SMART: PRESERVING DIAGNOSTIC SEMANTICS FOR UNSUPERVISED DOMAIN ADAPTATION IN WHOLE SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While pre-trained foundation models and Multiple Instance Learning (MIL) have advanced computational pathology, their clinical deployment remains severely hindered by cross-institutional domain shifts. Although Unsupervised Domain Adaptation (UDA) offers a practical solution to such discrepancies, directly applying general UDA paradigms to gigapixel Whole Slide Images (WSIs) often leads to negative transfer. This is because aggregated slide-level representations inherently entangle domain-specific confounders with task-relevant diagnostic semantics, exacerbating pseudo-label confirmation bias in self-training and marginalizing hard target cases under asymmetric domain shifts. To address these challenges, we propose Semantic Memory-Augmented Robust Transfer (SMART), a novel UDA framework that jointly optimizes across feature, label, and distribution dimensions. At the feature dimension, a convergence-aware variational information bottleneck dynamically purifies representations by filtering domain-specific confounders while preserving diagnostic semantics. Leveraging this purified latent space, we mitigate self-training confirmation bias at the label dimension by retrieving and incorporating reliable non-parametric source priors. Finally, at the distribution dimension, we address asymmetric domain shifts through an entropy-guided cross-difficulty mixup strategy. Extensive evaluations on real-world cohorts demonstrate consistent improvements over state-of-the-art methods. Designed as a plug-and-play framework, SMART can be readily integrated into existing bag-based MIL methods, substantially improving cross-institutional WSI classification. Code will be released.

1 INTRODUCTION

Digital pathology has significantly advanced clinical diagnostics, with deep learning algorithms achieving notable success in analyzing Whole Slide Images (WSIs) (Cui & Zhang, 2021). Recently, pathology-specific foundation models have further driven this progress, offering powerful feature extraction capabilities (Chen et al., 2024; Lu et al., 2024). However, because these models cannot directly achieve high performance in downstream slide-level tasks, the current optimal paradigm employs frozen foundation models to extract patch features and trainable Multiple Instance Learning (MIL) aggregators to aggregate them into slide-level representations (Lu et al., 2021; Xu et al., 2024). Despite the strong performance of models built under this paradigm on single-domain benchmarks, their clinical deployment remains hindered by domain shifts (Stacke et al., 2020; Thiringer et al., 2026). Variations across medical institutions driven by distinct scanning devices, tissue staining procedures, and patient demographics inevitably introduce distribution discrepancies (Cui et al., 2023). As shown in Fig. 1(d), cross-institutional shifts produce clear visual style discrepancies. More importantly, Fig. 1(a) shows that target samples from both diagnostic classes largely concentrate along the manifold of a single source class, rather than aligning with their respective source-class structures, revealing substantial semantic feature misalignment.

A straightforward approach to rectifying this feature misalignment would be to fine-tune models directly on the target domain; however, this is largely restricted by the high cost of data annotation (Koohbanani et al., 2021). In real-world scenarios, compiling a comprehensive, well-annotated dataset for every new target institution is cost-prohibitive and often practically infeasible due to the rarity of

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.

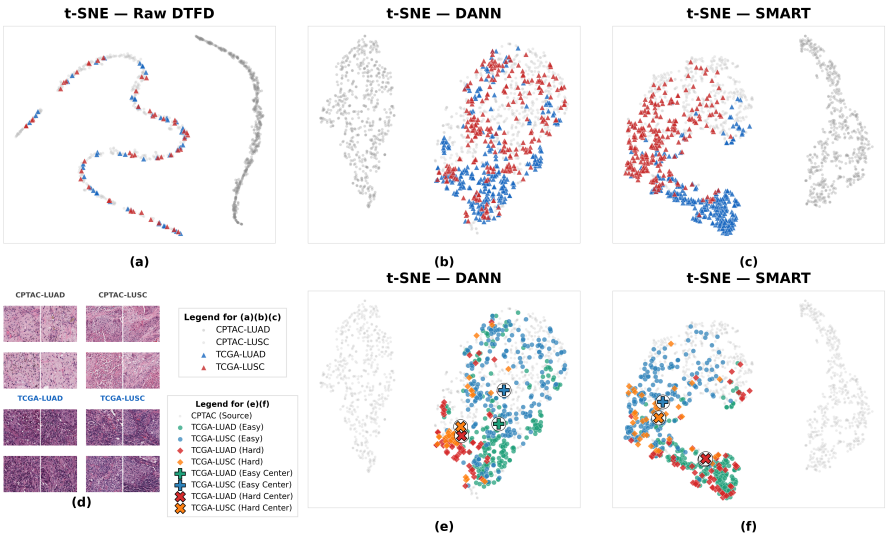


Figure 1: t-SNE visualization of slide-level features using different unsupervised domain adaptation methods for the NSCLC subtype (LUAD vs. LUSC) classification task. Source domain: CPTAC, target domain: TCGA. DTFD, a double-tier feature distillation-based MIL framework, is adopted as the slide-level classification backbone. (a) Source-only baseline. For visual clarity, only 20% of the target-domain points are displayed in (a), while all source-domain points are retained. Feature alignment using (b, e) DANN and (c, f) our proposed SMART. (d) Representative pathological patches from both domains, illustrating the visual domain shifts. In (e) and (f), target samples are explicitly partitioned into "easy" and "hard" subsets based on the standard Shannon entropy derived from the source-only model's predictions, with cross markers denoting their respective class centers.

certain diseases (Lu et al., 2022). Unsupervised Domain Adaptation (UDA) provides a pragmatic solution to this predicament. By effectively transferring diagnostic knowledge from a labeled source domain, UDA circumvents the need for target-domain labels. Consequently, developing efficient UDA strategies is essential for translating WSI analysis algorithms into reliable clinical practice (Shou et al., 2025; Dimitriou et al., 2019).

Despite its potential, UDA research in computational pathology remains relatively limited, primarily because directly applying existing UDA paradigms to MIL frameworks faces three major challenges. First, image-space adaptation methods such as traditional stain normalization (Vahadane et al., 2016) or recent generative models like CycleDiffusion (Chen et al., 2025) incur prohibitive computational overhead on gigapixel WSIs. More importantly, these approaches are incompatible with modern offline feature extraction pipelines powered by pre-trained foundation models (e.g., CONCH (Lu et al., 2024)). Thus, performing feature-level domain adaptation within the embedding space has emerged as a highly pragmatic pathway.

Second, even within this pathway, naively applying standard feature-level UDA (e.g., DANN (Ganin et al., 2016)) to MIL often causes significant negative transfer. Standard feature-level UDA relies on aligning the global feature distributions between the source and target domains. However, unlike natural images with distinct object-centric foregrounds, WSIs contain thousands of patches, including extensive non-diagnostic tissue regions and preparation artifacts. When MIL aggregates these patch-level features into slide-level features, critical task-relevant diagnostic semantics inevitably become entangled with domain-specific confounders (e.g., scanner-induced color biases). As illustrated in Fig. 1 (b), while conventional UDA (e.g., DANN) forcibly aligns the source and target distributions, the features of different target classes (TCGA-LUAD and TCGA-LUSC) collapse into a severely mixed cluster. This lack of class separability indicates that the alignment is dominated by domain confounders at the expense of diagnostic semantics. Consequently, this process effectively erases diagnostic semantics and compromises the model's discriminative power in the target domain.

Third, the entanglement of slide-level features drastically exacerbates confirmation bias in self-training UDA, where erroneous pseudo-labels are repeatedly reinforced during model training. Due

to the weakly supervised nature of MIL, the model's slide-level predictions on the target domain are inherently highly uncertain under distribution shifts. Relying on these noisy predictions to generate slide-level pseudo-labels causes conventional teacher-student self-training paradigms to amplify errors, ultimately risking model collapse (Qu et al., 2022; Arazo et al., 2020; Zhang et al., 2022). Furthermore, the inherent intra-domain heterogeneity in real-world pathological targets creates highly asymmetric domain shifts. As shown in Fig. 1 (e), while the class centers of the "easy" target samples remain separated, the centers of the "hard" samples collapse towards each other. This indicates that traditional UDA optimization takes an algorithmic shortcut by favoring low-entropy "easy" slides. Conversely, "hard" target slides experiencing severe shifts (e.g. rare subtypes or local artifacts) are marginalized. Consequently, while the model may learn discriminative features for easy samples, the slide-level features of hard samples remain conflated, critically undermining the model's true robustness in complex clinical scenarios.

To address these challenges, we build upon the general adversarial learning and teacher-student UDA paradigms to propose the Semantic Memory-Augmented Robust Transfer (SMART) framework. By orchestrating a joint optimization across three critical dimensions: features, labels, and distributions, SMART successfully mitigates negative transfer (Fig. 1 (c)) and clearly separates both easy and hard target samples (Fig. 1 (f)), thereby achieving robust and efficient cross-domain WSI classification. At the feature dimension, we design a Convergence-aware Residual Variational Information Bottleneck (CR-VIB) to act as a slide-level representation purifier. CR-VIB filters domain-specific confounders while preserving essential diagnostic semantics through residual pathways, thereby yielding a more domain-invariant feature manifold. At the label dimension, building upon this purified latent space, we introduce Semantic Anchor Retrieval Augmentation (SARA) to rectify confirmation bias in self-training. Inspired by pathologists consulting reference cases, SARA queries a source-domain memory bank to construct a non-parametric, similarity-weighted diagnostic prior from Top-K semantic neighbors, thereby producing more reliable pseudo-labels for target-domain self-training. At the distribution dimension, we propose Robust Entropy-Aware Latent Mixup (REALM) to address asymmetric domain shifts. REALM uses prediction uncertainty to identify difficult target samples and regularizes them through latent interpolation with reliable easy prototypes, improving robustness to challenging cases.

The main contributions of this work are summarized as follows:

- We are the first to systematically integrate pathology foundation models into the feature extraction pipeline for Unsupervised Domain Adaptation. By constructing a benchmark that closely reflects real-world clinical scenarios, we establish a new standard for future research.
- We propose Semantic Memory-Augmented Robust Transfer (SMART) framework, which constructs a comprehensive adaptation paradigm across feature, label, and distribution dimensions by integrating three core modules: (1) CR-VIB, featuring a convergence-aware residual design for intelligent feature deconfounding and semantic purification without information destruction; (2) SARA to inject reliable source-domain diagnostic priors across domains; and (3) REALM, which employs uncertainty-guided cross-difficulty latent mixup to mitigate the marginalization of hard target samples.
- Extensive evaluations on six WSI cohorts demonstrate that SMART outperforms existing state-of-the-art methods. Furthermore, comprehensive ablation studies validate the individual contribution of each component within SMART.
- Designed as a plug-and-play framework, SMART can be easily integrated into existing bag-based MIL methods to significantly enhance their cross-domain classification performance.

2 Related Works

Multiple Instance Learning in Whole Slide Images. Given the high annotation costs associated with gigapixel WSIs, weakly supervised Multiple Instance Learning (MIL) has emerged as a prevalent paradigm in computational pathology (Araujo et al., 2019; Chikontwe et al., 2020). In this formulation, a WSI is treated as a bag comprising patch-level instances. Typically, an encoder extracts patch features, which are subsequently pooled by an aggregator into a unified slide-level feature for downstream diagnostic tasks (Campanella et al., 2019).

Early explorations of instance fusion relied on non-parametric operations like mean- or max-pooling. However, since diagnostically relevant regions are often extremely sparse, these pooling mechanisms can dilute critical semantics with background noise (Lu et al., 2021). Consequently, learnable aggregation mechanisms were developed. Attention-based architectures (e.g., AB-MIL (Ilse et al., 2018), CLAM (Lu et al., 2021)) assign learnable attention scores to individual instances for adaptive feature weighting, significantly enhancing the model's discriminative power. To further leverage the intrinsic hierarchical nature of WSIs, DSMIL (Li et al., 2021) introduces a multi-scale dual-stream architecture, while DTFD-MIL (Zhang et al., 2022) proposes a two-tier aggregation paradigm via pseudo-bag decoupling.

To relax the traditional assumption of instance-level independence, recent works integrate self-attention (e.g., TransMIL (Shao et al., 2021)) or Graph Neural Networks (e.g., GTMIL (Zheng et al., 2022), WiKG (Li et al., 2024)) to model complex spatial dependencies within a slide. Furthermore, emerging multimodal approaches inject textual priors into the feature space to simultaneously enhance diagnostic accuracy and clinical interpretability (Lu et al., 2023; Guo et al., 2024). Despite these architectural advancements, state-of-the-art MIL frameworks still fundamentally rely on the i.i.d. assumption across domains. Consequently, they suffer from a severe performance drop when confronted with cross-institutional distribution shifts, underscoring the urgent need for robust domain adaptation strategies.

Unsupervised Domain Adaptation. Unsupervised Domain Adaptation (UDA) aims to learn domain-invariant representations, transferring models from a source domain to an unlabeled target domain (Wang & Deng, 2018; Long et al., 2015). Standard workflows initially focused on global feature alignment. Early domain discrepancy-based methods utilized metrics like Maximum Mean Discrepancy (MMD (Long et al., 2015; Tzeng et al., 2014)) or Wasserstein distance (Shen et al., 2018; Courty et al., 2016) to minimize cross-domain statistical divergences (Sun & Saenko, 2016). Concurrently, adversarial methods (e.g., DANN (Ganin et al., 2016), ADDA (Tzeng et al., 2017)) introduced domain discriminators to force the feature extractor to learn domain-invariant representations.

While global alignment is a fundamental baseline, it often struggles to capture fine-grained target semantics (Kang et al., 2019; Chen et al., 2019). Consequently, pseudo-label-based self-training has emerged as the mainstream paradigm. However, this approach is highly susceptible to confirmation bias, where initial prediction noise accumulates during training (Arazo et al., 2020; Zou et al., 2019; Rizve et al., 2021). To mitigate the propagation of pseudo-label errors, various rectification strategies have been proposed, including centroid-distance-based confidence weighting and consistency regularization combined with domain mixup (Liang et al., 2020; Na et al., 2021).

While these general UDA strategies have proven effective in standard vision tasks, adapting them to the computational pathology domain and gigapixel WSIs introduces specific hurdles. Image-space adaptation attempts to harmonize visual styles at the pixel level, progressing from patch-level consistency frameworks like Unsupervised Stain Adaptation (Reisenbüchler et al., 2024) to recent generative techniques such as Stain-Guided CycleDiffusion (Chen et al., 2025). However, these pixel-level transformations incur prohibitive computational overhead and remain fundamentally incompatible with the offline feature extraction pipelines necessitated by modern frozen foundation models. Consequently, recent efforts have shifted towards feature-space adaptation. Examples include alignment-based strategies on slide graphs (Shou et al., 2025) and retrieval-based methods leveraging optimal transport and external memory (Cui et al., 2023).

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let $\mathcal{D}_s = \{(X_i^s, Y_i^s)\}_{i=1}^{N_s}$ and $\mathcal{D}_t = \{X_j^t\}_{j=1}^{N_t}$ denote the labeled source and unlabeled target domains, respectively, where $Y_i^s \in \{1, \dots, C\}$. Due to the gigapixel resolution of WSIs, each slide X^d (where $d \in \{s, t\}$ denotes the domain indicator) is partitioned into a bag of M_k instances, denoted as $X_k^d = \{x_{k,1}^d, x_{k,2}^d, \dots, x_{k,M_k}^d\}$. Following standard computational pathology pipelines, a pre-trained foundation model $\mathcal{F}$ maps the raw patches into a feature space, yielding instance-level features $H_k^d = \{h_{k,1}^d, h_{k,2}^d, \dots, h_{k,M_k}^d\}$. To capture the global context, an MIL aggregator $\mathcal{A}_\theta$ aggregates these instance features into a slide-level feature $v_k^d = \mathcal{A}_\theta(H_k^d)$. The core challenge in cross-domain

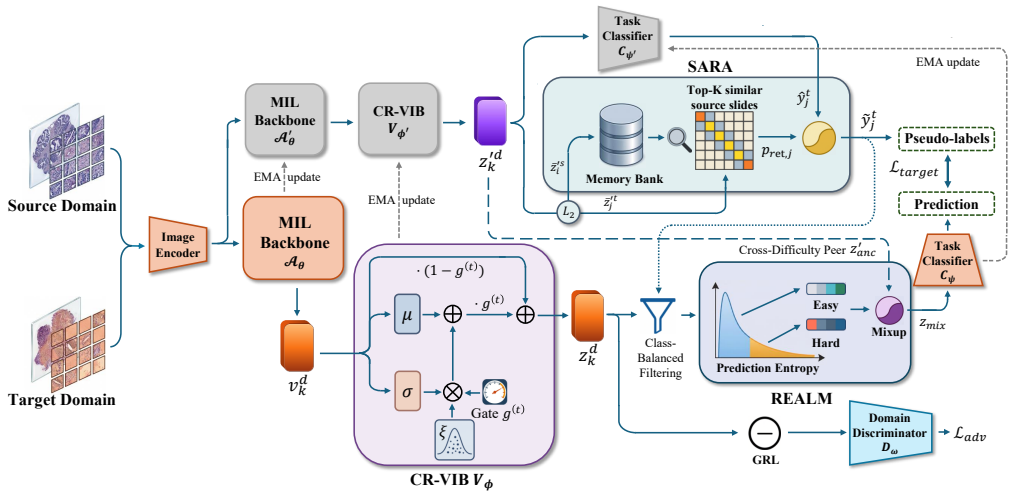


Figure 2: Pipeline of the proposed SMART framework. Building upon the adversarial teacher-student paradigm, it integrates three key components: (1) CR-VIB for convergence-aware feature purification, (2) SARA for pseudo-label rectification via source-domain retrieval, and (3) REALM for distribution robust optimization via cross-difficulty latent mixup.

WSI classification arises from the distribution shift between the source feature space $P(v^s)$ and the target feature space $P(v^t)$. To mitigate this, our goal is to learn a mapping that purifies the entangled representation v^d_k into a domain-invariant yet task-discriminative latent representation z^d_k via UDA, thereby minimizing the expected classification risk on target slides.

3.2 Framework Overview

SMART is a unified UDA framework designed for cross-domain WSI classification, built upon an adversarial teacher-student architecture. Let $\Theta = \{\theta, \phi, \psi\}$ denote the parameters of the student branch, which consists of three components: an MIL backbone $\mathcal{A}_\theta$, a CR-VIB layer V_ϕ for representation purification, and a task classifier C_ψ. In addition, a domain discriminator D_ω equipped with a Gradient Reversal Layer is applied to the student representations for adversarial domain alignment. To provide stable pseudo-supervision for the target domain $\mathcal{D}_t$, we maintain a teacher model with parameters $\Theta' = \{\theta', \phi', \psi'\}$. The teacher is strictly updated via an Exponential Moving Average (EMA) of the corresponding student parameters at each training step. As shown in Fig. 2, this self-training pipeline links our three core modules sequentially: First, in the upper teacher branch, the teacher's CR-VIB layer $V_{\phi'}$ dynamically filters task-irrelevant noise from the aggregated features of both source and target domains, yielding purified teacher representations z'^d_k. Using the purified representations from the target domain z'^t_j, the teacher's classifier $C_{\psi'}$ generates initial soft predictions $\hat{y}^t_j$ for the unlabeled target bags. Second, the SARA module rectifies $\hat{y}^t_j$ by retrieving Top-K source-domain semantic neighbors from a dynamically constructed memory bank comprising the purified source representations z'^s_i, thereby producing refined pseudo-labels $\tilde{y}^t_j$. Finally, in the lower student branch, the student model processes input slides through its own modules to extract purified representations z^d_k for both source and target domains. The student is then optimized by: (1) minimizing the supervised classification risk on $\mathcal{D}_s$, (2) confusing D_ω to align cross-domain global distributions, and (3) learning from the target domain via the REALM module. Specifically, REALM applies Soft-DRO reweighting and cross-difficulty latent mixup on the student's target features z^t_j, guided by the refined pseudo-labels $\tilde{y}^t_j$, to ensure robust generalization against hard examples. Detailed mathematical formulations of these modules are provided in the following subsections.

3.3 Convergence-aware Residual Variational Information Bottleneck (CR-VIB)

While symmetrically deployed in both teacher ($V_{\phi'}$) and student (V_ϕ) branches, we formulate the CR-VIB module using the student branch for notational simplicity. Taking the student's aggregated slide-level feature v^d_k as input, our CR-VIB module extracts a domain-invariant and task-discriminative

latent representation z_k^d. In standard adversarial UDA, directly aligning slide-level representations can entangle diagnostic semantics with domain-specific confounders, leading to negative transfer. To mitigate this problem, we utilize the Variational Information Bottleneck (VIB) to filter out these irrelevant confounders through KL-based information compression (Alemi et al., 2016; Achille & Soatto, 2018). However, directly applying a standard VIB with static compression is suboptimal for WSIs, where diagnostic signals are inherently weak and sparse (Campanella et al., 2019). Such premature compression may discard task-relevant semantics before the classifier has sufficiently learned them. We therefore further develop a Convergence-aware Residual VIB (CR-VIB), which dynamically adjusts the compression strength according to the model's learning status while preserving information through a residual pathway.

Specifically, the input representation v_k^d is first mapped to a mean μ_k^d and standard deviation σ_k^d. Instead of employing a static penalty, we introduce a dynamic convergence-aware gate $g^{(t)} \in [\epsilon, 1.0]$ at epoch t, where ϵ is a small positive constant to prevent complete module deactivation. The magnitude of $g^{(t)}$ is determined by a monotonically decreasing function with respect to the exponential moving average of the source-domain classification loss, denoted as $\mathcal{L}_{EMA}^{(t-1)}$. Specifically, during the early training phase when the loss remains high, $g^{(t)}$ is dynamically lowered to temporarily relax the information compression. As the source classifier progressively converges and $\mathcal{L}_{EMA}^{(t-1)}$ decays, $g^{(t)}$ smoothly and adaptively scales up, eventually saturating at 1.0 to execute full VIB purification. Using this adaptive gate, the stochastically sampled latent representation z_{vib}^d and the final residual output z^d are computed as:

$$z_{vib,k}^d = \mu_k^d + g^{(t)} \cdot \sigma_k^d \odot \xi, \quad \xi \sim \mathcal{N}(0, I), \quad z_k^d = (1 - g^{(t)}) v_k^d + g^{(t)} \cdot z_{vib,k}^d \tag{1}$$

This residual connection ensures that under conditions of incomplete source convergence ($g^{(t)} \to \epsilon$), the module preserves maximal information capacity by approximating a deterministic identity mapping ($z^d \approx v^d$), protecting the fragile semantics. Once fundamental discriminative boundaries are established and convergence improves ($g^{(t)} \to 1$), it smoothly transitions into standard compression. Correspondingly, the KL divergence penalty is scaled synchronously:

$$\mathcal{L}_{VIB} = \beta_{base} \cdot g^{(t)} \cdot \frac{1}{N} \sum_{d \in \{s,t\}} \sum_{k=1}^{N_d} D_{KL}\left(\mathcal{N}(\mu_k^d, (\sigma_k^d)^2) \parallel \mathcal{N}(0, I)\right) \tag{2}$$

where $N = N_s + N_t$ denotes the total number of samples across both domains, and β_{base} is a trade-off hyperparameter. By jointly gating the forward noise injection and the backward KL penalty, CR-VIB prevents the feature space from entangling with domain-specific confounders, yielding a transferable feature manifold while preserving essential diagnostic semantics.

3.4 Semantic Anchor Retrieval Augmentation (SARA)

To mitigate confirmation bias, where erroneous target predictions can be recursively reinforced through pseudo-label-based self-training, we introduce Semantic Anchor Retrieval Augmentation (SARA). SARA dynamically rectifies the target-domain slide's initial prediction $\hat{y}_j^t$ leveraging the purified representations $z_k'^d$ extracted from both the source and target domains by the teacher's CR-VIB module. Specifically, at each epoch, the teacher constructs a memory bank $\mathcal{M} = \{(\bar{z}_i'^s, Y_i^s)\}_{i=1}^{N_s}$ using its L_2-normalized source representations $\bar{z}_i'^s = z_i'^s / \|z_i'^s\|_2$. Given a target query identically normalized as $\bar{z}_j'^t = z_j'^t / \|z_j'^t\|_2$, SARA computes cosine similarities $s_{j,i} = (\bar{z}_j'^t)^\top \bar{z}_i'^s$ between the query and all entries in $\mathcal{M}$ to retrieve the Top-K nearest semantic neighbors, denoted as $\mathcal{N}_K(j)$. To translate the retrieved neighbor labels into a continuous probability distribution, we apply a temperature-scaled softmax over the similarities, yielding the retrieval-augmented probability $p_{ret,j}^{(c)}$ for class c:

$$p_{ret,j}^{(c)} = \frac{\sum_{i \in \mathcal{N}_K(j)} \mathbb{1}(Y_i^s = c) \exp(s_{j,i}/\tau_{ret})}{\sum_{i \in \mathcal{N}_K(j)} \exp(s_{j,i}/\tau_{ret})} \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function and τ_{ret} is the temperature hyperparameter. The final rectified pseudo-label $\tilde{y}_j^t$ is computed as a combination of the teacher's prediction and the retrieval distribution:

$$\tilde{y}_j^t = \eta_{ret}^{(t)} \hat{y}_j^t + (1 - \eta_{ret}^{(t)}) p_{ret,j} \tag{4}$$

where $\eta_{ret}^{(t)} \in (0,1)$ decays monotonically over time. As the cross-domain latent space aligns during training, this decay progressively shifts reliance toward the source semantic priors $p_{ret,j}$, effectively anchoring the target predictions and preventing error amplification. Ultimately, this refined soft distribution $\tilde{y}_j^t$ serves as the definitive pseudo-label output of the teacher branch, which is subsequently passed to the student model for robust self-training.

3.5 Robust Entropy-Aware Latent Mixup (REALM)

To establish a robust optimization objective for the target domain, we first construct a class-balanced, high-confidence target subset $\tilde{\mathcal{T}}$. Let $p_j = \max_c \tilde{y}_j^{t(c)}$ and $c_j = \arg\max_c \tilde{y}_j^{t(c)}$ denote the confidence and predicted class of the teacher's rectified pseudo-labels. For each class c, we collect a candidate set $\mathcal{V}_c = \{j \mid c_j = c, p_j \geq \tau_{conf}\}$. To prevent majority classes from dominating the self-training process, we compute a dynamic class budget $B_c = \max(\min_k |\mathcal{V}_k|, \lfloor \rho |\mathcal{V}_c| \rfloor)$, with truncation ratio ρ. By selecting the Top-B_c confident samples from each $\mathcal{V}_c$, we form the balanced subset $\tilde{\mathcal{T}} = \bigcup_{c=1}^{C} \tilde{\mathcal{T}}_c$.

Guided by the rectified pseudo-labels $\tilde{y}_j^t$ of the balanced subset $\tilde{\mathcal{T}}$, REALM utilizes the student's latent target features z_j^t and the teacher's stable target features $z_j'^t$ to perform Soft Distributionally Robust (Soft-DRO) reweighting and cross-difficulty feature interpolation. We first quantify the hardness of each sample in $\tilde{\mathcal{T}}$ using the normalized predictive entropy of the teacher model: $u_j = -\frac{1}{\log C} \sum_{c=1}^{C} \hat{y}_j^{t(c)} \log \hat{y}_j^{t(c)}$. Based on a predefined hardness ratio γ_{hard}, we sort the samples in $\tilde{\mathcal{T}}$ in descending order of their entropy u_j. The top γ_{hard} proportion of samples with the highest entropy are partitioned into the hard group $\mathcal{H}$, while the remaining samples constitute the easy group $\mathcal{E}$. To prevent the model from ignoring challenging samples, we formulate a Soft-DRO strategy, assigning a weight to each sample using a temperature-scaled softmax:

$$w_j = |\tilde{\mathcal{T}}| \cdot \frac{\exp(u_j/\tau_{dro})}{\sum_{k \in \tilde{\mathcal{T}}} \exp(u_k/\tau_{dro})} \tag{5}$$

where τ_{dro} regulates the focus on worst-case samples, and the weights are scaled to maintain an expectation of 1. To further calibrate the representation space, REALM forces hard samples to align with high-confidence core prototypes while encouraging easy samples to explore broader decision boundaries. We maintain two class-aware memory banks, structured as FIFO queues, to store the teacher's target representations and their rectified soft labels: a hard bank $\mathcal{B}_{\mathcal{H}} = \{(z_k'^t, \tilde{y}_k^t)\}_{k \in \mathcal{H}}$ and an easy bank $\mathcal{B}_{\mathcal{E}} = \{(z_k'^t, \tilde{y}_k^t)\}_{k \in \mathcal{E}}$. For a given student target feature z_j^t, if $j \in \mathcal{H}$, we sample an easy target prototype $(z_{anc}'^t, \tilde{y}_{anc}^t)$ belonging to the same pseudo-class from $\mathcal{B}_{\mathcal{E}}$. Conversely, if $j \in \mathcal{E}$, we sample a hard target prototype from $\mathcal{B}_{\mathcal{H}}$. By pairing the gradient-tracking student feature with a detached teacher prototype, we compute the latent mixup as:

$$z_{mix} = \lambda z_j^t + (1-\lambda) z_{anc}'^t, \quad y_{mix} = \lambda \tilde{y}_j^t + (1-\lambda) \tilde{y}_{anc}^t \tag{6}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ is the interpolation coefficient controlled by the hyperparameter α. Finally, both the original features z_j^t and the interpolated features z_{mix} are fed into the task classifier C_ψ to obtain the target classification probabilities. These predictions, alongside the DRO weights w_j and the augmented labels y_{mix}, are used to construct the robust self-training loss for the target domain.

3.6 Loss Function

For the target domain, we combine the DRO-weighted self-training loss and the latent mixup consistency loss. Optimized via Kullback-Leibler (KL) divergence to leverage the dark knowledge in soft pseudo-labels, the target loss $\mathcal{L}_{target}$ is defined as:

$$\mathcal{L}_{target} = \frac{1}{|\tilde{\mathcal{T}}|} \sum_{j \in \tilde{\mathcal{T}}} \left(w_j D_{KL}(\tilde{y}_j^t \parallel \sigma(C_\psi(z_j^t))) + \lambda_{mix} D_{KL}(y_{mix} \parallel \sigma(C_\psi(z_{mix}))) \right) \tag{7}$$

where C_ψ is the task classifier, σ denotes the softmax function, and λ_{mix} weights the latent mixup consistency term.

To jointly optimize the student branch parameters Θ and the domain discriminator parameters ω, the overall objective $\mathcal{L}_{total}$ is formulated as a weighted combination of four components:(1) the standard Cross-Entropy loss $\mathcal{L}_{src}$ for the labeled source domain; (2) the domain adversarial loss $\mathcal{L}_{adv}$ applied to the purified representations z_k^d via a Gradient Reversal Layer (GRL); (3) the target-domain loss $\mathcal{L}_{target}$; and (4) the convergence-aware KL-divergence penalty $\mathcal{L}_{VIB}$ from the CR-VIB layer.

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate SMART on three clinical tasks across six WSI cohorts, comprising five adaptation directions: (1) NSCLC subtyping with bidirectional adaptation between CPTAC and TCGA-NSCLC; (2) Brain tumor classification with bidirectional adaptation between IPD-Brain and TCGA-GBMLGG; and (3) PLN status prediction in Cervical Cancer, adapting unidirectionally from an in-house Cervical WSI dataset to TCGA-CESC. Detailed dataset descriptions are provided in the Supplementary Material. For the data partitioning, each source domain is randomly split into a training and a validation set with an 8:2 ratio to establish the source model. Each unlabeled target domain is randomly partitioned into an adaptation (training) set and a test set with a 5:5 ratio. To mitigate the impact of sample selection randomness and ensure robust evaluation, this target-domain partitioning was repeated across five random splits. Models are evaluated on their respective test sets, and we report the mean and standard deviation across all five splits.

Implementation Details. For WSI preprocessing, we remove non-tissue background regions and crop the remaining tissue into 512×512 patches at $20\times$ magnification. To ensure a fair comparison, all evaluated methods extract patch-level features using the frozen pre-trained CONCH (Lu et al., 2024). For feature aggregation, our method and all baselines utilize DTFD-MIL (Zhang et al., 2022) as the shared backbone, with the sole exception of DETA (Shou et al., 2025) which employs its own architecture. Detailed configurations are provided in the Supplementary Material. All experiments are implemented using PyTorch 2.3.0 and conducted on an NVIDIA A100 Tensor Core GPU.

Table 1: Results on four cross-domain WSI classification directions (all metrics in %). Best results are **bolded** and second-best are underlined.

Method	CPTAC-NSCLC → TCGA-NSCLC			TCGA-NSCLC → CPTAC-NSCLC			TCGA-GBMLGG → IPD-Brain			IPD-Brain → TCGA-GBMLGG		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
CLAM (Lu et al., 2021)	42.89 $\pm$1.84	50.53 $\pm$1.79	33.56 $\pm$0.79	90.03 $\pm$2.63	48.16 $\pm$2.78	33.48 $\pm$1.74	82.19 $\pm$2.82	61.93 $\pm$1.50	32.67 $\pm$1.39	81.10 $\pm$0.42	45.65 $\pm$1.17	40.89 $\pm$1.31
DTFD-MIL (Zhang et al., 2022)	57.68 $\pm$1.23	50.53 $\pm$1.79	33.56 $\pm$0.79	88.11 $\pm$2.42	55.40 $\pm$2.38	48.55 $\pm$1.89	74.17 $\pm$1.76	53.44 $\pm$1.14	26.05 $\pm$3.07	79.08 $\pm$1.10	42.64 $\pm$1.37	28.54 $\pm$0.87
DANN (Ganin et al., 2016)	51.41 $\pm$3.16	50.53 $\pm$1.79	33.56 $\pm$0.79	95.99 $\pm$1.91	90.12 $\pm$4.80	89.97 $\pm$4.98	81.16 $\pm$2.71	66.57 $\pm$5.71	43.98 $\pm$6.42	80.33 $\pm$1.55	63.43 $\pm$2.00	43.92 $\pm$1.96
CDAN (Long et al., 2018)	54.45 $\pm$7.47	50.53 $\pm$1.79	33.56 $\pm$0.79	86.95 $\pm$26.38	85.56 $\pm$23.93	82.99 $\pm$29.66	79.36 $\pm$2.24	65.05 $\pm$1.94	42.71 $\pm$2.76	81.92 $\pm$2.31	66.21 $\pm$1.96	44.64 $\pm$2.57
MK-MMD (Long et al., 2015)	50.50 $\pm$1.11	50.99 $\pm$2.58	34.67 $\pm$3.04	70.88 $\pm$21.44	66.22 $\pm$18.90	63.46 $\pm$22.33	75.12 $\pm$7.62	61.91 $\pm$11.13	33.30 $\pm$11.23	81.50 $\pm$5.39	66.19 $\pm$8.18	40.64 $\pm$7.86
SRDC (Tang et al., 2020)	50.36 $\pm$0.23	50.53 $\pm$1.79	33.56 $\pm$0.79	12.92 $\pm$15.27	49.27 $\pm$2.52	32.99 $\pm$1.15	63.30 $\pm$2.12	48.79 $\pm$2.80	17.20 $\pm$4.23	67.28 $\pm$3.69	40.52 $\pm$7.10	21.09 $\pm$4.16
MCD (Saito et al., 2018)	50.65 $\pm$0.49	50.53 $\pm$1.79	33.56 $\pm$0.79	56.86 $\pm$12.70	45.43 $\pm$2.01	31.23 $\pm$0.95	81.24 $\pm$3.10	66.04 $\pm$9.30	37.63 $\pm$11.11	80.97 $\pm$0.63	66.48 $\pm$1.90	41.94 $\pm$2.13
DeepCORAL (Sun & Saenko, 2016)	57.67 $\pm$1.23	50.53 $\pm$1.79	33.56 $\pm$0.79	66.93 $\pm$30.25	63.41 $\pm$22.99	56.14 $\pm$29.08	75.45 $\pm$2.36	54.77 $\pm$3.73	25.03 $\pm$4.28	79.16 $\pm$1.19	42.64 $\pm$2.07	32.58 $\pm$1.94
DRL (Kim et al., 2026)	55.57 $\pm$3.45	50.53 $\pm$1.79	33.56 $\pm$0.79	92.29 $\pm$11.86	76.22 $\pm$24.62	69.71 $\pm$33.52	70.20 $\pm$5.05	45.66 $\pm$15.85	19.28 $\pm$8.14	72.94 $\pm$6.37	43.25 $\pm$14.93	25.00 $\pm$9.77
RAM-MIL (Cui et al., 2023)	57.48 $\pm$1.34	50.53 $\pm$1.79	33.56 $\pm$0.79	90.52 $\pm$2.22	47.24 $\pm$2.72	32.55 $\pm$1.48	77.21 $\pm$1.81	56.38 $\pm$2.25	28.39 $\pm$2.27	80.44 $\pm$0.92	58.30 $\pm$0.81	31.05 $\pm$0.39
DETA (Shou et al., 2025)	45.25 $\pm$6.57	50.56 $\pm$1.85	33.65 $\pm$0.93	38.51 $\pm$5.43	49.61 $\pm$1.94	33.15 $\pm$0.87	65.44 $\pm$3.92	47.46 $\pm$1.08	14.14 $\pm$2.01	61.77 $\pm$9.95	36.97 $\pm$16.79	15.58 $\pm$10.16
SMART (Ours)	**95.58** $\pm$0.73	**86.69** $\pm$2.23	**86.53** $\pm$2.30	**98.20** $\pm$0.77	**94.10** $\pm$2.74	**94.07** $\pm$2.78	**86.10** $\pm$1.95	**68.09** $\pm$2.69	**50.28** $\pm$3.84	**85.08** $\pm$1.03	**68.44** $\pm$1.27	**48.28** $\pm$2.01

4.2 COMPARISONS WITH STATE-OF-THE-ART

We benchmark SMART against three categories of state-of-the-art methods: (1) Source-only baselines (CLAM (Lu et al., 2021) and DTFD-MIL (Zhang et al., 2022)); (2) General UDA methods (DANN (Ganin et al., 2016), CDAN (Long et al., 2018), MK-MMD (Long et al., 2015), SRDC (Tang et al., 2020), MCD (Saito et al., 2018), DeepCORAL (Sun & Saenko, 2016), DRL (Kim et al., 2026)); and (3) Pathology-specific UDA methods (RAM-MIL (Cui et al., 2023) and DETA (Shou et al., 2025)).

As shown in Table 1, direct generalization methods suffer from severe performance degradation due to cross-institutional domain shifts. For instance, CLAM and DTFD-MIL achieve only 42.9% and 57.7% AUC on the CPTAC $\rightarrow$ TCGA-NSCLC task, respectively. While general UDA methods attempt to align feature distributions, their performance fluctuates, often resulting in negative transfer. This indicates that directly applying domain adaptation methods designed for natural images fails to accommodate the unique paradigms of pathological images. Furthermore, existing pathology-specific UDA methods struggle to yield substantial and consistent improvements across these tasks. In contrast, our SMART framework substantially outperforms all baselines across these tracks. Notably, on the challenging CPTAC $\rightarrow$ TCGA-NSCLC benchmark, SMART achieves a remarkable AUC of 95.58%, effectively closing the performance gap caused by domain shift. See Section A.2 of the Supplementary Material for an analysis of why baselines suffer from mode collapse on this task and how SMART prevents it. As detailed in Supplementary Fig. 3, adaptation from the in-house dataset to TCGA-CESC presents a formidable clinical challenge. Specifically, training a MIL model

(DTFD) directly on the target domain completely fails to learn meaningful representations, yielding a random-guess AUC of 47.45%. This result highlights the practical utility of our domain adaptation strategy, particularly in scenarios where the target domain inherently lacks sufficient data volume or exhibits complex distributional characteristics. Despite these challenging settings, SMART achieves the highest AUC and significantly outperforms the evaluated UDA baselines. This demonstrates SMART's exceptional robustness and its capability to generalize effectively in highly complex, real-world pathological diagnostic tasks where standard supervised training remains challenging.

Furthermore, designed as a plug-and-play framework, SMART can be easily integrated into existing bag-based MIL methods to enhance their cross-domain classification performance. To demonstrate this flexibility, we evaluated SMART using CLAM as an alternative backbone. As summarized in Table 2 and fully detailed in Table 8 in the Supplementary Material, SMART consistently outperforms all competing UDA methods with CLAM backbone across four transfer directions.

Table 2: Average performance across four transfer directions using CLAM as the MIL backbone.

Method	Average Performance (%)		
	AUC	ACC	Macro-F1
CLAM	74.05	51.57	35.15
DANN	<u>75.78</u>	63.56	<u>50.63</u>
MCD	74.63	<u>67.26</u>	49.40
DeepCORAL	71.54	54.83	36.37
RAM-MIL	72.64	57.71	36.46
Ours	**90.59**	**81.38**	**70.68**

Table 3: Ablation study on CPTAC-NSCLC → TCGA-NSCLC and TCGA-GBMLGG → IPD-Brain.

Method	AUC (%)	ACC (%)	Macro-F1 (%)
CPTAC-NSCLC → TCGA-NSCLC			
w/o CR-VIB	64.46 $\pm$20.42	56.25 $\pm$14.30	52.54 $\pm$17.48
w/o SARA	93.95 $\pm$1.92	86.22 $\pm$1.64	86.14 $\pm$1.59
w/o REALM	94.85 $\pm$2.05	85.84 $\pm$3.63	85.66 $\pm$3.67
Ours (Full)	**95.58** $\pm$0.73	**86.69** $\pm$2.23	**86.53** $\pm$2.30
TCGA-GBMLGG → IPD-Brain			
w/o CR-VIB	85.66 $\pm$1.83	66.39 $\pm$3.17	47.57 $\pm$3.36
w/o SARA	84.95 $\pm$3.05	66.04 $\pm$1.78	49.63 $\pm$4.14
w/o REALM	84.39 $\pm$1.94	67.64 $\pm$3.67	45.59 $\pm$3.19
Ours (Full)	**86.10** $\pm$1.95	**68.09** $\pm$2.69	**50.28** $\pm$3.84

4.3 Ablation Studies

We ablate SMART's components on two transfer directions: CPTAC → TCGA (NSCLC) and TCGA → IPD (Brain Tumor) (Table 3). Removing CR-VIB leads to a drastic performance drop, most notably in the CPTAC → TCGA task where the AUC decreases from 95.58% to 64.46% with a standard deviation of ($\pm$20.42%). This confirms that without feature purification, forcing global alignment causes the model to overfit to domain confounders, resulting in severe negative transfer. Disabling SARA consistently degrades performance across datasets, indicating that injecting source-domain semantic priors is crucial to mitigate confirmation bias induced by uncalibrated pseudo-labels. Furthermore, excluding REALM reduces the F1 score in the heterogeneous five-class TCGA → IPD task (from 50.28% to 45.59%). This demonstrates REALM's efficacy in handling hard examples and preventing their marginalization under asymmetric shifts. These results validate that the full model achieves optimal stability and performance by jointly optimizing across feature, label, and distribution.

5 Conclusion

The clinical deployment of computational pathology models is severely bottlenecked by cross-institutional domain shifts. To overcome this challenge, we propose Semantic Memory-Augmented Robust Transfer (SMART), an Unsupervised Domain Adaptation framework that achieves robust cross-domain WSI classification through joint optimization across feature, label, and distribution dimensions. Specifically, the CR-VIB filters domain-specific confounders while preserving diagnostic semantics. Building on this purified latent space, SARA mitigates self-training confirmation bias via non-parametric source retrieval, while REALM prevents the marginalization of hard samples through entropy-guided cross-difficulty mixup. Extensive evaluations on real-world cohorts demonstrate that SMART, as a plug-and-play architecture, significantly outperforms SOTA baselines. By maintaining strong robustness under asymmetric shifts and data-scarce conditions, SMART provides a pragmatic, scalable paradigm for deploying reliable whole slide image analysis in diverse clinical environments.

AI USE STATEMENT

In this work, we used AI to assist with editing, organizing, and polishing author-written text for grammar, clarity, and readability, and with manuscript formatting. The authors review the AI-assisted edits and take responsibility for the final content of this work, including all text, claims, and other materials.

ETHICS STATEMENT

This study uses publicly available WSI datasets and a non-public, in-house cervical cancer cohort for secondary computational analysis. No new specimens were collected, and no participants were recruited for this study. The datasets were accessed and used in accordance with applicable data-access requirements and institutional authorizations.

REPRODUCIBILITY STATEMENT

The main text describes the SMART architecture, training objectives, unsupervised domain adaptation protocol, and evaluation metrics. Appendix A.1 describes the datasets and transfer tasks, while Appendix A.3 provides data partitioning, model selection, implementation, and optimization details. The appendix also reports hyperparameter analyses, component ablations, and additional evaluations across cohorts and MIL backbones. Together, these descriptions and results document the procedures and settings used in the reported experiments.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 4(11):e21, 2019.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

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.

Hsi-Ling Chen, Chun-Shien Lu, and Pau-Choo Chung. Sgcd: Stain-guided cyclediffusion for unsupervised domain adaptation of histopathology image classification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pp. 1081–1090. PMLR, 2019.

Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 519–528. Springer, 2020.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021.

Yufei Cui, Ziquan Liu, Yixin Chen, Yuchen Lu, Xinyue Yu, Xue Steve Liu, Tei-Wei Kuo, Miguel Rodrigues, Chun Jason Xue, and Antoni Chan. Retrieval-augmented multiple instance learning. *Advances in Neural Information Processing Systems*, 36:24859–24878, 2023.

Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 189–199. Springer, 2024.

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

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4893–4902, 2019.

Seonghwi Kim, Sung Ho Jo, Wooseok Ha, and Minwoo Chae. Distributionally robust classification for multi-source unsupervised domain adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RyAyH8ufGu.

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, 2021.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11323–11332, 2024.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

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.

Ming Y Lu, Richard J Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical image analysis*, 76:102298, 2022.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19764–19775, 2023.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1094–1103, 2021.

Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022.

Daniel Reisenbüchler, Lucas Luttner, Nadine S Schaadt, Friedrich Feuerhake, and Dorit Merhof. Unsupervised latent stain adaptation for computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 755–765. Springer, 2024.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732, 2018.

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.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yuntao Shou, Xiangyong Cao, Peiqiang Yan, Qiao Hui, Qian Zhao, and Deyu Meng. Graph domain adaptation with dual-branch encoder and two-level alignment for whole slide image-based survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19925–19935, 2025.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8725–8735, 2020.

Erik Thiringer, Fredrik K Gustafsson, Kajsa Ledesma Eriksson, and Mattias Rantalainen. Scanner-induced domain shifts undermine the robustness of pathology foundation models. *arXiv preprint arXiv:2601.04163*, 2026.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

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.

Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5982–5991, 2019.

A APPENDIX

A.1 DETAILED DATASET DESCRIPTIONS

The original benchmark comprises three clinical tasks and five directed transfers across six cohorts. Table 1 reports four directions: bidirectional NSCLC and brain tumor adaptation. Figure 3 reports the fifth direction, in-house cervical cancer to TCGA-CESC. The EBRAINS extension in Appendix A.9 adds one cohort and four brain tumor transfer directions, bringing the complete evaluation to seven cohorts and nine directed transfers across the same three clinical tasks. A quantitative summary of the six cohorts in the original benchmark, including diagnostic classes and slide distributions, is provided in Table 4.

A.1.1 TASK 1: NON-SMALL CELL LUNG CANCER (NSCLC) SUBTYPING

This task focuses on the binary classification of the two most common NSCLC subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). For this bidirectional adaptation task, we utilize two large-scale public repositories: CPTAC-NSCLC and TCGA-NSCLC.

Table 4: Summary of the six cohorts, three clinical tasks, and class distributions in the original benchmark.

Clinical Task	Dataset	Diagnostic Classes	Num Classes	Slides per Class	Total Slides
NSCLC Subtyping	CPTAC-NSCLC	LUAD	2	681	1,342
		LUSC		661	
	TCGA-NSCLC	LUAD	2	532	1,040
		LUSC		508	
Brain Tumor (GBMLGG)	TCGA-GBMLGG	IDH1-mutant astrocytoma	5	378	1,526
		IDH1-mutant oligodendroglioma		326	
		IDH1-mutant glioblastoma		66	
		IDH1-wild-type astrocytoma		126	
		IDH1-wild-type glioblastoma		630	
	IPD-Brain	IDH1-mutant astrocytoma	5	109	448
		IDH1-mutant oligodendroglioma		97	
		IDH1-mutant glioblastoma		25	
		IDH1-wild-type astrocytoma		8	
		IDH1-wild-type glioblastoma		209	
Cervical Cancer PLN Status	In-house Cervical	PLN Negative	2	421	1,012
		PLN Positive		591	
	TCGA-CESC	PLN Negative	2	217	322
		PLN Positive		105	

A.1.2 TASK 2: BRAIN TUMOR (GBMLGG) CLASSIFICATION

This task involves a fine-grained, 5-class classification problem for gliomas, categorized based on IDH1 mutation status and histological subtypes. The five classes include: IDH1-mutant astrocytoma (IDH1-Mut Astro), IDH1-mutant oligodendroglioma (IDH1-Mut Oligo), IDH1-mutant glioblastoma (IDH1-Mut GBM), IDH1-wild-type astrocytoma (IDH1-WT Astro), and IDH1-wild-type glioblastoma (IDH1-WT GBM). We evaluate bidirectional adaptation between the TCGA-GBMLGG and IPD-Brain cohorts.

A.1.3 TASK 3: PELVIC LYMPH NODE (PLN) STATUS PREDICTION IN CERVICAL CANCER

This task aims to predict Pelvic Lymph Node (PLN) metastasis directly from primary cervical tumor WSIs, formulated as a binary classification problem: PLN Negative (PLN-) versus PLN Positive (PLN+). Unlike the previous tasks, we formulate this as a practical unidirectional adaptation scenario (In-house $\rightarrow$ TCGA-CESC) to validate the model's capacity to generalize from a large, well-annotated cohort to a highly complex cohort.

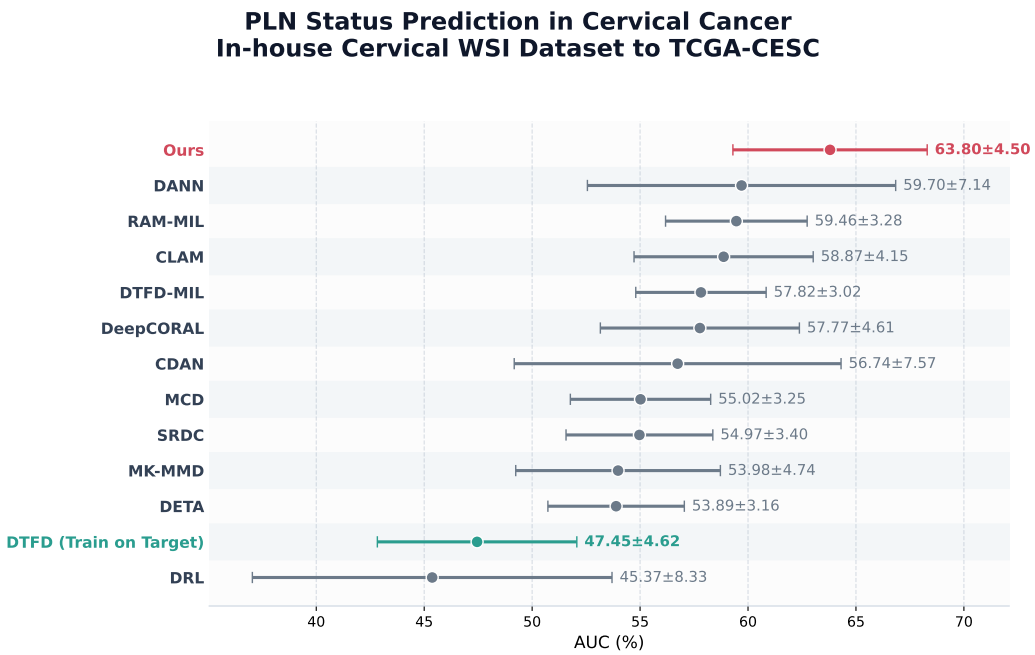


Figure 3: Performance of PLN Status Prediction in Cervical Cancer, from an in-house Cervical WSI Dataset to TCGA-CESC.

A.2 Detailed Analysis of Mode Collapse and Semantic Entanglement in Standard UDA

A.2.1 Quantitative Characterization of Domain Shift

We further characterize the domain discrepancy in the CPTAC-NSCLC $\rightarrow$ TCGA-NSCLC task using the Fréchet feature distance (FFD), which compares the means and covariances of patch-level features. Under the same feature extractor, the FFD is 0.7133 between CPTAC-NSCLC and TCGA-NSCLC, compared with 0.3708 between TCGA-GBMLGG and IPD-Brain. This comparison indicates a larger feature-distribution discrepancy for the evaluated NSCLC cohorts.

To examine the repeated accuracy values in the main comparison, we compute class-normalized confusion statistics on the target-test sets (Table 5). CLAM, DTFD-MIL, DANN, CDAN, SRDC, MCD, DeepCORAL, DRL, and RAM-MIL predict every target slide as LUSC on CPTAC-NSCLC $\rightarrow$ TCGA-NSCLC. Their LUAD recall is therefore zero, despite an overall accuracy of $50.53 \pm 1.79\%$. In contrast, SMART achieves recalls of $78.38 \pm 5.21\%$ for LUAD and $94.69 \pm 3.10\%$ for LUSC. These results identify a single-class prediction failure rather than successful discrimination between the target classes.

Table 5: Class-normalized confusion statistics on CPTAC-NSCLC $\rightarrow$ TCGA-NSCLC. Entries are percentages; SMART results are reported as mean $\pm$ standard deviation across five splits.

Model group	Ground-truth class	Predicted LUAD	Predicted LUSC
Collapsed baselines	LUAD	0.00	100.00
	LUSC	0.00	100.00
SMART	LUAD	78.38 ± 5.21	21.62 ± 5.21
	LUSC	5.31 ± 3.10	94.69 ± 3.10

We additionally train a target-supervised reference using only the labeled TCGA target-adaptation subset and evaluate it on the same held-out target-test subset. Across the five splits, this reference

achieves an AUC of $96.91 \pm 0.80\%$, compared with $95.58 \pm 0.73\%$ for SMART. The strong supervised result shows that the target classification task remains learnable, while the proximity of SMART to this reference demonstrates effective adaptation without target-domain training labels. The target-supervised model is reported as a separate reference and is not included in the fully unsupervised comparison.

A.2.2 Interpretation of Mode Collapse and Semantic Entanglement

Theoretical Causes in Pathological Images. We attribute this severe negative transfer to the fundamental mismatch between existing alignment mechanisms and the unique characteristics of WSIs. In computational pathology, actual diagnostic signals (e.g. tumor morphology) are extremely sparse and heavily overshadowed by pervasive domain-specific confounders (e.g. varying staining protocols, scanner color profiles and tissue preparation artifacts). For general UDA methods (e.g., DANN, CDAN), enforcing direct adversarial alignment on this highly noisy representation space forces the network to seek the easiest path to fool the domain discriminator. It achieves this by aggressively erasing not only the domain variations but also the fragile diagnostic semantics, causing the target representations to collapse into an entangled, semantically meaningless state.

Crucially, this vulnerability extends to recent pathology-specific UDA methods (e.g., RAM-MIL, DETA), which explains their identical mode collapse or severe degradation. Although methods like DETA attempt to bridge the domain gap via adversarial feature perturbations and dual-branch coupling, this underlying vulnerability remains unaddressed because they still operate directly on an unpurified latent space. These approaches typically rely on instance-level retrieval or graph structures constructed directly from raw feature embeddings. Because the distance metrics driving these mechanisms (such as the cost matrix in RAM-MIL's Optimal Transport formulation) rely on these uncalibrated embeddings, feature distances under extreme domain shifts become dominated by scanner and staining differences rather than actual pathological semantics. Consequently, retrieval mechanisms (as in RAM-MIL) fetch domain-correlated neighbors instead of semantically relevant ones, while graph constructions (as in DETA) establish erroneous edges between morphologically distinct but similarly stained patches. This "garbage-in, garbage-out" aggregation amplifies domain biases, trapping these specialized models in the same semantic entanglement.

Visual Verification via t-SNE. This theoretical analysis is strongly corroborated by the feature space visualized using t-SNE (Figure 1).

- **Raw DTFD (Fig. 1a):** Without adaptation, the source and target domains exhibit a clear domain shift. While the source features maintain some structure, the direct application to the target domain fails to map target samples into proper semantic clusters.
- **DANN (Fig. 1b):** This visualization illustrates the mode collapse phenomenon. Although DANN forces the features into a seemingly unified global distribution, the pathological classes (denoted by red and blue triangles) are completely entangled. The diagnostic decision boundary is erased by aggressive adversarial alignment, visually explaining why the model defaults to a majority-class guess.
- **SMART (Fig. 1c):** In contrast, our SMART framework successfully preserves the semantic integrity during adaptation. By dynamically purifying features (via CR-VIB) and preventing the marginalization of hard target samples (via REALM), SMART establishes a clear, linearly separable decision boundary between the red and blue classes in the target domain.

Quantitative Proof via Ablation. To substantiate our theoretical claims, we refer to the ablation study presented in Table 3 of the main paper. The empirical data perfectly aligns with our hypothesis regarding the necessity of feature purification. Specifically, on the CPTAC-NSCLC $\rightarrow$ TCGA-NSCLC task, removing the CR-VIB module causes a catastrophic performance drop, with AUC decreasing from 95.58% to 64.46% and ACC decreasing from 86.69% to 56.25%. More importantly, the substantial standard deviation observed without CR-VIB (± 20.42 in AUC) strongly indicates severe training instability. This confirms that without explicit decoupling and purification provided by CR-VIB, the network frequently succumbs to the exact same semantic entanglement and mode collapse that affects existing baselines. Therefore, CR-VIB acts as the critical gatekeeper, ensuring that downstream semantic retrieval (SARA) operates on a domain-agnostic and morphologically meaningful manifold.

A.3 Supplementary Experimental Protocol

A.3.1 Data Partitioning and Model Selection

All data partitions are constructed at the patient level, and no patient's WSIs are shared across subsets. The source domain is divided into training and validation subsets at an 8:2 ratio. The target domain is divided into an unlabeled adaptation subset and a held-out test subset at a 5:5 ratio. Target-domain partitioning is repeated across five random splits, and results are reported as mean and standard deviation across these splits.

Each SMART run is initialized from a source-domain checkpoint selected using source-validation data only. Adaptation then uses labeled source-training data and unlabeled target-adaptation data for a fixed 100 epochs. We evaluate the terminal epoch-100 checkpoint without early stopping. A single globally fixed hyperparameter configuration is shared across all five original adaptation directions and is retained in the additional cohort and backbone experiments. Target-test samples are excluded from adaptation, and target labels are not used for training, checkpoint selection, or hyperparameter selection.

For the main baseline comparisons, we follow the hyperparameter settings of the corresponding original implementations whenever they are compatible with the fully unsupervised adaptation protocol adopted in our benchmark. All DTFD-MIL-based adaptation methods share the same source-pretrained DTFD-MIL checkpoint. Methods with distinct architectures, such as DETA, use their own source-only initialization under the same source split. The same initialization principle is applied in the CLAM-backbone experiments. Notably, the default implementation of DRL uses a subset of target-domain labels for hyperparameter selection and early stopping, which introduces target supervision and is therefore not directly compatible with the standard UDA setting considered here. For a fair comparison, we disable all target-label-based model selection, retain the default hyperparameter configuration, and evaluate the terminal checkpoint after the fixed training schedule. Consequently, the DRL results reported in the main comparison do not use target-domain labels for training, hyperparameter selection, or checkpoint selection.

Hyperparameter sensitivity experiments are conducted after fixing the main configuration and are not used to choose task-specific settings. Likewise, analyses that require target ground-truth labels, including representation separability, retrieval purity, pseudo-label quality, and hard-slide geometry, are performed only after the corresponding training runs and checkpoints have been fixed.

A.3.2 Optimization and Scheduling Details

Network Initialization. Following the standard unsupervised domain adaptation protocol, our SMART framework and all evaluated baselines are initialized with weights pre-trained exclusively on the source domain. This ensures a fair benchmark and prevents early mode collapse during cross-domain alignment.

Warm-up Phase. Due to the gigapixel nature of WSIs and the weak supervisory signals in MIL, directly applying self-training at initialization is prone to overfitting to pseudo-label noise. To establish a well-structured feature space and initialize a reliable EMA teacher, the student model is trained solely using source-domain supervision and the domain adversarial objective for the first $T_{warm} = 30$ epochs. During this phase, the target-domain supervision modules (SARA and REALM) are deactivated ($\lambda_{tgt} = 0$), preventing the injection of premature pseudo-labels.

Phase Transition and Weight Scheduling. In the subsequent self-training phase (Epochs 31 to 100), the target supervision weight $\lambda_{tgt}^{(t)}$ ramps up linearly over the first 10 epochs to its maximum value of 0.1, preventing gradient perturbations that could destabilize the feature space. The domain adversarial weight $\lambda_{adv}^{(t)}$ follows a standard exponential schedule: $\lambda_{adv}^{(p)} = \frac{2}{1+\exp(-10 \cdot p)} - 1$, where p linearly scales from 0 to 1 over the total training iterations.

Optimizer and Teacher Update. The entire network is optimized end-to-end using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. A multi-step learning rate scheduler is employed, multiplying the learning rate by a factor of 0.2 at epoch 50. The total training process spans 100 epochs. Concurrently, the teacher model parameters Θ' are updated

at each training step via $\Theta' \leftarrow \alpha_{ema}\Theta' + (1-\alpha_{ema})\Theta$, where the momentum coefficient α_{ema} is set to 0.999 to guarantee stable pseudo-label generation.

Module-Specific Hyperparameters. In the CR-VIB module, the base KL-divergence weight β_{base} is set to 10^{-3}, and the minimum activation ratio ϵ is set to 0.01. In the SARA module, the source-domain memory bank is reconstructed by the teacher model at the beginning of each self-training epoch, and the number of retrieved semantic neighbors is set to $K = 5$. As the cross-domain latent spaces align during training, the weight parameter $\eta_{ret}^{(t)}$ is linearly decayed from 0.7 to a lower bound of 0.5, progressively increasing the contribution of the source semantic prior $p_{ret,j}$. In the REALM module, the initial confidence threshold τ_{conf} is set to 0.7 and the truncation ratio ρ is set to 0.5 for constructing the balanced target subset. The hardness ratio γ_{hard} is set to 0.5, and the Mixup parameter α is set to 0.2. The DRO temperature τ_{dro} is set to 0.05, and the mixup consistency loss weight λ_{mix} is set to 0.2.

A.4 Hyperparameter Sensitivity Analysis

We evaluate our framework's robustness across two distinct clinical scenarios: CPTAC-NSCLC → TCGA-NSCLC task and TCGA-GBMLGG → IPD-Brain task (Table 6).

Table 6: Sensitivity analysis of key hyperparameters on the CPTAC-NSCLC → TCGA-NSCLC and TCGA-GBMLGG → IPD-Brain tasks.

CPTAC-NSCLC → TCGA-NSCLC

SARA Top-K		**Hardness Ratio (γ_{hard})**		**Mixup (α)**	
K	AUC	γ_{hard}	AUC	α	AUC
1	94.36 $\pm$1.76	0.2	91.57 $\pm$5.62	0.05	93.46 $\pm$5.10
3	95.32 $\pm$1.22	0.3	95.63 $\pm$0.85	0.1	95.00 $\pm$2.21
5	95.58 $\pm$0.73	0.5	95.58 $\pm$0.73	0.2	95.58 $\pm$0.73
10	94.21 $\pm$1.13	0.7	95.70 $\pm$0.70	0.5	95.61 $\pm$1.02
15	94.90 $\pm$1.66	0.8	95.81 $\pm$0.38	1.0	95.74 $\pm$0.86

TCGA-GBMLGG → IPD-Brain

SARA Top-K		**Hardness Ratio (γ_{hard})**		**Mixup (α)**	
K	AUC	γ_{hard}	AUC	α	AUC
1	85.16 $\pm$2.33	0.2	85.61 $\pm$1.16	0.05	85.71 $\pm$1.78
3	85.81 $\pm$1.71	0.3	85.77 $\pm$1.84	0.1	84.79 $\pm$2.58
5	86.10 $\pm$1.95	0.5	86.10 $\pm$1.95	0.2	86.10 $\pm$1.95
10	85.28 $\pm$2.03	0.7	85.67 $\pm$1.72	0.5	85.87 $\pm$2.08
15	85.38 $\pm$2.71	0.8	85.66 $\pm$1.80	1.0	85.42 $\pm$0.99

Semantic Memory Retrieval (SARA Top-K). A small retrieval size ($K = 1$) provides limited contextual coverage, whereas increasing the number of retrieved neighbors introduces a broader set of source-domain references. Across both tasks, performance remains relatively stable over a moderate range of K, while overly large retrieval sets may introduce less relevant semantic neighbors.

Hard-Sample Mining (Hardness Ratio γ_{hard}). A very low hardness ratio ($\gamma_{hard} = 0.2$) degrades performance on CPTAC-NSCLC → TCGA-NSCLC. Beyond this setting, performance remains stable across a broad range of moderate-to-high hardness ratios ($\gamma_{hard} = 0.3$–0.8) on both tasks, indicating low sensitivity to this hyperparameter.

Interpolation Strength (Mixup α). The Mixup parameter α controls the interpolation between target representations and cross-difficulty prototypes. A very small α favors interpolation coefficients close to either endpoint, resulting in less balanced feature mixing. Performance remains stable across a broad range of α values, with only modest fluctuations on the brain tumor task, indicating limited sensitivity to this hyperparameter.

A.5 Generalizability Across Foundation Models and MIL Backbones

A.5.1 Foundation Model Substitution

To demonstrate that the robustness of SMART is not strictly bound to a specific foundation model, we conducted supplementary UDA experiments using UNI2 (Chen et al., 2024) as the image encoder instead of CONCH. As detailed in Table 7, we evaluated our framework against state-of-the-art baselines on the challenging bidirectional brain tumor classification tasks (IPD-Brain ↔ TCGA-GBMLGG). Despite the complete shift in the underlying latent space, SMART consistently maintains its superiority. It achieves the highest AUCs (84.20% and 87.90%) and macro-F1 scores (51.11% and 54.39%) in both adaptation directions, significantly outperforming recent competitors. This empirical evidence confirms that the mechanisms within SMART are foundation-model-agnostic, further proving its flexibility as a plug-and-play architecture for digital pathology.

Table 7: Performance on IPD-Brain → TCGA-GBMLGG and TCGA-GBMLGG → IPD-Brain tasks. In this experiment, UNI2 (Chen et al., 2024) is utilized as the image encoder instead of CONCH. Best results are **bolded** and second-best are underlined.

Method	IPD-Brain → TCGA-GBMLGG			TCGA-GBMLGG → IPD-Brain		
	AUC (%)	ACC (%)	Macro-F1 (%)	AUC (%)	ACC (%)	Macro-F1 (%)
DTFD-MIL	81.34 ±2.80	68.41 ±2.03	47.59 ±2.38	86.27 ±2.71	73.28 ±2.71	50.30 ±2.53
DANN	78.87 ±3.16	64.17 ±6.34	42.84 ±5.81	81.78 ±4.49	67.47 ±5.52	44.66 ±2.16
DeepCORAL	81.43 ±2.80	68.96 ±2.18	47.21 ±2.59	85.13 ±1.67	74.71 ±3.15	48.89 ±3.01
MCD	78.64 ±5.56	65.77 ±6.71	40.29 ±5.12	84.04 ±4.42	**78.01** ±1.85	49.37 ±3.12
RAM-MIL	82.60 ±2.84	69.59 ±1.77	46.60 ±2.28	84.96 ±2.33	74.26 ±2.25	50.37 ±2.30
SMART (Ours)	**84.20** ±3.16	**70.01** ±2.54	**51.11** ±1.99	**87.90** ±3.24	77.12 ±2.95	**54.39** ±3.71

A.5.2 CLAM Backbone

Table 8 reports the complete four-direction comparison using CLAM as the MIL backbone, complementing the average results in Table 2.

Table 8: Results on four cross-domain WSI classification directions using CLAM as the MIL backbone (all metrics in %). Best results are **bolded** and second-best are underlined.

Method	CPTAC-NSCLC → TCGA-NSCLC			TCGA-NSCLC → CPTAC-NSCLC			TCGA-GBMLGG → IPD-Brain			IPD-Brain → TCGA-GBMLGG		
	AUC	ACC	Macro-F1	AUC	ACC	Macro-F1	AUC	ACC	Macro-F1	AUC	ACC	Macro-F1
CLAM	42.89 ±1.84	50.53 ±1.79	33.56 ±0.79	90.03 ±2.63	48.16 ±2.78	33.48 ±1.74	82.19 ±2.82	61.93 ±1.50	32.67 ±1.39	81.10 ±0.42	45.65 ±1.17	40.89 ±1.31
DANN	50.00 ±0.00	50.53 ±1.79	33.56 ±0.79	94.40 ±6.89	87.71 ±8.87	87.46 ±9.27	81.45 ±2.80	66.49 ±2.73	42.40 ±4.06	77.27 ±2.19	49.53 ±5.32	39.09 ±5.64
MCD	50.47 ±0.45	50.53 ±1.79	33.56 ±0.79	84.98 ±29.20	83.25 ±24.13	80.92 ±29.27	83.16 ±4.29	68.56 ±5.60	41.22 ±6.49	79.90 ±1.45	**66.71** ±3.10	41.91 ±3.87
DeepCORAL	59.68 ±8.72	50.53 ±1.79	33.56 ±0.79	65.48 ±29.08	57.22 ±22.47	45.44 ±28.84	79.40 ±3.21	59.16 ±5.62	29.46 ±6.71	81.61 ±0.86	52.42 ±6.38	37.03 ±3.48
RAM-MIL	38.40 ±2.09	50.53 ±1.79	33.56 ±0.79	89.28 ±2.55	47.84 ±3.10	33.80 ±2.49	81.48 ±2.86	65.95 ±0.98	36.50 ±1.09	81.40 ±0.84	66.55 ±1.48	42.00 ±1.26
SMART (Ours)	**94.87** ±1.45	**90.09** ±2.61	**90.06** ±2.61	**99.00** ±1.01	**96.57** ±2.65	**96.56** ±2.67	**86.22** ±1.57	**72.65** ±1.68	**49.72** ±2.17	**82.25** ±1.82	66.21 ±2.48	**46.40** ±2.84

A.6 Robustness to Limited Target-Domain Data

To evaluate adaptation when only a small unlabeled cohort is available from a new institution, we vary the amount of target-adaptation data across the four lung and brain transfer directions. Percentages are defined relative to the complete target cohort. The standard 50% setting uses the full target-adaptation split, while the other 50% is held out for testing. The 1%, 5%, 10%, and 25% settings are sampled exclusively from the adaptation split, and the target-test set remains unchanged across all fractions within each split. All SMART runs use the same globally fixed configuration and the terminal epoch-100 checkpoint.

As shown in Table 9, SMART exceeds the source-only AUC in all four directions when 5% of the complete target cohort is available for adaptation. On CPTAC-NSCLC → TCGA-NSCLC, even the 1% setting, corresponding to 10 WSIs, improves AUC from $57.68 \pm 1.23\%$ to $81.24 \pm 4.36\%$.

Table 9: AUC (%) with limited unlabeled target-domain data. Percentages refer to the complete target cohort; 50% is the full adaptation split. The other 50% is held out for testing and remains unchanged across all adaptation fractions within each split.

Transfer	CPTAC-NSCLC → TCGA-NSCLC	TCGA-NSCLC → CPTAC-NSCLC	TCGA-GBMLGG → IPD-Brain	IPD-Brain → TCGA-GBMLGG
Source-only	57.68 ± 1.23	88.11 ± 2.42	74.17 ± 1.76	79.08 ± 1.10
1%	81.24 ± 4.36	90.04 ± 3.50	74.72 ± 1.90	79.00 ± 1.48
5%	87.66 ± 2.88	97.67 ± 0.58	74.56 ± 2.65	81.13 ± 0.57
10%	88.72 ± 5.89	95.57 ± 3.42	81.57 ± 1.14	83.03 ± 0.37
25%	90.21 ± 5.26	**98.74 ± 1.20**	83.82 ± 1.22	**85.13 ± 0.65**
50%	**95.58 ± 0.73**	98.20 ± 0.77	**86.10 ± 1.95**	85.08 ± 1.03

Source-only denotes the source-trained DTFD-MIL model. All other columns report SMART. Bold indicates the highest mean in each column.

At 10%, the AUC gains over source-only are 31.04, 7.46, 7.40, and 3.95 percentage points for CPTAC-NSCLC → TCGA-NSCLC, TCGA-NSCLC → CPTAC-NSCLC, TCGA-GBMLGG → IPD-Brain, and IPD-Brain → TCGA-GBMLGG, respectively.

The benefit depends on both the task and the available adaptation cohort. At 1%, the two brain tumor directions show little improvement over source-only, whereas larger adaptation subsets provide clearer gains. Performance is not strictly monotonic with sample fraction: the 25% setting slightly exceeds the 50% setting on TCGA-NSCLC → CPTAC-NSCLC and IPD-Brain → TCGA-GBMLGG. Overall, these results demonstrate useful adaptation with limited target data while retaining a separate, unseen target-test cohort.

A.7 Mechanistic Analyses

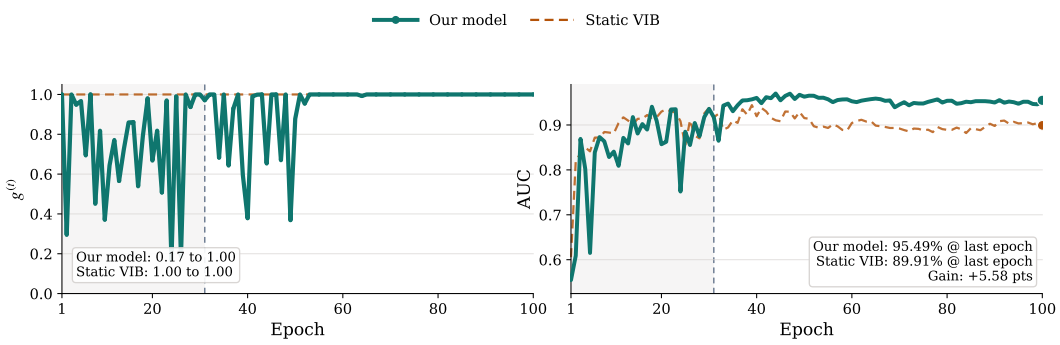
We examine intermediate representations and supervision signals to complement the downstream performance comparisons. All diagnostics requiring target ground-truth labels are conducted post hoc, after fixing the corresponding training runs and checkpoints.

A.7.1 Validation of CR-VIB Design

Comparison with Static VIB. To examine whether the improvement of CR-VIB arises merely from introducing a variational bottleneck or from its convergence-aware design, we compare three variants while keeping the remaining SMART components and training protocol unchanged: removing the VIB module entirely (w/o VIB), using a conventional Static VIB, and using our convergence-aware CR-VIB. The results on two representative transfer directions are reported in Table 10.

Table 10: Comparison of different VIB designs on two representative transfer directions. All metrics are reported in %. Bold indicates the highest mean in each task.

Method	AUC	Macro-F1
CPTAC-NSCLC → TCGA-NSCLC		
w/o VIB	64.46 ±20.42	52.54 ±17.48
Static VIB	92.90 ±4.00	79.36 ±11.00
CR-VIB (Ours)	**95.58** ±0.73	**86.53** ±2.30
TCGA-GBMLGG → IPD-Brain		
w/o VIB	85.66 ±1.83	47.57 ±3.36
Static VIB	84.99 ±1.50	50.06 ±4.92
CR-VIB (Ours)	**86.10** ±1.95	**50.28** ±3.84



(a) Dynamics of convergence-aware gate $g^{(t)}$ (b) Target-domain AUC

Figure 4: Training dynamics on the CPTAC-NSCLC → TCGA-NSCLC task. (a) The evolution of the convergence-aware gate $g^{(t)}$ across training epochs. (b) The target-domain AUC over training epochs. The plots compare our proposed model (solid green line) against the Static VIB baseline (dashed orange line).

On the challenging CPTAC-NSCLC → TCGA-NSCLC transfer, removing VIB leads to severe performance degradation and high variability. Introducing a Static VIB substantially recovers performance, increasing AUC from 64.46% to 92.90% and Macro-F1 from 52.54% to 79.36%. However, CR-VIB further improves these metrics to 95.58% and 86.53%, respectively, while markedly reducing the variability across splits. This indicates that the benefit does not arise solely from introducing information bottleneck regularization; adapting the compression process to model convergence is particularly important under severe domain shifts.

On TCGA-GBMLGG → IPD-Brain, the differences are more moderate. Static VIB slightly decreases AUC relative to removing VIB while improving Macro-F1, whereas CR-VIB achieves the highest mean AUC and Macro-F1. Together, these results suggest that convergence-aware compression provides its clearest advantage under challenging shifts while maintaining robust performance across different clinical scenarios.

Training Dynamics. To further examine the training behavior underlying the differences between static and convergence-aware compression, we visualize the training dynamics on the CPTAC-NSCLC → TCGA-NSCLC task within a single run (Figure 4). As shown in Figure 4(a),Static VIB applies a constant, full compression intensity ($g^{(t)} = 1.0$). Consequently, its target-domain AUC (Figure 4(b)) peaks early but gradually degrades to a suboptimal level. This confirms that forcing an immediate KL penalty suppresses weak yet essential diagnostic semantics before the classifier can sufficiently learn them. In contrast, our proposed CR-VIB dynamically adapts to the model's learning status. During early epochs, the convergence-aware gate $g^{(t)}$ actively fluctuates (dropping as low as 0.17) to temporarily relax the compression, safely preserving weak unlearned signals. As the model converges, $g^{(t)}$ smoothly stabilizes at 1.0 to execute full domain-invariant feature purification. This dynamic preservation-then-purification mechanism enables our model to achieve a substantially higher and more stable final AUC, effectively harmonizing domain alignment with discriminative power.

A.7.2 Diagnostic Relevance and Pseudo-Label Quality of SARA

Retrieval relevance. We measure the diagnostic relevance of source slides retrieved by SARA using Purity@5, defined as the fraction of the five retrieved source neighbors whose ground-truth class matches that of the target query. The random-retrieval baseline is computed under the corresponding source-class distribution.

SARA retrieval consistently exceeds the random baseline across all five adaptation directions (Table 11), indicating that the retrieved source neighbors carry meaningful diagnostic correspondence rather than reflecting the underlying source-class distribution alone. For example, on CPTAC-NSCLC → TCGA-NSCLC, Purity@5 reaches $78.33 \pm 5.03\%$, substantially higher than the random baseline of $49.98 \pm 0.08\%$. Similar improvements are observed in both directions of the brain tumor task, where SARA achieves Purity@5 values of $64.00 \pm 2.73\%$ and $66.95 \pm 1.01\%$, compared with random

Table 11: Diagnostic relevance of source retrieval across five adaptation directions. Purity@5 (%) is the fraction of the five retrieved source slides whose ground-truth class matches that of the target query. The random baseline accounts for the corresponding source-class distribution.

Transfer	SARA Purity@5	Random Purity@5
CPTAC-NSCLC → TCGA-NSCLC	78.33 ± 5.03	49.98 ± 0.08
TCGA-NSCLC → CPTAC-NSCLC	94.68 ± 1.95	50.00 ± 0.00
TCGA-GBMLGG → IPD-Brain	64.00 ± 2.73	30.31 ± 0.04
IPD-Brain → TCGA-GBMLGG	66.95 ± 1.01	30.16 ± 0.21
In-house Cervical → TCGA-CESC	54.57 ± 3.71	47.10 ± 0.05

baselines of $30.31 \pm 0.04\%$ and $30.16 \pm 0.21\%$, respectively. These results demonstrate that SARA retrieves diagnostically relevant cross-domain source neighbors, providing informative source-derived priors for target pseudo-label refinement.

Pseudo-label quality. We compare training variants without and with SARA at matched epochs 40 and 100 on the target-adaptation slides. The evaluated distributions are the soft pseudo-labels used to supervise the student under each variant. We report pseudo-label AUC, negative log-likelihood (NLL), and the correct-high-confidence rate. NLL is the mean negative log-probability assigned to the true class and penalizes confidently incorrect pseudo-labels. The correct-high-confidence rate is the fraction of all target-adaptation slides whose predicted class is correct and whose confidence reaches the training threshold τ_{conf}.

Table 12: Matched-epoch pseudo-label diagnostics on the target-adaptation slides. CPTAC → TCGA denotes the NSCLC task, and TCGA → IPD denotes the brain tumor task. Higher AUC and correct-high-confidence rate and lower NLL are preferable.

Transfer	Epoch	Variant	AUC (%)	NLL	Correct-high-conf. (%)
CPTAC → TCGA	40	w/o SARA	92.09 ± 2.18	1.4326 ± 0.3640	78.8 ± 5.2
		w/ SARA	92.22 ± 3.93	1.0741 ± 0.2087	81.0 ± 4.3
CPTAC → TCGA	100	w/o SARA	93.69 ± 2.24	0.7806 ± 0.1702	84.8 ± 1.6
		w/ SARA	94.86 ± 1.69	0.6754 ± 0.1878	84.1 ± 2.9
TCGA → IPD	40	w/o SARA	84.41 ± 1.98	1.2994 ± 0.0767	53.1 ± 1.9
		w/ SARA	85.64 ± 1.24	1.2406 ± 0.0960	56.3 ± 1.8
TCGA → IPD	100	w/o SARA	85.36 ± 2.29	1.2174 ± 0.0911	53.5 ± 1.7
		w/ SARA	86.84 ± 2.07	1.1583 ± 0.1089	57.4 ± 2.1

The correct-high-confidence rate uses all target-adaptation slides as its denominator, rather than only the selected high-confidence subset.

SARA consistently improves pseudo-label quality across both tasks and training stages (Table 12). Across all four matched-epoch comparisons, SARA yields higher pseudo-label AUC and lower NLL, indicating more discriminative and better-calibrated soft supervision. On CPTAC-NSCLC → TCGA-NSCLC at epoch 40, NLL decreases from 1.4326 ± 0.3640 to 1.0741 ± 0.2087, while the correct-high-confidence rate increases from $78.8 \pm 5.2\%$ to $81.0 \pm 4.3\%$. On TCGA-GBMLGG → IPD-Brain at epoch 100, AUC increases from $85.36 \pm 2.29\%$ to $86.84 \pm 2.07\%$, accompanied by an increase in the correct-high-confidence rate from $53.5 \pm 1.7\%$ to $57.4 \pm 2.1\%$. Overall, the correct-high-confidence rate improves in three of the four comparisons and remains comparable in the remaining setting. These results demonstrate that retrieval-augmented supervision consistently improves both the ranking quality and probabilistic reliability of pseudo-labels, providing cleaner and more reliable supervision for target-domain adaptation.

A.7.3 Complementarity and Representation Geometry of REALM

Factorial component analysis. REALM combines entropy-based Soft-DRO reweighting with cross-difficulty latent mixup. We conduct a 2×2 factorial ablation with both mechanisms disabled, Soft-DRO only, mixup only, and both enabled. The mixup-only variant assigns every sample a weight of one.

Table 13: Factorial ablation of Soft-DRO and cross-difficulty mixup using macro-F1 (%). In the mixup-only variant, all sample weights are fixed to one. Bold indicates the highest mean in each row.

Transfer	No REALM	Soft-DRO only	Mixup only	Full REALM
CPTAC-NSCLC → TCGA-NSCLC	85.66 ± 3.67	82.38 ± 3.23	84.04 ± 7.34	$\mathbf{86.53 \pm 2.30}$
TCGA-GBMLGG → IPD-Brain	45.59 ± 3.19	47.56 ± 2.52	44.53 ± 2.96	$\mathbf{50.28 \pm 3.84}$

Full REALM achieves the highest mean macro-F1 on both transfer tasks (Table 13), outperforming the stronger isolated component by 2.49 percentage points on CPTAC-NSCLC → TCGA-NSCLC and 2.72 points on TCGA-GBMLGG → IPD-Brain. Importantly, neither constituent alone reproduces the performance of the complete module. On the NSCLC task, Soft-DRO and cross-difficulty mixup individually reduce macro-F1 relative to disabling REALM, whereas jointly enabling them increases performance from $85.66 \pm 3.67\%$ to $86.53 \pm 2.30\%$. A similar complementary effect is observed on the brain tumor task: Soft-DRO alone provides a modest improvement and mixup alone does not improve the baseline, yet their combination raises macro-F1 from $45.59 \pm 3.19\%$ to $50.28 \pm 3.84\%$.

The corresponding factorial interaction gains are $+5.77$ and $+3.78$ percentage points on the two tasks, respectively, revealing a consistent positive interaction between difficulty-aware reweighting and cross-difficulty interpolation. These results show that REALM is not simply an additive combination of two independently beneficial operations; rather, its effectiveness depends on coupling sample reweighting with latent-space interpolation, allowing the two mechanisms to compensate for the limitations of either component used in isolation.

Geometry of hard target slides. We further compare target representations with and without REALM on CPTAC-NSCLC → TCGA-NSCLC. A normalized-entropy ranking from the source-initialized teacher defines the same hard/easy partition for both variants: the highest-entropy 50% of target slides are designated as hard, and the remainder as easy. For each diagnostic class, an easy-class centroid is computed from its ground-truth easy slides. For each hard slide, we measure cosine similarity to the correct-class easy centroid and to the nearest incorrect-class easy centroid.

We report three statistics. The hard-slide centroid margin is the correct-class similarity minus the highest incorrect-class similarity. The nearest wrong-class centroid similarity measures proximity to the most confusable incorrect-class region. The positive-margin rate is the proportion of hard slides whose correct-class similarity exceeds their similarity to every incorrect-class centroid.

Table 14: Post-hoc geometry of hard target slides on CPTAC-NSCLC → TCGA-NSCLC. The two variants use the same hard/easy partition derived from the source-initialized teacher. Larger centroid margins and positive-margin rates and smaller nearest wrong-class centroid similarities are preferable.

Variant	Hard-slide centroid margin	Nearest wrong-class similarity	Positive-margin rate (%)
w/o REALM	0.2157 ± 0.0730	0.7034 ± 0.0847	86.14 ± 3.46
w/ REALM	0.2406 ± 0.1021	0.6504 ± 0.1299	87.16 ± 1.84

With REALM, the mean hard-slide centroid margin increases from 0.2157 to 0.2406, while the similarity to the nearest incorrect-class centroid decreases from 0.7034 to 0.6504. The positive-margin rate also increases from 86.14% to 87.16% (Table 14). Taken together, these consistently directed changes indicate that REALM reshapes the representations of difficult target samples toward a more discriminative class-relative geometry: hard slides become better aligned with their corresponding class structure while being separated from the most confusable incorrect-class regions.

This geometric evidence provides a representation-level explanation for the performance gains observed in the REALM ablations.

Table 15: Comparison between random and difficulty-aware interpolation on the CPTAC-NSCLC → TCGA-NSCLC task.

Method	AUC (%)	ACC (%)	Macro-F1 (%)
Random Mixup	$93.13_{\pm 1.72}$	$84.72_{\pm 1.64}$	$84.48_{\pm 1.66}$
Ours (Full)	$\mathbf{95.58}_{\pm 0.73}$	$\mathbf{86.69}_{\pm 2.23}$	$\mathbf{86.53}_{\pm 2.30}$

Validation of Difficulty-Aware Interpolation. To isolate the contribution of the proposed difficulty-aware interpolation strategy, we replace cross-difficulty pairing in REALM with random interpolation on the CPTAC-NSCLC → TCGA-NSCLC task (Table 15). In this Random Mixup variant, interpolation partners are sampled uniformly without considering target-sample difficulty.

REALM consistently outperforms Random Mixup across all three evaluation metrics, improving AUC from $93.13 \pm 1.72\%$ to $95.58 \pm 0.73\%$, accuracy from $84.72 \pm 1.64\%$ to $86.69 \pm 2.23\%$, and macro-F1 from $84.48 \pm 1.66\%$ to $86.53 \pm 2.30\%$, corresponding to gains of 2.45, 1.97, and 2.05 percentage points, respectively. These consistent improvements show that the benefit does not arise from latent interpolation alone, but from explicitly structuring the interpolation according to target-sample difficulty. Unlike random pairing, which treats all potential interpolation partners indiscriminately, our cross-difficulty strategy deliberately couples difficult target samples with more reliable representations, providing more informative interpolation trajectories for adaptation. Together with the factorial analysis in Table 13, these results further demonstrate that difficulty-aware interpolation is an essential component of REALM rather than a replaceable generic mixup operation.

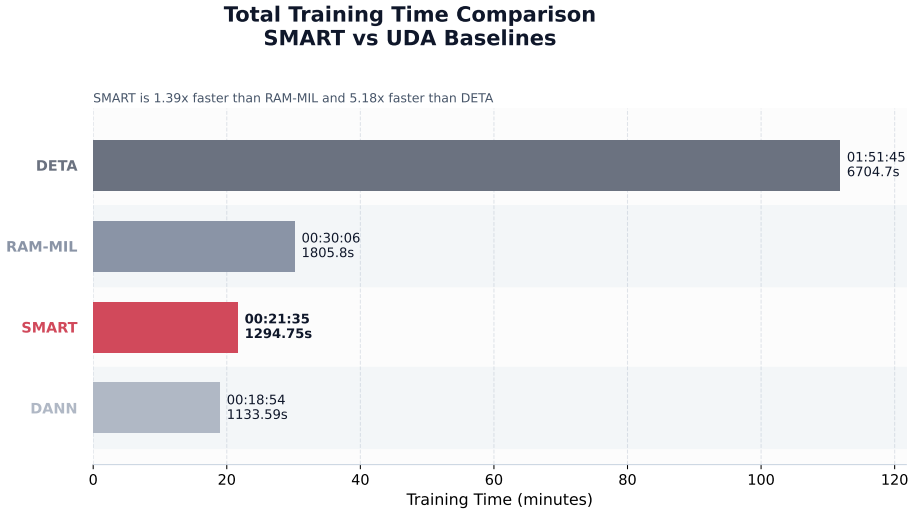


Figure 5: Total training time comparison on the IPD-Brain → TCGA-GBMLGG task. The bar chart displays the training duration of our proposed SMART framework alongside representative UDA baselines (DANN, RAM-MIL, and DETA).

A.8 Computational Efficiency Analysis

To evaluate the practical scalability of our framework, we compare the total training time of SMART against representative UDA baselines (Figure 5). As illustrated, SMART achieves highly efficient convergence, completing the training process in just 21.5 minutes (1294.75s). In stark contrast, graph-based approaches like DETA incur substantial computational overhead (6704.70s), as they necessitate expensive graph construction and alignment across both source and target domains.

Consequently, SMART operates 5.18× faster than DETA and 1.39× faster than the recent RAM-MIL baseline. While the naive DANN architecture is marginally faster (1133.59s) due to its simplistic design, previous results establish that it yields substantially lower diagnostic performance. This comparison demonstrates that SMART strikes an optimal balance, delivering state-of-the-art cross-domain performance while maintaining a lightweight and highly scalable training footprint suitable for real-world clinical deployment.

A.9 Extended Generalizability across Cohorts and MIL Backbones

A.9.1 Three-Cohort Evaluation of Brain Tumor Classification

We extend the brain tumor evaluation to EBRAINS, complementing the original TCGA-GBMLGG and IPD-Brain cohorts. EBRAINS diagnoses are harmonized to the same five IDH1-status/histology classes: IDH1-mutant astrocytoma, IDH1-mutant oligodendroglioma, IDH1-mutant glioblastoma, IDH1-wild-type astrocytoma, and IDH1-wild-type glioblastoma. Using DTFD-MIL and the same globally fixed SMART configuration, we evaluate the four additional pairwise transfer directions involving EBRAINS. Together with the original two brain tumor directions, these experiments cover all six directed pairwise transfers among the three cohorts.

Table 16: Generalization to EBRAINS with DTFD-MIL. All metrics are percentages. Bold indicates the higher mean within each source-only/SMART pair.

Transfer	Method	AUC	ACC	Macro-F1
EBRAINS → IPD-Brain	Source-only	77.44 ± 2.10	64.52 ± 1.34	37.92 ± 3.29
	SMART	$\mathbf{84.72 \pm 2.19}$	$\mathbf{71.58 \pm 3.68}$	$\mathbf{51.86 \pm 4.72}$
IPD-Brain → EBRAINS	Source-only	79.49 ± 1.65	39.34 ± 1.65	24.35 ± 0.70
	SMART	$\mathbf{85.56 \pm 1.80}$	$\mathbf{75.09 \pm 2.18}$	$\mathbf{47.80 \pm 1.63}$
EBRAINS → TCGA-GBMLGG	Source-only	83.54 ± 0.79	69.43 ± 1.28	$\mathbf{51.68 \pm 2.08}$
	SMART	$\mathbf{85.95 \pm 1.09}$	$\mathbf{71.45 \pm 0.86}$	50.44 ± 3.77
TCGA-GBMLGG → EBRAINS	Source-only	85.60 ± 1.40	69.48 ± 1.98	48.55 ± 2.50
	SMART	$\mathbf{88.18 \pm 1.37}$	$\mathbf{71.18 \pm 1.35}$	$\mathbf{54.68 \pm 2.30}$

SMART consistently improves cross-cohort transferability across all four EBRAINS-related transfer directions (Table 16). Specifically, SMART achieves consistent gains in both AUC and accuracy for every source-target pair, demonstrating its robustness under diverse cohort shifts. The largest improvement is observed on IPD-Brain → EBRAINS, where accuracy increases from $39.34 \pm 1.65\%$ to $75.09 \pm 2.18\%$ and macro-F1 improves from $24.35 \pm 0.70\%$ to $47.80 \pm 1.63\%$. Meanwhile, SMART also provides substantial improvements in class-balanced performance for most transfer settings, with macro-F1 gains in three out of four directions. These results demonstrate that SMART is not tailored to a specific dataset pair, but provides a generally applicable adaptation strategy across heterogeneous WSI cohorts.

A.9.2 Transformer-Based and Graph-Based MIL Backbones

To examine compatibility with different aggregation mechanisms, we integrate SMART with TransMIL and WiKG, representing transformer-based and graph-based MIL architectures, respectively. Each adapted model is compared with its corresponding source-only backbone using the same data splits and adaptation protocol.

As reported in Table 17, SMART consistently improves transfer performance across different MIL aggregation paradigms. Among the 24 backbone–task–metric combinations, SMART achieves improvements in 22 cases, demonstrating that its effectiveness is not tied to a specific MIL architecture. In particular, accuracy and macro-F1 are consistently improved across all eight backbone–task settings, indicating more reliable class-balanced decision boundaries under heterogeneous domain shifts.

Together with the DTFD-MIL and CLAM experiments, these results demonstrate that SMART provides a broadly applicable adaptation strategy across diverse MIL architectures, including attention-based, transformer-based, and graph-based aggregators.

Table 17: Generalization across transformer-based and graph-based MIL backbones. All metrics are percentages. The lung tasks use CPTAC-NSCLC and TCGA-NSCLC, and the brain tasks use TCGA-GBMLGG and IPD-Brain. Bold indicates the higher mean within each source-only/SMART pair.

Transfer	Method	AUC	ACC	Macro-F1
TransMIL				
CPTAC → TCGA (NSCLC)	Source-only	62.69 ± 2.26	50.53 ± 1.79	33.56 ± 0.79
	SMART	$\mathbf{87.51 \pm 3.42}$	$\mathbf{85.50 \pm 3.03}$	$\mathbf{85.46 \pm 3.00}$
TCGA → CPTAC (NSCLC)	Source-only	87.07 ± 2.50	53.59 ± 3.51	46.11 ± 3.60
	SMART	$\mathbf{98.77 \pm 1.30}$	$\mathbf{96.48 \pm 2.29}$	$\mathbf{96.47 \pm 2.30}$
TCGA → IPD (Brain)	Source-only	82.98 ± 2.61	63.63 ± 1.15	30.69 ± 1.12
	SMART	$\mathbf{84.93 \pm 4.50}$	$\mathbf{71.40 \pm 1.93}$	$\mathbf{53.88 \pm 5.13}$
IPD → TCGA (Brain)	Source-only	$\mathbf{80.94 \pm 1.17}$	56.81 ± 1.36	35.21 ± 1.49
	SMART	79.11 ± 1.40	$\mathbf{67.28 \pm 1.84}$	$\mathbf{45.66 \pm 2.78}$
WiKG				
CPTAC → TCGA (NSCLC)	Source-only	57.05 ± 1.80	50.53 ± 1.79	33.56 ± 0.79
	SMART	$\mathbf{87.72 \pm 2.75}$	$\mathbf{83.62 \pm 2.34}$	$\mathbf{83.54 \pm 2.31}$
TCGA → CPTAC (NSCLC)	Source-only	49.45 ± 3.27	46.75 ± 2.07	31.85 ± 0.96
	SMART	$\mathbf{98.73 \pm 2.15}$	$\mathbf{97.65 \pm 1.55}$	$\mathbf{97.64 \pm 1.56}$
TCGA → IPD (Brain)	Source-only	74.81 ± 4.12	54.86 ± 2.70	31.45 ± 3.32
	SMART	$\mathbf{82.13 \pm 3.44}$	$\mathbf{67.28 \pm 5.55}$	$\mathbf{49.46 \pm 6.30}$
IPD → TCGA (Brain)	Source-only	$\mathbf{79.44 \pm 1.25}$	49.66 ± 1.07	38.03 ± 1.28
	SMART	77.24 ± 1.65	$\mathbf{65.92 \pm 2.34}$	$\mathbf{44.81 \pm 2.53}$

A.10 Extended Ablation Studies

To investigate whether individual components contribute consistently under different domain shifts, we extend the component ablation study to all five original adaptation directions and evaluate class-balanced performance using macro-F1 (Table 18). Full SMART achieves the best macro-F1 in four out of five transfer directions and remains highly competitive in the remaining setting, where the difference from the best-performing variant is within 0.10 percentage points.

CR-VIB has the largest impact on CPTAC-NSCLC → TCGA-NSCLC, where removing it reduces macro-F1 from $86.53 \pm 2.30\%$ to $52.54 \pm 17.48\%$. Across the other transfer directions, removing CR-VIB, SARA, and REALM decreases macro-F1 by an average of 1.57, 1.53, and 3.42 percentage points, respectively. These results indicate that the three components provide complementary benefits under heterogeneous domain shifts. In particular, CR-VIB is critical for mitigating severe feature-level discrepancies, while SARA and REALM further improve adaptation by exploiting semantic source priors and enhancing robustness to difficult target samples.

Table 18: Component ablations across all five original adaptation directions, evaluated using macro-F1 (%). Bold indicates the highest mean in each row.

Transfer	w/o CR-VIB	w/o SARA	w/o REALM	Full SMART
CPTAC-NSCLC → TCGA-NSCLC	52.54 ± 17.48	86.14 ± 1.59	85.66 ± 3.67	$\mathbf{86.53 \pm 2.30}$
TCGA-NSCLC → CPTAC-NSCLC	91.82 ± 7.70	92.73 ± 2.33	92.55 ± 3.16	$\mathbf{94.07 \pm 2.78}$
TCGA-GBMLGG → IPD-Brain	47.57 ± 3.36	49.63 ± 4.14	45.59 ± 3.19	$\mathbf{50.28 \pm 3.84}$
IPD-Brain → TCGA-GBMLGG	$\mathbf{48.38 \pm 2.56}$	45.87 ± 1.79	44.94 ± 1.02	48.28 ± 2.01
In-house Cervical → TCGA-CESC	54.73 ± 5.04	54.43 ± 6.66	52.02 ± 4.60	$\mathbf{56.14 \pm 4.34}$