STABILIZING SLIDE REPRESENTATIONS FOR CONTINUAL WHOLE-SLIDE IMAGE LEARNING: FROZEN ATTENTION AND PROTOTYPE REHEARSAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Class-incremental whole-slide image (WSI) classification must learn newly arriving cancer types without catastrophic forgetting of previously learned classes and without access to data from earlier tasks. Compact class prototypes offer an exemplar-free way to rehearse earlier classes, but a prototype is valid only in the representation space that produced it. Freezing only the patch encoder does not stabilize slide-level representations: a trainable multiple-instance learning attention aggregator maps the same patch bag to different embeddings as its attention changes across tasks, making stored prototypes stale. We introduce **PACE** (Prototype-Anchored Continual Expansion), which fixes a pretrained gated-attention *aggregation map* so that identical patch features yield identical embeddings across tasks, and rehearses K class-conditional K-means centroids in this fixed space; only an expanding linear classifier adapts to the new cancer types. PACE retains no slides, patch bags, or pseudo-samples and needs no task identity at inference. Across seven public cohorts, 11 class-incremental curricula, and 21 continual baselines, PACE attains the highest mean end-of-stream average balanced accuracy (AACC) on eight curricula and ties on one; its margins over the strongest competitor reach 10.5 points, and slide-retaining methods lead on the other two curricula, where classes are smallest or most heterogeneous. A matched ablation shows that fixing the aggregation map, not its initialization, is what makes rehearsal effective, and that one prototype per class ($K{=}1$) recovers 82% of the AACC gain achieved with $K{=}16$. For continual prognosis, a prototype-free extension with frozen cohort-specific Cox heads also attains the highest macro-averaged concordance index among the evaluated continual methods under all five orderings of a six-cohort stream, though its margins over the strongest competitor (0.8–2.5 points) are smaller than in diagnosis and within the standard deviation across folds. Stabilizing the slide representation at the aggregation map thus makes compact, slide-free prototype rehearsal effective for continual WSI diagnosis and, without prototypes, extends to cohort-incremental prognosis.

1 INTRODUCTION

Whole-slide images (WSIs) are central to computational pathology for diagnosis and prognosis (Louis et al., 2016; Campanella et al., 2019; Lu et al., 2021; Kather et al., 2020; Echle et al., 2021). In clinical deployment, a model must learn newly recognized or rare cancer types, organ sites, and outcome cohorts without degrading performance on previously learned tasks, a failure known as catastrophic forgetting (McCloskey & Cohen, 1989). Class-incremental WSI models (Rebuffi et al., 2017; Bilal et al., 2025) must learn categories that arrive in disjoint tasks and, at inference, discriminate over the cumulative label space without task identity. Doing so at gigapixel scale is demanding: each WSI yields thousands of patch features under a single slide-level label, and retaining past slides or patch bags for rehearsal is costly in storage and raises data governance and privacy concerns. Attention-based multiple-instance learning (MIL) handles the scale by encoding patches with a pretrained model, pooling the feature bag into a slide representation, and classifying it (Ilse et al., 2018; Shao et al., 2021; Zhang et al., 2022).

Freezing the patch encoder stabilizes patch features but not the slide-level mapping: while the MIL aggregator remains trainable, an unchanged feature bag can receive different attention weights and

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

drift through the slide-representation space from one task to the next (Fig. 1). Prior continual WSI methods trace much of the old-task degradation to this component and mitigate it by distillation, replay, or constrained adaptation of a still-trainable aggregator (Li et al., 2025; Huang et al., 2023; Meseguer et al., 2024). Aggregation is thus an established intervention point, but one question remains open: *does fixing the aggregation map make compact prototype rehearsal effective in class-incremental WSI classification?*

A stored prototype is a static summary of the space that produced it. If the aggregation map changes as new cancer types arrive, earlier prototypes no longer correspond to slides embedded later, and restoring that correspondence would require retaining the original slides for recomputation. We introduce **PACE** (Prototype-Anchored Continual Expansion), which removes aggregation-induced coordinate drift as a source of prototype invalidation. Beyond the frozen patch encoder that is standard in WSI-MIL, PACE also fixes the attention aggregator—a Gated ABMIL module initialized from the PC-108 pan-cancer prior (Shao et al., 2025)—so an unchanged feature bag maps to the same representation at every task and stored prototypes remain valid without recomputation. Learning newly arriving cancer types requires adapting only an expanding linear classifier. Because its predictions still shift as new types are learned, PACE rehearses K class-conditional K-means centroids per earlier class in the fixed slide space. Each component has one role: the fixed aggregator anchors the representation, the expanding classifier adapts, and prototype rehearsal mitigates forgetting of earlier classes. No prior WSIs, patch bags, or pseudo-samples are retained, and inference needs neither task identity nor the prototype memory.

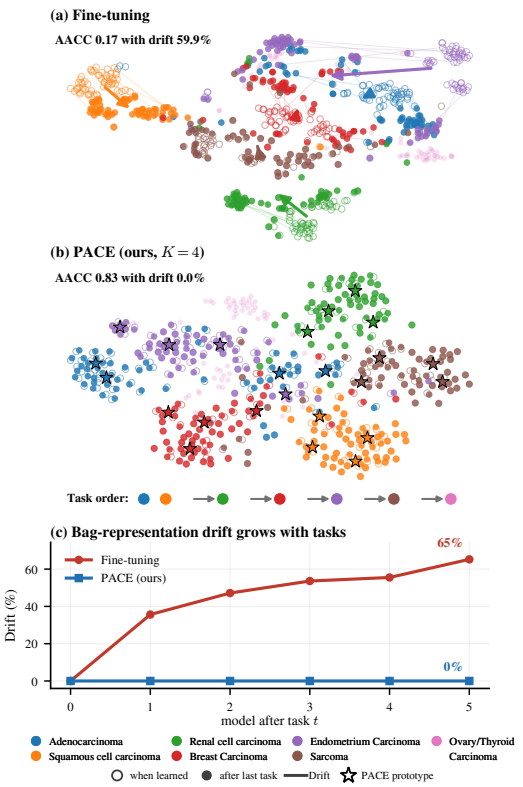

Figure 1: The same slides drift under fine-tuning but not under PACE (CHTN CANCER FAMILY curriculum; 7 classes, 6 tasks; one run per method; panel headers report that run's average balanced accuracy, AACC). **(a)** Fine-tuning (per-panel t-SNE; pretrained PC-108 attention-based aggregator, trained throughout continual learning): each slide moves from its position when its class was learned (hollow) to its position under the final model (filled). The mean relative drift in embedding space, Δ_{emb} (Eq. (3)), is 59.9%. **(b)** PACE (K=4; the same aggregator, frozen throughout continual learning): the two positions coincide (Δ_{emb}=0%); stars mark earlier-class prototypes. **(c)** $\Delta_{\text{emb}}(t)$ of the base-task slides after each task (t=0: base task). It rises to 65% under fine-tuning and stays at exactly 0 under PACE.

Our contributions are as follows:

- **We identify and quantify aggregation-induced slide-representation drift as a continual-learning stability target.** The same-slide diagnostic (Fig. 1) re-embeds identical feature bags after successive tasks, exposing large drift under fine-tuning (59.9%) and none under PACE by construction. Aggregation drift is not the sole cause of forgetting—the expanding classifier drifts too—which is why PACE pairs the fixed map with rehearsal.
- **We introduce PACE, a fixed-coordinate MIL design that keeps compact prototype rehearsal valid across tasks.** PACE fixes both the patch encoder and the Gated ABMIL aggregator, so stored class-conditional centroids stay in the coordinates of later-embedded slides while only the linear classifier expands; no slide, patch bag, or pseudo-sample is retained.
- **Matched ablations show that fixing the aggregator, rather than its initialization, is what makes prototype rehearsal effective.** On CANCER FAMILY (seven cancer families over six tasks), across PC-108 and Xavier attention initializations, one prototype per class raises AACC by 0.55–0.58 with a fixed aggregator but by only 0.02–0.22 with a trainable one, and already recovers 82% of the gain reached with K=16 (Sec. 6.1).

- **PACE leads or ties on nine of 11 diagnostic curricula and on every ordering of a prognosis extension while retaining no slides.** Against 21 continual baselines, its lead reaches 10.5 points and four lead margins exceed run-to-run sd, while slide-retaining methods remain stronger on the two curricula with the smallest or most heterogeneous classes. The prototype-free prognosis extension attains the highest macro-averaged C-index among the evaluated continual methods under every cohort ordering, with margins (0.8–2.5 points) within the standard deviation across folds (Secs. 6.1, 6.2).

2 Related work

Slide representation in WSI MIL. Weakly supervised MIL represents a WSI as a variable-size bag of patch features and maps it to a slide-level prediction through a learned aggregation function (Ilse et al., 2018; Carbonneau et al., 2018). Gated-attention MIL forms the slide embedding as an attention-weighted sum of patch features (Ilse et al., 2018); TransMIL and DTFD-MIL instead model instance correlations or distill across pseudo-bags (Shao et al., 2021; Zhang et al., 2022). WSI feature bags have low-rank structure (Xiang et al., 2023), stabilizing attention aids non-continual prediction (Ye et al., 2026), and learned aggregation supports MIL prognosis (Chen et al., 2021; 2022; Liu et al., 2024). These works treat aggregation as part of representation learning, but none asks whether fixing the bag-to-slide aggregation map makes compact prototype rehearsal effective across continual tasks.

Fixed representations and prototypes in class-incremental learning. Generic class-incremental learning stabilizes a growing label space through parameter or output regularization, exemplar or logit replay, classifier-bias correction, and parameter-isolated updates (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Rebuffi et al., 2017; Buzzega et al., 2020; Wu et al., 2019; Zhao et al., 2020; Liang & Li, 2024). Closest to PACE, feature-space methods reduce dependence on raw exemplars: SimpleCIL classifies by class means in a frozen pretrained representation (Zhou et al., 2025); FeTrIL freezes the feature extractor and uses class prototypes to translate new-class features toward earlier classes (Petit et al., 2023); ACIL recursively updates an analytic classifier without retaining historical examples (Zhuang et al., 2022); and PASS stores one mean prototype per class for feature-space augmentation, together with distillation and self-supervision (Zhu et al., 2021). These methods assume that each input already has one feature vector and therefore do not study the learned MIL map between patch features and slide-level class statistics. PACE transfers the fixed-space principle to this bag-to-slide boundary: it fixes both the patch encoder and the aggregator and rehearses K class-conditional centroids through one expanding classifier to adapt to new classes while mitigating forgetting of earlier ones.

Continual learning for WSI diagnosis. WSI-specific methods differ in what they retain and where they adapt. ConSlide replays stored regions as synthetic bags (Huang et al., 2023); PaGMIL selects patches with pathological priors and routes each slide by thumbnail prompt to a task-specific head (Zheng et al., 2025); QPMIL-VL guides vision–language aggregation with a buffer-free learned prototype pool (Gou et al., 2025); CoMEL uses bag prototypes for instance pseudo-labeling with orthogonal low-rank adaptation (Lee et al., 2025); CAMP learns task-specific knowledge in a generative classifier (Nguyen et al., 2026); and MergeSlide merges task-adapted vision–language models and selects the task by prompt alignment before predicting the class (Bui et al., 2026). In these methods, "prototype" often denotes a prompt, retrieval key, or pseudo-labeling anchor rather than a stored class centroid, and several route predictions through task-specific components rather than one expanding head.

Closest to PACE's stability question, MICIL combines rehearsal with output- and embedding-level distillation from the previous model while keeping the aggregation pathway trainable (Meseguer et al., 2024), and Li et al. (2025) further show, by recombining attention and classifier components across tasks, that much of the old-task degradation follows the attention component, which they constrain by distillation while retaining patch features as pseudo-bags. That aggregation matters is therefore established. PACE differs in fixing the aggregator outright rather than constraining a trainable one, which removes aggregation-induced coordinate drift exactly and leaves classifier drift to prototype rehearsal (Table 4 compares where each method intervenes).

Continual WSI prognosis. Continual survival methods differ in objective and assumptions: ConSurv combines multimodal WSI–genomic modeling with mixture-of-experts and feature-constrained replay (Yu et al., 2026), and MergeSurv merges task-specific survival models without retaining task

data (Tran et al., 2026). PACE's prototype-free extension is deliberately narrower—a fixed unimodal aggregator with frozen cohort-specific Cox heads—testing the fixed-aggregation principle under prognosis rather than the full diagnostic method.

3 PROBLEM FORMULATION

Class-incremental WSI classification. We study class-incremental continual learning: task boundaries and arriving classes are known during training, but task identity is unavailable at inference. Task $t \in \{1, \dots, T\}$ provides $\mathcal{D}_t = \{(X_i, y_i)\}_{i=1}^{N_t}$ with $y_i \in \mathcal{C}_t$, the class sets are disjoint ($\mathcal{C}_p \cap \mathcal{C}_q = \varnothing$ for $p \neq q$), and the cumulative label space is $\mathcal{Y}_{\leq t} = \bigsqcup_{s=1}^{t} \mathcal{C}_s$. Once task s is complete, $\mathcal{D}_s$ is no longer accessible; a learner may carry model parameters and a method-specific memory $\mathcal{M}_{\leq t-1}$ into task t within the protocol's budget. PACE's memory is at most K labeled slide-space centroids per observed class (Sec. 4.3), never a WSI, patch bag, or pseudo-sample, and is not accessed at inference.

Each WSI is a variable-size bag $X_i = \{h_{i,j}\}_{j=1}^{n_i}$ of patch features $h_{i,j} = g(p_{i,j}) \in \mathbb{R}^{d_h}$ from a frozen encoder g; the subscript t indexes the model after training through task t. At stage t the aggregation map $f_{\psi_t} : \bigcup_{n \geq 1} (\mathbb{R}^{d_h})^n \to \mathbb{R}^d$ comprises every post-encoder operation that yields the slide representation—any feature projection, attention scoring, and pooling—and an expanding linear classifier produces logits over the cumulative label space:

$$z_i^{(t)} = f_{\psi_t}(X_i), \qquad \ell_t(X_i) = W_t z_i^{(t)} + b_t, \qquad W_t \in \mathbb{R}^{|\mathcal{Y}_{\leq t}| \times d}, \;\; b_t \in \mathbb{R}^{|\mathcal{Y}_{\leq t}|}. \tag{1}$$

The current-task loss is softmax cross-entropy over $\mathcal{Y}_{\leq t}$; a continual method may add rehearsal or regularization terms that use only its permitted state. At inference, prediction is made over all observed classes,

$$\widehat{y}_t(X) = \arg\max_{c \in \mathcal{Y}_{\leq t}} [\ell_t(X)]_c, \tag{2}$$

and the objective is to learn $\mathcal{C}_t$ while retaining $\mathcal{Y}_{\leq t-1}$ under these constraints.

Aggregation-induced drift. Although g is fixed, a change in ψ can give $f_{\psi_t}(X) \neq f_{\psi_s}(X)$ for the same bag at stages $s < t$, and changes to (W_t, b_t) (the classifier weights and bias at stage t) can independently affect performance on earlier tasks. To measure movement of the slide representation directly, for a given stage t, let $\mathcal{Q}$ be a fixed set of feature bags whose classes were introduced before that stage, with t_X the task that introduced the class of $X \in \mathcal{Q}$; the same-slide relative drift is

$$\Delta_{\text{emb}}(t) = \frac{1}{|\mathcal{Q}|} \sum_{X \in \mathcal{Q}} \frac{\|f_{\psi_t}(X) - f_{\psi_{t_X}}(X)\|_2}{\|f_{\psi_{t_X}}(X)\|_2}, \tag{3}$$

which isolates aggregation-induced movement only when the bags and all evaluation-time operations are identical and deterministic.

Cohort-incremental WSI prognosis. As an extension, the prognosis study is a separate multi-head, task-incremental setting with cohort boundaries and cohort identity known during training and inference. Step $t \in \{1, \dots, T_{\text{surv}}\}$ provides a cohort $\mathcal{S}_t$ of bags with follow-up times and event indicators; a shared aggregation map f_{ϕ_t} feeds a cohort-specific linear risk head $\omega_t \in \mathbb{R}^d$, trained with the Cox partial likelihood (Cox, 1972) (Eq. (18), Sec. E.3) and frozen once its cohort is learned. Completed cohorts' bags are unavailable except to methods with a permitted memory; PACE retains none. At stage $t \geq s$, cohort s is scored by $\omega_s^\top f_{\phi_t}(X)$ on its fixed test partition, giving a concordance matrix $R[t, s]$ (Harrell et al., 1982) whose changes across stages, with fixed bags and deterministic evaluation, arise only from the shared aggregator.

Research questions. Diagnosis is evaluated by end-of-stream average balanced accuracy (AACC) and forgetting, and prognosis by the final macro-averaged C-index (definitions: Sec. 5). We address four research questions: **(RQ1)** whether a trainable aggregator yields nonzero $\Delta_{\text{emb}}(t)$ for identical bags despite a frozen patch encoder; **(RQ2)** whether prototype rehearsal reduces forgetting more under fixed than trainable aggregation in matched controls, controlling for initialization; **(RQ3)** whether compact rehearsal suffices, quantified as the fraction of the rehearsal-induced AACC gain that a single prototype per class (K=1) recovers relative to the largest count in the validation grid (K=16); and **(RQ4)** whether the fixed-representation survival extension matches or exceeds the best final macro-averaged C-index among continual baselines under each cohort ordering, with order sensitivity given by the range across orderings.

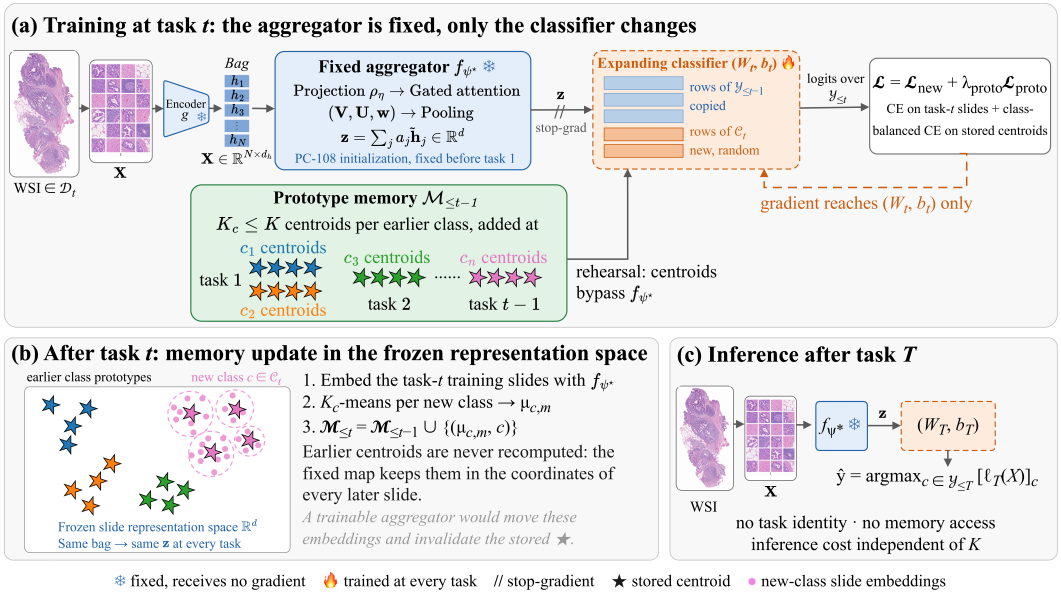

Figure 2: Overview of PACE: a fixed aggregation map with prototype rehearsal at an expanding classifier. **(a)** At task t, a frozen encoder g and the fixed aggregation map $f_{\psi^\star}$ (projection, gated attention, pooling; PC-108 initialization, fixed before task 1) produce the slide embedding $z \in \mathbb{R}^d$; only the expanding classifier (W_t, b_t) is trained, on task-t slides plus stored centroids of earlier classes, which bypass $f_{\psi^\star}$ (Eqs. (8)–(10)). **(b)** After task t, K_c-means on the task-t embeddings adds $K_c \leq K$ centroids per new class to the memory (Eq. (7)); earlier centroids are never recomputed (Eq. (6)). **(c)** Inference uses $f_{\psi^\star}$ and the cumulative classifier over $\mathcal{Y}_{\leq T}$, with no task identity or memory access. Solid blue boxes are fixed; dashed orange boxes are trained.

4 METHOD

4.1 PACE FRAMEWORK

PACE uses the standard Gated Attention-Based MIL (Gated ABMIL) architecture (Ilse et al., 2018), which maps a variable-size patch-feature bag to an attention-pooled slide representation and then to class logits. Rather than introduce a new pooling operator, PACE fixes the complete post-encoder aggregation map before the continual stream—initialized from the PC-108 pan-cancer prior (Shao et al., 2025) and frozen before task 1—and couples it to prototype rehearsal at an expanding linear classifier (Fig. 2). With the aggregation map fixed and evaluated deterministically (Sec. 4.2), slides from all tasks and stored centroids then share one coordinate system, so rehearsal constrains the classifier's earlier decision regions instead of compensating for the slide-representation drift that a trainable aggregation map would induce.

At task t, PACE expands the classifier to the cumulative label space and updates only (W_t, b_t), using current-task slides together with any stored centroids of earlier classes. It then embeds the task-t training slides with the fixed aggregation map, summarizes each new class by K-means, and keeps the resulting labeled centroids for later tasks; the WSIs, patch bags, and pseudo-samples are not kept. Inference uses the fixed aggregation map and the cumulative classifier, without task identity or access to the memory. We next (1) specify the fixed slide-embedding function and its invariance conditions (Sec. 4.2), (2) define the multi-prototype memory and its update (Sec. 4.3), and (3) present classifier expansion and the full rehearsal objective (Sec. 4.4); the survival extension, inference, and cost follow.

4.2 FIXED SLIDE-EMBEDDING FUNCTION

Writing any post-encoder feature projection as $\widetilde{h}_j = \rho_\eta(h_j) \in \mathbb{R}^d$ (ρ_η is the identity and $d_h = d$ when no projection is used), Gated ABMIL scores and pools the projected instances as

$$q_j = \tanh(V\widetilde{h}_j + b_V) \odot \sigma(U\widetilde{h}_j + b_U), \qquad e_j = w^\top q_j + b_w, \tag{4}$$

$$a_j = \frac{\exp(e_j)}{\sum_{l=1}^{n} \exp(e_l)}, \qquad z(X) = \sum_{j=1}^{n} a_j \widetilde{h}_j, \tag{5}$$

with $V, U \in \mathbb{R}^{L\times d}$, $w, b_V, b_U \in \mathbb{R}^L$, and $b_w \in \mathbb{R}$. We write $\psi = \{\eta, V, U, w, b_V, b_U, b_w\}$ for the complete learnable post-encoder parameters. PACE fixes $\psi = \psi^\star$ throughout the diagnostic stream, and the prognostic model fixes its independently instantiated $\phi = \phi^\star$. For an identical feature bag X and any stages s, t, deterministic evaluation with the complete map fixed gives

$$a_j^{(t)}(X) = a_j^{(s)}(X), \qquad f_{\psi_t}(X) = f_{\psi_s}(X) = f_{\psi^\star}(X) = z(X). \tag{6}$$

Intuitively, once the map is frozen, the same slide always pools to the same point, so a class summary computed in this space stays valid as later tasks arrive. The guarantee requires identical patch features and preprocessing, no stochastic tile subsampling or test-time augmentation, and deterministic evaluation. It fixes the coordinates of the slide embedding, not the logits or predictions, and does not make a stored centroid representative under distribution shift.

4.3 Multi-prototype class memory

PACE stores K-means centroids as class summaries (K=1 is the class mean), since a single mean can conceal separated regions of a class; whether more centroids help is assessed empirically by the K ablation (Sec. 6.1). After the classifier is optimized for task t, PACE embeds only that task's training slides: for each new class $c \in \mathcal{C}_t$ it collects $\mathcal{Z}_t^c = \{f_{\psi^\star}(X) : (X, y) \in \mathcal{D}_t,\ y = c\}$, runs K_c-means with $K_c = \min(K, |\mathcal{Z}_t^c|)$ into clusters $\{\mathcal{Z}_{t,m}^c\}_{m=1}^{K_c}$, and stores

$$\mu_{c,m} = \frac{1}{|\mathcal{Z}_{t,m}^c|} \sum_{z\in\mathcal{Z}_{t,m}^c} z, \qquad \mathcal{M}_{\le t} = \mathcal{M}_{\le t-1} \cup \{(\mu_{c,m}, c) : c \in \mathcal{C}_t,\ m \le K_c\}, \qquad \mathcal{M}_{\le 0} = \varnothing. \tag{7}$$

Intuitively, several centroids let a class occupy a few separate regions of the space instead of collapsing to one blurred point. Earlier centroids are never recomputed; Eq. (6) keeps them expressed in the same coordinates as later slide embeddings. The memory holds labeled, patient-derived statistics but no pixels or patch bags (prototype atlases: Sec. D).

4.4 Classifier expansion and prototype rehearsal

At task t the classifier expands from $\mathcal{Y}_{\le t-1}$ to $\mathcal{Y}_{\le t}$: old rows are copied, $W_t[c] = W_{t-1}[c]$ and $b_t[c] = b_{t-1}[c]$, new rows are randomly initialized, the aggregation map stays fixed, and gradients update only (W_t, b_t). With $\mathrm{CE}(v, y)$ the softmax cross-entropy of logits v over $\mathcal{Y}_{\le t}$,

$$\mathcal{L}_{\text{new}}^{(t)} = \frac{1}{|\mathcal{D}_t|} \sum_{(X,y)\in\mathcal{D}_t} \mathrm{CE}\big(W_t f_{\psi^\star}(X) + b_t, y\big), \tag{8}$$

$$\mathcal{L}_{\text{proto}}^{(t)} = \frac{1}{|\mathcal{Y}_{\le t-1}|} \sum_{c\in\mathcal{Y}_{\le t-1}} \frac{1}{K_c} \sum_{m=1}^{K_c} \mathrm{CE}\big(W_t \mu_{c,m} + b_t, c\big), \tag{9}$$

$$\mathcal{L}^{(t)} = \mathcal{L}_{\text{new}}^{(t)} + \lambda_{\text{proto}} \mathcal{L}_{\text{proto}}^{(t)}, \qquad \lambda_{\text{proto}} \ge 0, \quad \mathcal{L}_{\text{proto}}^{(1)} = 0. \tag{10}$$

Intuitively, the prototype term replays each earlier class as a few labeled points, so the expanding classifier keeps a decision region for it without revisiting any slides. In minibatch training, current-task batches are paired with prototype batches sampled uniformly over earlier classes and then over their centroids, the stochastic counterpart of Eq. (9); λ_{proto} and K are selected on validation data (Sec. 5). Prototypes for $\mathcal{C}_t$ are built only after this optimization and enter rehearsal from task $t{+}1$; the complete task loop is given as pseudo-code in Sec. A.2.

Extension: cohort-incremental survival prognosis. The prognosis model is trained independently: one complete aggregation map $f_{\phi^\star}$ is fixed across cohorts, a head ω_t is initialized for cohort t, optimized alone with Eq. (18), and frozen when the cohort is complete; no prototype memory is used because no shared classifier expands. Under the fixed-input conditions of Sec. 3, $\omega_s^\top f_{\phi^\star}(X)$ is identical at every stage $t \ge s$, so $R[t, s] = R[s, s]$: forgetting and backward transfer are exactly zero by construction, which also precludes positive backward transfer.

Inference and cost. For diagnosis, Eq. (2) uses the frozen map $f_{\psi^\star}$ and the cumulative classifier, requiring neither task identity nor prototype access; for prognosis, the known cohort identity selects its frozen head, $r_s(X) = \omega_s^\top f_{\phi^\star}(X)$. The prototype memory grows with classes, not slides—at most 32 KB per class at K=16 and d=512 (float32)—and inference cost is independent of K (full cost accounting: Sec. B.2).

5 EXPERIMENTAL SETUP

Diagnostic benchmarks and curricula. We evaluate continual diagnosis on seven public WSI cohorts. The primary benchmark is the Cooperative Human Tissue Network (CHTN) (National Cancer Institute, 2025): $10{,}359$ H&E WSIs from $8{,}246$ patients over 17 cancer and histology classes. The cross-cohort benchmark combines CAMELYON16/17 (Ehteshami Bejnordi et al., 2017; Litjens et al., 2018), TCGA-NSCLC/RCC/BRCA (Cancer Genome Atlas Research Network, 2014; 2013; Cancer Genome Atlas Network, 2012), and UBC-OCEAN (Asadi-Aghbolaghi et al., 2024) ($4{,}423$ WSIs, 16 classes); all splits are patient-stratified $70/15/15$. From these we build 11 deterministic class-incremental curricula, fixed before any result was inspected: eight on CHTN—CANCER FAMILY, FREQUENT-FIRST, RARE-FIRST, GROUPED INCREMENTS, DIAGNOSTIC TRIAGE, PRIMARY SITE, ORGAN SUBTYPE, and METASTATIC SITE—and three re-partitioning the cross-cohort classes into 6, 4, and 3 tasks (CROSS-COHORT-6/4/3) (Huang et al., 2023). Curriculum names, task/class counts, and compositions are detailed in Table 1 and Sec. E.

Baselines and comparison protocol. We compare with 21 continual-learning methods adapted to WSI-MIL: EWC and LwF (regularization) (Kirkpatrick et al., 2017; Li & Hoiem, 2017); iCaRL, ER, ER-ACE, DER++, A-GEM, and GDumb (replay) (Rebuffi et al., 2017; Rolnick et al., 2019; Caccia et al., 2022; Buzzega et al., 2020; Chaudhry et al., 2019; Prabhu et al., 2020); BiC and WA (bias correction) (Wu et al., 2019; Zhao et al., 2020); ACIL, SimpleCIL, FeTrIL, and PASS (feature/prototype) (Zhuang et al., 2022; Zhou et al., 2025; Petit et al., 2023; Zhu et al., 2021); LoRA and InfLoRA (parameter-efficient) (Hu et al., 2022; Liang & Li, 2024); and the WSI-specific MICIL, ConSlide, CoMEL, MergeSlide, and AdvMIL (Meseguer et al., 2024; Huang et al., 2023; Lee et al., 2025; Bui et al., 2026; Li et al., 2025); fine-tuning and joint training bracket the range. All methods share the frozen features, the PC-108-initialized Gated ABMIL aggregator, the expanding-classifier protocol, the splits, and validation-based tuning where their formulation permits (deviations: Sec. B.2). Beyond their continual-learning mechanisms, the one difference is that baselines adapt the aggregator while PACE keeps it fixed; SimpleCIL and FeTrIL, which also freeze it, are its closest baselines (Sec. 7). The PC-108 prior was pretrained on a separate corpus of roughly 24,000 slides disjoint from every evaluation cohort, so the shared initialization introduces no test-set leakage. Replay methods store 10 WSIs per class ($\approx$50k patch features); PACE stores none.

Survival-extension benchmark. Beyond classification, a prognosis study streams six cohorts under five orderings: five TCGA cohorts (TCGA-BRCA/KIRC/BLCA/UCEC/COAD-READ) scored on overall survival and the external SurGen colorectal cohort ($1{,}020$ WSIs) on recurrence-free survival (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2013; Myles et al., 2025). A separate fixed Gated ABMIL map ($f_{\phi^\star}$, UNI-v2 features) feeds cohort-specific Cox heads (Eq. (18)) that are frozen after their cohort; we adapt the 13 compatible baselines (protocol and results: Sec. E.3).

Evaluation metrics and implementation details. For diagnosis, we report AACC ($\uparrow$), the balanced accuracy on each task's test set after the final task, averaged over tasks with equal weight and with predictions over all observed classes, and forgetting ($\downarrow$), the mean drop from each earlier task's best accuracy to its final accuracy. For prognosis, we report the per-cohort C-index (Harrell et al., 1982) and, as the primary criterion, its macro-average over cohorts after the final step ($\uparrow$). Formal definitions, BWT, and intransigence are in Sec. B.1, and memory footprint in Sec. B.2. Results are mean $\pm$ sd over five runs (survival: five predefined patient-level folds, not independent replicates); we run no significance tests: the highest mean is reported as best (ties at three decimals), and a margin within run-to-run sd is read as parity rather than a win. Validation alone drives tuning and early stopping; test partitions are used once. WSIs are processed at $20\times$ with TRIDENT (Zhang et al., 2025; Vaidya et al., 2025), with frozen UNI features (Chen et al., 2024) for diagnosis and UNI-v2 for prognosis, pooled by Gated ABMIL (Ilse et al., 2018). Optimization uses AdamW with cosine decay (Loshchilov & Hutter, 2019; 2017); configurations, grids, and cost accounting are in Secs. B.2–B.3.

6 RESULTS

RQ1 is answered by Fig. 1: with the encoder frozen, the same-slide drift Δ_{emb} reaches 59.9% under fine-tuning and is 0 under PACE by construction. Sec. 6.1 compares PACE with baselines and closes with the ablation that addresses RQ2–RQ3; Sec. 6.2 addresses RQ4.

Table 1: Continual diagnosis on the CHTN curricula. Final average balanced accuracy (AACC, ↑), mean ± sd over five runs, at the 10-WSI budget (rehearsal methods store 10 WSIs per class, ≈50K patch features; PACE stores no slides). Curricula (tasks/classes): CF Cancer family (6/7); FF Frequent-first and RF Rare-first (13/14 each; the same 14 classes in opposite frequency order); GI Grouped increments (5/14); DT Diagnostic triage (7/17); OS Organ subtype (7/14); MS Metastatic site (7/8); PS Primary site (11/12). **Bold**/underline: best/second-best continual method per column, joint training excluded. PACE (ours) is the shaded row; joint training and fine-tuning are the upper and lower reference bounds.

	Cancer type				Tissue / site		Primary & metastatic morphology	
Method	**CF**	**FF**	**RF**	**GI**	**DT**	**OS**	**MS**	**PS**
Reference bounds								
Joint training	0.880±0.007	0.755±0.018	0.754±0.022	0.765±0.010	0.688±0.033	0.698±0.026	0.711±0.050	0.797±0.024
Fine-tuning	0.167±0.036	0.112±0.039	0.078±0.002	0.323±0.032	0.158±0.026	0.107±0.010	0.143±0.097	0.115±0.023
No Memory								
ACIL (Zhuang et al., 2022)	0.573±0.084	0.538±0.035	0.515±0.108	0.531±0.014	0.616±0.024	0.476±0.032	0.642±0.045	0.437±0.082
FeTrIL (Petit et al., 2023)	0.775±0.038	0.635±0.025	0.676±0.070	0.666±0.029	0.603±0.034	0.473±0.030	0.566±0.101	0.689±0.022
MergeSlide (Bui et al., 2026)	0.545±0.095	0.364±0.059	0.406±0.013	0.328±0.020	0.297±0.133	0.352±0.138	0.400±0.037	0.484±0.048
MICIL (Meseguer et al., 2024)	0.183±0.007	0.117±0.023	0.077±0.000	0.322±0.054	0.178±0.020	0.113±0.009	0.314±0.043	0.155±0.011
SimpleCIL (Zhou et al., 2025)	0.648±0.066	0.532±0.042	0.549±0.163	0.549±0.022	0.457±0.048	0.420±0.071	0.531±0.004	0.527±0.071
Regularization								
EWC (Kirkpatrick et al., 2017)	0.172±0.007	0.102±0.023	0.078±0.001	0.369±0.060	0.228±0.033	0.131±0.020	0.143±0.071	0.127±0.038
InfLoRA (Liang & Li, 2024)	0.263±0.010	0.279±0.054	0.159±0.004	0.333±0.030	0.244±0.030	0.163±0.022	0.330±0.001	0.305±0.023
LwF (Li & Hoiem, 2017)	0.169±0.003	0.116±0.042	0.078±0.001	0.267±0.025	0.156±0.016	0.137±0.022	0.143±0.098	0.143±0.060
Prompt / LoRA / Prototype								
LoRA (Hu et al., 2022)	0.174±0.005	0.081±0.002	0.077±0.000	0.293±0.021	0.144±0.008	0.109±0.008	0.170±0.014	0.092±0.001
CoMEL (Lee et al., 2025)	0.168±0.001	0.107±0.021	0.077±0.001	0.246±0.093	0.113±0.034	0.090±0.013	0.300±0.005	0.113±0.033
PASS (Zhu et al., 2021)	0.503±0.122	0.416±0.082	0.285±0.213	0.605±0.051	0.411±0.146	0.432±0.159	0.248±0.068	0.392±0.124
Memory / Rehearsal — 10 WSIs per class (≈ 50K patch features)								
AGEM (Chaudhry et al., 2019)	0.172±0.004	0.135±0.050	0.081±0.004	0.406±0.033	0.277±0.040	0.244±0.020	0.145±0.003	0.151±0.021
BiC (Wu et al., 2019)	0.617±0.088	0.458±0.042	0.301±0.066	0.481±0.070	0.345±0.016	0.329±0.055	0.614±0.024	0.378±0.072
ConSlide (Huang et al., 2023)	0.416±0.026	0.339±0.082	0.286±0.080	0.542±0.030	0.408±0.007	0.390±0.035	0.264±0.043	0.363±0.071
DER++ (Buzzega et al., 2020)	0.526±0.046	0.417±0.072	0.297±0.097	0.558±0.060	0.399±0.046	0.383±0.058	0.301±0.016	0.513±0.027
ER-ACE (Caccia et al., 2022)	0.488±0.166	0.429±0.028	0.398±0.114	0.564±0.061	0.539±0.039	0.434±0.082	0.517±0.022	0.452±0.085
GDumb (Prabhu et al., 2020)	0.571±0.131	0.410±0.043	0.428±0.045	0.502±0.053	0.267±0.003	0.282±0.020	0.463±0.191	0.323±0.003
iCaRL (Rebuffi et al., 2017)	0.712±0.078	0.656±0.025	0.645±0.050	0.642±0.050	0.578±0.026	0.556±0.032	**0.701**±0.002	0.666±0.045
ER (Rolnick et al., 2019)	0.557±0.018	0.424±0.075	0.199±0.062	0.581±0.016	0.397±0.041	0.363±0.026	0.606±0.015	0.485±0.101
WA (Zhao et al., 2020)	0.495±0.059	0.413±0.076	0.188±0.070	0.455±0.078	0.266±0.139	0.307±0.068	0.574±0.014	0.411±0.166
AdvMIL (Li et al., 2025)	0.703±0.006	**0.708**±0.005	0.537±0.027	0.692±0.009	0.579±0.008	0.555±0.023	0.685±0.088	0.624±0.040
PACE	**0.779**±0.070	0.652±0.105	**0.679**±0.041	**0.711**±0.057	**0.662**±0.025	**0.661**±0.014	0.643±0.042	**0.723**±0.077

6.1 Primary benchmark: continual diagnosis (Continual WSIs classification)

End-of-stream accuracy. Tables 1 and 5 report mean AACC at the 10-WSI budget across the eight CHTN curricula and the cross-cohort benchmark. PACE attains the highest mean AACC on six of the eight CHTN curricula and on two of the three cross-cohort curricula, tying AdvMIL on the third. Across its six CHTN leads, the mean margin over the strongest competitor is 3.5 points, and the largest is 10.5, on Organ subtype (0.661 vs. 0.556 for iCaRL). PACE trails on the other two CHTN curricula: on Frequent-first, AdvMIL leads 0.708 to 0.652, and on Metastatic site, both iCaRL (0.701) and AdvMIL (0.685) exceed PACE's 0.643. The baselines fall into two groups: methods that regularize or adapt a trainable aggregator without stored slides (EWC, LwF, LoRA, InfLoRA, MICIL, CoMEL), together with A-GEM, stay within 0.20 AACC of fine-tuning on every CHTN curriculum, whereas the exemplar-free feature-space methods (FeTrIL, ACIL, SimpleCIL) and the strongest replay methods (iCaRL, AdvMIL) form the competitive set.

Cross-cohort benchmark and task granularity. On the cross-cohort benchmark (Table 5), PACE leads by 3.1 points on Cross-cohort-6 (0.853 vs. 0.822 for iCaRL) and by 1.6 on Cross-cohort-4, and ties AdvMIL on Cross-cohort-3 (0.843). These three curricula split the same 16 classes into 6, 4, and 3 tasks, isolating the effect of task granularity: across them PACE's accuracy varies by only 1.0 point, against 34 for fine-tuning, so its gap to joint training barely depends on how often the classifier expands.

Arrival order and retention. Frequent-first and Rare-first present the same 14 classes in opposite orders of different cancer types, with frequent cancer class first and rare cancer classes second, respectively. Reversing the order moves PACE by 2.7 points (0.652 vs. 0.679) and FeTrIL by 4.1, but drops AdvMIL from 0.708 to 0.537. PACE leads Rare-first (at parity with FeTrIL) and is third on Frequent-first: arrival order changes which method leads, while PACE itself is comparatively order-insensitive. Forgetting alone can mislead: ER-ACE forgets less than PACE on Cancer family yet reaches only 0.488 AACC against PACE's 0.779 (Table 10).

Freezing, prototype count, and replay weight (RQ2–RQ3). On Cancer family we jointly ablate K, λ_{proto}, the aggregator regime (frozen at initialization; trained on the base task, then frozen; trained throughout), and the initialization of the attention parameters V, U, w (PC-108 prior or Xavier;

the PC-108 projection stays fixed in every regime); each regime has its own $\lambda_{\text{proto}}=0$ control (grids: Figs. 4, 5, Tables 7–9). Two observations follow. (i) *One prototype per class is already effective, and memory trades off against replay pressure*: at $\lambda_{\text{proto}}=10$, $K=1$ lifts AACC from 0.173 without replay to 0.738, 82% of the gain reached at $K=16$ (0.859), and cuts forgetting from 0.984 to 0.291 (0.112 at $K=16$); at $\lambda_{\text{proto}}=0.1$, larger K brings no consistent gain. (ii) *Freezing, not initialization, makes rehearsal effective, and neither freezing nor rehearsal suffices alone*: all six no-rehearsal controls sit at the fine-tuning level (0.167–0.173 AACC) whether or not the aggregator is fixed, and adding rehearsal ($K=1$, $\lambda_{\text{proto}}=10$) raises AACC by 0.55–0.58 in the four fixed-aggregator regimes but by only 0.02–0.22 in the two trained-throughout regimes, where no $(K, \lambda_{\text{proto}})$ pair reaches 0.55 even at $K=1024$ (Sec. D). In the fixed regimes, Xavier and PC-108 attention differ by at most 0.026 AACC for $K \leq 16$ at $\lambda_{\text{proto}}=10$ (e.g., 0.717 vs. 0.738 at $K=1$; Tables 7, 8).

6.2 Extension: continual prognosis (continual survival) (RQ4)

On cohort-incremental survival, with frozen per-cohort Cox heads, PACE has zero forgetting by construction and the highest mean C-index under all five orderings (0.584–0.603; Table 6). Its margin over the runner-up (0.8–2.5 points) is smaller than in diagnosis, as fine-tuning (0.549–0.585) nearly matches joint training (0.624). We attribute this to the frozen PC-108 aggregator, a pan-cancer classification prior whose slide-level representation is better aligned with diagnosis than prognosis, leaving little room for continual method to improve.

7 Discussion

The diagnostic results answer the question of Sec. 1: compact prototype rehearsal becomes effective once the aggregation map is fixed. The ablation over aggregator regimes and prototype counts shows why (Sec. 6.1; Figs. 4, 5; Tables 7–9). Fixing the map alone leaves AACC at the fine-tuning level (0.173); adding one prototype per class then raises it to 0.738, whereas a map trained throughout never reaches 0.55. Even randomly initialized and then frozen attention (Xavier V, U, w; the PC-108 projection is kept) is nearly as effective as the PC-108 prior (0.717 vs. 0.738). This suggests that the learned attention adds little here: fixed attention only reweights the pretrained patch features (Eq. (5)), so the class evidence comes mainly from the pretrained encoder, reported as the most consequential design choice in slide classification (Wölflein et al., 2025). The classifier can then carry continual learning, as it does on frozen pretrained features in natural images (Zhou et al., 2025; McDonnell et al., 2023), provided the map stays fixed.

Frozen features with one class mean, as in FeTrIL and SimpleCIL (which share PACE's initialized aggregator), are the closest baselines to PACE: FeTrIL is the runner-up on three curricula, yet PACE's mean is higher on all eleven. The K ablation attributes part of this gap to the multi-centroid memory (0.738 at $K=1$ vs. 0.859 at $K=16$); qualitatively, the interpretability maps show a class's stored centroids capturing distinct within-class tissue morphologies (Sec. D).

PACE does not lead where classes are smallest or most heterogeneous: AdvMIL is stronger on Frequent-first, and both AdvMIL and iCaRL on Metastatic site. This is consistent with, though not evidence for, a centroid summary being limited by class size; the cap K_c (Sec. 4.3) makes that limit explicit for tail classes. On survival, PACE still leads under every ordering but by smaller margins, reflecting a pretrained representation better suited to diagnosis than to prognosis (Sec. 6.2).

8 Conclusion

We asked whether fixing the aggregation map makes compact prototype rehearsal effective in class-incremental WSI classification. It does: PACE fixes the gated-attention aggregator and rehearses K centroids per class at an expanding classifier. Across 11 curricula and 21 baselines, it has the highest mean AACC on eight and ties on one, with four margins beyond run-to-run variation. Extended without prototypes to cohort-incremental prognosis, it has the highest macro-averaged C-index under all five orderings. A matched ablation on one curriculum shows that freezing the attention, not its initialization, is what makes rehearsal effective, and that one prototype per class already recovers 82% of the gain of $K=16$. Slide-retaining methods remain stronger where classes are smallest or most heterogeneous. Within that limit, one design serves both tasks: with the aggregation map fixed, compact, slide-free rehearsal carries continual WSI diagnosis and frozen Cox heads carry cohort-incremental prognosis.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. It uses only previously collected, de-identified datasets obtained under their respective access terms—CHTN, TCGA (NSCLC, RCC, BRCA, KIRC, BLCA, UCEC, COAD-READ), CAMELYON16/17, UBC-OCEAN, and SurGen (Sec. 5); no new data were collected. These data contain sensitive health information (cancer diagnoses and survival outcomes); we used no direct identifiers or demographic attributes, and no re-identification was attempted. PACE retains no slides or patch features between tasks, only class-conditional centroids of slide embeddings; these are still derived from patient data, and we make no formal privacy claim for them. The models are research prototypes for studying continual learning: they are not medical devices and have not been clinically validated, and the survival risk scores were evaluated for ranking only (C-index; Sec. 6.2), so neither the diagnostic nor the prognostic outputs should inform patient care. Balanced accuracy accounts for class imbalance; fairness across patient subgroups was not evaluated. A foreseeable risk is premature clinical deployment, which we caution against.

REPRODUCIBILITY STATEMENT

The paper and supplement are organized for end-to-end reproduction. The class- and cohort-incremental protocols are formalized in Sec. 3; every PACE component is specified by Eqs. (4)–(10) and Eq. (18), with the invariance and gradient-isolation derivations in Sec. A and training, memory-construction, and inference pseudo-code in Sec. A.2. All datasets were previously collected and are obtainable under their respective access terms (Sec. 5); cohort compositions, the 11 deterministic curricula with their patient-stratified $70/15/15$ splits, and the five survival orderings are documented in Secs. E and E.3, and metric definitions in Sec. B.1. Preprocessing (TRIDENT at $20\times$, frozen UNI/UNI-v2 features, the PC-108 prior), training schedules, and the complete search grids for PACE and all 21 baselines are given in Sec. 5 and Secs. B.2–B.3 (Tables 2 and 3); the full ablation grids are in Sec. D. Diagnostic results are mean $\pm$ sd over five runs and survival results over the five released patient-level folds; validation data alone drive model selection and early stopping, and test partitions are used once.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

A large language model was used as a writing and editing assistant: it revised the wording and structure of the manuscript and of its figure and table captions for clarity and consistency, checked that the numbers quoted in the text and captions match the reported tables (including simple derived quantities such as margins and percentages of gains), helped resolve LaTeX compilation issues, and drafted this and the preceding statements from author-provided facts. It also provided small, localized code fixes (e.g., resolving syntax errors and missing imports), which the authors tested and which did not change the method or the experimental logic. It did not contribute to research ideation, method design, experimental design, data processing, model training, or figure generation, and it did not perform statistical analysis or interpret the scientific results. The authors reviewed every LLM-assisted edit against their experimental records and take full responsibility for the manuscript.

REFERENCES

Maryam Asadi-Aghbolaghi, Hossein Farahani, Allen Zhang, Ardalan Akbari, Sirim Kim, Ashley Chow, Sohier Dane, OCEAN Challenge Consortium, OTTA Consortium, David G Huntsman, et al. Machine learning-driven histotype diagnosis of ovarian carcinoma: insights from the ocean ai challenge. *medRxiv*, pp. 2024–04, 2024.

Mohsin Bilal, Manahil Raza, Youssef Altherwy, Anas Alsuhaibani, Abdulrahman Abduljabbar, Fahdah Almarshad, Paul Golding, Nasir Rajpoot, et al. Foundation models in computational pathology: A review of challenges, opportunities, and impact. *arXiv preprint arXiv:2502.08333*, 2025.

Doanh C Bui, Ba Hung Ngo, Hoai Luan Pham, Khang Nguyen, Maï K Nguyen, and Yasuhiko Nakashima. Mergeslide: Continual model merging and task-to-class prompt-aligned inference for lifelong learning on whole slide images. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4859–4868. IEEE, 2026.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33: 15920–15930, 2020.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2022.

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

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418): 61, 2012.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013.

Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern recognition*, 77:329–353, 2018.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021.

Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696, 2021.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAMELYON16 consortium, Meyke Hermsen, Quirine F Manson, 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.

Jiaxiang Gou, Luping Ji, Pei Liu, and Mao Ye. Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 3158–3166, 2025.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Yanyan Huang, Weiqin Zhao, Shujun Wang, Yu Fu, Yuming Jiang, and Lequan Yu. 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.

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

Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Byung Hyun Lee, Wongi Jeong, Woojae Han, Kyoungbun Lee, and Se Young Chun. Continual multiple instance learning with enhanced localization for histopathological whole slide image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23232–23242, 2025.

Xianrui Li, Yufei Cui, Jun Li, and Antoni B. Chan. Advancing multiple instance learning with continual learning for whole slide imaging. *arXiv preprint arXiv:2505.10649*, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91:103020, 2024.

Wenzhuo Liu, Xinjian Wu, Fei Zhu, Mingming Yu, Chuang Wang, and Cheng-Lin Liu. Class incremental learning with self-supervised pre-training and prototype learning. *arXiv preprint arXiv:2308.02346*, 2023.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

David N Louis, Michael Feldman, Alexis B Carter, Anand S Dighe, John D Pfeifer, Lynn Bry, Jonas S Almeida, Joel Saltz, Jonathan Braun, John E Tomaszewski, et al. Computational pathology: a path ahead. *Archives of pathology & laboratory medicine*, 140(1):41–50, 2016.

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

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Pablo Meseguer, Rocío Del Amor, and Valery Naranjo. Micil: Multiple-instance class-incremental learning for skin cancer whole slide images. *Artificial intelligence in medicine*, 152:102870, 2024.

Craig Myles, In Hwa Um, Craig Marshall, David Harris-Birtill, and David J Harrison. Surgen: 1020 h&e-stained whole-slide images with survival and genetic markers. *GigaScience*, 14:giaf086, 2025.

National Cancer Institute. Cooperative Human Tissue Network (CHTN). `https://www.chtn.org`, 2025. Accessed: 2026.

Anh Tien Nguyen, Keunho Byeon, Kyungeun Kim, Boram Song, Seoung Wan Chae, and Jin Tae Kwak. CAMP: Continuous and adaptive learning model in pathology. *npj Artificial Intelligence*, 2:2, 2026. doi: 10.1038/s44387-025-00051-5. URL `https://www.nature.com/articles/s44387-025-00051-5`.

Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3911–3920, 2023.

Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pp. 524–540. Springer, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54219–54238, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Vu Minh Tran, Doanh C. Bui, Maï K. Nguyen, and Khang Nguyen. Mergesurv: Merging-based continual learning for survival analysis on whole-slide images. *arXiv preprint arXiv:2607.04747*, 2026.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

Georg Wölflein, Dyke Ferber, Asier R. Meneghetti, Omar S. M. El Nahhas, Daniel Truhn, Zunamys I. Carrero, David J. Harrison, Ognjen Arandjelović, and Jakob Nikolas Kather. A good feature extractor is all you need for weakly supervised pathology slide classification. In *Computer Vision – ECCV 2024 Workshops*, volume 15638 of *Lecture Notes in Computer Science*, pp. 68–87. Springer, 2025.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 374–382, 2019.

Jinxi Xiang, Xiyue Wang, Jun Zhang, Sen Yang, Xiao Han, and Wei Yang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *International Conference on Learning Representations (ICLR)*, 2023.

Linfeng Ye, Shayan Mohajer Hamidi, Zhixiang Chi, Guang Li, Mert Pilanci, Takahiro Ogawa, Miki Haseyama, and Konstantinos N. Plataniotis. ASMIL: Attention-stabilized multiple instance learning for whole-slide imaging. In *International Conference on Learning Representations (ICLR)*, 2026.

Dianzhi Yu, Conghao Xiong, Yankai Chen, Wenqian Cui, Xinni Zhang, Yifei Zhang, Hao Chen, Joseph J. Y. Sung, and Irwin King. Consurv: Multimodal continual learning for survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27899–27907, 2026. doi: 10.1609/aaai.v40i33.40013.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

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

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13208–13217, 2020.

Weixi Zheng, Aoling Huang, Jingping Yuan, Haoyu Zhao, Zhou Zhao, Yongchao Xu, and Thierry Géraud. Pathological prior-guided multiple instance learning for mitigating catastrophic forgetting in breast cancer whole slide image classification. In *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10890868.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5871–5880, 2021.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. Acil: Analytic class-incremental learning with absolute memorization and privacy protection. *Advances in Neural Information Processing Systems*, 35:11602–11614, 2022.