KMP-MIL: Knowledge Memory Pool Multiple Instance Learning with Foundation Model for Continual Whole Slide Image Classification

Lingling Yuan[1,3], Zhaoxia Yin[3], Yan Han[1], Jinghua Zhang[4], Marcin Grzegorzek[5], and Chen Li(✉)[2,1]

[1] Group of Microscopic Image and Medical Image Analysis, College of Medicine and Biological Information Engineering, Northeastern University, China
[2] Software College, Northeastern University, China
[3] Shanghai Key Laboratory of Multidimensional Information Processing, School of Information and Electronic Engineering, East China Normal University, China
[4] College of Computer Science and Software Engineering, Hohai University, China
[5] Institute of Medical Informatics, University of Lübeck, Germany
lichen@bmie.neu.edu.cn

Abstract. Continual learning (CL) for pathological whole slide image (WSI) classification is challenging because gigapixel images, sparse diagnostic cues, and cross-organ domain shifts exacerbate catastrophic forgetting. In addition, existing methods often rely on rehearsal buffers, which are difficult to deploy in clinical practice under privacy and governance constraints. We propose knowledge memory pool multiple instance learning (KMP-MIL), a rehearsal-free continual MIL framework that couples a frozen pathology foundation model (FM) with parameterized memory-based adaptation. KMP-MIL introduces: 1) a growing KMP that incrementally stores compact task-specific memory units instead of replayed samples, and 2) query-based memory retrieval with prototype-conditioned feature calibration (PFC) and textual calibration (TC) in a shared vision-language space to improve adaptation. The framework supports task-CL with task identity (ID) and class-CL with task-agnostic inference through a universal memory pool that retrieves task-relevant knowledge without task ID. Experiments on sequential cross-organ WSI classification across breast, lung, kidney, and esophagus datasets show strong and stable performance. KMP-MIL achieves 89.08% average accuracy in task-CL and competitive class-CL results with reduced forgetting. It is also stable under reverse-order evaluation and storage-efficient, with persistent memory growing from 0.80 MB to 3.21 MB over four tasks, about 3.14% of a replay buffer storing 10 WSIs. Our code is available at https://github.com/Lingling-Yuan/KMP-MIL.

Keywords: Whole slide image · Continual learning · Foundation model.

1 Introduction

Whole slide images (WSIs) provide cellular-resolution evidence that is critical for tumor diagnosis and subtyping. However, a single slide can reach gigapixel-level

resolution, while diagnostic cues are sparse and unevenly distributed across the tissue [19]. These properties make dense annotation over WSIs impractical, due to high cost and engineering burden [3, 25]. As a result, clinical-grade WSI classification commonly uses weakly supervised multiple instance learning (MIL). In MIL, each WSI is split into patches and treated as a bag, and the model learns from slide-level labels. Aggregation modules, such as attention mechanisms or transformers, combine patch features to produce a slide-level prediction without pixel-level annotations [16, 9, 22]. In real deployments, data collection rarely stays fixed. Systems often grow incrementally, expanding from pilot sites to more hospitals, and from a single organ to multiple organs [23, 7]. With changing clinical needs and collaborations, the model must continuously incorporate data from new sources, forming a sequential task stream as shown in Fig. 1A. Therefore, models need continual learning (CL), rather than one-time offline training, to keep performance on earlier tasks while learning new ones [6, 18].

When organ-specific tasks arrive sequentially, CL faces greater catastrophic forgetting risk. Large differences in tissue structure, staining appearance, and scanning protocols across organs and cancer types can cause the model to overwrite useful cues from earlier tasks when learning later tasks, worsening forgetting in cross-organ settings [24]. Existing pathology CL methods often rely on rehearsal buffers that store representative slides or sample-level embeddings, and combine replay with regularization or distillation to stabilize MIL aggregation across tasks [8, 17, 13, 12, 11]. However, compliance and data governance can restrict centralized storage and replay of historical WSIs, limiting the applicability of buffer-based methods. This motivates rehearsal-free CL methods that retain knowledge without storing or replaying past slides or sample-level features.

Motivated by these constraints, we study rehearsal-free continual WSI classification with sequential organ-specific tasks, while considering that task identity (ID) may be unavailable at inference. We propose *knowledge memory pool multiple instance learning* (KMP-MIL), as shown in Fig. 1. KMP-MIL couples a frozen pathology foundation model (FM) with parameterized memory-based adaptation in a vision-language space. Instead of replaying historical WSIs, it maintains a growing *knowledge memory pool* (KMP) of compact task-specific memory units. For each slide, a WSI-level query retrieves Top-K units to provide knowledge-conditioned context. The retrieved units are used for *prototype-conditioned feature calibration* (PFC) before MIL aggregation, while *textual calibration* (TC) uses class text prototypes to calibrate bag representations and improve semantic alignment. KMP-MIL supports both task-CL and class-CL. In task-CL, retrieval is restricted to the task-specific KMP segment when task ID is available; in class-CL, the model queries a universal pool built from all seen tasks for task-agnostic prediction. Our main contributions are as follows:

- We propose KMP-MIL, which uses a growing KMP to store compact task-specific memory units, thereby avoiding the replay of historical WSIs.
- We design a query-based mechanism for WSI-specific Top-K memory selection, and combine it with PFC and TC in a shared vision-language space to improve adaptation and cross-organ semantic alignment.

– We evaluate KMP-MIL on sequential cross-organ WSI classification across breast, lung, kidney, and esophagus datasets under both task-CL and class-CL settings, including reverse-order evaluation. KMP-MIL achieves strong performance with reduced forgetting and good storage efficiency.

2 Proposed Method

We study rehearsal-free continual WSI classification with sequentially introduced organ-specific tasks. The learning stream contains T tasks indexed by $t \in \{1, \ldots, T\}$, and each task provides training data from a new organ domain. We propose KMP-MIL, which maintains a growing KMP for rehearsal-free CL and applies PFC and TC to improve cross-domain robustness.

2.1 Training Stage

Knowledge Memory Pool (KMP). For each task t, we maintain a task-specific $\mathrm{KMP}_t = \{(k_m^t, p_m^t)\}_{m=1}^M$, as shown in Fig. 1B. Each memory unit contains a learnable key $k_m^t \in \mathbb{R}^C$ in the aligned visual-text space and prompt tokens $p_m^t \in \mathbb{R}^{L \times D}$ for WSI aggregation, where C denotes the visual-text embedding dimension, and L and D denote the prompt length and the hidden dimension of the frozen text tower E_{txt}, respectively. As tasks arrive, the memory pool grows as $\mathrm{KMP}_{1:t} = \bigcup_{t'=1}^{t} \mathrm{KMP}_{t'}$, enabling rehearsal-free CL by storing compact parameterized memory units instead of historical WSIs. For a WSI bag $\mathcal{X}_i$, we extract patch embeddings h_{ij} by the frozen pathology FM image encoder E_{img} and form $H_i = [h_{i1}, \ldots, h_{iN_i}]^\top \in \mathbb{R}^{N_i \times C}$. We obtain a WSI-level query $q_i \in \mathbb{R}^C$ by max pooling over patch embeddings, using it only as a lightweight retrieval router to preserve sparse diagnostic cues without adding trainable query modules. To inject task knowledge into retrieval, we use E_{txt} to encode a task-level descriptor d^t that captures organ-specific histological patterns and diagnostic morphological cues. The resulting embedding initializes the newly added keys in KMP_t, providing semantic priors that improve early routing stability by reducing ambiguity between current-task keys and historical keys. In parallel, class text embeddings are used as class prototypes for TC. We ℓ_2-normalize q_i and all keys, so cosine similarity is equivalent to an inner product. The retrieval score for the m-th key in task t is:

$$s_{i,m}^t = \cos(q_i, k_m^t) = q_i^\top k_m^t, \quad m \in \{1, \ldots, M\}, \tag{1}$$

where higher $s_{i,m}^t$ indicates the query is more consistent with the memory unit. We select Top-K indices $\mathcal{I}_i^t = \mathrm{TopK}(\{s_{i,m}^t\}_{m=1}^M)$ from current-task keys and use $\mathcal{L}_{\mathrm{match}} = 1 - \frac{1}{BK} \sum_{i=1}^{B} \sum_{m \in \mathcal{I}_i^t} s_{i,m}^t$ to align queries with retrieved keys, where K is the retrieval size and B is the batch size. For calibration and aggregation, we convert the Top-K retrieval similarities into normalized weights $w_{i,m}$ using softmax, so that more relevant memory units receive larger weights while non-selected units are assigned zero weight. To stabilize retrieval, we encourage each

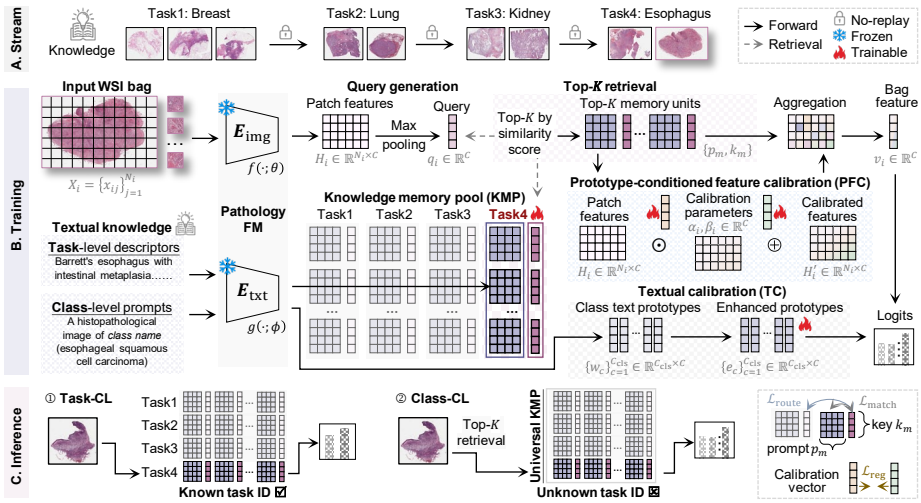

Fig. 1. Overview of KMP-MIL for rehearsal-free continual WSI classification. A: Data stream. Sequential stream of four organ-specific WSI tasks. **B: Training.** Frozen E_{img} extracts patch features; query q retrieves Top-K KMP units for PFC and prototype-guided aggregation, and TC-enhanced prototypes from E_{txt} are used for classification. **C: Inference.** Task-CL retrieves within the task-specific KMP using task ID, while class-CL retrieves from a universal KMP without task ID.

query to prefer current-task keys over past-task keys using a routing loss:

$$\mathcal{L}_{\text{route}} = \frac{1}{B}\sum_{i=1}^{B} \text{softplus}\left(\frac{m_i^{\text{past}} - m_i^{\text{cur}} + \mu}{\tau}\right), \tag{2}$$

where $m_i^{\text{cur}} = \max_m q_i^\top k_m^t$ and $m_i^{\text{past}} = \max_{t'<t,\, m} q_i^\top k_m^{t'}$, and μ and τ are the margin and temperature, respectively. For the first task, we set $\mathcal{L}_{\text{route}} = 0$. The KMP loss is $\mathcal{L}_{\text{KMP}} = \mathcal{L}_{\text{match}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}}$. The retrieved memory units provide knowledge-conditioned context for downstream MIL aggregation, enabling the model to focus on task-relevant patterns without accessing past training data. During continual training, only the newly added memory units in KMP_t are updated, while memory units from previous tasks are frozen and reused.

Prototype-conditioned Feature Calibration (PFC). To mitigate domain shift across sequential organ tasks, we calibrate patch features using the Top-K memory units retrieved from KMP. Given a query q_i, we reuse the KMP retrieval outputs, including the Top-K index set $\mathcal{I}_i^t$ and the normalized retrieval weights $w_{i,m}$. Inspired by FiLM [20], we propose the PFC module, which attaches learnable calibration parameters to each KMP memory unit, as shown in Fig. 1B. Specifically, for task t, we learn two channel-wise parameter pools $\{\alpha_m^t\}_{m=1}^M$ and $\{\beta_m^t\}_{m=1}^M$ with $\alpha_m^t, \beta_m^t \in \mathbb{R}^C$, where each pair (α_m^t, β_m^t) serves as an affine modulator in the shared embedding space. Because these parameters are indexed by retrieved KMP units, the modulation is compact and retrieval-aware, which is

suitable for rehearsal-free CL. For each bag, we fuse the retrieved calibration parameters using the retrieval weights $\alpha_i = \sum_{m \in \mathcal{I}_i^t} w_{i,m} \alpha_m^t, \beta_i = \sum_{m \in \mathcal{I}_i^t} w_{i,m} \beta_m^t$, and calibrate the patch embedding matrix $H_i \in \mathbb{R}^{N_i \times C}$ as:

$$H_i' = H_i \odot \alpha_i + \mathbf{1}_{N_i} \beta_i^\top, \tag{3}$$

where $\alpha_i, \beta_i \in \mathbb{R}^C$ are broadcast to all N_i patches, $\mathbf{1}_{N_i} \in \mathbb{R}^{N_i}$ is an all-ones vector, and $\odot$ denotes channel-wise multiplication. To avoid overly aggressive modulation, we regularize the fused scaling and bias by encouraging α_i to stay close to the identity transform and β_i to stay close to zero:

$$\mathcal{L}_{\text{reg}} = \frac{1}{BC} \sum_{i=1}^{B} \sum_{c=1}^{C} (\alpha_{i,c} - 1)^2 + \frac{1}{BC} \sum_{i=1}^{B} \sum_{c=1}^{C} (\beta_{i,c})^2 . \tag{4}$$

The calibrated patch embeddings H_i' are then aggregated by a prototype-guided MIL module using multi-prototype cosine-weighted pooling, where cosine similarities between patch features and retrieved KMP prototypes are normalized across patches to compute the bag feature v_i.

Textual Calibration (TC). To align prediction with the shared vision-language embedding space, we build class text prototypes using a textual template ensemble. For each class c, we construct multiple textual descriptions by combining template sentences with class-name variants, and encode them with E_{txt} to obtain a base class text prototype w_c. We then introduce a residual TC module for CL. Specifically, as shown in Fig. 1B, a learnable calibration vector $u_c \in \mathbb{R}^C$ is added to the base prototype to form the enhanced prototype $e_c = w_c + \gamma u_c$, where γ is a residual scaling factor. This residual design preserves the semantic prior of w_c while allowing task-adaptive refinement under CL. Given the bag representation v_i, we ℓ_2-normalize v_i and e_c and compute class logits by cosine similarity $\hat{y}_{i,c}$. We optimize the classification branch with $\mathcal{L}_{\text{cls}}$ and a class similarity regularizer to reduce prototype collapse $\mathcal{L}_{\text{sim}} = \frac{1}{C_{\text{cls}}(C_{\text{cls}}-1)} \sum_{a \neq b} \left(e_a^\top e_b + 1 \right)$, where C_{cls} is the number of classes, and a and b index different class prototypes.

Training Objective. We optimize only the new KMP_t, PFC calibration parameters α_m^t and β_m^t, and TC prototype residual u_c. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{KMP}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{sim}} \mathcal{L}_{\text{sim}}$, where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss.

2.2 Inference Stage

As shown in Fig. 1C, KMP-MIL uses the same forward pipeline in both evaluation settings, but differs in retrieval scope. For a test WSI, we compute a query q_i and retrieve Top-K memory units from KMP to guide feature calibration and prototype-guided aggregation, producing the bag representation v_i and scaled cosine logits with TC-enhanced class prototypes. By reusing stored memory units from previous tasks, the model leverages knowledge at test time without storing or replaying historical WSIs, alleviating forgetting. In task-CL, retrieval is restricted to the task-specific pool KMP_t to match the known organ domain. In class-CL, task ID is unavailable, so we retrieve from a universal pool $\text{KMP}_{1:t}$

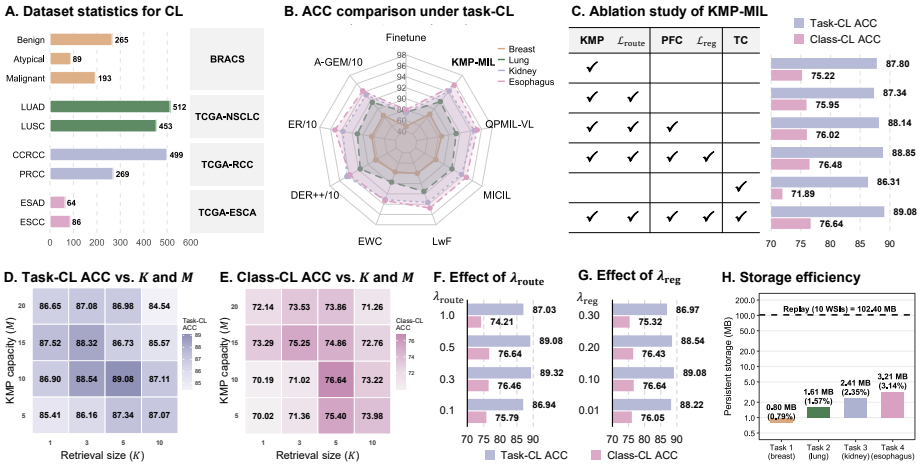

Fig. 2. Dataset statistics and related experimental results for continual WSI learning. A: Dataset statistics of the four sequential organ-specific WSI tasks used for CL. **B:** ACC comparison across different CL methods under task-CL. **C:** Ablation study of KMP-MIL under task-CL and class-CL. **D-E:** Sensitivity analysis of K and M under task-CL and class-CL. **F-G:** Sensitivity analysis of λ_{route} and λ_{reg} under task-CL and class-CL. **H:** Storage efficiency comparison between KMP-MIL and replay-based methods across sequential tasks.

formed by all seen tasks, enabling task-agnostic inference while reusing learned memory units and calibration parameters.

3 Experiments

3.1 Experimental Setup

Datasets. We used four organ-specific WSI datasets covering breast, lung, kidney, and esophagus. Unless otherwise specified (i.e., reverse-order evaluation), the sequential task order is breast → lung → kidney → esophagus. The breast dataset is from BRACS [1] with three categories, and the other three datasets are from TCGA[1], including non-small cell lung carcinoma (NSCLC), renal cell carcinoma (RCC), and esophageal carcinoma (ESCA), each with two categories. Detailed statistics are shown in Fig. 2A. We cropped non-overlapping 256×256 patches at $10\times$ magnification using CLAM [16].

Implementation Details. All experiments were conducted on an NVIDIA RTX 4090 GPU. We trained the model sequentially for 20 epochs per task and reported the mean and standard deviation (std) over 10-fold cross-validation with patient-level splits. Due to variable-sized MIL bags, we used batch size 1 with gradient accumulation over 16 iterations per optimizer step. For a fair

[1] `https://portal.gdc.cancer.gov`

Table 1. Performance comparison under class-CL (task ID unavailable). Results are reported as mean ± std, with the best performance highlighted in **bold**.

Method	Performance (↑)			Forgetting (↓)		
	ACC	AUC	PR-AUC	$\mathcal{F}_{\text{ACC}}$	$\mathcal{F}_{\text{AUC}}$	$\mathcal{F}_{\text{PR-AUC}}$
Finetune	15.83±1.14	68.83±1.80	67.53±1.65	41.71±2.68	22.63±0.94	21.65±0.76
A-GEM/10 [4]	31.49±12.55	78.05±8.96	74.95±8.24	12.14±5.27	12.76±8.07	11.94±7.21
ER/10 [21]	73.88±3.86	91.20±2.19	88.29±2.26	14.91±2.50	5.47±2.10	5.61±2.08
DER++/10 [2]	75.79±3.13	**93.74±1.56**	90.48±1.97	12.81±2.26	3.00±1.32	3.50±1.63
EWC [10]	13.07±6.17	84.08±4.28	81.73±3.94	11.25±2.92	3.03±0.92	3.05±0.64
LwF [14]	51.10±4.17	93.61±0.96	90.15±1.17	21.10±3.39	3.80±0.49	2.48±0.72
MICIL [17]	20.74±7.25	90.77±1.68	87.74±1.31	15.53±5.70	4.06±0.88	4.54±1.05
QPMIL-VL [5]	75.55±3.38	92.49±1.22	90.05±1.37	8.18±1.74	2.91±0.60	2.75±0.50
Ours (KMP-MIL)	**76.64±2.27**	93.56±0.98	**90.64±0.93**	**7.15±1.72**	**2.13±0.62**	**2.08±0.54**

comparison, all methods were evaluated using the pre-trained CONCH encoder E_{img} [15] to extract patch features. Trainable parameters were optimized with Adam (lr=5e−2, weight decay=5e−4). Unless otherwise stated, we set $M = 10$, $K = 5$, $\mu = 0.05$, $\tau = 0.07$, $\lambda_{\text{route}} = 0.5$, $\lambda_{\text{reg}} = 0.1$, $\lambda_{\text{sim}} = 0.5$, and $\gamma = 0.5$.

Evaluation Metrics. We evaluated KMP-MIL under task-CL and class-CL using classification metrics: accuracy (ACC), area under the ROC curve (AUC), and area under the precision-recall curve (PR-AUC). To measure old-task degradation, we used the corresponding forgetting rates, $\mathcal{F}_{\text{ACC}}$, $\mathcal{F}_{\text{AUC}}$, and $\mathcal{F}_{\text{PR-AUC}}$, when the base metrics were reported. These rates are especially informative in class-CL, where unavailable task IDs make cross-task confusion central.

3.2 Results

Comparison with SOTA. We compared KMP-MIL with representative CL baselines under both task-CL and class-CL settings, including rehearsal-based methods (A-GEM [4], ER [21], and DER++ [2]), regularization methods (EWC [10] and LwF [14]), and pathology-oriented CL methods (MICIL [17] and QPMIL-VL [5]). Under task-CL, KMP-MIL achieves the best ACC on three of four organs, namely breast, lung, and esophagus, while remaining competitive on kidney, as shown in Fig. 2B. It reaches 73.91% on breast, 91.06% on lung, and 96.71% on esophagus, and attains the highest average ACC across organs at 89.08%, surpassing QPMIL-VL at 88.59%. Under class-CL, where task ID is unavailable at inference, KMP-MIL achieves the best overall balance between performance and forgetting, as shown in Table 1. It obtains the highest ACC and PR-AUC, while maintaining a highly competitive AUC that is close to the best reported result. In addition, KMP-MIL yields the lowest forgetting metrics, including $\mathcal{F}_{\text{ACC}}$, $\mathcal{F}_{\text{AUC}}$, and $\mathcal{F}_{\text{PR-AUC}}$.

Ablation Study. The ablation results in Fig. 2C show that KMP is the core contributor in both task-CL and class-CL, as it preserves task-specific knowledge through compact memory units. Starting from the KMP-only variant, adding $\mathcal{L}_{\text{route}}$ improves class-CL by 0.73% while slightly reducing task-CL by 0.46%, indicating reduced routing ambiguity under task-agnostic inference. Adding PFC further improves task-CL and class-CL by 0.80% and 0.07%, respectively, and

Table 2. Performance comparison under reverse-order CL (sequence: esophagus → kidney → lung → breast) for both task-CL and class-CL settings. Results are reported as mean ± std, with the best performance highlighted in **bold**.

Method	Task-CL setting		Class-CL setting			
	ACC (↑)	AUC (↑)	ACC (↑)	AUC (↑)	$\mathcal{F}_{\text{ACC}}$ (↓)	$\mathcal{F}_{\text{AUC}}$ (↓)
Finetune	64.58±3.04	88.29±2.21	12.19±3.20	66.37±2.46	28.97±3.26	5.51±2.44
A-GEM/10 [4]	81.27±2.39	89.28±1.20	36.32±9.39	75.57±5.43	21.94±4.99	4.82±2.00
ER/10 [21]	84.83±2.07	92.78±1.33	70.32±3.36	93.88±2.04	19.38±3.47	2.15±1.17
DER++/10 [2]	85.18±1.94	93.13±1.02	70.14±3.56	93.60±1.63	18.90±3.42	2.36±0.51
EWC [10]	78.72±1.88	90.25±2.01	13.97±5.94	82.58±3.93	26.79±4.20	4.81±1.23
LwF [14]	88.35±2.87	**95.90±1.14**	49.03±4.74	93.90±1.22	25.37±4.83	5.32±1.43
MICIL [17]	87.73±1.03	93.20±1.29	23.44±8.32	92.54±1.28	22.66±4.03	3.21±1.84
QPMIL-VL [5]	86.27±2.93	92.92±0.82	69.48±2.33	92.21±1.82	19.62±3.20	2.15±0.91
Ours (KMP-MIL)	**88.46±2.44**	95.79±0.91	**71.90±3.77**	**94.11±1.18**	**16.16±2.93**	**1.40±0.78**

$\mathcal{L}_{\text{reg}}$ brings additional gains of 0.71% and 0.46%, suggesting that retrieval-aware calibration mitigates cross-organ feature shift while avoiding over-modulation. TC alone performs worse than KMP-only by 1.49% in task-CL and 3.33% in class-CL, showing that textual prototypes alone cannot replace memory retrieval. When combined with KMP-MIL, TC provides complementary semantic alignment, yielding the best overall results.

Hyperparameter Sensitivity Analysis. We analyzed the key hyperparameters of KMP-MIL, including KMP capacity M, retrieval size K, and the weights of $\mathcal{L}_{\text{route}}$ and $\mathcal{L}_{\text{reg}}$ in Fig. 2D-G. M and K balance memory cost and retrieval coverage, where too small values may miss task knowledge and overly large values may introduce redundant memories. KMP-MIL remains stable across different settings, with strong balanced performance around $M = 10$, $K = 5$, $\lambda_{\text{route}} = 0.5$, and $\lambda_{\text{reg}} = 0.1$. λ_{route} improves task-agnostic routing in class-CL, while λ_{reg} stabilizes PFC by preventing overly aggressive feature modulation.

Reverse-order Evaluation. We further evaluated KMP-MIL under a reverse-order CL setting, as shown in Table 2. KMP-MIL achieves the best ACC in both task-CL and class-CL, reaching 88.46% and 71.90%, respectively. Under class-CL, KMP-MIL also achieves the best AUC at 94.11% and the lowest forgetting metrics, including $\mathcal{F}_{\text{ACC}}$ of 16.16% and $\mathcal{F}_{\text{AUC}}$ of 1.40%. These results show that KMP-MIL is less sensitive to this reversed task order and can effectively reduce forgetting in this more challenging CL scenario.

Rehearsal-free Storage Efficiency. To evaluate the rehearsal-free advantage of our method in continual WSI learning, we further compared the persistent storage cost of KMP-MIL with replay-based methods. As shown in Fig. 2H, as tasks are added sequentially, the cumulative storage of KMP-MIL increases only from 0.80 MB to 3.21 MB. In contrast, a replay-based method requires about 102.40 MB when storing 10 WSIs in the buffer. This means that after four tasks, KMP-MIL uses only about 3.14% of the replay storage budget, which greatly reduces storage cost in CL. These results highlight the rehearsal-free nature of KMP-MIL: even without replaying historical samples, it can still retain and transfer knowledge effectively with very small persistent storage overhead.

4 Conclusion

In this paper, we proposed KMP-MIL as a rehearsal-free continual MIL framework for WSI classification. It couples a frozen pathology FM with a growing KMP. By retrieving compact task-specific memory units and combining PFC with TC, KMP-MIL reuses knowledge without storing or replaying historical WSIs, thereby reducing catastrophic forgetting in cross-organ CL. KMP-MIL supports both task-CL and class-CL settings, including task-agnostic inference when task ID is unavailable. Experiments show stable performance, reduced forgetting, robustness under reverse-order evaluation, and favorable storage efficiency. One limitation is that the current KMP uses a fixed memory budget per task and grows linearly with the number of tasks. Adaptive memory allocation, memory merging, pruning, compression, Top-K task-origin tracing, and visualization of retrieved memory units, feature shifts, and prototype shifts are future directions for longer task streams and better interpretability.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant No. 82220108007).

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

References

1. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., et al.: BRACS: A dataset for breast carcinoma subtyping in H&E histology images. Database **2022**, baac093 (2022)
2. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: A strong, simple baseline. In: Advances in Neural Information Processing Systems. vol. 33, pp. 15920–15930 (2020)
3. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019)
4. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with A-GEM. In: Proceedings of the International Conference on Learning Representations (2019)
5. Gou, J., Ji, L., Liu, P., Ye, M.: Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3158–3166 (2025)
6. Hao, X., Ni, W., Jiang, X., Tan, W., Yan, B.: Addressing imbalance for class incremental learning in medical image classification. In: Proceedings of the ACM International Conference on Multimedia. pp. 2467–2476 (2024)
7. Hartman, D.J., Pantanowitz, L., McHugh, J.S., Piccoli, A.L., O'Leary, M.J., et al.: Enterprise implementation of digital pathology: Feasibility, challenges, and opportunities. Journal of Digital Imaging **30**(5), 555–560 (2017)
8. Huang, Y., Zhao, W., Wang, S., Fu, Y., Jiang, Y., et al.: ConSlide: Asynchronous hierarchical interaction transformer with breakup-reorganize rehearsal for continual whole slide image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21349–21360 (2023)

9. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning. vol. 80, pp. 2127–2136 (2018)
10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017)
11. Kumari, P., Reisenbüchler, D., Bozorgpour, A., Schaadt, N.S., Feuerhake, F., et al.: Attention-based generative latent replay: A continual learning approach for WSI analysis. arXiv preprint (2025)
12. Kumari, P., Reisenbüchler, D., Luttner, L., Schaadt, N.S., Feuerhake, F., et al.: Continual domain incremental learning for privacy-aware digital pathology. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. 15012, pp. 34–44. Springer (2024)
13. Li, X., Cui, Y., Li, J., Chan, A.B.: Advancing multiple instance learning with continual learning for whole slide imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20800–20809 (2025)
14. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018)
15. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
16. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
17. Meseguer, P., del Amor, R., Naranjo, V.: MICIL: Multiple-instance class-incremental learning for skin cancer whole slide images. Artificial Intelligence in Medicine **152**, 102870 (2024)
18. Nguyen, A.T., Byeon, K., Kim, K., Song, B., Chae, S.W., et al.: CAMP: Continuous and adaptive learning model in pathology. npj Artificial Intelligence **2**(1), 2 (2026)
19. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The Lancet Oncology **20**(5), e253–e261 (2019)
20. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
21. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T.P., Wayne, G.: Experience replay for continual learning. In: Advances in Neural Information Processing Systems. vol. 32, pp. 348–358 (2019)
22. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., et al.: TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
23. Treanor, D., Williams, B.: The Leeds Guide to Digital Pathology. The Leeds Teaching Hospitals NHS Trust and University of Leeds, Leeds (2019)
24. Wu, X., Xu, Z., Tong, R.K.Y.: Continual learning in medical image analysis: A survey. Computers in Biology and Medicine **182**, 109206 (2024)
25. Yuan, L., Gu, Y., Han, Y., Du, T., Grzegorzek, M., et al.: Exploring transparency in pathological image analysis: A comprehensive review of explainable artificial intelligence (XAI) techniques. Computer Science Review **60**, 100863 (2026)