SEMANTIC PRIOR-GUIDED REINFORCED REGION REVIEWING FOR WHOLE SLIDE IMAGE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole slide image (WSI) analysis is challenged by the sparsity of diagnostically relevant lesions among numerous irrelevant patches, which can dilute discriminative evidence and disperse visual attention. Vision-language models (VLMs) provide pathology semantic prior for this task. However, directly fusing such information with visual features may be unreliable because LLM-generated descriptions can be coarse, incomplete, or noisy. To address these challenges, we propose the Prior-Guided Semantic Multi-Scale Hypergraph Learning Model (PSMG), which uses LLM-generated descriptions as semantic priors for region-level WSI modeling. In the visual architecture, we first perform scale-specific hypergraph encoding to capture high-order tissue context at each magnification. Based on this, we design a semantic-guided region modeling strategy that translates textual pathology priors into region-level visual modulation. This strategy uses textual priors to highlight semantically relevant candidate regions for subsequent visual reasoning. Subsequently, we formulate a slide-level feedback mechanism to review semantically guided regions. Through task-driven region refinement, the feedback mechanism further enhances regions that contribute to prediction. Finally, we introduce a cross-magnification alignment strategy that combines unified hypergraph reasoning with knowledge distillation to reduce diagnostic inconsistency caused by scale-specific representation differences. Experiments on three public WSI benchmarks demonstrate strong classification performance, while qualitative heatmaps show more concentrated responses around annotated lesion regions. The anonymized implementation is available at `https://anonymous.4open.science/r/psmg-anonymous`.

1 INTRODUCTION

Histopathological examination is the clinical gold standard for cancer diagnosis, and whole slide images (WSIs) provide gigapixel-level digital representations of tissue morphology (Campanella et al., 2019; Xu et al., 2024). However, diagnostically relevant lesion regions are often sparse among numerous irrelevant patches, which can dilute discriminative evidence and pose a major challenge for effective WSI analysis.

Existing WSI methods mainly improve visual representation learning and region aggregation. MIL-based approaches learn patch importance from slide-level supervision (Lu et al., 2021; Li et al., 2021; Zhang et al., 2024; Tang et al., 2024), while graph and hypergraph methods further model spatial and higher-order tissue relations (Hou et al., 2022; Zheng et al., 2022; Chan et al., 2023; Di et al., 2022; Shi et al., 2024b). Multi-magnification methods additionally combine global tissue organization with fine-grained morphology (Hou et al., 2022; Shi et al., 2024a). These advances improve contextual modeling, but region importance is still inferred mainly from visual features and slide labels. As a result, the model may attend to visually salient yet diagnostically weak regions. Meanwhile, differences in representation granularity across magnifications can also lead to inconsistent scale-specific evidence.

Recently, vision-language models (VLMs) and large language models (LLMs) have introduced pathology-related textual knowledge into WSI analysis (Lu et al., 2024; Huang et al., 2023; Lu et al., 2023; Shi et al., 2024a). Despite this progress, two limitations remain. First, representative methods incorporate textual representations into prediction through vision–language alignment

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

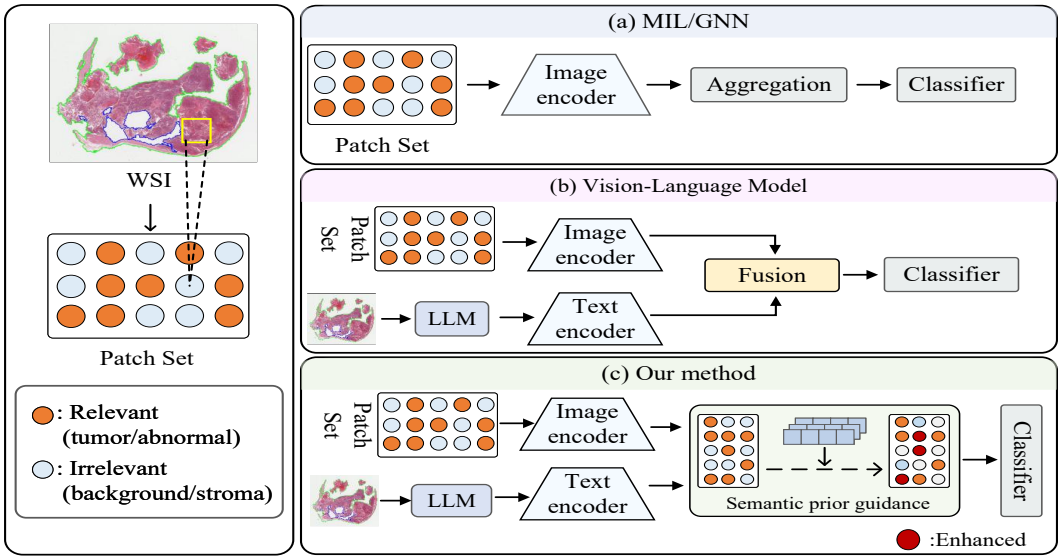


Figure 1: Comparison of WSI Classification Paradigms. (a) MIL/GNN-based visual modeling. (b) Direct vision-language fusion. (c) Our semantic prior-guided visual region modeling.

or fusion (Lu et al., 2023; Shi et al., 2024a). However, LLM-generated descriptions are typically global and coarse, whereas diagnostic evidence in WSIs is sparse and localized. Directly using such descriptions for prediction may introduce irrelevant or incomplete semantic cues (Asgari et al., 2025). Second, recent language-guided methods prioritize regions mainly through text-conditioned relevance or cross-modal attention (Li et al., 2024a; Guo et al., 2025). However, semantic relevance does not necessarily indicate task discriminability. Moreover, the value of a candidate region may depend on which regions have already been selected. We therefore treat region reviewing as a sequential subset selection problem and evaluate the reviewed subset using slide-level task feedback. To the best of our knowledge, we are the first in WSI analysis to apply such task-feedback-driven reviewing to candidate regions identified by slide-specific semantic priors.

To address the above issues, we propose the Prior-Guided Semantic Multi-Scale Hypergraph Learning Model (PSMG). PSMG uses LLM-generated descriptions as semantic priors for region-level visual modeling, reducing the risk that noisy textual descriptions dominate the prediction. First, to capture high-order tissue context, we encode scale-specific hypergraphs. Second, to avoid propagating coarse global text through direct vision–language fusion and improve cross-magnification semantic guidance, we develop Semantic-Guided Region Modeling (SGRM). It uses LLM-generated descriptions as region-level priors, modulates relevant low-magnification regions, and links them to corresponding high-magnification child patches through parent–child mapping, enabling coarse-to-fine guidance. Then, we design Reinforced Semantic Reviewing (RSR) to further evaluate the task utility of semantically guided candidates. RSR sequentially selects candidate regions conditioned on the current reviewed subset and uses slide-level classification feedback to optimize the resulting selection policy. Finally, to alleviate cross-magnification inconsistency caused by differences in representation granularity and attended tissue regions, we design Cross-Magnification Hypergraph Alignment (CMHA). It constructs a unified hypergraph over the refined multi-magnification regions to model high-order cross-magnification interactions and aligns scale-specific predictions through knowledge distillation. The main contributions of this work are summarized as follows:

- We propose a coarse-to-fine semantic guidance strategy. It uses slide-specific LLM descriptions as region-level priors and associates semantically relevant low-magnification regions with their high-magnification child patches.
- We propose a task-feedback-driven reinforced reviewing strategy. It formulates candidate-region reviewing as sequential subset selection and optimizes the policy using slide-level classification feedback.

- We propose a cross-magnification hypergraph alignment strategy. It models high-order interactions across scales and performs asymmetric prediction alignment using a stop-gradient high-magnification target.

2 Related Work

2.1 Visual WSI Modeling

Weakly supervised WSI classification is commonly formulated as Multiple Instance Learning (MIL), where each slide is treated as a bag of patch instances with only slide-level labels. Attention-based MIL improves bag aggregation by learning instance-level importance scores (Ilse et al., 2018; Lu et al., 2021), while later methods further enhance feature aggregation or reduce unreliable attention through hard instance mining and causal intervention (Li et al., 2021; Zhang et al., 2022; Tang et al., 2023; Lin et al., 2023). However, these methods mainly infer region importance from visual features and slide-level supervision, without explicit pathological semantic guidance. Graph-based methods further model tissue context by constructing spatial or feature-space graphs among patches (Chen et al., 2021; Chan et al., 2023; Zheng et al., 2022; Li et al., 2024b; Lu et al., 2022). For multi-scale WSI analysis, HIPT (Chen et al., 2022) and H^2-MIL (Hou et al., 2022) explore hierarchical or heterogeneous representations. Hypergraph neural networks (Feng et al., 2019) can further model higher-order relationships among multiple correlated regions. Moreover, these visual models still lack explicit pathological semantic guidance, making them susceptible to irrelevant or non-lesion tissue regions.

2.2 Vision-Language Models in Pathology

Vision-language models have recently introduced textual knowledge into computational pathology. Early pathology-specific VLMs, such as PLIP (Huang et al., 2023) and CONCH (Lu et al., 2024), learn image-text aligned representations from large-scale pathology data, enabling transferable visual-semantic embeddings. Moreover, Quilt-1M (Ikezogwo et al., 2023) and Quilt-LLaVA provide large-scale image-text pairs and generated pathology descriptions for visual-language learning. More recently, ViLa-MIL (Shi et al., 2024a) incorporates generated descriptions into dual-scale WSI classification through vision-language fusion. These studies demonstrate the effectiveness of textual knowledge in pathology analysis. However, LLM-generated pathology descriptions may be coarse-grained, incomplete, or hallucinated, and may not precisely localize the visual evidence required for slide-level prediction (Asgari et al., 2025; Zhao et al., 2025).

2.3 Reinforcement Learning for WSI Analysis

Reinforcement learning has been explored in WSI analysis to enable sequential and task-driven region selection. RLogist learns a multi-resolution observation policy that navigates toward informative regions, reducing the need for exhaustive high-magnification processing (Zhao et al., 2023). Beyond optimizing observation paths, MuRCL uses online rewards to progressively update discriminative feature sets for more effective slide-level aggregation (Zhu et al., 2022). PAMIL further integrates policy learning with dynamic instance sampling, allowing previously observed instances and classification feedback to guide subsequent selection and representation learning (Zheng et al., 2024). Despite their different objectives, these methods learn selection policies primarily from visual instances and task rewards, without examining whether regions identified by slide-specific textual semantics are truly discriminative.

3 Method

3.1 Overview

As illustrated in Fig. 2, PSMG consists of scale-specific hypergraph encoding, Semantic-Guided Region Modeling (SGRM), Reinforced Semantic Reviewing (RSR), and Cross-Magnification Hypergraph Alignment (CMHA). SGRM grounds slide-specific textual priors to multi-scale regions, RSR further refines these regions using task feedback, and CMHA coordinates cross-scale diagnostic evidence for final prediction.

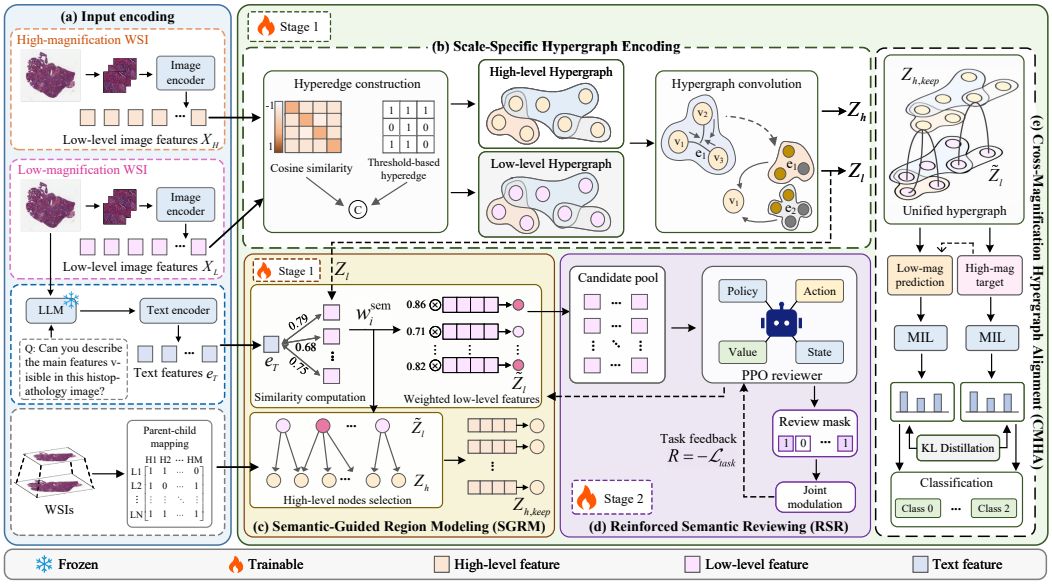


Figure 2: Overview of the proposed PSMG framework. (a) Input encoding. (b) Scale-specific hypergraph encoding for high-order tissue context modeling. (c) SGRM for text-guided feature modulation and parent–child-based high-magnification node association. (d) RSR uses a PPO-based reviewer to refine semantically enhanced regions with task feedback. (e) CMHA for unified cross-magnification hypergraph reasoning, prediction alignment, and classification.

3.2 Input Encoding

Given a WSI $\mathcal{W}$, we extract non-overlapping patches from low- and high-magnification views, denoted as $\mathcal{P}_l = \{\mathbf{p}_{l,i}\}_{i=1}^{N_l}$ and $\mathcal{P}_h = \{\mathbf{p}_{h,j}\}_{j=1}^{N_h}$. A pretrained DINO (Caron et al., 2021) encoder $\Phi_I(\cdot)$ extracts patch features:

$$\mathbf{x}_{l,i} = \Phi_I(\mathbf{p}_{l,i}), \quad \mathbf{x}_{h,j} = \Phi_I(\mathbf{p}_{h,j}), \tag{1}$$

which are stacked as $\mathbf{X}_l \in \mathbb{R}^{N_l \times d}$ and $\mathbf{X}_h \in \mathbb{R}^{N_h \times d}$. Since each high-magnification patch is spatially located within a low-magnification region, we define a parent–child mapping $\pi(j) = i$, indicating that $\mathbf{p}_{h,j}$ belongs to $\mathbf{p}_{l,i}$.

For each $\mathcal{W}_l$, a pathology-oriented description is generated offline by Quilt-LLaVA (Seyfioglu et al., 2024) using the fixed prompt: "Can you describe the main features visible in this histopathology image?". We adopt greedy decoding with a temperature of zero to reduce generation randomness, and the generated descriptions remain fixed throughout training and inference. Each description, denoted as $\mathcal{T}$, is tokenized and truncated to a maximum length of 77 tokens before being encoded by the QuiltNet text encoder $\Phi_T(\cdot)$:

$$\mathbf{e}_T = \Phi_T(\mathcal{T}), \quad \mathbf{e}_T \in \mathbb{R}^{d_T}. \tag{2}$$

The resulting textual embedding is used exclusively as a region-level semantic prior and is not directly fed into the final classifier.

3.3 Scale-Specific Hypergraph Encoding

For each scale $s \in \{l, h\}$, we construct a hypergraph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$. The hyperedge centered at $v_{s,i}$ combines distance-based and k-nearest neighbors:

$$e_{s,i} = \{v_{s,j} \mid \tilde{d}(v_{s,i}, v_{s,j}) < \theta\} \cup \mathcal{N}_k(v_{s,i}), \tag{3}$$

where $\tilde{d}(\cdot,\cdot)$ is the normalized feature distance. The incidence matrix is defined by $\mathbf{H}_s(i,m) = \mathbb{I}(v_{s,i} \in e_{s,m})$.

The hypergraph convolution is

$$\mathrm{HGConv}_{\Theta}(\mathbf{X}, \mathcal{E}) = \sigma\big(\mathbf{H}\Psi(\mathbf{H}^{\top}\mathbf{X}\mathbf{W}_v)\mathbf{W}_e + \mathbf{X}\big), \tag{4}$$

where $\Theta = \{\mathbf{W}_v, \mathbf{W}_e\}$ and $\Psi(\cdot)$ denotes stochastic suppression in the hyperedge space. The resulting representations are $\mathbf{Z}_s = \mathrm{HGConv}_{\Theta_s}(\mathbf{X}_s, \mathcal{E}_s)$, $s \in \{l, h\}$.

3.4 Semantic-Guided Region Modeling (SGRM)

To avoid the potential noise and instability introduced by directly fusing LLM-generated text with visual features, SGRM uses textual descriptions as semantic priors for region modeling. This module aims to identify semantically relevant low-magnification tissue regions and guide the subsequent high-magnification region association in a coarse-to-fine manner.

The textual description of each WSI is first encoded into a semantic embedding $\mathbf{e}_{\mathcal{T}}$ by the text encoder. The embedding is then projected into the visual feature space:

$$\tilde{\mathbf{e}}_{\mathcal{T}} = \mathbf{W}_T \mathbf{e}_{\mathcal{T}}, \tag{5}$$

where $\mathbf{W}_T$ denotes a learnable projection matrix. For each low-magnification region representation $\mathbf{z}_{l,i}$, its semantic relevance to the textual prior is measured by cosine similarity:

$$s_i = \cos(\tilde{\mathbf{e}}_{\mathcal{T}},\ \mathbf{z}_{l,i}). \tag{6}$$

To obtain stable semantic modulation weights, the similarity scores are normalized within each WSI and then passed through a sigmoid function:

$$w_i^{\text{sem}} = \sigma\left(\alpha \cdot \frac{s_i - \mu_s}{\sigma_s + \epsilon}\right), \tag{7}$$

where μ_s and σ_s denote the mean and standard deviation of similarity scores across low-magnification regions. The parameter α controls the sharpness of semantic weighting. The resulting weight is applied to the low-magnification representation through residual modulation:

$$\bar{\mathbf{z}}_{l,i} = \mathbf{z}_{l,i} \cdot (1 + \beta_{\text{sem}} w_i^{\text{sem}}), \tag{8}$$

where β_{sem} controls the modulation strength.

Based on s_i, the top-K_l low-magnification regions are retained as semantic anchors:

$$\begin{aligned} K_l &= \max(1, \mathrm{round}\,(\eta_l |\mathcal{V}_{l,\text{valid}}|)), \\ \mathcal{V}_{l,\text{sem}} &= \mathrm{TopKIdx}\left(\{s_i\}_{i \in \mathcal{V}_{l,\text{valid}}}, K_l\right), \end{aligned} \tag{9}$$

where η_l denotes the low-magnification retention ratio. Their high-magnification child nodes are retrieved through the parent–child mapping:

$$\mathcal{V}_{h,\text{cand}} = \{j \in \mathcal{V}_h \mid \pi(j) \in \mathcal{V}_{l,\text{sem}}\}. \tag{10}$$

Finally, the selected multi-magnification representations are

$$\begin{aligned} \tilde{\mathbf{Z}}_l &= \{\bar{\mathbf{z}}_{l,i} \mid i \in \mathcal{V}_{l,\text{sem}}\}, \\ \mathbf{Z}_{h,\text{keep}} &= \{\mathbf{z}_{h,j} \mid j \in \mathcal{V}_{h,\text{cand}}\}. \end{aligned} \tag{11}$$

3.5 Reinforced Semantic Reviewing (RSR)

To refine semantically guided regions with task feedback, we formulate RSR as a Markov decision process (MDP) that sequentially selects a small subset of low-magnification regions for additional enhancement.

We first restrict the action space to the top-M semantic anchors:

$$\mathcal{C} = \text{Top-M}\left(\{s_i\}_{i \in \mathcal{V}_{l,\text{sem}}}\right), \qquad M \le |\mathcal{V}_{l,\text{sem}}|. \tag{12}$$

At step t, the state $S_t = \{o_i^t\}_{i \in \mathcal{C}}$ is composed of

$$o_i^t = [z_{l,i},\ s_i,\ w_i^{\text{sem}},\ m_i^t,\ \tau_t], \qquad i \in \mathcal{C}, \tag{13}$$

where m_i^t is the selection mask and $\tau_t = t/\max(T-1,1)$ is the normalized step index. The policy selects an unreviewed candidate at each step,

$$a_t \sim \pi_\psi(a_t \mid S_t), \qquad a_t \in \mathcal{C} \setminus \mathcal{A}_{<t}, \tag{14}$$

for

$$T = \min(\max(1, \mathrm{round}(\rho|\mathcal{C}|)), |\mathcal{C}|, T_{\max}) \tag{15}$$

steps, where previously selected candidates are masked during sampling. Repeated selections are avoided by masking already selected regions during policy sampling.

After T reviewing steps, let $r_i = \mathbb{I}(i \in \mathcal{A}_{1:T})$ indicate whether a valid low-magnification region i is selected. Its final representation is

$$\hat{\mathbf{z}}_{l,i} = \mathbf{z}_{l,i}\left(1 + \beta_{\mathrm{sem}} w_i^{\mathrm{sem}} + \beta_{\mathrm{rl}} r_i\right), \quad i \in \mathcal{V}_{l,\mathrm{sem}}, \tag{16}$$

where β_{sem} and β_{rl} control semantic modulation and reviewing enhancement, respectively. Accordingly, the reviewed low-magnification representation is $\tilde{\mathbf{Z}}_l^{rev} = \{\hat{\mathbf{z}}_{l,i} \mid i \in \mathcal{V}_{l,\mathrm{sem}}\}$, while the retained high-magnification representation $\mathbf{Z}_{h,\mathrm{keep}}^{rev}$ remains unchanged.

3.6 Cross-Magnification Hypergraph Alignment (CMHA)

To reduce prediction inconsistency across magnifications, CMHA performs asymmetric cross-magnification prediction alignment. The high-magnification prediction serves as a stop-gradient target, while the low-magnification prediction is optimized to match it. This forms an online distillation process within joint multi-scale training. It operates on the SGRM representations in Stage 1 and their RSR-refined counterparts in Stage 2:

$$Z_u^{(0)} = [Z_l^*; Z_{h,\mathrm{keep}}], \tag{17}$$

where $Z_l^* = \tilde{Z}_l$ in Stage 1 and $Z_l^* = \tilde{Z}_l^{\mathrm{rev}}$ in Stage 2.

A unified hypergraph with hyperedge set $\mathcal{E}_u$ is dynamically constructed to obtain

$$Z_u = \mathrm{HGConv}_{\Theta_u}(Z_u^{(0)}, \mathcal{E}_u). \tag{18}$$

We split Z_u by magnification and obtain scale-specific predictions:

$$p_l = \mathrm{MIL}_l(Z_{u,l}), \qquad p_h = \mathrm{MIL}_h(Z_{u,h}). \tag{19}$$

The scale-specific classification losses and cross-scale alignment loss are

$$\mathcal{L}_{\mathrm{cls}}^s = \ell_{\mathrm{cls}}(p_s, y), \qquad s \in \{l, h\}, \tag{20}$$

$$\mathcal{L}_{\mathrm{ka}} = D_{\mathrm{KL}}\left(\mathrm{sg}(p_{h,\tau}) \parallel p_{l,\tau}\right), \tag{21}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. The high-magnification prediction is used as a fixed target within the alignment loss, and gradients are propagated only through the low-magnification prediction for this term. The task objective used in both stages is

$$\mathcal{L}_{\mathrm{task}} = \lambda_l \mathcal{L}_{\mathrm{cls}}^l + \lambda_h \mathcal{L}_{\mathrm{cls}}^h + \gamma \mathcal{L}_{\mathrm{ka}}. \tag{22}$$

3.7 Training and Inference

PSMG is trained in two stages. Stage 1 optimizes the scale-specific hypergraph encoder, SGRM and CMHA by minimizing $\mathcal{L}_{\mathrm{task}}$. In Stage 2, the scale-specific hypergraph encoders, SGRM, and CMHA are frozen, and only RSR is optimized with PPO using the terminal reward $R = -\mathcal{L}_{\mathrm{task}}$, computed from the post-review predictions on training slides. Validation and test labels are never used for policy optimization. During inference, the learned RSR policy selects regions solely based on the observed state, and the high-magnification MIL head $\mathbf{p}_h$ produces the slide-level prediction.

4 Experiments

4.1 Datasets

We evaluate our method on three public datasets, including Camelyon16 (Bejnordi et al., 2017), TCGA_Lung (Weinstein et al., 2013), and BRACS (Brancati et al., 2022). For all datasets, we conduct five-fold cross-validation and report Accuracy (ACC), Area Under the ROC Curve (AUC), and F1-score as mean±std. Detailed dataset information is provided in the supplementary materials.

Table 1: Comparison (presented in %) with representative WSI analysis methods on all datasets under five-fold cross-validation.

Method	Camelyon16			TCGA_Lung			BRACS		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Mean pooling	76.6±2.1	75.5±2.6	68.2±2.7	85.2±2.3	91.8±1.9	84.1±2.4	67.5±2.1	74.1±1.9	58.3±2.8
Max pooling	82.0±3.8	85.4±2.1	81.7±2.6	87.0±2.3	91.5±1.9	84.5±2.3	65.1±2.4	71.8±2.2	54.2±3.1
ABMIL	85.4±2.7	84.2±2.2	77.3±2.7	88.4±2.6	84.0±1.4	77.4±2.0	74.8±2.6	87.5±1.8	68.2±3.1
CLAM_SB	85.4±3.2	89.9±2.9	80.4±3.6	89.1±2.9	90.7±2.1	87.3±1.7	77.3±2.4	89.5±1.4	71.1±3.5
CLAM_MB	86.6±2.8	90.2±2.0	81.5±3.1	88.0±1.8	91.1±3.0	88.4±1.8	75.1±2.4	89.5±1.6	69.9±3.6
I^2HGC	81.6±2.5	92.3±2.6	81.1±2.1	88.7±2.6	95.8±1.7	88.7±2.7	74.3±2.9	89.2±2.3	70.5±1.8
HDMIL	94.9±2.6	**98.7±1.7**	94.7±2.6	89.2±2.9	96.2±1.2	89.1±2.9	76.6±2.3	88.3±2.9	70.6±2.4
GRASP	92.5±2.2	96.9±1.5	91.8±2.6	90.1±1.6	95.5±1.4	85.8±1.9	79.2±2.1	88.7±3.9	68.7±2.2
ViLa-MIL	88.2±3.9	95.1±1.7	87.7±3.8	85.4±2.1	89.5±1.8	84.8±2.2	75.1±3.5	87.6±3.9	68.6±3.0
FOCUS	88.9±2.3	95.6±3.3	88.3±2.6	88.7±3.3	94.8±2.0	88.6±3.1	73.3±2.0	86.7±3.1	64.0±2.8
DyKo	92.7±1.5	97.9±3.8	92.3±1.5	87.1±2.9	94.3±1.4	87.0±3.0	76.2±2.8	88.1±3.5	67.4±2.9
PSMG	**97.8±1.1**	97.9±1.2	**96.0±2.1**	**94.1±1.5**	**97.2±1.0**	**92.8±2.6**	**83.1±1.2**	**91.2±1.7**	**77.6±2.5**

Table 2: Ablation results (presented in %) of structural modeling and cross-magnification knowledge alignment.

Model	$\mathcal{L}_{ka}$	Camelyon16			TCGA_Lung			BRACS		
		ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
GCN	✓	88.7±2.1	91.2±1.5	81.4±3.1	82.1±2.6	89.5±1.8	76.1±3.3	72.2±2.4	84.6±2.2	63.5±3.4
GAT	✓	90.5±1.8	92.8±1.2	83.6±2.8	83.6±2.0	90.7±1.3	77.8±2.9	74.3±2.1	86.5±1.7	66.2±3.1
Hypergraph	×	92.1±2.4	94.1±1.0	85.2±3.6	84.9±2.2	91.4±1.5	79.2±3.5	76.4±1.9	88.6±2.0	68.8±2.9
PSMG-Stage 1	✓	**95.9±1.5**	**96.6±0.6**	**93.0±2.7**	**89.8±2.2**	**94.6±1.4**	**88.5±2.3**	**81.2±2.2**	**90.1±1.1**	**75.1±2.8**

4.2 Implementation

We use frozen DINO features extracted from 256×256 patches at 10× and 20× magnifications. PSMG is optimized with Adam under patient-level five-fold cross-validation. Unless otherwise stated, we set $\beta_{\text{sem}} = 0.2$, $\tau = 1.5$, $\gamma = 1$, and use a 5% reviewing ratio over the top-256 semantic candidates. Complete optimization, preprocessing, and PPO settings are provided in the Supplementary Material.

4.3 Paradigm Comparison

We compare PSMG with state-of-the-art WSI analysis methods, including Max-pooling, Mean-pooling, ABMIL (Ilse et al., 2018), CLAM (Lu et al., 2021), I^2HGC (Han et al., 2025), HDMIL (Dong et al., 2025), GRASP (Mirabadi et al., 2025), ViLa-MIL (Shi et al., 2024a), FOCUS (Guo et al., 2025), and DyKo (Li et al., 2026). For fair comparison, all methods are evaluated under identical data splits and evaluation metrics. Methods originally designed for other downstream tasks are adapted by modifying only the task-specific prediction head and loss function, while keeping their core modeling components unchanged. For all reimplemented baselines, we replace their original patch representations with exactly the same frozen DINO features. Detailed baseline settings are provided in the supplementary material.

As shown in Table 1, PSMG consistently outperforms pooling and conventional MIL baselines, and achieves clear gains over most graph-, hypergraph-, and vision–language-based methods. These results support the effectiveness of combining high-order structural modeling, semantic region guidance, and task-feedback-driven refinement. On Camelyon16, HDMIL achieves a slightly higher AUC, likely benefiting from its hierarchical patch pre-screening for localized metastatic evidence.

4.4 Ablation Studies

4.4.1 Structural Modeling and Cross-Scale Alignment

We first examine whether higher-order structural modeling provides more informative visual representations for subsequent semantic guidance. As shown in Table 2, the hypergraph variant outperforms the GCN and GAT variants across the three datasets, suggesting that higher-order model-

Table 3: SGRM ablation results (presented in %) on three WSI classification benchmarks.

Variant	Camelyon16			TCGA_Lung			BRACS		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
w/o Text	87.0±9.8	89.0±10.1	84.3±5.1	87.4±2.3	92.6±1.9	81.1±5.8	75.1±3.5	83.5±2.4	61.2±4.1
Direct Fusion	86.4±7.6	89.2±8.8	83.1±6.2	87.6±2.1	92.3±1.5	82.0±4.9	74.8±2.9	84.1±2.0	62.3±3.7
Low-only	91.2±3.4	92.0±2.5	88.4±4.0	87.9±1.8	91.8±1.4	82.6±3.1	75.8±2.7	86.2±1.8	65.5±3.2
Weight-Prop	92.8±2.9	93.4±1.8	90.1±3.5	88.1±1.6	92.4±1.2	83.9±2.6	76.5±2.4	87.7±2.1	67.1±2.8
Text-Select	94.2±2.2	94.1±4.0	91.9±3.2	88.4±1.3	92.8±1.1	84.7±2.3	75.9±2.5	88.2±1.6	68.0±3.3
PSMG-Stage 1	**95.9±1.5**	**96.6±0.6**	**93.0±2.7**	**89.8±2.2**	**94.6±1.4**	**88.5±2.3**	**81.2±2.2**	**90.1±1.1**	**75.1±2.8**

Table 4: RSR ablation results on three WSI classification benchmarks. Values are reported in percentage (%) as mean±std.

Variant	Camelyon16			TCGA_Lung			BRACS		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
PSMG-Stage 1	95.9±1.5	96.6±0.6	93.0±2.7	89.8±2.2	94.6±1.4	88.5±2.3	81.2±2.2	90.1±1.1	75.1±2.8
Random Review	96.1±1.2	96.8±0.8	93.3±2.1	90.1±1.9	94.9±1.1	89.0±2.0	81.5±2.0	90.3±1.0	75.5±2.5
TopK	96.8±1.4	97.3±0.9	94.6±2.4	91.2±1.6	95.6±1.2	90.2±2.2	82.1±1.9	90.6±0.9	76.2±2.4
MLP-Gate	97.0±1.2	97.4±0.8	94.9±2.1	92.4±1.5	96.3±0.9	91.2±2.0	82.6±1.4	90.9±1.0	77.0±2.2
Vis-PPO	96.5±1.1	97.1±0.7	94.1±1.9	92.0±1.4	96.1±1.0	90.9±1.8	82.4±1.5	90.8±0.8	76.8±2.1
PSMG-Stage 2	**97.8±1.1**	**97.9±1.2**	**96.0±2.1**	**94.1±1.5**	**97.2±1.0**	**92.8±2.6**	**83.1±1.2**	**91.2±1.7**	**77.6±2.5**

ing captures tissue dependencies beyond pairwise relations. PSMG-stage 1 achieves the strongest overall results among the evaluated configurations, supporting the use of multi-scale hypergraph reasoning and cross-scale alignment for visual representation learning.

4.4.2 Semantic Prior Modeling

Table 3 compares several region-guidance strategies. The lower performance of w/o Text suggests that generated descriptions provide complementary semantic information. However, Direct Fusion remains inferior to full SGRM, indicating that global textual descriptions are more effective as region-level guidance than as direct classification features under the evaluated settings. The remaining variants further show that soft low-magnification modulation combined with parent–child association provides a stronger coarse-to-fine guidance strategy than direct propagation or hard selection. Additional experiments under corrupted and mismatched textual priors are provided in the supplementary material, further examining the dependence of PSMG on generated descriptions.

4.4.3 Task-Driven Semantic Reviewing

As shown in Table 4, PSMG-Stage 2 consistently improves over PSMG-Stage 1, showing that task feedback provides complementary information beyond semantic guidance. Random Review yields only marginal gains, indicating that arbitrary enhancement is ineffective. TopK performs better by favoring semantically relevant candidates, but remains below the full reviewer. MLP-Gate further improves performance by learning task-adaptive region scores, showing that trainable candidate weighting is beneficial. However, it still treats candidates independently and does not model the evolving reviewed subset. Vis-PPO introduces sequential selection but lacks semantic guidance in the candidate evaluation. In contrast, the full RSR combines semantic priors with sequential task-feedback-driven reviewing, leading to the most consistent improvements across datasets.

4.5 Visualization

As shown in Fig. 3(a), PSMG produces more compact intra-class clusters and clearer inter-class separation than the representative baselines, indicating more discriminative slide-level representations. In Fig. 4, its responses are more concentrated within annotated lesion regions, with fewer isolated activations on irrelevant tissue. PSMG also preserves responses to relatively small lesion areas while maintaining coherent localization for larger lesions, consistent with the quantitative localization results in the Supplementary Material. These observations support that semantic guidance

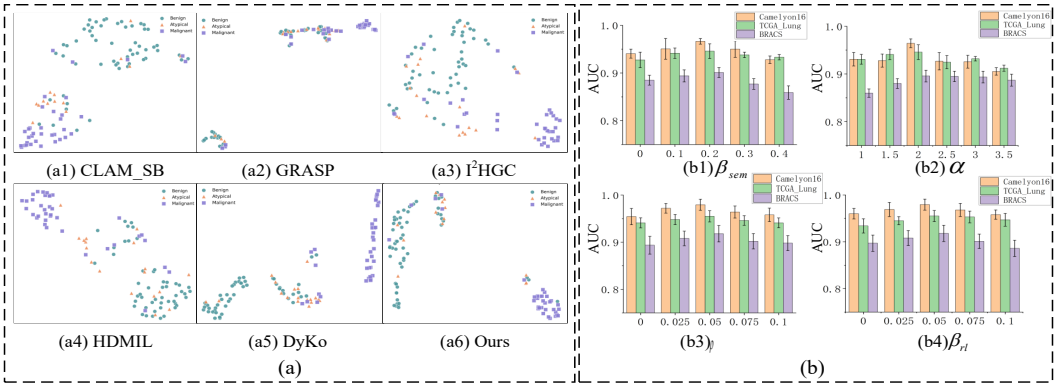


Figure 3: (a) Slide-level UMAP of different methods on BRACS. (b) Parameter sensitivity analysis of key hyperparameters in PSMG on three WSI benchmarks.

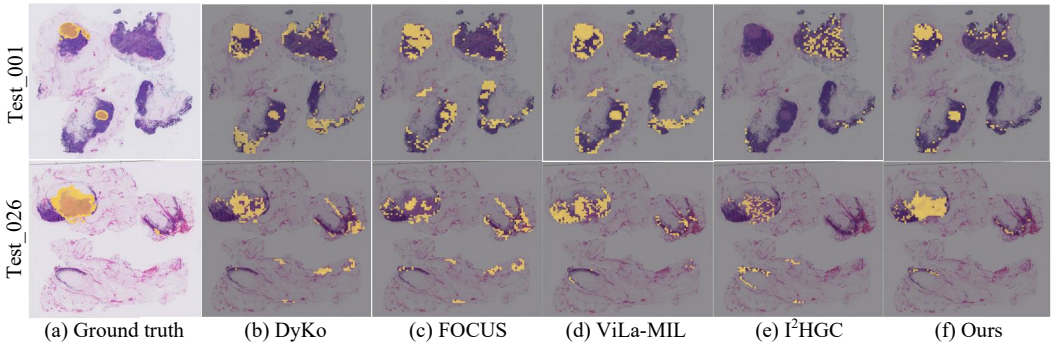


Figure 4: Heatmap comparison with representative methods on Camelyon16. (a) Ground-truth lesion annotations. (b)–(f) Heatmaps generated by DyKo, FOCUS, ViLa-MIL, I^{2}HGC, and PSMG, respectively.

and reinforced reviewing jointly improve both slide-level representation learning and lesion-focused region identification.

4.6 Parameter Sensitivity Analysis

Fig. 3(b) reports the sensitivity analysis of four key hyperparameters in PSMG. The AUC first increases and then decreases as these parameters become larger, indicating that both semantic modulation and reinforced reviewing require moderate strengths. Too small values provide insufficient guidance, while overly large values may over-amplify noisy semantic priors or reviewed regions. Camelyon16 and TCGA_Lung achieve the best performance with $\alpha = 2.0$ and $\beta_{sem} = 0.2$, whereas BRACS benefits from a sharper semantic weighting with $\alpha = 3.0$ due to its finer-grained three-class setting. For the RSR stage, $\rho = 0.05$ and $\beta_{rl} = 0.05$ provide the best trade-off across all datasets.

5 Conclusion

We presented PSMG, a semantic prior-guided framework for weakly supervised WSI classification. It uses LLM-generated descriptions as region-level priors, combines multi-scale hypergraph reasoning with task-feedback-driven reviewing, and aligns cross-magnification evidence. Experiments on three public benchmarks demonstrate strong classification performance and more lesion-focused localization, supporting the use of semantic priors as guidance for visual reasoning rather than direct fusion signals.

AI USE STATEMENT

In this work, we used generative AI tools for research brainstorming, experimental planning, code development and debugging, literature organization, manuscript editing, and language polishing. We have not used generative AI tools for generating, fabricating, modifying, or replacing experimental data or reported results, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for improving the clarity and organization of the manuscript. We have reviewed all AI-assisted work. AI-generated research ideas and methodological suggestions were independently examined by the authors, AI-assisted code was manually reviewed and tested for correctness, and all reported experimental results were obtained from the implemented models and actual experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available whole slide image benchmarks and does not involve the collection of new patient data or direct interaction with human participants. We follow the original data access and usage conditions of Camelyon16, TCGA Lung, and BRACS.

REPRODUCIBILITY STATEMENT

We provide the model formulation, training objectives, and inference procedure in Sec. 3. Experimental settings and baseline comparison protocols are described in Sec. 4. Detailed dataset statistics, preprocessing, feature extraction, optimization settings, hyperparameters, five-fold data splits, textual-prior generation, and additional experimental protocols are provided in the Supplementary Material. We also report fold-wise mean and standard deviation and provide statistical significance analyses. An anonymized implementation is available at `https://anonymous.4open.science/r/psmg-anonymous`.

REFERENCES

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *NPJ digital medicine*, 8(1):274, 2025.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

Gianpaolo Bontempo, Federico Bolelli, Angelo Porrello, Simone Calderara, and Elisa Ficarra. A graph-based multi-scale approach with knowledge distillation for wsi classification. *IEEE Transactions on Medical Imaging*, 43(4):1412–1421, 2023.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15661–15670, 2023.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-level graphs. In *Medical Image Computing and Computer Assisted Intervention*, pp. 339–349. Springer, 2021.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, 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.

Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5800–5815, 2022.

Jiuyang Dong, Junjun Jiang, Kui Jiang, Jiahan Li, and Yongbing Zhang. Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30818–30828, June 2025.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019.

Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15590–15600, June 2025.

Xiangmin Han, Huijian Zhou, Zhiqiang Tian, Shaoyi Du, and Yue Gao. Inter-intra hypergraph computation for survival prediction on whole slide images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):6006–6021, 2025. doi: 10.1109/TPAMI.2025.3557391.

Wentai Hou, Lequan Yu, Chengze Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. H2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 933–941, 2022.

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

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

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136, 2018.

Bin Li, 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.

Hao Li, Ying Chen, Yifei Chen, Rongshan Yu, Wenxian Yang, Liansheng Wang, Bowen Ding, and Yuchen Han. Generalizable whole slide image classification with fine-grained visual-semantic interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11398–11407, 2024a.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11323–11332, 2024b.

Junjian Li, Hulin Kuang, Jin Liu, Hailin Yue, Mengshen He, and Jianxin Wang. Universal-to-specific: Dynamic knowledge-guided multiple instance learning for few-shot whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26614–26623, June 2026.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023.

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

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

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Ober, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis*, 80:102486, 2022.

Ali Khajegili Mirabadi, Graham AD Archibald, Amirali Darbandsari, Alberto Contreras-Sanz, Ramin Nakhli, Maryam Asadi, Allen W Zhang, Blake Gilks, Peter Colin Black, Gang Wang, Hossein Farahani, and Ali Bashashati. GRASP: GRAph-structured pyramidal whole slide image representation. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025. URL https://openreview.net/forum?id=k6xlOfZnTC.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13183–13192, 2024.

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11248–11258, 2024a.

Jun Shi, Tong Shu, Kun Wu, Zhiguo Jiang, Liping Zheng, Wei Wang, Haibo Wu, and Yushan Zheng. Masked hypergraph learning for weakly supervised histopathology whole slide image classification. *Computer Methods and Programs in Biomedicine*, 253:108237, 2024b.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4078–4087, 2023.

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

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Honglin Xu, Naoto Usuyama, Jaspreet Bagber, Sheng Zhang, Rajesh Rao, Tristan Naumann, Clifford Wong, Zelalem Gero, Javier González, Yu Gu, et al. Whole slide image classification via iterative patch sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11637–11648, 2024.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European conference on computer vision*, pp. 125–143. Springer, 2024.

Boxuan Zhao, Jun Zhang, Deheng Ye, Jian Cao, Xiao Han, Qiang Fu, and Wei Yang. Rlogist: fast observation strategy on whole-slide images with deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3570–3578, 2023.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19677–19701, 2025.

Tingting Zheng, Kui Jiang, and Hongxun Yao. Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8028–8037, 2024.

Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burber, Margrit Liu, Daniel Bethmann, Konstantinos Syrigos, David L Rimm, Elias Zerhouni, Rohit Bhargava, et al. Graph transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11):3003–3015, 2022.

Zhonghang Zhu, Lequan Yu, Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE Transactions on Medical Imaging*, 42(5):1337–1348, 2022.

A Experimental Details

The main paper reports ACC, AUC, and F1-score comparisons under five-fold cross-validation. Here, we provide additional dataset and baseline details, statistical significance and ROC analyses, quantitative and qualitative localization results, text-prior robustness evaluation, scale ablation, and computational efficiency analysis.

A.1 Implementation Details

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 5090 GPU. The model was implemented in PyTorch 2.1.2 with CUDA 11.8 using Python 3.10. Training was performed with a batch size of one WSI for a maximum of 200 epochs. We used the Adam optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 5×10^{-3}. A cosine annealing scheduler was adopted, with the minimum learning rate set to 5×10^{-6}. Early stopping was applied according to the validation AUC with a patience of 20 epochs and a minimum improvement threshold of 1×10^{-4}. The checkpoint with the highest validation AUC was selected for final evaluation. The test set was used only for reporting final performance and was not involved in model selection or hyperparameter tuning.

For Semantic-Guided Region Modeling (SGRM), the semantic modulation strength was set to $\beta_{\text{sem}} = 0.2$. The semantic weighting sharpness was set to $\alpha = 2.0$ for Camelyon16 and TCGA_Lung, and $\alpha = 3.0$ for BRACS. For Cross-Magnification Hypergraph Alignment (CMHA), the temperature of the KL-based cross-scale knowledge alignment was set to $\tau = 1.5$, with the corresponding loss weight $\gamma = 1$.

For Reinforced Semantic Reviewing (RSR), the reviewer operated on the top-256 semantically relevant candidate regions with a reviewing ratio of 5%. PPO was optimized using a clipping coefficient of 0.2, an entropy coefficient of 0.01, and a value-function coefficient of 0.5.

All experiments were conducted using patient-level five-fold cross-validation with fixed data splits and random seeds. For each fold, slide-specific pathology descriptions were independently generated from the corresponding slide images using the same fixed prompt. No slide labels, pathology reports, or information from other folds were provided during description generation. The generated descriptions remained fixed throughout model training and inference. Final results are reported as mean$\pm$std over the five folds.

For WSI preprocessing, non-overlapping patches of size 256×256 pixels were extracted at each magnification. During hypergraph construction, each node was connected to its $K = 4$ nearest neighbors, and the normalized feature-distance threshold was set to $\theta = 0.1$.

A.2 Datasets

We evaluate PSMG on three public WSI benchmarks: Camelyon16, TCGA_Lung, and BRACS. Camelyon16 is a breast cancer lymph node metastasis dataset containing 400 H&E-stained WSIs, including 161 tumor slides and 239 normal slides. TCGA_Lung is collected from the Genomic Data Commons (GDC) portal and contains two major lung cancer subtypes, i.e., lung adenocarcinoma (LUAD) with 541 cases and lung squamous cell carcinoma (LUSC) with 513 cases. BRACS is a breast cancer WSI benchmark with multiple histopathological categories. It is used for 3-class breast cancer tissue classification, containing 525 patient-level WSIs (249 benign, 89 atypical, and 187 malignant tissue images).

We selected Camelyon16, TCGA-Lung, and BRACS because they cover complementary WSI classification settings, including binary and multi-class tasks, different organs, and varying lesion distributions and diagnostic granularity. Together, they provide a comprehensive evaluation of the robustness and generalizability of PSMG.

A.3 Data Preprocessing and Feature Extraction

We followed the WSI preprocessing pipeline of CLAM (Lu et al., 2021) and the multi-scale processing protocol of Bontempo et al. (2023). Otsu thresholding (Otsu, 1979) was first applied to separate foreground tissue regions from the slide background. The detected tissue regions were then partitioned into regular, non-overlapping patches of size 256×256 pixels at both 10X and 20X magnifications. Accordingly, the patch extraction stride was also set to 256 pixels. The same tissue detection and patch extraction procedure was applied at both magnifications to ensure consistent multi-scale processing. For each retained patch, a pretrained DINO image encoder (Caron et al., 2021) was used to extract a 384-dimensional visual feature. The image encoder was kept frozen throughout model training, and all patch features were extracted offline. For every patch, we retained its slide identifier, magnification level, and spatial coordinates in the WSI coordinate system.

A.4 Baseline Methods and Adaptation Protocol

To provide a complete description of the paradigm comparison in the main paper, we summarize the compared baselines and the adaptation protocol used in our experiments. We compare PSMG with representative WSI classification methods from four categories.

Pooling-based methods. Mean-pooling and Max-pooling are used as simple aggregation baselines. They aggregate all patch-level features within a WSI into a slide-level representation without learnable region selection.

MIL-based methods. We include ABMIL (Ilse et al., 2018) and CLAM (Lu et al., 2021), which are widely used weakly supervised WSI classification methods. ABMIL learns attention weights over instances, while CLAM further introduces clustering-constrained attention learning for instance-level discrimination.

Table 5: Paired significance testing for AUC. Each cell reports mean paired ΔAUC (PSMG − baseline). Significance markers are based on BH-FDR adjusted q-values, where $q < 0.05$ (*), $q < 0.01$ (**), and ns denotes not significant.

Baseline	Camelyon16	TCGA_Lung	BRACS
Mean pooling	$+0.224^{**}$	$+0.054^{**}$	$+0.171^{**}$
Max pooling	$+0.125^{**}$	$+0.057^{**}$	$+0.194^{**}$
ABMIL	$+0.137^{**}$	$+0.132^{**}$	$+0.037^{*}$
CLAM_MB	$+0.077^{**}$	$+0.061^{*}$	$+0.017^{*}$
CLAM_SB	$+0.080^{**}$	$+0.065^{*}$	$+0.017^{*}$
DyKo	0.000^{ns}	$+0.029^{*}$	$+0.031^{*}$
FOCUS	$+0.023^{ns}$	$+0.024^{*}$	$+0.045^{*}$
GRASP	$+0.010^{ns}$	$+0.017^{ns}$	$+0.025^{*}$
HDMIL	-0.008^{ns}	$+0.010^{ns}$	$+0.029^{*}$
I^2HGC	$+0.056^{*}$	$+0.014^{ns}$	$+0.020^{*}$
ViLa-MIL	$+0.028^{*}$	$+0.077^{*}$	$+0.036^{*}$

Graph- and hypergraph-based methods. We compare with I^2HGC (Han et al., 2025) and other representative structure-aware WSI methods. These methods model spatial or semantic relations among patches to improve slide-level representation learning. This group is used to examine whether explicit pathological semantic guidance can further improve structure-based WSI modeling.

Recent WSI and vision-language methods. We further include HDMIL (Dong et al., 2025), GRASP (Mirabadi et al., 2025), ViLa-MIL (Shi et al., 2024a), FOCUS (Guo et al., 2025), and DyKo (Li et al., 2026). These methods represent recent advances in hierarchical WSI modeling, region mining, and vision-language WSI analysis.

All methods are evaluated using the same patient-level five-fold splits, patch encoder, preprocessing protocol, and evaluation metrics. Single-scale methods use 20X patch features, whereas methods natively supporting multi-scale inputs use 10X and 20X features.

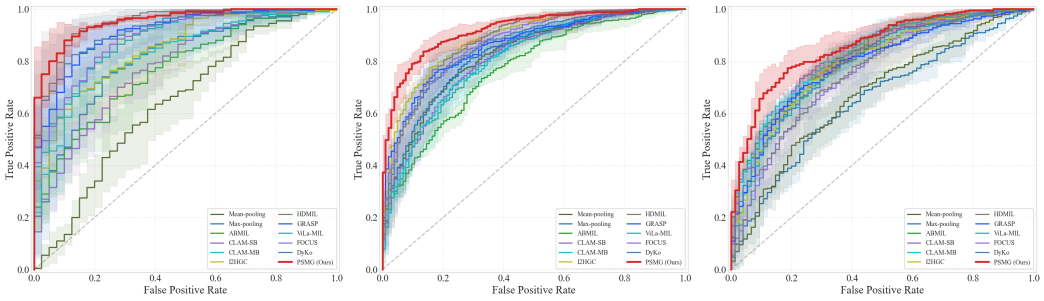
A.5 Statistical Significance Analysis

To evaluate the consistency of AUC improvements across folds, we perform two-sided paired t-tests between PSMG and each baseline using fold-wise AUC values under the same data splits. To account for multiple baseline comparisons, the resulting p-values are adjusted using the Benjamini–Hochberg false discovery rate (BH-FDR) correction. Table 5 reports the mean paired ΔAUC between PSMG and each baseline, with significance markers determined by the BH-FDR adjusted q-values.

The significance analysis provides additional evidence for the effectiveness of PSMG. PSMG achieves statistically supported AUC improvements over most conventional pooling, MIL, and CLAM-based baselines across datasets, confirming that semantic-prior-guided multi-scale region modeling is more effective than purely visual patch aggregation. The consistent gains over ViLa-MIL further support our motivation that LLM-generated descriptions are better used as semantic priors for region-level modulation rather than as direct classification cues. Notably, PSMG shows significant improvements over all compared methods on BRACS, suggesting that semantic-guided modulation and task-feedback reviewing are particularly beneficial for fine-grained pathology classification. For several recent strong WSI methods on Camelyon16 and TCGA_Lung, the AUC differences are smaller and some comparisons are not statistically significant, reflecting the strong competitiveness of existing visual-region mining methods on tasks with relatively localized or visually distinctive evidence.

A.6 Threshold-Independent Discrimination Analysis

To evaluate threshold-independent discrimination, we plot ROC curves for all methods. For the binary Camelyon16 and TCGA_Lung datasets, ROC curves are computed using positive-class scores, while for the multi-class BRACS dataset, macro-averaged one-vs-rest ROC/AUC is used.



(a) Camelyon16 (b) TCGA_Lung (c) BRACS

Figure 5: ROC curves of different methods on three WSI benchmarks.

Table 6: Quantitative localization evaluation on representative Camelyon16.

Method	Test_001 (Micro)			Test_026 (Macro)		
	IoU	Recall	Precision	IoU	Recall	Precision
HDMIL	0.514	0.785	0.612	0.685	0.842	0.783
FOCUS	0.556	0.812	0.648	0.642	0.805	0.751
DyKo	0.598	0.834	0.681	0.721	0.879	0.804
PSMG	**0.672**	**0.895**	**0.734**	**0.796**	**0.926**	**0.852**

The ROC curves in Fig. 5 show that PSMG achieves consistently strong discrimination across the three WSI benchmarks. On TCGA_Lung and BRACS, PSMG maintains higher true positive rates than most competing methods over a wide range of false positive rates, indicating a better sensitivity–specificity trade-off. The advantage is particularly evident on BRACS, where the macro-averaged one-vs-rest ROC reflects the model's ability to distinguish fine-grained benign, atypical, and malignant categories. On Camelyon16, PSMG remains highly competitive, although HDMIL obtains a slightly higher AUC, suggesting that strong region-mining methods are also effective for localized metastatic evidence.

B Attention Localization Analysis

B.1 Quantitative Tumor Localization Evaluation

PSMG combines semantic-guided region modulation with task-feedback-driven reviewing. We examine whether the resulting model responses overlap with annotated lesion regions. Therefore, beyond slide-level classification performance, we further investigate whether the learned attention responses are consistent with pathological evidence. We quantitatively evaluate the agreement between model attention regions and tumor annotations on representative Camelyon16 cases. Following the qualitative visualization in the main paper, we select two test slides with different metastatic patterns, including micro-metastasis (Test001) and macro-metastasis (Test026). For each method, the attention heatmap is binarized using the same threshold and compared with the ground-truth tumor annotation using IoU, recall, and precision.

As shown in Table 6, PSMG achieves better agreement with tumor annotations on both micro-metastasis and macro-metastasis cases. Qualitatively, the reviewed regions of PSMG are more concentrated on pathologically relevant areas with fewer false positive activations. Meanwhile, the improved recall indicates that PSMG shows higher overlap with annotated lesion regions.

B.2 Additional Qualitative Localization Results

To further illustrate how semantic guidance and reinforced reviewing affect region selection across different pathological appearances, we provide additional qualitative examples on three independent

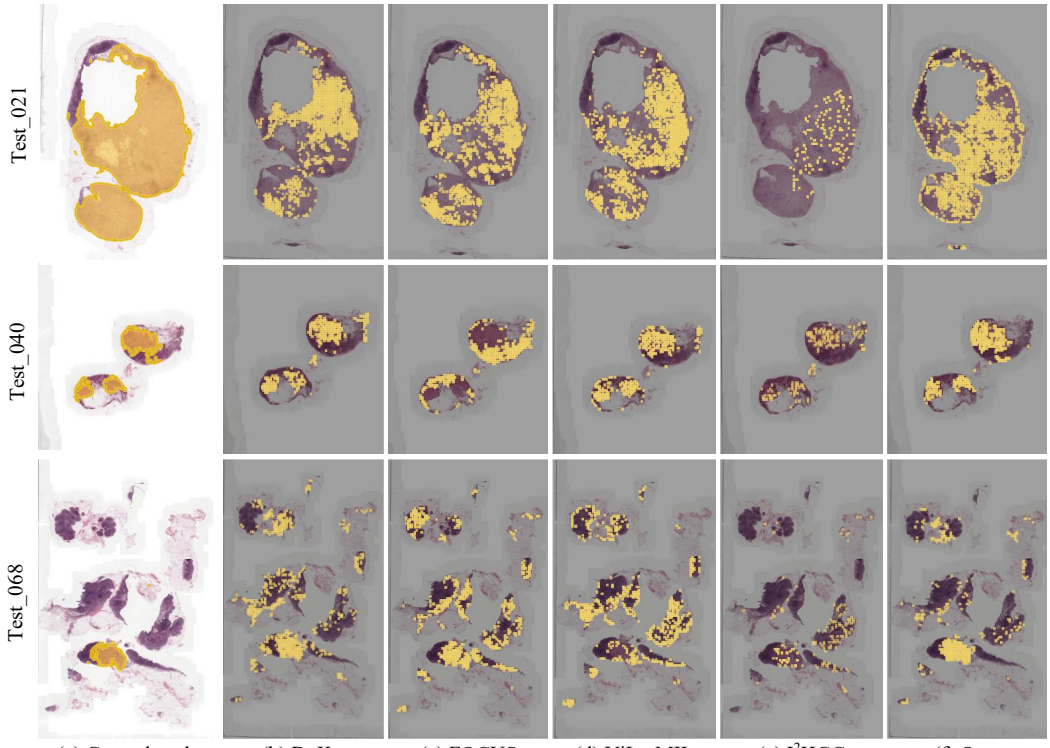


Figure 6: Additional qualitative localization results on Camelyon16 test cases: (a) Ground truth, (b) DyKo, (c) FOCUS, (d) ViLa-MIL, (e) I^2HGC, (f) Ours.

Table 7: Test-time textual-prior sensitivity analysis on three WSI benchmarks. Values are reported in percentage (%) as mean±std over five-fold cross-validation using fixed trained PSMG checkpoints.

Variant	Camelyon16			TCGA_Lung			BRACS		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
No text	96.9±1.5	97.1±1.4	94.8±2.3	91.5±2.2	95.4±1.6	89.8±2.9	77.2±2.7	88.4±2.5	70.2±3.4
Mismatched text	97.2±1.4	97.5±1.3	95.2±2.2	92.4±1.9	96.3±1.3	90.9±2.7	79.1±2.4	89.8±2.2	72.8±3.1
Generic text	97.4±1.3	97.6±1.3	95.5±2.1	92.8±1.7	96.5±1.2	91.4±2.6	79.5±2.2	90.1±2.0	73.2±2.9
Mask 50%	97.5±1.3	97.7±1.3	95.6±2.1	93.1±1.7	96.7±1.1	91.8±2.6	80.6±1.8	90.5±1.9	74.5±2.7
Mask 25%	97.6±1.2	97.8±1.2	95.8±2.1	93.6±1.6	97.0±1.1	92.3±2.6	81.8±1.4	90.9±1.8	76.0±2.6
Normal LLM text	**97.8±1.1**	**97.9±1.2**	**96.0±2.1**	**94.1±1.5**	**97.2±1.0**	**92.8±2.6**	**83.1±1.2**	**91.2±1.7**	**77.6±2.5**

Camelyon16 cases in Fig. 6. These cases cover different tumor morphologies and metastatic patterns. These additional examples show similar localization behavior to the quantitatively evaluated cases, with more concentrated responses within tumor regions and fewer irrelevant activations.

C Analysis of Textual Semantic Priors

The semantic prior provides pathological guidance for region modulation in SGRM. To examine whether the model relies on a fixed textual description or learns robust semantic representations, we evaluate the model under different test-time textual priors while keeping all other components unchanged.

C.1 Test-Time Text Prior Sensitivity

To examine the dependence of PSMG on textual semantic priors, we perturb only the text supplied to SGRM at test time while keeping the trained model and all other settings fixed. We compare the

Table 8: AUC (%) results of PSMG under different single-scale and multi-scale input settings. Values are mean±std over five folds.

Scale setting	Camelyon16	TCGA_Lung	BRACS
5X	80.6 ± 3.8	85.4 ± 2.9	72.5 ± 3.3
10X	93.2 ± 2.1	92.6 ± 1.9	84.4 ± 2.4
20X	95.1 ± 1.8	94.1 ± 1.6	86.8 ± 2.2
5X+10X	94.3 ± 2.0	93.5 ± 1.7	85.7 ± 2.1
5X+20X	96.6 ± 1.5	95.8 ± 1.3	89.1 ± 1.9
10X+20X (Ours)	$\mathbf{97.9 \pm 1.2}$	$\mathbf{97.2 \pm 1.0}$	$\mathbf{91.2 \pm 1.7}$

Table 9: Computational resource comparison. Param. and Mem. denote model parameters (M) and peak GPU memory (MB), respectively.

Method	Param.	Mem.
PSMG-Stage1	2.57	149.91
PSMG (Full)	2.77	231.24
FOCUS	1.51	132.88
ViLa-MIL	1.73	166.56
DyKo	3.11	333.23
HDMIL	0.50	100.49

normal slide-specific LLM description with five variants: no text, mismatched text, generic text, and descriptions with 25% or 50% token masking.

As shown in Table 7, the normal slide-specific description consistently achieves the best performance on all three datasets. Removing the textual prior causes the largest degradation, indicating that semantic guidance provides complementary information for region modeling. Mismatched and generic descriptions perform better than the no-text setting but remain inferior to the normal description, showing that the relevance and specificity of the textual prior affect candidate-region guidance. When 25% or 50% of the text is masked, performance decreases gradually rather than sharply, suggesting that PSMG does not depend on every textual detail. The effect of text perturbation is relatively small on Camelyon16 but more pronounced on BRACS, where finer-grained class distinctions may benefit more from slide-specific semantic cues. Overall, these results show that PSMG benefits from informative textual priors while retaining substantial predictive capability from visual representations.

D Ablation studies

D.1 Single scale vs multi scale

As shown in Table 8, single-scale inputs provide incomplete pathological evidence. The 5X setting performs the worst across all datasets, indicating that low-magnification tissue context alone is insufficient for discriminative WSI classification. The 10X and 20X settings achieve better performance, with 20X being the strongest single-scale input, suggesting the importance of fine-grained morphological details.

The multi-scale results further show that performance gains do not simply come from adding another scale. Although 5X+10X improves over 5X and 10X, it remains inferior to 20X, indicating that high-magnification evidence is essential. In contrast, both 5X+20X and 10X+20X outperform the corresponding single-scale variants, demonstrating the complementarity between contextual tissue structure and local morphological details. Among all settings, 10X+20X achieves the best AUC on all three datasets, suggesting that medium-magnification tissue context and high-magnification details can be more effectively aligned by PSMG for cross-scale hypergraph reasoning and semantic-guided region modeling.

D.2 Efficiency Analysis

Table 9 reports the model size and peak GPU memory of representative methods. PSMG introduces additional computational overhead due to multi-scale hypergraph reasoning and reinforced region reviewing. Compared with PSMG-Stage1, the full model uses slightly more parameters and GPU memory because it includes the task-feedback-driven reviewing process. Nevertheless, its overall resource requirement remains moderate. Although PSMG is not the lightest model, it uses fewer parameters than DyKo and requires substantially less peak memory. These results indicate that PSMG achieves improved WSI classification with acceptable resource overhead.