size filtering. This adapter produces region outputs from class masks without GT prompts.

MedSAM execution checks. All thirty TMA outputs were checked for unique identifiers, H&E and FICTURE input hashes, mask dimensions, areas, finite scores, and class membership. Per-image loading receipts confirm that all 916 released LoRA tensors match the supplied checkpoint and that adapter modules execute during inference. The fixed input protocol is identical across the ten runs. These checks establish correct execution, not biological validity of every predicted region. Because experts annotated only selected high-confidence regions, unmatched predictions are not automatically treated as false positives.

D EXTENDED RESULTS AND UNCERTAINTY

The results follow the main comparisons: complete systems, proposal geometry, and recognition on fixed candidates. Class-wise results and statistical sensitivity analyses follow their corresponding comparison.

D.1 COMPLETE-SYSTEM OUTPUTS

GPT-6 B classified all 340 MOSST candidates from ten TMAs using H&E and Cell Map close and wider views with the same twelve-class prompt. We reused 79 valid responses after checking model settings, prompt text, and image hashes, and obtained 261 additional responses. Every candidate has a valid twelve-class score vector and a recorded Top-1 label. The source file `gpt6_all340_predictions.csv` lists all candidate IDs, labels, and scores. One-to-one matching gives 37/75 correctly classified GT regions at IoU .10 and 13/75 at IoU .50. No GT geometry selects candidates for this full-output evaluation.

The separate H&E-only Gemma control holds the recognizer and prompt fixed across MOSST and SAM2 automatic, yielding 28/75 versus 6/75 correct at IoU .10 and 11/75 versus 3/75 at IoU .50. The candidate sets contain 340 and 294 masks. Comparing GPT-6 B with this control changes

both model and recognition input. The fixed-candidate GPT-6 result of 42/63 remains a separate GT-paired experiment.

Other executed systems use different checkpoints, prompting, or training, and therefore provide separate context rather than a swap-in comparison. MedSAM3 LoRA correctly names 7/75 GT regions at IoU .10 and 4/75 at IoU .50. Grounded MedSAM reaches 5/75 and 1/75, and MedPLIB reaches 8/75 and 0/75. These numbers describe the tested implementations, not the best possible adaptation of their architectures (Appendix Table 8 and Appendix C.2).

Table 8: **Complete-output comparison.** This table includes additional configurations beyond the main-text table. Loc. and Correct use separate maximum-cardinality one-to-one matchings. Loc. uses geometry alone. Correct additionally requires class agreement. All counts are out of 75 GT regions using one-to-one matching. Both SAM2 + Gemma and MOSST + Gemma recognize from H&E only. MOSST + GPT-6 uses H&E and Cell Map under Full-12. "Regions" is the number of masks returned, not an accuracy score. The shaded row is MOSST with GPT-6. Bold marks the largest matched count. This evaluates all outputs, unlike Table 2.

Pipeline	Regions	Loc. .10	Correct .10	Correct .50
SAM3, text	26	5	0	0
Grounded SAM2	120	5	5	1
Grounded MedSAM	120	5	5	1
SAM2 + CONCH	294	19	7	1
SAM2 + Gemma	294	19	6	3
LISA	11,703	11	5	1
GLaMM	134	5	5	1
MedPLIB	1,530	23	8	0
MedSAM3 LoRA	161	20	7	4
Gemma localization + SAM2	13	6	4	0
MOSST + Gemma	340	54	28	11
MOSST + GPT-6 (B)	340	**54**	**37**	**13**

D.1.1 Medical-model output diagnostics

The released MedSAM3 adapter obtains seven correctly classified GT regions at IoU .10 and four at .50. These outcomes demonstrate the difficulty of this tissue vocabulary under direct prompting. They do not establish that medical adaptation is ineffective on its intended tasks. The historical checkpoint has different weights and inference settings, so its contrast with the released adapter is not a controlled fine-tuning ablation.

Grounded MedSAM produces twelve near-whole-core masks per TMA. All 66 within-TMA mask pairs have IoU at least .90, despite their different query labels. Its five correct IoU-.10 matches become 29 under the permissive containment rule, illustrating how broad masks can inflate that criterion without delineating individual structures. The corresponding permissive counts are nine for the historical checkpoint and eleven for released MedSAM3. These pair counts measure redundancy and are distinct from the sequential mask-removal counts in Table 10.

D.2 Proposal overlap and boundary agreement

Before tissue naming, the selected MOSST masks make 63/75 GT regions accessible under the permissive rule in Eq. 7. One-to-one matching yields 54/75 at IoU $\geq$ .10 and 14/75 at IoU $\geq$ .50 (Table 9). Watershed reaches 59/75 at IoU .10 with 400 masks, while the quality-ranked pool reaches 20/75 at IoU .50 with 397 masks. MOSST therefore does not dominate every geometric criterion. Six GT regions pass only through containment: three contain small candidates and three lie inside one broad candidate. Removing containment reduces accessibility to 57/75. On these 57 GT regions, with representative candidates unchanged, correct Top-1 counts are 28 (A), 37 (B), 34 (C), 34 (D), and 25 (E). The B-minus-A gain remains nine labels. IAMSAM reaches 62/75 regions under the permissive access rule with 108 masks, but saved whole-core examples include broad masks spanning unrelated tissue (Appendix Figure 17). These counts and examples show why

finding an overlapping region, matching a boundary, and producing a useful candidate are distinct assessments.

Table 9: **How many GT regions do candidate masks reach?** Each overlap count is out of 75 GT regions, with at most 40 masks returned per TMA. The two IoU columns pair candidate masks and GT regions one-to-one. "Access." counts a GT region if any mask meets Eq. 7, allowing reuse. SLIC and SLICO (Achanta et al., 2012) each collapsed to one mask per TMA in this execution; their outputs are recorded in Table 10. The shaded row is MOSST; bold marks the largest count in each metric column.

Proposal method	Masks	IoU $\geq$ .50	IoU $\geq$ .10	Access.
Watershed	400	6	**59**	**68**
SAM3, grid	400	15	50	57
SAM3, multiscale	400	14	36	40
SAM2, automatic	294	9	19	53
IAMSAM	108	11	18	62
Expression components	288	13	42	50
Quality-ranked pool	397	**20**	47	55
Ours, fixed selection	340	14	54	63

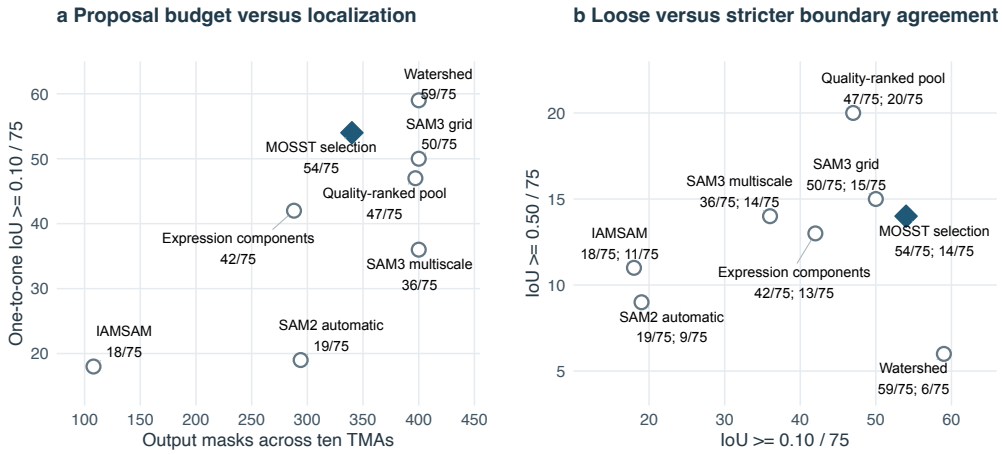

Figure 10: **Proposal count and boundary overlap.** (a) Total masks versus one-to-one GT matches at IoU $\geq$ .10. (b) Matches at IoU $\geq$ .10 versus IoU $\geq$ .50. Each point represents one proposal policy, capped at 40 masks per TMA and evaluated against 75 GT regions across ten TMAs. The blue diamond denotes MOSST.

Table 10: Additional proposal diagnostics at the same per-TMA cap as Table 9. Mask AR averages recall at IoU thresholds .50, .55, ..., .95. Mean best IoU averages each GT region's best overlap. Near duplicates counts masks removed by a sequential IoU-.90 deduplication pass; the SLIC/SLICO execution collapsed to one mask per TMA.

Method	Mask AR	Mean best IoU	Near duplicates
SLICO, multiscale	0.0107	0.0395	0
Watershed, grid	0.0147	0.2393	0
SAM3, grid	0.0747	0.2678	0
SAM3, multiscale	0.0707	0.2160	0
SAM2, automatic	0.0467	0.1327	32
IAMSAM, automatic	0.0600	0.1469	3
SLIC, multiscale	0.0107	0.0395	0
Expression components	0.0560	0.2470	0
Quality-ranked pool	0.0773	0.2843	0
Ours, fixed pool	0.0533	0.2957	0

D.3 Fixed-candidate recognition by class and TMA

The aggregate encoder comparison is reported in main-text Table 2. The following tables describe heterogeneity within the GPT-6 recognition experiment. Table 11 reports B by class. Restricted scores use a different negative population and are not a matched-input gain over Full-12. Table 12 reports the recorded A–D breakdown by TMA.

Table 11: Current GPT-6 B class-wise ranking results. N^+ counts positive GT regions among the 63 accessible regions. AP is the AUPRC summary. Full-12 has twelve eligible classes; Restricted excludes hGranuloma because it has no negatives.

Class	N^+	Full AUROC	Full AP	Restricted AUROC	Restricted AP
airway	8	0.9875	0.9464	0.9661	0.9306
alveoli	8	1.0000	1.0000	0.9861	0.9659
arteriole	8	0.9750	0.7888	0.9589	0.7644
epitheloid_macrophage	3	0.8389	0.1799	0.7500	0.8333
fibroblast_foci	9	0.8405	0.6490	0.8140	0.7180
hGranuloma	3	0.9750	0.6984	N/A	N/A
interstitial_muscle	1	1.0000	1.0000	1.0000	1.0000
lymphoid_aggregate	5	0.9828	0.7962	1.0000	1.0000
mucous_plug	5	0.8379	0.2733	0.8600	0.7056
sGranuloma	8	0.9545	0.8024	0.9875	0.9861
tertiary_lymphoid_follicle	2	1.0000	1.0000	1.0000	1.0000
venule	3	0.8000	0.1639	0.6667	0.3869

Table 12: Per-TMA Top-1 counts for the A–D protocol. All arms share the same accessible GT regions; reused candidates retain identical predictions. This table reports per-TMA counts for the same input conditions as Table 3.

TMA	Accessible GT	A	B	C	D
TMA07	4	2/4	3/4	3/4	2/4
TMA24	2	2/2	1/2	0/2	0/2
TMA29	7	2/7	3/7	3/7	4/7
TMA30	11	8/11	8/11	8/11	8/11
TMA31	3	2/3	2/3	2/3	2/3
TMA34	1	1/1	1/1	1/1	1/1
TMA36	7	2/7	4/7	0/7	1/7
TMA39	18	7/18	11/18	11/18	12/18
TMA41	7	4/7	6/7	6/7	4/7
TMA42	3	3/3	3/3	3/3	3/3
Total	63	33/63	42/63	37/63	37/63

D.3.1 Class ranks within candidates

Table 13: **Within-region GPT-6 ranks.** Top-k counts a GT class among the k highest-scoring tissue classes for the same 63 accessible GT regions. MRR is the mean reciprocal rank of the GT class. Ties keep the recorded Top-1 label first and then follow the original class-list order. Two E rows have a tied highest score; the other conditions have none. These quantities measure recognition given a GT-paired candidate, not complete-pipeline segmentation.

Input	Top-1	Top-2	Top-3	MRR
A: H&E	33/63	41/63	43/63	.6489
B: H&E + Cell Map	42/63	51/63	54/63	.7777
C: H&E + FICTURE	37/63	48/63	52/63	.7281
D: H&E + both maps	37/63	51/63	56/63	.7447
E: Cell Map only	27/63	44/63	48/63	.6275

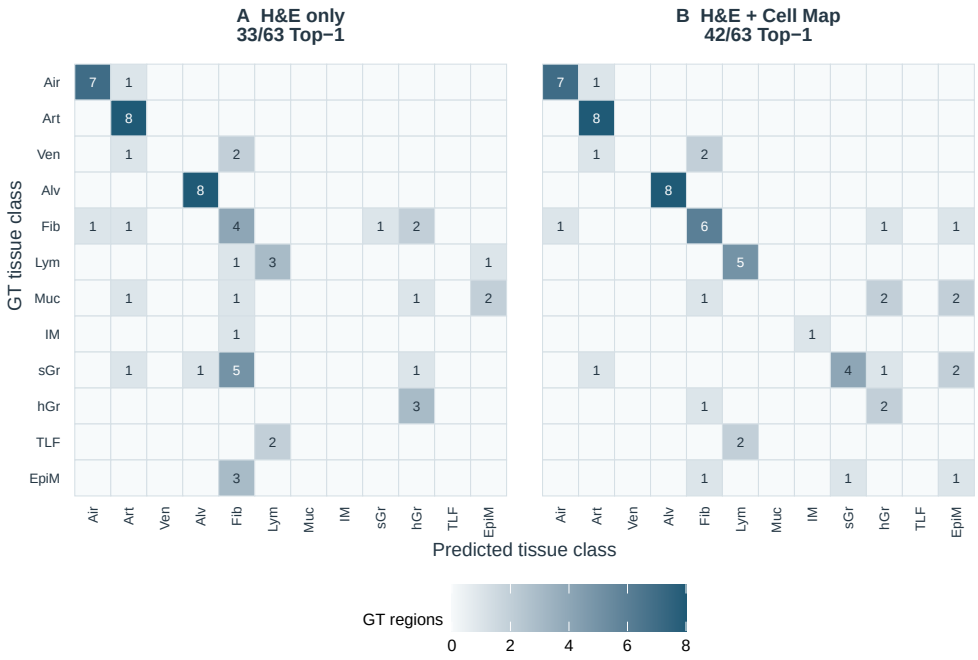

Figure 11: **GPT-6 confusion matrices.** H&E alone and H&E with Cell Map are compared on the same 63 GT-region assignments. Rows indicate GT classes and columns indicate predicted classes. Both panels share a count scale. Air, airway. Art, arteriole. Ven, venule. Alv, alveoli. Fib, fibroblast foci. Lym, lymphoid aggregate. Muc, mucous plug. IM, interstitial muscle. sGr, cellular granuloma. hGr, hyalinized granuloma. TLF, tertiary lymphoid follicle. EpiM, epithelioid macrophage.

D.4 Paired uncertainty and significance analyses

The 63 GT regions map to 56 unique candidates. We computed percentile intervals by resampling candidates 1,000 times with replacement, preserving all GT rows associated with each sampled candidate and applying the same resample to every input condition. The table in the main text gives marginal candidate-bootstrap intervals. The ranking intervals use the sorted group order and seed 20260924 in the anonymous code supplement. For B, resampling TMAs gives AUROC [.895, .977] and AP [.607, .870]. Paired differences below better address whether one input's ranking metric exceeds another's. For the Top-1 B-minus-A difference, we additionally resampled ten whole TMAs with replacement 10,000 times, preserving all paired GT outcomes and shared candidate predictions within each sampled TMA. In each replicate, the net change in correct labels was divided by the total number of sampled GT assignments. Python `random.Random` used sorted TMA IDs and seed 20260924. The observed difference was 9/63 (14.3 percentage points), with a 95% percentile interval of 2.4–21.9 percentage points. Five TMAs improved, one declined, and four were unchanged. This exploratory interval accounts for grouping within a TMA but does not establish independence between patients. The candidate-bootstrap intervals do not account for dependence among candidates within the same TMA. Resampling the ten TMAs instead gives B–A AUROC and AP difference intervals of [.039, .099] and [.025, .229], and D–B intervals of [-.022, .022] and [-.027, .106], respectively. Because only ten TMAs are available and some resamples omit rare classes, these are sensitivity checks rather than independent-cohort confidence bounds.

Table 14: **Paired class-ranking differences.** Each value is the second input minus the first. Intervals are 95% percentile intervals from 1,000 paired candidate-cluster bootstrap resamples (seed 20260924), with class metrics recomputed in every resample. A class without both positive and negative GT regions in a resample is omitted from that replicate's macro average.

Inputs	ΔAUROC [95% CI]	ΔAP [95% CI]
A → B	.0677 [.037, .114]	.1684 [.042, .240]
A → C	.0613 [.031, .101]	.1605 [.034, .222]
B → D	.0035 [-.016, .026]	.0319 [-.048, .094]
C → D	.0099 [-.005, .029]	.0398 [-.017, .106]

D.4.1 Exploratory paired Top-1 contrasts

Table 15: **Exploratory paired GPT-6 Top-1 contrasts.** Each row compares the same 63 accessible GT-region outcomes. "Gain" counts incorrect-to-correct changes from the first input to the second; "loss" counts the reverse. Two-sided exact McNemar p values are unadjusted for multiple comparisons and treat GT regions as independent, although they share 56 unique candidate images across ten TMAs. These nominal tests do not account for shared candidates or within-TMA clustering.

Inputs	Gain	Loss	Exact p
A → B	10	1	.0117
A → C	8	4	.3877
A → D	8	4	.3877
A → E	9	15	.3075
B → C	0	5	.0625
B → D	2	7	.1797
B → E	3	18	.0015

E Diagnostic experiments

These analyses examine sensitivity to proposal budgets, matching rules, class lists, and input presentation. GPT-6 vocabulary results are reported separately from the local Gemma experiments.

E.1 Candidate budget and selection sensitivity

Figure 12 tests whether a method's recall depends on how many masks it returns. Each cap is applied per TMA, with native quality ranking for model proposals, area ranking for expression regions, and the original order for our fixed pool. Curves therefore compare delivered policies, rather than assigning a common optimized selector to every generator. A cap is not a guarantee of equal actual counts: our pool saturates at 340, while grid methods continue to 800 at a cap of 80. The IAMSAM adaptation uses cell centers mapped to H&E coordinates for its tissue filter, without GT tissue-region prompts. It is not a strictly H&E-only input control.

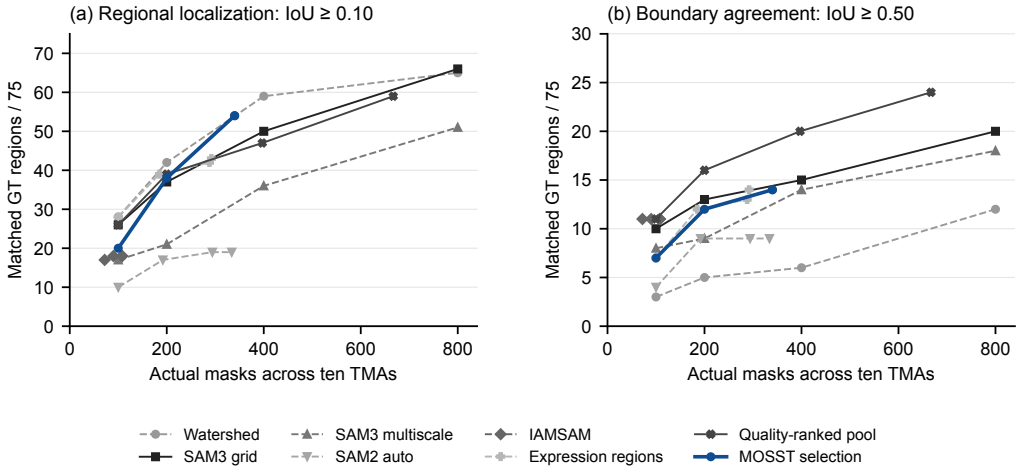

Figure 12: **Proposal-budget sensitivity.** Curves show one-to-one matches among 75 GT regions at IoU $\geq .10$ and $\geq .50$, using per-TMA limits of 10, 20, 40, and 80 masks. Coincident points indicate an exhausted proposal pool. SLIC/SLICO results are reported in Table 10.

At cap 40, the fixed policy obtains 54 IoU-.10 matches versus 47 for quality ranking, but 14 versus 20 at IoU .50. This is a coverage–boundary tradeoff, not uniform superiority. Grid sampling at cap 80 obtains 66 IoU-.10 matches with 800 masks. More masks can improve recall while increasing downstream work. No matched wall-clock speedup follows from these counts. Existing-pool packaging time is not proposal-generation time.

Grid sampling, expression components, and quality ranking change proposal construction or selection as well as the number of masks. These comparisons do not isolate the contribution of one proposal branch. The SLIC/SLICO runs collapsed to one mask per TMA and should be interpreted as execution diagnostics rather than representative performance of the method family.

E.2 Sensitivity to the geometric matching rule

Under the permissive geometric criterion, all-output GPT-6 B correctly classifies 43/75 GT regions with maximum-cardinality one-to-one assignment. For the historical baselines, the corresponding counts are SAM3 5/75, Grounded SAM2 29/75, SAM2–CONCH 11/75, SAM2–Gemma 11/75, ours–Gemma 31/75, LISA 35/75, GLaMM 31/75, MedPLIB 21/75, and Gemma localization–SAM2 6/75. These changes relative to Table 8 illustrate sensitivity to containment. They are not additional correctly segmented regions at IoU .50.

E.3 GPT-6 with a TMA-restricted class list

Table 16 isolates the provision of a TMA-specific vocabulary within GPT-6 condition B.

Table 16: GPT-6 condition B with Full-12 versus TMA-restricted classes. Restricted requires a separate inference and receives GT-derived vocabulary information. Its ranking metrics exclude observations from TMAs where a class is not offered, so the AUROC/AP difference is not a matched-population effect.

Vocabulary	Top-1	Correct / 75	AUROC	AP	Eligible classes
Full-12	42/63	42/75	.9327	.6915	12
Restricted	50/63	50/75	.9081	.8446	11

E.4 GPT-6 COMPOSITION-TEXT CONTROL

We added an exploratory cell-type composition control inspired by the ordered-frequency representation in LLMiniST (Wei et al., 2025). This adapts a cell-neighborhood method to candidate regions and adds H&E, rather than reproducing native LLMiniST. The 56 candidates, outlined H&E images, twelve class definitions, shared decision rules, response schema, and GPT-6 model settings match A and B. Only the cell-type input is changed.

Cell counts are summarized in three disjoint zones: inside the candidate, outside it within the close crop, and additional background in the wider crop. Types are listed by decreasing frequency with percentages and zone totals. All observed types are retained. Counts use the archived cell-center annotations and masks associated with the original Cell Map inputs. No gene list or GT tissue label enters the request. The existing cell-map response field explicitly records text evidence in this condition. This retains coarse regional context and supplies exact abundance, so it is not a pure spatial-shuffling experiment. Calls were made after A–E with the same model identifier, which does not guarantee an unchanged API backend.

All 56 responses were valid and mapped to the same 63 GT assignments. F gives 35/63 correct, macro AUROC .9190, and macro AP .6300 (Table 3). Compared with F, B gains seven correct labels, or 11.1 percentage points. In 10,000 paired whole-TMA resamples with seed 20260924, the 95% difference interval is [3.2, 25.0] percentage points. B–F AUROC and AP intervals are [-.0032, .0316] and [-.0086, .0956]. Classes without positives or negatives in a resample are omitted equally for both conditions. Five TMAs favor B and five tie. These exploratory intervals condition on recorded responses and do not establish patient independence or inference stability.

E.5 GEMMA INPUT-PRESENTATION CONTROLS

Table 17: **Gemma representation controls.** Same candidates and Full-12 classes. Rendering variants add the modified Cell Map to H&E. This is a local-model experiment, not a GPT ablation.

Input	Top-1	AUROC	AP
H&E only	19/63	.7188	.2966
H&E + composition text	31/63	.8647	.4901
H&E + Cell Map	24/63	.7605	.3776
Larger points ($r = 6$)	23/63	.7810	.4030
Continuous map ($\sigma = 6$)	25/63	.7838	.3537

Gemma performs better with cell-composition text than with the Cell Map image in this experiment. Enlarging cell points or smoothing the map does not consistently improve performance. The preference for text here differs from the GPT-6 result in Appendix E.4. Neither experiment establishes a model-independent advantage for image or text representations.

E.5.1 CELL-POINT RETENTION

Table 18 tests reduced cell observations through random retention. Point enlargement and continuous rendering instead test readability with the original cells retained. Mean correct counts across three seeds are 20.3, 21.3, and 23.0 at 25%, 50%, and 75% retention, respectively, versus 24 with the complete map. This trend is consistent with information loss, but the seed spread and mixed ranking metrics do not establish a monotonic effect for every target. The three seeds quantify variation from point sampling, not variation across tissue samples.

Table 18: Three random-retention seeds per fraction. Full-map Gemma obtains 24/63, AUROC .7605, and AP .3776. The same 56 candidate predictions map to 63 GT regions in each condition. AUROC and AP are Full-12 macro summaries.

Condition	Top-1	AUROC	AP
25% cells, seed 1	20/63	0.7294	0.3139
25% cells, seed 2	18/63	0.6981	0.2940
25% cells, seed 3	23/63	0.7499	0.3118
50% cells, seed 1	22/63	0.7480	0.3334
50% cells, seed 2	22/63	0.7582	0.3308
50% cells, seed 3	20/63	0.6971	0.2816
75% cells, seed 1	23/63	0.7643	0.3848
75% cells, seed 2	24/63	0.6989	0.3051
75% cells, seed 3	22/63	0.7692	0.3114

E.5.2 Gemma joint versus independent class queries

Table 19 compares joint and independent queries using Gemma. The joint condition requests twelve support scores in one call. The binary condition instead asks whether the target supports one specified tissue class, gives only that class definition, and records evidence for, against, and missing. The same H&E and Cell Map images and the same representative mapping are used. Twelve sets of 56 valid calls yield 672 responses. This changes the query and response format together. It diagnoses query formulation rather than isolating competition as the sole causal variable.

Table 19: **Gemma joint and independent class assessment.** Both protocols evaluate the same class-positive and class-negative GT regions. Each class is ranked across targets; independent-query scores are not combined into a calibrated multiclass predictor.

	GT regions		Joint scores		Independent query	
Class	Pos.	Neg.	AUROC	AP	AUROC	AP
airway	8	55	0.7898	0.6518	0.9841	0.8292
alveoli	8	55	0.6886	0.3968	0.9170	0.5019
arteriole	8	55	0.7807	0.5327	0.8295	0.4284
epitheloid_macrophage	3	60	0.8028	0.1899	0.6472	0.0681
fibroblast_foci	9	54	0.5833	0.2065	0.6389	0.1971
hGranuloma	3	60	0.8917	0.1875	0.6528	0.0749
interstitial_muscle	1	62	1.0000	1.0000	0.3871	0.0256
lymphoid_aggregate	5	58	0.9534	0.6548	0.8379	0.4310
mucous_plug	5	58	0.4310	0.0794	0.6172	0.1535
sGranuloma	8	55	0.8375	0.3340	0.8705	0.3955
tertiary_lymphoid_follicle	2	61	0.9426	0.2500	0.8443	0.1000
venule	3	60	0.4250	0.0476	0.9139	0.2292
Macro	—	—	0.7605	0.3776	0.7617	0.2862

Independent queries improve airway and venule ranking but reduce AP for several classes, including lymphoid aggregate and the single interstitial-muscle positive. The aggregate AP decrease rules out a blanket claim that twelve separate binary questions are a better recognition protocol. A per-class AP near one with only one or two positives remains weak evidence of stable class performance.

E.5.3 Gemma vocabulary sensitivity

Restricted Gemma uses a separate inference with the GT-derived TMA vocabulary, as does Restricted GPT. Table 20 retains every TMA, including single-class cases where ranking is undefined. Table 21 records the changed positive/negative populations. This tests the value of privileged vocabulary knowledge, not a deployable improvement to Full-12 recognition.

Table 20: Gemma H&E + Cell Map recognition, Full-12 and Restricted. Correct counts are over accessible GT; the full-cohort counts are 24/75 and 42/75. Pooled Total metrics are not averages of the TMA rows.

	Full-12			Restricted		
TMA	Correct	AUROC	AP	Correct	AUROC	AP
TMA07	1/4	0.7083	0.6667	3/4	1.0000	1.0000
TMA24	2/2	N/A	N/A	2/2	N/A	N/A
TMA29	1/7	0.7188	0.5228	5/7	0.7708	0.6190
TMA30	5/11	0.6285	0.5247	7/11	0.8802	0.7404
TMA31	0/3	0.5000	0.3889	1/3	0.4167	0.5556
TMA34	1/1	N/A	N/A	1/1	N/A	N/A
TMA36	0/7	0.4875	0.3710	3/7	0.5458	0.3798
TMA39	7/18	0.8269	0.7534	12/18	0.8493	0.6845
TMA41	6/7	0.8333	0.7857	6/7	1.0000	1.0000
TMA42	1/3	0.7500	0.7778	2/3	0.8333	0.6667
Total	24/63	0.7605	0.3776	42/63	0.8462	0.7036

Table 21: Gemma class-wise vocabulary analysis. P/N records positives and negatives; the missing Restricted hGranuloma metric reflects zero negatives. Class sets remain twelve, while macro ranking uses twelve and eleven eligible classes.

	Full-12			Restricted		
Class	P/N	AUROC	AP	P/N	AUROC	AP
airway	8/55	0.7898	0.6518	8/35	0.9107	0.8427
alveoli	8/55	0.6886	0.3968	8/27	0.8565	0.6888
arteriole	8/55	0.7807	0.5327	8/38	0.8273	0.5109
epitheloid_macrophage	3/60	0.8028	0.1899	3/4	0.6667	0.8095
fibroblast_foci	9/54	0.5833	0.2065	9/23	0.7512	0.6005
hGranuloma	3/60	0.8917	0.1875	3/0	N/A	N/A
interstitial_muscle	1/62	1.0000	1.0000	1/17	0.9706	0.5000
lymphoid_aggregate	5/58	0.9534	0.6548	5/20	0.9400	0.8769
mucous_plug	5/58	0.4310	0.0794	5/20	0.6550	0.2883
sGranuloma	8/55	0.8375	0.3340	8/10	0.8812	0.8552
tertiary_lymphoid_follicle	2/61	0.9426	0.2500	2/5	1.0000	1.0000
venule	3/60	0.4250	0.0476	3/11	0.8485	0.7667

F SPATIAL EXAMPLES AND FAILURE CASES

The examples below distinguish recognition errors, missing candidates, and uncertain annotation boundaries. GT regions cover confidence-annotated tissue only. Predictions outside those regions remain unverified.

F.1 RECOGNITION CHANGES AND FAILURE STAGES

Figure 13 shows both corrections and a reversal after adding molecular input. Figure 14 distinguishes failure to propose a region from assigning the wrong class to an available candidate.

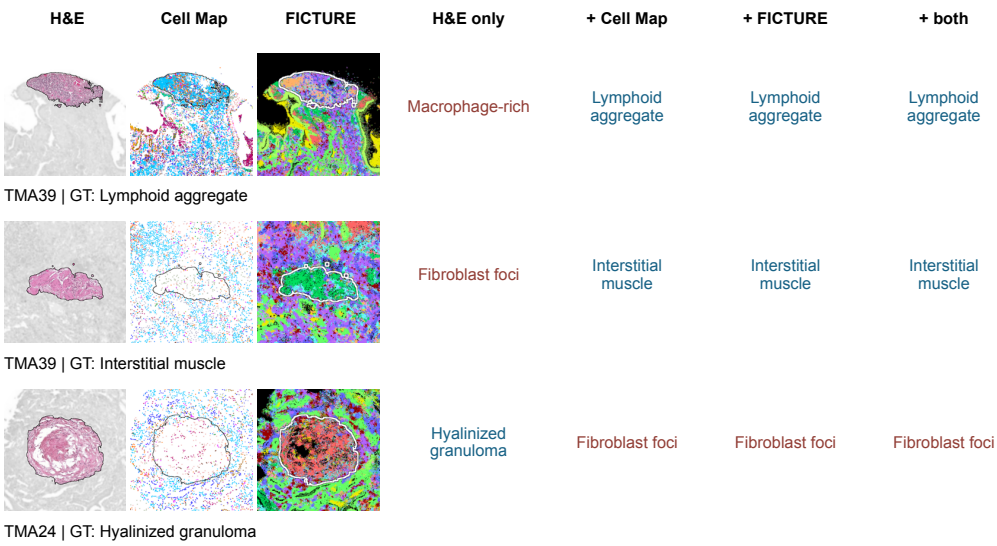

Figure 13: **Changes in GPT-6 classification with molecular input.** The same candidate is evaluated under H&E alone, H&E with Cell Map, H&E with FICTURE, and all three modalities. Blue labels match GT and red labels differ. Rows show two corrections and one reversal. Macrophage-rich denotes epitheloid_macrophage.

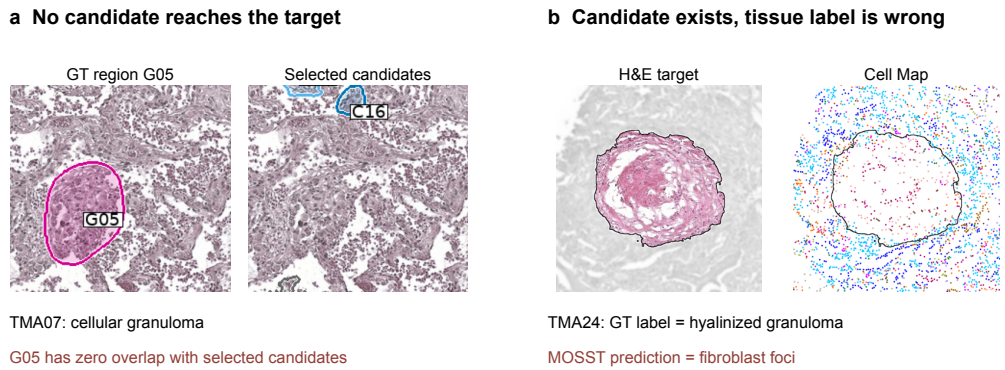

Figure 14: **Proposal and recognition errors.** (a) No selected candidate overlaps the cellular-granuloma GT region in TMA07. (b) A candidate overlaps the hyalinized-granuloma GT region in TMA24, but GPT-6 B classifies it as fibroblast foci. Images within each panel share coordinates.

F.2 NESTED PREDICTIONS AND ANNOTATION BOUNDARIES

Figure 5 in the main text maps all selected TMA39 predictions to H&E. Overlapping masks with different predicted classes remain separate because distinct tissue structures can be nested.

Nested predictions. A candidate labeled airway can contain a smaller candidate labeled mucous plug. We preserve both masks and their labels rather than forcing one class per pixel. In the TMA39 example, H&E_P09 is correctly labeled airway against GT region airway_1. Independent_H&E_R2 is labeled mucous plug, but that location has no separate mucous-plug GT annotation. Its identity is therefore unverified, even though it lies inside the annotated airway. This example illustrates output structure, not two confirmed correct predictions.

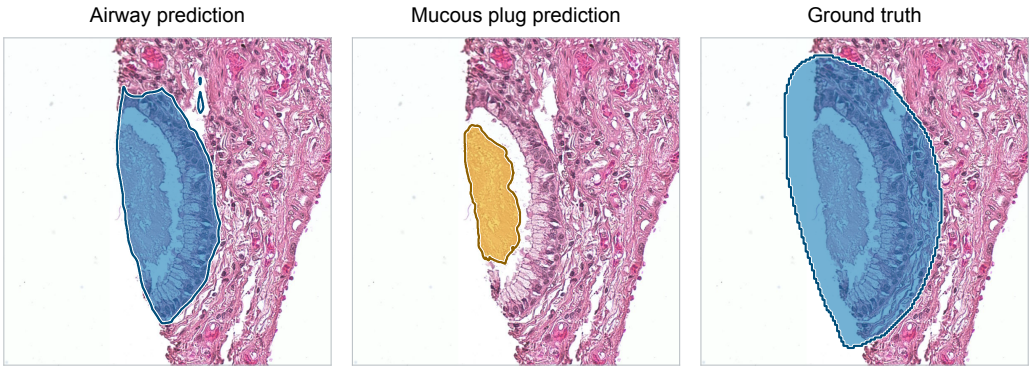

Figure 15: **Nested airway and mucous-plug predictions.** Matched H&E views show the GPT-6 B airway mask (blue), mucous-plug mask (gold), and airway GT region (blue). A mucous plug can lie within an airway lumen, so the masks retain separate labels. The GT boundary extends beyond the predicted airway contour. Without independent boundary review, this difference cannot establish which contour is more accurate. The mucous-plug prediction has no separate GT annotation and remains unverified.

F.3 LUMINAL REGIONS WITH SPARSE CELL OBSERVATIONS

Figure 16 shows recognition outcomes in regions where the Cell Map provides few plotted cells. Sparse observations may limit the available molecular evidence, but do not establish that the tissue is acellular.

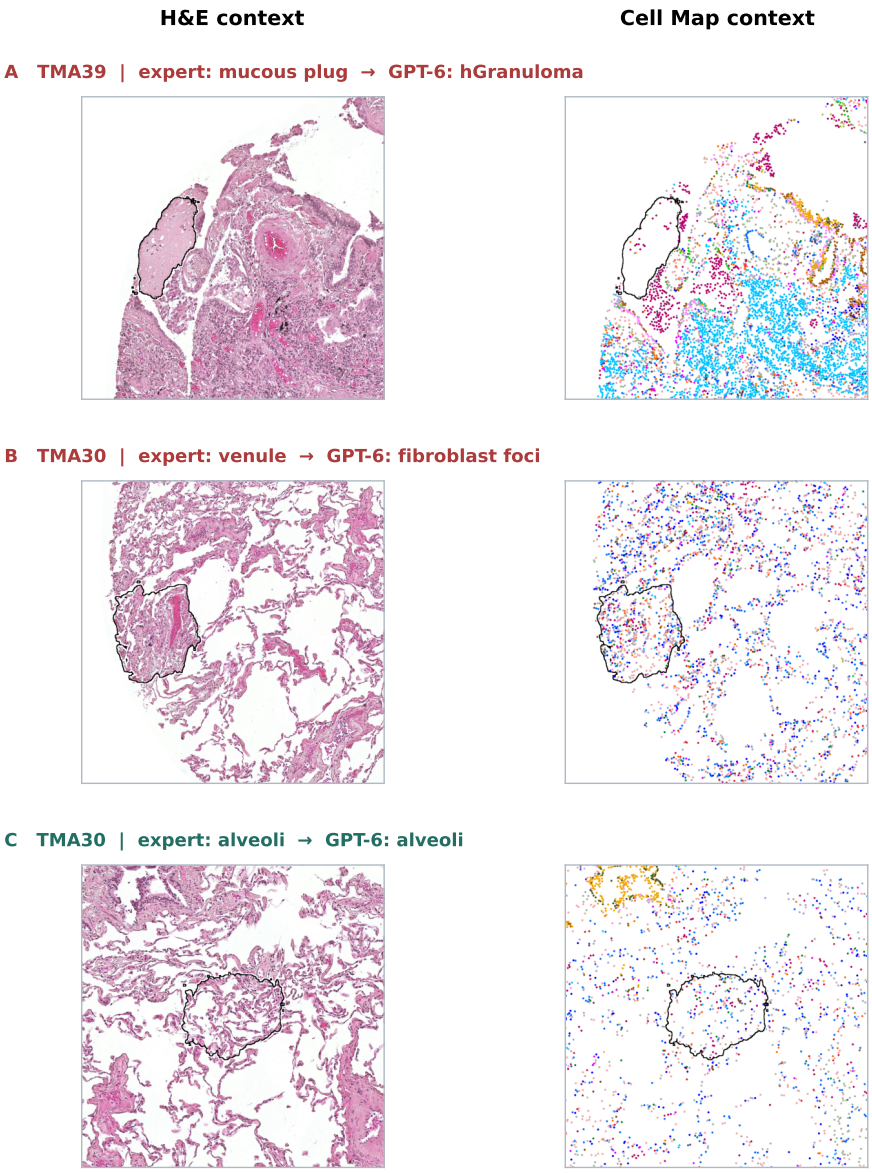

Figure 16: **Recognition in luminal regions.** Wider H&E and Cell Map views show three candidates with black mask outlines. GPT-6 B classifies the TMA39 mucous plug as hGranuloma, the TMA30 venule as fibroblast foci, and the TMA30 alveolar region as alveoli. Red and green labels indicate incorrect and correct classifications, respectively. Sparse Cell Map points do not establish that tissue is acellular.

F.4 WHOLE-TMA CANDIDATE SHAPES

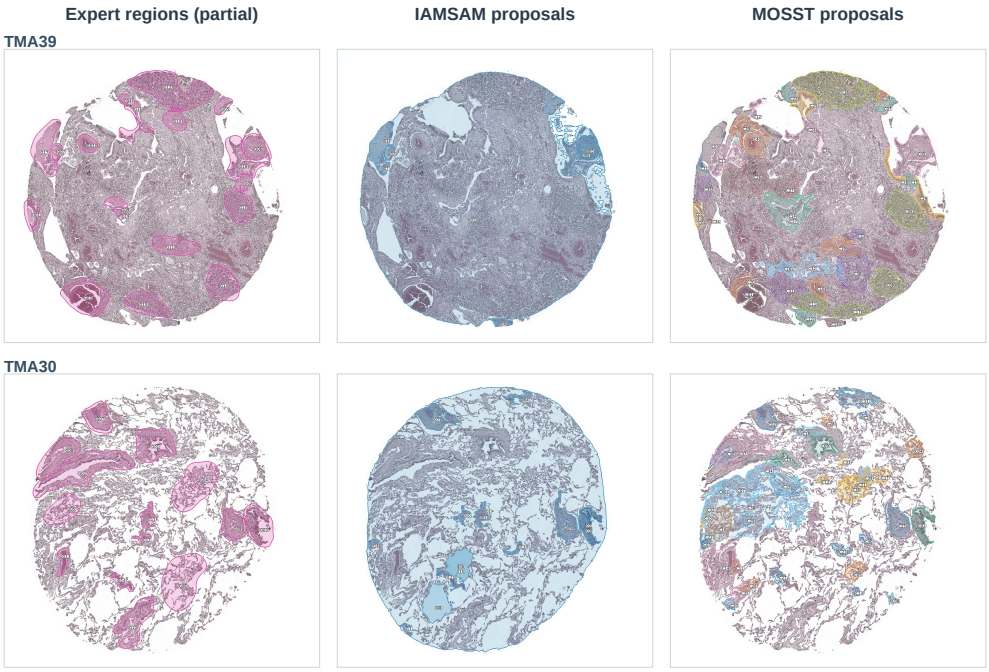

Figure 17: **IAMSAM and MOSST proposals.** GT regions, IAMSAM masks, and selected MOSST masks are shown at matched coordinates for TMA39 and TMA30. GT covers confidence-annotated regions only. These examples illustrate boundary differences and do not constitute a blinded quality assessment.

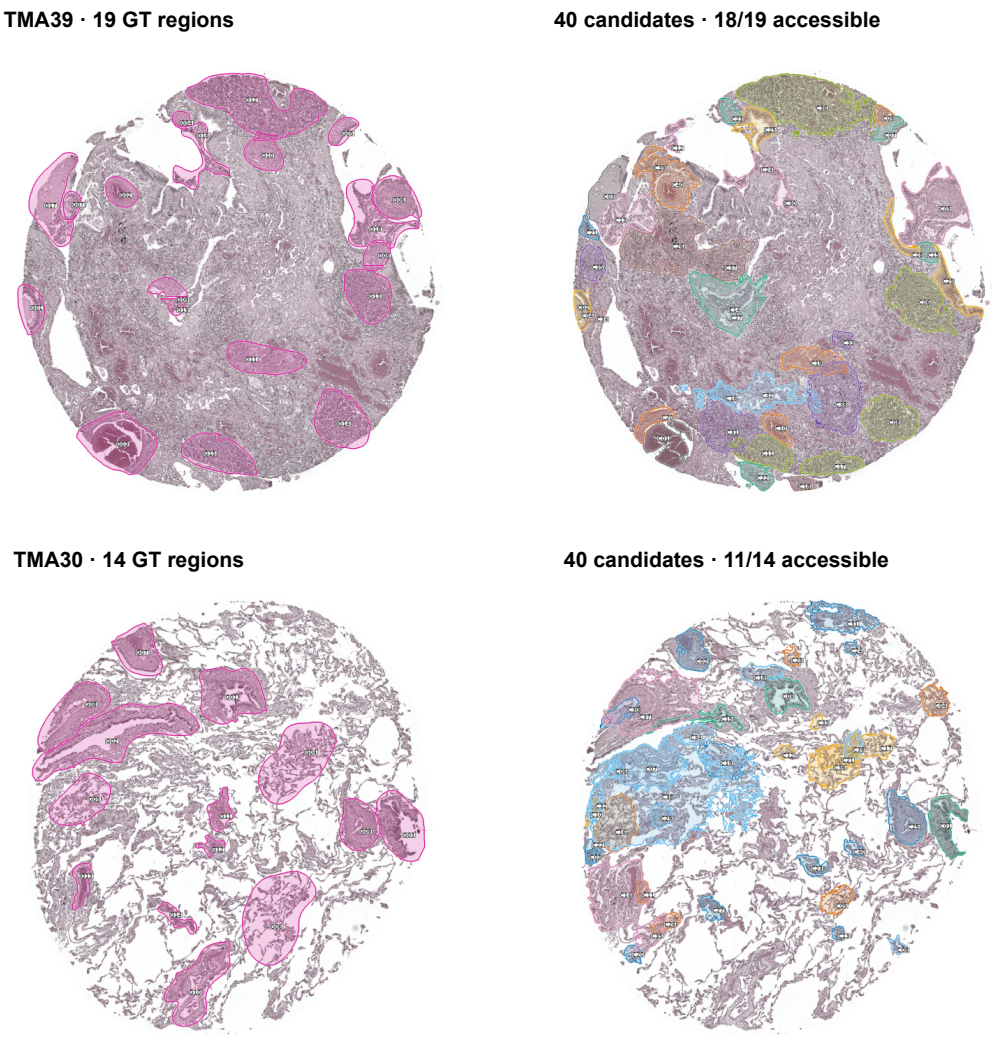

Figure 18: **GT regions and selected candidates.** TMA39 and TMA30 are shown in matched H&E coordinates, with GT regions on the left and candidates on the right. Colors distinguish masks. Each TMA has 40 candidates, with 18/19 and 11/14 accessible GT regions, respectively.