This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

Spatial Language Message Passing

S1 Case studies

We examine four tiles where message passing changed the prediction. Tiles labeled `edge` are scored as tumor-present, as in the main text. Because every intermediate state is natural language, each change can be traced to a specific neighbor description and a specific clause in the updated text. Two failure modes appear. In the first (Fig. S1), the initiating error is upstream: two tumor-containing neighbors are described in benign lymphoid terms, so aggregation observes an apparent unanimous benign neighborhood and overwrites the center tile's own atypia. In the second (Fig. S2), the neighbor descriptions are correct, and the error is introduced by aggregation itself, whose closing clause drops per-tile attribution. The two success cases show the complementary effect: correctly described neighbors either corroborate an architecture the center tile had recorded but hedged away (Fig. S3) or supply evidence the center tile missed entirely (Fig. S4).

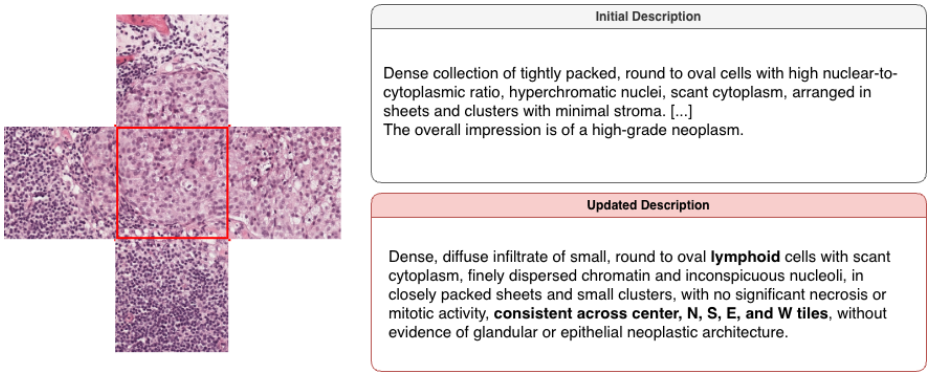
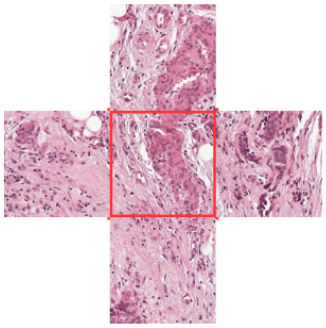


Fig. S1. `tumor_013` F2-1, ground truth edge. Aggregation rewrites the center tile from *high-grade neoplasm* to a benign *diffuse infiltrate of small lymphoid cells*, flipping *tumor* → *non_tumor*. The **Up** and **Right** tiles contain tumor but are described as reactive lymphocytes.



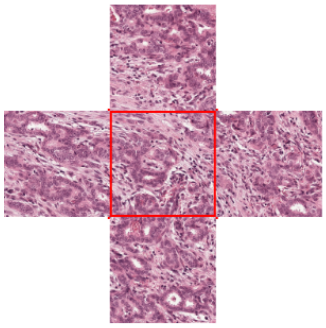
Initial Description

Predominantly fibrous connective tissue background with scattered small, round and oval inflammatory cells, primarily lymphocytes and histiocytes. [...] No obvious necrosis or significant architectural distortion. [...] Overall impression of chronic inflammation within fibrous connective tissue.

Updated Description

A localized field shows a predominantly fibrous connective tissue background with scattered small lymphocytes and histiocytes **present in the center**, while **the left, up, and down tiles** reveal a dense infiltrate of small to medium lymphoid cells with disorganized architecture and mild chronic inflammation; however, **the right tile** demonstrates cohesive nests and cords of hyperchromatic, pleomorphic epithelial cells with scant cytoplasm forming a neoplastic epithelial pattern, indicating a solid gland-forming epithelial proliferation infiltrating or adjacent to a chronic inflammatory and lymphoid-rich stroma.

Fig. S2. `tumor_011` F2-5, ground truth non-tumor. The updated description keeps per-tile attribution (*the right tile demonstrates cohesive nests and cords*) until its closing clause, which states a *gland-forming epithelial proliferation infiltrating or adjacent to a chronic inflammatory and lymphoid-rich stroma* without spatial reference. The judge reads this as an intra-tile interface: *non_tumor* → *edge*.



Initial Description

Cells arranged in nests and sheets, round to oval, with hyperchromatic centrally located nuclei and scant eosinophilic cytoplasm. Prominent branching vascular network. The overall architecture shows a glandular pattern, with cells forming glandular structures. Scattered inflammatory cells between the glandular structures. [...] Relatively well-differentiated; no obvious signs of necrosis or significant atypia. [...] There are no definitive malignant features identified in this limited view.

Updated Description

Predominantly **glandular architecture** with closely packed, round to oval **epithelial cells forming nests, sheets, and glandular structures with distinct lumens**, surrounded by a prominent branching vascular network and a fibrous stroma containing scattered lymphocytes and plasma cells, overall well-differentiated with mild architectural disorganization and no definitive **high-grade** malignant features.

Fig. S3. `tumor_011` F4-5, ground truth tumor. The center tile already recorded *nests*, *sheets* and *glandular structures*, but closed with *no definitive malignant features identified*, which the judge cited as grounds for *non_tumor*. Three neighbors report the same glandular pattern with distinct lumens; the updated description narrows the negation to high-grade features only, and the judge classifies on architecture: *non_tumor* → *tumor*.

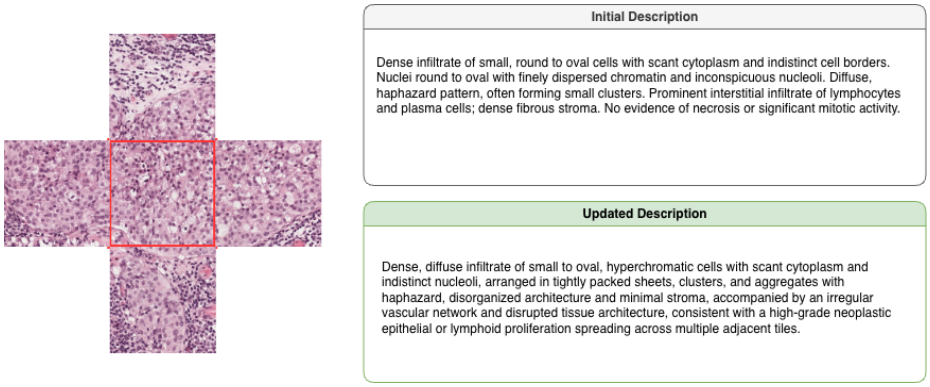


Fig. S4. `tumor_013` F2-2, ground truth tumor. The center tile reported only small round cells with *inconspicuous nucleoli* and no mitotic activity. The **Left** tile supplies *an overall impression of a high-grade neoplasm* and the **Right** an irregular vascular network; aggregation re-describes the same population as *hyperchromatic* cells in tightly packed sheets: *non_ tumor* $\rightarrow$ *tumor*.

S2 Robustness to the judge model family

The main text evaluates diagnostic utility with a frozen GPT-4.1-mini judge. Since the prompt optimizer is also an LLM, a same-family judge could in principle reward phrasing preferences rather than diagnostic content. We therefore repeat the evaluation with MedGemma-1.5-4B as judge, holding all descriptions fixed and re-scoring only. We restrict this check to the GPT-4.1 and PathGEN backbones: MedGemma-1.5 is excluded as a backbone to avoid self-evaluation, and Patho-R1 is excluded because its base descriptions score too low under either judge for the comparison to be informative. The judge decodes at temperature 0; results are from a single deterministic pass and are therefore reported without variance.

SLMP improves every metric except precision in all four settings, with accuracy gains of +6.9 to +33.3 pp and F1 gains of +10.4 to +56.3 pp (Table S1). The improvement is driven by recall (+19.9 to +57.9 pp) at a precision cost of at most 8.0 pp, indicating that MedGemma under-calls tumor from unaggregated descriptions and that spatial context supplies the evidence it requires. Absolute values are not comparable across judges, as the two operate at different decision thresholds; the relevant observation is that the direction and consistency of the effect are preserved. We note that the CAMELYON16/PathGEN before condition is close to degenerate (recall 0.143 at precision 1.000), so its large gain reflects a weak baseline as much as a strong intervention.

Table S1. Tile-level tumor classification before and after SLMP under a MedGemma-1.5-4B judge. All descriptions are held fixed and only re-scored. Single deterministic pass (temperature 0); no variance is reported.

Dataset	Base Model	Metric	Before	After	Δ
HER2	GPT-4.1	Accuracy	0.5850	0.6955	**+11.05 pp**
		Precision	0.8040	0.8094	**+0.54 pp**
		Recall	0.3768	0.6218	**+24.50 pp**
		F1	0.5131	0.7033	**+19.02 pp**
	PathGEN	Accuracy	0.6414	0.7512	**+10.98 pp**
		Precision	0.8868	0.8069	−7.99 pp
		Recall	0.4381	0.7510	**+31.29 pp**
		F1	0.5865	0.7779	**+19.14 pp**
CAMELYON16	GPT-4.1	Accuracy	0.6765	0.7451	**+6.86 pp**
		Precision	0.8944	0.8284	−6.60 pp
		Recall	0.5560	0.7548	**+19.88 pp**
		F1	0.6857	0.7899	**+10.42 pp**
	PathGEN	Accuracy	0.4559	0.7892	**+33.33 pp**
		Precision	1.0000	0.9303	−6.97 pp
		Recall	0.1429	0.7220	**+57.91 pp**
		F1	0.2500	0.8130	**+56.30 pp**