Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LG-MIL: Language-Guided Multiple Instance Learning for Coordinate-Aligned Attention Analysis in Whole-Slide Images

Mohammad Mehedi Hassan

m_hassan@kcg.edu

The Kyoto College of Graduate Studies for Informatics

Md. Rezaul Bashar

Western Sydney University

Research Article

Keywords: Multiple Instance Learning, Whole-Slide Images, Computational Pathology, Vision–Language Models

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11763838/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

LG-MIL: Language-Guided Multiple Instance Learning for Coordinate-Aligned Attention Analysis in Whole-Slide Images

Mohammad Mehedi Hassan[1*] and Md. Rezaul Bashar[2]

[1*]Department of Information Systems, The Kyoto College of Graduate Studies for Informatics, Sakyo Ward, Kyoto, 606-8225, Kyoto, Japan.
[2]School of Engineering, Western Sydney University, Penrith, NSW, 2751, State, Australia.

*Corresponding author(s). E-mail(s): m_hassan@kcg.ac.jp;
Contributing authors: m.bashar@westernsydney.edu.au;

Abstract

Language-Guided Multiple Instance Learning (LG-MIL) is introduced as a weakly supervised framework for whole-slide image (WSI) analysis. The proposed approach integrates visual and textual information to support interpretable slide-level prediction. Patch-level features extracted from WSIs are combined with class-level textual prompts encoded in a shared multimodal space. To enable semantically informed aggregation under weak supervision, text embeddings guide patch weighting through a cross-modal attention mechanism. A coordinate retrofit strategy is further employed to recover per-patch $(\boldsymbol{x}, \boldsymbol{y})$ locations at inference, enabling the generation of attention heatmaps, top-$\boldsymbol{K}$ tile visualizations, and saliency-based summaries for spatial interpretability. To regulate the sharpness and semantic alignment of attention distributions, the framework incorporates two interpretable controls: the attention temperature $\boldsymbol{\tau}$ and the contrastive loss weight $\boldsymbol{\alpha}$. Experiments on patient-exclusive validation and test splits demonstrate competitive slide-level classification performance, with measurable variation across random initializations and hyperparameter settings. Saliency analyses further investigate the relationship between model confidence and attention dispersion, while ablation studies show that attention temperature influences both attention concentration and predictive performance. Lower temperature values produce more concentrated attention distributions, whereas higher values result in more diffuse attention patterns. Overall, LG-MIL provides a reproducible pipeline for language-guided, coordinate-aligned attention analysis in WSIs, enabling transparent investigation of model behavior under weak

supervision. The framework highlights the potential of text guidance, spatial alignment, and saliency quantification for advancing interpretability in large-scale WSI modeling without requiring dense or clinical annotations.

Keywords: Multiple Instance Learning, Whole-Slide Images, Computational Pathology, Vision–Language Models

1 Introduction

Accurate localization of pathological regions in medical images is a critical requirement for reliable computer-aided diagnosis and clinical decision support systems [1, 2]. Although state-of-the-art deep learning approaches [3] provide excellent performance for disease classification, many clinically relevant problems also require spatially informed explanations of where abnormal patterns are present. However, it is extremely expensive, time-consuming, and in many cases impractical to obtain dense pixel-level or region-level annotations from human experts. As a result, most available medical imaging datasets provide only weak supervision in the form of image-level or case-level labels, which limits the development of accurate localization models.

Weakly supervised localization techniques [4] offer a promising alternative for overcoming this limitation. However, since these models are based only on vision, they may also be misled by spurious correlations or irrelevant parts of the input, thereby compromising the transparency of the models. At the same time, rich semantic information about disease characteristics, morphological patterns, and contextual cues often exists in textual sources. These text-based descriptions represent expert knowledge that has the potential to direct visual models to more semantically meaningful regions but is currently underleveraged as a supervisory signal to localizations.

To this end, this research introduces a language-guided attention framework for weakly supervised whole-slide image analysis and attention-based localization. The goal is to leverage the connection between vision and meaning by injecting class-level text prompts into the attention mechanism. The proposed framework is intended to direct weakly supervised models to more semantically meaningful regions of images, as supported by expert knowledge. The goal of this work is to develop a weakly supervised framework that operates in a multimodal space and facilitates anomaly localization using language-guided attention. By mapping visual features to semantic information provided in text prompts, this framework aims to increase the interpretability of deep learning models through spatial attention maps that facilitate transparent inspection of model behavior. Such interpretability tools may support downstream analysis workflows in computational pathology by enabling more transparent inspection of model behavior. Moreover, this study aims to investigate whether language guidance can influence attention allocation and improve interpretability relative to vision-only weakly supervised baselines, particularly in settings where dense annotations are unavailable.

2 Related Works

Weakly supervised learning (WSL) is one of the most popular paradigms adopted in the field of computational pathology owing to the unavailability of pixel-level or region-level annotations for whole-slide images (WSIs) [4–6]. The initial methods focused on the application of multiple instance learning (MIL) to WSI analysis. This involved the representation of a whole slide as a bag of patch features. The labels were provided only on a whole slide basis [7–9]. Attention-based MIL proposed a mechanism to estimate the importance of the patches in the bag. This allowed the application of coarse localization using weak labels [10]. Later, more variants were proposed to enhance the quality of the representations by the application of gating functions, transformers, and hierarchical pooling [11, 12]. However, all the aforementioned methods are purely visual-based and may not provide interpretable attention maps.

The techniques for weakly supervised localization try to focus on discriminative regions without relying on dense annotations. For WSIs, localization can be performed by attention heatmaps [13], class activation mapping (CAM) [14], and transformer relevance scores [15]. Nevertheless, vision-based localization techniques might suffer from overfitting to data-specific correlations and non-semantic visual patterns [16]. In addition, the localization problem remains ill-defined in the context of weakly supervised learning. The attention patterns can vary substantially for different random seeds and training paths, even when the classification performance is consistent [17]. Hence, additional supervision can be useful in guiding attention towards more semantically relevant regions. Significant recent breakthroughs have been made in cross-modal learning tasks to leverage the potential of text-image representation for better alignment and interpretability. Models like Contrastive Language-Image Pre-Training (CLIP) [18] have been successful in learning contrastive objectives to align the image and text embeddings for zero-shot recognition tasks and text-informed attention mechanisms. Subsequent works have also been successful in leveraging text-informed attention mechanisms for natural images [19, 20], proving the point that text offers an extremely rich context for better image interpretability. Text-image fusion has also been explored for tasks like medical report generation, multimodal retrieval, and text-image embedding learning [21, 22]. However, relatively few studies have explored text-guided attention mechanisms for WSI-level localization under weakly supervised settings, where text can serve as an interpretable source of guidance for attention distributions. For accurate spatially grounded interpretability, it is necessary that patch-level features are correctly associated with their coordinates in WSIs. In most pipelines of weakly supervised methods, spatial information is not utilized during tiling, which makes localization and heatmap visualization challenging [23]. Some methods that have utilized this information or have provided post-processing capabilities for coordinate recovery have proven effective in accurate visualization of heatmaps, spatial saliency maps, and even top K region detection for analysis [24, 25]. However, few methods have utilized this information in conjunction with text-conditioned attention, which makes this a relevant area of research.

Across weak supervision, MIL, cross-modal learning, and coordinate-aligned interpretability, most existing approaches, as discussed above, lack mechanisms to incorporate semantics from textual descriptions into patch-level attention, and many

ignore the need for spatial alignment for faithful localization. The proposed LG-MIL framework addresses these gaps by (i) incorporating class-level text prompts into a cross-modal attention mechanism, and (ii) providing coordinate-aligned localization outputs such as WSI heatmaps, top-K tiles, and quantitative saliency metrics. Together, these contributions combine language guidance, MIL interpretability, and coordinate-aware WSI analysis within a weakly supervised framework for whole-slide image analysis.

3 Methodology

In the proposed approach, a multimodal learning framework is followed in which precomputed UNI visual features and frozen CLIP text embeddings are integrated under a weakly supervised scheme. Precomputed patch features from whole-slide images, extracted using the UNI foundation model [26], serve as the visual input, while class-level textual prompts are encoded using a pretrained CLIP [18] text encoder to obtain 512-dimensional semantic text embeddings. A cross-modal attention mechanism is formulated to allow the text embedding to guide patch-level attention weights, enabling the model to emphasize prompt-conditioned regions of the slide during prediction. Weak supervision is employed in the form of slide-level labels, and a contrastive regularization term encourages semantic consistency between the aggregated visual features and the corresponding class prompt. The framework is designed to be data-agnostic and applicable to large-scale whole-slide imaging datasets. For empirical evaluation, the dataset is partitioned into training, validation, and test sets on a per-patient basis to prevent information leakage. Performance evaluation is performed using slide-level classification metrics, saliency-based analyses, and visual inspection of text-guided attention maps. Ablation studies are conducted to analyze the influence of the attention temperature τ and contrastive loss weight α on model behavior.

In the proposed framework (Fig. 1), class-level text prompts are first encoded into a semantic embedding, while each whole-slide image is processed via patch extraction and a visual encoder to produce patch-level features. The text embedding guides a cross-modal attention module that highlights regions of the slide most relevant to the prompt. The attended patch features are aggregated through a MIL pooling operation to form a slide-level representation used for classification under weak supervision. During inference on validation/test slides, coordinate retrofit is applied to recover the (x, y) locations of attended patches, enabling heatmap generation, extraction of top-K tiles with coordinates, and computation of saliency metrics without requiring pixel-level annotations.

3.1 Dataset Construction and Preprocessing

We curated a subset of whole-slide images (WSIs) from the publicly available TCGA-BRCA [27] collection. Each WSI is associated with a slide-level binary label derived from publicly available pathology diagnosis metadata. The binary task was defined using the TCGA-BRCA metadata field `diagnoses.primary_diagnosis`. Slides with primary diagnoses containing "lobular carcinoma" were assigned to Class 0, whereas slides with primary diagnoses containing "infiltrating duct carcinoma" were assigned

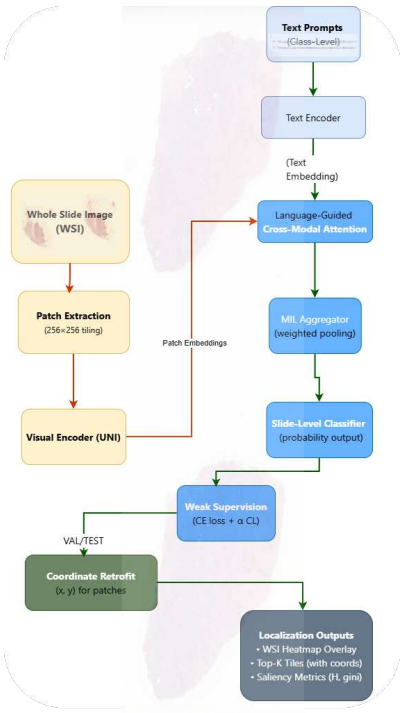


Fig. 1 Overall framework of the proposed language-guided MIL method, where class-level text prompts guide cross-modal attention over projected patch embeddings, followed by MIL pooling and slide-level prediction under weak supervision, with coordinate retrofit applied at inference to recover spatial locations for heatmap overlays, Top-K tiles, and saliency metrics.

Table 1 Dataset composition and class distribution.

Split	Class 0	Class 1	Total
Train	45 (23.9%)	143 (76.1%)	188
Validation	4 (20.0%)	16 (80.0%)	20
Test	4 (20.0%)	16 (80.0%)	20
Total	53 (23.2%)	175 (76.8%)	228

to Class 1. The corresponding class-level prompts were: Class 0, "Histopathology shows invasive lobular carcinoma of the breast," and Class 1, "Histopathology shows invasive ductal carcinoma of the breast." To ensure strict patient-level independence, all slides were grouped by patient barcode (TCGA-XX-YYYY), and then partitioned into train, validation, and test splits such that no patient appears in more than one split. Each WSI was tiled at level-0 into fixed-size 256×256 patches with a stride of 256 pixels. Tiles dominated by background were removed by applying a white-pixel threshold of 220 and a minimum tissue ratio of 30%. Patch-level features were obtained from

a precomputed UNI feature [26] directory, storing one feature vector per patch. For interpretability, we constructed a coordinate-enabled (COORD) subset by retrospectively calibrating the (x, y) location of each patch on the original WSI. For localization analysis, a coordinate-enabled subset of the validation and test slides was constructed by retrospectively associating patch features with valid (x, y) coordinates. This subset was used exclusively for localization visualizations and saliency analyses.

3.2 Patch Feature Bags

For each slide, we assemble a bag of patch features (B),

$$B = \{p_1, p_2, \ldots, p_N\}, \qquad p_i \in \mathbb{R}^D, \tag{1}$$

where each p_i corresponds to one image patch extracted from the whole slide. Since a slide is too large to process at once, it is broken into many smaller patches. Here, $\mathbb{R}$ denotes the set of real numbers. Each patch p_i is represented as a feature vector of dimension D, i.e., $\mathbb{R}^D$ denotes a D-dimensional real-valued vector space, and N is the number of valid tissue patches. These bags are the inputs to the Multiple Instance Learning (MIL) model.

For slides included in the coordinate-enabled localization subset, the corresponding patch features are associated with a coordinate matrix,

$$C = \{(x_1, y_1), \ldots, (x_N, y_N)\}, \tag{2}$$

enabling geometric reconstruction during localization.

3.3 Language-Guided MIL Model

Our model is constructed around a language-guided cross-modal attention mechanism, integrating text prompts with patch features for slide-level predictions.

3.3.1 Text Encoder

Each class is associated with a user-defined textual description (class prompt). The text encoder embeds this description into a vector:

$$t_{\text{cls}} = \text{TextEncoder}(\text{prompt}). \tag{3}$$

3.3.2 Visual Projection

Each precomputed UNI patch feature $p_i \in \mathbb{R}^{1024}$ is mapped to the 512-dimensional multimodal representation space using a two-layer projection network:

$$v_i = W_2 \,\text{ReLU}\,(W_1 p_i + b_1) + b_2, \tag{4}$$

where $W_1 \in \mathbb{R}^{512\times1024}$ and $W_2 \in \mathbb{R}^{512\times512}$ are learned projection matrices. The resulting representation $v_i \in \mathbb{R}^{512}$ is subsequently used by the text-conditioned attention module.

3.3.3 Class-Guided Cross-Modal Attention

The projected patch embeddings interact with the class-level text representation through a learned text-gated attention mechanism. Let $v_i \in \mathbb{R}^d$ denote the projected visual embedding of patch i, and let $t_c \in \mathbb{R}^d$ denote the frozen CLIP embedding of the prompt associated with class c.

The class-level text embedding is first projected into the attention space:

$$\tilde{t}_c = W_t t_c + b_t. \tag{5}$$

The visual and textual gating representations are then computed as

$$h_i^{(v)} = \tanh(W_V v_i + b_V), \tag{6}$$

and

$$h_c^{(t)} = \sigma(W_U \tilde{t}_c + b_U), \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The text representation is shared across all patches within the bag. Class-conditioned interaction is introduced through element-wise gating between the visual and textual representations:

$$a_{ic} = w_a^\top \left(h_i^{(v)} \odot h_c^{(t)} \right) + b_a, \tag{8}$$

where $\odot$ denotes element-wise multiplication and a_{ic} is the attention logit assigned to patch i when conditioned on class prompt c.

The attention logits are normalized across all N patches using a temperature-scaled softmax:

$$w_{ic} = \frac{\exp(a_{ic}/\tau)}{\sum_{j=1}^{N} \exp(a_{jc}/\tau)}. \tag{9}$$

Here, $\tau > 0$ controls the selectivity of the attention distribution. Lower values of τ produce sharper and more concentrated attention distributions, whereas higher values produce smoother and more diffuse distributions.

3.3.4 MIL Aggregation and Classification

The slide representation is the weighted sum:

$$z_{\text{bag}}^{(c)} = \sum_{i=1}^{N} w_{ic} v_i. \tag{10}$$

$z_{\text{bag}}^{(c)}$ is the slide representation obtained when the attention mechanism is conditioned on class prompt c.

3.3.5 Class-Wise Prompt Inference

During inference, the ground-truth class label is not used to select a class prompt. Instead, the slide is evaluated independently using each class-level prompt. For each

Algorithm 1 Language-Guided Attention MIL Training

Require: Bag $B = \{p_i\}_{i=1}^{N}$, slide label y, label-associated class prompt $\texttt{prompt}_y$, attention temperature τ, and contrastive weight α

Ensure: Loss $\mathcal{L}$, bag representation z_{bag}

$t_y \leftarrow \text{TextEncoder}(\texttt{prompt}_y)$
for $i = 1$ **to** N **do**
    $v_i \leftarrow \text{VisualProjection}(p_i)$
end for
$a_i \leftarrow f_{\text{attn}}(v_i, t_y)\ \forall i$
$w_i \leftarrow \text{softmax}_i(a_i/\tau)$
$z_{\text{bag}} \leftarrow \sum_{i=1}^{N} w_i\, v_i$
$\boldsymbol{\ell} \leftarrow \text{Classifier}(\text{stopgrad}(z_{\text{bag}}))$
$\bar{t}_c \leftarrow \text{Normalize}(\text{TextEncoder}(\text{prompt}_c)), \quad c = 0, \ldots, C-1$
$\mathcal{L} \leftarrow \text{CE}(\ell, y) + \alpha\, \mathcal{L}_{\text{contrastive}}\left(z_{\text{bag}}, \{\bar{t}_c\}_{c=0}^{C-1}, y\right)$
return $(\mathcal{L}, z_{\text{bag}}, \{w_i\}_{i=1}^{N})$

class c, the slide is evaluated using the corresponding class prompt, producing a class-conditioned score s_c. The class-conditioned score is defined as the logit corresponding to class c,

$$\ell^{(c)} = g_\phi\left(z_{\text{bag}}^{(c)}\right), \tag{11}$$

$$s_c = \ell_c^{(c)}, \tag{12}$$

where $\ell_c^{(c)}$ denotes the class-c logit obtained from the prompt-conditioned forward pass using class prompt c.

The class-wise scores are converted into probabilities:

$$p_c = \frac{\exp(s_c)}{\sum_k \exp(s_k)}. \tag{13}$$

For binary classification, the predicted class is determined as

$$\hat{y} = \mathbb{I}(p_1 \geq \theta), \tag{14}$$

where θ is selected on the validation set by maximizing balanced accuracy and then applied unchanged to the test set. For localization analysis, attention maps are generated using the prompt associated with the predicted class rather than the ground-truth label.

3.4 Training Objective

During training, the class-level prompt corresponding to the slide label is used for both attention guidance and contrastive alignment. Given a bag of patch features and a text prompt, the algorithm computes text-conditioned attention over patches, aggregates them into a bag representation, and optimizes a combined cross-entropy and contrastive objective. The stop-gradient operation prevents the cross-entropy term

from propagating through the bag representation. Consequently, the cross-entropy term updates the classifier, while the contrastive term provides the optimization signal for the visual-projection and attention pathways. During training, slide-level labels are used to select the corresponding class prompt for attention guidance and contrastive alignment. During validation and testing, class prompts are evaluated independently as described in the class-wise inference procedure above. The objective combines two terms.

3.4.1 Cross-Entropy Loss (CE)

Standard cross-entropy between predicted logits and ground-truth slide labels.

3.4.2 Class-Prototype Contrastive Loss (Weight α)

The pretrained CLIP text encoder is kept frozen and produces one text prototype for each class. Let $\{\bar{t}_c\}_{c=0}^{C-1}$ denote the ℓ_2-normalized class-prompt embeddings, and let $\bar{z}_b$ denote the normalized bag representation of slide b:

$$\bar{z}_b = \frac{z_b}{\|z_b\|_2}, \qquad \bar{t}_c = \frac{t_c}{\|t_c\|_2}. \tag{15}$$

The similarity between slide b and class-prompt prototype c is computed as

$$q_{bc} = \frac{\bar{z}_b^\top \bar{t}_c}{\tau_{\text{con}}}, \tag{16}$$

where τ_{con} is the contrastive temperature. The class-prototype contrastive loss is then defined as

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{M}\sum_{b=1}^{M} \log \frac{\exp(q_{by_b})}{\sum_{c=0}^{C-1}\exp(q_{bc})}, \tag{17}$$

where M is the minibatch size and y_b is the slide-level class label. The contrastive temperature was set to $\tau_{\text{con}} = 0.05$.

3.4.3 Total Loss

$$\mathcal{L} = \text{CE} + \alpha \cdot \mathcal{L}_{\text{contrastive}}. \tag{18}$$

We optimize with Adam (learning rate 10^{-4}) for 10 epochs. Short training is sufficient because feature extraction is precomputed.

3.5 Localization and Coordinate-Aligned Outputs

During inference, attention weights w_i for each patch in a slide are calculated by the text-conditioned attention mechanism defined in Algorithm 1. The model evaluates all class prompts independently using the class-wise inference procedure, and attention maps are generated using the prompt associated with the predicted class. The attention weights quantify the relative model-assigned attention to individual patches under the selected prompt and are interpreted here as model-behavior signals rather than causal measures of patch importance.

3.5.1 Attention Weight Export

For each slide, the normalized attention weights are exported and stored as a NumPy array. In addition, the top-K patches with the highest attention scores (default $K = 200$) are recorded in a CSV table. When spatial metadata are available, the corresponding patch coordinates (x, y) are included for each entry, enabling spatial localization of high-attention regions on the whole slide image.

3.5.2 WSI Attention Heatmap

Heatmaps are generated to visualize the spatial distribution of model attention across the whole slide image. First, a blank canvas is initialized using a low-resolution thumbnail representation of the WSI to preserve the global slide structure. Next, for all the patches for which the spatial coordinates (x, y) are known, a Gaussian blurred disc is rendered at the respective coordinates. The intensity of each rendered disc is proportional to the corresponding normalized attention weight w_i. The smoothing operation spatially spreads each patch-level attention weight over a local neighborhood for visualization. Once all the patches have been projected onto the canvas, the attention weights are normalized to ensure consistent intensity scaling for all the slides. Finally, the attention weights are overlaid over the thumbnail of the WSI to create a heatmap.

3.6 Top-K Tile Extraction

For the highest-weight patches, we crop 256×256 image tiles directly from the original SVS whole-slide image using the stored patch coordinates. Coordinates provide the physical coordinates corresponding to the patches' positions in the image. In other words, by knowing the coordinates of a specific patch, we can extract the regions of interest from the full resolution slides.The extracted patches correspond to regions receiving the highest attention weights under the selected prompt. The tiles of interest are saved to enable visualization of the regions that have been considered important by the model.

3.6.1 Debug Panel

A set of diagnostics is provided for each WSI in the form of a three-panel visualization. The first panel shows the WSI thumbnail, the second panel shows the spatial distribution of the top-K attended patch coordinates, and the third panel presents the attention heatmap overlay. These visualizations are intended to facilitate qualitative analysis of model behavior and interpretability, and are not designed for clinical decision-making.

3.7 Saliency Metrics (Model-Behavior Only)

Saliency metrics are computed to quantify the distribution of attention across patches and to characterize the model's focus within each slide. These metrics are used solely for analyzing model behavior and interpretability, and are not intended for clinical or diagnostic decision-making.

3.7.1 Entropy and Normalized Entropy

Entropy H represents the concentration or dispersion of the attention weights, whereas the normalized entropy H_{norm} normalizes the entropy by a factor of $\log(N)$, which allows for the comparison of the entropy across slides with varying numbers of patches.

$$H = -\sum_i w_i \log w_i, \qquad H_{\text{norm}} = H/\log(N). \tag{19}$$

3.7.2 Gini Coefficient

The Gini coefficient quantifies the inequality of the attention weights across patches. Given normalized attention weights $\{w_{(i)}\}_{i=1}^{N}$ sorted in ascending order,

$$G = \frac{2}{N}\sum_{i=1}^{N} i\, w_{(i)} - \frac{N+1}{N}, \tag{20}$$

A higher Gini value indicates that attention is concentrated on fewer patches, while a lower value corresponds to a more uniform distribution.

3.7.3 Coverage Curves

Coverage curves measure how many top-ranked patches (K) are required to accumulate specified fractions of the total attention mass (50%, 80%, 90%, and 95%), indicating how concentrated the model's focus is within the slide.

3.7.4 Top-K Mass

The Top-K mass measures the cumulative attention weight assigned to the K highest-ranked patches, indicating how much of the model's total attention is concentrated within the most influential regions of the slide.

$$\text{top-K_sum} = \sum_{i \in \text{topK}} w_i. \tag{21}$$

3.8 Ablation Study ($\tau \times \alpha$ Grid)

To analyze the influence of key hyperparameters, we perform an ablation study over combinations of the temperature parameter τ and the contrastive loss weight α. The temperature parameter controls the sharpness of the attention distribution, while α regulates the contribution of the contrastive alignment between slide representations and text prompts.

The method is evaluated under configurations with $\tau \in \{0.5, 1.0, 2.0\}$ and $\alpha \in \{0.0, 0.3, 0.7, 1.0\}$. For each (τ, α) setting, model performance is measured using validation and test classification metrics, including AUC and accuracy.

3.9 Seed Stability Analysis

To evaluate sensitivity to random initialization, we use the selected configuration ($\tau = 1.0$, $\alpha = 1.0$) and repeat training using six random seeds: $\{1, 2, 3, 4, 5, 42\}$. For each run, the checkpoint is selected using validation AUC. A seed-specific decision threshold is then selected on the validation set by maximizing balanced accuracy and applied unchanged to the corresponding test predictions. Classification performance is summarized using validation and test AUC, accuracy, and balanced accuracy.

4 Experiments and Results

In this section, we evaluate the proposed Language-Guided Attention MIL framework on the TCGA-BRCA dataset to assess its effectiveness in slide-level classification. We start by explaining our experimental setting. Next, we compare our model to a standard attention-based MIL baseline. Finally, we perform ablation studies that examine the importance of each component in our model. We also provide some quantitative and qualitative results that emphasize the benefits of using text-guided attention in the MIL process.

4.1 Classification Performance

Table 2 summarizes the classification behavior of the proposed language-guided MIL (LG–MIL) framework on patient-exclusive splits. For the canonical seed-42 model, the decision threshold selected on the validation set was $\theta = 0.639444$. This threshold was applied unchanged to the held-out test set. To quantify uncertainty arising from the limited test-set size, 95% confidence intervals for test AUC, accuracy, and balanced accuracy were estimated using 10,000 stratified bootstrap resamples. The confidence intervals for the held-out test metrics were relatively wide, reflecting the limited size and class imbalance of the test set. Accordingly, the reported point estimates should be interpreted together with both bootstrap uncertainty and the multi-seed variability reported in the stability analysis.

Using precomputed UNI patch features and a 10-epoch optimization schedule, the model was evaluated using a class-wise prompt inference protocol. During inference, all class prompts were evaluated independently, and the final prediction was obtained from the resulting class-wise scores. As the goal of this work is interpretability and weakly supervised localization rather than incremental improvements in classification accuracy, we emphasize downstream attention behavior in subsequent analyses.

Table 2 Classification performance on patient-exclusive validation and test splits for the canonical seed-42 LG–MIL model. Test-set 95% confidence intervals were estimated using 10,000 stratified bootstrap resamples.

Split	AUC	Accuracy	Balanced Accuracy
Validation	0.8750	0.9000	0.8438
Test	0.9062 (0.7500–1.0000)	0.8500 (0.7000–1.0000)	0.8125 (0.5625–1.0000)

Table 3 Comparison with attention-based MIL, CLAM-SB, and neutral-prompt LG–MIL baselines.

Method	Val AUC	Test AUC	Val Acc	Test Acc
Attention MIL	0.7969	0.8750	0.700	0.850
CLAM-SB	0.5469	0.8750	0.650	0.900
LG–MIL (Neutral)	0.8281	**0.9219**	0.800	0.850
LG–MIL (Pathology)	**0.8750**	0.9062	**0.900**	0.850

To assess whether the classification task is inherently trivial and whether semantic prompts contribute beyond generic language conditioning, we compare LG–MIL with three baselines: a standard attention-based MIL model, CLAM-SB, and an LG–MIL variant using neutral class prompts. The neutral-prompt variant uses the same prompt for both classes, thereby removing class-specific semantic information while preserving the language-guided architecture. As shown in Table 3, the standard attention-based MIL and CLAM-SB baselines both achieved test AUC values of 0.8750, while the neutral-prompt LG–MIL variant achieved a test AUC of 0.9219. The pathology-prompt LG–MIL model achieved a validation AUC of 0.8750 and a test AUC of 0.9062, with validation and test accuracies of 0.900 and 0.850, respectively.

The comparison in Table 3 shows that the pathology-prompt LG–MIL model achieved the highest validation AUC among the evaluated methods and remained competitive on the held-out test set. However, the neutral-prompt LG–MIL variant achieved a slightly higher test AUC (0.9219) than the pathology-prompt model (0.9062). Thus, class-specific pathology prompts do not provide a consistent classification advantage over neutral language conditioning in this experiment. The contribution of semantic prompting should therefore not be inferred from classification performance alone. Its influence is examined further through attention behavior, spatial localization, and prompt-conditioned analyses.

4.2 Localization Behavior

Localization quality is assessed through coordinate-aligned visualizations made possible by our retrofit pipeline. For all validation and test slides with associated coordinate files, the model produces WSI-aligned heatmaps by projecting patch-level attention weights back onto slide coordinates. In addition, the model exports top-K tiles (256×256) for the most influential patches and creates a diagnostic figure with three panels showing the WSI thumbnail, the spatial distribution of top-K coordinates, and the heatmap overlay (Fig. 2). Localization outputs are generated using the predicted-class prompt obtained through the class-wise prompt inference procedure. Visual inspection of the validation and test outputs indicates that the spatial alignment is preserved without apparent rotation, drift, or coordinate mismatch. Attention patterns are generally diffuse across large tissue regions while remaining geometrically aligned with the original slide coordinates.

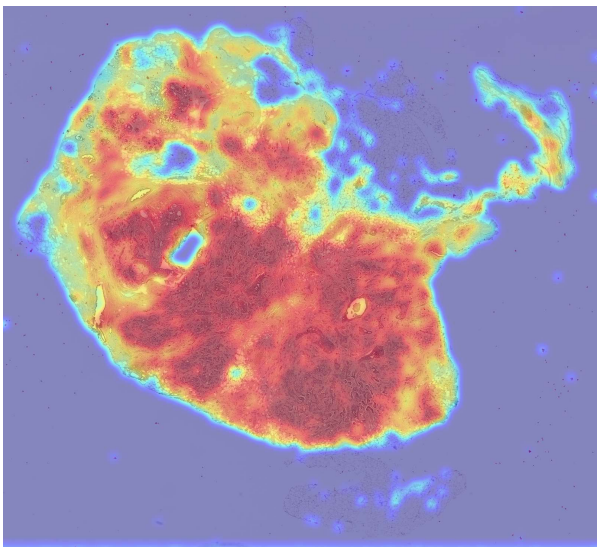


Fig. 2 Representative coordinate-aligned attention heatmap generated using the predicted-class prompt. The visualization maps patch-level attention weights to their corresponding locations on the whole-slide image. The heatmap represents model-assigned attention and does not constitute ground-truth pathological localization.

4.3 Saliency and Attention Analysis

The following metrics were calculated in order to quantify attention behavior: entropy, normalized entropy, Gini coefficient, attention mass coverage, and top-K mass. For convenience, we state the definitions of each metric using Equations (22)–(24).

$$H = -\sum_{i=1}^{N} w_i \log w_i, \tag{22}$$

$$H_{\text{norm}} = \frac{H}{\log N}, \tag{23}$$

$$\text{top-K_sum} = \sum_{i \in \text{top-K}} w_i, \tag{24}$$

Fig. 3 illustrates the distribution of normalized entropy and Gini coefficients across test slides. The normalized entropy values are generally high, whereas the Gini coefficients remain relatively low to moderate, indicating that attention is typically distributed across broad tissue regions rather than concentrated on only a few patches.

4.3.1 Correlation Between Confidence and Saliency

Subsequently, correlations were calculated for the correlation of model confidence with the saliency measures for both the validation set and the test set. This is shown in Table 4. The correlations were computed using confidence scores obtained from the class-wise prompt inference procedure. Under the class-wise prompt inference protocol, no strong confidence–saliency relationship was observed. Test-set correlations between

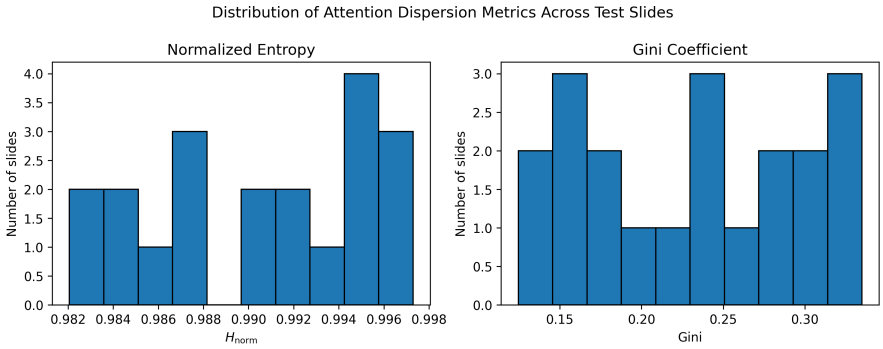


Fig. 3 Histogram of normalized entropy across test slides and distribution of Gini coefficients across slides. The distributions indicate that attention is generally diffuse across slides, reflected by high normalized entropy and relatively low-to-moderate Gini coefficients.

Table 4 Spearman correlations between confidence and saliency metrics using class-wise prompt inference.

Pair	Split	r	p
$\text{prob}_{\text{true}}$ vs. H_{norm}	Test	-0.036	0.880
$\text{prob}_{\text{true}}$ vs. Gini	Test	+0.014	0.955
$\text{prob}_{\text{true}}$ vs. H_{norm}	Val	+0.221	0.349
$\text{prob}_{\text{true}}$ vs. Gini	Val	-0.232	0.326

true-class probability and normalized entropy or Gini coefficient were close to zero, while validation-set correlations were weak and not statistically significant.

4.4 Ablation Study: Attention Temperature (τ) and Contrastive Weight (α)

We performed a 3×4 ablation study over the hyperparameters $\tau \in \{0.5, 1.0, 2.0\}$ and $\alpha \in \{0.0, 0.3, 0.7, 1.0\}$ to study their effects on model behavior and classification performance. As shown in Table 5, performance varied across the evaluated configurations. The highest validation AUC (0.8750) was obtained for $\tau = 0.5$ with $\alpha = 0.0$ and for $\tau = 1.0$ with $\alpha \in \{0.3, 0.7, 1.0\}$. The highest test AUC (0.9062) was obtained for $\tau = 0.5$ with $\alpha = 0.0$, for $\tau = 1.0$ with $\alpha \in \{0.7, 1.0\}$, and for $\tau = 2.0$ with $\alpha = 0.0$.

The inclusion of $\alpha = 0$ serves as a no-contrastive-alignment control. Under the stop-gradient training design, setting $\alpha = 0$ removes the contrastive optimization signal to the visual-projection and attention pathways; consequently, these pathways receive no gradient-based updates, while the classifier is still optimized by the cross-entropy objective. In contrast, configurations with $\alpha > 0$ optimize the visual-projection and attention pathways through the class-prototype contrastive objective, while the cross-entropy term updates the classifier. Thus, the α ablation evaluates the effect of

contrastive alignment within this decoupled training design rather than the presence or absence of the text-conditioned attention mechanism itself.

Notably, the $\alpha = 0$ configurations retained competitive classification performance despite the absence of gradient-based optimization of the visual-projection and attention pathways. This suggests that the precomputed UNI representations contain substantial discriminative information for the present binary task and indicates that classification performance alone should not be interpreted as evidence for the benefit of learned semantic attention.

For configurations with $\alpha > 0$, performance was generally lower at $\tau = 2.0$ than at $\tau = 0.5$ or $\tau = 1.0$. Because attention is computed using $\mathrm{softmax}(a_i/\tau)$, lower values of τ produce sharper and more concentrated attention distributions, whereas higher values produce smoother and more diffuse distributions. Thus, τ directly controls the concentration of the normalized attention weights and is also associated with differences in classification performance under the evaluated settings.

Overall, the ablation indicates that both τ and α affect the behavior of the model, but their effects are not monotonic. The contrastive weight determines whether and to what extent the visual-projection and attention pathways are optimized, while the attention temperature directly controls the concentration of the resulting attention distribution.

Table 5 Ablation over attention temperature (τ) and contrastive weight (α).

τ	α	Val AUC	Test AUC	Val Acc	Test Acc
0.5	0.0	**0.8750**	**0.9062**	**0.900**	0.800
0.5	0.3	0.8125	0.8750	0.750	0.800
0.5	0.7	0.7969	0.8750	0.750	0.800
0.5	1.0	0.8125	0.8750	0.750	0.800
1.0	0.0	0.8594	**0.9062**	**0.900**	0.800
1.0	0.3	**0.8750**	0.8906	**0.900**	**0.850**
1.0	0.7	**0.8750**	**0.9062**	**0.900**	**0.850**
1.0	1.0	**0.8750**	**0.9062**	**0.900**	**0.850**
2.0	0.0	0.8594	**0.9062**	**0.900**	0.800
2.0	0.3	0.8125	0.8594	0.700	**0.850**
2.0	0.7	0.7812	0.8594	0.800	**0.850**
2.0	1.0	0.7812	0.8594	0.800	**0.850**

4.5 Seed Stability Analysis

We repeated the selected configuration ($\tau = 1.0$, $\alpha = 1.0$) using six random seeds $\{1, 2, 3, 4, 5, 42\}$ to assess sensitivity to initialization. As summarized in Table 6, validation AUC was 0.7734 ± 0.0583, while test AUC was 0.8542 ± 0.0703. Validation balanced accuracy was 0.7969 ± 0.0431, and test balanced accuracy was 0.8021 ± 0.0920.

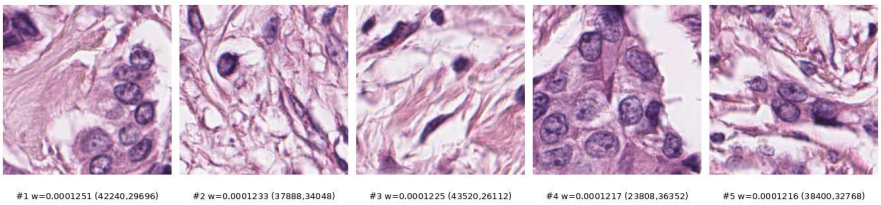


Fig. 4 Top-5 attention-ranked tiles extracted from a representative validation slide. Each tile corresponds to one of the highest-attention patches generated using the predicted-class prompt. Coordinates are shown for qualitative inspection of model behavior and do not represent ground-truth pathological annotations.

The results show noticeable seed-dependent variability, particularly for test accuracy and test AUC. This variability indicates that the classification behavior is sensitive to random initialization under the current weakly supervised training setting. Consequently, single-seed performance should be interpreted together with the multi-seed summary rather than as a complete characterization of model robustness. The multi-seed analysis revealed noticeable variability in classification performance across random initializations, with greater variability observed for some test-set metrics. This reinforces the importance of reporting aggregate performance across seeds rather than relying on a single initialization. Seed-stability experiments showed measurable variation in classification performance across random initializations, while attention-related behavior also remained sensitive to training stochasticity.

4.6 Qualitative Analyses

For selected validation slides (Fig. 4), we visualize the top-5 attention-ranked patches, each showing the raw 256×256 crop, its attention weight w_i, and corresponding coordinate (x_i, y_i). The diversity across tiles reflects natural heterogeneity in tissue appearance among regions receiving high model-assigned attention. We also compare heatmaps generated with $\tau = 0.5$ and $\tau = 2.0$ for the same slide (Fig. 5). Because attention is computed using $\text{softmax}(a_i/\tau)$, lower temperature values produce sharper and more selective attention distributions, whereas higher temperature values yield smoother and more diffuse attention patterns. Slide-level diagnostic panels (Fig. 6) consistently demonstrated spatial alignment and coordinate mapping across validation and test sets.

To quantify the sensitivity of the attention mechanism to class-level language prompts, we performed a prompt-swap analysis across all validation and test slides. Each slide was evaluated twice while keeping the visual input fixed and changing only the class prompt. Prompt sensitivity was quantified using Top-K Jaccard overlap and Spearman correlation between the complete patch-attention distributions.

Across validation slides, the median Top-50, Top-100, and Top-200 Jaccard overlaps were 0.7857, 0.7778, and 0.8182, respectively. The median Spearman correlation between the complete attention distributions was 0.9910. Similar behavior was observed on the test set, where the corresponding median Jaccard overlaps were 0.7544, 0.7857, and 0.8182, with a median Spearman correlation of 0.9904. These

Table 6 Seed stability for the selected configuration ($\tau = 1.0$, $\alpha = 1.0$) using class-wise prompt inference. Each checkpoint was selected using validation AUC, and a seed-specific decision threshold was selected on the validation set and applied unchanged to the test set.

	Validation			Test		
Seed	AUC	Acc	BAcc	AUC	Acc	BAcc
1	0.7812	0.700	0.8125	0.8906	0.500	0.6875
2	0.7812	0.700	0.8125	0.8125	0.800	0.7812
3	0.7656	0.650	0.7812	0.8594	0.850	0.9062
4	0.7344	0.700	0.8125	0.9219	0.850	0.9062
5	0.7031	0.700	0.7188	0.7344	0.700	0.7188
42	0.8750	0.900	0.8438	0.9062	0.850	0.8125
Mean $\pm$ SD	0.7734 ± 0.0583	0.7250 ± 0.0880	0.7969 ± 0.0431	0.8542 ± 0.0703	0.7583 ± 0.1393	0.8021 ± 0.0920

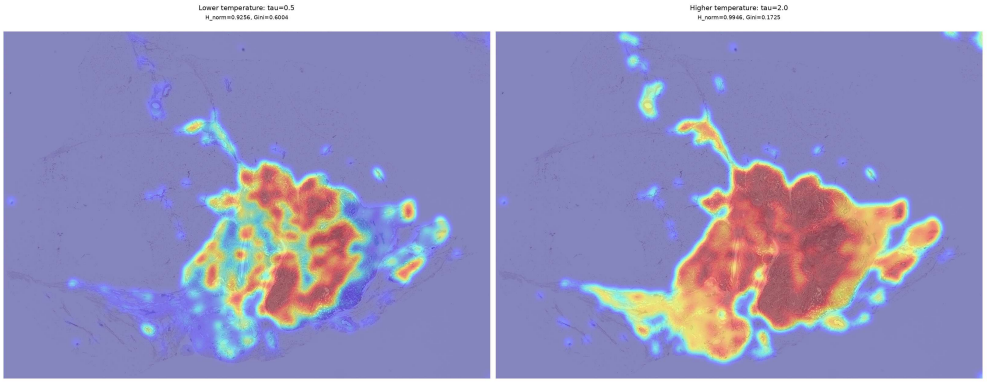


Fig. 5 Controlled attention-temperature comparison for the same validation slide using the same checkpoint and a fixed predicted-class prompt. Lower temperature ($\tau = 0.5$) produces a more concentrated attention distribution ($H_{\text{norm}} = 0.9256$, $G = 0.6004$), whereas higher temperature ($\tau = 2.0$) produces a more diffuse distribution ($H_{\text{norm}} = 0.9946$, $G = 0.1725$). The Top-200 patch ranking remains unchanged (Jaccard = 1.000), indicating that temperature primarily alters attention concentration rather than the ordering of attended regions.

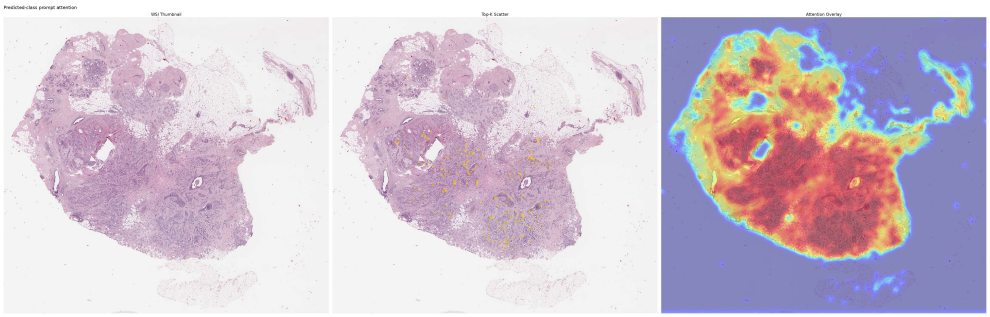


Fig. 6 Diagnostic visualization showing the WSI thumbnail, spatial distribution of the top-K attended patches, and the corresponding attention heatmap overlay generated using the predicted-class prompt.

results indicate that changing the class prompt produces measurable changes among the highest-ranked patches while preserving the overall ordering of patch attention scores to a large extent. The prompt-conditioned effect therefore appears to be localized rather than producing a global reorganization of the attention distribution. This interpretation is also consistent with the small Jensen–Shannon divergence observed between prompt-conditioned attention distributions.

4.7 Effect of Stop-Gradient Decoupling

To examine the effect of decoupling the classifier from the language-guided attention pathway, we compared the canonical stop-gradient model against an otherwise identical variant in which classification gradients were allowed to propagate through the bag representation.

Table 7 Prompt-swap sensitivity across validation and test slides. Values are reported as medians.

Split	Top-50 Jaccard	Top-100 Jaccard	Top-200 Jaccard	Spearman ρ
Validation	0.7857	0.7778	0.8182	0.9910
Test	0.7544	0.7857	0.8182	0.9904

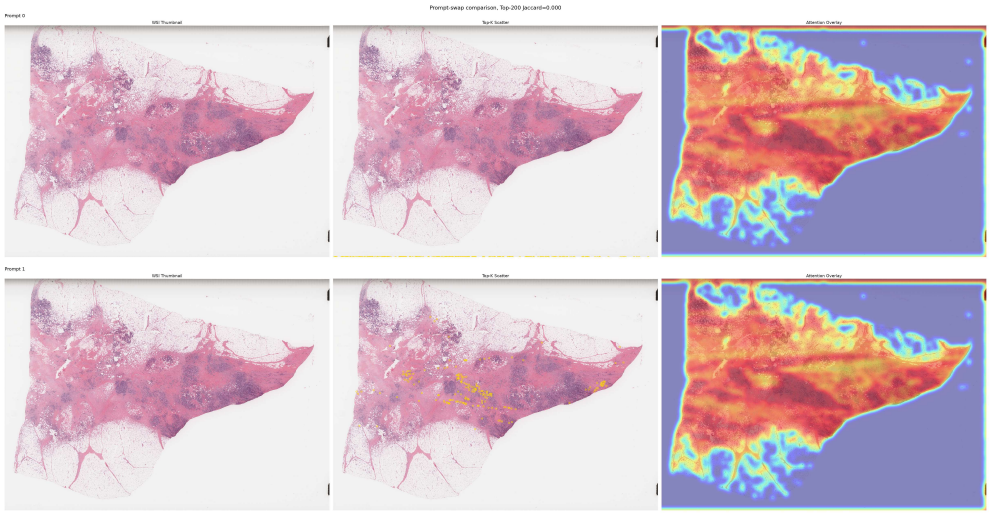


Fig. 7 Prompt-swap localization on the same whole-slide image. Top row: localization produced using the class-0 prompt; bottom row: localization produced using the class-1 prompt. The slide is held constant while only the prompt is changed. In this representative example, the Top-200 patch sets have a Jaccard index of 0.681. Full-dataset analysis showed median Top-200 Jaccard overlap of 0.8182 for both validation and test sets, indicating that prompt substitution modifies a subset of highly ranked regions while preserving much of the global attention ordering.

Both variants used the same patient-exclusive splits, random seed (42), attention temperature ($\tau = 1.0$), contrastive weight ($\alpha = 1.0$), contrastive temperature ($\tau_{\text{con}} = 0.05$), learning rate (10^{-4}), and 10-epoch training schedule. Model selection was based on validation AUC, and the classification threshold was selected on the validation set by maximizing balanced accuracy and then applied unchanged to the test set.

The stop-gradient configuration achieved higher validation AUC and higher validation accuracy than the no-stop-gradient variant under the controlled comparison. The same trend was also observed on the held-out test set. Based on validation performance, the stop-gradient configuration was retained as the canonical LG–MIL training design.

Table 8 Effect of stop-gradient decoupling on classification performance.

Variant	Val AUC	Val Acc	Val BAcc	Test AUC	Test Acc	Test BAcc
StopGrad	0.8750	0.9000	0.8438	0.9062	0.8500	0.8125
No StopGrad	0.7656	0.7000	0.8125	0.7969	0.7000	0.7188

5 Discussion

The current study proposes a language-guided multiple-instance learning (LG-MIL) approach to perform WSIs analysis using weak supervision. This model incorporates class-specific textual representations and image patches' visual representation using cross-modality attention, while a coordinate-retrofit mechanism enables mapping of attention weights to their corresponding slide locations. This process leads to predictions of slide labels, along with attention heatmaps, top-K tiles, and saliency values. In this part of the paper, the main behavioral attributes of the proposed model, the role of different hyperparameters, implications of the observed trends, and associated challenges will be considered. The proposed LG–MIL framework achieved strong slide-level classification performance across patient-exclusive validation and test sets. Although performance varied across hyperparameter settings and random seeds, competitive AUC values were obtained under several configurations. These results suggest that the pre-extracted patch features contain discriminative information for the present binary task. The strong classification performance shifts the emphasis of the investigation toward understanding how the attention distribution is learned under weak supervision. Correlation analysis revealed only weak relationships between model confidence and attention-dispersion metrics. The observed correlations between confidence, normalized entropy, and Gini coefficient were small and not statistically significant, suggesting that confidence and attention concentration behave largely independently in the current dataset. The importance of the coordinate-retrofit module was evident in its contribution to spatially aligned attention maps for validation and testing slides. The localization output produced consistent results in terms of alignment between attention locations and patch geometry. The three-panel diagnostic visualization facilitates visual inspection of coordinate consistency and can help identify obvious tiling or mapping errors. The attention temperature τ played an important role in controlling the sharpness of the attention distribution. Lower values of τ produced more selective attention maps, whereas higher values generated smoother and more diffuse attention distributions. The ablation study further showed that larger values of τ were associated with reduced classification performance compared with $\tau = 0.5$ and $\tau = 1.0$. The ablation study demonstrated that τ influenced both attention behavior and classification performance, whereas the influence of the contrastive weight α was comparatively limited under the evaluated settings. This provides users with a simple and intuitive mechanism for adjusting spatial selectivity in weakly supervised settings. The six-seed stability analysis revealed that slide-level performance remained reasonably stable across seeds, although both validation and test AUC exhibited measurable variability. Different random initializations may converge to functionally equivalent classifiers with

different internal attention configurations, reinforcing the value of reporting behavioral stability rather than relying solely on classification metrics when interpretability is emphasized.

There are a number of practical difficulties that are related to the integration of the text information, which can be overcome by using a number of preprocessing, domain adaptation, and evaluation mechanisms. Ethical issues related to patient privacy, handling, and demographic biases are also important factors that are related to the integration of text information, which can be overcome by the use of a number of mechanisms. Different limitations should be acknowledged, all relating to the model's weak supervision, dataset composition, or engineering choices, and none implying clinical interpretation. First, the evaluation dataset is constrained to slides for which coordinate-enabled features were available. This restriction ensures that localization outputs can be mapped to valid slide coordinates, but may limit dataset diversity, particularly for the validation and test splits. Second, the model achieves strong classification performance on the present dataset, which may limit the ability to fully analyze trade-offs between predictive performance and interpretability. In scenarios where classification difficulty is higher, the impacts of the attention temperature τ and the contrastive weight α on decision performance may surface more clearly. Third, the model's weak supervision inherently allows for a variety of valid attention configurations, even with sweep operations and variance calculation, as attention maps are still under-constrained with respect to slide-level labels. Fourth, the model uses class-level text prompts, which are manually defined, potentially leading to prompt dependence; however, using prompt ensembling, learnable prompts, or domain-specific language models could alleviate this dependence. Fifth, using a fixed top-K ($K = 200$) approach may introduce cross-slide comparability issues for slides with substantially different numbers of patches; adaptive-K approaches could potentially yield more consistent cross-slide comparisons. Sixth, as pre-computed patch features are used, dependence on upstream encoder models and staining variations is inherent; no encoder-level ablation or normalization was performed. Finally, although designed for model-behavior analysis and algorithmic interpretability, the system is for model analysis purposes only, with all visualizations (tiles, heatmaps, saliency curves) reflecting model signals and not for clinical use or interpretation.

A further limitation is the asymmetry between training and inference. During training, attention is conditioned on the prompt associated with the ground-truth slide label, whereas inference evaluates all class prompts independently. Although this design enables class-wise prompt-conditioned scoring at test time, future work should investigate class-wise training objectives in which competing class prompts are jointly evaluated during optimization.

The proposed research addresses a fundamental limitation in current medical imaging AI systems by reducing reliance on dense expert annotations while improving spatial interpretability. By anchoring the visual attention to class-level semantic prompts, the framework is intended to generate localization outputs that are conditioned on class-level semantic descriptions and can support qualitative inspection of model behavior and domain knowledge. This may facilitate more transparent inspection of AI model behavior. Additionally, the modular architecture may also be

applicable to other disease domains, although this requires validation on additional datasets and tasks. This increases the overall scope of the proposed architecture in the medical AI community. Overall, this study illustrates that language conditioning, weak supervision, and coordinate-aligned localization can be effectively integrated to create a flexible and reproducible method for model analysis. The system can perform extensive interpretability analysis in the absence of pixel-level annotations. Hence, this can create a more transparent evaluation process for a weakly supervised MIL pipeline.

6 Conclusion

The proposed work presents a language-guided multiple-instance learning framework for weakly supervised whole-slide image analysis and attention-based localization. The model incorporates language-informed attention to bridge between visual and semantic cues. The framework uses class-level text descriptions to guide attention allocation and support interpretable analysis of model behavior under weak supervision. This contributes toward the development of transparent and data-efficient medical image analysis methods and may provide a foundation for broader applications across diseases and multimodal datasets. By integrating class-prompt embeddings into cross-modal attention, the method produces slide-level predictions together with spatially grounded interpretability artifacts, including WSI heatmaps, top-K tiles, and quantitative saliency metrics. The coordinate-retrofitting component enables alignment of the attention weights with the corresponding (x, y) locations in the original whole-slide image, thereby supporting spatial visualization of model attention. On the evaluated dataset, the framework achieved strong slide-level classification performance across validation and test sets, as well as across multiple ablation and seed-stability experiments. The experimental results enabled analysis of both predictive performance and attention behavior under different training configurations. Ablation experiments demonstrated that the attention temperature τ had a greater influence on model behavior and predictive performance than the contrastive weight α, highlighting the importance of temperature-controlled attention distributions.

Seed-stability experiments showed that classification performance remained reasonably stable across random initializations, while saliency metrics exhibited greater variability, reflecting the under-constrained nature of attention in weakly supervised MIL. In summary, this research presents a reproducible framework for examining model behavior at a patch level with limited supervision based on a language guidance paradigm. This framework's modularity, packaging, and diagnostic capabilities enable a transparent investigation of attention mechanisms that can inform further research on prompt-robust models, multi-scale analysis, and related areas.

Statements and Declarations

Competing Interests

The authors declare that they have no competing interests.

Funding

The authors received no financial support for the research, authorship, or publication of this article.

Data Availability

Whole-slide images were obtained from the TCGA-BRCA cohort via the Genomic Data Commons. Patch-level features were extracted using precomputed UNI embeddings, a self-supervised foundation model for computational pathology. All results, including coordinate-aligned localization outputs and saliency metrics, were generated through fully scripted pipelines and can be reproduced from the original data sources, subject to TCGA access policies and the UNI research license.

Acknowledgments

The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Ethics statement

This study used publicly available, de-identified data from The Cancer Genome Atlas (TCGA)-BRCA project. No new human participants were recruited, and no identifiable patient information was accessed. Therefore, this study did not require additional ethical approval or informed consent from the authors' institution.

Author Contributions

Mohammad Mehedi Hassan: Conceptualization, Methodology, Software, Writing—Original Draft Preparation. Md. Rezaul Bashar: Investigation, Data Curation, Visualization, Supervision, Project Administration, Writing—Review & Editing.

References

[1] Rodríguez, J.H., Fraile, F.J.C., Conde, M.J.R., Llorente, P.L.G.: Computer aided detection and diagnosis in medical imaging: a review of clinical and educational applications. TEEM '16, pp. 517–524. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/3012430.3012567

[2] Halalli, B., Makandar, A.: Computer aided diagnosis - medical image analysis techniques. In: Kuzmiak, C.M. (ed.) Breast Imaging. IntechOpen, London (2017). Chap. 5. https://doi.org/10.5772/intechopen.69792

[3] Lee, S.J., Rho, M.: Multimodal deep learning applied to classify healthy and disease states of the human microbiome. Scientific Reports **12**(1), 824 (2022) https://doi.org/10.1038/s41598-022-04773-3

[4] Xu, J., Cao, R., Luo, P., Mu, D.: Active localization learning for weakly supervised instance segmentation. Expert Systems with Applications **276**, 126962 (2025) https://doi.org/10.1016/j.eswa.2025.126962

[5] Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.A.: Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Transactions on Cybernetics **50**(9), 3950–3962 (2020) https://doi.org/10.1109/TCYB.2019.2935141

[6] Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Scl-wc: cross-slide contrastive learning for weakly-supervised whole-slide image classification. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Curran Associates Inc., Red Hook, NY, USA (2022). https://doi.org/10.52202/068431-1309

[7] Barbosa, D., Ferreira, M., Junior, G.B., Salgado, M., Cunha, A.: Multiple instance learning in medical images: A systematic review. IEEE Access **12**, 78409–78422 (2024)

[8] Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4055–4064 (2023). IEEE

[9] Tavolara, T.E., Gurcan, M.N., Niazi, M.K.K.: Contrastive multiple instance learning: An unsupervised framework for learning slide-level representations of whole slide histopathology images without labels. Cancers **14**(23) (2022) https://doi.org/10.3390/cancers14235778

[10] Peled, S., Maruvka, Y.E., Freiman, M.: Psa-mil: A probabilistic spatial attention-based multiple instance learning for whole slide image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1211–1220 (2026)

[11] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**, 555–570 (2021) https://doi.org/10.1038/s41551-020-00682-w

[12] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16123–16134 (2022). https://doi.org/10.1109/CVPR52688.2022.01567

[13] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference

on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136 (2018). https://proceedings.mlr.press/v80/ilse18a.html

[14] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

[16] Almubarak, H.A.: Self-supervised learning for data augmentation in histopathology image segmentation. Scientific Reports **15**(1), 39596 (2025) https://doi.org/10.1038/s41598-025-23280-9

[17] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning for medical imaging. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA (2019)

[18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021). https://proceedings.mlr.press/v139/radford21a.html

[19] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Curran Associates Inc., Red Hook, NY, USA (2022). https://doi.org/10.52202/068431-1723

[20] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR, ??? (2022). https://proceedings.mlr.press/v162/li22n.html

[21] Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3922–3931 (2021)

[22] Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision–language processing. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, pp. 1–21. Springer, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20059-5_1 . https://doi.org/10.1007/978-3-031-20059-5_1

[23] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole-slide images. Nature Medicine **25**, 1301–1309 (2019) https://doi.org/10.1038/s41591-019-0508-1

[24] Jabin, A., Kirar, J.S., Ahmad, S.: Ai-based methods for modelling whole-slide imaging data in cancer diagnosis and transcriptome profile prediction. BMC Artificial Intelligence **1**, 16 (2025) https://doi.org/10.1186/s44398-025-00018-1

[25] Sharma, A., Liu, B., Rantalainen, M.: A method for spatial interpretation of weakly supervised deep learning models in computational pathology. Scientific Reports **15**, 19804 (2025) https://doi.org/10.1038/s41598-025-04043-y

[26] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., *et al.*: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**, 850–862 (2024) https://doi.org/10.1038/s41591-024-02857-3

[27] Lingle, W., Erickson, B.J., Zuley, M.L., Jarosz, R., Bonaccio, E., Filippini, J., Net, J.M., Levi, L., Morris, E.A., Figler, G.G., Elnajjar, P., Kirk, S., Lee, Y., Giger, M., Gruszauskas, N.: The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA). The Cancer Imaging Archive (TCIA). Dataset (2016). https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP