REPATH: ZERO-SHOT WHOLE-SLIDE TUMOR SEGMENTATION VIA INTRA-SLIDE RELATIONAL AFFINITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot tumor segmentation of whole-slide images (WSIs) via vision–language models (VLMs) faces a fundamental obstacle: the gigapixel resolution of WSIs forces patch-level processing, with each patch independently scored against fixed text embeddings. This absolute patch-level scoring produces two failure modes: *prompt dependency* (sensitivity to prompt wording) and *spatial inconsistency* (adjacent patches of the same tissue receiving conflicting labels), both reflecting a mismatch between slide-level histopathological interpretation and independent patch-wise classification. We propose RePath, which reformulates inference from per-patch scoring against fixed text embeddings to relational comparison among patches within the slide, built on three design principles. First, slide-specific anchor selection draws tumor, normal, and non-tissue anchor patches from the slide itself, using text prompts only to seed the selection. Second, intra-slide relational affinity recasts isolated image-text similarity as the Jensen-Shannon Divergence (JSD) between each patch's class distribution and the anchor distributions, yielding tumor, normal, and non-tissue affinity maps. Third, Differential Affinity Integration (DAI), inspired by clinical differential diagnosis, contrasts the tumor affinity against the normal and non-tissue affinities to produce a coherent segmentation mask. On 1,647 test WSIs across four organs and three VLM backbones, RePath improves Dice by 13.0 and IoU by 14.5 over the vanilla zero-shot baseline and, on average, matches or exceeds few-shot supervised methods trained with 10 labeled WSIs, without annotation or training.

1 INTRODUCTION

Tumor segmentation in whole-slide images (WSIs) underlies cancer staging, prognosis, and treatment planning (Campanella et al., 2019; Bera et al., 2019; 2020; Omar et al., 2024; Hassell et al., 2025). However, the gigapixel resolution of WSIs makes pixel-level annotation prohibitively expensive and prevents standard vision architectures, such as CNNs (Krizhevsky et al., 2012) and ViTs (Dosovitskiy et al., 2021), from processing a slide at once, forcing computational pipelines onto small image patches. Existing supervised and weakly supervised methods (Shi et al., 2022; Zhang et al., 2023a; Chan et al., 2019; Han et al., 2022; Li et al., 2022; Pan et al., 2023; Zhang et al., 2023b; Tang et al., 2025; Zhang et al., 2022b) rely on large curated datasets, but inter-institutional variability in tissue preparation, staining, and imaging (Lin et al., 2025; Wu et al., 2025) makes such labels prohibitively expensive to collect, limiting their scalability and generalization. These constraints motivate *zero-shot tumor segmentation in WSIs*, namely localizing tumor regions in any new slide without dataset-specific training or pixel-level supervision.

Recent histopathology Vision–Language Models (VLMs) (Huang et al., 2023; Lu et al., 2024; Sun et al., 2025; Ding et al., 2025; Lu et al., 2023; Zhang et al., 2025) are a natural candidate for this zero-shot setting via image–text alignment, but they inherit the same gigapixel constraint. *These VLMs are trained at the patch level, and extending them to gigapixel-scale segmentation remains challenging*—even recent histopathology multimodal LLMs (Liang et al., 2025; Chen et al., 2025) that scale VLMs to slide-level reasoning target captioning and VQA rather than dense pixel-level prediction. Zero-shot WSI segmentation is therefore forced to operate as patch-tiled inference, with each patch independently scored against fixed text embeddings, reducing each decision to an absolute score against text and discarding the slide-level context that diagnostic reasoning relies on. This exposes two failure modes of zero-shot WSI segmentation. First, prompt dependency is the sensitivity

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.

of absolute scores to the wording of the prompt. Second, spatial inconsistency arises because each decision ignores its neighbors, so adjacent patches of the same tissue receive conflicting labels. Both reflect a mismatch between slide-level histopathological reasoning (Gurcan et al., 2009) and patch-wise classification, where isolated patches share no information.

In this work, we propose RePath (**Re**lational **Path**ology inference), an inference-only framework for zero-shot WSI tumor segmentation that re-evaluates each patch by how closely its class distribution matches those of anchor patches drawn from the same slide, rather than by its absolute score against text. RePath addresses the two diagnosed failures at their structural source through three design principles. First, slide-specific anchor selection draws anchor patches from the slide itself, using text prompts only to seed the selection, so that the references inherit each slide's tissue, staining, and preparation. Second, intra-slide relational affinity recasts isolated image–text similarity as the Jensen–Shannon Divergence (JSD) between each patch's class distribution and the anchor distributions, a symmetric and bounded measure that re-evaluates patch identity rather than smoothing existing predictions, motivated by the observation that patches of the same class share strongly correlated class distributions (Kang et al., 2025). Third, Differential Affinity Integration (DAI), inspired by the contrastive reasoning in clinical differential diagnosis, reframes direct tumor classification as a contrast of the tumor affinity against the normal and non-tissue affinities, suppressing residual confounders. Together, these moves shift inference from per-patch classification against fixed text embeddings to relational comparison within the slide, addressing prompt dependency and spatial inconsistency at their common source, without dataset-specific tuning.

Our contributions are summarized as follows.

- **Reformulation.** We show that prompt dependency and spatial inconsistency share a structural source – absolute, patch-independent scoring against fixed text embeddings – and that both are mitigated by reformulating inference as intra-slide relational comparison. Inference formulation thus matters as much as representation quality for VLM-based WSI analysis.
- **Principled framework.** We propose RePath, which reformulates WSI tumor segmentation as intra-slide relational affinity analysis through slide-specific anchor selection, JSD-based affinity over class distributions, and differential-diagnosis-inspired DAI, with text prompts seeding the anchors but not determining the final decision.
- **Empirical validation.** On 1,647 test WSIs across four organs and three backbones, RePath improves Dice by 13.0 and IoU by 14.5 over the vanilla zero-shot baselines, on average matches or exceeds 10-WSI few-shot baselines without annotation, and generalizes across organs and centers where few-shot baselines often degrade due to single-cohort overfitting.

2 Related Work

2.1 Vision-language Histopathology Model

Large-scale Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), have demonstrated that contrastive learning on image–text pairs yields highly transferable representations. Building on this paradigm, histopathology VLMs (Huang et al., 2023; Lu et al., 2024; Sun et al., 2025; Ikezogwo et al., 2023; Ding et al., 2025; Lu et al., 2023; Zhang et al., 2025)—including PLIP (Huang et al., 2023), CONCH (Lu et al., 2024), PathGen-CLIP (Sun et al., 2025), and Patho-CLIP (Zhang et al., 2025)—address annotation scarcity through large-scale image–text corpora from biomedical literature, social media, or multimodal LLM-generated supervision. Their strong zero-shot generalization is valuable in data-scarce clinical settings. We leverage their pretrained semantic knowledge for zero-shot WSI tumor segmentation.

2.2 Enhancing Classification in Vision-language Foundation Models

VLMs can facilitate slide-level segmentation through patch-level classification. Recent studies enhance VLM classification in both zero-shot (Radford et al., 2021; Pratt et al., 2023; Guo et al., 2023; Udandarao et al., 2023; Stojnić et al., 2024) and few-shot (Zhou et al., 2022b;a; Khattak et al., 2023; Zhu et al., 2023; Yao et al., 2023; Zhang et al., 2022a; Silva-Rodriguez et al., 2024) settings. For zero-shot, CuPL (Pratt et al., 2023) uses LLM-generated category descriptions, while CALIP (Guo et al., 2023) introduces parameter-free visual-textual attention. For few-shot, CoOp (Zhou et al.,

2022b) optimizes task-specific soft prompts while keeping the backbone frozen, KgCoOp (Yao et al., 2023) regularizes learnable prompts toward hand-crafted descriptions to reduce overfitting, Tip-Adapter (Zhang et al., 2022a) constructs a non-parametric key-value cache from few-shot samples, and CLAP (Silva-Rodriguez et al., 2024) revisits linear probing with class-adaptive prototypes. Closest to ours are methods that exploit unlabeled test images without parameter updates: Udandarao et al. (2023) builds an external support set as anchors, Stojnić et al. (2024) propagates labels over a test-image graph, and Kang et al. (2025) exploits correlated class distributions of same-class tokens within a single image. RePath instead draws anchors from the slide being segmented, measures affinity by JSD between per-patch class distributions rather than feature similarity or propagation (Appendices G.8, G.10), and contrasts three category-wise affinities through DAI.

2.3 Histopathology Segmentation

Histopathology segmentation has evolved from localized patch analysis to models incorporating global WSI context (Shi et al., 2022; Zhang et al., 2023a; Chan et al., 2019; Han et al., 2022; Li et al., 2022; Pan et al., 2023; Zhang et al., 2023b; Tang et al., 2025; Zhang et al., 2022b). HoVer-Net (Graham et al., 2019) predicts pixel-to-centroid distances for nuclear instance segmentation, while HookNet (van Rijthoven et al., 2021) integrates multi-resolution WSI context. Recent training-free open-vocabulary segmentation (OVSS) methods modify CLIP's ViT attention for dense prediction, including SCLIP (Wang et al., 2024), ClearCLIP (Lan et al., 2024), and NACLIP (Hajimiri et al., 2025). However, these methods refine visual tokens *within a single image*, whereas WSI segmentation independently encodes patches using their CLS embeddings; thus, the inter-patch inconsistency we target cannot be directly addressed by such intra-image attention reformulations. We instead address this inter-patch regime training-free through anchor-based affinity over patch-level class distributions.

3 Traditional VLM Inference for WSI Segmentation

In standard zero-shot WSI segmentation, a Vision-Language Model (VLM) typically employs a patch-tiled inference pipeline to handle the gigapixel resolution of the images. A slide is first decomposed into a set of patches $\{x_i\}_{i=1}^{N}$, from which an image encoder f_I extracts visual features $\mathbf{z}_i$. Simultaneously, a text encoder f_T generates a set of embeddings $\{\mathbf{w}_c\}_{c=1}^{C}$ corresponding to C target class prompts. These components are used to estimate a patch-level class distribution $\mathbf{s}_i$ by applying a Softmax function over the cosine similarities:

$$\mathbf{s}_i = \text{Softmax}\left(\frac{\rho(\mathbf{z}_i, \mathbf{W})}{\tau}\right), \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C]$ represents the matrix of text embeddings, $\rho(\cdot, \cdot)$ denotes the cosine similarity, and τ is a temperature scaling factor. The final segmentation mask M is then constructed by selecting the class with the maximum probability $s_{i,c}$ from each distribution, $M(x_i) = \arg\max_c s_{i,c}$.

However, this approach inevitably produces two failure modes. The first is *prompt dependency*, where the decision boundary is fixed by static text embeddings, making zero-shot predictions sensitive to the wording of the text prompt. The second is *spatial inconsistency*, where each patch is scored in isolation, resulting in fragmented masks that miss organized malignant structures.

4 Methodology

Our goal is to revisit zero-shot WSI tumor segmentation by shifting inference from per-patch classification against fixed text embeddings to inter-patch relational comparison within the slide. As described in Sec. 3, traditional VLM inference processes patches in isolation, inevitably yielding both prompt dependency and spatial inconsistency. To address these failure modes through reformulation rather than dataset-specific tuning, we propose RePath, a framework that augments patch-wise VLM inference with intra-slide patch-affinity analysis. Building on CDAM's (Kang et al., 2025) observation that same-category patches exhibit highly correlated class distributions, we evaluate each patch against slide-specific anchors using distributional similarity. This mitigates spatial inconsistency via relational comparison rather than explicit spatial modeling.

As illustrated in Fig. 1, RePath comprises three components. (i) *Slide-specific anchor selection* identifies representative tumor, normal, and non-tissue patches from within each slide. (ii) *Relational*

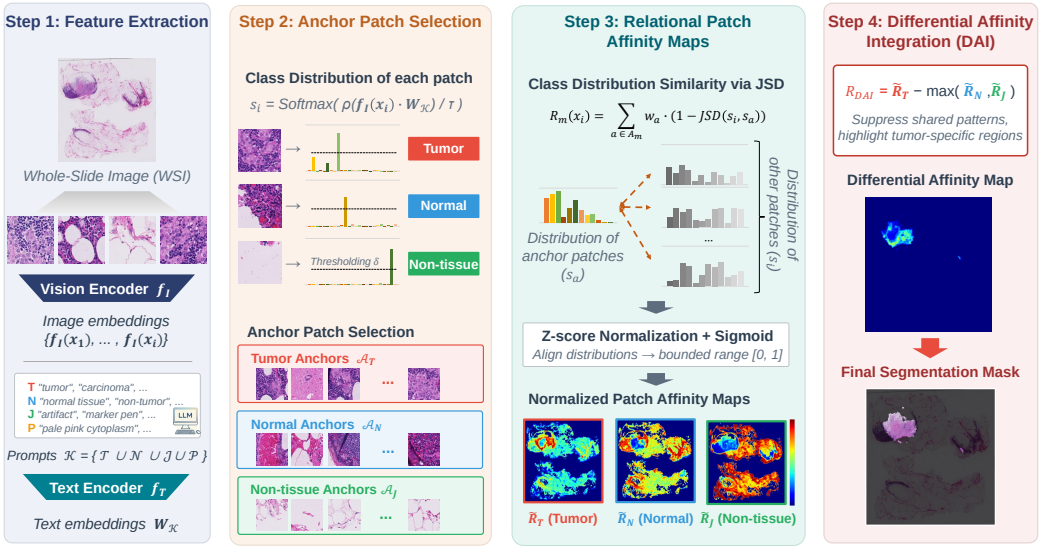

Figure 1: Overview of RePath for WSI tumor segmentation. The training-free pipeline comprises: 1) extracting visual and text features from WSI patches, 2) slide-specific anchor selection for tumor, normal, and non-tissue categories, 3) computing relational affinity maps using Jensen-Shannon Divergence (JSD) between patches and anchors, and 4) Differential Affinity Integration (DAI) to produce the tumor segmentation mask.

patch-affinity map construction via JSD scores every patch against these anchors using Jensen–Shannon Divergence over class distributions. (iii) *Differential Affinity Integration (DAI)* contrasts tumor affinity against normal and non-tissue affinities to produce the final segmentation mask.

4.1 Slide-specific Anchor Selection

Pathological appearance varies across slides in tissue type, staining, and preparation, so a fixed external reference such as a text prompt cannot adapt to each slide's specific visual context. To overcome this, we select *slide-internal* anchor sets $\mathcal{A}_m$ for each tissue category $m \in \{T, N, J\}$, corresponding to Tumor, Normal, and Non-tissue, respectively. The selection process begins by enriching the semantic density of the VLM's latent space. While the primary keywords for Tumor ($\mathcal{T}$) and Normal ($\mathcal{N}$) are predefined by the task, we leverage a Large Language Model (LLM) to generate auxiliary keyword sets:

- **Non-tissue Set ($\mathcal{J}$):** Keywords representing artifacts and non-biological elements such as *marker pen ink* or *folded tissue*.
- **Pathological Property Set ($\mathcal{P}$):** Keywords describing staining characteristics and tissue morphology, such as *pale pink cytoplasm* or *irregular nuclear contours*.

Let $\mathcal{K} = \{\mathcal{T} \cup \mathcal{N} \cup \mathcal{J} \cup \mathcal{P}\}$ be the union of these sets. For each patch x_i, we derive a high-dimensional class distribution $\mathbf{s}_i \in \mathbb{R}^{|\mathcal{K}|}$ by computing the similarity between its image embedding $f_I(x_i)$ and the expanded text embeddings $\mathbf{W}_{\mathcal{K}}$:

$$\mathbf{s}_i = \text{Softmax}\left(\rho(f_I(x_i), \mathbf{W}_{\mathcal{K}})/\tau\right). \tag{2}$$

The Pathological Property Set ($\mathcal{P}$) stabilizes the class distribution $\mathbf{s}_i$ by providing auxiliary morphological descriptors. This enrichment prevents overconfident assignments to primary categories when visual features are ambiguous, ensuring a more balanced distribution for reliable anchor patch identification.

A patch x_i is a candidate for category $m \in \{T, N, J\}$ if its summed probability $\sum_{k \in m} s_{i,k}$ both dominates the other categories and exceeds a confidence threshold δ, filtering out ambiguous or low-confidence samples.

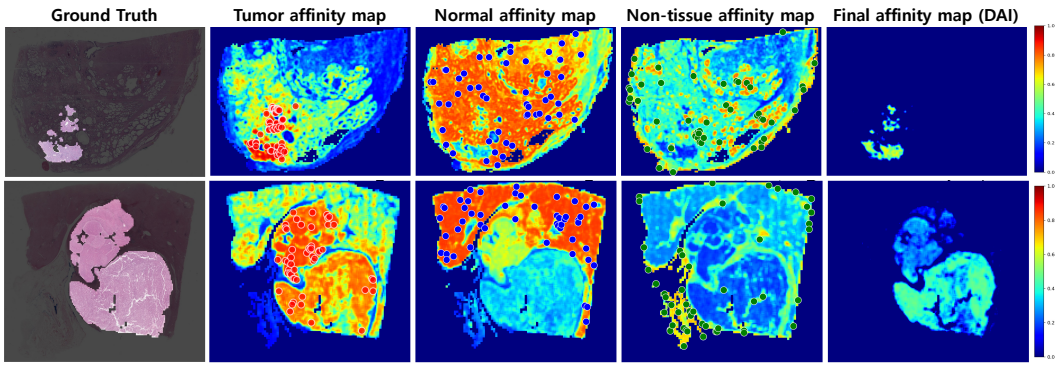

Figure 2: Visualization of relational patch affinity maps. Comparison of the original WSI, ground truth (GT), and individual patch affinity maps for Tumor, Normal, and Non-tissue categories. Colored dots indicate selected anchor patches for each category. The final affinity map demonstrates effective background suppression and precise tumor localization through the differential affinity integration.

From these candidates, we form each anchor set $\mathcal{A}_m$ with at most N_{max} patches. For Tumor ($\mathcal{A}_T$), we select the top-N_{max} candidates by confidence to prioritize representative tumor patterns. For Normal ($\mathcal{A}_N$) and Non-tissue ($\mathcal{A}_J$), we uniformly sample candidates, as these categories span diverse morphologies (stroma, epithelium, fat for normal; air, blur, ink for non-tissue) that top-confidence selection may underrepresent. When fewer than N_{max} candidates are available, all are used.

4.2 Relational Patch Affinity Map Construction via JSD

Unlike traditional methods that rely on isolated similarity scores, RePath computes three patch-to-patch relational affinity maps denoted as R_T, R_N, and R_J. Crucially, this affinity is a *re-evaluation* of patch identity through distributional comparison, not a post-hoc spatial smoothing of the existing patch predictions; we directly compare against neighborhood-smoothing baselines in the Appendix. For each category $m \in \{T, N, J\}$, we measure the relational similarity between the class distributions of all patches $\{\mathbf{s}_i\}_{i=1}^{N}$ and the selected anchors $\{\mathbf{s}_a\}_{a\in\mathcal{A}_m}$ using Jensen-Shannon Divergence (JSD). We adopt JSD because it is symmetric, bounded in [0, 1], and well-defined for distributions with non-overlapping support—properties that KL divergence lacks and that are critical when comparing sparse, high-dimensional class distributions across heterogeneous tissue types.

To prioritize high-confidence anchors while maintaining robustness, we employ temperature-scaled softmax weighted pooling:

$$R_m(x_i) = \sum_{a\in\mathcal{A}_m} w_a \cdot (1 - \text{JSD}(\mathbf{s}_i, \mathbf{s}_a)), \quad w_a = \frac{\exp((1 - \text{JSD}(\mathbf{s}_i, \mathbf{s}_a))/\beta)}{\sum_{a'\in\mathcal{A}_m} \exp((1 - \text{JSD}(\mathbf{s}_i, \mathbf{s}_{a'}))/\beta)} \tag{3}$$

where β is the temperature controlling the influence of each anchor. This calculation is performed for every patch $i \in \{1, \ldots, N\}$, effectively generating category-specific affinity maps that capture the internal relational structure of WSI.

For consistent scale across categories, we Z-score normalize the raw affinity scores $\mathcal{R} = \{R_T, R_N, R_J\}$ using their global mean and standard deviation, then apply a sigmoid to obtain bounded refined affinity maps $\tilde{R}_T, \tilde{R}_N$, and $\tilde{R}_J$ for subsequent integration stage.

4.3 Differential Affinity Integration (DAI)

To produce the final coherent WSI segmentation, we introduce Differential Affinity Integration (DAI), which contrasts the tumor affinity map against the normal and non-tissue affinity maps. The design of DAI is inspired by the contrastive aspect of clinical *differential diagnosis*, where pathologists establish a malignant reading by ruling out benign mimics and processing artifacts. Mirroring this reasoning, DAI suppresses shared histological patterns and emphasizes tumor-specific characteristics

Table 1: Comparison of WSI tumor segmentation. All results are reported as IoU (↑) / Dice (↑). For few-shot methods, we report the average performance across 10 trials using a 10-WSI shot setting. The top three performances within each backbone are highlighted in 1st, 2nd, and 3rd. Zero-shot methods and RePath use a single LLM-generated prompt set; statistics over 10 prompt sets are reported in Fig. 3 and Table A13.

Category	Method	Annotation	Training	PAIP Prostate		PAIP HCC		PAIP COLON		Camelyon 16		Average	
				IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)
Backbone VLM: CONCH													
Few-shot (10-WSI)	CLAP (Silva-Rodriguez et al., 2024)	✓	✓	42.5	55.4	67.0	77.2	60.3	73.0	34.7	43.9	51.1	62.4
	Tip-Adapter (Zhang et al., 2022a)	✓	✗	50.6	63.9	69.5	79.5	56.9	70.3	34.1	43.5	52.8	64.3
	KgCoOp (Yao et al., 2023)	✓	✓	49.6	62.6	71.0	80.5	60.5	72.9	40.2	50.7	55.3	66.7
	CoOp (Zhou et al., 2022b)	✓	✓	52.1	65.0	73.5	82.4	61.7	73.8	40.6	52.1	57.0	68.3
Zero-shot	Vanilla	✗	✗	51.1	65.4	63.0	74.4	52.3	66.2	33.2	42.5	49.9	62.1
	CuPL (Pratt et al., 2023)	✗	✗	38.6	52.0	61.3	71.9	51.1	65.1	22.6	30.2	43.4	54.8
	CALIP (Guo et al., 2023)	✗	✗	47.9	61.9	64.5	75.2	56.7	70.1	24.4	31.2	48.4	59.6
	RePath (Ours)	✗	✗	67.4	78.1	78.1	85.5	63.0	75.0	45.9	52.7	63.6	72.8
Backbone VLM: PathGen-CLIP													
Few-shot (10-WSI)	CLAP (Silva-Rodriguez et al., 2024)	✓	✓	38.0	50.8	61.7	73.4	55.1	68.9	24.3	31.8	44.8	56.2
	Tip-Adapter (Zhang et al., 2022a)	✓	✗	34.2	46.3	51.5	64.6	52.9	66.8	26.4	34.5	41.3	53.1
	KgCoOp (Yao et al., 2023)	✓	✓	41.2	54.3	64.6	75.8	56.9	70.3	29.7	38.1	48.1	59.6
	CoOp (Zhou et al., 2022b)	✓	✓	44.2	57.5	66.1	76.9	58.5	71.3	30.9	39.7	49.9	61.4
Zero-shot	Vanilla	✗	✗	32.9	45.4	43.5	57.2	52.9	67.0	15.8	21.8	36.3	47.9
	CuPL (Pratt et al., 2023)	✗	✗	24.3	35.6	47.0	60.4	42.6	56.8	12.0	16.6	31.5	42.4
	CALIP (Guo et al., 2023)	✗	✗	26.7	37.8	42.6	56.2	43.4	57.6	15.4	20.9	32.0	43.1
	RePath (Ours)	✗	✗	45.9	57.9	66.9	76.5	56.3	69.5	32.2	40.2	50.3	61.0
Backbone VLM: Patho-CLIP-B													
Few-shot (10-WSI)	CLAP (Silva-Rodriguez et al., 2024)	✓	✓	36.1	48.9	64.1	75.6	58.3	71.1	23.9	31.4	45.6	56.8
	Tip-Adapter (Zhang et al., 2022a)	✓	✗	35.3	47.8	65.2	76.0	56.0	69.1	16.5	22.7	43.3	53.9
	KgCoOp (Yao et al., 2023)	✓	✓	40.7	53.8	67.9	78.5	59.8	72.4	27.3	35.6	48.9	60.1
	CoOp (Zhou et al., 2022b)	✓	✓	43.5	57.0	69.1	79.4	59.3	71.9	26.5	34.7	49.6	60.8
Zero-shot	Vanilla	✗	✗	25.5	37.0	53.8	65.6	48.4	62.3	11.5	16.2	34.8	45.3
	CuPL (Pratt et al., 2023)	✗	✗	19.4	29.5	42.9	56.0	44.6	59.0	10.4	14.6	29.3	39.8
	CALIP (Guo et al., 2023)	✗	✗	23.3	34.1	61.4	72.5	47.8	62.0	11.2	15.8	35.9	46.1
	RePath (Ours)	✗	✗	45.2	57.5	72.6	81.3	60.1	72.7	24.2	30.5	50.5	60.5

by leveraging the relative differences between $\tilde{R}_T, \tilde{R}_N$, and $\tilde{R}_J$. The refined tumor score is then calculated by subtracting the maximum background interference:

$$R_{DAI}(x_i) = \tilde{R}_T(x_i) - \max(\tilde{R}_N(x_i), \tilde{R}_J(x_i)) \tag{4}$$

where $\tilde{R}_m$ denotes the normalized affinity map for domain m. This dual-subtraction strategy filters out normal tissue structures and non-biological artifacts that may visually resemble tumor patterns. The final segmentation mask M is obtained by thresholding the differential affinity map:

$$M(x_i) = \mathbb{I}(R_{DAI}(x_i) > 0) \tag{5}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The zero threshold follows naturally from the construction of R_{DAI}: since the three affinity maps are z-score normalized to a common distribution before the sigmoid transform, $R_{DAI}(x_i) > 0$ indicates that a patch's tumor affinity exceeds the stronger of its normal and non-tissue affinities in standardized units. As illustrated in Fig. 2, this strategy suppresses background noise and produces tumor masks closely aligned with the ground truth.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on four WSI tumor segmentation benchmarks across four organs: **Camelyon 16 (CM-16)** (Ehteshami Bejnordi et al., 2017) for breast lymph-node metastasis, restricted to tumor-bearing WSIs (111/48 train/test), and three **PAIP** (Kim et al., 2023; 2021) organs—Liver HCC (101/402), Prostate (120/479), and Colon (180/718)—each split 2:8 into train/test WSIs. Evaluation uses test WSIs, while few-shot baselines randomly sample 10 training WSIs per trial. For cross-center generalization, we additionally use **Camelyon 17 (CM-17)** (Bandi et al., 2018), restricted to 50 WSIs (10 per center) with public tumor masks; each center serves in turn as the few-shot training set and the other four as test sets. Further dataset and split details are provided in the Appendix.

Table 2: Multi-center generalization results on Camelyon17 using CONCH backbone. For few-shot methods, each column denotes the average performance on the remaining four centers after training on 10 WSIs from the specified center. Our RePath performs zero-shot inference without center-specific training. All results are reported as IoU (↑) / Dice (↑). The best and second-best results in each column are highlighted in **bold** and underline, respectively.

Category	Method	Annotation	Training	Trained on Center $i \rightarrow$ Tested on Others (IoU / Dice)					
				Center 1	Center 2	Center 3	Center 4	Center 5	Avg.
Few-shot (10-WSI)	Tip-Adapter (Zhang et al., 2022a)	✓	✗	30.8 / 38.5	29.4 / 36.1	32.6 / **41.2**	28.4 / 35.8	32.6 / 40.8	30.8 / 38.5
	CLAP (Silva-Rodriguez et al., 2024)	✓	✓	16.1 / 21.1	24.6 / 31.1	15.2 / 20.6	19.2 / 25.5	18.5 / 24.9	18.7 / 24.6
	CoOp (Zhou et al., 2022b)	✓	✓	23.5 / 29.6	31.6 / 39.0	21.5 / 27.4	21.4 / 27.6	27.6 / 35.6	25.1 / 31.8
	KgCoOp (Yao et al., 2023)	✓	✓	17.5 / 22.8	30.4 / 37.5	14.0 / 19.2	21.4 / 28.0	22.4 / 29.4	21.1 / 27.4
Zero-shot	Vanilla	✗	✗	29.4 / 37.2	31.4 / 38.8	29.6 / 37.9	29.6 / 37.4	32.0 / 40.5	30.4 / 38.4
	RePath (Ours)	✗	✗	**34.0 / 38.6**	**38.5 / 43.2**	**35.0** / 40.1	**38.2 / 43.5**	**37.2 / 42.7**	**36.6 / 41.6**

Table 3: Ablation study of RePath components. Using CONCH, we evaluate the contribution of Differential Affinity Integration (DAI) with relation patch affinity maps for normal tissue region ($\tilde{R}_N$) and non-tissue region ($\tilde{R}_J$), alongside pathological property keyword set ($\mathcal{P}$). All results are reported as IoU (↑) / Dice (↑). The best and second-best results in each column are highlighted in **bold** and underline, respectively.

Method	Components			PAIP Prostate		PAIP HCC		PAIP COLON		Camelyon 16		Avg.	
	$\tilde{R}_N$	$\tilde{R}_J$	$\mathcal{P}$	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)
RePath (Ours)	✓	✓	✓	**67.4**	**78.1**	78.1	85.5	63.0	75.0	**45.9**	**52.7**	**63.6**	**72.8**
(a) *w/o* $\tilde{R}_N$ in DAI	✗	✓	✓	33.9	44.9	57.0	68.3	62.0	74.1	15.0	20.4	42.0	51.9
(b) *w/o* $\tilde{R}_J$ in DAI	✓	✗	✓	66.7	77.3	76.0	83.7	**65.8**	**76.9**	32.1	39.0	60.2	69.2
(c) *w/o* DAI	✗	✗	✓	41.3	52.3	65.0	74.6	65.4	76.6	22.0	28.4	48.4	58.0
(d) *w/o* $\mathcal{P}$	✓	✓	✗	65.8	77.0	**81.5**	**88.3**	58.8	71.9	43.8	52.6	62.5	72.4

Implementation details. We use three histopathology VLM backbones: CONCH (ViT-B/16) (Lu et al., 2024), PathGen-CLIP (Sun et al., 2025), and Patho-CLIP-B (Zhang et al., 2025). All slides are pre-processed with the CLAM framework (Lu et al., 2021), using 512×512 patches for CM-16/CM-17 and 256×256 for PAIP. The softmax temperature τ is set per backbone to match each model's softmax scale (0.05 for CONCH, 0.09 for Patho-CLIP-B, 0.005 for PathGen-CLIP), and is held fixed across all datasets. For RePath, we use a fixed anchor budget $N_{max} = 50$, confidence threshold $\delta = 0.2$, and pooling temperature $\beta = 0.5$ across all datasets and backbones without further tuning. Class-specific keyword sets are generated with the Gemini 2.5 Pro API (Comanici et al., 2025). All experiments run on a single NVIDIA RTX A5000 GPU; further details are deferred to the Appendix.

5.2 Quantitative Comparison Results

Quantitative Comparison on WSI Tumor Segmentation. Table 1 compares RePath against zero-shot baselines and four 10-WSI few-shot methods: CLAP (Silva-Rodriguez et al., 2024), Tip-Adapter (Zhang et al., 2022a), KgCoOp (Yao et al., 2023), and CoOp (Zhou et al., 2022b). Without any annotation or training, RePath obtains the best average IoU on all three VLM backbones (63.6 on CONCH, 50.3 on PathGen-CLIP, 50.5 on Patho-CLIP-B), improving over the strongest few-shot competitor by 6.6, 0.4, and 0.9 IoU, indicating that intra-slide affinity provides a reliable inference paradigm independent of backbone or annotation. On CONCH, RePath additionally ranks first on every dataset (e.g., 67.4 / 78.1 IoU/Dice on Prostate and 78.1 / 85.5 on HCC). Notably, even the few-shot 10-WSI setting still requires pixel-level annotation over thousands of patches per slide, a cost that RePath bypasses entirely. In the Appendix, RePath further surpasses weakly supervised MIL baselines on Camelyon16 (21.4–39.6 vs. 45.9 IoU), while dense supervision reaches 57.9 IoU.

Generalizability Across Multi-cohort Evaluation. Table 2 evaluates RePath across five independent cohorts of Camelyon 17. Few-shot methods optimized on a single center often exhibit instability, with CLAP and KgCoOp underperforming vanilla CONCH on every center, and CoOp on four of five, due to center-specific over-fitting. In contrast, RePath surpasses all baselines in IoU on every held-out center and in both metrics on average, without additional training, providing a robust and scalable approach that maintains high precision across varying clinical environments.

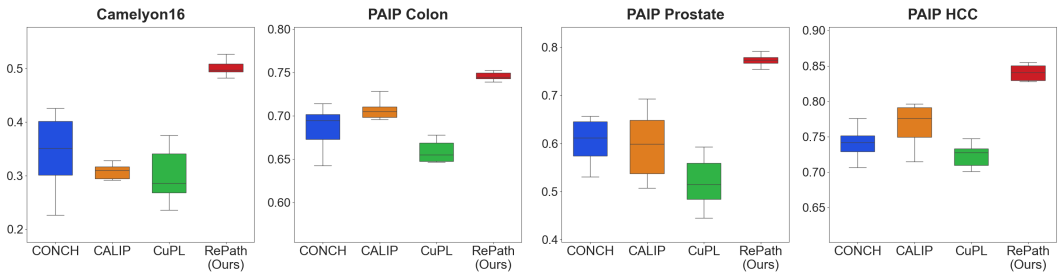

Figure 3: Prompt dependency analysis across four organ datasets. We report the mean and standard deviation of Dice scores over 10 different prompt sets generated by an LLM. RePath consistently achieves superior performance with significantly lower variance compared to zero-shot baselines, demonstrating its robustness against prompt variations through intra-slide patch affinity analysis.

Table 4: Impact of tumor anchor patch selection strategies ($\mathcal{A}_T$). We compare RePath with two sampling strategies from Ground Truth (GT) regions (Random vs. High-confidence) across four datasets. Results are reported as IoU (↑) / Dice (↑). The best and second-best results in each column are highlighted in **bold** and underline, respectively.

Strategy	PAIP Prostate		PAIP HCC		PAIP COLON		Camelyon 16	
	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)	IoU (↑)	Dice (↑)
GT (Random)	65.7	75.8	**80.6**	**87.5**	**66.6**	**77.7**	42.9	50.8
GT (High-confidence)	**69.5**	**79.5**	80.3	87.4	63.7	75.8	**47.1**	**54.5**
RePath (Ours)	67.4	78.1	78.1	85.5	63.0	75.0	45.9	52.7

5.3 Ablation Studies and Analysis

Effectiveness of RePath Components. Table 3 evaluates each component across four datasets. The full RePath achieves the highest overall performance (Avg IoU 63.6 / Dice 72.8), with removing $\tilde{R}_N$ causing the largest drop. Notably, $\tilde{R}_J$ alone (a) performs worse than no DAI (c), as non-tissue affinity suppresses artifacts but cannot resolve tumor–normal ambiguity without $\tilde{R}_N$, potentially perturbing tissue scores. Removing $\tilde{R}_J$ or P occasionally improves HCC or Colon, where large, cohesive tumors with distinct boundaries benefit less from additional artifact suppression or fine-grained morphological prompts; nevertheless, the full framework remains the most globally robust. On Camelyon16, an additional ablation in the Appendix shows that random anchors and cosine affinity reduce IoU from 45.9 to 7.6 and 41.9, respectively, supporting both anchor selection and JSD.

Robustness to Hyperparameters. Our hyperparameters (N_{max}, δ) are selected once, as described in Sec. 5.1, and reused unchanged across all datasets and backbones in Table 1. On CONCH, performance is stable across the anchor budget (Avg IoU 63.2–63.6 for $N_{max} \in \{30, 50, 70, 90\}$). For δ, the working range depends on backbone confidence calibration. CONCH and PathGen-CLIP tolerate $\delta \in [0, 0.4]$, while Patho-CLIP-B has a narrower range of $[0, 0.2]$. We therefore use $\delta = 0.2$ across all backbones, avoiding per-dataset tuning. Detailed results are provided in the Appendix.

Robustness against Prompt Dependency. Fig. 3 compares RePath against zero-shot baselines (CONCH, CuPL (Pratt et al., 2023), CALIP (Guo et al., 2023)) over 10 LLM-generated prompt sets per category. While patch-wise baselines fluctuate substantially with prompt wording, RePath attains the highest mean and lowest variance in Dice, since its relational patch affinity prioritizes inter-patch structure over absolute image-text alignment.

Impact of Anchor Patch Selection Strategies. Table 4 compares RePath against an oracle that draws tumor anchors ($\mathcal{A}_T$) from GT regions via Random or High-confidence sampling. Random GT sampling is unstable, even underperforming RePath on Prostate and Camelyon16, whereas High-confidence sampling performs best on average, indicating that prioritizing representative patterns matters even within expert labels. By adopting the same high-confidence mechanism, RePath closely approaches this oracle across all datasets, automating expert-level anchor selection without annotation.

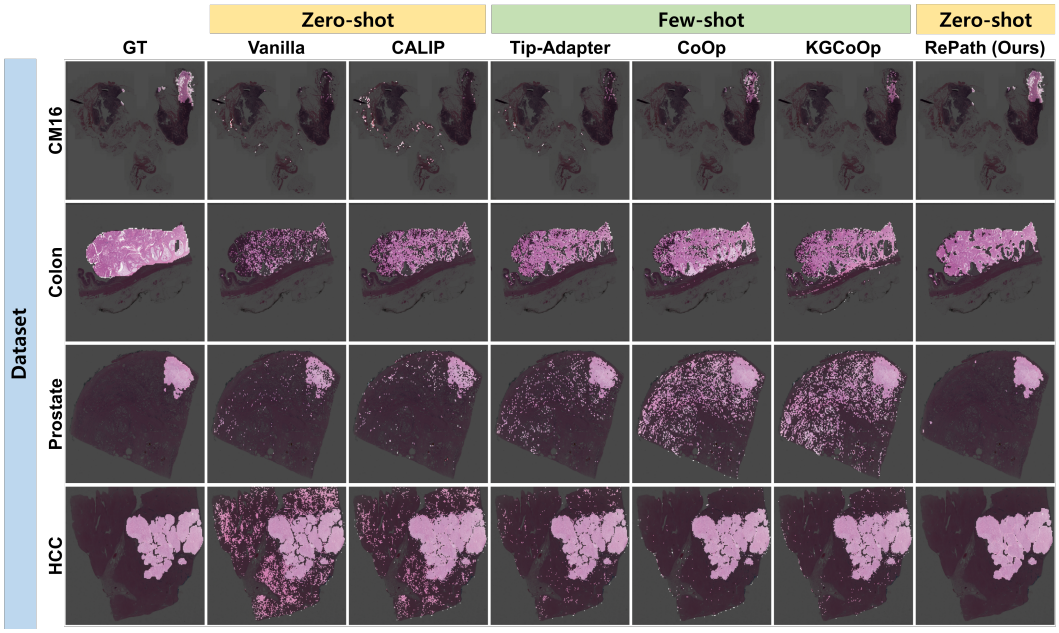

Figure 4: Qualitative comparison of WSI tumor segmentation. Compared with zero-shot (Vanilla CONCH, CALIP) and 10-shot (Tip-Adapter, CoOp, KgCoOp) baselines, RePath produces smoother, more contiguous masks across four organ datasets. RePath reduces spatial inconsistency and fragmentation from independent patch-wise classification, improving boundary alignment without supervision.

Qualitative Analysis. Fig. 4 compares RePath with zero-shot and few-shot (10-WSI) baselines. Both baseline groups suffer from spatial inconsistency and fragmented predictions due to independent patch-wise classification, whereas RePath produces smoother, more contiguous masks aligned closely with ground-truth boundaries without manual supervision. This stems from our intra-slide patch affinity, which leverages global morphological context to resolve local ambiguities. Notably, the lighter tumor region in the Colon example remains challenging for RePath and other CONCH-based zero-shot methods, suggesting a limitation in recognizing this morphology with the underlying VLM.

6 Discussion

While RePath advances zero-shot WSI segmentation, several aspects warrant discussion. First, despite Camelyon16's smaller tumor scale than PAIP, size-stratified analysis (Appendix) shows RePath improves over vanilla CONCH by +10.7 IoU on small tumors ($\leq$500 patches), with 91.2% detection. Second, affinity analysis adds modest inference overhead ($\sim$0.24 sec/WSI on a single NVIDIA RTX A5000 GPU, excluding embedding extraction), with detailed analysis in the Appendix. Third, on Camelyon17 cross-center evaluation, few-shot baselines overfit to their training center and fail to generalize, while training-free analysis remains robust across all centers, suggesting a paradigm-level advantage under clinical distribution shift.

7 Conclusion

We introduced RePath, a novel framework that shifts zero-shot gigapixel WSI tumor segmentation from per-patch classification to inter-patch relational comparison within the slide. By grounding predictions in patch-to-patch distributional similarity and contrasting tumor signals against normal and non-tissue patterns, RePath achieves spatially coherent segmentation without any manual annotation, outperforming zero-shot baselines and, on average across four organs and three VLM backbones, matching few-shot supervised methods. More broadly, our results suggest that inference formulation matters as much as representation quality for VLM-based WSI analysis, and that similar relational strategies may benefit other slide-level tasks such as grading, subtyping, or biomarker prediction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the execution and analysis of experiments and the interpretation of experimental results. Gemini 2.5 Pro was also used to generate the histopathology-specific keywords used by RePath. We did not use generative AI to formulate mathematical claims or proofs, and other required disclosure tasks are not applicable to this work. Additionally, generative AI tools were used for software code assistance, literature discovery and organization, and drafting, editing, and polishing portions of the manuscript.

All AI-assisted work was reviewed by the authors. AI-assisted code and experimental outputs were verified through execution and evaluation, generated keywords were manually reviewed before use, and AI-assisted references and manuscript content were checked by the authors for correctness. We take responsibility for the final content of this work, including text, claims, code, and experimental results produced with the aid of generative AI.

ETHICS STATEMENT

This study uses publicly available, de-identified histopathology datasets and does not involve new data collection or direct interaction with human subjects.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sec. 5.1 provides the core implementation settings, including preprocessing, backbone-specific parameters, and RePath hyperparameters. Appendix A details the datasets and evaluation splits; Appendices B and C provide the training and implementation settings for all few-shot and zero-shot baselines, respectively; and Appendix D describes the complete prompt construction and keyword-generation procedure. We will release the source code and complete generated keyword sets upon publication.

REFERENCES

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715, 2019.

Kaustav Bera, Ian Katz, and Anant Madabhushi. Reimagining t staging through artificial intelligence and machine learning image processing approaches in digital pathology. *JCO Clinical Cancer Informatics*, 4:1039–1050, 2020.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10662–10671, 2019.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5134–5143, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier

with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical image analysis*, 80:102485, 2022.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 31(11):3749–3761, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAMELYON16 consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 746–754, 2023.

Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2: 147–171, 2009.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, 2025.

Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Zeyan Xu, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis*, 80:102487, 2022.

Lewis A Hassell, Marika L Forsythe, Ami Bhalodia, Thanh Lan, Tasnuva Rashid, Astin Powers, Marilyn M Bui, Arlen Brickman, Qiangqiang Gu, Andrey Bychkov, et al. Toward optimizing the impact of digital pathology and augmented intelligence on issues of diagnosis, grading, staging, and classification. *Modern Pathology*, 38(7):100765, 2025.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

Dong Un Kang, Hayeon Kim, and Se Young Chun. Class distribution-induced attention map for open-vocabulary semantic segmentations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Kyungmo Kim, Kyoungbun Lee, Sungduk Cho, Dong Un Kang, Seongkeun Park, Yunsook Kang, Hyunjeong Kim, Gheeyoung Choe, Kyung Chul Moon, Kyu Sang Lee, et al. Paip 2020: Microsatellite instability prediction in colorectal cancer. *Medical Image Analysis*, 89:102886, 2023.

Yoo Jung Kim, Hyungjoon Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Choyeon Hong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, et al. Paip 2019: Liver cancer segmentation challenge. *Medical image analysis*, 67:101854, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European conference on computer vision*, pp. 143–160. Springer, 2024.

Yi Li, Yiduo Yu, Yiwen Zou, Tianqi Xiang, and Xiaomeng Li. Online easy example mining for weakly-supervised gland segmentation from histology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 578–587. Springer, 2022.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22718–22727, 2025.

Siyu Lin, Haowen Zhou, Mark Watson, Ramaswamy Govindan, Richard J Cote, and Changhuei Yang. Impact of stain variation and color normalization for prognostic predictions in pathology. *Scientific reports*, 15(1):2369, 2025.

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

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19764–19775, 2023.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Mohamed Omar, Mohammad K Alexanderani, Itzel Valencia, Massimo Loda, and Luigi Marchionni. Applications of digital pathology in cancer: a comprehensive review. *Annual Review of Cancer Biology*, 8(1):245–268, 2024.

Wentao Pan, Jiangpeng Yan, Hanbo Chen, Jiawei Yang, Zhe Xu, Xiu Li, and Jianhua Yao. Human-machine interactive tissue prototype learning for label-efficient histopathology image segmentation. In *International Conference on Information Processing in Medical Imaging*, pp. 679–691. Springer, 2023.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15691–15701, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Jiangbo Shi, Tieliang Gong, Chunbao Wang, and Chen Li. Semi-supervised pixel contrastive learning framework for tissue segmentation in histopathological image. *IEEE Journal of Biomedical and Health Informatics*, 27(1):97–108, 2022.

Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195: 105637, 2020.

Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 23681–23690, 2024.

Vladan Stojnić, Yannis Kalantidis, and Giorgos Tolias. Label propagation for zero-shot classification with vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23209–23218. IEEE, 2024.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, et al. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, pp. 94611–94653, 2025.

Qingchen Tang, Lei Fan, Maurice Pagnucco, and Yang Song. Prototype-based image prompting for weakly supervised histopathological image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30271–30280, 2025.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11343–11352. IEEE, 2024.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2725–2736, 2023.

Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi. Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical image analysis*, 68:101890, 2021.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European conference on computer vision*, pp. 315–332. Springer, 2024.

Yuxin Wu, Xiongjun Xu, Yuan Cheng, Xiuming Zhang, Fanxi Liu, Zhenhui Li, Lei Hu, Anant Madabhushi, Peng Gao, Zaiyi Liu, et al. Beex is an open-source tool that evaluates batch effects in medical images to enable multicenter studies. *Cancer research*, 85(2):218–230, 2025.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6757–6767, 2023.

Jingwei Zhang, Ke Ma, Saarthak Kapse, Joel Saltz, Maria Vakalopoulou, Prateek Prasanna, and Dimitris Samaras. SAM-Path: A segment anything model for semantic segmentation in digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops*, pp. 161–170. Springer Nature Switzerland, 2023a. doi: 10.1007/978-3-031-47401-9_16.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022a.

Shaoteng Zhang, Jianpeng Zhang, and Yong Xia. Transws: Transformer-based weakly supervised histology image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 367–376. Springer, 2022b.

Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 109–118. Springer, 2023b.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. *arXiv preprint arXiv:2505.11404*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15659–15669, 2023.

A Dataset Details

We provide detailed descriptions and split statistics for the four WSI tumor segmentation benchmarks used in this work.

Camelyon16 (CM-16) (Ehteshami Bejnordi et al., 2017). A benchmark dataset for breast cancer metastasis detection in sentinel lymph nodes, originally comprising 270 training and 130 test WSIs with pixel-level annotations. Since our task targets tumor segmentation rather than slide-level detection, we evaluate exclusively on tumor-bearing slides, yielding 111 training and 48 test WSIs. For few-shot baselines, we randomly sample 10 WSIs from the training set per trial.

Pathology Artificial Intelligence Platform (PAIP) (Kim et al., 2023; 2021). A WSI diagnostic resource spanning multiple organs. We use the Liver (HCC), Prostate, and Colon subsets, comprising 503, 599, and 898 WSIs respectively. For each organ, we partition the WSIs into train/test using a 2:8 ratio (101/402 for HCC, 120/479 for Prostate, 180/718 for Colon). Following the same protocol as Camelyon16, we randomly sample 10 WSIs from the training set per trial for few-shot baselines.

Camelyon17 (CM-17) (Bandi et al., 2018). A multi-center WSI dataset of breast cancer sentinel lymph nodes collected from five Dutch medical centers, with 1,000 WSIs from 200 patients in total. Pixel-level tumor masks are publicly released for only 50 WSIs (10 per center), and we restrict our evaluation to this annotated subset. We use Camelyon17 to assess cross-center generalization under a leave-one-center-out protocol: for each center, its 10 WSIs serve as the few-shot training set while the 40 WSIs from the remaining four centers are used for testing, repeated across all five centers.

B Training Details of Few-shot Baselines

We provide the detailed experimental setups for the few-shot baselines, including Tip-Adapter (Zhang et al., 2022a), CoOp (Zhou et al., 2022b), KgCoOp (Yao et al., 2023), and CLAP (Silva-Rodriguez et al., 2024). To ensure a fair comparison, all few-shot experiments were conducted over 10 independent trials. For each trial, we used seeds from 1 to 10 to randomly sample $k = 10$ whole-slide images (WSIs) from the designated training set of each dataset to serve as the few-shot training samples.

B.1 Tip-Adapter

While Tip-Adapter (Zhang et al., 2022a) typically utilizes 1 to 16-shot samples in standard image classification, the gigapixel resolution of WSIs results in tens of thousands of patches even from a single slide. To maintain efficiency and prevent memory issues while capturing representative histological features, we randomly selected 100 tumor patches and 100 normal patches from the 10 training WSIs. For the hyperparameters, we set the scale factor $\alpha = 1.0$ and the residual ratio $\beta = 5.5$ for all experiments.

B.2 CoOp

We optimized a separate CoOp (Zhou et al., 2022b) model for each organ dataset. We utilized 16 learnable context tokens and formulated the task as a binary classification between "normal" and "tumor" classes. The learnable prompts were optimized for 20 epochs with a learning rate of 1×10^{-4} while keeping the VLM backbone frozen. The training was conducted using all patches extracted from the 10 training slides to ensure the learned prompts captured the diverse morphological variations across the slides.

B.3 KgCoOp

KgCoOp (Yao et al., 2023) extends CoOp by adding a regularization term that minimizes the discrepancy between learnable prompts and hand-crafted text descriptions, preventing the learned context from drifting too far from meaningful language semantics. We followed the same training protocol as CoOp (20 epochs, learning rate 1×10^{-4}, 16 learnable tokens, backbone frozen) with the addition of the knowledge-guided regularization loss. The hand-crafted prompts used for regularization were

generated using the same organ-specific keyword sets as RePath to ensure a fair comparison. Training was performed on all patches from the 10 training WSIs per trial.

B.4 CLAP

CLAP (Silva-Rodriguez et al., 2024) revisits linear probing for few-shot VLM adaptation by constructing class-adaptive prototypes that combine visual and textual information. Instead of learning a single linear classifier, CLAP builds per-class prototypes from the few-shot support set and refines them with the text encoder's class embeddings. We followed the default hyperparameters from the original implementation and used all available patches from the 10 training WSIs as the support set, consistent with the other few-shot baselines.

C Implementation Details of Zero-shot Baselines

We describe the implementation details for the zero-shot baselines, including CuPL (Pratt et al., 2023) and CALIP (Guo et al., 2023). To ensure consistency, all baseline evaluations were conducted using the same VLM backbones as RePath.

C.1 CuPL

For CuPL (Pratt et al., 2023), we generated customized descriptive prompts using the Gemini 2.5 Pro API (Comanici et al., 2025). Following the methodology of CuPL, we utilized three LLM-prompt templates designed to extract specific visual characteristics for "normal tissue" and "tumor" classes, incorporating organ-specific information. To align these descriptions with the histopathological domain, the generation process was strictly constrained to focus on microscopic morphological features as observed in H&E stained pathology slides. By explicitly excluding extraneous information such as clinical symptoms or macroscopic findings, we ensured that the LLM produced only visual descriptors suitable for patch-wise classification. We produced 10 distinct descriptive sentences per template, resulting in 30 prompts per class, with the sampling temperature set to 0.99 to ensure diversity.

C.2 CALIP

We implemented CALIP (Guo et al., 2023) using its parameter-free attention mechanism to enhance the zero-shot interaction between visual and textual embeddings. For the attention temperature, we set $\alpha_t = 2.0$ and $\alpha_s = 2.0$ for the textual and visual branches, respectively. To combine the resulting logits, we employed an ensemble weight of $\beta_1 = 1.0$ for the original CLIP logit, $\beta_2 = 5.0$ for the visual-guided logit, and $\beta_3 = 0.1$ for the textual-blended logit. We observed that a higher weight for the visual-guided logit (β_2) was particularly effective for capturing the complex spatial features inherent in WSI patches.

D Text Prompt Preparation for Histopathology VLMs

We detail the systematic process of generating text prompts to provide rich pathological context to the Vision-Language Models (VLMs). To ensure robust representation of each class, we employ an ensemble-based embedding approach that combines diverse prompt templates with domain-specific keywords generated by a Large Language Model (LLM).

D.1 Prompt Templates and Embedding Generation

To generate text embeddings, we adopted 23 prompt templates established in prior histopathology VLM research (e.g., CONCH (Lu et al., 2024)). These templates are specifically designed to provide pathological context to the text encoder. For each keyword generated via the LLM, we populate the `CLASSNAME` placeholder across all 23 templates. The final text embedding for a specific keyword is obtained by averaging the individual embeddings of these 23 prompted sentences. This procedure is applied consistently to all baseline models that require text prompts, including Vanilla histopathology VLMs (CONCH (Lu et al., 2024), PathGen-CLIP (Sun et al., 2025), and Patho-CLIP-B (Zhang et al.,

2025)), CALIP (Guo et al., 2023), Tip-Adapter (Zhang et al., 2022a), and our proposed RePath. The templates are listed below:

- a photomicrograph showing {CLASSNAME}.
- a photomicrograph of {CLASSNAME}.
- an image of {CLASSNAME}.
- an image showing {CLASSNAME}.
- an example of {CLASSNAME}.
- {CLASSNAME} is shown.
- this is {CLASSNAME}.
- there is {CLASSNAME}.
- a histopathological image showing {CLASSNAME}.
- a histopathological image of {CLASSNAME}.
- a histopathological photograph of {CLASSNAME}.
- a histopathological photograph showing {CLASSNAME}.
- shows {CLASSNAME}.
- presence of {CLASSNAME}.
- {CLASSNAME} is present.
- an H&E stained image of {CLASSNAME}.
- an H&E stained image showing {CLASSNAME}.
- an H&E image showing {CLASSNAME}.
- an H&E image of {CLASSNAME}.
- {CLASSNAME}, H&E stain.
- {CLASSNAME}, H&E.
- Pathological verification showing {CLASSNAME}.
- An H&E-stained slide confirming {CLASSNAME}.

D.2 Generation of Keyword Sets for $\mathcal{K}$

To construct the comprehensive keyword set $\mathcal{K} = \{\mathcal{T} \cup \mathcal{N} \cup \mathcal{J} \cup \mathcal{P}\}$, we utilized the LLM to generate distinct categories of keywords that capture both diagnostic and morphological nuances. Specifically, we employed the Gemini 2.5 Pro API (Comanici et al., 2025) to obtain domain-specific keywords for the Tumor ($\mathcal{T}$), Normal ($\mathcal{N}$), Non-tissue ($\mathcal{J}$), and Pathological Property ($\mathcal{P}$) sets.

We observed that naive keyword generation often yields non-visual clinical information, such as prognosis or treatment protocols, which cannot be captured from image patches and consequently degrades VLM performance. To address this, we defined a specialized system prompt for the LLM to act as a Board-Certified Pathologist and Computer Vision Researcher. To ensure the generation of keywords strictly focused on histopathological visual characteristics, we incorporated specific instructions into the system prompt as shown in Fig. A5. This "Visual Anchor" mindset forces the model to decompose high-level diagnoses into pixel-level attributes (Color, Texture, and Shape) and maintain a patch-level perspective (e.g., $40\times$ magnification) while avoiding non-visual diagnostic terms.

Table A1: Target organs and tumor types for keyword generation using LLM.

Dataset	Organ	Tumor Type (Target)
Camelyon16	Lymph node (Breast)	Metastatic carcinoma
PAIP Colon	Colon	Colorectal adenocarcinoma
PAIP Prostate	Prostate	Prostate acinar adenocarcinoma
PAIP HCC	Liver	Hepatocellular carcinoma

Target-specific Sets ($\mathcal{T}$ and $\mathcal{N}$). For each dataset, we utilized the primary tumor types listed in Table A1 to generate 10 distinct keywords for both Tumor ($\mathcal{T}$) and Normal ($\mathcal{N}$) classes. To ensure contextual specificity and high-quality visual anchors, we designed a structured generation prompt for the LLM, as illustrated in Fig. A6. This prompt instructs the LLM to provide precise diagnostic terms for $\mathcal{T}$ (e.g., "Colorectal adenocarcinoma" or clinical abbreviations) and specific micro-anatomical terms for $\mathcal{N}$ (e.g., "Normal Colonic Crypts" or organ-specific parenchyma). By avoiding generic terms like "Tumor" or "Tissue," we encourage the model to identify more discriminative features that naturally align with the organ's histopathological context.

Non-tissue Set ($\mathcal{J}$). The Non-tissue Set ($\mathcal{J}$) comprises keywords representing slide-level artifacts and background noise, which are universally present across all histopathology datasets. To effectively isolate these non-biological elements from the affinity analysis, we utilized the LLM to define five fixed categories of artifacts, as illustrated in Fig. A7. These include: (1) "background debris, stain artifact", (2) "air bubble, dust particles", (3) "torn tissue, folded tissue", (4) "white background, empty space", and (5) "marker pen ink, glass edge". This set allows the model to explicitly assign low affinity to non-informative regions, thereby enhancing the robustness of tumor segmentation.

Pathological Property Set ($\mathcal{P}$). The Pathological Property Set ($\mathcal{P}$) is designed to capture a wide spectrum of tissue characteristics, such as color and morphology, which are independent of the diagnostic status (malignant vs. benign). For each organ and tumor type listed in Table A1, we generated a total of 30 distinct keywords to enrich the visual context. As illustrated in Fig. A8, the generation logic is divided into two categories: (1) Staining (15 keywords), which describe H&E staining intensities and color properties of the cytoplasm and nucleus (e.g., "Pale pink cytoplasm"), and (2) Morphology (15 keywords), which describe physical cell shapes and stroma arrangements (e.g., "Spindle-shaped cells"). By providing these agnostic morphological descriptors, we can stabilize the class distribution and improve the reliability of the relational affinity analysis.

Example Keyword Set. To illustrate the output of our LLM-based generation pipeline, Table A2 presents representative keywords from one of the prompt sets used for Camelyon16 (lymph node, metastatic carcinoma). For brevity, we show 5 examples per category for $\mathcal{T}$, $\mathcal{N}$, and $\mathcal{P}$, while $\mathcal{J}$ is reported in full as it consists of only 5 fixed categories shared across all datasets. The complete keyword sets across all datasets and organ–tumor combinations will be released together with our code.

Table A2: Representative keywords from one prompt set for Camelyon16 (metastatic carcinoma in lymph nodes). Tumor ($\mathcal{T}$) and Normal ($\mathcal{N}$) contain 10 keywords each; the Non-tissue set ($\mathcal{J}$) contains 5 fixed categories shared across datasets; the Pathological Property set ($\mathcal{P}$) contains 30 keywords (15 Staining + 15 Morphology). The "#" column indicates the number of examples shown out of the total set size.

Category	#	Example Keywords
Tumor ($\mathcal{T}$)	5/10	Metastatic Carcinoma; Nodal Metastasis of Carcinoma; Infiltrating Nests of Malignant Epithelial Cells; Atypical Cells with Pleomorphic Nuclei and Prominent Nucleoli; Carcinoma Cells with High N/C Ratio
Normal ($\mathcal{N}$)	5/10	Benign Lymphoid Tissue; Normal Lymph Node Parenchyma; Lymphoid Follicle with Germinal Center; Paracortical T-cell Zone; Subcapsular Sinus of Lymph Node
Non-tissue ($\mathcal{J}$)	5/5	background debris, stain artifact; air bubble, dust particles; torn tissue, folded tissue; white background, empty space; marker pen ink, glass edge
Pathological Property ($\mathcal{P}$):		
Staining	5/15	Intensely eosinophilic cytoplasm; Pale pink cytoplasm; Hyperchromatic blue nuclei; Vesicular nuclei with pale chromatin; Golden-brown pigment granules
Morphology	5/15	Spindle-shaped cells; Sheets of cohesive cells; Gland-like or acinar structures; Irregular nuclear contours; Dense collagenous stroma

E From the CDAM Observation to a Zero-Shot WSI Tumor Segmentation Framework

We adopt Jensen–Shannon Divergence over class distributions as our core distance metric, building on the empirical observation of CDAM (Kang et al., 2025) that patches of the same category share strongly correlated class distributions.

Around this core, RePath builds a complete framework for zero-shot whole-slide image (WSI) tumor segmentation with histopathology vision–language models (VLMs). The framework is composed of slide-internal anchor selection, an LLM-augmented keyword space tailored to histopathology, JSD-based relational affinity scoring, and Differential Affinity Integration inspired by the clinical practice of differential diagnosis. Together, these design choices target the structural failure modes of patch-tiled VLM inference—prompt dependency and inter-patch spatial inconsistency—in a setting (gigapixel histopathology, training-free, zero-shot) that the CDAM observation alone does not address.

F Why Training-Free Open-Vocabulary Segmentation Methods Are Not Directly Applicable to WSI Segmentation

A growing family of training-free open-vocabulary segmentation (OVSS) methods—including SCLIP (Wang et al., 2024), NACLIP (Hajimiri et al., 2025), and ClearCLIP (Lan et al., 2024)—achieve dense prediction by manipulating CLIP's internal attention. We show that this family shares a structural assumption that is incompatible with WSI segmentation.

Shared mechanism of OVSS. Given a single image, a ViT vision encoder produces M visual tokens $\{\mathbf{h}_j\}_{j=1}^{M}$ (e.g., $M = 14 \times 14$) and a CLS token. Standard CLIP self-attention at the final layer L computes

$$\mathrm{Attn}^{(L)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}^{(L)}{\mathbf{K}^{(L)}}^{\top}}{\sqrt{d}}\right)\mathbf{V}^{(L)}, \tag{6}$$

where $\mathbf{Q}^{(L)}, \mathbf{K}^{(L)}, \mathbf{V}^{(L)} \in \mathbb{R}^{(M+1)\times d}$. OVSS methods replace $\mathrm{Attn}^{(L)}$ with a function ϕ of these projections to produce spatially discriminative tokens,

$$\hat{\mathrm{Attn}}^{(L)} = \phi\Big(\mathbf{Q}^{(L)}, \mathbf{K}^{(L)}, \mathbf{V}^{(L)}\Big), \tag{7}$$

with ϕ realized through correlative self-attention (SCLIP), neighbor-aware attention with a Gaussian prior (NACLIP), or residual stream and FFN removal (ClearCLIP). All require the M visual tokens to be *co-present within a single forward pass*. The refined tokens are then classified against text embeddings $\mathbf{w}_c$ to yield a dense $\sqrt{M} \times \sqrt{M}$ map within that image.

WSI segmentation operates at a different granularity. A gigapixel WSI cannot be fed into a ViT as one image. The slide is decomposed into N patches $\{x_i\}_{i=1}^{N}$ ($N \sim 10^3$–10^4), each processed independently through the image encoder f_I:

$$\mathbf{z}_i = f_I(x_i) \in \mathbb{R}^d, \quad i = 1, \ldots, N, \tag{8}$$

where $\mathbf{z}_i$ is the CLS embedding of patch x_i. The internal visual tokens within each patch are discarded after pooling, and the slide-level mask assigns a single label per patch:

$$M(x_i) = \arg\max_c \; g(\mathbf{z}_i, \mathbf{w}_c), \tag{9}$$

where g is a scoring function over CLS embeddings.

Structural incompatibility. The function ϕ in Eq. 7 alters the standard self-attention that feeds the CLS token, *degrading the CLS embedding that CLIP-based WSI segmentation relies on*. ClearCLIP explicitly removes the residual stream and FFN, while SCLIP and NACLIP replace $\mathbf{QK}^{\top}$ with self-correlations or neighbor priors; in all cases, the CLS attention pattern that aggregates global image semantics is broken. Applying OVSS to a WSI patch therefore yields a CLS embedding that is *worse than vanilla CLIP* for patch-level image–text matching, the standard mode of zero-shot WSI tumor segmentation.

RePath instead constructs relational affinity directly over the available CLS embeddings $\{\mathbf{z}_i\}_{i=1}^{N}$ via JSD on per-patch class distributions, operating at the inter-patch granularity needed to address slide-level consistency.

G Additional Quantitative Results

G.1 Ablation Study on Anchor Threshold δ

We conducted an ablation study to investigate the impact of the anchor threshold δ on the segmentation performance. The parameter δ serves as a critical confidence filter for selecting anchor patches during the construction of the Relational Patch Affinity Map. To ensure the generalizability of our findings, we evaluated the performance across four datasets using three state-of-the-art histopathology VLMs: CONCH (ViT-B/16) (Lu et al., 2024), PathGen-CLIP (Sun et al., 2025), and Patho-CLIP-B (Zhang et al., 2025).

As shown in Table A3, we evaluated the impact of the anchor threshold δ by varying it from 0.0 to 0.6. Across all three VLMs, $\delta = 0.2$ consistently yielded robust performance, achieving the highest average IoU and Dice scores for PathGen-CLIP and Patho-CLIP-B (50.3% and 50.5% IoU, respectively) and near-optimal performance for CONCH. While increasing δ to 0.4 provides marginal gains for CONCH, it leads to a significant performance degradation for the other VLMs—most notably in Patho-CLIP-B, where scores drop sharply from $\delta = 0.4$. These results demonstrate that $\delta = 0.2$ provides the optimal balance for selecting reliable anchor patches while filtering out low-confidence noise across diverse histopathology datasets.

Table A3: Ablation study on the anchor threshold δ across three histopathology VLMs. Segmentation performance (IoU and Dice, %) is reported across four datasets and their average.

VLM	δ	**PAIP Prostate**		**PAIP HCC**		**PAIP Colon**		**Camelyon 16**		**Avg.**	
		IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
CONCH	0	64.1	74.8	74.7	82.8	61.6	73.9	43.4	50.9	61.0	70.6
	0.2	67.4	78.1	78.1	85.5	63.0	75.0	45.9	52.7	63.6	72.8
	0.4	69.1	79.7	78.5	85.9	63.6	75.4	45.7	53.2	64.2	73.6
	0.6	68.8	79.2	78.2	85.6	63.9	75.5	43.4	50.7	63.6	72.8
PathGen-CLIP	0	36.2	47.7	64.9	74.4	52.1	66.0	29.9	37.4	45.8	56.4
	0.2	45.9	57.9	66.9	76.5	56.3	69.5	32.2	40.2	50.3	61.0
	0.4	36.6	48.0	66.7	76.4	57.3	70.4	28.5	35.1	47.3	57.5
	0.6	36.7	48.1	66.2	76.1	58.0	70.9	26.0	31.4	46.7	56.6
Patho-CLIP-B	0	34.3	46.6	71.4	80.2	55.6	69.1	24.9	31.2	46.6	56.8
	0.2	45.2	57.5	72.6	81.3	60.1	72.7	24.2	30.5	50.5	60.5
	0.4	38.6	50.1	41.5	48.3	52.1	64.9	6.2	7.1	34.6	42.6
	0.6	0.0	0.0	0.0	0.0	2.1	2.7	0.0	0.0	0.5	0.7

G.2 Ablation Study on Anchor Budget N_{max}

We further ablate the per-category anchor budget N_{max} on the CONCH backbone across all four datasets. As shown in Table A4, performance is stable across $N_{max} \in \{30, 50, 70, 90\}$, with average IoU varying within 0.4 points (63.2–63.6) and average Dice within 0.5 points (72.3–72.8). We adopt $N_{max} = 50$ in all main experiments as a balanced default.

G.3 Robustness to the Choice of Keyword-Generation LLM

The keyword set $\mathcal{K}$ in our main experiments is generated using the Gemini 2.5 Pro API. A natural concern is whether RePath's gains depend on the specific LLM used: if the framework relies on a strong commercial LLM, its applicability would be limited by API cost, access restrictions, and reproducibility constraints. To verify that RePath is agnostic to the keyword-generation LLM, we replace Gemini 2.5 Pro (Comanici et al., 2025) with Gemma 4 (26B)[1], an open-source model executed

[1] https://ai.google.dev/gemma/docs/core/model_card_4

Table A4: Ablation study on the per-category anchor budget N_{max} (CONCH backbone) across four datasets. All results are reported as IoU (↑) / Dice (↑).

N_{max}	PAIP Prostate		PAIP HCC		PAIP Colon		Camelyon 16		Avg.	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
30	66.8	77.5	78.0	85.4	62.7	74.8	46.0	53.2	63.4	72.7
50 **(Default)**	67.4	78.1	78.1	85.5	63.0	75.0	45.9	52.7	63.6	72.8
70	67.1	77.6	78.1	85.5	63.0	75.0	45.6	52.4	63.5	72.6
90	67.2	77.7	78.1	85.5	62.4	74.3	44.9	51.7	63.2	72.3

locally via Ollama[2], while keeping the system prompt, generation logic (Sec. D.2), and all RePath hyperparameters identical.

Table A5: Effect of the keyword-generation LLM on RePath performance (CONCH backbone). Gemma 4 (26B) is an open-source model run locally; Gemini 2.5 Pro is a commercial API. All other components, including the system prompt, keyword categories, and RePath hyperparameters, are kept identical. Results are reported as IoU (↑) / Dice (↑).

Keyword LLM	Camelyon 16	PAIP Prostate	PAIP HCC	PAIP Colon	Avg.
Gemini 2.5 Pro (API)	45.9 / 52.7	67.4 / 78.1	78.1 / 85.5	63.0 / 75.0	63.6 / 72.8
Gemma 4 26B (open-source)	45.4 / 52.3	67.8 / 78.4	77.2 / 84.8	62.8 / 74.8	63.3 / 72.6

As shown in Table A5, switching from Gemini 2.5 Pro to Gemma 4 incurs only a marginal change in average performance (-0.3 IoU / -0.2 Dice), with per-dataset differences within ± 0.9 IoU. Notably, Gemma 4 even slightly surpasses Gemini on Prostate. This indicates that RePath's effectiveness arises from its intra-slide affinity formulation rather than the specific lexical content of the keyword set, and confirms that the framework can be reproduced end-to-end with freely available, locally executable LLMs without sacrificing segmentation quality.

G.4 Analysis of Computation Costs

We evaluate the computational efficiency of RePath compared to existing zero-shot and few-shot methods. All measurements were conducted using the Camelyon16 dataset, reporting the average inference time per Whole-Slide Image (WSI) in seconds. To ensure a fair comparison of the algorithmic overhead, we exclude the time required for image embedding extraction.

As summarized in Table A6, while RePath requires slightly more inference time than baseline zero-shot methods due to its intra-slide patch affinity analysis, it remains highly feasible within clinical workflows. Notably, few-shot methods involve significant hidden costs that are often overlooked: the time required for expert pathologists to annotate training WSIs and the subsequent model fine-tuning process. Even with a minimal setup of 10 WSIs, the preparation phase for few-shot learning can take hours or days, whereas RePath is ready for immediate deployment without any prior training or labeling.

G.5 Effect of Similarity Metrics

We evaluated the effectiveness of our JSD-based affinity by comparing it with a cosine similarity baseline. In this setup, cosine similarity was employed to measure patch relationships in place of JSD. While the cosine baseline achieves an average IoU of 56.6 and Dice of 67.4, our JSD-based approach yields superior results with an average IoU of 63.6 and Dice of 72.8. This performance gap indicates that JSD captures more discriminative patch affinities, providing a more robust measure for training-free segmentation than standard vector-based similarity.

[2] `https://ollama.com`

Table A6: Comparison of computational costs on the Camelyon16 dataset. All metrics are reported in seconds (sec) per WSI, excluding image embedding extraction time. Preparation costs (Labeling and Training) are required only once per dataset/center for few-shot methods.

Category	Method	Preparation Cost		Test Time
		Labeling	Training (sec)	(sec/WSI)
Few-shot (10-WSI)	CLAP (Silva-Rodriguez et al., 2024)	High (10-WSI)	80.38	0.01
	CoOp (Zhou et al., 2022b)	High (10-WSI)	49.65	0.01
Zero-shot	Vanilla	None	–	0.01
	CALIP (Guo et al., 2023)	None	–	0.01
	RePath (Ours)	None	–	0.24

G.6 Quantitative AUROC Analysis for WSI Tumor Segmentation

To evaluate the effectiveness of the proposed relational affinity mechanism, we conducted a quantitative comparison using the CONCH (ViT-B/16) backbone as the baseline. RePath consistently outperformed the vanilla zero-shot inference across all four benchmark datasets in terms of AUROC. Specifically, in the PAIP Prostate and HCC datasets, RePath improved the baseline AUROC from 77.7% and 84.9% to 88.7% and 91.6%, respectively. Similar performance gains were observed in the PAIP Colon dataset, where the AUROC increased from 77.7% to 83.7%. Notably, on the Camelyon16 dataset, RePath achieved a significant improvement, raising the AUROC from 74.2% to 88.2%. These results demonstrate that our method effectively leverages intra-slide patch relationships to refine classification boundaries, providing more robust and accurate tumor localization compared to independent patch-wise predictions.

G.7 Robustness against Uniform Sampling for Normal Anchor Patch Selection ($\mathcal{A}_N$)

To verify the stability of RePath, we conducted a robustness analysis by employing a uniform sampling strategy for the selection of normal anchor patches ($\mathcal{A}_N$). Instead of using high-confidence reference patches, we randomly sampled a fixed number of normal patches and evaluated the impact on segmentation performance. To ensure statistical reliability, the experiment was repeated over 10 independent trials using the CONCH (Lu et al., 2024) backbone across four datasets. As shown in Table A7, RePath demonstrates remarkably stable mIoU and Dice scores with negligible standard deviations across all organs. This indicates that our relational patch affinity mechanism is robust to the stochasticity of anchor selection, successfully capturing the discriminative contrast between tumor and normal regions.

Table A7: Robustness analysis of RePath on CONCH backbone using uniform sampling for normal anchor ($\mathcal{A}_N$) selection. We report the mean and standard deviation of IoU and Dice (%) across 10 independent trials.

Method	**PAIP Prostate** IoU / Dice	**PAIP HCC** IoU / Dice	**PAIP Colon** IoU / Dice	**Camelyon 16** IoU / Dice
RePath (Ours)	67.3±0.2 / 77.9±0.2	78.0±0.1 / 85.4±0.1	63.0±0.0 / 75.0±0.0	45.8±0.4 / 52.7±0.4

G.8 Comparison with Classical Smoothing and Label Propagation Baselines

To address the concern that RePath's gains may stem from simple spatial smoothing rather than distributional affinity analysis, we compare against five classical smoothing and label propagation baselines. All methods operate on the *same* adjacency matrix and initial VLM predictions as RePath, ensuring a fair comparison of the post-processing mechanism itself.

The baselines are: (1) **Adjacency Smoothing**, which performs graph convolution via the adjacency matrix ($\tilde{\mathbf{s}} = \mathbf{A} \cdot \mathbf{s}$); (2) **Label Propagation (Binary)**, which iteratively propagates labels via $\mathbf{F}^{(t+1)} = \alpha \mathbf{A}\mathbf{F}^{(t)} + (1-\alpha)\mathbf{F}^{(0)}$; (3) **LP Multi-class**, which extends binary LP to three classes (tumor,

normal, non-tissue) and applies $R_T - \max(R_N, R_J)$ for final prediction—the closest structural analog to RePath's DAI; (4) **Majority Vote**, which assigns each patch the majority label among its spatial neighbors; and (5) **CRF**, which combines unary potentials (VLM predictions) with pairwise potentials penalizing label disagreement between neighbors. For each baseline, we report the best result across hyperparameter sweeps.

Table A8: Comparison with classical smoothing and label propagation baselines on the same slide adjacency graph (CONCH backbone). All methods use identical VLM predictions as input. Results are from an independent run with a different LLM-generated prompt set and thus differ slightly from Table 1. Best result per column in **bold**.

Method	Camelyon 16		PAIP Prostate		PAIP HCC		PAIP Colon		Avg	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Vanilla	32.6	42.5	51.3	65.3	64.9	76.0	52.5	66.4	50.3	62.6
Adj. Smoothing (best iter)	28.8	35.7	53.1	65.6	70.3	79.5	**64.0**	74.7	54.0	63.8
Label Prop. (best α)	34.9	44.0	50.8	64.7	73.7	82.0	56.5	68.5	54.0	64.8
LP Multi-class (best α)	41.2	51.6	50.7	64.1	72.8	81.8	55.5	68.4	55.0	66.5
Majority Vote	32.3	39.5	51.3	64.1	70.6	79.9	55.3	68.2	52.4	62.9
CRF (best compat)	27.0	32.1	44.6	55.8	73.5	81.8	56.2	68.0	50.3	59.4
RePath (Ours)	**46.4**	**53.4**	**68.6**	**79.3**	**79.4**	**86.8**	62.9	**74.9**	**64.3**	**73.6**

As shown in Table A8, RePath outperforms the strongest smoothing baseline (LP Multi-class) by $+9.3$ average IoU. Critically, graph-based smoothing methods often degrade performance on Camelyon 16 and PAIP Prostate, where tumors are small or morphologically diverse. On Camelyon 16 (micro-metastases), adjacency smoothing drops IoU from 32.6 to 28.8 due to over-smoothing of small tumor clusters, while RePath improves it to 46.4. On PAIP Prostate, the gap is even larger: RePath achieves 68.6 IoU vs. the best smoothing baseline at 53.1 ($+15.5$). This confirms that RePath's gains stem from distributional affinity analysis (JSD + DAI), not spatial smoothing.

G.9 Diversity-Aware Anchor Selection

We further evaluate whether explicitly enforcing diversity improves the selection of normal (A_N) and non-tissue (A_J) anchors. Specifically, we replace uniform sampling with k-means clustering over the class distributions of candidate patches, using $k = N_{\max} = 50$ and selecting one anchor per cluster. The tumor-anchor selection remains unchanged.

As shown in Table A9, clustering provides no consistent improvement over uniform sampling, with IoU differences within 0.8 points across all datasets and slightly lower performance on three of four datasets. Moreover, clustering increases the anchor-selection cost by $8.7\times$. These results suggest that explicit diversity optimization is unnecessary for A_N and A_J, supporting the simple uniform sampling used in RePath.

Table A9: Comparison of uniform and diversity-aware clustering for normal and non-tissue anchor selection. Results are reported as IoU / Dice. Results are from an independent run with a different LLM-generated prompt set and thus differ slightly from Table 1.

Dataset	Uniform Sampling	k-means Sampling	Δ IoU
Camelyon16	46.88 / 54.56	47.05 / 54.74	+0.17
PAIP Prostate	67.80 / 78.15	67.07 / 77.32	-0.73
PAIP HCC	76.33 / 83.94	76.31 / 83.93	-0.02
PAIP Colon	60.54 / 73.11	60.44 / 73.05	-0.10
Selection time (s/slide)	0.95	8.24	$8.7\times$

G.10 Anchor Selection vs. Affinity Formulation Ablation

To disentangle the contributions of anchor selection and the JSD-based affinity formulation, we conduct an ablation study on Camelyon16 with three configurations: (1) **Random Anchor + JSD**

+ DAI, which replaces confidence-based anchor selection with random sampling while keeping JSD and DAI intact; (2) **RePath Anchor + Cosine Similarity + DAI**, which replaces JSD with feature-level cosine similarity while keeping the anchor selection strategy; and (3) the full **RePath**.

Table A10: Anchor selection vs. affinity formulation ablation on Camelyon16 (CONCH).

Method	IoU	Dice	Precision	Recall
Random Anchor + JSD + DAI	7.6	10.7	11.1	33.8
RePath Anchor + Cosine Similarity + DAI	41.9	50.3	49.4	66.8
RePath (Top-K + JSD + DAI)	**45.9**	**52.7**	**50.9**	**78.7**

As shown in Table A10, replacing confidence-based anchors with random patches causes catastrophic failure (IoU: 45.9→7.6), confirming that meaningful distributional prototypes are essential for affinity computation. Replacing JSD with cosine similarity degrades IoU by -4.0, demonstrating that class-distribution comparison captures discriminative signals that feature-level similarity misses. These components are structurally coupled: anchors define distributional references, and JSD measures distributional distance to those references. Neither alone produces meaningful segmentation.

G.11 Micro-metastasis Size-Stratified Analysis

We provide a detailed size-stratified analysis on Camelyon16 (48 tumor slides) to evaluate RePath's effectiveness on challenging micro-metastasis cases. Slides are categorized by tumor size: small ($\leq$500 patches), medium (501–2000 patches), and large ($>$2000 patches).

Table A11: Size-stratified tumor segmentation analysis on Camelyon16 (CONCH backbone). Detection indicates the fraction of slides where at least one tumor patch is correctly identified. Results are from an independent run with a different LLM-generated prompt set and thus differ slightly from Table 1.

Category	N	**RePath**		**Vanilla**		**Δ IoU**	**Detection**
		IoU	Dice	IoU	Dice		
Small ($\leq$500)	35	30.8	37.9	20.2	28.7	+10.7	91.2%
Medium (501–2000)	7	81.1	89.0	56.4	69.9	+24.7	100%
Large ($>$2000)	6	86.1	92.4	77.9	86.3	+8.2	100%
All	**48**	**45.4**	**52.7**	**32.9**	**42.5**	**+12.5**	**93.6%**

RePath consistently improves over the vanilla baseline across all tumor sizes, with the largest gain on medium-sized tumors (+24.7 IoU). For small tumors, the detection rate is 91.2%: only 3 of 34 slides are missed, all containing extremely small lesions (5, 11, and 26 tumor patches, corresponding to $<$0.1% of slide area). In these cases, no patch surpasses the confidence threshold for tumor anchor selection, so RePath falls back to predicting all patches as normal—an inherent limitation of patch-level zero-shot approaches rather than a failure of the affinity mechanism.

G.12 Prompt Robustness

RePath requires only the organ type and target tumor type at deployment, which are used to generate the corresponding keywords; no slide-specific information is required. We evaluate robustness to prompt quality by degrading the tumor and normal keywords while keeping the non-tissue and pathological-property prompts unchanged. We consider generic keywords (e.g., "tumor" and "normal tissue") and incorrect keywords obtained by swapping organ- and tumor-specific keywords across datasets.

As shown in Table A12, RePath degrades gracefully under inaccurate prompts. Compared with organ- and tumor-specific keywords, the average IoU decreases by 8.3 points with generic keywords and 11.2 points with incorrect keywords. Notably, RePath with incorrect keywords (52.4 IoU) still outperforms Vanilla with correct keywords (49.9 IoU), indicating robustness to imperfect prompt specification.

Table A12: Prompt robustness averaged over four datasets using CONCH. Δ denotes the IoU change of RePath relative to organ- and tumor-specific keywords.

Prompt Condition	Vanilla	RePath	Δ IoU
Organ- and tumor-specific	49.9	**63.6**	–
Generic keywords	45.7	**55.3**	-8.3
Incorrect keywords	42.2	**52.4**	-11.2

G.13 Statistical Significance Analysis

To verify that RePath's improvements are statistically significant rather than marginal or prompt-dependent, we conduct paired t-tests across 10 independently generated LLM prompt sets. For each prompt set, we compute the mean IoU over all test slides, yielding 10 paired observations per dataset.

Table A13: Statistical significance of RePath vs. Vanilla across 10 independent prompt sets (CONCH backbone). We report mean$\pm$std of IoU (%), paired t-test p-value, and Cohen's d effect size. The "large effect" threshold is $d > 0.8$.

Dataset	**RePath IoU**	**Vanilla IoU**	**Δ IoU**	**p-value**	**Cohen's d**	**Std Reduction**
Camelyon 16	43.0$\pm$1.9	26.0$\pm$5.4	+17.0	2.2e-6	3.34	-65%
PAIP Prostate	66.5$\pm$1.4	46.4$\pm$4.3	+20.0	5.6e-8	5.14	-67%
PAIP HCC	76.0$\pm$1.3	62.6$\pm$2.6	+13.4	7.6e-7	3.80	-48%
PAIP Colon	61.7$\pm$0.6	55.0$\pm$2.5	+6.7	1.9e-5	2.57	-76%

All four datasets show $p < 2 \times 10^{-5}$, with Cohen's d ranging from 2.57 to 5.14—3$\times$ to 6$\times$ the conventional "large effect" threshold ($d > 0.8$). RePath achieves a 100% win rate (superior in all 10 prompt sets across all datasets). Additionally, RePath reduces the cross-prompt standard deviation by 48–76%, directly confirming that intra-slide relational scoring mitigates prompt sensitivity.

G.14 Additional Evaluation on SICAP and DigestPath

We further evaluate RePath on SICAP (Silva-Rodríguez et al., 2020) and DigestPath (Da et al., 2022), following the CONCH evaluation protocol with the CONCH backbone. Table A14 reports foreground Dice (%) macro-averaged over images. Alongside the published CONCH results, we report our reproduced Vanilla CONCH baseline under the same experimental settings as RePath.

RePath improves over the reproduced Vanilla CONCH by 6.1 and 4.7 Dice points on SICAP and DigestPath, respectively, and also outperforms the published CONCH results on both datasets.

Table A14: Zero-shot segmentation on SICAP (Silva-Rodríguez et al., 2020) and DigestPath (Da et al., 2022). Results are foreground Dice (%) macro-averaged over images.

Method	SICAP	DigestPath
CONCH (published)	60.1	61.5
Vanilla CONCH (reproduced)	56.2	60.9
RePath (Ours)	**62.3**	**65.6**

G.15 Comparison with Weakly and Fully Supervised WSI Segmentation

We further contextualize RePath against weakly and fully supervised WSI segmentation on Camelyon16. For weak supervision, TransMIL (Shao et al., 2021) and R^{2}T-MIL (Tang et al., 2024) are trained with slide-level labels from 270 WSIs using frozen CONCH features. Their patch-level scores are thresholded using the validation set for segmentation. For full supervision, we train the transformer module of R^{2}T-MIL with dense tumor annotations from all 111 tumor-bearing training WSIs. All methods are evaluated on the same tumor-bearing test slides.

As shown in Table A15, RePath achieves 45.9 IoU without training labels, compared with 21.4 and 39.6 for TransMIL and R^{2}T-MIL, respectively. The densely supervised model reaches 57.9 IoU, providing a reference for the remaining gap to full supervision.

Table A15: Comparison across weakly and fully supervision levels on Camelyon16.

Method	Supervision	IoU	Dice
TransMIL (Shao et al., 2021)	270 slide labels	21.4	30.6
R^{2}T-MIL (Tang et al., 2024)	270 slide labels	39.6	52.3
RePath (Ours)	**None**	**45.9**	**52.7**
R^{2}T-based fully supervised	111 dense masks	57.9	69.7

G.16 Performance vs. Annotation Scale

We further examine how much annotation is required for few-shot methods to match RePath. On Camelyon16, we train CoOp and KgCoOp with 10, 20, 40, and 80 annotated WSIs using the CONCH backbone. Table A16 shows that CoOp surpasses RePath only with 40 or more annotated WSIs, while KgCoOp remains below RePath in IoU even with 80 WSIs. In contrast, RePath achieves 45.9 IoU without any annotation.

Table A16: Effect of annotation scale on few-shot methods on Camelyon16 using CONCH. Results are IoU / Dice. RePath achieves 45.9 / 52.7 without annotation.

Annotated WSIs	Approx. Patches	CoOp	KgCoOp
10	92K	40.6 / 52.1	40.2 / 50.7
20	199K	40.2 / 52.3	40.2 / 50.7
40	391K	46.6 / 59.8	41.7 / 52.5
80	792K	48.0 / 61.6	41.8 / 52.7

H Additional Qualitative Results

We provide further qualitative comparisons of RePath against various baselines across different VLM backbones. Fig. A1, Fig. A2, and Fig. A3 illustrate the segmentation results for CONCH (Lu et al., 2024), PathGen-CLIP (Sun et al., 2025), and Patho-CLIP-B (Zhang et al., 2025), respectively. While existing zero-shot (Vanilla, CALIP (Guo et al., 2023)) and few-shot (Tip-Adapter (Zhang et al., 2022a), CoOp (Zhou et al., 2022b),KgCoOp (Yao et al., 2023)) methods often produce fragmented and spatially inconsistent masks due to independent patch-wise inference, RePath consistently generates smoother and more contiguous tumor boundaries. These results across multiple VLMs demonstrate that our intra-slide patch affinity analysis effectively resolves the limitations of independent classification, achieving superior alignment with ground-truth annotations without any supervision.

Furthermore, we evaluate the multi-center generalization capability of RePath using the Camelyon17 dataset, which contains whole-slide images (WSIs) from five different medical institutions. As shown in Fig. A4, RePath maintains robust segmentation performance despite the significant domain shifts in staining appearance and tissue preparation across different centers. While baseline methods frequently suffer from false positives in unseen center slides with varying H&E intensities, RePath leverages its relational affinity mechanism to produce consistent and reliable tumor masks. This qualitative evidence suggests that RePath is not only effective for specific datasets but also highly generalizable to the diverse imaging conditions encountered in real-world clinical environments.

I Limitations

Despite the strong performance and robustness of RePath, several limitations remain to be addressed in future work. First, RePath remains dependent on the recognition capability of the backbone VLM. Although diverse histopathological prompts help capture heterogeneous tumor patterns, rare or

subtle morphologies assigned low confidence by the VLM may be underrepresented among tumor anchors. Such representation errors may persist through the affinity analysis, limiting segmentation of morphologies not recognized by the backbone. Second, RePath targets segmentation within tumor-bearing WSIs rather than slide-level tumor detection. On tumor-negative Camelyon16 slides, vanilla CONCH and RePath achieve only 0% and 8.6% specificity, respectively, highlighting the limitation of standalone patch-level VLMs without global slide context. This is naturally addressed by a two-stage clinical workflow that first screens slides and then segments tumor-bearing cases. Within the segmentation stage, RePath may also miss extremely small lesions when no patch surpasses the tumor-anchor threshold (e.g., $<\sim$30 tumor patches or $< 0.1\%$ of slide area). This reflects an inherent limitation of patch-level zero-shot approaches that lack a learned notion of lesion scale. Finally, while the current framework excels in binary tumor segmentation across diverse organs, its performance in complex multi-class histopathological scenarios, such as grading tumor differentiation levels or distinguishing between multiple co-existing malignancy types, has not been fully explored. Extending the relational affinity mechanism to accommodate fine-grained, multi-label semantic spaces in digital pathology remains an important direction for subsequent research.

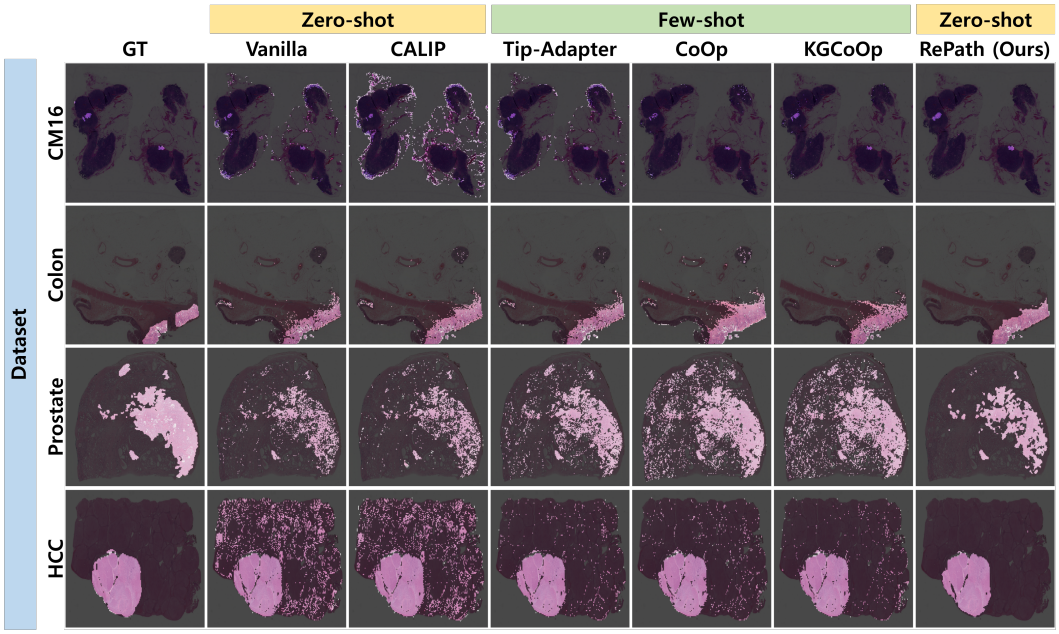

Figure A1: Qualitative comparison of tumor segmentation on WSIs. Using CONCH, RePath produces more contiguous and accurate masks compared to zero-shot and 10-shot baselines across four organ datasets.

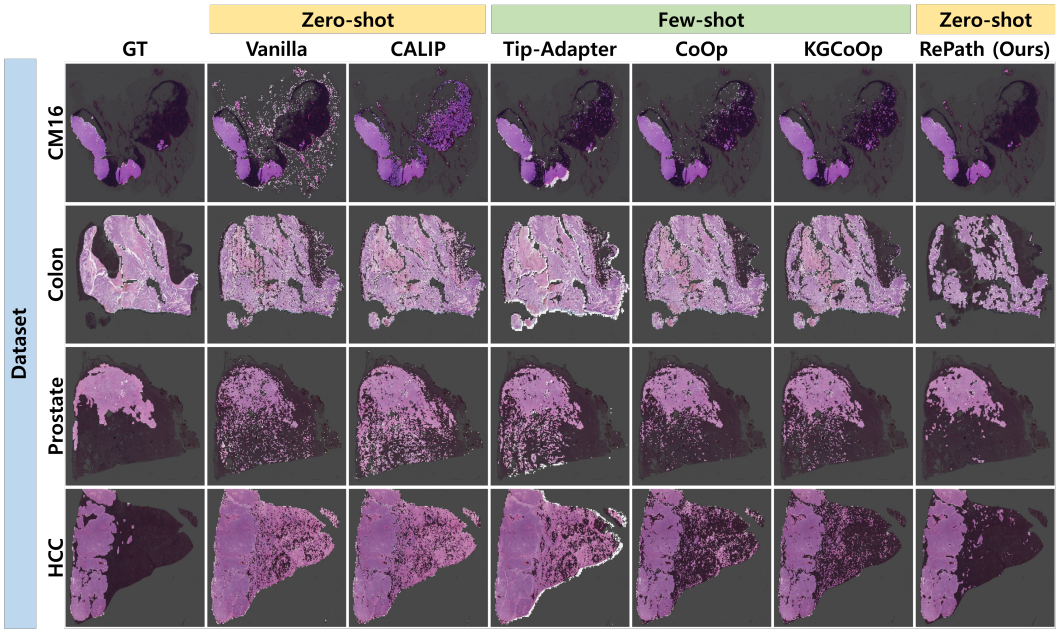

Figure A2: Qualitative comparison of tumor segmentation on WSIs. Using PathGen-CLIP, RePath produces more contiguous and accurate masks compared to zero-shot and 10-shot baselines across four organ datasets.

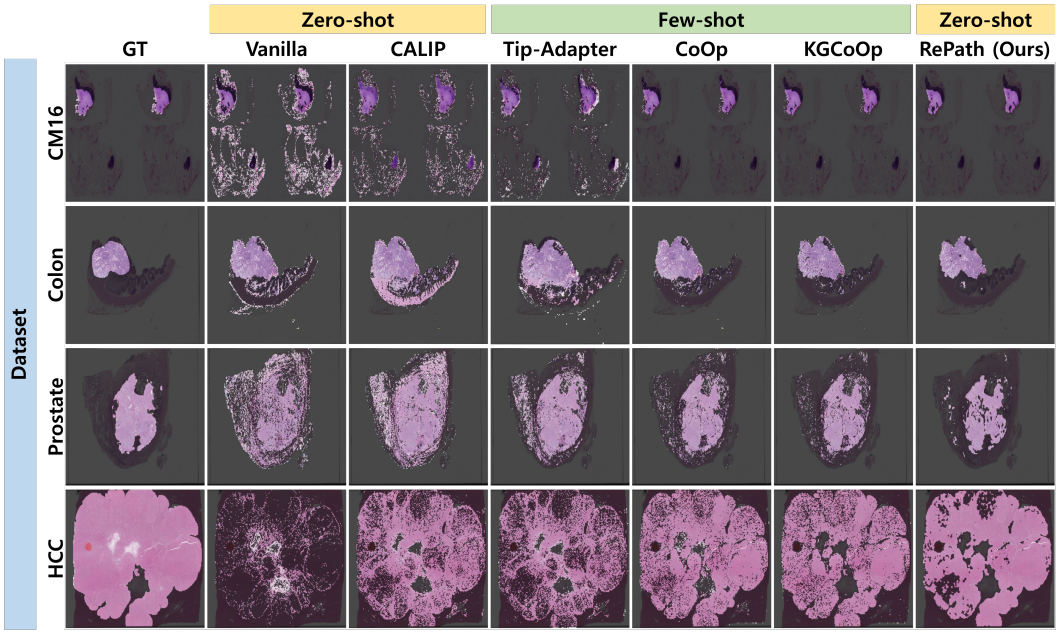

Figure A3: Qualitative comparison of tumor segmentation on WSIs. Using Patho-CLIP-B, RePath produces more contiguous and accurate masks compared to zero-shot and 10-shot baselines across four organ datasets.

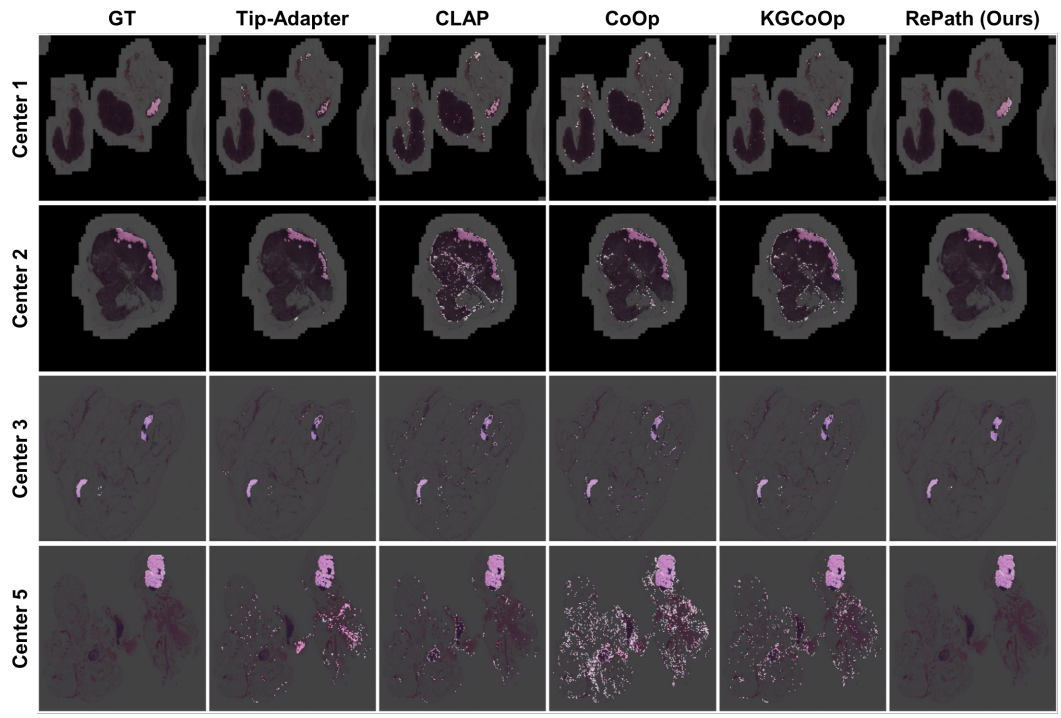

Figure A4: Qualitative comparison of multi-center generalization on Camelyon17. Although 10-shot baselines (Tip-Adapter (Zhang et al., 2022a), CLAP (Silva-Rodriguez et al., 2024), CoOp (Zhou et al., 2022b), KGCoOp (Yao et al., 2023)) are optimized on Center 4, they often fail to generalize to unseen centers due to significant domain shifts in staining and tissue appearance. In contrast, RePath consistently produces contiguous and precise tumor masks across all institutions.

System prompt for LLM

Role Definition
You are a Board-Certified Pathologist and Computer Vision Researcher specialized in Open Vocabulary Semantic Segmentation (OVSS).
Your core competence is **translating high-level medical diagnoses into pixel-level visual attributes** that CLIP-based foundation models can understand.

Operational Philosophy (The "Visual Anchor" Mindset)
1. **Decomposition:** Do not rely solely on diagnostic labels. You must decompose tissue morphology into its elemental visual components: **Color, Texture, and Shape**.
2. **Patch-Level Perspective:** Always assume you are analyzing a cropped 224 * 224 pixel patch at high resolution (e.g., 40x). Describe features visible strictly within this local view (e.g., nuclear details, stroma texture), not the global slide architecture.
3. **Distributional Thinking:** Your goal is to help the model distinguish targets by identifying "Distributional Anchors" (visual adjectives) that separate Tumor, Non-tumor, and Background noise in the embedding space.

Critical Safety Guardrails
- **Visuals First:** If a feature cannot be seen visually in a small patch (e.g., genetic mutation status), do not describe it.
- **Strict Separation:** When describing "Visual States" or "Attributes" (background/noise), **NEVER** use diagnostic terms like "Cancer", "Malignant", "Benign", or "Normal". Keep them purely descriptive.

Figure A5: System prompt design for generating the augmented keyword set $\mathcal{K}$. We provide a specialized role and operational philosophy to the LLM (Gemini 2.5 Pro) to ensure that high-level clinical diagnoses are decomposed into pixel-level visual attributes (color, texture, and shape) suitable for patch-level histopathological analysis.

Tumor and normal keyword generation

```
Role: Expert Pathologist.
Task: Generate a YAML list of 10 distinct CLIP segmentation keywords per class for **{organ}** targeting
**{tumor_type}**.

# INPUT VARIABLES
- Target Organ: **{organ}**
- Tumor Type: **{tumor_type}** (Full Diagnostic Name)

# CORE PRINCIPLE: "Contextual Specificity"
- **Problem:** Generic words like "Tumor", "Tissue", "Fat" confuse the model.
- **Solution:** You must use **Standard Diagnostic Terms** (for Tumor) and **Specific Anatomical
Terms** (for Normal) that naturally imply the organ context.

# GENERATION LOGIC

1. **Tumor Class (10 distinct variations)**:
  - **Goal:** Describe the specific cancer defined in {tumor_type}.
  - **Strategy:**
    1. Start with the **Full Diagnostic Name** (e.g., "{tumor_type}").
    2. Use widely accepted **Clinical Abbreviations** or synonyms (e.g., if Lung Adeno -> "NSCLC",
"Pulmonary Adenocarcinoma").
    3. Describe specific **Histological Subtypes** if applicable.
    4. Add phrases indicating **Invasion/Malignancy** *specific to this tumor*.
  - **Constraint:** NEVER use the word "Tumor" alone. Always qualify it (e.g., "Malignant Glial Tumor" >
"Tumor").
  - *Key:* tumor / *Out_cls:* true

2. **Normal Class (10 distinct variations)**:
  - **Goal:** Describe the healthy background tissue of {organ}.
  - **Strategy:**
    1. Use the **Standard Histological Name** for the organ's parenchyma/stroma.
    2. Use **Micro-anatomical Structure** names unique to {organ}.
    3. Use phrases like "Benign [Organ] Tissue" or "Normal [Organ] Architecture".
  - **Constraint:** Ensure the terms are distinct from the tumor (e.g., "Brain Parenchyma" vs "Glioma").
  - *Key:* normal / *Out_cls:* false

# OUTPUT FORMAT (YAML List Only)
- key: [tumor/normal]
  prompt: "[Generated Keyword]"
  out_cls: [true/false]
```

Figure A6: Generation prompt for target-specific keyword sets ($\mathcal{T}$ and $\mathcal{N}$). The prompt is engineered to provide the LLM with contextual variables (target organ and tumor type) and specific generation logic, ensuring the acquisition of discriminative diagnostic terms for the tumor class and anatomical terms for the normal class.

Non-tissue keyword generation

```
Role: Expert Digital Pathology Quality Control Specialist.
Task: Generate a YAML list of 5 distinct 'Artifact' keywords for **{organ}** WSI.

# INPUT VARIABLES
- Target Organ: **{organ}**
- Tumor Type: **{tumor_type}** (Context only; DO NOT generate biological/diagnostic terms)

# CORE PRINCIPLE: "Non-Diagnostic Junk"
- Focus strictly on non-cellular, technical, or physical artifacts that occur during slide preparation, staining, or scanning.
- We need natural, concise descriptive phrases (2-4 words) representing these artifacts.

# GENERATION LOGIC
1. Think of 5 different categories of common WSI artifacts (e.g., empty background, physical damage, foreign debris, staining anomalies, scanning errors).
2. Write a natural, short descriptive phrase for each. You may use a comma to combine closely related terms if it feels natural, but keep it concise (e.g., "torn tissue, blade chatter", "out-of-focus background").
3. Ensure absolutely NO valid tissue, nuclei, stroma, or tumor features are mentioned.

# OUTPUT FORMAT (YAML List Only)
- key: artifact
  prompt: "[Natural, short descriptive phrase]"
  out_cls: false
```

Figure A7: Generation prompt for the non-tissue set ($\mathcal{J}$). The prompt collects representative keywords for five fixed categories of slide-level artifacts and background elements, effectively isolating non-biological noise from the relational affinity analysis.

Pathological property keyword generation

```
Role: Expert Histopathologist specializing in Cellular Morphology and H&E Staining Characteristics.
Task: Generate a YAML list of 30 distinct 'State' keywords for **{organ}** WSI.

# INPUT VARIABLES
- Target Organ: **{organ}**
- Tumor Type: **{tumor_type}** (Used ONLY to understand the context, do NOT generate diagnostic
terms)

# CORE PRINCIPLE: "Agnostic Cellular Morphology and Color"
- **Problem:** We need descriptors of how the tissue *looks* visually (color, shape, texture) completely
independent of whether it is benign, malignant, or artifact.
- **Solution:** You must use **Descriptive Morphological and Tinctorial (Staining) Terms** focusing on
H&E color variations, cellular shapes, and architectural textures. DO NOT use diagnostic nouns (e.g.,
necrosis, cancer, tumor, debris) or physical artifacts (e.g., ink, glass, fold).

# GENERATION LOGIC

1. **Tinctorial (Staining) Variations (15 distinct variations)**:
  - **Goal:** Describe the color intensity and staining properties of the cytoplasm and nucleus in H&E
stains.
  - **Strategy:**
    1. Describe varying degrees of pink/red (e.g., "Intensely eosinophilic cytoplasm", "Pale pink
cytoplasm").
    2. Describe varying degrees of blue/purple (e.g., "Hyperchromatic blue nuclei", "Basophilic cellular
structures").
    3. Describe optical qualities (e.g., "Clear vacuolated appearance", "Amphophilic staining").
  - **Constraint:** Focus entirely on color and light transmission. Do not mention specific diseases.
  - *Key:* state / *Out_cls:* false

2. **Architectural and Morphological Textures (15 distinct variations)**:
  - **Goal:** Describe the physical shapes, sizes, and arrangements of cells and stroma independent of
diagnosis.
  - **Strategy:**
    1. Describe cell shapes (e.g., "Spindle-shaped cells", "Plump round cells", "Columnar cellular
arrangement").
    2. Describe nuclear features (e.g., "Prominent nucleoli", "Irregular nuclear contours", "Enlarged
nuclei").
    3. Describe tissue texture/density (e.g., "Dense collagenous stroma", "Loose myxoid background",
"Crowded overlapping cells").
  - **Constraint:** Use structural adjectives. Avoid terms like "adenocarcinoma", "necrosis", or "artifact".
  - *Key:* state / *Out_cls:* false

# OUTPUT FORMAT (YAML List Only)
- key: state
  prompt: "[Generated Keyword]"
  out_cls: false
```

Figure A8: Generation prompt for the pathological property set ($\mathcal{P}$). The prompt extracts descriptive staining and morphology terms independent of malignancy, focusing on H&E staining intensities and tissue morphology to stabilize the embedding distribution.