Confusion-Aware Evidence Calibration for Few-shot Whole Slide Image Classification

Peng Li[1,2], Xuefeng Yan[1]✉, Gaoxiang Ma[3], Hao Tang[2], Mingqiang Wei[1], and Jing Qin[2]

[1] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
fzjm_313@nuaa.edu.cn
[2] The Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
[3] Clinical Metabolomics Center, China Pharmaceutical University, Nanjing, China

Abstract. While fine-grained prompt-to-patch matching has significantly advanced few-shot whole-slide image (WSI) classification, it introduces a critical vulnerability: the massive aggregation of local interactions inherently amplifies alignment noise. Such noise stems from both intrinsic pathological ambiguity, where shared local morphologies trigger spurious non-target responses, and extrinsic textual unreliability, arising from incomplete or noisy LLM-generated prompts. Together, these factors severely blur inter-class boundaries. To tackle this, we propose Confusion-Aware Evidence Calibration (CAEC), a robust framework designed to explicitly rectify alignment noise. Specifically, CAEC maintains a dynamic confusion memory module to capture recurring misleading activations, which guides a sample-adaptive calibrator to suppress spurious text-guided responses. Furthermore, we introduce a morphology evidence compensator that adaptively integrates pure visual features to anchor predictions when text guidance is weak. Evaluated on the TCGA-RCC and TCGA-NSCLC benchmarks, CAEC achieves state-of-the-art performance, demonstrating remarkable robustness in data-scarce settings.

Keywords: Computational pathology · Few-shot Learning · Confusion Calibration.

1 Introduction

Whole slide images (WSIs) serve as the clinical gold standard for tumor diagnosis, but their gigapixel resolution renders direct end-to-end deep learning computationally prohibitive. Therefore, Multiple Instance Learning (MIL) [8] has emerged as the dominant paradigm, mapping unannotated local patches to slide-level diagnostic outcomes [11, 16, 20, 24, 14]. While effective, standard MIL is data-intensive. This requirement conflicts with clinical realities, where privacy constraints and the natural rarity of specific cancer subtypes severely limit dataset sizes [5, 10]. Therefore, developing robust few-shot classification methods for WSIs remains a pressing imperative for real-world automated diagnosis.

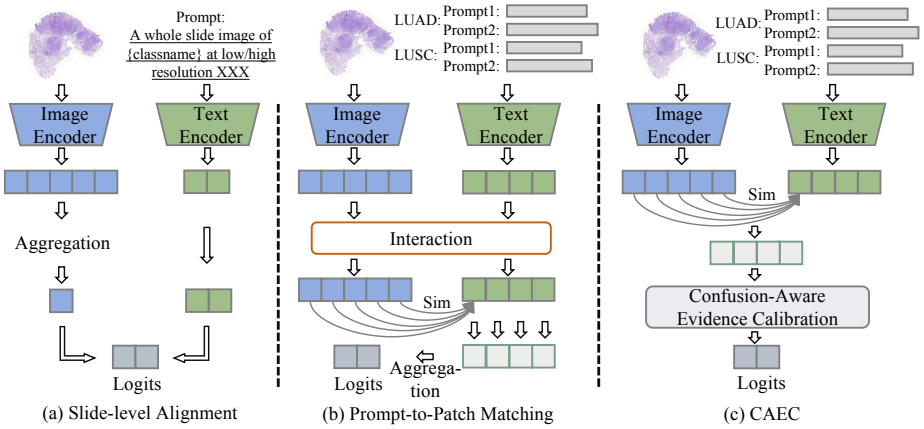


Fig. 1. Comparison of CAEC with existing methods. (a) Slide-level alignment aggregates patches globally. (b) Prompt-to-patch matching improves local sensitivity but accumulates mismatches. (c) CAEC performs confusion-aware evidence calibration on aggregated patch–prompt evidence.

Recent advances in vision–language models (VLMs) [19, 13, 2] have demonstrated strong few-shot transferability by learning aligned visual and textual representations from web-scale image–text pairs. In computational pathology, domain-specific VLMs [6, 7, 15] further improve this paradigm by leveraging large-scale pathology image–text supervision to produce more domain-aligned representations. Meanwhile, prompt learning [26, 27] has emerged as an efficient adaptation strategy for frozen pre-trained models, enabling task-specific customization through a small set of learnable prompts without extensive backbone fine-tuning. Building on these developments, early VLM-based WSI classification methods [18, 21, 3, 4] mainly integrate prompt learning into the MIL framework through holistic slide-level alignment, where patch features are first aggregated into a global slide representation and then matched with textual prompts (Fig. 1 (a)). Although effective, this macroscopic alignment may weaken fine-grained multimodal correspondence by smoothing out localized phenotypic variations. To better preserve local pathology cues, more recent methods have shifted toward a prompt-to-patch matching paradigm [25, 22], which directly matches fine-grained prompts with patch-level features and aggregates the resulting local similarities into slide-level predictions (Fig. 1 (b)).

Despite improved sensitivity to local morphology, this paradigm introduces a fundamental challenge in WSI analysis: *slide-level predictions are formed by aggregating thousands of patch–prompt interactions, during which alignment noise can accumulate and be amplified, thereby blurring inter-class boundaries.* Such noise arises from two factors. First, different pathological categories may share local morphological patterns, and prompts with different wording can encode overlapping visual cues. Consequently, non-target prompts may produce spuriously

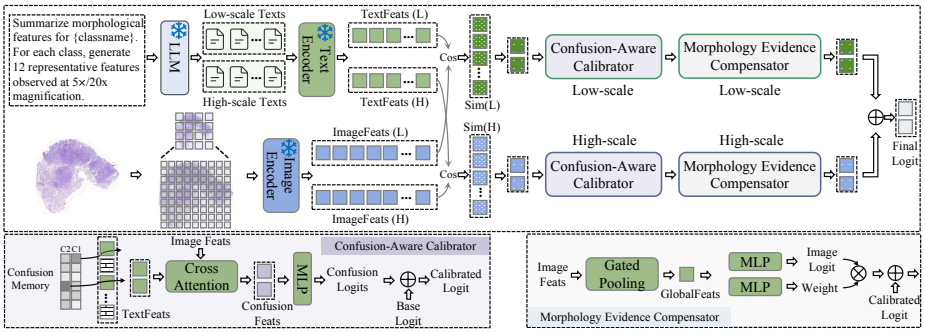


Fig. 2. Overview of CAEC. Multi-scale vision-language features are first matched to generate base logits. Then, a Confusion-Aware Calibrator queries the confusion memory to suppress alignment noise, while a Morphology Evidence Compensator extracts global image features to supplement pure visual evidence. Logits from both scales are finally fused for WSI classification. (Note: Modules across both scales share weights, except for the logit-predicting MLP in the Calibrator).

high responses during patch-level matching, which then accumulate through aggregation. Second, while LLM-generated prompt banks provide valuable semantic priors, their relevance is not uniform across samples. Incomplete, noisy, or sample-mismatched prompts can weaken vision–language correspondence and cause text-guided matching to underutilize discriminative visual evidence.

To address this challenge, we propose **CAEC** (**C**onfusion-**A**ware **E**vidence **C**alibration), a novel framework for few-shot WSI classification that explicitly calibrates aggregated patch-prompt evidence (Fig. 1(c)). Specifically, CAEC consists of three components: (1) a Confusion Memory Module, which maintains a category-to-prompt confusion matrix as a running prior of frequently over-activated non-target prompts; (2) a Confusion-Aware Calibrator, which extracts confusion patterns from visual–text interactions and predicts residual calibration logits to suppress spurious evidence; and (3) a Morphology Evidence Compensator, which adaptively supplements morphology-driven visual cues to compensate for uninformative language priors, thereby enhancing the overall robustness of the model.

2 Method

The overall pipeline of CAEC is illustrated in Fig. 2. Given an input WSI, we extract multi-scale visual patch features and encode LLM-generated morphological descriptions into scale-specific textual features. The framework then aligns visual tokens with textual features to compute base logits via semantic similarity aggregation. To mitigate alignment noise, a Confusion Memory Module dynamically records over-activated non-target prompts. Using this prior, the Confusion-Aware Calibrator performs cross-attention between visual features and the memory to predict residual calibration logits, suppressing spurious evidence in the

base predictions. Meanwhile, a Morphology Evidence Compensator yields pure image-based logits by aggregating patch tokens via gated pooling, compensating for potentially incomplete or mismatched language priors. Finally, these calibrated and compensated logits across both scales are fused to yield the final WSI prediction.

2.1 Feature Extraction and Base Logit Generation

Visual and Textual Feature Extraction. Following [22], a WSI is represented as multi-scale bags of patch embeddings: $X^{(s)} = \{\mathbf{x}_n^{(s)}\}_{n=1}^{N_s}$, where $s \in \{h, l\}$ denotes high/low magnifications (e.g., 20× and 5×). These D-dimensional embeddings are extracted offline via a frozen pathology VLM. Meanwhile, we generate scale-specific prompts by querying the LLMs with a predefined template (see Fig. 2). Following CoOp [27], L learnable context tokens $\mathbf{v}_1, \ldots, \mathbf{v}_L$ are prepended to each raw prompt $\mathcal{P}_j^{(s)}$ and encoded by a frozen text transformer to yield prompt embeddings: $\mathbf{t}_j^{(s)} = f_t\Big([\mathbf{v}_1, \ldots, \mathbf{v}_L;\ \text{tokens}(\mathcal{P}_j^{(s)})]\Big) \in \mathbb{R}^D$.

Base Logits Generation. Base logits $z_c^{(s)}$ are computed via cosine similarities $\text{sim}_{n,j}^{(s)}$ between visual tokens $X^{(s)}$ and prompt embeddings $T^{(s)} = \{\mathbf{t}_j^{(s)}\}_{j=1}^{Q_s}$. Following [22], we average the top-k_s scoring patches (whose indices are denoted as Ω_j) for each prompt, and then average across all prompts $\mathcal{P}_c^{(s)}$ belonging to class c, scaled by the standard VLM logit scalar γ:

$$z_c^{(s)} = \frac{\gamma}{k_s |\mathcal{P}_c^{(s)}|} \sum_{j \in \mathcal{P}_c^{(s)}} \sum_{n \in \Omega_j} \text{sim}_{n,j}^{(s)}. \tag{1}$$

2.2 Confusion Memory Module

As discussed, aggregating prompt-patch interactions can amplify alignment noise due to shared local morphologies, blurring inter-class boundaries. To mitigate this, we introduce a persistent confusion memory matrix $\mathbf{M}^{(s)} \in \mathbb{R}^{C \times Q_s}$ to track historically misleading prompt activations and use it to calibrate logits. The element $M_{c,j}^{(s)}$ quantifies the accumulated confusion, i.e., the degree to which the j-th prompt erroneously aligns with class c. We initialize $M_{c,j}^{(s)}$ to 0 if class c is the true target of prompt j, and to a small constant ω otherwise.

The memory is dynamically updated based on aggregated prompt responses $q_j^{(s)}$. If a non-target prompt frequently yields high aggregated scores for a specific category, it serves as a direct indicator of cross-class morphological confusion. For a training slide with ground-truth class y, we establish a *class anchor* $a_y^{(s)} = \frac{1}{r} \sum_{q \in \mathcal{S}_y^{(s)}} q$ by averaging its top-r target responses (denoted as set $\mathcal{S}_y^{(s)}$). We then evaluate the confusion of non-target prompts ($\phi^{(s)}(j) \neq y$) against this anchor. Intuitively, non-target responses approaching or exceeding $a_y^{(s)}$ act as misleading signals. The update strength $u_j^{(s)}$ is thus formulated as

$u_j^{(s)} = \max\Big(0,\ q_j^{(s)} - a_y^{(s)} + m\Big)$, where m is a margin. The memory row for class y is then updated via an exponential moving average with a learning rate η:

$$M_{y,j}^{(s)} \leftarrow (1-\eta) M_{y,j}^{(s)} + \eta\ u_j^{(s)}. \qquad \forall j \text{ s.t. } \phi^{(s)}(j) \neq y \tag{2}$$

2.3 Confusion-Aware Calibrator

Given the constructed confusion memory, directly applying the global confusion scores as static penalties is sub-optimal, as it ignores whether these confusing patterns actually manifest in the current slide. Instead of image-agnostic suppression, we use these learned confusion priors to explicitly aggregate their corresponding visual features, thereby generating a sample-adaptive calibration logit based on the actual visual evidence of confusion.

For class c at scale $s \in \{h, l\}$, we retrieve the top-K confusing prompts $\{\mathbf{t}_{c,k}^{(s)}\}_{k=1}^{K}$ from $\mathbf{M}^{(s)}$. Using these as queries and patch embeddings $X^{(s)}$ as keys/values, we employ cross-attention to aggregate the confusion-conditioned visual feature $\mathbf{f}_{c,k}^{(s)}$. The feature $\mathbf{f}_{c,k}^{(s)}$ is processed by an MLP $g_s(\cdot)$ to compute a logit offset. We then average these offsets across the K prompts and apply a hyperbolic tangent activation function scaled by a learnable parameter α_s:

$$\delta_c^{(s)} = \alpha_s \tanh\left(\frac{1}{K}\sum_{k=1}^{K} g_s(\mathbf{f}_{c,k}^{(s)})\right). \tag{3}$$

The term $\delta_c^{(s)}$ functions as a sample-adaptive calibration signal. Finally, the base prediction $\mathbf{z}_c^{(s)}$ is refined by $\hat{\mathbf{z}}_c^{(s)} = \mathbf{z}_c^{(s)} + \delta_c^{(s)}$.

2.4 Morphology Evidence Compensator

As discussed, incomplete or mismatched LLM prompts can weaken the correspondence between visual and textual cues, causing text-guided matching to underutilize discriminative visual evidence. To address this, we introduce a Morphology Evidence Compensator to adaptively integrate pure morphological evidence based on its representation reliability. Specifically, at each scale s, we aggregate the visual patch tokens $\mathbf{x}_n^{(s)}$ into a global representation $\mathbf{g}^{(s)}$ via gated attention pooling [21, 25]. This feature is mapped to image-based logits $\mathbf{v}^{(s)} = h_{\text{img}}(\mathbf{g}^{(s)}) \in \mathbb{R}^C$ via an MLP h_{img}. Since pure visual representation quality varies across heterogeneous WSIs, naively adding these logits could introduce bias. Thus, we predict a slide-specific reliability gate $r^{(s)} = \sigma\big(h_{\text{gate}}(\mathbf{g}^{(s)})\big) \in (0, 1)$ to dynamically modulate the visual contribution. Finally, the gated visual logits are fused with the memory-calibrated logits $\hat{\mathbf{z}}^{(s)}$ across both scales for the final WSI prediction: $\mathbf{z} = \sum_{s\in\{h,l\}} \left(\hat{\mathbf{z}}^{(s)} + r^{(s)}\mathbf{v}^{(s)}\right)$.

2.5 Training Objective

We supervise the final logits with standard cross-entropy loss. To stabilize training and encourage each component to be individually discriminative, we also apply auxiliary losses to the scale-wise calibrated logits and the compensated logits:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(\mathbf{z}, y) \; + \; \frac{1}{2} \sum_{s \in \{h,l\}} \Big(\mathcal{L}_{\mathrm{CE}}(\hat{\mathbf{z}}^{(s)}, y) + \mathcal{L}_{\mathrm{CE}}(\mathbf{v}^{(s)}, y) \Big). \tag{4}$$

3 Experiments

Datasets and Evaluation Protocol. We evaluate our method on two TCGA benchmarks: TCGA-RCC (860 WSIs) and TCGA-NSCLC (1,042 WSIs) [4]. Following [21], datasets are split 4:3:3 into training, validation, and test sets. For N-shot training, we randomly sample $N \in \{4, 8, 16\}$ slides per class. We report Accuracy (ACC), Area Under the Curve (AUC), and Macro-F1 score using 5-fold cross-validation, where the validation set is used for early stopping and model selection, and final metrics are computed on the test set.

Implementation Details. Following [22], WSIs are cropped into patches at 5× and 20× magnifications, with features extracted via TRIDENT [23] using the frozen VLM CONCH [15]. Scale-specific descriptions generated by GPT-4o [1] are prepended with $L = 16$ learnable context tokens. For base logits, we aggregate the top-100 patch similarities ($k_s = 100$) per prompt across scales. The confusion memory w, initialized to 10^{-5}, is updated via EMA ($\eta = 0.02$, $m = 0.01$) using the top-2 target responses ($r = 2$) as anchors to retrieve the top-3 confusing prompts ($K = 3$) for fine-grained calibration. The framework is trained end-to-end in PyTorch [17] on a single RTX 3090 GPU using Adam [9] (LR: 10^{-4}, weight decay: 10^{-5}, batch size: 1) for up to 80 epochs, with a 10-epoch early stopping patience.

Comparisons with State-of-the-Art. Table 1 presents the quantitative results, demonstrating our framework's state-of-the-art performance across most data-scarce settings. Notably, on the TCGA-NSCLC dataset, our method exhibits remarkable resilience. In the extreme 4-shot setting, it attains 93.64% AUC and 84.78% Macro-F1, outperforming the second-best method by 3.55% and 3.37%, respectively. For 4-shot TCGA-RCC, our method achieves the highest AUC (97.18%). Although F1 and ACC are slightly lower than HiVE, the superior threshold-independent AUC indicates stronger overall feature separability. These consistent improvements underscore the effectiveness of our proposed framework. On one hand, the model explicitly calibrates aggregated patch-prompt interactions to suppress spurious responses. Concurrently, when sample-mismatched prompts weaken vision-language correspondence, the morphology evidence compensator adaptively integrates pure morphological features. Working in synergy, these two mechanisms effectively sharpen inter-class boundaries and ensure overall diagnostic robustness.

[4] https://portal.gdc.cancer.gov/

Table 1. Results on TCGA-RCC and TCGA-NSCLC under 4, 8, 16-shot settings. The best results are in **bold**, and the second-best results are underlined.

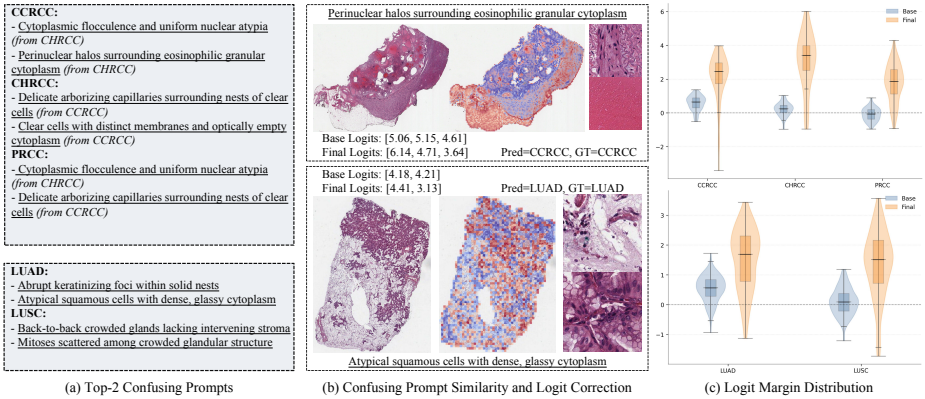
Dataset	Methods	TCGA-RCC			TCGA-NSCLC		
		AUC	F1	ACC	AUC	F1	ACC
16-shot	CLAMSB [16]	$98.48_{\pm 0.47}$	$90.09_{\pm 0.98}$	$91.32_{\pm 1.09}$	$95.37_{\pm 0.50}$	$89.27_{\pm 1.86}$	$89.29_{\pm 1.84}$
	TransMIL [20]	$97.88_{\pm 0.66}$	$87.74_{\pm 0.99}$	$89.61_{\pm 0.96}$	$93.94_{\pm 2.25}$	$86.36_{\pm 3.93}$	$86.41_{\pm 3.87}$
	WiKG [12]	$97.44_{\pm 0.49}$	$87.16_{\pm 1.09}$	$89.53_{\pm 0.65}$	$89.91_{\pm 1.87}$	$82.25_{\pm 2.69}$	$82.31_{\pm 2.65}$
	ViLa-MIL [21]	$95.46_{\pm 2.03}$	$87.20_{\pm 3.59}$	$89.22_{\pm 2.42}$	$90.51_{\pm 4.35}$	$84.70_{\pm 5.85}$	$84.87_{\pm 5.52}$
	MSCPT [4]	$97.65_{\pm 0.85}$	$89.30_{\pm 1.20}$	$91.53_{\pm 1.15}$	$91.73_{\pm 2.80}$	$84.81_{\pm 1.95}$	$83.55_{\pm 3.00}$
	FOCUS [3]	$98.25_{\pm 0.40}$	$91.06_{\pm 1.74}$	$92.25_{\pm 1.45}$	$95.17_{\pm 0.73}$	$87.64_{\pm 3.49}$	$87.69_{\pm 3.41}$
	MAPLE [25]	$98.45_{\pm 0.57}$	$90.60_{\pm 1.77}$	$91.94_{\pm 1.24}$	$89.89_{\pm 9.79}$	$83.86_{\pm 9.18}$	$83.91_{\pm 9.12}$
	HiVE [22]	$98.75_{\pm 0.33}$	$92.23_{\pm 1.30}$	$93.19_{\pm 1.27}$	$95.79_{\pm 0.82}$	$90.12_{\pm 1.12}$	$90.13_{\pm 1.12}$
	Ours	$\mathbf{98.89_{\pm 0.20}}$	$\mathbf{92.73_{\pm 1.01}}$	$\mathbf{93.88_{\pm 0.79}}$	$\mathbf{96.27_{\pm 0.93}}$	$\mathbf{90.64_{\pm 2.03}}$	$\mathbf{90.64_{\pm 2.02}}$
8-shot	CLAMSB [16]	$97.77_{\pm 0.52}$	$87.80_{\pm 1.40}$	$89.84_{\pm 1.02}$	$95.91_{\pm 1.14}$	$90.19_{\pm 1.47}$	$90.19_{\pm 1.47}$
	TransMIL [20]	$97.34_{\pm 0.53}$	$86.36_{\pm 1.86}$	$88.76_{\pm 1.41}$	$93.37_{\pm 1.53}$	$86.77_{\pm 2.97}$	$86.79_{\pm 2.95}$
	WiKG [12]	$96.56_{\pm 0.67}$	$83.71_{\pm 1.22}$	$86.20_{\pm 1.50}$	$90.37_{\pm 1.25}$	$82.72_{\pm 1.78}$	$82.76_{\pm 1.76}$
	ViLa-MIL [21]	$94.72_{\pm 0.96}$	$84.43_{\pm 1.75}$	$87.52_{\pm 1.28}$	$90.40_{\pm 2.56}$	$81.71_{\pm 3.64}$	$81.92_{\pm 3.57}$
	MSCPT [4]	$96.80_{\pm 1.15}$	$86.61_{\pm 1.85}$	$87.83_{\pm 1.75}$	$92.74_{\pm 2.10}$	$85.20_{\pm 2.50}$	$85.25_{\pm 2.45}$
	FOCUS [3]	$97.75_{\pm 0.65}$	$88.95_{\pm 1.13}$	$90.31_{\pm 1.10}$	$94.91_{\pm 2.36}$	$87.41_{\pm 2.24}$	$87.44_{\pm 2.24}$
	MAPLE [25]	$97.69_{\pm 0.56}$	$89.23_{\pm 1.96}$	$90.47_{\pm 2.15}$	$86.87_{\pm 13.24}$	$80.21_{\pm 11.67}$	$80.31_{\pm 11.62}$
	HiVE [22]	$98.07_{\pm 0.50}$	$89.45_{\pm 2.43}$	$91.47_{\pm 1.55}$	$95.77_{\pm 0.96}$	$90.31_{\pm 2.77}$	$90.32_{\pm 2.76}$
	Ours	$\mathbf{98.64_{\pm 0.23}}$	$\mathbf{91.70_{\pm 1.42}}$	$\mathbf{93.02_{\pm 1.39}}$	$\mathbf{96.13_{\pm 1.30}}$	$\mathbf{90.43_{\pm 2.38}}$	$\mathbf{90.45_{\pm 2.37}}$
4-shot	CLAMSB [16]	$95.36_{\pm 3.61}$	$82.85_{\pm 6.81}$	$85.35_{\pm 6.92}$	$86.97_{\pm 9.02}$	$77.75_{\pm 8.51}$	$77.88_{\pm 8.49}$
	TransMIL [20]	$94.28_{\pm 3.73}$	$81.89_{\pm 4.78}$	$83.80_{\pm 5.20}$	$80.69_{\pm 7.30}$	$72.69_{\pm 6.44}$	$72.95_{\pm 6.10}$
	WiKG [12]	$93.31_{\pm 2.86}$	$79.19_{\pm 5.02}$	$82.33_{\pm 4.76}$	$81.06_{\pm 4.03}$	$70.65_{\pm 3.87}$	$71.22_{\pm 3.21}$
	ViLa-MIL [21]	$88.82_{\pm 4.58}$	$74.58_{\pm 11.22}$	$79.30_{\pm 7.91}$	$77.94_{\pm 10.42}$	$71.89_{\pm 6.96}$	$71.99_{\pm 6.87}$
	MSCPT [4]	$95.30_{\pm 2.85}$	$84.20_{\pm 4.10}$	$86.36_{\pm 3.90}$	$86.60_{\pm 4.90}$	$75.41_{\pm 6.20}$	$79.03_{\pm 6.10}$
	FOCUS [3]	$96.47_{\pm 2.55}$	$84.84_{\pm 5.68}$	$86.59_{\pm 5.64}$	$88.17_{\pm 6.44}$	$78.31_{\pm 5.94}$	$78.46_{\pm 5.92}$
	MAPLE [25]	$94.81_{\pm 6.66}$	$84.62_{\pm 9.14}$	$85.81_{\pm 10.13}$	$88.14_{\pm 6.87}$	$79.18_{\pm 6.04}$	$79.27_{\pm 5.97}$
	HiVE [22]	$96.84_{\pm 0.84}$	$\mathbf{87.27_{\pm 2.97}}$	$\mathbf{89.26_{\pm 2.87}}$	$90.09_{\pm 4.87}$	$81.41_{\pm 4.46}$	$81.47_{\pm 4.41}$
	Ours	$\mathbf{97.18_{\pm 1.30}}$	$86.81_{\pm 3.54}$	$88.53_{\pm 3.55}$	$\mathbf{93.64_{\pm 0.71}}$	$\mathbf{84.78_{\pm 1.71}}$	$\mathbf{84.87_{\pm 1.61}}$

Ablation Study. Table 2 details the ablation study on TCGA-RCC. The baseline struggles in extreme low-data regimes (82.14% F1 at 4-shot). Integrating the Confusion-Aware Calibrator (CC) yields a +2.76% F1 improvement, validating its ability to suppress spurious patch-prompt responses. Furthermore, the Morphology Evidence Compensator (MEC) enhances robustness by providing morphology-driven visual cues when language priors are uninformative. By integrating these capabilities, the full framework effectively rectifies inter-class confusion and overcomes the fundamental challenge of alignment noise accumulation, achieving optimal performance across all settings.

Interpretability and Visualization. Fig. 3(a) displays the top-2 confusing prompts autonomously mined by our dynamic memory, such as CCRCC confusing with CHRCC's "perinuclear halos" or LUAD with LUSC's "keratinizing foci". Fig. 3(b) presents two representative WSIs where these confusing prompts yield spuriously high similarities with visual patches (exceeding or approaching the

Table 2. Ablation study of individual components on the TCGA-RCC dataset.

Model	4-shot		8-shot		16-shot	
	AUC	F1	AUC	F1	AUC	F1
Baseline	$94.90_{\pm 1.98}$	$82.14_{\pm 1.20}$	$98.01_{\pm 0.36}$	$86.96_{\pm 2.38}$	$98.29_{\pm 0.37}$	$90.34_{\pm 0.70}$
Baseline+CC	$96.01_{\pm 1.60}$	$84.90_{\pm 0.54}$	$98.20_{\pm 0.37}$	$89.05_{\pm 3.14}$	$98.33_{\pm 0.46}$	$91.05_{\pm 1.58}$
Baseline+MEC	$96.38_{\pm 1.32}$	$85.04_{\pm 3.27}$	$98.21_{\pm 0.22}$	$90.01_{\pm 1.21}$	$98.04_{\pm 0.34}$	$91.31_{\pm 1.23}$
Ours	$97.18_{\pm 1.30}$	$86.81_{\pm 3.54}$	$98.64_{\pm 0.23}$	$91.70_{\pm 1.42}$	$98.89_{\pm 0.20}$	$92.73_{\pm 1.01}$



(a) Top-2 Confusing Prompts (b) Confusing Prompt Similarity and Logit Correction (c) Logit Margin Distribution

Fig. 3. Interpretability of the proposed method. (a) Top-2 confusing prompts captured by the dynamic memory. (b) Confusing Prompt Similarity and Logit Correction. (c) Logit Margin Distribution (True Logit − Max Other Logit), demonstrating enhanced inter-class separability after calibration.

true class prompts), leading to incorrect baseline predictions (see Base Logits). Remarkably, our method recognizes this ambiguity and explicitly rectifies the prediction, restoring the true class as the dominant Final Logits. Furthermore, Fig. 3(c) illustrates the logit margin distribution. The clear upward shift from base to final margins confirms that our method effectively rectifies ambiguous samples across the decision boundary, globally sharpening inter-class separation.

4 Conclusion

In this paper, we propose Confusion-Aware Evidence Calibration (CAEC), a robust framework designed to tackle the inherent amplification of alignment noise in fine-grained vision-language WSI classification. To overcome the severe blurring of inter-class boundaries, our method unites two complementary mechanisms. First, a sample-adaptive calibrator, guided by a dynamic confusion memory, suppresses spurious responses caused by intrinsic pathological ambiguity. Concurrently, a morphology evidence compensator integrates pure visual features to anchor predictions, overcoming extrinsic textual unreliability. By explicitly mitigating these dual sources of noise, CAEC sharpens inter-class bound-

aries. Extensive experiments on the TCGA-RCC and TCGA-NSCLC datasets demonstrate that CAEC consistently achieves state-of-the-art performance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. T2322012, No. 62572240), and a project under the Collaborative Research with World-leading Research Groups scheme of The Hong Kong Polytechnic University (project no. G-SACF).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. 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)
4. 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)
5. Holub, P., Müller, H., Bíl, T., Pireddu, L., Plass, M., Prasser, F., Schlünder, I., Zatloukal, K., Nenutil, R., Brázdil, T.: Privacy risks of whole-slide image sharing in digital pathology. Nature Communications **14**(1), 2577 (2023)
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. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36**, 37995–38017 (2023)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Lee, J., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J.R., Chung, W.K., Weng, C.: Deep learning for rare disease: A scoping review. Journal of biomedical informatics **135**, 104227 (2022)
11. 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)

12. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: 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 (2024)
13. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
14. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: 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)
15. 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)
16. 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)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. 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)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
20. 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)
21. 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)
22. Wong, B., Kim, J.W., Fu, H., Yi, M.Y.: Few-shot learning from gigapixel images via hierarchical vision-language alignment and modeling. arXiv preprint arXiv:2505.17982 (2025)
23. 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)
24. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: 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)
25. Zhou, J., Shao, W., Yue, Y., Mu, W., Wan, P., Zhu, Q., Zhang, D.: Maple: Multi-scale attribute-enhanced prompt learning for few-shot whole slide image classification. arXiv preprint arXiv:2509.25863 (2025)
26. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)

27. 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)