Footprint-Guided Exemplar-Free Continual Histopathology Report Generation

Pratibha Kumari[1⋆@], Daniel Reisenbüchler[1⋆], Afshin Bozorgpour[1], Yousef Sadegheih[1], Priyankar Choudhary[2], and Dorit Merhof[1,3]

[1] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
[2] Indian Institute of Information Technology Bhopal, India
[3] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
⋆Equal contribution, @Correspondence (Pratibha.Kumari@ur.de)

Abstract. Rapid progress in vision-language modeling has enabled pathology report generation from gigapixel whole-slide images, but most approaches assume static training with simultaneous access to all data. In clinical deployment, however, new organs, institutions, and reporting conventions emerge over time, and sequential fine-tuning can cause catastrophic forgetting. We introduce an exemplar-free continual learning framework for WSI-to-report generation that avoids storing raw slides or patch exemplars. The core idea is a compact domain footprint built in a frozen patch-embedding space: a small codebook of representative morphology tokens together with slide-level co-occurrence summaries and lightweight patch-count priors. These footprints support generative replay by synthesizing pseudo-WSI representations that reflect domain-specific morphological mixtures, while a teacher snapshot provides pseudo-reports to supervise the updated model without retaining past data. To address shifting reporting conventions, we distill domain-specific linguistic characteristics into a compact style descriptor and use it to steer generation. At inference, the model identifies the most compatible descriptor directly from the slide signal, enabling domain-agnostic setup without requiring explicit domain identifiers. Evaluated across multiple public continual learning benchmarks, our approach outperforms exemplar-free and limited-buffer rehearsal baselines, highlighting footprint-based generative replay as a practical solution for deployment in evolving clinical settings.

Keywords: WSI report generation · Continual Learning · Style shift.

1 Introduction

Gigapixel whole slide images (WSIs) contain rich morphological information that underlies diagnostic and prognostic decisions in surgical pathology. In routine workflows, this information is summarized in free-text reports that describe salient findings and frequently follow semi-structured conventions (for example, diagnosis, grade, stage, and margin status). Automating report generation from

WSIs is therefore an important step toward reducing documentation burden and enabling more consistent reporting. Recent vision-language models have begun to translate WSIs into coherent diagnostic narratives by combining slide encoders with large language models (LLMs), enabling multi-scale reasoning over billions of pixels and producing clinically meaningful text [4,24,24].

Despite this progress, most WSI report-generation systems are developed in a static training regime that assumes simultaneous access to multiple datasets spanning different organs, institutions, scanners, and patient populations. In practice, deployment is inherently continual, with new datasets (domains) arriving over time. Repeated cumulative retraining on all observed data can become computationally costly and may be infeasible when prior data cannot be retained due to storage and governance constraints [21]. Naïve fine-tuning on incoming data can lead to catastrophic forgetting (CF), where performance degrades on previously learned domains [22,8,10]. This challenge is pronounced in report generation when domain shifts affect both the input distribution and the report language, requiring CL methods that preserve input-output drift over time.

A widely used mechanism to reduce forgetting is rehearsal, where past examples are interleaved with current-domain training [20,17,1]. In computational pathology, however, rehearsal is challenging because WSIs are naturally represented as variable-length sets of patch embeddings, making storage costly even when retaining only intermediate features rather than raw pixels; in addition, long-term retention of patient data may be restricted in some settings [11]. Beyond exemplar replay, prompt-based CL for LLMs learns task-specific soft prompts while keeping the backbone frozen, but it typically assumes task or domain identity at inference to apply the appropriate prompt sequence [18]. Relatedly, recent work on continual radiology report generation uses parameter-efficient tuning, learning a small set of domain-specific parameters while freezing the LLM backbone [23]; however, such strategies still maintain domain-specific components whose overhead grows as domains accumulate. In computational pathology, domain identifiers or metadata may be missing, inconsistent across sites, or insufficient to reflect shifts in reporting conventions, which can reduce robustness at deployment. Together, these considerations motivate a CL framework for WSI report generation that is storage-efficient, avoids retaining raw exemplars, and supports domain-agnostic inference.

In this work, we introduce a footprint-based generative replay framework for continual WSI report generation. Building on a HistoGPT-style vision-language generator [24], we address the core constraint of continual deployment: retaining prior-domain competence without keeping WSIs or large feature archives. Our key idea is to compress each domain into a compact domain footprint in a frozen embedding space, combining a representative codebook with lightweight slide-level composition statistics. These footprints enable exemplar-free rehearsal by synthesizing pseudo-WSIs, while pseudo reports are obtained from an immediate teacher to provide consistent supervision under the same next-token objective as current-domain training. Since latent generation operates in this frozen space, it avoids generator drift across domains. To further cope with shifts in reporting

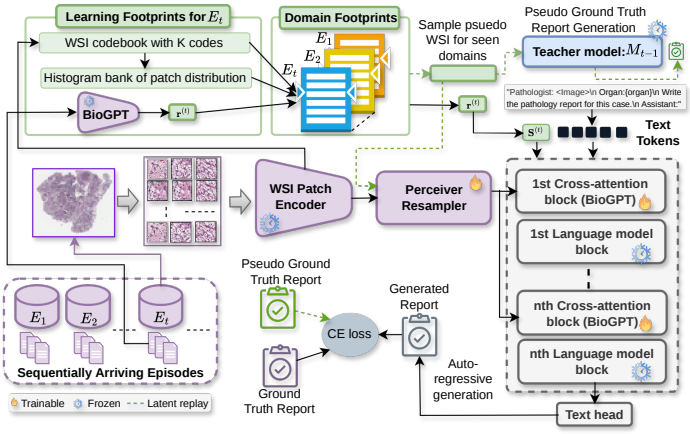

Fig. 1. Overview of the proposed continual WSI report generation framework.

conventions, we incorporate per-domain *report-style prototype* and condition the LLM through a lightweight style-prefix. At inference, we infer the most compatible footprint directly from slide content, enabling domain-agnostic generation. We evaluate our approach under domain-incremental scenarios, measuring report quality and retention across domains.

Contributions. We propose a CL framework for WSI report generation that (i) represents each domain with compact visual footprints in a frozen embedding space, enabling rehearsal without storing raw WSIs or features, (ii) enables exemplar-free replay via footprint-based pseudo-WSIs and immediate-teacher pseudo reports, with performance comparable to exemplar replay, and (iii) supports domain-agnostic, style-conditioned generation using per-domain report-style prototypes.

2 Method

We study continual WSI report generation in a domain-incremental setting. Training proceeds over domains, also termed as episodes $\{E_t\}_{t=1}^{T}$, where at episode t the model has access only to the current dataset $\mathcal{D}_t = \{(\mathbf{X}, y)\}$ and cannot revisit samples from earlier episodes $\{\mathcal{D}_1, \ldots, \mathcal{D}_{t-1}\}$. The goal is to learn each new domain while maintaining performance on previously learned domains. Our framework (Figure 1) has three components: (i) compact per-domain footprints that summarize morphology in a frozen embedding space, (ii) footprint-based replay that synthesizes pseudo-WSIs and pseudo reports to rehearse prior domains, and (iii) a per-domain report-style prototype injected into the language model to handle shifts in reporting conventions. The proposed CL strategy is

model-agnostic and can be attached to any WSI-to-report generator that consumes patch embeddings; we instantiate it with a HistoGPT architecture [24].

2.1 Report Generation for Giga-pixel Histopathology Images

A WSI x is represented as a set of patch embeddings $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ and N varies with tissue size and tiling. Patch embeddings are extracted with a frozen visual encoder. As in prior work, a Perceiver Resampler [6] maps the variable-length set $\mathbf{X}$ to a fixed-length latent sequence $\mathbf{Z} \in \mathbb{R}^{L_v \times D_{\text{LM}}}$, which serves as visual context for language generation. An autoregressive language model then generates report tokens $y = (y_1, \ldots, y_L)$ conditioned on $\mathbf{Z}$ via cross-attention, using an organ-specific prompt [24]. Training uses teacher forcing with next-token cross-entropy. Let q denote prompt tokens and y the report tokens. We minimize cross-entropy over the concatenated sequence $[q; y]$ while masking prompt positions, so the loss is computed only on report tokens.

2.2 Report Generation with Report-style Prototype

Reporting conventions can vary across domains (e.g., template, verbosity and phrasing). To provide a constant-size style cue, we compute a single report-style prototype for each domain t. Domain reports are encoded with a frozen text encoder, and the prototype $\mathbf{r}^{(t)} \in \mathbb{R}^{D_{text}}$ is obtained by mean-pooling token embeddings followed by ℓ_2 normalization. The prototype is computed once when a domain is first observed and is used to condition both real samples from $\mathcal{D}_t$ and future pseudo replay samples associated with domain t.

We condition the language model via style-prefix tokens. The tokenizer is extended with M dedicated style tokens $s = (s_1, \ldots, s_M)$ that are prepended to the textual input. A learned linear projection maps the prototype to prefix embeddings in the LM space, $\mathbf{S}^{(t)} = \phi(\mathbf{r}^{(t)}) \in \mathbb{R}^{M \times D_{\text{LM}}}$, where $\phi : \mathbb{R}^{D_{text}} \to \mathbb{R}^{M \times D_{\text{LM}}}$. During the forward pass, the input embeddings at the style-token positions are replaced by $\mathbf{S}^{(t)}$, while all other tokens use the standard embedding table. The resulting sequence $[s; q; y]$ is then processed by the autoregressive decoder with cross-attention to the visual latents $\mathbf{Z}$. Style-token positions are excluded from the loss by masking their labels, so training remains next-token cross-entropy on report tokens.

2.3 Domain Footprints

To support rehearsal without storing WSIs or per-slide patch features, we summarize each domain t with a compact domain footprint computed in the frozen patch-embedding space. For footprint construction, we subsample a bounded number of WSIs and, per WSI, a bounded number of patch embeddings, yielding a filtered set $\mathbf{X}^\star$. We learn a domain codebook by k-means clustering, $\mathbf{C}^{(t)} = \{\mathbf{c}_k^{(t)}\}_{k=1}^{K}$ with $\mathbf{c}_k^{(t)} \in \mathbb{R}^D$. Each patch embedding $\mathbf{x}_i \in \mathbf{X}^\star$ is quantized to its nearest codeword, $k_i = \arg\min_k \|\mathbf{x}_i - \mathbf{c}_k^{(t)}\|_2^2$, turning a WSI into a multi-set of discrete code indices with fixed storage per domain. To retain slide-level

composition, we compute for each training slide a normalized histogram $\mathbf{h} \in \mathbb{R}^K$ over code assignments, $h_k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[k_i = k]$, and store a compact, diversity-selected histogram bank $\mathcal{H}^{(t)}$ to preserve varied slide-level code compositions. We additionally store patch-count statistics $(\mu_N^{(t)}, \sigma_N^{(t)})$ to reproduce realistic slide sizes during replay. The domain footprint is $\mathcal{F}^{(t)} = \{\mathbf{C}^{(t)}, \mathcal{H}^{(t)}, \mu_N^{(t)}, \sigma_N^{(t)}, \mathbf{r}^{(t)}\}$, where $\mathbf{r}^{(t)}$ denotes the report-style prototype (Sec. 2.2).

2.4 Generative Replay with Pseudo Reports

At episode t, training interleaves supervised updates on $\mathcal{D}_t$ with replay sampled from footprints of past domains. To synthesize a pseudo WSI for a previous domain $j < t$, we sample a patch count $N \sim \mathcal{N}(\mu_N^{(j)}, (\sigma_N^{(j)})^2)$ and draw a histogram $\tilde{\mathbf{h}} \in \mathcal{H}^{(j)}$. We then sample code indices $k_1, \ldots, k_N$ from the categorical distribution defined by $\tilde{\mathbf{h}}$ and map them to codewords, yielding a synthetic patch set $\hat{\mathbf{X}}^{(j)} = \{\mathbf{c}_{k_i}^{(j)}\}_{i=1}^{N}$. Optionally, we add small Gaussian noise to the sampled codewords to increase diversity.

Because $\hat{\mathbf{X}}^{(j)}$ has no ground-truth report, we generate a pseudo target using an *immediate teacher*, i.e., a frozen snapshot of the model at the end of episode $t-1$. Conditioned on $\hat{\mathbf{X}}^{(j)}$ and the prompt, the teacher produces a pseudo report $\tilde{y}$, and the current model is trained on $(\hat{\mathbf{X}}^{(j)}, \tilde{y})$ with the same next-token cross-entropy as for real data. Both real and replay samples are conditioned using the report-style prototype of their domain. The t^{th} episode objective is

$$\mathcal{L}_t \;=\; \mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}_t}\Big[\mathcal{L}_{\mathrm{CE}}(y \mid \mathbf{X}, \mathbf{r}^{(t)})\Big] \;+\; \lambda\, \mathbb{E}_{(\hat{\mathbf{X}},\tilde{y},j)\sim\mathcal{R}_{<t}}\Big[\mathcal{L}_{\mathrm{CE}}(\tilde{y} \mid \hat{\mathbf{X}}, \mathbf{r}^{(j)})\Big]. \quad (1)$$

with $\mathcal{R}_{<t}$ denoting pseudo tuples $(\hat{\mathbf{X}}, \tilde{y}, j)$ from domains $j < t$, and $\mathcal{L}_{\mathrm{CE}}(\cdot; \mathbf{r})$ the next-token cross-entropy under style conditioning.

2.5 Domain-Agnostic Inference

At test time, explicit domain identifiers may be unavailable. We therefore select the most compatible footprint directly from the slide content and use its report-style prototype for conditioning. Given a test slide embedding set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, we score each domain t by its codebook reconstruction error in the frozen embedding space, $\mathrm{QE}(t; \mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} \min_{k\in\{1,\ldots,K\}} \|\mathbf{x}_i - \mathbf{c}_k^{(t)}\|_2^2$, and choose $\hat{t} = \arg\min_t \mathrm{QE}(t; \mathbf{X})$. Report generation is then conditioned using the selected prototype $\mathbf{r}^{(\hat{t})}$ via the style-prefix mechanism, enabling domain-agnostic generation without relying on dataset metadata.

3 Experiment and Results

3.1 Experimental Setup

Datasets. We evaluate WSI report generation on two public datasets. REG2025 [19] provides WSI-report pairs from 7 organs, with slides acquired across five scanners, arranged as an organ-shift (OS) sequence with one organ per episode.

Table 1. Details of CL datasets and train-test used in various experiments.

Source	Exp.	(#) Episodes information	#Train	#Test
REG2025	OS-R	(7) Breast→Bladder→Cervix→Colon →Lung→Prostate→Stomach	1746, 759, 511, 741, 636, 652, 1307	100 each
PathText	OS-P	(6) breast→lung→kidney→colorectal→uterus→thyroid	948, 820, 774, 485, 451, 403	100 each
Common organs (REG2025, PathText)	HS-C	(6) REG2025: (Breast→Colon→Lung)→ PathText: (breast→colorectal→ lung)	1746, 741, 636, 948, 485, 820	100 each
Disjoint organs (REG2025, PathText)	HS-D	(6) REG2025: (Bladder→Cervix→Prostate) PathText: (kidney→uterus→thyroid)	759, 511, 652 774, 451, 403	100 each

PathText [4], curated from multi-site TCGA [3], is re-partitioned into an OS benchmark with one organ per episode; we keep 6 organs with at least 500 pairs each. We also construct 2 hybrid 6-episode streams by considering common and disjoint organs across REG2025 and PathText, yielding heterogeneous shifts (HS), including dataset/source, site/scanner, organ, and larger report-style differences. Dataset statistics, experiment names, and episode composition are provided in Table 1.

Baselines. We compare against representative CL methods. Rehearsal-based: ER [20] and DER [2]. Regularization-based: EWC [7], SI [25], and LwF [12]. We additionally include ProgPrompt [18], which grows domain-specific soft prompts and selects the appropriate prompt using the domain identity. We also include the continual radiology report generation method, CMRG-LLM [23], which demands domain-specific sub-networks; we report results without their label-based alignment loss because disease labels are not available in our WSI data. Sequential fine-tuning (Naïve) and From-Scratch Training (FST) serve as lower-bound (LB), whereas Joint and Cumulative provide upper-bound (UB) references [11].

Implementation Details and Evaluation Metrics. WSIs are tiled into patches with CLAM [15] and encoded using frozen CONCH visual encoder [14]. All methods use the same HistoGPT backbone and prompt format. Models are trained for 10 epochs using a learning rate of 5×10^{-5}. ER/DER use a fixed exemplar budget of 10-50 samples, and regularization baselines follow standard settings from prior CL work [11]. (#) token for ProgPrompt is set to 50. For our method, we set M=4, K=64, $\mathcal{H}^{(t)}$=50 per domain for footprint construction, and generate 20 pseudo WSI-report pairs per domain. The report-style prototypes are computed from frozen BioGPT [16]. We follow REG2025 evaluation protocol [19] and compute the composite ranking score[4]. For a sequence of T episodes, we evaluate after each episode to obtain a $T \times T$ matrix of composite scores, from which we compute standard CL metrics: AVG, ILM, and BWT [13,5,9].

3.2 Results

Table 2 lists performance on four experiments (OS-R, OS-P, HS-C, HS-D) using AVG, ILM, and BWT computed from *Ranking_score*. For OS-R (Reg2025 organ shift), sequential fine-tuning on HistoGPT (Naïve and FST) suffer severe CF, evidenced by strongly negative BWT and low average performance

[4] *Ranking_score*: $0.15(\text{ROUGE} + \text{BLEU4}) + 0.4\,\text{key\_score} + 0.3\,\text{emb\_score}$.

Table 2. Best result in exemplar-based / exemplar-based with low size / exemplar-free categories indicated in red / blue / green, respectively. **Bold**: UB, $\mathcal{B}$: exemplar size.

	$\mathcal{B}$	Exp.→	OS-R			OS-P			HS-C			HS-D		
		Approach	AVG ↑	ILM ↑	BWT ↑	AVG ↑	ILM ↑	BWT ↑	AVG ↑	ILM ↑	BWT ↑	AVG ↑	ILM ↑	BWT ↑
Non-CL	-	naïve	0.3535	0.4065	-0.5086	0.3428	0.3423	-0.0308	0.2949	0.3545	-0.4239	0.2974	0.3505	-0.4376
		FST	0.3502	0.3942	-0.5026	0.3389	0.3414	-0.0303	0.2910	0.3524	-0.4230	0.2945	0.3483	-0.4427
		Joint	0.7381	-	-	**0.3734**	-	-	0.5687	-	-	**0.5702**	-	-
		Cumulative	**0.7426**	**0.7707**	**-0.0312**	0.3668	**0.3637**	**-0.0089**	**0.5692**	**0.6653**	**-0.0180**	0.5672	**0.6516**	**-0.0420**
CL	✓	ER ($\mathcal{B}$=50)	0.6714	0.7109	-0.0916	0.3549	0.3559	-0.0121	0.5251	0.6141	-0.0596	0.5308	0.6289	-0.0456
		DER ($\mathcal{B}$=50)	0.6620	0.7144	-0.0883	0.3524	0.3594	-0.0123	0.5392	0.6231	-0.0487	0.5480	0.6393	-0.0483
		ER ($\mathcal{B}$=10)	0.6100	0.6099	-0.2186	0.3461	0.3460	-0.0206	0.4866	0.5773	-0.1137	0.4502	0.5444	-0.1708
		ER ($\mathcal{B}$=20)	0.5992	0.6431	-0.1735	0.3527	0.3540	-0.0139	0.4943	0.5952	-0.0982	0.5049	0.6083	-0.0707
		DER ($\mathcal{B}$=10)	0.6299	0.6092	-0.2246	0.3493	0.3498	-0.0212	0.4747	0.5650	-0.1273	0.4699	0.5662	-0.1471
		DER($\mathcal{B}$=20)	0.6648	0.6778	-0.1251	0.3570	0.3563	-0.0112	0.5000	0.5989	-0.0915	0.5139	0.6032	-0.0761
	✗	SI	0.3211	0.3823	-0.0012	0.3284	0.3317	-0.0032	0.3409	0.3995	-0.0044	0.3416	0.4037	-0.0092
		LwF	0.3542	0.3998	-0.4972	0.3475	0.3433	-0.0219	0.2949	0.3615	-0.4324	0.2966	0.3543	-0.4244
		EWC	0.3567	0.4057	-0.4960	0.3425	0.3442	-0.0313	0.2879	0.3495	-0.4242	0.2951	0.3466	-0.4377
		ProgPrompt	0.6584	0.6680	0.0000	0.3624	0.3586	0.0000	0.5131	0.5750	0.0000	0.5265	0.5939	0.0000
		CMRG-LLM	0.2998	0.3008	0.0000	0.2327	0.2295	0.0000	0.2765	0.2911	0.0000	0.2148	0.2259	0.0000
		Our	0.6961	0.7319	-0.0686	0.3517	0.3544	-0.0130	0.5176	0.6094	-0.0850	0.5473	0.6424	-0.0413

(AVG and ILM). Joint and Cumulative provide upper bounds but require simultaneous access to all datasets, along with substantial storage and computation. Rehearsal methods are effective when the exemplar budget ($\mathcal{B}$) is large (DER ($\mathcal{B}$=50): 0.6620/0.7144/-0.0883) but, the performance degrades notably with smaller $\mathcal{B}$ (DER ($\mathcal{B}$=10): 0.6299/0.6092/-0.2246). Among exemplar-free baselines, regularization methods EWC, SI and LwF largely collapse to Naïve behavior (e.g., OS-R, ILM remains ≈0.40), suggesting that constraining parameter drift alone is insufficient for this report generation application. ProgPrompt yields zero forgetting by construction since the backbone is frozen and only domain-specific prompts are learned then frozen, resulting in BWT=0 on all sequences; however, restricted plasticity reduces average performance (OS-R, AVG/ILM: 0.6584/0.6680) and it requires domain identity at inference to activate the appropriate prompt. CMRG-LLM performs poorly in our WSI setup, consistent with the absence of explicit disease labels and limited vision-text interaction when the LLM is fully frozen. Our exemplar-free method achieves AVG/ILM/BWT: 0.6961/0.7319/-0.0686, substantially outperforming rehearsal methods having small buffer ($\mathcal{B}$=10, 20) and matching performance with rehearsal methods having large buffer. Thus, our method, which operates without storing exemplars or requiring domain identity at inference, demonstrates that footprint-based generative replay serves as an effective rehearsal alternative.

For OS-P (organ shift on PathText), absolute performance is consistently lower than OS-R across methods. This sequence is more challenging because reports frequently include clinical context that is only weakly grounded in slide morphology, which limits the maximum attainable performance even for non-CL UB. This trend is reflected in the smaller improvement from Naïve to Cumulative method on OS-P (ILM: 6.25%) relative to OS-R/HS-C/HS-D (89.59/87.67/89.91)%. Within this constrained regime, our method remains a practical alternative that requires neither domain identifiers nor real exemplars.

For HS-C/HS-D, performance degrades across methods due to the presence of heterogeneous shift that combine organ changes with variations in report style. In this setting, our method again outperforms other exemplar-free approaches while avoiding domain-identifiers. Rehearsal methods exhibit degradation with small $\mathcal{B}$

Table 3. Qualitative OS-R results. Two Bladder (E_1) test cases (GT vs. generated) after training to Stomach (E_T). Blue: GT-consistent; Red: mismatches.

Example 1	GT	Urinary bladder, transurethral resection; No tumor present Note) The specimen includes muscle proper
	Naïve	Stomach, endoscopic biopsy; Extranodal marginal zone B cell lymphoma of MALT type
	DER $\mathcal{B}$=10	Urinary bladder, transurethral resection; 1. No tumor present 2. Chronic granulomatous inflammation with foreign body reaction Note) The specimen includes muscle proper.
	DER $\mathcal{B}$=50	Urinary bladder, transurethral resection; Chronic nonspecific cystitis
	Ours	Urinary bladder, transurethral resection; No tumor present Note) The specimen includes muscle proper.
Example 2	GT	Urinary bladder, transurethral resection; Non-invasive papillary urothelial carcinoma, low grade Note) The specimen does not include muscle proper.
	Naïve	Stomach, endoscopic mucosal resection; Gastrointestinal stromal tumor
	DER $\mathcal{B}$=10	Urinary bladder, transurethral resection; 1. No tumor present 2. Chronic granulomatous inflammation with foreign body reaction Note) The specimen includes muscle proper.
	DER $\mathcal{B}$=50	Urinary bladder, transurethral resection; Non-invasive papillary urothelial carcinoma, low grade Note) The specimen includes muscle proper.
	Ours	Urinary bladder, transurethral resection; Non-invasive papillary urothelial carcinoma, low grade Note) The specimen includes muscle proper.

Table 4. Ablation study. FR: footprint-based generative replay, RS: Report-style.

Exp.→	OS-R			OS-P			HS-C			HS-D		
Method	AVG ↑	ILM ↑	BWT ↑	AVG ↑	ILM ↑	BWT ↑	AVG ↑	ILM ↑	BWT ↑	AVG ↑	ILM ↑	BWT ↑
naïve	0.3535	0.4065	-0.5086	0.3428	0.3423	-0.0308	0.2949	0.3545	-0.4239	0.2974	0.3505	-0.4376
Our ✗FR ✓RS	0.3918	0.4192	-0.4915	0.3451	0.3447	-0.0251	0.2953	0.3625	-0.4221	0.2985	0.3507	-0.4412
Our ✓FR ✗RS	0.6661	0.7087	-0.0930	0.3427	0.3489	-0.0155	0.5075	0.5990	**-0.0838**	0.5340	0.6300	-0.0577
Our ✓FR ✓RS	**0.6961**	**0.7319**	**-0.0686**	**0.3517**	**0.3544**	**-0.0130**	**0.5176**	**0.6094**	-0.0850	**0.5473**	**0.6424**	**-0.0413**

(e.g., DER drops by 7.3/3.9/87.9% with $\mathcal{B}$=50→20 on HS-C). Compared to the best low-buffer result (blue), our method improves HS-C (AVG/ILM/BWT) by (3.5/1.8/7.1)% and HS-D by (6.5/5.6/41.6)%. Thus, our method benefits from report-style prototype conditioning, which improves both average performance and retention under HS.

Table 3 presents qualitative results for two Bladder (E_1) cases evaluated after training up to Stomach (E_T) in OS-R experiment. (Naïve) shows pronounced cross-organ drift, generating Stomach-specific reports (e.g., MALT lymphoma or gastrointestinal stromal tumor) instead of Bladder findings. DER mitigates this drift but still introduces inconsistent clinical details. In contrast, our method preserves organ context and produces Bladder-consistent reports across both cases, demonstrating improved performance retention without storing past slides.

3.3 Ablation

Table 4 confirms that footprint-based generative replay (FR) is a key contributor: removing FR causes a large drop in AVG/ILM and more negative BWT across settings. The report-style prototype alone yields only a marginal improvement over naïve sequential fine-tuning, indicating a limited but consistent benefit in the continual setup. Adding style conditioning on top of generative replay yields consistent gains in AVG/ILM (OS-R: 4.5/3.3%, OS-P: 2.6/1.6%, HS-C: 2.0/1.7%, HS-D: 2.5/2.0%) and improves retention on most shifts by reducing BWT (OS-R: 26.2%, OS-P: 16.1%, HS-D: 28.4%), suggesting that domain-appropriate style prototypes, selected via domain-agnostic matching at inference, help preserve reporting conventions and mitigate CF under distribution shifts.

4 Conclusion

We presented an exemplar-free CL framework for WSI-to-report generation. The method summarizes each domain with compact visual footprints in a frozen embedding space, enabling replay via synthesized pseudo-WSIs and pseudo reports generated by an immediate teacher snapshot. We further incorporate per-domain report-style prototypes and footprint-matching strategy to support domain-agnostic style selection at inference. This framework enables continual adaptation under image and reporting-convention shifts without storing any WSI-report pairs. While the proposed framework is effective for exemplar-free replay across seen continual domains, future work will explore calibrated domain matching, rare-aware replay sampling, and topology-aware footprint variants to further improve robustness under increasing domain heterogeneity and underrepresented pathology patterns in continual pathology streams.

Acknowledgments. The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

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

References

1. Bhatt, R., Kumari, P., Mahapatra, D., Saddik, A.E., Saini, M.: Characterizing continual learning scenarios and strategies for audio analysis. arXiv preprint arXiv:2407.00465 (2024)
2. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems **33**, 15920–15930 (2020)
3. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature **487**(7407), 330–337 (Jul 2012)
4. Chen, P., Li, H., Zhu, C., Zheng, S., Shui, Z., Yang, L.: Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 546–556. Springer (2024)
5. Díaz-Rodríguez, N., Lomonaco, V., Filliat, D., Maltoni, D.: Don't forget, there is more than forgetting: new metrics for continual learning. arXiv preprint arXiv:1810.13166 (2018)
6. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
7. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
8. Kumari, P., Bozorgpour, A., Reisenbüchler, D., Jost, E., Crysandt, M., Matek, C., Merhof, D.: Domain-incremental white blood cell classification with privacy-aware continual learning. Scientific Reports **15**(1), 25468 (2025)

9. Kumari, P., Chauhan, J., Bozorgpour, A., Azad, R., Merhof, D.: Continual learning in medical imaging analysis: A comprehensive review of recent advancements and future prospects. arXiv preprint arXiv:2312.17004 (2023)
10. Kumari, P., Reisenbüchler, D., Bozorgpour, A., Schaadt, N.S., Feuerhake, F., Merhof, D.: Attention-based generative latent replay: A continual learning approach for wsi analysis. arXiv preprint arXiv:2505.08524 (2025)
11. Kumari, P., Reisenbüchler, D., Luttner, L., Schaadt, N.S., Feuerhake, F., Merhof, D.: Continual domain incremental learning for privacy-aware digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 34–44. Springer (2024)
12. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
13. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)
14. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
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. Nature Biomedical Engineering **5**(6), 555–570 (2021)
16. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), bbac409 (2022)
17. Pellegrini, L., Graffieti, G., Lomonaco, V., Maltoni, D.: Latent replay for real-time continual learning. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10203–10209. IEEE (2020)
18. Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A.: Progressive prompts: Continual learning for language models. arXiv preprint arXiv:2301.12314 (2023)
19. REG2025 Challenge Organizers: REG2025: REport Generation in Pathology using Pan-Asia Gigapixel Whole-Slide Images (WSIs). Grand Challenge (2025), `https://reg2025.grand-challenge.org/reg2025/`
20. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. Advances in Neural Information Processing Systems **32** (2019)
21. Sadegheih, Y., Kumari, P., Merhof, D.: Modality-agnostic brain lesion segmentation with privacy-aware continual learning. In: International Workshop on PRedictive Intelligence In MEdicine. pp. 1–13. Springer (2025)
22. Sadegheih, Y., Merhof, D., Kumari, P.: Towards modality-agnostic continual domain-incremental brain lesion segmentation. arXiv preprint arXiv:2601.13927 (2026)
23. Sun, Y., Khor, H.G., Wang, Y., Wang, Z., Zhao, H., Zhang, Y., Ma, L., Zheng, Z., Liao, H.: Continually tuning a large language model for multi-domain radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 177–187. Springer (2024)
24. Tran, M., Schmidle, P., Guo, R.R., Wagner, S.J., Koch, V., Lupperger, V., Novotny, B., Murphree, D.H., Hardway, H.D., D'Amato, M., et al.: Generating dermatopathology reports from gigapixel whole slide images with histogpt. Nature Communications **16**(1), 4886 (2025)
25. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International conference on machine learning. pp. 3987–3995. PMLR (2017)