DARE: Dual-Level Adversarial Learning with Domain-Aware Regularization for Whole Slide Image Classification

Yuxuan Jiang[1,2*] **Haichuan Dong**[1,2*] **Manning Wang**[1,2†]

[1] Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, China
[2] Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, China
{25211010025,25111010027}@m.fudan.edu.cn, mnwang@fudan.edu.cn

Abstract

Pathological whole slide images (WSIs) exhibit substantial domain shifts across hospitals due to variations in scanning devices, staining protocols, and tissue preparation procedures, leading to significant performance degradation when models are applied to data from different sources. Although unsupervised domain adaptation (UDA) has shown strong performance in transferring models from a labeled source domain to an unlabeled target domain in natural images, it remains relatively underexplored in computational pathology. The ultra-high resolution of WSIs and their pronounced inter-slide and intra-slide heterogeneity pose unique challenges for domain adaptation, often rendering existing UDA methods unstable and susceptible to negative transfer. To address these challenges, we propose Dual-level Adversarial learning with domain-aware REgularization (DARE), a plug-and-play UDA framework for WSI classification. DARE incorporates adaptive pseudo-labeling within a teacher-student architecture to provide stable supervision for the target domain, alongside adversarial alignment of both patch-level and slide-level embeddings. Together, these components enable the joint optimization of semantic representations at both levels for cross-domain alignment. We further introduce domain-aware attention masking to regularize attention and improve cross-domain generalization. Extensive experiments on four public WSI datasets demonstrate that DARE consistently outperforms state-of-the-art approaches across various domain shifts.

1 Introduction

Classification of pathological whole slide images (WSIs) plays an important role in clinical cancer detection, subtyping, and treatment response prediction, among others. Recently, deep learning has achieved remarkable success in WSI classification tasks, reaching or exceeding human-level performance [7, 32, 40, 33]. However, the performance of these models often degrades when deployed across different clinical centers [3]. This degradation is mainly caused by domain shifts between medical institutions, including variations in scanning devices, staining protocols, and tissue preparation procedures. As shown in Fig. 1a, noticeable staining differences can be observed between the CPTAC-NSCLC and TCGA-NSCLC datasets.

Due to the extremely high resolution of WSIs, reaching up to 100,000 $\times$ 100,000 pixels, they are typically divided into a large number of patches and modeled under a multiple instance learning (MIL) framework [24, 26, 29, 31], where a patch-level feature extractor is applied first to obtain

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

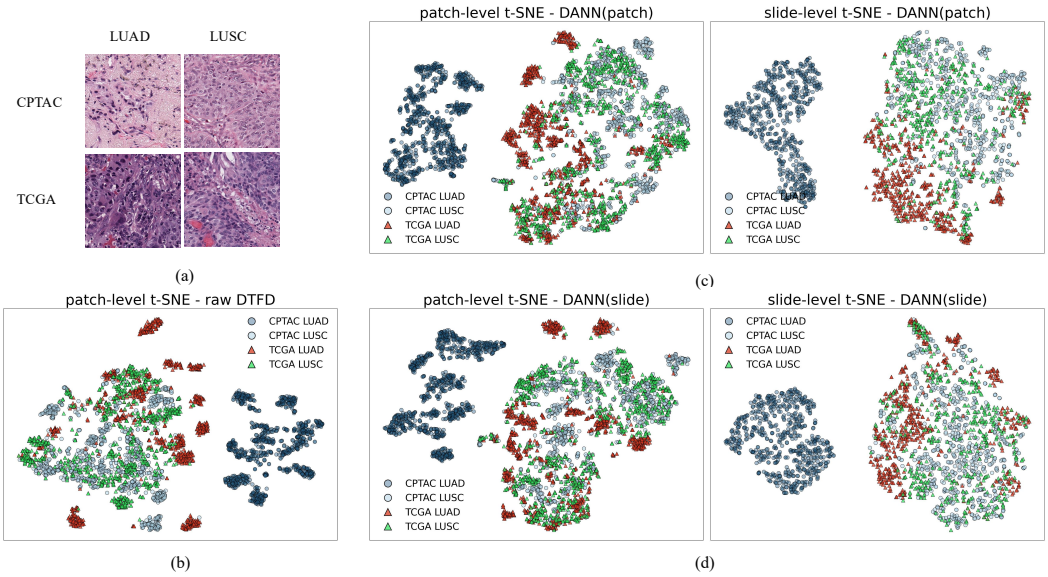


Figure 1: Illustration of domain discrepancy and DANN-based feature alignment on the CPTAC → TCGA (NSCLC) task. (a) Example patches. (b) t-SNE of original DTFD-aggregated slide-level features. (c) Applying DANN to patch-level features: (left) patch-level, (right) slide-level t-SNE. (d) Applying DANN to slide-level features: (left) patch-level, (right) slide-level t-SNE. CONCH is used for feature extraction.

local patch representations and an attention-based aggregator is then used to aggregate these patch features into a slide-level feature for classification under slide-level supervision [19, 36, 47]. In recent years, pathology foundation models have advanced rapidly, and the current mainstream approach is to use a frozen foundation model encoder for patch-level feature extraction. Although foundation models are generally considered more generalizable, they still cannot fully address the domain shift problem. We extract patch-level features from the CPTAC-NSCLC and TCGA-NSCLC datasets using CONCH [27] as the encoder and visualize their feature distributions in Fig. 1b. Substantial distribution shifts can be observed across the two datasets in the feature space. Such distribution shifts make it still difficult for models trained on one domain to generalize to another domain. This requires additional annotations from pathologists on the target domain data, followed by retraining on the labeled target data, which significantly increases both time and labor costs.

Unsupervised domain adaptation (UDA) offers a promising paradigm to address this issue. It enables a model trained on a labeled source domain to transfer to an unlabeled target domain [2, 34]. Many UDA methods have been developed for natural images, but they usually do not generalize well to WSI classification. Our experiments show that directly applying these methods to pathology often leads to instability or even negative transfer. There are some UDA studies in the computational pathology field, and they mainly operate at the image level. They focus on color normalization [28, 1, 43] or style transfer [37, 45]. However, these methods fail to capture high-level semantic structures in the feature space. Moreover, these image-level methods are not well aligned with the current paradigm of patch-level feature extraction using foundation models.

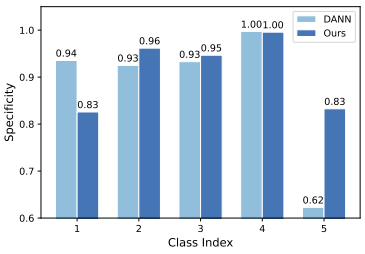


Figure 2: Specificity comparison on IPD → TCGA (GBMLGG).

To improve UDA performance for WSI classification under the current mainstream pipeline of extracting patch features using a frozen pathology foundation model, we design a novel framework to perform domain adversarial learning at the feature level. The framework is motivated by the specific characteristics of UDA in WSI classification. First, as shown in Fig. 1c and Fig. 1d, we apply the classical domain adversarial learning method DANN [10] at the patch level and at the slide level,

respectively, and visualize the resulting feature distributions. We observe that applying DANN at a single granularity yields poor feature separation at both levels: neither patch-level features nor slide-level features can be well separated by subtype in the target domain. WSIs naturally exhibit a two-level structure, where patch-level and slide-level features capture different granularities of information. Domain shifts can exist at both levels, so aligning only one may not be enough. Second, most regions within a WSI correspond to background or normal tissue, while diagnostically relevant regions are often sparse, and when aggregating patch features using the attention mechanism, the final prediction highly depends on a small subset of discriminative patches [14, 38, 15, 48]. The bias amplification effect in the attention mechanisms for aggregating patch-level features can harm generalization [25, 52]. In cross-domain settings, a domain-agnostic aggregator cannot distinguish whether a highly attended patch carries transferable semantics or domain-specific shortcuts. This calls for a domain-aware attention regularization strategy that can suppress domain-specific easy patches and encourage learning from harder, domain-invariant evidence. Third, we compare the per-class specificity of our method and slide-level DANN. As illustrated in Fig. 2, DANN suffers from poor specificity on the fifth class, revealing that global domain alignment does not guarantee class-wise alignment. Achieving category-level alignment across domains requires a mechanism that provides class-specific supervision in the target domain to separate the distributions of different categories.

Accordingly, we propose Dual-level Adversarial learning with domain-aware REgularization (DARE), a framework for WSI domain adaptation that integrates dual-level adversarial alignment, domain-aware instance selection, and adaptive pseudo-label supervision. Specifically, DARE performs patch-level adversarial learning to reduce local feature discrepancies between the source and target domains. Meanwhile, patch-level domain predictions are combined with attention scores to identify and mask highly attended target patches that exhibit strong source-domain characteristics. The retained patches are then re-aggregated into target slide representations and adversarially aligned with source slide representations. In this way, patch-level domain information guides local instance selection and determines which tissue evidence contributes to slide-level alignment, coupling adaptation across the two levels. Furthermore, adaptive soft pseudo-label supervision from the unmasked teacher guides the masked student to preserve class semantics while learning from the remaining evidence. Together, these components connect local feature alignment, domain-aware instance selection, global distribution alignment, and semantic preservation.

Our main contributions are summarized as follows:

- We propose a domain-aware attention masking strategy that combines patch importance with domain predictions to identify and suppress highly attended, source-specific patches. This strategy incorporates domain information into instance selection and encourages the model to learn from more transferable tissue evidence.
- We develop a cross-level adaptation mechanism in which patch-level domain predictions guide instance selection, and the retained instances are re-aggregated for slide-level adversarial alignment. Adaptive teacher-student pseudo-label supervision preserves class semantics during this process, coupling local instance selection with global domain adaptation.
- We evaluate DARE on four public datasets and demonstrate superior performance over state-of-the-art UDA methods, including under limited target-domain data. Experiments with alternative attention-based MIL backbones and pathology foundation encoders further demonstrate its plug-and-play applicability and compatibility with different feature representations.

2 Related Work

2.1 Multiple Instance Learning for WSI Classification

MIL for WSI classification follows a dual-level structure: patch-level features are first extracted from individual patches, and then aggregated into slide-level representations for final prediction. Most existing MIL methods are based on attention mechanisms for feature aggregation. For example, ABMIL [14] assigns an attention score to each patch and performs weighted aggregation to obtain slide-level features. DTFD-MIL [50] further improves MIL from the perspective of feature distribution modeling. These methods typically rely on pretrained ResNet [6] models for patch-level feature extraction. With the emergence of pathology foundation models, their use as feature extractors for

patch-level representations has become a dominant paradigm in WSI analysis. For example, HistAUG [3] uses CONCH [27] and UNI [5] for feature extraction and introduces spatial augmentation on top of attention-based frameworks to improve generalization. MAPLE [53] uses PLIP [13] as the patch encoder and proposes an entity-guided cross-attention mechanism.

Although these methods have achieved significant progress in single domain settings, most of them assume that training and test data follow the same distribution [54, 51]. They largely lack explicit modeling of cross-center distribution shifts, which limits their generalization in real-world clinical scenarios.

2.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation aims to learn models that generalize well to an unlabeled target domain using labeled source domain data and unlabeled target domain data. Existing methods can be broadly categorized into two groups: distribution alignment methods and self-training methods. Distribution alignment methods aim to reduce the discrepancy between source and target feature distributions. They can be further divided into two subcategories: methods based on statistical distance and methods based on adversarial learning. Statistical distance methods minimize explicit divergence measures, such as maximum mean discrepancy or optimal transport, to align the distributions [22, 41, 49, 8, 30]. Adversarial learning methods introduce a domain classifier and train the feature extractor in an adversarial manner to produce domain-invariant representations [10, 23, 35, 44, 4]. In computational pathology, RAM-MIL [9] is based on statistical distance and leverages optimal transport for nearest-neighbor retrieval to reduce slide-level domain shifts. DETA [39] is based on adversarial learning and employs an adversarial perturbation strategy to augment source domain features, performing slide-level adversarial alignment. However, these distribution alignment methods largely ignore the hierarchical patch-slide relationship inherent to MIL-based WSI analysis. In contrast, self-training methods generate pseudo-labels for target domain samples and iteratively refine the model using these pseudo-labels to improve target domain discrimination [17, 20, 18, 11]. Therefore, our method DARE integrates dual-level adversarial learning with self-training, explicitly aligning features at both patch and slide levels while introducing pseudo-labeling to provide supervision for the target domain.

3 Method

3.1 MIL Problem Formulation

For the source domain, we denote the dataset as $S = \{(I_i^s, Y_i^s)\}_{i=1}^{N_s}$, where each WSI $I_i^s = \{p_{i,j}^s\}_{j=1}^{n_i}$ is composed of n_i patches. The label $Y_i^s \in \{1, \ldots, C\}$ denotes the slide-level annotation. For the target domain, we have an unlabeled dataset $T = \{I_i^t\}_{i=1}^{N_t}$, which contains only WSIs without annotations. In attention-based MIL, a feature extractor $f(\cdot)$ is first applied to each patch to obtain patch-level features $h_{i,j}^D$, where $D \in \{s, t\}$. An attention-based aggregator $\mathrm{Agg}(\cdot)$ is then used to obtain slide-level features F_i^D by pooling patch-level features. Finally, a classifier $g(\cdot)$ operates on the slide features to produce the predictions $\hat{y}_i^D = g(F_i^D)$. Under the unsupervised domain adaptation setting, the model is trained using only slide-level labels from the source domain while leveraging unlabeled data from the target domain to learn representations with strong cross-domain generalization.

3.2 Method Overview

As shown in Fig. 3, the WSI is first divided into patches, and a pathology foundation model is used to extract patch-level features $h_{i,j}^D$. After a projection network $P(\cdot)$, these features are mapped into an embedding space, yielding $x_{i,j}^D$. Patch-level embeddings $x_{i,j}^D$ are aggregated by an attention network to obtain slide-level embeddings for the source and target domains, denoted as F_{slide}^D. For the source domain data, we follow the standard WSI classification pipeline. The source slide-level embedding F_{slide}^s is fed into a classifier shared with the student network to give a slide-level prediction $\hat{y}^s$.

In the proposed DARE framework, we first adopt a patch-level domain classifier Q_{ins} to adversarially align patch-level embeddings of the source and target domains. Meanwhile, for each target domain

patch, Q_{ins} outputs a probability that the patch belongs to the source domain, and the attention module produces an attention score. By jointly considering these two quantities, we select and mask the Top-K patches. The remaining patch embeddings are re-aggregated to obtain a masked slide-level embedding F^t_{mask} for the target domain. We further adopt a slide-level domain classifier Q_{slide} to adversarially align F^s_{slide} from the source domain and F^t_{mask} from the target domain. In parallel, we can also aggregate the target domain patch embeddings through standard attention without masking, producing an unmasked slide-level embedding F^t_{slide}. Finally, a teacher-student pseudo-labeling framework is adopted to provide supervisory signals for the target domain. The teacher classifier takes the unmasked embedding F^t_{slide} as input and generates pseudo-labels. The student classifier takes the masked embedding F^t_{mask} and learns to predict these pseudo-labels. Importantly, the source and target domains share the same student classifier: source slide embeddings F^s_{slide} are supervised by ground-truth labels, while masked target embeddings F^t_{mask} are supervised by pseudo-labels from the teacher. In inference, both domain classifiers and the attention masking module are removed, leaving only the attention-based MIL aggregator and the student classifier for prediction.

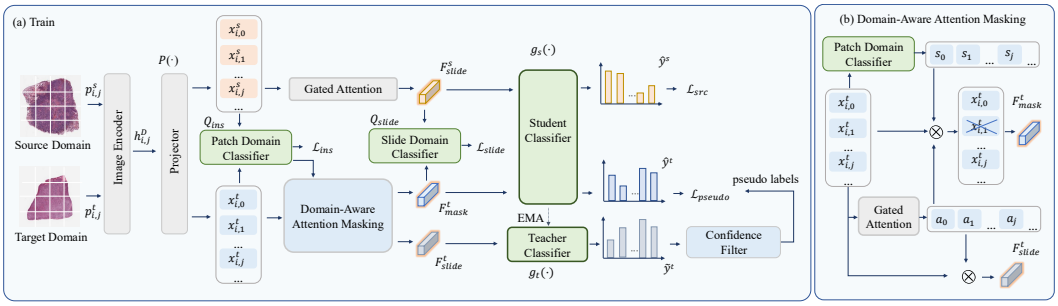


Figure 3: Overview of the DARE framework. (a) Training pipeline with multi-granularity domain alignment, attention masking, and pseudo-label learning. (b) Domain-aware attention masking: masking patches guided by domain and attention distributions.

3.3 Dual-Level Domain Adversarial Learning

The domain shift between the source and target domains is reflected not only in local cellular and tissue-level texture appearance at the patch level, but also in global contextual distributions of the tissue microenvironment at the slide level. To address this, we introduce a dual-level domain adversarial module.

Specifically, the patch-level domain classifier Q_{ins} distinguishes whether a patch embedding $x^D_{i,j}$ comes from the source or the target domain. Its adversarial loss $\mathcal{L}_{ins}$ is defined as:

$$\mathcal{L}_{ins} = \mathrm{CE}(Q_{ins}(x^D_{i,j}), d), \tag{1}$$

$\mathrm{CE}(\cdot)$ denotes the standard cross-entropy function, and $d \in \{0, 1\}$ is the domain label. To achieve cross-domain embedding alignment, we introduce an adversarial learning scheme between the projector $P(\cdot)$ and the domain classifier Q_{ins}. The projector $P(\cdot)$ is encouraged to produce domain-invariant representations. Formally, this process can be cast as a mini-max optimization problem: the domain classifier Q_{ins} aims to maximize domain classification accuracy, while $P(\cdot)$ aims to minimize this ability, thereby aligning the distributions of patch embeddings from the source and target domains. In practice, we insert a gradient reversal layer (GRL) after the patch embeddings. During the forward pass, the GRL acts as an identity mapping and passes embeddings to the domain classifier. During backpropagation, the GRL multiplies the gradients from the domain classifier by a negative coefficient, effectively reversing their direction before passing them to the projector. As a result, the projector is optimized to confuse the domain classifier, leading to more domain-invariant embeddings.

Then, we introduce a slide-level domain classifier Q_{slide} for slide-level embeddings and enforce adversarial alignment. For the target domain, we apply the domain-aware attention masking mechanism to patch embeddings to obtain an aggregated slide embedding F^t_{mask}. We align it with the source slide embedding F^s_{slide} by minimizing the slide-level adversarial loss $\mathcal{L}_{slide}$.

3.4 Domain-Aware Attention Masking

We revisit the role of attention under cross-domain settings and introduce a domain-aware attention masking strategy. Specifically, patches with both high attention scores and high source-domain probability from Q_{ins} are treated as domain-specific easy patches and are dynamically masked.

As illustrated in Fig. 3b, for a target domain WSI indexed by i, let $X_i^t = \{x_{i,1}^t, x_{i,2}^t, \ldots, x_{i,n_i}^t\}$ denote its patch-level embedding set, where n_i is the number of patches. A gated attention mechanism is applied to compute an attention score for each patch. The attention weights $a_{i,j} \in [0,1]$ and the aggregated slide-level embedding F_{slide}^t are defined as:

$$a_{i,j} = \frac{\exp\left(w^T\left(\tanh(Vx_{i,j}^t) \odot \sigma(Ux_{i,j}^t)\right)\right)}{\sum_{q=1}^{n_i} \exp\left(w^T\left(\tanh(Vx_{i,q}^t) \odot \sigma(Ux_{i,q}^t)\right)\right)}, \tag{2}$$

$$F_{slide}^t = \sum_{j=1}^{n_i} a_{i,j}\, x_{i,j}^t, \tag{3}$$

where w, V, and U are learnable parameters of the attention network, and $\odot$ denotes element-wise multiplication in the gating mechanism.

Meanwhile, for each target domain patch $x_{i,j}^t$, we use a patch-level domain classifier Q_{ins} to obtain a probability $p_{i,j}$ that it belongs to the source domain. We define a joint masking score by combining the attention weight and the domain probability as $m_{i,j} = p_{i,j} \times a_{i,j}$. We rank all patches in X_i^t by $m_{i,j}$ and mask the Top-K ones, where $K = \lceil r \cdot n_i \rceil$ and r is a predefined masking ratio. This yields the masked embedding set $\widetilde{X}_i^t$. The attention weights over the remaining patches in $\widetilde{X}_i^t$ are re-normalized to sum to one and aggregated into the masked slide-level embedding F_{mask}^t.

3.5 Teacher-Student Based Adaptive Pseudo-label Learning

For each target domain slide, the domain-aware attention masking module produces an unmasked slide embedding F_{slide}^t via standard attention aggregation, and a masked slide embedding F_{mask}^t by suppressing domain-specific easy patches and re-aggregating the remaining ones. F_{mask}^t flows into the student branch, passes through a student classifier $g_s(\cdot)$ shared with the source domain, and produces a prediction $\hat{y}^t$. F_{slide}^t flows into the teacher branch and passes through a teacher classifier $g_t(\cdot)$ to generate soft pseudo-labels $\tilde{y}^t$. The student is trained to match these pseudo-labels. The teacher receives no gradient updates; its parameters are an exponential moving average (EMA) of the student parameters.

To improve pseudo-label quality and training stability, we further adopt pseudo-label memory maintenance for building a global bank with ensemble predictions, adaptive threshold selection for filtering noisy labels via smoothly evolving per-class thresholds, and class-balanced weighting for mitigating class imbalance, as detailed below.

Pseudo-Label Memory Maintenance. We maintain a global pseudo-label bank that stores soft pseudo-labels for all target domain slides. At the beginning of each epoch, we use the teacher network's attention parameters from the end of the previous epoch to aggregate the patch features of every target domain slide, producing their unmasked slide embeddings F_{slide}^t. The teacher classifier then processes these embeddings to produce soft pseudo-labels, which are stored in the bank. The soft pseudo-label for each target slide is computed as:

$$\tilde{y}^t = Softmax(g_t(F_{slide}^t)), \tag{4}$$

where g_t denotes the teacher network's classifier.

Adaptive Threshold Selection. From the pseudo-label bank, we can obtain a soft pseudo-label $\tilde{y}_i$ and its confidence μ_i for every target sample. To filter out noisy pseudo-labels, we adopt a per-class adaptive threshold. For each class c, we collect the confidences of all samples predicted as class c into a set μ^c and compute its median as the initial threshold $\hat{\tau}_c$. To prevent abrupt fluctuations, we smooth $\hat{\tau}_c$ via an EMA and clip the update to a maximum change Δ from the previous threshold, resulting in the current threshold $\tau_c^{(t)}$.

$$\tilde{\tau}_c = \alpha \cdot \tau_c^{(t-1)} + (1-\alpha) \cdot \hat{\tau}_c, \tag{5}$$

$$\tau_c^{(t)} = \varphi\big(\tilde{\tau}_c,\ \max(\tau_{\min},\ \tau_c^{(t-1)} - \Delta),\ \min(\tau_{\max},\ \tau_c^{(t-1)} + \Delta)\big), \tag{6}$$

where φ denotes the clipping function, α is the smoothing coefficient, Δ is a hyperparameter, and $[\tau_{\min}, \tau_{\max}]$ prevents excessively low or high extreme threshold values. Based on the updated $\tau_c^{(t)}$, we select the sample set A_c whose confidences satisfy $\mu_i \geq \tau_c^{(t)}$. The selected set for class c is $A_c = \{i \mid \mu_i \geq \tau_c^{(t)}\}$, which is then passed to the class-balanced weighting and pseudo-label loss.

Class-Balanced Weighting of Pseudo Samples. After completing the selection for all classes, to account for potential class imbalance, we compute a class-specific normalized distillation weight w_c for class c based on the dynamic sample counts. The weight is inversely proportional to the number of accepted samples $|A_c|$ for that class and is normalized by the total number of classes C:

$$w_c = \frac{C \cdot |A_c|^{-1}}{\sum_{j=1}^{C} |A_j|^{-1}}. \tag{7}$$

Let $A = \bigcup_{c=1}^{C} A_c$ be the set of all accepted target samples, and $c(i)$ be the pseudo-label class of sample i. Under this mechanism, hard classes with fewer samples receive higher weights in the pseudo-supervision loss. The student network then aligns its predictions to the teacher-generated soft labels using KL divergence:

$$\hat{y}_i^t = Softmax\big(g_s(F_{mask,i}^t)\big)\,, \tag{8}$$

$$\mathcal{L}_{pseudo} = \frac{1}{|A|} \sum_{i \in A} w_{c(i)} \cdot \mathrm{KL}\big(\tilde{y}_i^t \parallel \hat{y}_i^t\big), \tag{9}$$

where g_s is the student network's classifier.

The final total training loss is:

$$\mathcal{L} = \lambda_{ins}\mathcal{L}_{ins} + \lambda_{slide}\mathcal{L}_{slide} + \lambda_{pseudo}\mathcal{L}_{pseudo} + \mathcal{L}_{src}, \tag{10}$$

where $\mathcal{L}_{src}$ denotes the supervised cross-entropy loss on the source domain.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on four public datasets, including TCGA-NSCLC, CPTAC-NSCLC, TCGA-GBMLGG, and IPD-Brain, and compare it with state-of-the-art approaches. TCGA-NSCLC and CPTAC-NSCLC are used for binary classification of Lung Adenocarcinoma (LUAD) versus Lung Squamous Cell Carcinoma (LUSC), while TCGA-GBMLGG and IPD-Brain are used for five-class classification of glioma subtypes. These datasets differ in both task settings and data characteristics, and exhibit substantial domain shifts in staining styles and patient populations. This setting provides a challenging benchmark for assessing cross-domain generalization.

Implementation and Evaluation Details. We adopt the data preprocessing pipeline of CLAM [25]. WSIs are partitioned into patches of size 512×512. Features for each patch are extracted in advance using the CONCH [27] encoder, resulting in 512-dimensional representations. Our method and the natural image based baselines are implemented within the DTFD-MIL [50] framework. Both the patch feature aggregators used in the source and target branches and the classifiers are based on the DTFD architecture. We report ACC, F1-score, and AUC in the main experiments, while the ablation studies use AUC as the primary metric. The source domain is divided into training and validation sets with a ratio of 8:2. For the target domain, we perform 5 random splits. In each split, the target domain data is divided into training and test sets with a ratio of 5:5.

4.2 Main Results

We report the performance of our method and competing approaches across different datasets in Table 1. Among the baselines, we include commonly used pathology models, DTFD [50] and CLAM [25], which are trained on the source domain and directly evaluated on the target domain. We also compare with recent domain adaptation methods specifically designed for pathology images, namely DETA [39], RAM-MIL [9], and HASD [21]. However, given the limited number of domain

Table 1: Experimental results across four datasets. For each setting, one dataset is used as the source domain and another as the target domain. The best and second-best results are highlighted in red and blue, respectively. Methods with the suffix "-P" are applied at the patch-level embedding space, while other adapted methods operate on slide-level embeddings by default.

methods	TCGA → IPD (GBMLGG)			IPD → TCGA (GBMLGG)			CPTAC → TCGA (NSCLC)			TCGA → CPTAC (NSCLC)		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
DTFD [50]	73.74±0.58	53.66±0.64	25.18±3.18	80.48±2.41	43.00±2.04	29.07±1.61	57.46±1.30	50.22±2.03	66.85±1.80	87.43±2.93	54.24±1.44	65.69±1.22
CLAM [25]	80.85±0.74	61.52±1.21	32.43±1.40	82.17±1.93	46.48±2.94	41.91±2.78	41.41±1.92	50.22±2.03	66.84±1.79	89.43±3.11	46.06±2.14	62.76±1.89
MMD [22]	75.18±9.65	64.45±9.93	36.44±8.70	81.63±5.67	64.73±8.74	41.06±8.03	50.50±1.11	50.69±2.82	67.04±2.11	80.60±24.00	74.33±20.72	75.16±12.50
MMD-P	59.18±5.84	44.92±3.03	13.45±1.10	57.43±3.55	43.31±4.61	17.34±7.92	50.77±1.14	51.11±2.07	65.02±4.89	69.84±17.90	67.66±17.99	70.16±15.44
DANN [10]	80.68±2.18	65.09±3.83	42.05±4.29	81.45±1.54	63.31±2.61	44.27±1.04	51.41±3.16	50.22±2.03	66.84±1.79	97.30±2.78	94.52±2.47	94.17±2.57
DANN-P	82.12±2.61	66.61±1.58	44.53±3.08	82.34±1.15	65.60±2.75	46.66±2.50	69.83±11.29	58.38±11.43	70.82±6.40	96.56±2.99	92.70±4.42	92.12±4.84
CDAN [23]	79.90±1.84	66.96±1.52	44.13±1.69	81.61±2.47	65.99±1.92	44.88±1.78	57.14±7.87	50.22±2.03	66.84±1.79	86.36±26.07	85.09±23.69	88.34±15.96
MCD [35]	80.71±5.28	67.05±8.49	38.00±8.16	80.91±1.77	64.91±4.88	41.04±4.24	50.25±0.15	50.22±2.03	66.84±1.79	44.27±21.14	44.26±2.40	61.15±1.73
SRDC [42]	64.84±1.91	49.19±2.44	17.72±3.80	69.57±3.73	45.57±6.80	24.42±3.73	50.36±0.23	50.22±2.03	66.84±1.79	17.76±15.83	46.75±2.42	49.90±28.01
CORAL [41]	76.27±3.65	55.80±2.16	25.90±1.99	80.72±2.28	44.25±1.79	33.73±1.55	57.45±1.29	50.22±2.03	66.84±1.79	57.60±32.12	58.17±20.69	68.74±11.84
DRL$_{w/o\ label}$ [16]	62.00±14.33	45.81±10.60	15.90±6.35	72.41±6.58	39.15±16.10	20.17±13.11	55.43±3.38	50.22±2.03	66.84±1.79	71.10±31.83	65.53±22.86	74.68±13.31
DRL$_{w/i\ label}$	72.40±5.65	51.97±4.87	20.21±6.85	77.69±1.26	61.08±5.64	38.10±8.42	51.67±3.42	50.22±2.03	66.84±1.79	99.97±0.06	99.13±0.58	99.47±0.45
SimProF [46]	79.13±2.98	50.17±11.58	41.53±5.34	80.99±2.43	52.30±7.12	44.96±5.52	78.41±13.82	70.31±12.79	77.58±7.49	92.76±2.77	87.67±12.62	88.63±9.71
DETA [39]	65.73±3.90	46.97±0.60	13.26±1.10	63.85±9.71	40.62±15.88	16.46±9.74	44.46±5.66	50.26±2.09	66.86±1.82	38.13±7.66	47.52±1.57	64.41±1.38
RAM-MIL [9]	76.78±1.31	56.60±1.85	28.58±2.08	81.55±2.16	58.50±0.97	31.21±0.45	57.29±1.42	50.22±2.03	66.84±1.79	89.34±2.30	44.93±1.83	61.86±1.63
ours	83.89±1.42	67.50±0.56	46.83±2.82	83.45±0.68	67.09±2.31	47.63±1.76	87.39±6.01	77.71±6.74	75.97±9.84	97.83±1.10	94.71±2.27	94.39±2.41

Table 2: AUC (%, mean ± standard deviation) using CLAM as the shared MIL backbone.

Method	TCGA → IPD (GBMLGG)	IPD → TCGA (GBMLGG)	CPTAC → TCGA (NSCLC)	TCGA → CPTAC (NSCLC)
DANN	80.49 ± 4.92	80.77 ± 2.94	50.00 ± 0.00	95.70 ± 7.58
DANN-P	72.64 ± 4.26	75.64 ± 1.85	62.24 ± 5.01	35.34 ± 7.02
MCD	80.33 ± 3.05	81.24 ± 1.62	50.43 ± 0.44	85.40 ± 29.44
CORAL	77.57 ± 2.43	82.28 ± 1.67	62.42 ± 10.77	60.36 ± 24.37
RAM-MIL	80.27 ± 1.20	82.42 ± 1.97	37.42 ± 2.29	88.67 ± 2.94
Ours	84.83 ± 1.65	83.01 ± 2.52	84.11 ± 9.92	99.57 ± 0.33

adaptation methods in computational pathology, we further adapt a range of representative approaches from natural image analysis to the pathology setting, including MMD [22], DANN [10], CDAN [23], MCD [35], SRDC [42], CORAL [41], DRL [16], and SimProF [46]. All methods are compared comprehensively with our proposed approach.

The original DRL method relies on target domain labels for hyperparameter tuning, which are not available in the UDA setting. Therefore, we report DRL results both without and with access to target domain labels for a fair comparison. Our method achieves the best performance on most metrics across the four adaptation settings, even when compared with DRL tuned using target domain labels. On CPTAC → TCGA (NSCLC), our method improves AUC by 8.98 points over the second-best method. Meanwhile, we observe that many methods exhibit clear negative transfer compared with direct inference without adaptation. In particular, methods originally designed for natural image domain adaptation do not generalize well to pathological images. Notably, on the challenging CPTAC → TCGA setting, several compared methods collapse to predicting a single class, resulting in the same ACC and F1. We provide a detailed analysis of this phenomenon in Appendix H.

To assess DARE's plug-and-play applicability, we evaluate it and the compared adaptation methods using CLAM [25] as the shared backbone. DARE achieves the highest AUC across all four transfer directions (Table 2), supporting its applicability beyond DTFD-MIL to attention-based MIL backbones that provide patch-importance scores.

4.3 Analysis of Feature Extraction Backbones

In the main experiments, all methods utilize patch features extracted by the CONCH encoder. To evaluate whether the proposed DARE generalizes well to other patch feature extractors, we further conduct experiments using the foundation model UNI2-h [5] as the encoder for feature extraction. The results, summarized in Table 3, show that DARE consistently outperforms the compared methods under this feature extractor, demonstrating its robustness to the choice of the underlying encoder.

Table 3: Experimental results based on the UNI2-h encoder on two cross-domain settings.

methods	TCGA → IPD (GBMLGG)			IPD → TCGA (GBMLGG)		
	AUC	ACC	F1	AUC	ACC	F1
DTFD	$84.92_{\pm 2.59}$	$73.24_{\pm 1.49}$	$49.61_{\pm 2.53}$	$85.22_{\pm 1.40}$	$71.84_{\pm 0.96}$	$48.35_{\pm 0.99}$
DANN	$84.17_{\pm 3.36}$	$69.04_{\pm 2.10}$	$46.36_{\pm 2.42}$	$82.52_{\pm 2.51}$	$67.80_{\pm 5.93}$	$46.89_{\pm 2.54}$
DANN-P	$81.24_{\pm 5.44}$	$72.79_{\pm 5.90}$	$48.46_{\pm 4.86}$	$84.36_{\pm 2.00}$	$71.11_{\pm 1.95}$	$48.17_{\pm 2.57}$
CORAL	$83.65_{\pm 2.70}$	$74.22_{\pm 2.64}$	$48.48_{\pm 1.59}$	$85.23_{\pm 1.35}$	$71.47_{\pm 1.11}$	$48.23_{\pm 1.33}$
MCD	$86.37_{\pm 4.47}$	$69.04_{\pm 2.72}$	$49.68_{\pm 6.11}$	$83.51_{\pm 2.70}$	$70.11_{\pm 3.80}$	$47.56_{\pm 2.53}$
RAM-MIL	$85.29_{\pm 2.52}$	$72.97_{\pm 1.40}$	$48.64_{\pm 1.79}$	$84.28_{\pm 1.57}$	$71.87_{\pm 1.03}$	$46.28_{\pm 0.95}$
ours	$86.80_{\pm 2.94}$	$75.64_{\pm 2.41}$	$51.21_{\pm 2.46}$	$86.17_{\pm 1.89}$	$72.23_{\pm 2.85}$	$53.05_{\pm 3.99}$

Table 4: Ablation study results on two datasets.

method	TCGA → IPD	IPD → TCGA
w/o domain-aware	82.90	81.78
w/o $\mathcal{L}_{ins}$	72.49	80.09
w/o $\mathcal{L}_{slide}$	75.92	76.52
w/o $\mathcal{L}_{pseudo}$	83.58	79.74
ours	83.89	83.45

4.4 Ablation Studies

We perform ablation studies on the TCGA-GBMLGG and IPD-Brain datasets. As shown in Table 4, we denote each ablation setting as follows: "w/o domain-aware" removes domain probability guidance masking; "w/o $\mathcal{L}_{ins}$" removes patch-level domain adversarial learning; "w/o $\mathcal{L}_{slide}$" removes slide-level domain adversarial learning; and "w/o $\mathcal{L}_{pseudo}$" removes pseudo-label supervision on the target domain. The full model achieves the best performance. Both patch-level and slide-level domain adversarial modules provide significant gains, highlighting the importance of multi-granularity embedding alignment. The domain-aware masking also yields consistent improvement. Notably, on the IPD → TCGA task, pseudo-labeling leads to substantial performance gains, reinforcing its effectiveness in providing supervision signals for the target domain. Overall, each module contributes positively to the overall performance, validating the effectiveness of every component in our framework.

4.5 Robustness to Limited Target Domain Data

In real-world clinical scenarios, some hospitals and clinical centers may face difficulties in obtaining large-scale datasets for model adaptation. To evaluate the data efficiency of our method under such resource-limited conditions, we conduct experiments with limited target data by varying the proportion of target domain data used for adaptation. As shown in Table 5, the proposed method delivers remarkable gains even with extremely limited target domain data. With only 5% of the target domain data, both ACC and F1 improve substantially over the performance without any adaptation, confirming the strong data efficiency of our approach in adapting to new domains with minimal unlabeled target samples.

Table 5: Robustness to Limited Target Domain Data. Ratio: proportion of target data used. Eval: without adaptation.

ratio	TCGA → IPD (GBMLGG)			IPD → TCGA (GBMLGG)			CPTAC → TCGA (NSCLC)			TCGA → CPTAC (NSCLC)		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
eval	$73.74_{\pm 0.58}$	$53.66_{\pm 0.64}$	$25.18_{\pm 3.18}$	$80.48_{\pm 2.41}$	$43.00_{\pm 2.04}$	$29.07_{\pm 1.61}$	$57.46_{\pm 1.30}$	$50.22_{\pm 2.03}$	$66.85_{\pm 1.80}$	$87.43_{\pm 2.93}$	$54.24_{\pm 1.44}$	$65.69_{\pm 1.22}$
5%	$80.38_{\pm 4.08}$	$59.29_{\pm 8.84}$	$42.68_{\pm 7.73}$	$82.11_{\pm 1.81}$	$65.99_{\pm 1.91}$	$45.90_{\pm 2.25}$	$74.99_{\pm 13.45}$	$68.00_{\pm 11.92}$	$70.22_{\pm 10.23}$	$92.99_{\pm 3.74}$	$79.83_{\pm 16.18}$	$82.94_{\pm 11.12}$
10%	$83.53_{\pm 3.41}$	$65.10_{\pm 4.34}$	$44.64_{\pm 2.73}$	$80.74_{\pm 3.34}$	$64.39_{\pm 3.96}$	$44.73_{\pm 3.80}$	$81.32_{\pm 9.36}$	$74.90_{\pm 9.15}$	$77.25_{\pm 8.53}$	$93.72_{\pm 6.32}$	$85.15_{\pm 13.33}$	$86.67_{\pm 10.64}$
15%	$82.91_{\pm 2.14}$	$65.09_{\pm 3.79}$	$47.22_{\pm 3.96}$	$81.98_{\pm 2.59}$	$65.12_{\pm 2.43}$	$45.64_{\pm 2.42}$	$87.20_{\pm 7.19}$	$78.26_{\pm 8.65}$	$76.72_{\pm 9.76}$	$91.27_{\pm 9.11}$	$87.54_{\pm 13.70}$	$88.83_{\pm 10.15}$

4.6 Visualization of the Feature Alignment

To demonstrate the effectiveness of the dual-level feature alignment, we visualize the feature distributions of the source and target domains of different methods. Fig. 4 presents the slide-level embeddings of three methods: (a) the original DTFD backbone, (b) DANN (slide-level) with domain adversarial training, and (c) our proposed method. The embedding space of the original DTFD exhibits two distinct manifolds with compact intra-cluster structures. The two target subtypes are nearly indistinguishable in the embedding space. DANN reduces the domain gap across domains. However, the features of the two subtypes in the target domain still overlap in several regions, making them difficult to distinguish. In contrast, our method reduces domain discrepancy while producing well-separated clusters for target domain subtypes, leading to clearer decision boundaries and improved discrimination over the baselines.

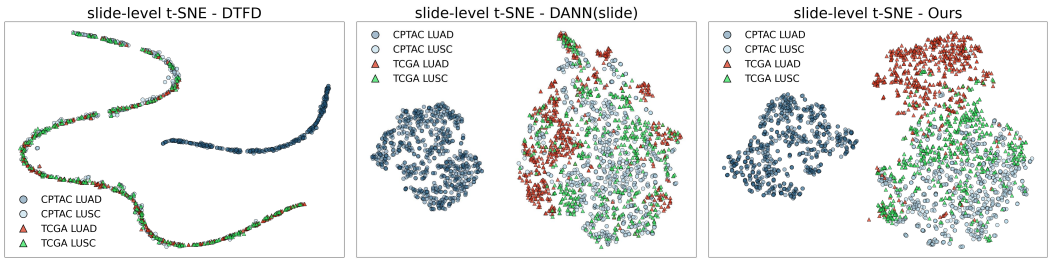


Figure 4: T-SNE visualization of source and target embedding distributions for the CPTAC → TCGA adaptation, comparing DTFD, DANN, and our method.

5 Conclusions

We propose DARE, a plug-and-play framework for unsupervised domain adaptation in computational pathology. The model addresses the hierarchical relationship between patches and slides through a dual-level domain adversarial mechanism for multi-granularity embedding alignment. We further introduce a domain-aware attention masking strategy that regularizes MIL attention and facilitates hard sample mining across domains. In addition, we design a teacher-student-based adaptive pseudo-labeling mechanism to provide supervision for target-domain data and enable class-wise alignment. We evaluate DARE on four datasets and demonstrate its effectiveness. Extensive comparisons with state-of-the-art methods show that DARE achieves superior performance in unsupervised domain adaptation for pathology images.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2026YFE0199900) and the National Natural Science Foundation of China (Grant No. 62471149). Additional support was provided by the CFFF Intelligent Computing Platform at Fudan University.

References

[1] Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM Van Der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2015.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1): 151–175, 2010.

[3] Sofiène Boutaj, Marin Scalbert, Pierre Marza, Florent Couzinie-Devy, Maria Vakalopoulou, and Stergios Christodoulidis. Controllable latent space augmentation for digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22165–22174, 2025.

[4] Aaron Chadha and Yiannis Andreopoulos. Improved techniques for adversarial discriminative domain adaptation. *IEEE Transactions on Image Processing*, 29:2622–2637, 2019.

[5] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[7] Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9101–9110, 2020.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[13] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

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

[15] Hassan Keshvarikhojasteh, Josien Pluim, and Mitko Veta. Multi-head attention-based deep multiple instance learning. *arXiv preprint arXiv:2404.05362*, 2024.

[16] 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`.

[17] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6092–6101, 2019.

[18] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International conference on machine learning*, pages 5468–5479. PMLR, 2020.

[19] Xintong Li, Chen Li, Md Mamunur Rahaman, Hongzan Sun, Xiaoqi Li, Jian Wu, Yudong Yao, and Marcin Grzegorzek. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review*, 55(6):4809–4878, 2022.

[20] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021.

[21] Jingsong Liu, Han Li, Chen Yang, Michael Deutges, Ario Sadafi, Xin You, Katharina Breininger, Nassir Navab, and Peter J Schüffler. Hasd: hierarchical adaption for pathology slide-level domain-shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 332–342. Springer, 2025.

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

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

[24] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

[25] 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.

[26] 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.

[27] 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*, 30:863–874, 2024.

[28] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.

[29] Faisal Mahmood, Daniel Borders, Richard J Chen, Gregory N McKay, Kevan J Salimian, Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging*, 39(11):3257–3267, 2019.

[30] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017.

[31] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):48, 2019.

[32] Linhao Qu, Siyu Liu, Xiaoyu Liu, Manning Wang, and Zhijian Song. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*, 67(20):20TR01, 2022.

[33] Jérôme Rony, Soufiane Belharbi, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *arXiv preprint arXiv:1909.03354*, 2019.

[34] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[35] 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*, pages 3723–3732, 2018.

[36] Massimo Salvi, U Rajendra Acharya, Filippo Molinari, and Kristen M Meiburger. The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128:104129, 2021.

[37] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pages 953–956. IEEE, 2019.

[38] Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5742–5749, 2020.

[39] 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*, pages 19925–19935, 2025.

[40] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021.

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

[42] 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*, pages 8725–8735, 2020.

[43] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[44] 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*, pages 7167–7176, 2017.

[45] Sophia J Wagner, Nadieh Khalili, Raghav Sharma, Melanie Boxberg, Carsten Marr, Walter De Back, and Tingying Peng. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 257–266. Springer, 2021.

[46] Meng-zhu Wang. Simprof: a simple probabilistic framework for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21153–21161, 2025.

[47] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern recognition*, 74:15–24, 2018.

[48] Hangchen Xiang, Junyi Shen, Qingguo Yan, Meilian Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis. *Medical Image Analysis*, 89:102890, 2023.

[49] Werner Zellinger, Bernhard A Moser, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Robust unsupervised domain adaptation for neural networks via moment alignment. *Information Sciences*, 483:174–191, 2019.

[50] 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*, pages 18802–18812, 2022.

[51] Yu-Xuan Zhang, Zhengchun Zhou, Weisha Liu, and Mingxing Zhang. Rethinking multi-instance learning through graph-driven fusion: A dual-path approach to adaptive representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28510–28518, 2026.

[52] Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European conference on computer vision*, pages 125–143. Springer, 2024.

[53] Junjie Zhou, Wei Shao, Yagao Yue, Wei Mu, Peng Wan, Qi Zhu, and Daoqiang Zhang. Maple: Multi-scale attribute-enhanced prompt learning for few-shot whole slide image classification. *arXiv preprint arXiv:2509.25863*, 2025.

[54] Wenhui Zhu, Xiwen Chen, Peijie Qiu, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *European conference on computer vision*, pages 333–351. Springer, 2024.

A Dataset Description

We conduct experiments on four datasets: TCGA-NSCLC, CPTAC-NSCLC, TCGA-GBMLGG, and IPD-Brain.

TCGA-NSCLC dataset contains 1,040 WSIs for binary classification of Lung Adenocarcinoma (LUAD) versus Lung Squamous Cell Carcinoma (LUSC).

CPTAC-NSCLC dataset contains 1,342 WSIs and is also used for LUAD versus LUSC classification. Compared with TCGA-NSCLC, it shows a clear shift in staining appearance and a noticeable domain gap in visual distribution.

TCGA-GBMLGG dataset contains 1,526 WSIs, integrating both lower-grade glioma and glioblastoma cases. Based on IDH1 mutation status and histopathological subtypes, samples are categorized into five molecular subtypes: IDH1-mutant astrocytoma, IDH1-mutant oligodendroglioma, IDH1-mutant glioblastoma, IDH1-wildtype astrocytoma, and IDH1-wildtype glioblastoma. In our experiments, it is used for a five-class classification task. Due to large morphological variability across subtypes and heterogeneous data sources, this dataset exhibits strong intrinsic distributional heterogeneity.

IPD-Brain dataset contains 448 WSIs and is collected from an Indian glioma cohort. It is used for the same five-class classification task. Compared with TCGA-GBMLGG, it shows a clear distribution shift in patient population. This cross-region variation further increases the difficulty of feature transfer and decision boundary alignment.

B Implementation Details

For our DARE, we follow the standard unsupervised domain adaptation protocol: we first pre-train a source model under the DTFD framework on the labeled source domain, and then load this pre-trained model as initialization for domain adaptation on the combined labeled source and unlabeled target data.

Specifically, we use the Adam optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. For the dual-level domain adversarial module, the loss weights default to $\lambda_{slide} = 0.5$ and $\lambda_{ins} = 0.7$. For the domain-aware attention masking module, the masking ratio r defaults to 0.1. For the teacher-student pseudo-labeling module, the exponential moving average smoothing coefficient for the adaptive threshold defaults to $\alpha = 0.8$, with a maximum allowed change $\Delta = 0.05$ and a clipping range $[\tau_{\min}, \tau_{\max}] = [0.7, 0.95]$. The teacher model parameters θ_t are updated via $\theta_t \leftarrow \alpha_t \theta_t + (1 - \alpha_t)\theta_s$, where θ_s denotes the student parameters and $\alpha_t = 0.99$ by default. The pseudo-label loss weight defaults to $\lambda_{pseudo} = 0.5$. To avoid division by zero when computing the class-specific normalized distillation weights w_c, we set a minimum of 4 accepted samples per class ($|A_c| \geq 4$). Pseudo-label learning is activated at the 40th epoch as a warm-up phase to mitigate confirmation bias from unreliable early-stage predictions. All experiments are conducted on an NVIDIA A100 Tensor Core GPU.

C Hyperparameter Analysis

We conduct hyperparameter experiments to explore the impact of different values for the masking ratio r on the CPTAC-NSCLC and TCGA-NSCLC datasets. As shown in Fig. 5, on the harder CPTAC $\rightarrow$ TCGA task, all three metrics improve substantially as r increases from 0.02 to 0.18, confirming that suppressing domain-specific easy patches effectively regularizes attention and facilitates hard sample mining. The peak AUC is reached at $r = 0.18$. When r continues to increase beyond 0.18, performance gradually declines (AUC drops from 0.8992 to 0.8844 at $r = 0.22$ and further to 0.8683 at $r = 0.26$), suggesting that over-aggressive masking begins to discard useful discriminative patches. In contrast, on the easier TCGA $\rightarrow$ CPTAC task, the model is robust to r and maintains high performance across the entire tested range from 0.02 to 0.26, indicating that the domain-aware masking does not hurt transfer when the domain gap is small. We use $r = 0.1$ as the default setting.

We further examine the sensitivity of the pseudo-label learning module to $\tau_{\min}$, $\tau_{\max}$, α, and Δ on the two GBMLGG transfer tasks. We vary one parameter at a time while keeping the remaining

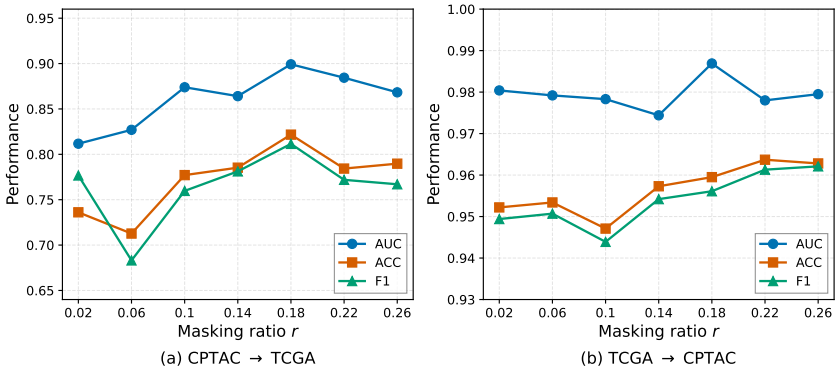


Figure 5: Sensitivity analysis of the masking ratio r on two domain adaptation tasks. (a) CPTAC $\rightarrow$ TCGA. (b) TCGA $\rightarrow$ CPTAC. Performance is measured by ACC, F1-score, and AUC.

Table 6: Sensitivity of the pseudo-label learning module to $\tau_{\min}$, $\tau_{\max}$, α, and Δ. Results are mean AUC (%). One parameter is varied at a time; default values are indicated in parentheses. The best result within each parameter group and transfer direction is shown in red.

Parameter	Value	TCGA-GBMLGG $\rightarrow$ IPD-Brain	IPD-Brain $\rightarrow$ TCGA-GBMLGG
$\tau_{\min}$	0.60	82.34	82.11
	0.64	84.72	82.81
	0.67	83.47	82.69
	0.70 (default)	83.89	83.45
	0.74	84.21	83.16
	0.77	84.26	82.45
	0.80	85.45	81.99
$\tau_{\max}$	0.80	81.35	81.39
	0.90	84.12	80.36
	0.93	83.89	83.12
	0.95 (default)	83.89	83.45
	0.99	83.89	83.02
α	0.75	84.33	82.72
	0.80 (default)	83.89	83.45
	0.85	83.06	82.27
	0.90	85.28	82.36
Δ	0.00	82.11	81.41
	0.05 (default)	83.89	83.45
	0.10	81.97	83.06

settings fixed. Table 6 reports the mean AUC over multiple random seeds, following the protocol used in the main experiments.

For $\tau_{\min}$, both transfer directions maintain competitive performance over a range of values. The default value of 0.70 achieves the highest AUC on IPD-Brain $\rightarrow$ TCGA-GBMLGG, whereas 0.80 performs best in the reverse direction. The results for $\tau_{\max}$ are stable between 0.93 and 0.99 on TCGA-GBMLGG $\rightarrow$ IPD-Brain; the default value of 0.95 performs best in the reverse direction. For α, different values favor the two transfer directions, and the default value of 0.80 provides a balanced choice. Setting Δ to 0.05 yields the highest AUC in both directions, supporting the use of adaptive threshold updates. Overall, the default configuration performs competitively across both transfer directions without tuning parameters separately for each task.

Table 7: Ablation study of pseudo-label learning on the two GBMLGG transfer tasks. Results are AUC (%, mean ± standard deviation). The best result in each column is shown in red.

Method	TCGA-GBMLGG → IPD-Brain	IPD-Brain → TCGA-GBMLGG
DARE (full)	$83.89_{\pm 1.42}$	$83.45_{\pm 0.68}$
w/o adaptive threshold selection	$82.11_{\pm 4.25}$	$81.41_{\pm 3.76}$
w/o class-balanced weighting	$79.07_{\pm 5.33}$	$78.84_{\pm 1.93}$
Cross-entropy with one-hot teacher labels	$82.71_{\pm 3.97}$	$82.58_{\pm 2.02}$

D Pseudo-label Learning Ablation

We assess three design choices in the pseudo-label learning module on the five-class TCGA-GBMLGG and IPD-Brain datasets: adaptive threshold selection, class-balanced weighting, and the KL-divergence-based pseudo-label loss. For the loss comparison, we replace the KL-divergence objective with cross-entropy against a one-hot label predicted by the teacher. All other settings are kept the same as in the full DARE model.

As shown in Table 7, removing adaptive threshold selection decreases AUC in both transfer directions. Removing class-balanced weighting causes a larger decrease, indicating its value for learning from target-domain pseudo-labels. Replacing the KL-divergence loss with cross-entropy against a one-hot teacher label also reduces AUC in both directions. The full model achieves the highest AUC across these comparisons.

E Visualization of Attention and Domain-Aware Masking

E.1 Visualization of Attention Redistribution

To examine how DARE affects the spatial distribution of attention, Fig. 6 presents representative target-domain slides for glioma and lung cancer subtyping. Each example includes the original H&E image and attention maps before and after adaptation. Colors indicate attention percentiles, reflecting the relative ranking of patch attention.

For glioma subtyping (Fig. 6a,b), DARE redistributes attention toward more cellular regions. In Fig. 6a, the upper-central tissue fragment receives lower relative attention after adaptation, whereas regions including the lower-left fragment receive greater attention. Local H&E inspection indicates that the former region is relatively sparsely cellular and contains neuron-like structures, while the latter regions exhibit more densely packed small, round nuclei. Similarly, in Fig. 6b, relative attention increases in the lower-left region containing dense small, round nuclei and decreases in portions of the right-sided tissue with more prominent neuron-like cells and lower cellularity. These patterns suggest reduced emphasis on some background brain parenchymal features and increased attention to potentially tumor-associated cellular regions.

For lung cancer subtyping (Fig. 6c,d), attention shifts toward regions exhibiting adenocarcinoma-associated morphology. In Fig. 6c, high-ranking attention shifts from portions of the peripheral tissue with preserved alveolar architecture toward internal regions containing irregular gland-like structures. In Fig. 6d, the upper fibrous region containing dark granular deposits receives lower relative attention, while the lower-left region with irregular gland-like structures receives greater attention.

Overall, these examples suggest that DARE redistributes relative attention toward potentially task-relevant morphology while reducing emphasis on some background tissue components.

E.2 Visualization of Domain-Aware Masking

To examine the behavior of domain-aware attention masking, Fig. 7 presents representative target-domain slides together with masked and retained patches of comparable attention. This comparison highlights how source-domain affinity contributes to patch selection beyond attention alone.

In Fig. 7a, both the masked and retained patches contain epithelial-like cell groups interspersed with stromal tissue. The masked patch exhibits relatively compact cellular groups adjacent to pale stromal and vascular-like spaces, whereas the retained patch contains epithelial-like groups separated by more conspicuous fibrous stroma. Despite these local differences, both patches contain substantial

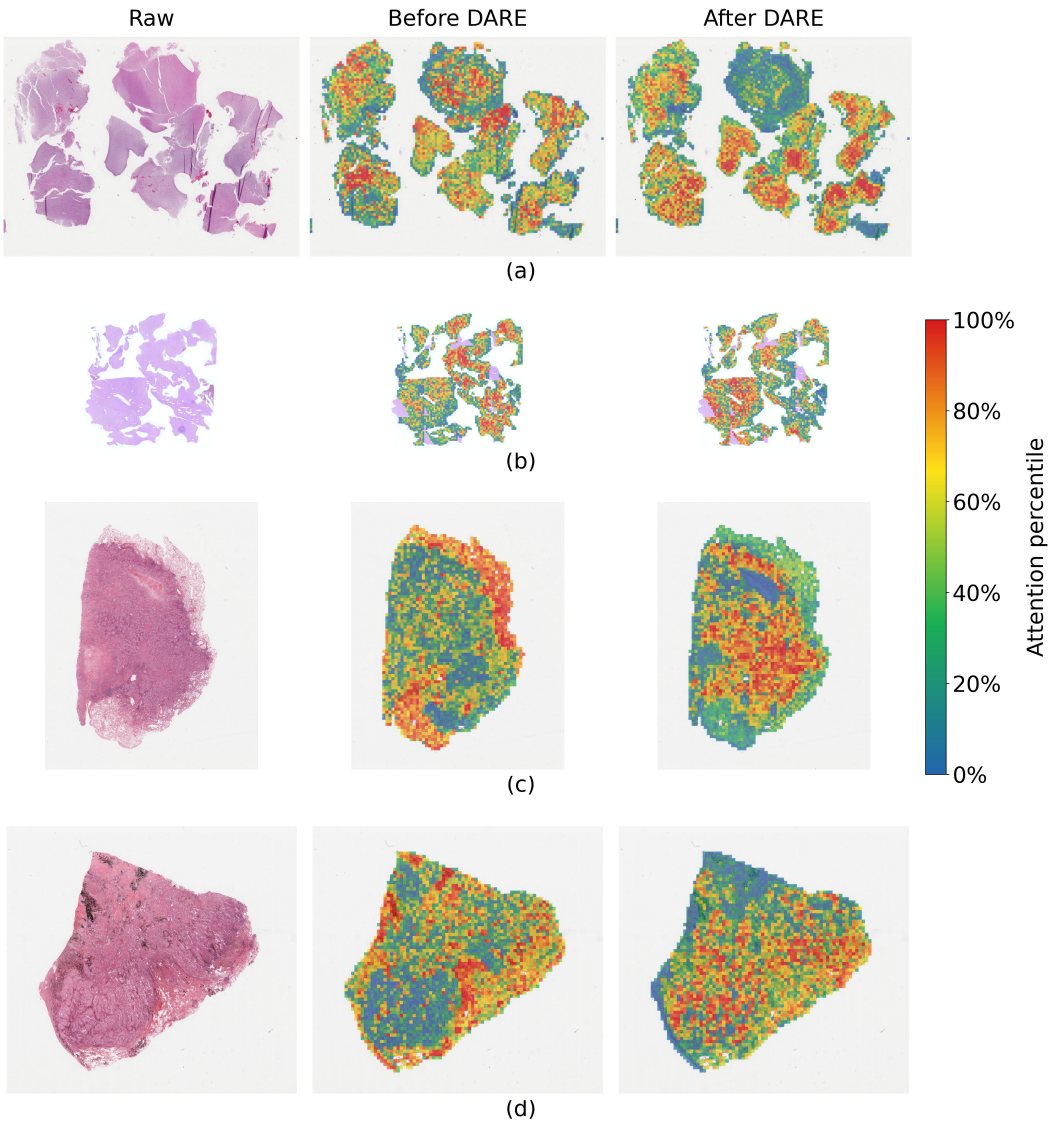


Figure 6: Attention maps before and after DARE for (a,b) glioma subtyping and (c,d) lung cancer subtyping.

cellular and architectural information. Their different masking outcomes illustrate that DARE does not simply remove cellular regions or exclude an entire tissue pattern. Instead, it selectively masks highly attended source-like regions while preserving other highly attended regions with overlapping morphological features.

A more direct morphological comparison is shown in Fig. 7b. Both patches exhibit gland-like epithelial structures with pale or vacuolated cytoplasm. The masked patch contains more closely arranged epithelial structures, while the retained patch shows more open luminal spaces bordered by epithelial cells. Importantly, recognizable glandular morphology remains present in the retained patch. This example illustrates how selective masking can reduce reliance on a subset of highly attended regions without eliminating the broader morphological pattern available to the slide-level classifier.

Together, these examples highlight the complementary roles of attention and domain information: attention identifies regions emphasized by the classifier, while source-domain affinity guides which of these regions are masked. By selectively withholding highly attended source-like patches, DARE is designed to discourage reliance on source-associated appearance cues, which may include staining characteristics. Meanwhile, the retained regions provide alternative morphological evidence for slide-level aggregation. Within the teacher-student framework, this masked view encourages the student to maintain predictions consistent with the complete-slide teacher while drawing on the remaining tissue evidence.

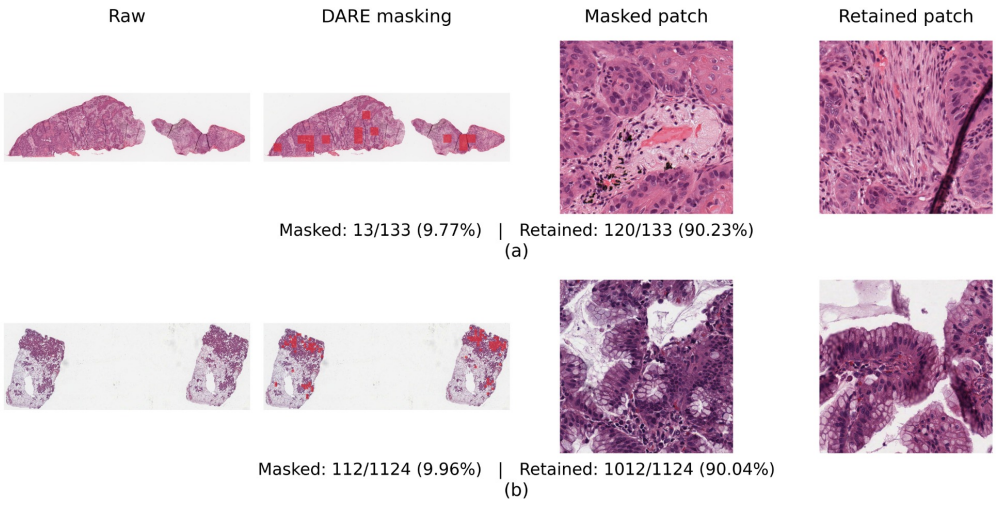


Figure 7: Domain-aware masking in lung cancer subtyping. Columns show the original H&E images, masking maps, and attention-matched masked and retained patches. Red overlays indicate masked regions.

F Domain Discrepancy and Adaptation Performance

We investigate how the degree of domain discrepancy relates to the performance gains achieved by DARE. In addition to the cross-cohort experiments reported in the main text, we conduct supplementary cross-center experiments on CAMELYON17, which comprises data from five hospital centers. These experiments extend our evaluation to a multi-center setting and allow us to examine adaptation performance under relatively mild domain shifts.

To quantify domain discrepancy, we adopt the Fréchet-distance formulation underlying the Fréchet Inception Distance (FID) [12], but compute it using CONCH embeddings instead of Inception features. We refer to this metric as the Fréchet Feature Distance (FFD). This choice allows us to measure domain mismatch in the representation space used by our downstream pathology models. Let $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_t$ denote the empirical means of the source- and target-domain embeddings, and let $\boldsymbol{\Sigma}_s$

Table 8: Domain discrepancy and adaptation performance. FFD is computed using CONCH embeddings. AUCs are averaged over the evaluated transfer directions within each setting. Gain denotes the AUC improvement over the source-only baseline in percentage points (pp).

Adaptation setting	FFD	Source-only AUC (%)	DARE AUC (%)	Gain (pp)
CAMELYON17 center pairs	0.0796–0.2469 (mean: 0.1591)	96.10	96.14	+0.04
TCGA-GBMLGG ↔ IPD-Brain	0.3708	77.11	83.67	+6.56
TCGA-NSCLC ↔ CPTAC-NSCLC	0.7133	72.45	92.61	+20.16

and $\boldsymbol{\Sigma}_t$ denote their covariance matrices. FFD is defined as

$$\mathrm{FFD}(s,t) = \|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_t - 2\left(\boldsymbol{\Sigma}_s^{1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_s^{1/2}\right)^{1/2}\right). \tag{11}$$

A larger FFD indicates a greater discrepancy between the two feature distributions under this Gaussian approximation.

Table 8 summarizes FFD and the corresponding average AUCs of the source-only baseline and DARE. Across CAMELYON17 center pairs, FFD ranges from 0.0796 to 0.2469, with a mean of 0.1591. Under these relatively mild shifts, the source-only baseline already achieves an average AUC of 96.10%, and DARE maintains comparable performance at 96.14%. The strong baseline performance leaves limited room for further improvement.

Larger gains are observed in the cross-cohort settings with greater feature-space discrepancy. For TCGA-GBMLGG and IPD-Brain, an FFD of 0.3708 is accompanied by an average AUC improvement of 6.56 percentage points. For TCGA-NSCLC and CPTAC-NSCLC, the FFD reaches 0.7133, and DARE improves the average AUC by 20.16 percentage points. Together, these results show that DARE maintains strong average performance under mild center shifts and provides larger gains in the evaluated settings with more pronounced domain mismatch.

FFD measures overall feature-distribution mismatch, whereas adaptation performance also depends on how the shift affects task-relevant features and class boundaries. Thus, the observed relationship does not establish a universal discrepancy threshold for effective adaptation. The CAMELYON17 FFD range of 0.0796–0.2469 represents the smallest-discrepancy regime covered by our experiments.

G Computational Efficiency Analysis

Table 9: Training time comparison.

Method	Total time (s)
DANN-P	1183.66
DANN	1202.09
DETA	6705.07
RAM-MIL	2084.88
Ours	1721.52

The training time comparison is presented in Table 9. Our method achieves a training time of 1721.52 seconds, which is significantly lower than that of RAM-MIL (2084.88 s) and DETA (6705.07 s), two representative UDA approaches designed for pathology. Although our method is slower than the DANN-based baselines, those natural-image domain adaptation methods are less effective on histopathological data, as shown in the main experiments (Table 1). Therefore, our method strikes a better balance between computational efficiency and domain adaptation performance in the pathology setting.

H Analysis of Mode Collapse

As noted in the main results (Table 1), several compared methods produce identical accuracy (50.22%) and F1-score (66.84%) on the CPTAC $\rightarrow$ TCGA (NSCLC) task. Mathematically, this specific metric combination indicates complete mode collapse: the adapted model loses all discriminative ability and degenerates into a trivial solution that predicts all target samples as the majority class.

We attribute this mode collapse to a deep inconsistency between prevailing alignment strategies and the distinct nature of whole slide image data. In WSIs, genuinely meaningful morphological signals, such as neoplastic patterns, are extremely sparse, whereas the representational space is crowded with domain-specific nuisance factors arising from differences in staining, scanner color response, and tissue handling. For generic unsupervised domain adaptation methods that rely on adversarial learning (e.g., DANN, CDAN), forcing a minimax game directly on such a heavily noisy latent space pushes the model to take the simplest route to deceive the domain classifier. Consequently, the model tends to shed not only domain-specific variations but also fragile diagnostic cues, driving the target representations into a tangled, semantically indiscernible state.

Notably, this vulnerability is not limited to generic methods but also exists in recently proposed pathology-specific UDA methods such as RAM-MIL and DETA. When RAM-MIL performs nearest-neighbor retrieval via optimal transport, the retrieved neighbors tend to be domain-similar rather than semantically similar instances. When DETA constructs graph structures in the feature space, it may establish spurious edges between patches that differ morphologically but share similar staining styles. These mechanisms, operating on input features already contaminated by strong domain bias, inadvertently amplify domain information and push the model further along the biased direction, eventually leading to the same degenerate solutions observed in other methods.

Our method avoids this pitfall. As shown in Fig. 4, we visualize the slide-level embeddings of three approaches: (a) the original DTFD backbone evaluated directly on the target domain, (b) DANN with slide-level adversarial training, and (c) our DARE. For DTFD, the two subtypes of target domain are completely inseparable. For DANN, a vague boundary remains poorly defined. In contrast, DARE produces a well-separated embedding space with a clear decision boundary between the two subtypes. This improvement stems from three synergistic designs. First, multi-granularity alignment jointly matches patch-level and slide-level feature distributions across domains, preventing the model from exploiting easy shortcuts that erase semantic information. Second, attention regularization encourages the model to focus on diagnostically relevant regions rather than domain-specific stains or artifacts. Third, the teacher-student pseudo-labeling module provides stable category-level supervision on the target domain, which guides the embedding space to preserve subtype discriminability even when domain shifts are severe. Together, these components enable DARE to recover meaningful structure without collapsing into a semantically indistinguishable state.

I Limitations and Future Work

Limitations. First, although our approach is more efficient than existing pathology-specific UDA methods, its training speed is slower than methods that perform adversarial alignment at only one level. As shown in Table 9, the dual-level adversarial design and the teacher-student pseudo-label learning framework introduce extra computational overhead compared to single-level adaptation methods. Second, similar to other MIL-based WSI classification models, the performance of our method is influenced by the choice of the feature encoder. As reported in Table 1 and Table 3, when switching from CONCH to the UNI2-h encoder, both our method and the compared baselines exhibit performance changes, with most methods achieving higher absolute scores under UNI2-h. While our method consistently achieves the best results across different encoders, the upper bound of overall performance is still tied to the quality of the underlying feature extractor.

Future Work. A practical concern in real-world deployment is data privacy: source domain data may not always be accessible due to institutional regulations. In future work, we plan to extend our framework to source-free domain adaptation for WSIs, where only a pre-trained source model and unlabeled target data are available. This would help alleviate privacy concerns in clinical scenarios where direct sharing of source domain slides is restricted.

J Broader Impact

Our work aims to improve the robustness of whole slide image classification under domain shifts commonly encountered across different medical centers. This line of research may help make pathology AI more accessible to hospitals that lack large annotated datasets, with the long-term potential to support more consistent diagnostic services. We acknowledge that any AI-assisted diagnostic tool carries risks if deployed without proper validation. Our method, like other data-driven approaches, may still be susceptible to biases in the training data. We therefore advocate for rigorous external validation and close collaboration with clinical experts before real-world deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and scope of this paper are claimed in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not contain theoretical studies.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide experimental details in Section 4 of the main paper and Section B of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and documentation will be publicly released upon paper publication. Public datasets are openly accessible.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental details in Section 4 of the main paper and Section B of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All reported metrics are averaged over multiple random seeds, and standard deviations are provided in Table 1 and Table 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide experimental details in Section B and Section G of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This study complies strictly with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss this in the Appendix J.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite all existing papers, models, and datasets used in this work, and respect their terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.