Addressing Tissue and Appearance Heterogeneity in Text-Guided Few-Shot WSI Classification

Yongcen Li[1], Zelin Xu[1], Xiaoyu Shi[2], Zhiyi Huang[1], Xinchen Ye[1], Miao Zhang[1], Zhihui Wang[1], Rui Xu[1,✉], Yen-Wei Chen[3], and Xudong Xing[2,✉]

[1] Dalian University of Technology, China
[2] Beijing Institute of Genomics, Chinese Academy of Sciences, China National Center for Bioinformation, China
[3] Ritsumeikan University, Japan

Correspondence to Rui Xu: xurui@dlut.edu.cn or Xudong Xing: xingxd@big.ac.cn

Abstract. In text-guided few-shot whole-slide image (WSI) classification, multiple instance learning (MIL) methods can leverage vision–language priors with very limited annotations, yet they still lag behind strongly supervised baselines. We empirically show that this gap is largely caused by slide-to-slide domain shifts induced by tissue-composition and appearance heterogeneity. Under few-shot supervision, these shifts lead to unstable key-instance selection and feature aggregation, thereby degrading the reliability of slide-level predictions. This problem is particularly pronounced at low magnification (5×): although 5× features provide broader global context, they are more sensitive to staining variation, artifacts, and background content than 20× features, making low-magnification cues less reliable across slides. To address this issue, we propose a Heterogeneity-Aware Text-guided MIL with two modules: the Tissue-Composition Balancing Module (TCBM), which performs tissue-balanced patch selection across tissue-related clusters to mitigate composition bias, and the Appearance-Robust Aggregation Module (ARAM), which constructs low-magnification contextual representations from stable 20× local structural patterns without using raw 5× appearance features, which are sensitive to staining variation and slide-to-slide domain shifts, thereby improving robustness to appearance heterogeneity. Code is available at: https://github.com/noviq123/HAT.

Keywords: Few-Shot Learning · Whole-Slide Image · Text-Guided MIL

1 Introduction

Whole-slide images (WSIs) have been widely adopted in computational pathology for disease diagnosis and analysis. However, the gigapixel resolution and inherent multi-scale structure of WSIs make dense annotation and end-to-end training costly. Consequently, multiple instance learning (MIL) has emerged as a prevalent paradigm [7, 12, 9, 16, 2], aggregating patch-level representations for slide-level prediction. However, most existing MIL approaches still depend on

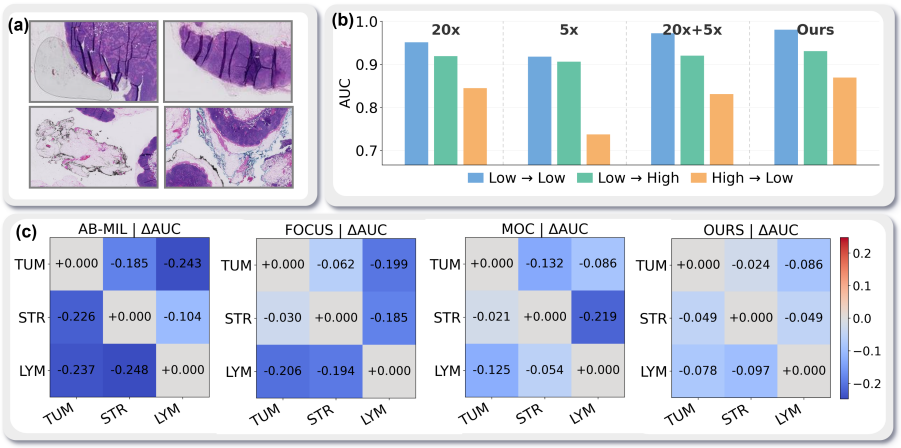

Fig. 1. Analysis of appearance and tissue-composition heterogeneity in few-shot WSI classification. (a) Representative slides with appearance perturbations, including staining variation, folds, debris, and scanning artifacts. (b) Test AUC under appearance distribution shift. Slides are divided into low- and high-appearance-perturbation groups. Each method setting (20×, 5×, 20×+5×, and HAT) is evaluated under three train/test combinations: Low→Low, Low→High, and High→Low. (c) Performance drop relative to the matched in-group setting (ΔAUC) under tissue-composition shift. Slides are divided into tumor-dominant (TUM), stroma-dominant (STR), and lymphocyte-rich (LYM) groups according to the dominant tissue type. Rows denote training groups and columns denote test groups. More negative values indicate larger degradation under out-group testing.

substantial amounts of WSI-level supervision, which is often impractical in clinical settings. This motivates data-efficient learning under limited supervision.

Recently, text-guided few-shot methods built upon pathology vision–language models [22, 6, 10, 11] compute image–text similarities between patch features and disease-related textual descriptions. These similarity scores guide MIL key-instance selection and aggregation, improving discrimination under extremely limited supervision [14, 23, 4, 20]. Many methods further incorporate multiple magnification modeling to exploit complementary cues across scales [17, 5]: the 20× view captures fine-grained diagnostic signals (e.g., cellular morphology), whereas the 5× view provides tissue-level context (e.g., architecture and spatial layout) for slide-level prediction. However, these methods still lag behind strongly supervised baselines trained with large-scale annotations.

A central bottleneck in text-guided few-shot WSI learning is the amplified inter-slide domain shift caused by tissue-composition and appearance heterogeneity. Across slides, the proportions of major tissue components (e.g., tumor, stroma, and lymphocyte-rich regions) can vary substantially, while appearance factors such as staining variation, folds, debris, and scanning artifacts further increase slide-to-slide inconsistency (Fig. 1a). Under few-shot supervision, models

are more vulnerable to such shifts and struggle to learn stable, transferable diagnostic patterns. To examine these effects, we conduct a controlled distribution-shift evaluation by training on one group and comparing performance on both in-group and out-of-group test slides. In Fig. 1b, when training on the High-appearance-perturbation group, performance drops under appearance distribution shift, with 5× features degrading more than 20× features. In Fig. 1c, AUC also drops consistently when transferring from one dominant tissue group to another, confirming substantial domain shift under tissue-composition variation. This issue is especially pronounced at low magnification (5×), which is more sensitive to staining variation, artifacts, and background noise. This further suggests that simply incorporating low-magnification inputs is insufficient, and motivates the need for constructing a more robust low-magnification proxy rather than directly using raw 5× features.

To address these issues, we propose Heterogeneity-Aware Text-guided MIL (HAT), a framework designed to reduce the effect of tissue-composition variation and appearance perturbations under extremely limited supervision, thereby improving the stability and generalization of slide-level predictions. Specifically, HAT contains two modules. The first is the Tissue-Composition Balancing Module (TCBM), which groups patches into tissue-related clusters based on feature-space similarity and performs tissue-balanced patch selection across these clusters to mitigate composition bias. The second is the Appearance-Robust Aggregation Module (ARAM), which avoids directly using raw 5× appearance features that are sensitive to staining variation and artifacts. Instead, it organizes local structural units from the more stable 20× view, selects representative structures, and aggregates them into a representation capturing global tissue context and spatial layout.

Our contributions are threefold. First, we identify tissue-composition and appearance heterogeneity as two major sources of slide-to-slide variation in text-guided few-shot WSI classification, and we demonstrate their effects through controlled experiments. Second, we propose two dedicated modules, TCBM and ARAM. TCBM mitigates composition bias through tissue-balanced patch selection across tissue-related clusters, while ARAM builds low-magnification contextual representations from stable high-magnification structural information without directly relying on raw 5× appearance features. Third, extensive experiments across multiple datasets, backbone networks, and few-shot settings demonstrate consistent improvements in both classification performance and prediction stability over existing approaches.

2 Methodology

2.1 Preliminaries

Let the dataset be denoted as $\mathcal{D} = \{(\mathcal{X}_i, y_i)\}_{i=1}^{I}$, where $\mathcal{X}_i \in \mathbb{R}^{H\times W\times 3}$ is the i-th whole-slide image and $y_i \in \{1, 2, \ldots, C\}$ is its slide-level label, with C classes in total. Due to the gigapixel resolution of WSIs, we partition the tissue foreground of each WSI into N_i non-overlapping patches $\{x_{i,j}\}_{j=1}^{N_i}$, where $x_{i,j} \in \mathbb{R}^{h\times w\times 3}$,

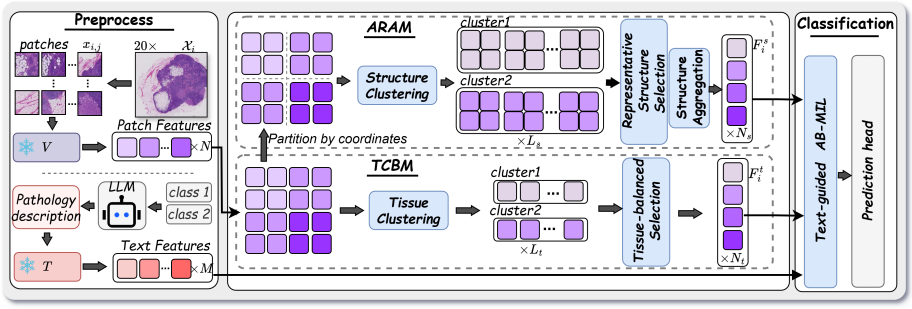

Fig. 2. Illustration of the proposed Heterogeneity-Aware Text-guided MIL (HAT).

and patch-level labels are unavailable. Given a pre-trained visual encoder $f(\cdot)$, we obtain the corresponding bag of instance representations

$$\mathcal{B}_i = \{\mathbf{F}_{i,j}\}_{j=1}^{N_i}, \quad \mathbf{F}_{i,j} = f(x_{i,j}) \in \mathbb{R}^D,$$

where D denotes the feature dimension.

We consider the C-way K-shot few-shot WSI classification setting, where the training set $\mathcal{D}_{\text{tr}} \subset \mathcal{D}$ contains only K labeled WSIs per class, and thus $|\mathcal{D}_{\text{tr}}| = K \times C$.

2.2 Tissue-Composition Balancing Module (TCBM)

TCBM mitigates aggregation bias from inter-slide tissue-composition variation through balanced instance selection. Given slide i with instance features $\mathcal{B}_i = \{\mathbf{F}_{i,j}\}_{j=1}^{N_i}$, we measure instance similarity using cosine distance

$$d(\mathbf{F}, \mathbf{F}') = 1 - \frac{\mathbf{F}^\top \mathbf{F}'}{\|\mathbf{F}\| \, \|\mathbf{F}'\|}. \tag{1}$$

Based on $d(\cdot)$, we perform tissue clustering by partitioning instances into L_t tissue-related clusters, where L_t denotes the number of clusters, with assignments $\ell_{i,j} \in \{1, \ldots, L_t\}$. Let $\mathcal{C}_{i,k} = \{j \mid \ell_{i,j} = k\}$ be the index set of instances in cluster k. We then perform tissue-balanced selection by selecting N_t instances in a cluster-balanced manner. Specifically, we select $q = \lceil N_t / L_t \rceil$ instances from each cluster. Let $\mathcal{S}_{i,k}$ denote the selected index subset from cluster k for slide i, defined as

$$\mathcal{S}_{i,k} = \arg \min_{\substack{\mathcal{S} \subseteq \mathcal{C}_{i,k} \\ |\mathcal{S}| = r_{i,k}}} \sum_{j \in \mathcal{S}} d(\mathbf{F}_{i,j}, \mathbf{c}_{i,k}), \qquad r_{i,k} = \min\big(q, |\mathcal{C}_{i,k}|\big), \tag{2}$$

where $\mathbf{c}_{i,k} \in \mathbb{R}^D$ is the cluster prototype in feature space (e.g., the medoid). The final selected set is

$$\mathcal{S}_i = \bigcup_{k=1}^{L_t} \mathcal{S}_{i,k}, \qquad |\mathcal{S}_i| \le N_t, \tag{3}$$

yielding $\mathcal{B}_i^t = \{\mathbf{F}_{i,j} \mid j \in \mathcal{S}_i\}$. By enforcing balanced selection across tissue-related clusters, TCBM reduces sensitivity to tissue-distribution fluctuations without requiring additional annotations. This cluster-balanced selection is designed to reduce sensitivity to tissue composition imbalance across slides, which is a major source of inter-slide domain shift in few-shot settings.

2.3 Appearance-Robust Aggregation Module (ARAM)

ARAM improves robustness to appearance heterogeneity by constructing a low-magnification context proxy from stable 20× local structures, rather than directly using unstable 5× appearance features. Given slide i with 20× instance features $\mathcal{B}_i = \{\mathbf{F}_{i,j}\}_{j=1}^{N_i}$ and coordinates $\mathbf{P}_i = \{\mathbf{p}_{i,j}\}_{j=1}^{N_i}$, we denote by $\mathbf{G}_{i,m}$ the m-th local structural unit organized as a 4×4 grid. We then form local structural units as

$$\mathbf{G}_{i,m} = \{\mathbf{F}_{i,\pi(m,u)}\}_{u=1}^{16} \in \mathbb{R}^{16\times D}, \qquad m = 1, \ldots, U_i, \tag{4}$$

where $\pi(m, u)$ maps each grid position to an instance index, and U_i is the number of valid grids with complete 4×4 neighborhoods. We use $\mathrm{vec}(\cdot)$ to denote flattening a 4×4 grid into a vector in $\mathbb{R}^{16D}$.

We perform structure clustering over structural units using a rotation- and reflection-invariant cosine distance. Let $\mathcal{T}$ be the set of rotations and flips, and define

$$d_s(\mathbf{G}, \mathbf{G}') = \min_{\tau\in\mathcal{T}} \left(1 - \frac{\langle \mathrm{vec}(\mathbf{G}),\, \mathrm{vec}(\tau(\mathbf{G}'))\rangle}{\|\mathrm{vec}(\mathbf{G})\|\, \|\mathrm{vec}(\tau(\mathbf{G}'))\|}\right). \tag{5}$$

Based on $d_s(\cdot)$, structural units are partitioned into L_s clusters, and we perform cluster-balanced selection by choosing $\lceil N_s/L_s \rceil$ representative units from each cluster. Each selected unit is aggregated into a structure-level representation by a token aggregator $\phi(\cdot)$:

$$\mathbf{F}_{i,m}^s = \phi(\mathbf{G}_{i,m}) \in \mathbb{R}^D, \tag{6}$$

where $\phi(\cdot)$ is a lightweight Transformer [19] encoder followed by mean pooling over the 16 tokens. The ARAM output is $\mathcal{B}_i^s = \{\mathbf{F}_{i,m}^s\}_{m=1}^{N_s}$, which provides low-magnification contextual cues with improved robustness to appearance perturbations. By aggregating representative structural units across the slide, ARAM captures tissue organization and spatial layout, thereby serving as a robust proxy for low-magnification contextual information. This avoids directly modeling raw 5× appearance features, which are unstable across slides due to staining and scanning variation, and instead builds a more robust low-magnification proxy from 20× structures.

2.4 Text-Guided MIL

For slide i, TCBM outputs $\mathcal{B}_i^t$ and ARAM outputs $\mathcal{B}_i^s$. We combine them to form the final instance set

$$\mathcal{B}_i^{\mathrm{all}} = \mathcal{B}_i^t \cup \mathcal{B}_i^s = \{\mathbf{F}_{i,n}\}_{n=1}^{\tilde{N}_i}, \qquad \tilde{N}_i = |\mathcal{B}_i^t| + |\mathcal{B}_i^s|. \tag{7}$$

Let $\mathbf{T}^{\text{high}} = \{\mathbf{t}^{\text{high}}_{c,m}\}$ and $\mathbf{T}^{\text{low}} = \{\mathbf{t}^{\text{low}}_{c,m}\}$ be the text-embedding banks, where $c \in \{1, \dots, C\}$ and $m \in \{1, \dots, M\}$. For an instance $\mathbf{F}$ and class c, the class-specific text prior is

$$s_c(\mathbf{F}) = \begin{cases} \dfrac{1}{M} \sum_{m=1}^{M} \dfrac{\mathbf{F}^\top \mathbf{t}^{\text{high}}_{c,m}}{\|\mathbf{F}\| \, \|\mathbf{t}^{\text{high}}_{c,m}\|}, & \mathbf{F} \in \mathcal{B}^t_i, \\ \dfrac{1}{M} \sum_{m=1}^{M} \dfrac{\mathbf{F}^\top \mathbf{t}^{\text{low}}_{c,m}}{\|\mathbf{F}\| \, \|\mathbf{t}^{\text{low}}_{c,m}\|}, & \mathbf{F} \in \mathcal{B}^s_i. \end{cases} \tag{8}$$

We use a learnable scorer $g(\cdot)$ within an attention-based MIL framework [7] to produce class-specific instance logits $u^{(c)}_{i,n} = g(\mathbf{F}_{i,n})_c$, and add the text prior before normalization:

$$\bar{\alpha}^{(c)}_{i,n} = \frac{\exp\big(u^{(c)}_{i,n} + \gamma \, s_c(\mathbf{F}_{i,n})\big)}{\sum_{k=1}^{\tilde{N}_i} \exp\big(u^{(c)}_{i,k} + \gamma \, s_c(\mathbf{F}_{i,k})\big)}, \qquad \mathbf{F}^{(c)}_i = \sum_{n=1}^{\tilde{N}_i} \bar{\alpha}^{(c)}_{i,n} \mathbf{F}_{i,n}, \qquad \mathbf{F}_i = \frac{1}{C} \sum_{c=1}^{C} \mathbf{F}^{(c)}_i, \tag{9}$$

where γ controls the strength of text guidance. Finally, we obtain the slide-level prediction by $\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}\mathbf{F}_i)$ and optimize the model using the cross-entropy loss.

3 Experiments

Datasets. We evaluate our method on three public pathology datasets: TCGA-NSCLC [18], CAMELYON [3], and BRACS [1]. TCGA-NSCLC contains two lung cancer subtypes, LUAD (525 slides) and LUSC (507 slides); CAMELYON includes 398 Normal and 271 Tumor slides; and BRACS comprises three categories, Benign (248), Atypical (89), and Malignant (193). Following the few-shot protocol, each dataset is split into training, validation, and test sets. We then randomly sample K slides per class from the training set to construct the K-shot training set.

Implementation Details. We use TRIDENT [21] to crop each WSI at 20× and 5× magnification into non-overlapping 256×256 patches. Pathology knowledge prompts are generated by a frozen large language model, ChatGPT-5.1, using the prompt: "Please provide concise morphological descriptions for {class name} at high and low resolution, using bullet points to list distinct morphological features for each resolution." For the controlled analysis in Fig. 1c, we use TIAToolbox [13] to obtain tissue-type assignments. We set $N_t = 512$, $N_s = 128$, $L_t = 16$, $L_s = 8$, and $\gamma = 0.5$, which are selected based on validation performance to balance representation diversity and computational efficiency. Both TCBM and ARAM employ CLARA-based k-medoids clustering [8, 15]. Selected structural units are aggregated by a 2-layer Transformer with 8 attention heads and a dropout rate of 0.1. We optimize the model using Adam with a learning rate of 0.02 and a weight decay of 1e-4.

Evaluation Metrics. We report balanced accuracy (BACC) and AUC, averaged over 5 runs with mean and standard deviation.

Table 1. Few-shot weakly-supervised learning results (%) on TCGA-NSCLC, BRACS and CAMELYON16 under 4-shot, 8-shot, and 16-shot settings using the CONCH encoder. Best results are in bold and second best are underlined.

Setting	Method	TCGA-NSCLC		BRACS		CAMELYON16	
		BACC	AUC	BACC	AUC	BACC	AUC
4-shot	AB-MIL [7]	76.8±2.7	85.7±1.5	54.5±1.6	75.3±2.2	63.2±1.5	64.4±2.6
	CLAM [12]	75.8±1.2	84.4±2.3	55.3±0.9	77.2±1.3	62.4±3.8	64.5±2.0
	DS-MIL [9]	76.1±1.5	84.8±1.1	55.6±2.6	77.0±1.7	62.5±2.2	66.3±2.1
	TransMIL [16]	73.1±3.6	81.5±2.4	56.5±2.0	76.4±1.4	63.3±3.6	66.5±5.0
	TOP [14]	77.1±5.9	86.0±4.8	58.0±6.7	77.9±8.7	64.8±2.2	68.2±1.4
	ViLa-MIL [17]	80.1±4.7	90.2±1.0	57.8±1.8	78.3±0.9	65.4±3.1	69.7±2.8
	MSCPT [5]	80.4±4.8	88.4±4.7	58.1±1.4	78.2±0.8	64.0±2.9	67.8±5.3
	FOCUS [4]	81.5±4.5	90.6±3.2	57.6±3.1	78.5±2.1	66.7±2.8	71.6±8.2
	MOC [20]	81.9±3.9	90.6±3.1	57.5±2.0	78.8±2.0	66.7±2.9	71.4±4.5
	HAT	**84.0±2.1**	**92.0±1.7**	**61.5±1.9**	**81.1±1.5**	**70.2±6.8**	**74.3±7.8**
8-shot	AB-MIL [7]	80.1±0.9	89.2±0.9	55.9±2.7	76.5±1.8	78.4±1.3	83.8±1.6
	CLAM [12]	80.6±1.5	90.3±1.5	56.4±2.2	77.5±1.0	79.6±5.1	85.2±3.3
	DS-MIL [9]	82.9±4.5	91.9±3.0	57.2±1.1	78.1±2.5	77.6±5.7	85.0±2.4
	TransMIL [16]	79.5±1.6	88.5±1.9	58.4±3.4	78.9±0.5	77.9±2.8	86.1±5.2
	TOP [14]	81.3±4.7	90.9±1.4	59.2±4.1	79.3±1.0	80.7±2.2	87.2±2.8
	ViLa-MIL [17]	82.9±4.5	91.9±3.0	58.1±3.5	80.5±2.0	80.3±6.4	87.4±7.9
	MSCPT [5]	84.0±3.0	91.6±3.3	59.3±2.8	81.1±1.1	82.1±2.3	86.9±1.7
	FOCUS [4]	83.3±7.4	92.7±3.3	59.9±2.7	79.7±1.9	84.5±5.5	89.1±3.6
	MOC [20]	85.6±1.9	92.6±2.2	59.9±2.6	80.9±1.3	83.3±7.9	87.5±4.6
	HAT	**87.7±1.3**	**94.0±0.8**	**65.1±1.9**	**83.5±0.9**	**87.1±2.5**	**91.3±3.1**
16-shot	AB-MIL [7]	84.3±1.6	93.0±1.0	59.3±3.2	79.9±0.9	79.9±1.5	85.1±2.5
	CLAM [12]	84.7±1.9	93.3±0.9	59.0±2.0	82.5±2.1	83.0±2.2	86.3±2.7
	DS-MIL [9]	87.9±4.1	92.8±3.8	60.0±1.0	81.0±2.2	83.3±3.0	86.9±0.9
	TransMIL [16]	86.8±2.7	92.5±2.0	60.2±2.7	81.4±1.5	80.3±2.4	88.4±1.8
	TOP [14]	88.9±2.4	94.3±0.9	61.2±3.4	80.2±1.9	85.4±3.0	89.2±0.4
	ViLa-MIL [17]	89.7±2.1	94.8±2.5	60.7±3.5	82.8±1.0	88.9±4.3	90.8±4.8
	MSCPT [5]	87.7±2.3	95.1±1.7	63.7±3.0	82.2±1.2	83.1±2.1	91.6±2.9
	FOCUS [4]	88.6±2.1	95.4±1.7	61.2±2.1	82.7±3.7	87.1±3.5	91.2±4.4
	MOC [20]	88.5±1.3	95.1±1.0	62.3±3.0	82.6±1.3	88.4±2.6	91.2±4.4
	HAT	**90.9±0.5**	**96.3±0.5**	**69.7±2.8**	**86.1±1.9**	**91.0±1.0**	**94.4±1.3**

3.1 Main Results

Table 1 shows that ViLa-MIL [17] and MSCPT [5] benefit from 20×+5× multi-magnification modeling, but their reliance on 5× appearance cues may make them more sensitive to appearance heterogeneity under few-shot supervision. FOCUS [4] and MOC [20] improve performance via representative patch selection, yet they can still be biased by tissue-composition heterogeneity when slide-level tissue proportions shift. In contrast, our method alleviates both issues and yields consistent improvements across datasets, with further gains over strong prior methods (e.g., BRACS 4-shot AUC: 78.8 → 81.1).

3.2 Ablation Study

We conduct two ablation studies to validate our design for handling inter-slide heterogeneity. First, under the 8-shot protocol with PLIP features, the full model achieves the best performance on all three datasets (Fig. 3a), indicating strong generalization and robustness. Second, Table 2 evaluates the contribution of each

Table 2. Ablation study of key modules of HAT under the 8-shot setting.

Ablation (Modules)			TCGA-NSCLC		BRACS	
TCBM	ARAM	Text	BACC	AUC	BACC	AUC
			80.1 ± 0.9	89.2 ± 0.9	55.9 ± 2.7	76.5 ± 1.8
✓			83.6 ± 4.2	91.7 ± 3.2	58.8 ± 7.7	79.2 ± 7.5
✓	✓		$\underline{85.3 \pm 5.8}$	$\underline{93.8 \pm 1.7}$	$\underline{62.3 \pm 5.9}$	$\underline{80.5 \pm 4.2}$
✓	✓	✓	$\mathbf{87.7 \pm 1.3}$	$\mathbf{94.0 \pm 0.8}$	$\mathbf{65.1 \pm 1.9}$	$\mathbf{83.5 \pm 0.9}$

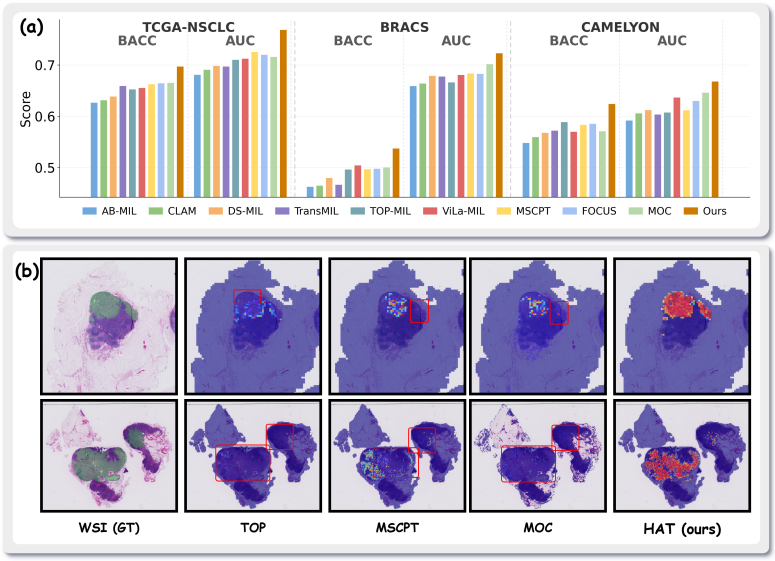

Fig. 3. (a) 8-shot results with PLIP encoder. (b) CAMELYON visualizations. Ground-truth regions are shown in green, predictions in red, and boxes indicate regions with incorrect predictions.

component (TCBM, ARAM, and Text). Adding TCBM improves AUC by +2.5 on TCGA-NSCLC and +2.7 on BRACS, suggesting that tissue-aware balancing reduces composition bias. Adding ARAM further brings gains of +2.1 on TCGA-NSCLC and +1.3 on BRACS, indicating improved robustness to appearance variation. Enabling text guidance yields the best overall performance and lower variance, showing that the two modules and language priors are complementary under few-shot supervision.

3.3 Visualization Results

Fig. 3b compares HAT with TOP [14], MSCPT [5], and MOC [20]. These methods show more prediction errors in challenging regions, whereas HAT aligns more closely with the ground-truth tumor regions, indicating improved robustness to inter-slide heterogeneity.

4 Conclusion

We identify tissue-composition variation and appearance perturbations as key sources of inter-slide heterogeneity that destabilize text-guided few-shot MIL. By combining tissue-aware balancing (TCBM) with robust low-magnification context constructed from stable 20× structures (ARAM), our framework consistently improves performance across datasets and few-shot settings.

5 Acknowledgments.

This work was supported by the National Key R&D Program of China (2024YFC3407800), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA0460203), the National Natural Science Foundation of China (82573343), and the Beijing Nova Program (20250484775).

6 Disclosure of Interests.

The authors have no competing interests to declare.

References

1. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
2. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 339–349. Springer (2021)
3. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., consortium, C., Hermsen, M., Manson, Q.F., 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)
4. Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., Chen, H.: Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15590–15600 (2025)
5. Han, M., Qu, L., Yang, D., Zhang, X., Wang, X., Zhang, L.: Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. IEEE Transactions on Medical Imaging (2025)
6. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

8. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons (1990)
9. Li, B., Li, Y., Eliceiri, K.W.: 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)
10. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
11. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: 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)
12. 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**(6), 555–570 (2021)
13. Pocock, J., Graham, S., Vu, Q.D., Jahanifar, M., Koohbanani, N.A., Shaban, M., Aepfelbacher, A., Rajpoot, N.: Tiatoolbox as an end-to-end library for advanced tissue image analytics. Communications Medicine **2**, 120 (2022). https://doi.org/10.1038/s43856-022-00186-5
14. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. Advances in Neural Information Processing Systems **36**, 67551–67564 (2023)
15. Schubert, E., Rousseeuw, P.J.: Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. Information Systems **101**, 101804 (2021). https://doi.org/10.1016/j.is.2021.101804
16. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
17. Shi, J., Li, C., Gong, T., Zheng, Y., Fu, H.: 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 (2024)
18. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Xiang, T., Li, Y., Zhang, Q., Li, X.: Moc: Meta-optimized classifier for few-shot whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 423–432. Springer (2025)
21. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating data processing and benchmarking of ai models for pathology. arXiv preprint arXiv:2502.06750 (2025)
22. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)

23. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)