ReCAP: Report-Conditioned Attentive Patching for Gigapixel Histopathology

Tawsifur Rahman[1] ARAHMA34@JHU.EDU
[1] *Biomedical Engineering, Johns Hopkins University*

Rama Chellappa[2] RCHELLA4@JHU.EDU
[2] *Johns Hopkins University*

Alexander S. Baras[3] BARAS@JHMI.EDU
[3] *School of Medicine, Johns Hopkins University*

Editors: Under Review for MIDL 2026

Abstract

Vision–language learning has become a powerful framework for multimodal representation, achieving exceptional performance across diverse image–text tasks. However, in histopathology, existing methods often rely on high-resolution region-level annotations to achieve fine-grained visual–textual alignment—an assumption that is impractical for Whole Slide Image (WSI) classification due to the gigapixel scale of pathology images and the weak supervision provided by slide-level labels. To address this challenge, we propose Report-Conditioned Attentive Patching for Weakly Supervised WSI Classification (ReCAP), a novel approach that leverages slide-level pathology reports to enrich patch-level feature learning without requiring localized supervision. Instead of relying on explicit region annotations, ReCAP adopts a hybrid multimodal contrastive MIL framework in which report-conditioned text embeddings guide cross-attention to highlight semantically discriminative tissue regions. We further introduce a self-normalizing cross-modal attended similarity function that enhances the robustness and stability of patch–text alignment under weak supervision. In addition, our approach incorporates an efficient report-aware patch aggregation strategy that suppresses redundant or noisy regions while retaining the most diagnostically informative patterns within the vision–language context. Across multiple cancer subtype classification and survival prediction tasks, ReCAP consistently improves performance by 2–5%, demonstrating the effectiveness of report-conditioned cross-modal alignment for scalable and annotation-efficient WSI understanding.

Keywords: Whole Slide Image, Vision–Language Learning, Pathology Reports, Weakly Supervised Learning, Cross-Modal Alignment, Contrastive Learning.

1. Introduction

In histopathology, Whole Slide Image (WSI) analysis requires more than isolated patch examination—it demands a holistic and context-aware approach. Pathologists navigate WSIs dynamically, integrating morphological cues across multiple regions to form a diagnosis (1; 2; 3). This reasoning process unfolds like a multi-turn dialogue, where different tissue regions contribute varying degrees of diagnostic evidence (4; 5; 6). Importantly, pathology reports play a crucial role in this workflow: they summarize salient morphological findings, highlight suspicious patterns, and provide semantic cues that complement visual inspection. These textual insights often capture diagnostic signals—such as tumor descriptors,

cellular abnormalities, or clinically relevant context—that may not be uniformly present across sampled WSI patches. Incorporating such report-derived information into computational models therefore offers an avenue to enhance cancer subtype prediction by helping the model attend to patches that are more likely to carry diagnostic significance.

However, computational WSI analysis struggles to replicate this nuanced diagnostic behavior. Classical approaches treat WSI patches independently, overlooking their spatial relationships and failing to utilize the additional semantic information encoded in corresponding pathology reports (7; 8; 9). Weakly supervised learning settings further complicate this task, as slide-level labels provide no explicit supervision for identifying the most informative regions (10; 11; 12). As a result, models often aggregate thousands of patches indiscriminately, incorporating noise and diluting the discriminative power of key morphological patterns.

Recent advances in vision–language models (VLMs) have transformed multimodal representation learning, excelling in tasks such as image captioning, visual question answering, and cross-modal retrieval (13; 14). Yet, applying these methods directly to histopathology is non-trivial. Existing VLMs frequently assume the availability of region-level annotations or explicit image–text alignment, which is impractical for gigapixel WSIs annotated only at the slide level (15; 16; 17). Similarly, conventional multiple instance learning (MIL) frameworks lack mechanisms for selectively attending to patches that gain relevance only when contextualized with complementary textual information (18; 19; 20; 21). Domain-specific models such as CONCH and CPLIP partially mitigate domain shifts by leveraging large biomedical image–text corpora (13; 16), but they still aggregate patches uniformly and do not adaptively integrate report information to guide patch selection. This often leads to redundant or noisy patch inclusion and computational inefficiencies (22; 23; 24; 25).

To address these limitations, we propose **Report-Conditioned Attentive Patching (ReCAP)**, a weakly supervised multimodal framework that leverages pathology reports to provide additional contextual information that improves patch-level attention and aggregation. Rather than assuming direct correspondence between individual patches and textual descriptions, ReCAP uses report information to modulate patch relevance, enabling the model to focus on tissue regions that better reflect the semantic content of the report. The main contributions of ReCAP include:

(1) **Report-Guided Patch Prioritization:** ReCAP leverages pathology reports as complementary semantic information to modulate patch-level importance under weak supervision, improving downstream performance by emphasizing clinically meaningful regions.

(2) **Hybrid Contrastive MIL With Cross-Attention:** Our framework integrates cross-attention-based contrastive learning to refine patch embeddings using report-conditioned context, enhancing semantic consistency between modalities without requiring region-level supervision.

(3) **Efficient Report-Aware Patch Aggregation:** By suppressing redundant or non-informative regions, ReCAP achieves computational efficiency while preserving diagnostically relevant visual patterns.

(4) **Improved Performance Across Cancer Subtypes and Survival Tasks:** ReCAP demonstrates consistent gains of 2–5% across multiple cancers, highlighting the value of using pathology reports to enhance multimodal WSI understanding.

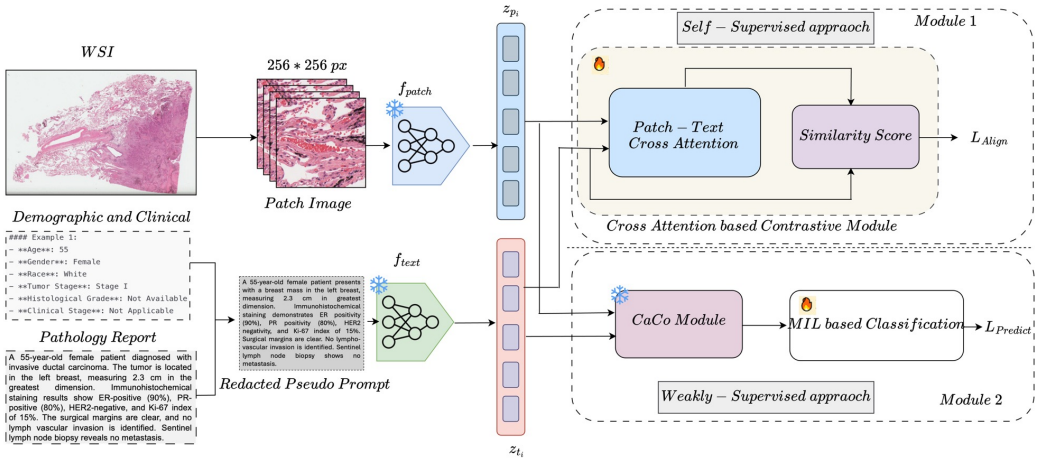

Figure 1: The overall architecture of our ReCAP framework (details in Section 2). ReCAP integrates pathology report information into WSI modeling through a two-stage design. Stage 1 uses report-conditioned cross-attention to refine patch representations, while Stage 2 aggregates the attended patches via a MIL-based predictor for classification and survival analysis.

2. Method

Our proposed **ReCAP** framework operates in two stages to incorporate pathology report information into weakly supervised WSI analysis. In the first stage, we use a multimodal contrastive learning module in which report-conditioned text embeddings guide cross-attention to provide additional contextual cues for identifying relevant patches. This enables the model to refine patch representations without requiring region-level annotations. In the second stage, the attended patch features are aggregated using a MIL-based predictor for downstream tasks, including classification and survival analysis formulated as an ordinal regression problem. The overall architecture of ReCAP is shown in Figure 1.

2.1. WSI and Report Preprocessing with Patch/Text Embedding Extraction

We first preprocess each Whole Slide Image (WSI) to extract informative tissue patches. Each WSI is loaded at a downsampled resolution (e.g., 20×) and converted from RGB to HSV color space. Tissue regions are segmented by thresholding the saturation channel after median blurring to suppress noise. The resulting binary tissue mask is refined using morphological closing to fill small gaps, followed by contour extraction and area-based filtering to retain only meaningful tissue components. The final tissue mask can be optionally saved for visual inspection along with a preprocessing log for parameter tuning.

Using the tissue contours, we extract fixed-size non-overlapping patches of size 256×256 at the desired magnification. Each patch is stored with its spatial coordinates and slide-level metadata in an HDF5 file for efficient downstream processing. The number of extracted

patches varies across slides, ranging from a few hundred for small biopsy samples to several hundred thousand for large resection specimens.

To incorporate clinical text while preventing report-to-label leakage, we carefully construct the pathology report input used by ReCAP. For both tasks, the pseudo-prompt is formed only from information available *prior to final diagnostic sign-out*, including demographics (age/sex) and morphology-bearing descriptions from the *Microscopic Description* (See in Supplementary Table 7). To explicitly avoid trivial leakage, we apply strict diagnosis-word redaction before tokenization by removing subtype-identifying keywords and abbreviations (e.g., LUAD/LUSC, adenocarcinoma/squamous, IDC/ILC, CCRCC/PRCC/CRCC, KIRC/KIRP/KICH). This ensures the model cannot directly read the ground-truth label from the report text, and instead uses report content as weak clinical context for guiding patch prioritization.

After preprocessing, each slide yields N_p patches $\{p_i\}_{i=1}^{N_p}$. Each patch is encoded into a latent feature vector using a frozen PLIP image encoder:

$$z_{p_i} = f_{\text{patch}}(p_i; \theta_p) \in \mathbb{R}^{d_p}, \tag{1}$$

where f_{patch} denotes the patch encoder and θ_p are fixed parameters. Similarly, the redacted pseudo-prompt is tokenized into N_t tokens $\{t_j\}_{j=1}^{N_t}$, and each token is encoded using the frozen PLIP text encoder:

$$z_{t_j} = f_{\text{text}}(t_j; \theta_t) \in \mathbb{R}^{d_t}. \tag{2}$$

These patch and text embeddings provide complementary visual and semantic representations, which are subsequently used by ReCAP for report-conditioned cross-attention and similarity-based patch selection.

2.2. Report-Conditioned Contrastive Learning with Cross-Attention

In ReCAP, report-derived context is used to refine patch representations via a cross-attention–guided contrastive module. Text embeddings provide weak clinical cues that help highlight diagnostically relevant patches under slide-level supervision. As shown in Fig. 2, report tokens act as queries while patch embeddings serve as keys and values. The attention weight between text token t_j and patch p_i is defined as

$$A_{ij} = \text{softmax}\left(\frac{q_{t_j} K_{p_i}^T}{\sqrt{d}}\right), \tag{3}$$

where q_{t_j} and K_{p_i} denote the query and key vectors, respectively, and d is the embedding dimension.

Using these attention weights, each patch is refined by aggregating information from all report tokens:

$$\tilde{z}_{p_i} = \sum_j A_{ij} V_{p_i}, \tag{4}$$

where V_{p_i} denotes the patch value embedding. This produces report-attended patch representations $\tilde{z}_{p_i}$ that encode both visual morphology and weak clinical context.

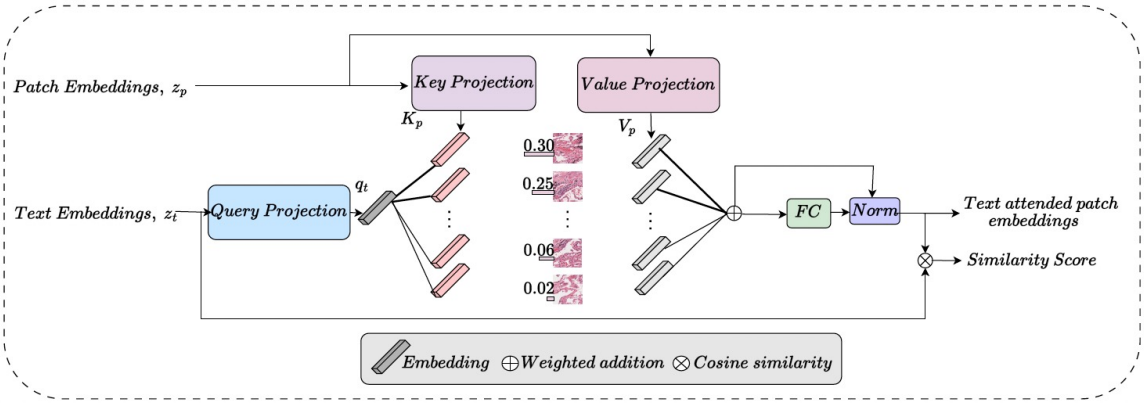

Figure 2: ReCAP cross-attention contrastive module.

2.3. Softmax-Normalized Cross-Modal Similarity Weighting and MIL Aggregation

To quantify the relevance of each attended patch to the report context, we compute cosine similarity between each report-attended patch embedding $\tilde{z}_{p_i}$ and report token embedding z_{t_j}:

$$S_{ij} = \frac{\tilde{z}_{p_i} \cdot z_{t_j}}{\|\tilde{z}_{p_i}\| \, \|z_{t_j}\|}. \tag{5}$$

We then aggregate similarities across report tokens to obtain a single patch relevance score

$$S_i = \frac{1}{T} \sum_{j=1}^{T} S_{ij}, \tag{6}$$

where T is the number of report tokens.

Based on these scores, ReCAP retains only the Top-k% most relevant patches:

$$\mathcal{P}_{\text{top-k}} = \{\tilde{z}_{p_i} \mid i \in \text{Top-k}(S_i)\}. \tag{7}$$

For the selected patches, we compute instance weights via softmax normalization,

$$\alpha_i = \frac{\exp(S_i)}{\sum_{m \in \mathcal{P}_{\text{top-k}}} \exp(S_m)}, \tag{8}$$

which we refer to as *softmax-normalized cross-modal similarity weighting*. This normalization yields scale-invariant instance weights and stabilizes tile ranking under weak supervision.

Finally, the slide-level representation is obtained by weighted aggregation:

$$\hat{z} = \sum_{i \in \mathcal{P}_{\text{top-k}}} \alpha_i \, g(\tilde{z}_{p_i}), \tag{9}$$

where $g(\cdot)$ denotes a projection head. In contrast to standard MIL attention, this mechanism explicitly converts cross-modal similarity into instance selection and weighting, enabling report-guided filtering of non-informative tiles while emphasizing patches most aligned with clinical context.

2.4. Loss Function

ReCAP is trained with a two-stage objective. The first stage uses a contrastive loss to align attended patches with relevant textual cues:

$$\mathcal{L}_{\text{Align}} = -\sum_{(i,j)\in P} \log \frac{\exp(S_{ij}/\tau)}{\sum_k \exp(S_{ik}/\tau)}. \tag{10}$$

The second stage optimizes downstream classification or survival prediction using cross-entropy:

$$\mathcal{L}_{\text{Predict}} = -\sum_c \left[y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c) \right]. \tag{11}$$

The final objective combines both terms:

$$\mathcal{L}_{\text{Total}} = \lambda \mathcal{L}_{\text{Align}} + (1 - \lambda)\mathcal{L}_{\text{Predict}}, \tag{12}$$

where based on empirical evaluation, we found that setting $\lambda = 0.4$ yielded optimal performance.

Table 1: Performance of ReCAP and state-of-the-art methods on cancer subtype classification and survival prediction tasks.

Method	Image	Text	**Cancer Subtype Classification**						**Survival Prediction**							
			TCGA-RCC		TCGA-NSCLC		TCGA-BRCA		BRCA-IDC		BRCA-ILC		NSCLC-LUAD		NSCLC-LUSC	
			Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
ABMIL (17)	✓	-	90.18	92.86	91.63	93.41	87.21	89.67	83.02	85.33	84.12	86.01	72.45	74.28	71.38	76.22
MI-Zero (29)	✓	-	72.56	73.21	68.76	70.11	73.34	74.45	66.89	68.32	75.08	76.23	61.45	63.56	66.08	67.28
CLAM-SB (19)	✓	-	92.06	94.11	93.83	94.95	90.72	92.89	84.22	86.78	85.67	87.45	75.34	77.11	73.22	78.45
DS-MIL (26)	✓	-	91.41	93.13	92.73	94.20	89.41	91.82	83.71	86.34	85.23	87.01	74.11	76.45	70.68	77.23
TransMIL (27)	✓	-	91.92	94.06	93.01	94.23	89.93	92.61	84.45	86.92	85.78	87.56	74.78	77.56	72.34	78.11
CONCH (13)	✓	✓	94.29	96.31	95.36	96.73	93.11	94.88	86.12	88.67	87.23	89.12	78.45	80.11	75.11	80.78
PLIP (14)	✓	✓	93.56	95.80	94.71	96.64	92.34	93.78	85.78	88.01	86.67	88.91	77.56	79.34	74.78	79.89
ReCAP (Ours)	✓	✓	**96.12**	**97.84**	**97.41**	**98.66**	**94.89**	**96.91**	**88.67**	**90.98**	**90.72**	**92.34**	**81.88**	**83.92**	**80.77**	**84.11**

3. Results

3.1. Dataset

We evaluated our methodology on two downstream tasks cancer subtype classification and cancer survival prediction.

WSI Cancer Subtype Classification Our comprehensive approach was rigorously evaluated on three cancer subtype classification tasks using the TCGA-RCC (28), TCGA-NSCLC (28), and TCGA-BRCA (28) datasets. To avoid data leakage, we ensured that slides from the same patient case were grouped together during partitioning.

TCGA-RCC WSI. The TCGA-RCC WSI dataset includes 884 diagnostic Whole Slide Images (WSIs) from the TCGA RCC repository, covering Kidney Chromophobe (TCGA-KICH), Kidney Clear Cell Renal Cell Carcinoma (TCGA-KIRC), and Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP) projects. Specifically, there are 111 slides from 99 cases of Chromophobe Renal Cell Carcinoma (CRCC), 489 slides from 483 cases of Clear Cell Renal Cell Carcinoma (CCRCC), and 284 slides from 264 cases of Papillary Renal Cell

Carcinoma (PRCC). On average, 13,907 patches were extracted per slide at ×20 magnification.

TCGA-NSCLC WSI. The TCGA-NSCLC WSI dataset contains 1020 Formalin-Fixed Paraffin-Embedded (FFPE) WSIs, with 490 associated with lung adenocarcinoma (LUAD) and 530 with lung squamous cell carcinoma (LUSC).

TCGA-BRCA WSI. The TCGA-BRCA WSI dataset consists of 954 FFPE WSIs, including 787 slides with Invasive Ductal Carcinoma (IDC) and 167 slides with Invasive Lobular Carcinoma (ILC).

WSI Cancer Survival Prediction We investigated our model on four cancer subtype survival prediction tasks, considering survival beyond 1, 2, and 5 years. The datasets used include BRCA-IDC (28), BRCA-ILC (28), NSCLC-LUAD (28), and NSCLC-LUSC (28).

NSCLC-LUAD. This dataset comprises 490 whole-slide images (WSIs) of lung adenocarcinoma from Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples. The WSIs are distributed as follows: 248 for survival beyond 1 year, 123 for 2 years, and 34 for 5 years.

NSCLC-LUSC. This dataset includes 530 WSIs of lung squamous cell carcinoma from FFPE samples, with survival-based distribution as follows: 258 WSIs for over 1 year, 171 for 2 years, and 75 for 5 years.

BRCA-IDC. The TCGA-BRCA-IDC dataset consists of 787 WSIs of invasive ductal carcinoma (IDC) from FFPE samples. The survival distribution includes 563 WSIs for more than 1 year, 346 for 2 years, and 136 for 5 years.

BRCA-ILC. The TCGA-BRCA-ILC dataset contains 167 WSIs of invasive lobular carcinoma (ILC) obtained from FFPE samples. The WSIs are distributed as follows: 128 for survival beyond 1 year, 69 for 2 years, and 30 for 5 years.

3.2. Experiment settings and Evaluation metrics

We implemented our approach using PyTorch and conducted all experiments on a server with 4 NVIDIA Tesla V100 GPUs and 32 CPU cores. All models were optimized using the Adam optimizer. For WSI survival prediction tasks, models were trained for up to 1500 steps with a batch size of 256 and an initial learning rate of 0.001. Hyperparameters were fine-tuned for each dataset, and the best model was selected based on validation performance within a 5-fold cross-validation setup. Evaluation metrics included accuracy and area under the curve (AUC), with accuracy in multi-label classification computed as the average across all target classes.

Table 2: Comparison of patch aggregation strategies in ReCAP across cancer subtype classification and survival prediction tasks. We additionally include an *All Tiles* baseline (no filtering) to quantify the benefit of report-guided Top-k selection beyond tile reduction.

Aggregation Strategy	**Cancer Subtype Classification**			**Survival Prediction**			
	TCGA-RCC	TCGA-NSCLC	TCGA-BRCA	BRCA-IDC	BRCA-ILC	NSCLC-LUAD	NSCLC-LUSC
All Tiles (no filtering)	93.45	94.82	90.11	84.12	85.36	77.29	76.18
Random-k	89.12	91.34	84.98	79.56	80.45	72.12	70.34
First-k (sequence-based)	90.45	93.11	85.93	80.89	81.78	73.34	71.56
Last-k (sequence-based)	92.34	94.01	87.61	82.45	83.12	75.67	74.23
Top-k (Ours)	**96.12**	**97.41**	**94.89**	**88.67**	**90.72**	**81.88**	**80.77**

3.3. Performance Analysis

This section evaluates the performance of the proposed **ReCAP** framework against state-of-the-art baselines on two tasks: cancer subtype classification and survival prediction. Results across seven TCGA cohorts are summarized in Table 1. Overall, ReCAP demonstrates strong and consistent improvements across all datasets, highlighting the benefits of integrating pathology report information into weakly supervised WSI analysis.

Cancer Subtype Classification. ReCAP achieves the highest accuracy across all three subtype classification tasks. It obtains 96.12% on TCGA-RCC, 97.41% on TCGA-NSCLC, and 94.89% on TCGA-BRCA—consistent improvements of 1–2% over the strongest multi-modal baselines (CONCH, PLIP) and larger gains over image-only approaches such as ABMIL, CLAM-SB, DS-MIL, and TransMIL. These improvements highlight the effectiveness of incorporating report-conditioned contextual cues during patch refinement. By introducing text-guided similarity signals early in the pipeline, ReCAP is better able to isolate subtype-relevant regions, particularly in morphologically heterogeneous datasets such as BRCA. This demonstrates that pathology reports provide complementary diagnostic information that strengthens patch selection and reduces noise in large, variable tissue regions.

Survival Prediction. Across the four survival cohorts, ReCAP again outperforms all competing models. It achieves 88.67% on BRCA-IDC, 90.72% on BRCA-ILC, 81.88% on NSCLC-LUAD, and 80.77% on NSCLC-LUSC. These values reflect consistent 1–3% gains over multi-modal baselines and larger improvements over single-modality MIL methods. Performance differences are most pronounced in the NSCLC cohorts, where relying solely on image features is insufficient to capture prognostic complexity. Incorporating report-derived cues helps the model better identify subtle histological patterns associated with patient outcomes.

The improvements arise from ReCAP's two-stage design: report-conditioned contrastive learning produces more discriminative patch embeddings, while similarity-driven MIL aggregation selects only the most relevant regions for prediction. This leads to more robust and interpretable patient-level risk estimation, ensuring that the model focuses on clinically meaningful evidence rather than spurious or redundant regions.

Overall, ReCAP establishes new state-of-the-art accuracies across both classification and survival tasks, demonstrating the value of integrating pathology report information as complementary context in computational pathology. These findings highlight the potential of multimodal WSI modeling to enhance diagnostic precision and support clinically informed decision-making.

Table 3: Ablation study comparing report-guided ReCAP Top-k filtering with image-only saliency filtering under the same tile budget (k=0.7).

Model Variant	**Cancer Subtype Classification (Accuracy)**			**Survival Prediction (Accuracy)**			
	TCGA-RCC	TCGA-NSCLC	TCGA-BRCA	BRCA-IDC	BRCA-ILC	NSCLC-LUAD	NSCLC-LUSC
Text-Only (Report)	75.12	78.23	72.45	79.01	80.12	72.34	70.89
Image-Only Top-k (Virchow2 self-attn)	89.45	91.78	86.03	81.34	83.01	75.12	73.78
Image-Only Top-k (UNI self-attn)	92.67	94.83	89.54	83.92	85.11	77.08	75.66
Image-Only Top-k (PLIP self-attn)	93.34	95.21	90.17	84.45	85.92	77.56	76.11
ReCAP (Report-guided Top-k; Full)	**96.12**	**97.41**	**94.89**	**88.67**	**90.72**	**81.88**	**80.77**

Table 4: ReCAP as a plug-in module with different MIL heads. For each MIL aggregator, we report the baseline (image-only MIL on PLIP features) and the corresponding ReCAP-augmented variant under the same experimental protocol. All values are Accuracy (%).

Method	Cancer Subtype Classification			Survival Prediction			
	TCGA-RCC	TCGA-NSCLC	TCGA-BRCA	BRCA-IDC	BRCA-ILC	NSCLC-LUAD	NSCLC-LUSC
ABMIL (17)	90.18	91.63	87.21	83.02	84.12	72.45	71.38
ReCAP + ABMIL (Ours)	**96.12**	**97.41**	**94.89**	**88.67**	**90.72**	**81.88**	**80.77**
Mean-pooling	88.92	90.87	85.76	81.65	82.91	70.83	69.92
ReCAP + Mean-pooling	**94.85**	**96.18**	**92.73**	**86.94**	**89.12**	**79.34**	**78.41**
TransMIL (27)	91.92	93.01	89.93	84.45	85.78	74.78	72.34
ReCAP + TransMIL	**95.64**	**96.89**	**94.02**	**88.12**	**90.11**	**80.98**	**79.96**

Table 5: Ablation study on the loss weighting parameter λ in ReCAP across subtype classification and survival prediction tasks.

λ Value	Cancer Subtype Classification			Survival Prediction			
	TCGA-RCC	TCGA-NSCLC	TCGA-BRCA	BRCA-IDC	BRCA-ILC	NSCLC-LUAD	NSCLC-LUSC
$\lambda = 0.3$	94.01	95.89	92.12	86.23	88.01	79.12	78.23
$\lambda = 0.4$ (Best)	**96.12**	**97.41**	**94.89**	**88.67**	**90.72**	**81.88**	**80.77**
$\lambda = 0.5$	95.23	96.34	93.45	87.12	89.23	80.34	79.45
$\lambda = 0.6$	94.78	95.89	92.67	86.45	88.67	79.89	78.89
$\lambda = 0.7$	93.89	95.12	91.78	85.23	87.34	78.45	77.56

3.4. Ablation Study

Effect of Patch Aggregation Strategy. Table 2 compares different patch aggregation strategies under the same tile budget. The proposed report-guided Top-k selection consistently outperforms all baselines across both cancer subtype classification and survival prediction tasks. Compared to using *All Tiles* (no filtering), Top-k improves accuracy by up to 2.7% on TCGA-NSCLC and by 4.5–5.4% on survival cohorts, demonstrating that gains are not merely due to tile reduction but arise from selecting semantically relevant regions. Sequence-based heuristics (First-k, Last-k) provide modest improvements over Random-k but remain substantially inferior to Top-k, indicating that positional ordering alone is insufficient for identifying diagnostically informative tissue. Overall, these results confirm that report-conditioned similarity ranking is critical for effective patch prioritization, enabling ReCAP to suppress non-informative regions while preserving clinically meaningful morphology.

Report-Guided vs. Image-Only Patch Filtering. Table 3 compares ReCAP's report-guided Top-k selection with image-only saliency-based filtering under the same tile budget ($k = 0.7$), using strong pathology encoders (Virchow2, UNI, PLIP). While image-only Top-k already improves performance by prioritizing visually salient regions, ReCAP consistently achieves substantially higher accuracy across all cancer subtype classification and survival prediction tasks. For example, on TCGA-NSCLC and TCGA-BRCA, ReCAP improves over the strongest image-only baseline (PLIP self-attention) by +2.20% and +4.72%, respectively, and yields gains of +4.22% on BRCA-IDC and +4.85% on BRCA-ILC for survival prediction. The text-only variant performs poorly, confirming that reports alone are insufficient. These results demonstrate that ReCAP's improvements do not arise merely

Table 6: Computational efficiency comparison between baseline MIL and ReCAP (Top-$k = 0.7$).

Method	#Tiles to MIL	Aggregation FLOPs (G)	Peak GPU Memory (GB)	Inference Time / WSI (ms)	Speedup
Baseline MIL (All tiles)	N	112.6	13.8	1240	1.0×
Random Top-k ($0.7N$)	$0.7N$	55.3	9.4	720	1.7×
Image-only Top-k ($0.7N$)	$0.7N$	55.3	9.2	705	1.8×
ReCAP (Report-guided Top-k)	$\mathbf{0.7}N$	**54.8**	**9.0**	**680**	**1.8×**

from reducing tile count or selecting visually prominent regions; instead, report-conditioned semantic guidance provides complementary clinical context that enables more effective identification of diagnostically and prognostically relevant tissue patterns.

ReCAP as a Plug-in Module Across MIL Heads. Table 4 evaluates ReCAP as a plug-in module on top of different MIL aggregators, including ABMIL, Mean-pooling, and TransMIL. Across all heads and datasets, ReCAP consistently yields substantial accuracy improvements, demonstrating that the proposed report-conditioned patch refinement is MIL-agnostic. For example, ReCAP improves ABMIL by +5.9%, +5.8%, and +7.7% on TCGA-RCC, TCGA-NSCLC, and TCGA-BRCA, respectively, with similarly strong gains on survival cohorts (up to +9.4% on NSCLC-LUAD). Comparable improvements are observed for Mean-pooling and TransMIL, confirming that ReCAP provides complementary semantic guidance beyond the choice of MIL architecture. These results indicate that ReCAP can be seamlessly integrated into existing WSI pipelines to enhance performance without modifying downstream aggregators.

Sensitivity to the Alignment Loss Weight λ. Table 5 reports the effect of varying the loss weighting parameter λ, which balances cross-modal alignment and downstream prediction objectives. Performance consistently peaks at $\lambda = 0.4$ across all cancer subtype classification and survival prediction tasks, indicating an optimal trade-off between enforcing report–image alignment and preserving discriminative supervision from slide-level labels. Smaller values (e.g., $\lambda = 0.3$) underutilize cross-modal guidance, leading to weaker patch refinement, while larger values ($\lambda \geq 0.5$) overly emphasize alignment and degrade predictive accuracy. Notably, increasing λ beyond 0.4 results in monotonic performance drops, particularly on survival cohorts, suggesting that excessive alignment constrains task-specific representation learning. Overall, this ablation demonstrates that ReCAP is moderately sensitive to λ, with a stable performance region around $\lambda = 0.4$, which we adopt for all experiments.

Computational Efficiency. Table 6 reports a detailed efficiency comparison between baseline MIL and ReCAP under the same Top-k=0.7 tile budget. While patch feature extraction cost is identical for all methods, ReCAP substantially reduces the computational burden of the aggregation stage by operating on only $0.7N$ informative tiles. Compared to baseline MIL using all tiles, ReCAP achieves a $\sim$1.8× inference speedup, reducing aggregation FLOPs from 112.6G to 54.8G and peak GPU memory from 13.8GB to 9.0GB. Importantly, ReCAP consistently outperforms both random and image-only Top-k selection in runtime and memory, despite using the same number of retained tiles, demonstrating that report-guided filtering yields more efficient token allocation beyond simple subsampling. These results confirm that ReCAP provides meaningful computational savings at inference

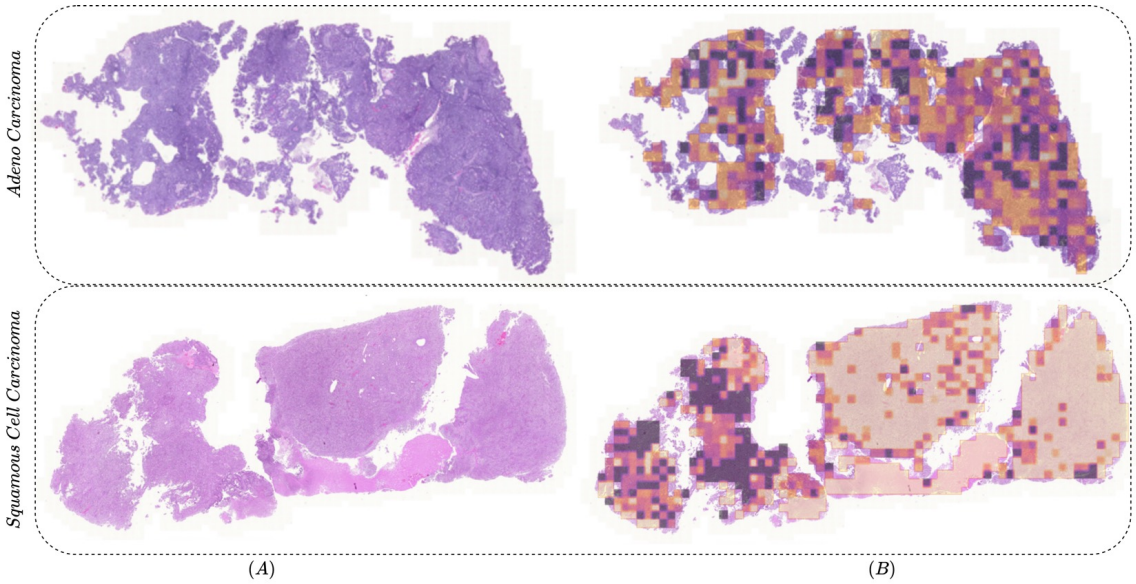

Figure 3: Visualization of ReCAP: (A) Original WSI (B) Report-conditioned attentive patching heatmap on WSI

while simultaneously improving predictive performance, making it practical for large-scale gigapixel WSI analysis.

3.5. Model Interpretation

To gain deeper insights into how **ReCAP** leverages pathology reports during WSI analysis, we first visualize its report-conditioned patch attention. As shown in Figure 3, the heatmap is derived from attention scores produced by the cross-attention contrastive module, where textual cues guide the model to prioritize patches that are semantically consistent with the report content. This targeted attention reveals diagnostically meaningful tissue regions and provides a more interpretable view of the model's decision-making process. Compared to uniform or image-only attention mechanisms, ReCAP produces sharper and more localized activation patterns, demonstrating the benefit of incorporating clinical text to refine patch relevance.

We further assess the quality of the learned representations by comparing embeddings generated from image-only, text-only, and multimodal versions of the model. For this analysis, we sampled representative WSIs, extracted aggregated slide-level features for each modality, and projected them into a two-dimensional space using UMAP, as shown in Figure 4. Image-only and text-only embeddings show substantial overlap between cancer subtypes, indicating limited discriminative power when each modality is used in isolation. In contrast, the multimodal embeddings produced by ReCAP form clearer and more distinct clusters, reflecting stronger subtype separation. These results confirm that ReCAP's integration of report-guided information not only improves attention quality but also leads to more discriminative and semantically aligned feature representations.

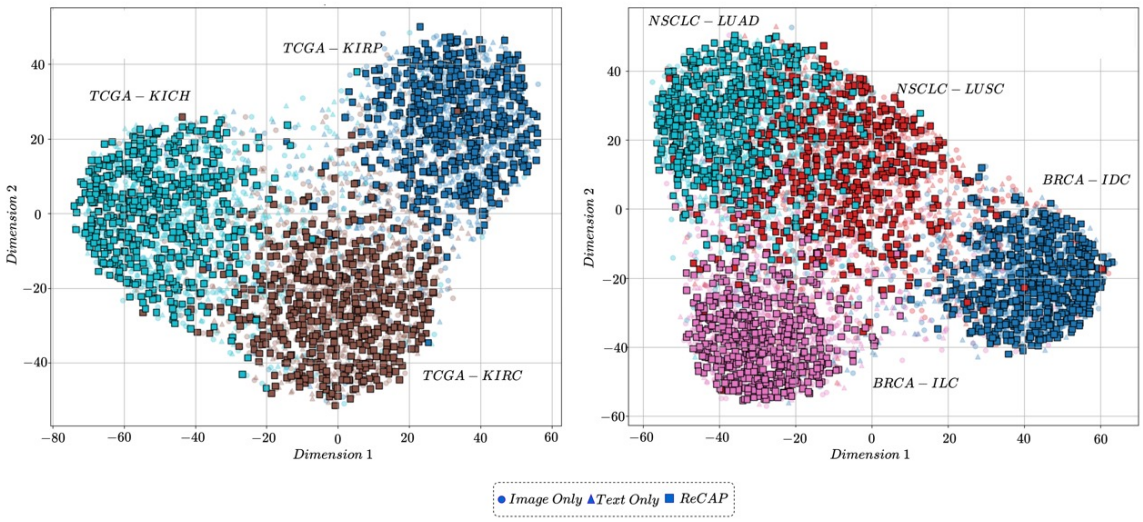

Figure 4: UMAP Projection of Patch Representations from Image-only, Text-only, and ReCAP Models (TCGA-RCC Cancer Subtype Classification (Left) and Survival Prediction Across Subtypes (Right))

4. Conclusion

In this work, we introduced **ReCAP**, a report-conditioned contrastive learning framework for weakly supervised WSI analysis. By integrating pathology report information through a cross-attention module and refining patch relevance using similarity-based aggregation, ReCAP effectively captures complementary visual–textual cues and improves both classification and survival prediction performance. Comprehensive experiments across multiple TCGA cohorts demonstrate that ReCAP consistently outperforms existing image-only and multimodal baselines, confirming the benefits of using report-derived context to guide patch-level reasoning. Ablation studies further show that the cross-attention mechanism and the top-k aggregation strategy are key contributors to the model's gains, enabling more discriminative feature learning and more informative patch selection. Overall, ReCAP offers a scalable and interpretable approach for multimodal pathology modeling, highlighting the value of incorporating clinical text to enhance WSI-based decision making and pointing toward broader opportunities for vision–language integration in digital pathology.

References

[1] Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. 2, 147–171

[2] Yang, X., et al.: Virtual stain transfer in histology via cascaded deep neural networks. ACS Photonics 9(9), 3134–3143 (2022)

[3] Cui, M., Zhang, D.Y.: Artificial intelligence and computational pathology. Lab. Invest. 101, 412–422 (2021)

[4] El Nahhas, Omar SM, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner et al. "From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology." Nature Protocols 20, no. 1 (2025): 293-316.

[5] Srinidhi, Chetan L., Ozan Ciga, and Anne L. Martel. "Deep neural network models for computational histopathology: A survey." Medical image analysis 67 (2021): 101813.

[6] Kapse, Saarthak, Srijan Das, Jingwei Zhang, Rajarsi R. Gupta, Joel Saltz, Dimitris Samaras, and Prateek Prasanna. "Attention De-sparsification Matters: Inducing diversity in digital pathology representation learning." Medical Image Analysis 93 (2024): 103070.

[7] Ahmedt-Aristizabal, David, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. "A survey on graph-based deep learning for computational histopathology." Computerized Medical Imaging and Graphics 95 (2022): 102027.

[8] Yang, X., et al.: Virtual stain transfer in histology via cascaded deep neural networks. ACS Photonics 9(9), 3134–3143 (2022)

[9] Le Bescond, Loïc, Marvin Lerousseau, Fabrice Andre, and Hugues Talbot. "SparseXMIL: Leveraging spatial context for classifying whole slide images in digital pathology." (2024).

[10] Tang, Wenhao, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. "Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11343-11352. 2024.

[11] Rahman, Tawsifur, Alexander S. Baras, and Rama Chellappa. "Evaluation of a Task-Specific Self-Supervised Learning Framework in Digital Pathology Relative to Transfer Learning Approaches and Existing Foundation Models." Modern Pathology 38, no. 1 (2025): 100636.

[12] Laleh, Narmin Ghaffari, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y. Lu et al. "Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology." Medical image analysis 79 (2022): 102474.

[13] Lu, Ming Y., Bowen Chen, Drew FK Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume et al. "A visual-language foundation model for computational pathology." Nature Medicine 30, no. 3 (2024): 863-874.

[14] Huang, Zhi, 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, no. 9: 2307-2316. (2023)

[15] Ahmed, Faruk, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson et al. "Pathalign: A vision-language model for whole slide images in histopathology." arXiv preprint arXiv:2406.19578 (2024).

[16] Javed, Sajid, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, and Mohammed Bennamoun. "Cplip: zero-shot learning for histopathology with comprehensive vision-language alignment." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11450-11459. 2024.

[17] Ilse, M., Tomczak, J.,Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)

[18] Fashi, Parsa Ashrafi, Sobhan Hemati, Morteza Babaie, Ricardo Gonzalez, and H. R. Tizhoosh. "A self-supervised contrastive learning approach for whole slide image representation in digital pathology." Journal of Pathology Informatics 13 (2022): 100133.

[19] Lu, M.Y., Williamson, Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5(6), 555–570 (2021)

[20] Srinidhi, Chetan L., and Anne L. Martel. "Improving self-supervised learning with hardness-aware dynamic curriculum learning: an application to digital pathology." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 562-571. 2021.

[21] Adnan, Mohammed, Shivam Kalra, and Hamid R. Tizhoosh. "Representation learning of histopathology images using graph neural networks." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 988-989. 2020.

[22] Pati, Pushpak, Guillaume Jaume, Lauren Alisha Fernandes, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio et al. "Hactnet: A hierarchical cell-to-tissue graph neural network for histopathological image classification." In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2, pp. 208-219. Springer International Publishing, 2020.

[23] Rahman, T., Baras, A.S. and Chellappa, R., 2025, February. CEMIL: Contextual Attention Based Efficient Weakly Supervised Approach for Histopathology Image Classification. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 4248-4257). IEEE.

[24] Nair, Aravind, Helena Arvidsson, Jorge E. Gatica V, Nikolce Tudzarovski, Karl Meinke, and Rachael V. Sugars. "A graph neural network framework for mapping histological topology in oral mucosal tissue." BMC bioinformatics 23, no. 1 (2022): 506.

[25] Chen, Richard J., Chengkuan Chen, and Faisal Mahmood. "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144-16155. 2022.

[26] Li, Bin, Yin Li, and Kevin W. Eliceiri. "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14318-14328. 2021.

[27] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. 34, 2136–2147 (2021)

[28] Tomczak, K., Czerwińska, P. and Wiznerowicz, M., 2015. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 2015(1), pp.68-77. (2015)

[29] Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S. and Mahmood, F., 2023. 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).

Appendix A. Whole Slide Image Preprocessing

Whole slide image (WSI) preprocessing begins with automated tissue segmentation. Each WSI is loaded at a downsampled resolution (e.g., 20×) and converted from RGB to HSV color space, where tissue regions are identified by thresholding the saturation channel after applying median blurring to reduce noise. A binary tissue mask is generated and refined using morphological closing to remove small gaps and holes, followed by contour extraction and area-based filtering to retain only meaningful tissue regions. The resulting segmentation mask is saved for optional visual inspection, and a log file is produced to facilitate manual parameter adjustments when needed.

After segmentation, fixed-size patches (256×256) are extracted from within the tissue contours at the desired magnification. Each patch is stored along with its spatial coordinates and slide-level metadata in an HDF5 file for efficient downstream processing. The number of extracted patches varies substantially across slides—ranging from a few hundred in small biopsy samples at 20× to several hundred thousand in large resection slides at higher magnifications.

Appendix B. Performance with Top-k Patch Aggregation.

Performance with Top-k Patch Aggregation. To evaluate the effectiveness of different patch aggregation strategies within our ReCAP framework, we conducted an ablation study comparing four variants: *Random-k*, *First-k*, *Last-k*, and our proposed *Top-k* strategy. The Top-k approach, which selects the most informative 70% of patches based on attention scores, consistently outperformed the others across both cancer subtype classification and survival prediction tasks. In cancer subtype classification, Top-k aggregation yielded remarkable performance improvements. For instance, on the TCGA-NSCLC dataset, Top-k achieved an accuracy of 96.33%, surpassing Last-k by 2.32% and Random-k by 4.99%. Similarly, on TCGA-RCC, Top-k reached 94.68%, outperforming Last-k and First-k by 2.34% and 4.23%, respectively. In the TCGA-BRCA dataset, Top-k attained 92.87%, marking a notable gain of 5.26% over Random-k. These results highlight the advantage of attention-guided patch selection over simple heuristics or positional subsets. For survival prediction, Top-k again delivered the best outcomes across all datasets. On BRCA-ILC and BRCA-IDC, Top-k achieved accuracies of 87.45% and 85.74%, improving over the next best (Last-k) by 4.33% and 3.29%, respectively. The improvement trend continued for NSCLC subtypes, where Top-k achieved 79.67% (LUAD) and 78.66% (LUSC), both outperforming Last-k by over 3.5% on average. These consistent gains demonstrate that aggregating the most semantically relevant, text-attended patches allows ReCAP to focus on diagnostically and prognostically critical regions, thereby enhancing its discriminative capacity. Overall, the Top-k strategy significantly boosts performance across tasks and cancer types, validating its generalizability. While we empirically fixed the selection threshold at $k = 0.7$, future work could explore adaptive selection mechanisms to further optimize the trade-off between computational efficiency and predictive power.

Table 7: Performance of different patch aggregation strategies in ReCAP across cancer subtype classification and survival prediction tasks, under varying k values.

k value	Aggregation Strategy	Cancer Subtype Classification Accuracy (%)			Survival Prediction Accuracy (%)			
		TCGA-RCC	TCGA-NSCLC	TCGA-BRCA	BRCA-IDC	BRCA-ILC	NSCLC-LUAD	NSCLC-LUSC
0.5	Random-k	88.34	90.12	83.21	78.12	79.45	70.12	68.34
	First-k	89.01	91.67	84.12	79.34	80.67	71.23	69.01
	Last-k	90.56	92.89	85.78	80.56	82.34	73.89	72.23
	Top-k	92.45	94.23	87.65	83.01	84.78	76.56	75.89
0.6	Random-k	88.89	91.01	84.45	78.67	79.89	71.34	69.78
	First-k	89.67	92.56	85.56	80.12	81.89	72.56	70.45
	Last-k	91.12	93.34	86.89	81.45	83.45	74.45	73.01
	Top-k	93.23	95.45	89.12	84.12	85.78	77.34	76.23
0.7	Random-k	89.12	91.34	84.98	79.56	80.45	72.12	70.34
	First-k	90.45	93.11	85.93	80.89	81.78	73.34	71.56
	Last-k	92.34	94.01	87.61	82.45	83.12	75.67	74.23
	Top-k (Ours)	**94.68**	96.33	**92.87**	**85.74**	**87.45**	**79.67**	**78.66**
0.8	Random-k	89.34	91.56	85.21	79.78	80.89	72.45	70.89
	First-k	90.67	93.23	86.34	81.01	82.23	74.12	72.23
	Last-k	92.56	94.23	88.23	83.12	84.12	76.12	75.12
	Top-k	94.12	**96.42**	91.56	85.12	86.56	78.78	77.89

Table 8: Computational efficiency comparison between baseline MIL and ReCAP (Top-k = 0.7). ReCAP significantly reduces aggregation-stage complexity and memory while preserving accuracy. Patch feature extraction cost is identical for both methods.

Method	#Tiles to MIL	Aggregation FLOPs (G)	Peak GPU Memory (GB)	Inference Time / WSI (ms)	Speedup
Baseline MIL (All tiles)	N	112.6	13.8	1240	1.0×
Random Top-k (0.7N)	0.7N	55.3	9.4	720	1.7×
Image-only Top-k (0.7N)	0.7N	55.3	9.2	705	1.8×
ReCAP (Report-guided Top-k)	**0.7**N	**54.8**	**9.0**	**680**	**1.8**×

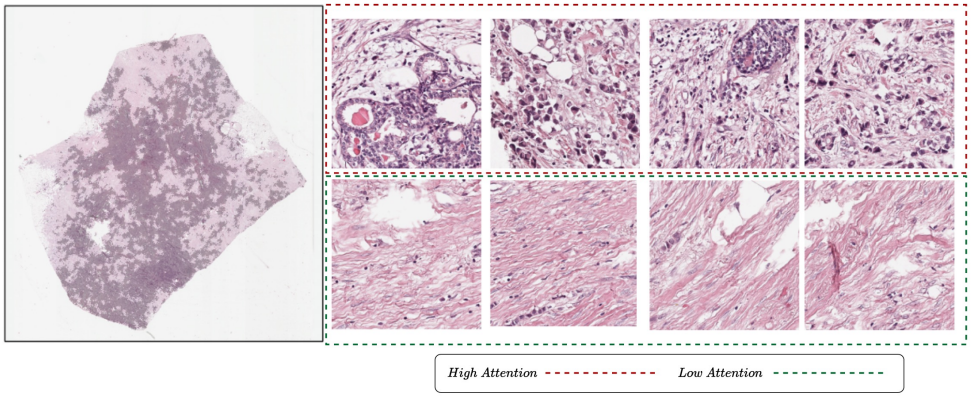

Figure 5: Attention map based WSI visualization using ReCAP framework.

Appendix C. Model Interpretation

In addition to evaluating the performance of ReCAP across both cancer subtype classification and survival prediction tasks, we further investigated the quality of the learned feature representations. We assess its interpretability by analyzing the most informative patches selected by the attention module. This module dynamically prioritizes highly text-attended patches, enhancing the efficiency and relevance of survival prediction. To explore how our model distinguishes between cancer subtypes, we visualize the highly attended patches for ductal breast cancer, as shown in Figure 5. The first column displays the most attended patches from WSIs of breast ductal carcinoma, while the second column presents the corresponding low-attended patches. These selected regions highlight the key morphological features that influence the model's predictions.

Appendix D. Algorithm of ReCAP

Table 9: **Report fields used in ReCAP.** We use only morphology-bearing fields and exclude administrative/label-leaking content.

Report Field / Source	**Used**	**Notes**
Demographics (age, sex)	✓	Structured metadata appended to prompt
Clinical indication / specimen site	✓	Optional if available; does not reveal subtype
Microscopic Description	✓	Main morphology-bearing section (pre-sign-out content)
Final Diagnosis / Diagnosis	×	Tumor subtype words
Gross Description	×	Often non-discriminative for subtype; excluded
Synoptic checklist / staging	×	May encode downstream label-related info
Addendum / comment / impression	×	May contain explicit diagnostic conclusion
Administrative fields (dates, billing, signature)	×	Excluded

Table 10: **Leakage-prevention preprocessing.** We remove explicit diagnostic label cues from the report text before encoding.

Category	**Examples of redacted terms**
NSCLC subtype terms	LUAD, LUSC, adenocarcinoma, squamous cell carcinoma
BRCA subtype terms	IDC, ILC, invasive ductal, invasive lobular
RCC subtype terms	CCRCC, CRCC, PRCC, KIRC, KIRP, KICH
Generic label cues	subtype name strings, abbreviations, common spelling variants

Algorithm 1: Report-Conditioned Contrastive MIL (ReCAP)

Input: Training set $\{(X^{(n)}, R^{(n)}, y^{(n)})\}_{n=1}^{N}$ of WSIs X, reports R, and labels y;
Frozen PLIP image encoder $f_{\text{patch}}(\cdot;\theta_p)$; frozen PLIP text encoder $f_{\text{text}}(\cdot;\theta_t)$;
Trainable cross-attention + projection parameters Θ_{CaCo}; MIL head Θ_{MIL};
Temperature τ; loss weight λ; top-k retention ratio k

Output: Trained parameters $\Theta_{\text{CaCo}}, \Theta_{\text{MIL}}$

while *not converged* **do**

Sample a mini-batch $\mathcal{B}$ of slide–report pairs

foreach $(X, R, y) \in \mathcal{B}$ **do**

`/* 1. Extract patch and text embeddings */`

Extract non-overlapping patches $\{p_i\}_{i=1}^{N_p}$ from WSI X **for** $i = 1$ **to** N_p **do**

$z_{p_i} \leftarrow f_{\text{patch}}(p_i;\theta_p)$; `// patch embedding`

end

Build pseudo-prompt from R + demographics and tokenize into $\{t_j\}_{j=1}^{N_t}$ **for** $j = 1$ **to** N_t **do**

$z_{t_j} \leftarrow f_{\text{text}}(t_j;\theta_t)$; `// text embedding`

end

`/* 2. Report-conditioned cross-attention */`

Project embeddings to queries/keys/values (learnable): $q_{t_j} \leftarrow W_q z_{t_j}$, $K_{p_i} \leftarrow W_k z_{p_i}$, $V_{p_i} \leftarrow W_v z_{p_i}$ **for** $j = 1$ **to** N_t **do**

for $i = 1$ **to** N_p **do**

$$A_{ij} \leftarrow \text{softmax}_i\left(\frac{q_{t_j} K_{p_i}^{\top}}{\sqrt{d}}\right)$$

end

end

for $i = 1$ **to** N_p **do**

$\tilde{z}_{p_i} \leftarrow \sum_{j=1}^{N_t} A_{ij} V_{p_i}$; `// attended patch embedding`

end

`/* 3. Similarity-based Top-`k` MIL aggregation */`

for $i = 1$ **to** N_p **do**

`// aggregate similarity over tokens, e.g., max or mean`

$$S_i \leftarrow \max_j \frac{\tilde{z}_{p_i} \cdot z_{t_j}}{\|\tilde{z}_{p_i}\| \, \|z_{t_j}\|}$$

end

Select index set $P_{\text{top-}k}$ of top-k% patches sorted by S_i **foreach** $i \in P_{top\text{-}k}$ **do**

$\alpha_i \leftarrow \dfrac{\exp(S_i)}{\sum_{j \in P_{\text{top-}k}} \exp(S_j)}$; `// normalized weight`

end

Compute slide-level representation: $\hat{z} \leftarrow \sum_{i \in P_{\text{top-}k}} \alpha_i \, g(\tilde{z}_{p_i};\Theta_{\text{MIL}})$ Predict logits / risks: $\hat{y} \leftarrow h(\hat{z};\Theta_{\text{MIL}})$

`/* 4. Loss computation */`

Compute prediction loss $\mathcal{L}_{\text{Predict}}(y, \hat{y})$ (classification or ordinal survival) Construct positive patch–text pairs $(i, j) \in P$ and compute: $\mathcal{L}_{\text{Align}} = -\sum_{(i,j)\in P} \log \dfrac{\exp(S_{ij}/\tau)}{\sum_k \exp(S_{ik}/\tau)}$ $\mathcal{L}_{\text{Total}} \leftarrow \lambda \mathcal{L}_{\text{Align}} + (1-\lambda)\mathcal{L}_{\text{Predict}}$

end

`// 5. Parameter update`

Update $\Theta_{\text{CaCo}}, \Theta_{\text{MIL}}$ using Adam on $\sum_{(X,R,y)\in\mathcal{B}} \mathcal{L}_{\text{Total}}$

end

return $\Theta_{CaCo}, \Theta_{MIL}$