Spectral-Semantic Consensus for Few-Shot Whole Slide Image Classification

Hanyu Du[1], Peilin Zhou[1], Weizhuo Shi[1], Zhikai Wei[1], Yajie Chen[4]✉, Bo Du[1,2,3], and Yongchao Xu[1,2,3]✉

[1] School of Computer Science, Wuhan University
[2] National Engineering Research Center for Multimedia Software, Wuhan University
[3] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Institute of Artificial Intelligence, Wuhan University
{hanyu.du, yongchao.xu}@whu.edu.cn
[4] School of Computer Science and Technology, Wuhan University of Science and Technology
yajiechen@wust.edu.cn

Abstract. Few-shot Whole Slide Image (WSI) classification is challenged by extreme label scarcity and gigapixel resolution. While Vision–Language Models (VLMs) offer a promising route via text-guided retrieval, they often suffer from discriminative sparsity: rigid selection can either overlook key diagnostic regions or include excessive background. We propose Spectral-Semantic Consensus (SSC), a patch selection framework that couples distributed budget prompt retrieval with spectral knowledge projection selection. SSC consists of two complementary modules: (1) a Distributed Budget Prompt Consensus (DBPC) module that employs shared learnable prompts with distributed selection budget to improve the coverage of heterogeneous, diagnostically relevant patches; (2) a Spectral Knowledge Projection (SKP) module that selects patches by their alignment with subspace spanned by principal components (spectral components) derived from textual priors, thus reducing task-irrelevant regions. To adapt aggregation of the two modules, we further introduce an Uncertainty-guided Fusion mechanism that calibrates module contributions using retrieval uncertainty, emphasizing spectral alignment when semantic matching becomes unreliable. By aligning visual evidence selection with the spectral geometry of domain knowledge, SSC provides a robust and principled paradigm for data-efficient computational pathology. Experiments on TCGA-NSCLC, CAMELYON16, and UBC-OCEAN demonstrate consistent and substantial gains over strong baselines in few-shot regimes. Code is available at: https://github.com/Hanyu1124/ssc.

Keywords: Few-Shot Learning · Whole Slide Image · Multiple instance learning.

1 Introduction

Computational pathology increasingly relies on deep learning to analyze Whole Slide Images (WSIs) for diagnosis and biomarker discovery. However, WSIs are

gigapixel-scale, making dense annotations costly and motivating the widespread use of Multiple Instance Learning (MIL), where supervision is available only at the slide level [3,10,4,8,19]. In clinically relevant scenarios such as rare diseases and novel subtypes, WSI classification further becomes a few-shot weakly supervised problem [17,27,18,6]. This setting is particularly challenging due to the discriminative sparsity: diagnostically informative regions often occupy less than 5% of a slide. Thus, evidence selection must be both precise and comprehensive under a low signal-to-noise ratio.

VLMs have recently been introduced to inject textual priors into WSI analysis [7,22,24,26]. Works such as QUILT [9] and MI-Zero [14] establish patch–text alignment via large-scale pretraining. Later methods (e.g., ViLa-MIL [22], FOCUS [7]) adapt these priors to WSIs by handling resolution heterogeneity and redundancy. However, current text-guided selection remains limited by two issues. First, coarse-grained semantic collapsing: prompt sets are often averaged into a single class prototype, which suppresses semantic variance and underretrieves heterogeneous pathological patterns. Second, rigid semantic matching: selecting patches solely by maximal text–image similarity tends to miss diagnostically relevant regions that deviate from canonical "textbook" appearances, leading to high false negatives or background contamination.

To address these gaps, we propose Spectral-Semantic Consensus (SSC), a text-guided patch selection framework for few-shot WSI classification. Considering independent semantics in each prompt and visual appearance deviation from formal pathological descriptions, we introduce two complementary modules to identify informative patches under discriminative sparsity: a learnable prompt guided patch selection with distributed selection budget for semantic coverage and a spectral projection module for knowledge subspace alignment consistency. The selected patches are integrated by an uncertainty-guided fusion strategy that adaptively balances the two modules based on retrieval uncertainty.

Our contributions are threefold:

1. Patch selection based on Distributed Budget Prompt Consensus (DBPC). We introduce DBPC, which uses shared learnable prompts with distributed budget and a consensus-based retrieval strategy to broaden evidence coverage and reduce sensitivity to any single query under few-shot supervision.

2. Patch selection based on Spectral Knowledge Projection (SKP). We propose SKP, which selects patches by the score of their projection onto the subspace spanned by spectral components derived from textual knowledge.

3. Uncertainty-guided Fusion of DBPC and SKP. An uncertainty-guided fusion strategy balances two modules based on DBPC retrieval uncertainty of the selected patches. SSC consistently improves over strong MIL and VLM baselines in label-scarce regimes.

2 Methodology

We study few-shot WSI classification under the Multiple Instance Learning (MIL) setting. A gigapixel slide $X \in \mathbb{R}^{H\times W\times 3}$ is partitioned into N non-

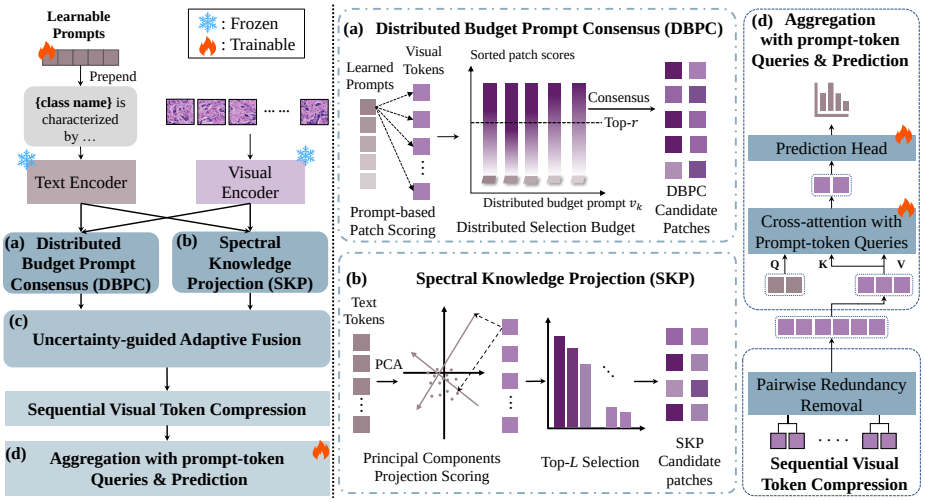

Fig. 1. Overview of SSC. (a) DBPC retrieves diverse patches with distributed selection budget over prompts. (b) SKP selects patches via projection onto a principal semantic subspace of text tokens. (c) Uncertainty-guided Fusion balances DBPC and SKP selections. (d) Cross-modal attention aggregates compressed tokens for slide prediction.

overlapping tissue patches $\{x_i\}_{i=1}^{N}$, where only slide-level labels $Y \in \{0, \ldots, C-1\}$ is given. Each patch is encoded by a frozen pretrained backbone $f(\cdot)$ to obtain a bag of embeddings $B = \{b_i\}_{i=1}^{N}$, where $b_i = f(x_i) \in \mathbb{R}^D$. Our goal is to identify a compact set of informative patches and aggregate them to predict Y.

SSC is a text-guided patch slection for WSI classification (Figure 1). Given patch embeddings B, our method selects informative tokens via two parallel routes: the Distributed Budget Prompt Consensus module and the Spectral Knowledge Projection module. These outputs are adaptively fused by the uncertainty of DBPC selected patches, and further compressed by SVTC [21], then aggregated with the prompt queries through cross-modal attention to produce the slide-level prediction.

2.1 Distributed Budget Prompt Consensus (DBPC)

In few-shot MIL, the main challenge is unstable supervision: a single retrieval query can overfit spurious textures and miss heterogeneous diagnostic cues. DBPC addresses this by evenly allocating the selection budget across multiple prompts, followed by aggregating patch scores over the selected prompts (consensus). Reducing to single-prompt similarity when only one prompt selects the patch, improving evidence coverage and training stability. We construct a text sequence $P_{=}[t_{\mathrm{SOT}}, \mathbf{V}, \mathbf{E}_{\mathrm{text}}]$, where $\mathbf{V} = \{v_k\}_{k=1}^{K}$ are K learnable prompts, $\mathbf{E}_{\mathrm{text}}$ is the fixed token sequence of a class-specific description. The frozen text

encoder $\mathcal{T}$ outputs token features $T = \mathcal{T}(P) \in \mathbb{R}^{M\times D}$. We use the K outputs corresponding to $\mathbf{V}$ as query tokens $Q = \{q_k\}_{k=1}^{K} \in \mathbb{R}^{K\times D}$.

Given patch features $B = \{b_i\}_{i=1}^{N}$, we compute cosine similarities $S_{i,k} = \cos(b_i, q_k)$. For each query q_k, we select the top-r patches with

$$r = \max\left(1, \left\lfloor \frac{L}{K} \right\rfloor\right). \tag{1}$$

We compute a per-patch consensus confidence by averaging similarities over the selected prompts:

$$s_i^{\mathrm{DBPC}} = \frac{\sum_{k=1}^{K} \mathbf{1}_{i,k} S_{i,k}}{\sum_{k=1}^{K} \mathbf{1}_{i,k} + \epsilon}, \tag{2}$$

where $\mathbf{1}_{i,k}$ indicates whether patch i is selected by prompt k. If a patch j is not selected by any prompt, we set $s_j^{\mathrm{DBPC}} = 0$. Each prompt selects r patches and gives these patches confidence scores $S_{i,k}$. In the adaptive fusion stage, patches in DBPC are selected based on the consensus confidence score s_i^{DBPC} ranking. The total number of patches scored at this stage is $\leq L$. The learnable prompts, also denoted Q, serve as queries reused for subsequent cross-modal aggregation.

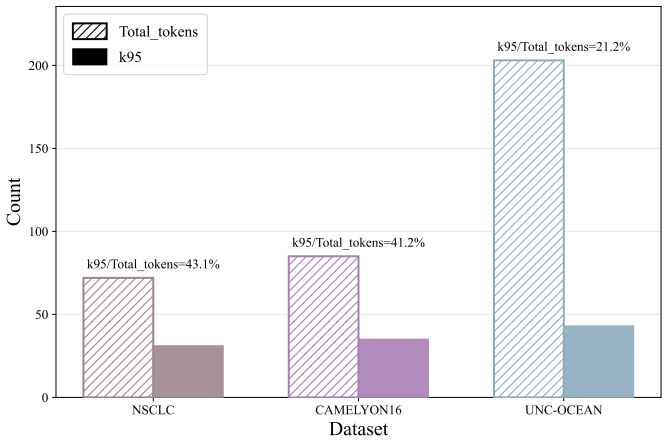

Fig. 2. Spectral heterogeneity of text embeddings. Comparison of the principal components of text-token embeddings across three datasets. For each dataset, the left bar denotes the total number of tokens (Total tokens), while the right bar denotes the number of principal components required to explain 95% of the variance (k95). The annotation above reports the ratio $\frac{k95}{\text{Total tokens}}$.

2.2 Spectral Knowledge Projection (SKP)

DBPC relies on explicit text–image similarity and may respond strongly to background regions or artifacts. SKP complements DBPC with a spectral-based

patch selection: it scores patches by their projection energy onto a principal semantic subspace induced by knowledge text tokens. This design is motivated by the anisotropy of pretrained text embeddings, which concentrates in a low-dimensional manifold [5,11]. As shown in Figure 2, pathology text embeddings exhibit distinct concentration patterns, motivating us to model textual knowledge priors with a low-dimensional spectral structure. The resulting subspace serves as a global structural prior, reducing task-irrelevant responses without requiring keyword-level matches.

Knowledge principal subspace. We build a content–token mask on text tokens by discarding PAD/SOT/EOT while keeping learnable context tokens. Let the retained tokens be $X = [x_1, \ldots, x_{M_t}]^\top \in \mathbb{R}^{M_t \times D}$.

To stabilize cross-modal matching, we whiten X before getting the principal components. With mean $\mu = \frac{1}{M_t}\sum_{m=1}^{M_t} x_m$ and covariance $\Sigma = \frac{1}{\max(1, M_t-1)}(X - \mathbf{1}\mu^\top)^\top(X - \mathbf{1}\mu^\top)$, we perform eigendecomposition $\Sigma = H\Lambda H^\top$ and define the whitening matrix:

$$W = H\Lambda_\epsilon^{-1/2}H^\top, \qquad \Lambda_\epsilon^{-1/2} = \mathrm{diag}\Big((\max(\lambda_1, \epsilon))^{-1/2}, \ldots, (\max(\lambda_D, \epsilon))^{-1/2}\Big),$$

where ϵ clamps small eigenvalues. The whitened tokens are $X_w = (X - \mathbf{1}\mu^\top)W$. We apply SVD, $X_w = USV^\top$, and retain the top-d_s components $V_\rho \in \mathbb{R}^{D \times d_s}$, which explain a variance ratio ρ.

Patch scoring and selection. Given patch embeddings $\{b_i\}_{i=1}^N$, SKP scores each patch by its projection energy onto the subspace spanned by V_ρ:

$$\tilde{b}_i = (b_i - \mu)W, \qquad s_i^{\mathrm{SKP}} = \|\tilde{b}_i V_\rho\|_2^2. \tag{3}$$

Patches with larger s_i^{SKP} are more aligned with the principal textual semantics. In this module, patches are ranked by s_i^{SKP}, and we keep the top-L patches.

2.3 Uncertainty-guided Fusion, Aggregation, and Objective

We adaptively fuse DBPC and SKP based on the DBPC retrieval uncertainty. Let $\mathbf{p} = [s_i^{\mathrm{DBPC}}]_{i=1}^N$ be the raw DBPC scores over N patches and $\boldsymbol{\pi} = \mathrm{softmax}(\mathbf{p})$. The normalized entropy is

$$\mathcal{H}(\boldsymbol{\pi}) = -\frac{1}{\log N}\sum_{i=1}^{N} \pi_i \log(\pi_i + \epsilon). \tag{4}$$

This yields module ratios $\alpha_{\mathrm{DBPC}} = 1 - \mathcal{H}(\boldsymbol{\pi})$ and $\alpha_{\mathrm{SKP}} = \mathcal{H}(\boldsymbol{\pi})$. We select $\lfloor \alpha_{\mathrm{DBPC}} L \rfloor$ top-ranked patches by s_i^{DBPC} and $\lfloor \alpha_{\mathrm{SKP}} L \rfloor$ by s_i^{SKP}. We take the union of the two sets and thus keep at most L patches. SVTC then compresses the resulting tokens [21].

We use Q as queries to attend to the selected visual tokens. A linear classifier produces the slide logits. The loss is given by

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{div}}, \tag{5}$$

Table 1. Balanced ACC (%) in few-shot WSI classification (4/8/16-shot) on TCGA-NSCLC, CAMELYON16, and UBC-OCEAN. Best and second-best results are marked in bold and underline.

Datasets	Methods	Balanced ACC (%)		
		4-shot	8-shot	16-shot
TCGA-NSCLC	TRANS-MIL[20]	71.2±6.1	77.2±4.2	84.6±3.9
	ILRA-MIL[25]	69.3±10.2	74.8±7.1	84.3±3.1
	WIKG-MIL[12]	66.7±9.6	76.4±3.5	79.8±7.2
	Vila-MIL[22]	78.0±8.2	84.2±5.0	**87.9±3.2**
	AMD-MIL[13]	70.9±8.8	77.9±5.5	85.1±4.9
	TDA-MIL [19]	71.7±8.4	80.9±4.4	86.6±4.7
	SSC (Ours)	**80.1±7.2**	**84.9±2.7**	87.8±2.8
CAMELYON16	TRANS-MIL[20]	55.1±4.7	58.0±5.5	68.0±13.3
	ILRA-MIL[25]	53.6±4.2	65.2±10.6	76.2±10.0
	WIKG-MIL[12]	53.0±4.7	65.3±8.4	76.4±10.8
	Vila-MIL[22]	61.6±1.0	62.2±2.7	68.1±5.8
	AMD-MIL[13]	60.3±10.4	78.0±7.0	86.7±5.3
	TDA-MIL [19]	64.8±11.5	74.6±9.2	77.6±6.6
	SSC (Ours)	**73.5±12.2**	**82.5±8.1**	**90.4±4.1**
UBC-OCEAN	TRANS-MIL[20]	60.0±6.8	69.8±7.3	78.1±2.2
	ILRA-MIL[25]	63.2±6.1	68.9±5.7	76.8±4.0
	WIKG-MIL[12]	63.8±7.0	71.2±6.6	77.6±6.0
	Vila-MIL[22]	60.4±5.7	70.9±2.7	77.1±4.1
	AMD-MIL[13]	63.1±7.8	70.9±5.5	79.1±3.9
	TDA-MIL [19]	61.9±8.7	70.0±5.7	77.1±4.0
	SSC (Ours)	**73.7±6.1**	**77.0±5.3**	**80.4±2.6**

where $\mathcal{L}_{\text{CE}}$ is a standard cross-entropy. To reduce redundancy among the K query tokens, with $\hat{\mathbf{q}}_i$ indicates ℓ_2-normalized i-th query token, we define a temperature-scaled similarity distribution

$$A_{ij} = \frac{\exp(\hat{\mathbf{q}}_i \cdot \hat{\mathbf{q}}_j^\top / \tau)}{\sum_{l=1}^{K} \exp(\hat{\mathbf{q}}_i \cdot \hat{\mathbf{q}}_l^\top / \tau)},$$

then minimize the normalized row-wise entropy

$$\mathcal{L}_{\text{div}} = \frac{1}{K} \sum_{i=1}^{K} \frac{-\sum_{j=1}^{K} A_{ij} \log(A_{ij} + \epsilon)}{\log K}. \tag{6}$$

3 Experiments

Datasets. We validate the proposed framework on three distinct histopathology benchmarks spanning lung, breast, and ovarian malignancies. We use TCGA-NSCLC [23] for binary subtyping, distinguishing between lung adenocarcinoma

Table 2. F1 Score and AUC in the same experiment settings as Table 1. Names of compared methods are abbreviated by omitting the "-MIL" suffix.

Datasets	Methods	F1 Score (%)			AUC (%)		
		4-shot	8-shot	16-shot	4-shot	8-shot	16-shot
TCGA-NSCLC	TRANS	70.8±6.2	76.9±4.3	84.3±4.2	77.9±7.2	85.5±5.1	93.1±2.3
	ILRA	67.9±11.9	73.6±8.7	84.2±3.2	79.3±8.6	86.3±6.3	91.8±2.8
	WIKG	62.6±15.5	76.0±3.8	79.2±7.8	78.2±8.6	85.7±4.2	91.0±4.6
	Vila	77.8±8.4	84.2±5.0	**87.9±3.2**	84.5±10.6	91.2±5.1	94.8±1.9
	AMD	69.2±12.6	77.7±5.6	84.9±5.1	79.9±7.9	86.8±6.0	93.6±2.3
	TDA	69.8±11.8	80.6±4.5	86.2±5.2	81.3±9.1	89.4±5.9	**95.3±1.3**
	SSC	**79.8±7.4**	**84.8±2.7**	87.8±2.8	**87.7±8.3**	**91.9±3.1**	94.9±2.2
CAMELYON16	TRANS	51.1±8.4	54.0±9.9	64.2±18.3	58.9±6.2	59.9±7.3	74.4±13.0
	ILRA	50.9±5.3	63.1±12.7	73.8±11.8	55.1±4.8	74.5±8.2	85.7±7.7
	WIKG	47.4±8.0	60.9±14.4	73.2±16.3	56.6±7.1	74.4±8.6	85.3±7.3
	Vila	40.4±3.1	47.5±8.7	57.9±13.3	50.4±4.9	54.1±5.9	62.8±9.9
	AMD	56.8±14.9	77.1±7.9	86.0±5.9	67.3±8.4	85.4±6.0	93.5±3.6
	TDA	63.6±13.9	74.3±9.6	76.5±8.6	68.6±11.5	82.8±6.1	85.8±3.8
	SSC	**67.8±18.2**	**81.7±8.2**	**89.9±4.2**	**74.4±1.8**	**88.8±5.4**	**95.0±2.6**
UBC-OCEAN	TRANS	56.2±5.8	65.1±7.0	74.8±3.0	86.1±2.7	91.3±2.2	94.1±1.9
	ILRA	58.4±7.0	64.5±5.8	72.8±4.9	87.9±3.0	90.2±2.3	93.2±1.1
	WIKG	58.9±6.7	65.1±5.3	74.6±5.8	89.1±3.0	92.1±2.1	94.3±2.1
	Vila	55.7±6.2	67.9±3.2	74.5±4.2	83.7±3.7	89.7±2.4	93.9±1.7
	AMD	58.6±6.1	65.8±4.4	75.2±4.1	87.9±3.6	91.7±2.0	94.9±1.4
	TDA	56.9±6.5	66.1±7.3	74.9±3.6	86.4±3.2	91.1±1.5	94.3±1.3
	SSC	**70.2±6.7**	**74.7±4.9**	**78.3±2.7**	**92.6±2.3**	**93.3±1.9**	**95.6±0.8**

(541 slides) and squamous cell carcinoma (512 slides). For metastasis detection, we utilize CAMELYON16 [2], which comprises 399 whole slide images (159 tumor-positive and 240 normal). To assess multi-class performance, UBC-OCEAN [1] is employed, covering five ovarian carcinoma subtypes: high-grade serous (221), endometrioid (122), clear cell (98), along with low-grade serous and mucinous carcinomas (43 each). All datasets are stratified into training, validation, and testing sets with a 6:2:2 ratio. Consistent with standard Few-Shot Learning (FSL) protocols, we conduct experiments by randomly sampling $k \in \{4, 8, 16\}$ support slides per class.

Implementation Details. WSIs are pre-processed using the CLAM [15] toolset and extract 512×512 patches at 20× or 40× magnification. For pathology-specific knowledge synthesis, we utilize a large language model (Gemini 3). Both visual and textual modalities are embedded into a $D = 512$ dimensional space using the CONCH [16] encoder. In the SSC framework, we initialize 5 learnable prompts and set maximum token budget 8192 before SVTC. The model is optimized via AdamW ($LR = 1 \times 10^{-4}$) for up to 80 epochs, with early stopping applied based on the best validation performance. The weight λ in the training objective is 0.2.

Evaluation Metrics. Classification performance is assessed using Balanced Accuracy, F1-score, and AUC. To ensure robust evaluation, we implement a ten-fold cross-validation protocol, using unique data splits for each fold. We report the mean and standard deviation across all folds to represent the final performance.

Results and Discussion. Tables 1 and 2 show that SSC improves performance consistently across Balanced ACC, F1 and AUC, with the largest margins in the low-shot regime. On CAMELYON16 [2], where slide-level labels are determined by rare, small positive foci amid overwhelmingly irrelevant instances, SSC outperforms the strongest competing baseline by +8.7 BACC, +4.2 F1 and +5.8 AUC in the 4-shot setting, and the advantage remains clear at 8-shot (+4.5 BACC, +4.6 F1, +3.4 AUC) and 16-shot (+3.7 BACC, +3.9 F1, +1.5 AUC). On the five-class UBC-OCEAN [1], SSC shows even larger gains under 4-shot supervision (e.g., +9.9 BACC and +11.3 F1 over the best baseline), while still maintaining measurable improvements as supervision increases (up to +1.3 BACC, +3.1 F1, +0.7 AUC at 16-shot). On TCGA-NSCLC, SSC achieves consistent wins in the low-shot settings (e.g., +2.1 BACC, +2.0 F1, +3.2 AUC at 4-shot) and remains competitive at 16-shot where multiple baselines approach saturation (less than 0.1 BACC/F1 and 0.4 AUC of the best results). Overall, SSC improves data efficiency without sacrificing performance as supervision increases, and the trends are consistent across metrics and datasets.

Ablation Study. As shown in Fig 3, results (mean with error bar) on BACC, F1, and AUC are reported for $m = 4$, 8, 16 shots under three settings. BASELINE setting indicates using the embedding of *a histopathology of {class name}* to directly select patches while keeping the following SVTC [21]. DBPC consistently improves over BASELINE, while adding SKP yields further gains, with the highest margins under the most constrained budget ($m = 4$) and sustained benefits in larger shots. These results indicate that both components are complementary and jointly enhance robustness across metrics and selection regimes.

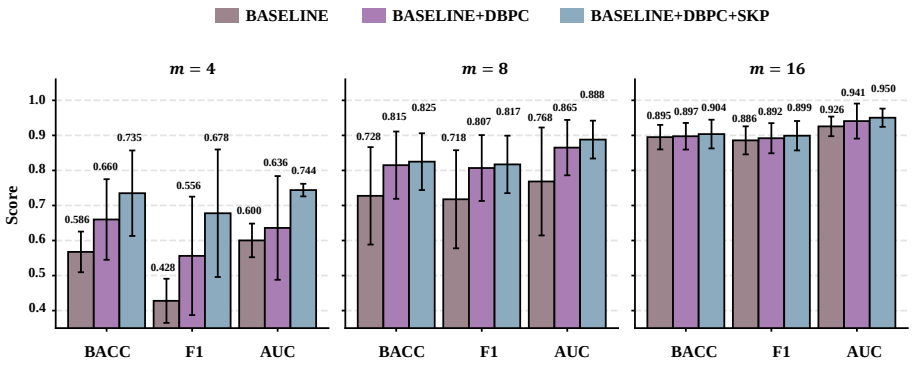

Fig. 3. Ablation of prompt components across different number of shots m.

4 Conclusion

We present SSC, a novel framework for few-shot WSI classification that uses knowledge subspace projection score selection to complement prompt-based consensus retrieval. SSC unifies two modules via uncertainty-guided fusion to adapt selection based on prompt consensus retrieval uncertainty, capturing wider range of patches with subtle diagnostic clues under extreme few-shot scenarios. Extensive experiments validate the effectiveness of this assess-then-fuse principle and achieve competitive results in data-efficient pathology diagnosis, while the interpretability of the principal subspace remains a direction for future work.

Acknowledgments. This work was partially supported by the National Key Research and Development Program of China (2023YFC2705704), the National Natural Science Foundation of China under Grant 62222112, and Grant 621761 86, the Innovative Research Group Project of Hubei Province under Grant 2024 AFA017, and the New Cornerstone Science Foundation through the XPLORER PRIZE. This work was also supported by the WHU-Kingsoft Joint Lab. The numerical calculations in this paper have been performed on the supercomputing system at the Supercomputing Center of Wuhan University.

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

References

1. Asadi-Aghbolaghi, M., Farahani, H., Zhang, A., Akbari, A., Kim, S., Chow, A., Dane, S., Consortium, O.C., Consortium, O., Huntsman, D.G., Gilks, C.B., Ramus, S., Köbel, M., Karnezis, A.N., Bashashati, A.: Machine learning-driven histotype diagnosis of ovarian carcinoma: Insights from the ocean ai challenge. medRxiv:2024.04 (2024)
2. Bejnordi, B.E., Veta, M., Diest, P.J.v., Ginneken, B.v., Karssemeijer, N., Litjens, G., Laak, J.A.W.M.v.d., the CAMELYON16 Consortium: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
3. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (2019)
4. Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H.: Multiple instance learning with center embeddings for histopathology classification. In: Med. Image Comput. Comput.-Assist. Interv. pp. 519–528 (2020)
5. Ethayarajh, K.: How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: EMNLP-IJCNLP. pp. 55–65 (2019)
6. Fu, K., Luo, X., Qu, L., Wang, S., Xiong, Y., Maglogiannis, I., Gao, L., Wang, M.: Fast: A dual-tier few-shot learning paradigm for whole slide image classification. Adv. Neural Inf. Process. Syst. pp. 105090–105113 (2024)

7. 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: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 15590–15600 (2025)
8. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Proc. AAAI Conf. Artif. Intell. pp. 933–941 (2022)
9. 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. Adv. Neural Inf. Process. Syst. pp. 37995–38017 (2023)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proc. Int. Conf. Mach. Learn. pp. 2127–2136 (2018)
11. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the sentence embeddings from pre-trained language models. arXiv preprint arXiv:2011.05864 (2020)
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: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 11323–11332 (2024)
13. Ling, X., Ouyang, M., Wang, Y., Chen, X., Yan, R., Chu, H., Cheng, J., Guan, T., Tian, S., Liu, X., He, Y.: Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In: Proc. ACM Int. Conf. Multimedia. p. 2795–2803 (2024)
14. 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: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 19764–19775 (2023)
15. 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. Nat. Biomed. Eng. **5**(6), 555–570 (2021)
16. Lu, M., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. Nat. Med. **30**(3), 863–874 (2024)
17. Qu, L., luo, x., Fu, K., Wang, M., Song, Z.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. In: Adv. Neural Inf. Process. Syst. vol. 36, pp. 67551–67564 (2023)
18. Qu, L., Yang, D., Huang, D., Guo, Q., Luo, R., Zhang, S., Wang, X.: Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In: Proc. Eur. Conf. Comput. Vis. pp. 196–212 (2024)
19. Reisenbüchler, D., Deng, R., Matek, C., Feuerhake, F., Merhof, D.: Top-down attention-based multiple instance learning for whole slide image analysis. In: Med. Image Comput. Comput.-Assist. Interv. pp. 651–660 (2025)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. pp. 2136–2147 (2021)
21. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., J. Kim, H., Soran, B., Krishnamoorthi, R., Elhoseiny, M., Chandra, V.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)

22. 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: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 11248–11258 (2024)
23. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia pp. 68–77 (2015)
24. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.H., Willens, S., Olguin, F.M., Nirschl, J.J., Joel, N., Diehn, M., Yang, S., Li, R.: A vision-language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)
25. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: Int. Conf. Learn. Represent. (2023)
26. Xu, Y., Wang, Y., Zhou, F., Ma, J., Jin, C., Yang, S., Li, J., Zhang, Z., Zhao, C., Zhou, H., Li, Z., Lin, H., Wang, X., Wang, J., Chan, R.C.K., Liang, L., Zhang, X., Chen, H.: A multimodal knowledge-enhanced whole-slide pathology foundation model. Nat. Commun. **16**, 11406 (2025)
27. Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. pp. 3583–3592 (2025)