arXiv:2607.14703v1 [cs.CV] 16 Jul 2026

Pretraining Multiple Instance Learning Networks with Multi-Teacher Distillation from Pathology Slide Foundation Models

Mingxi Fu[1,*] **Jiawen Li**[1,*] **Renao Yan**[2] **Jiali Hu**[3]
Qiehe Sun[4] **Tian Guan**[1,†] **Yonghong He**[1,4,5,†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]University of Washington [3]City University of Hong Kong (Dongguan)
[4]Medical Optical Technology R&D Center, Research Institute of Tsinghua
[5]Jinfeng Laboratory
guantian@sz.tsinghua.edu.cn heyh@sz.tsinghua.edu.cn
[*]Equal contribution [†]Corresponding authors

Abstract

Multiple instance learning (MIL) has become the main paradigm for whole-slide image (WSI) analysis in computational pathology. However, existing MIL aggregators are still typically trained from scratch for each downstream task, relying on limited slide-level labels to learn both aggregation mechanisms and downstream discriminative representations simultaneously. As a result, they often suffer from unstable optimization, overfitting, and limited transferability. Similar to pretrained ResNet and Vision Transformer models in natural image learning, MIL also requires reusable pretrained initialization. However, high-quality slide-level pretraining data remain scarce, and MIL models are usually lightweight and weakly supervised, making large-scale pretraining difficult in practice. To address this challenge, we propose a distillation-based pretraining framework for MIL, which leverages two slide-level foundation models, TITAN and CARE, as teachers to transfer their representational knowledge into a diverse set of MIL architectures. To effectively balance supervision from different teachers, we further introduce an angular dispersion normalized distillation loss. The distilled weights are then used as initialization for downstream adaptation. We conduct systematic evaluations on 15 benchmark datasets under both linear probing and full-parameter fine-tuning, and further validate its advantages in few-shot scenarios. Experimental results show that pretraining generally improves MIL aggregators over from scratch training, especially in linear-probing and few-shot settings, while maintaining the computational efficiency of lightweight MIL models. Code is available at `https://github.com/fu0201/MIL_Pretrained`.

1 Introduction

Multiple instance learning (MIL) has been widely adopted as the mainstream framework for whole-slide image (WSI) analysis in computational pathology. Under this framework, patch features produced by a frozen tile encoder are treated as instances, and the entire slide is modeled as a bag whose label supervises learning [50]. Driven by this formulation, a rich spectrum of MIL aggregators has emerged, including attention-based pooling [17, 30], Transformer-based architectures [42], graph neural networks [26, 11], and state-space models [52]. These methods have continuously advanced instance aggregation and slide-level generalization. Despite this architectural diversity, a fundamental limitation persists: existing MIL aggregators are invariably trained from scratch on

Preprint.

each downstream task, relying solely on a handful of slide-level labels to simultaneously learn both discriminative instance representations and an effective aggregation strategy. Under such weak supervision and limited data, aggregators frequently overfit to spurious correlations and fail to acquire stable, transferable slide representations [50, 41].

A natural solution is to pretrain MIL aggregators before adapting them to downstream WSI tasks. In modern computer vision, pretraining followed by downstream fine-tuning has become the standard paradigm for visual representation learning. Representative backbones such as ResNet [15] and Vision Transformer (ViT) [10] are typically pretrained on large-scale image corpora and then adapted to downstream tasks through fine-tuning or linear-probing as shown in Figure 1. Studies have shown that such pretrained initialization substantially improves transferability, convergence speed, and data efficiency, thereby providing a strong and reusable starting point for a wide range of vision tasks [24, 23]. This paradigm has increasingly reshaped computational pathology as well. At the tile level, self-supervised and vision–language foundation models, such as UNI [7], Virchow [43], CONCH [29], and MUSK [47], have dramatically improved patch-level feature quality through pretraining on large-scale histopathology images. More recently, at the slide level, models such as TITAN [9], Prov-GigaPath [49], and CARE [51] have further demonstrated that pretraining entire slide encoders on large-scale high-resolution pathology images can yield powerful and general-purpose slide representations. Beyond foundation models themselves, Shao et al. [41] also provided direct evidence that pretrained MIL aggregators consistently outperform their randomly initialized counterparts in downstream transfer.

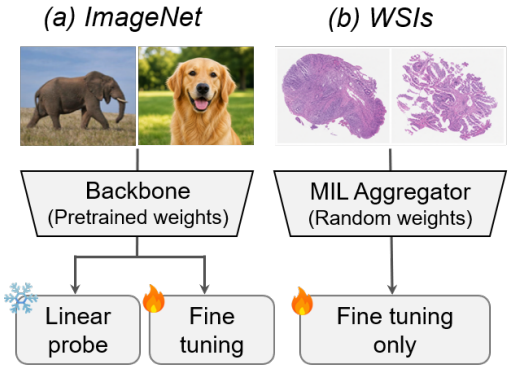

Figure 1: Pretrained backbones support freezing or fine-tuning in natural image analysis, whereas MIL models for WSI are typically trained from scratch.

However, directly pretraining MIL aggregators still faces several challenges. First, large-scale slide-level datasets remain intrinsically scarce. Unlike low-resolution tile images, which can be densely cropped from WSIs and thereby scaled to hundreds of millions or even billions of samples, publicly available WSI cohorts are typically limited to thousands or tens of thousands of slides, far below the data scale required for robust foundation-model pretraining [9, 49]. Second, conventional MIL aggregators are usually designed as lightweight task-specific modules rather than high-capacity pretraining targets. Most MIL aggregators contain only a few million parameters and are mainly supervised by bag-level labels, providing insufficient optimization signals for learning generalizable slide-level representations [19]. Recent studies have attempted to mitigate this issue by scaling up slide encoders and pretraining them on large WSI cohorts [40]. However, in digital pathology, where downstream tasks are highly heterogeneous and annotated data are often limited, directly fine-tuning large slide encoders can be computationally expensive and optimization-unstable, thereby restricting their practical applicability across diverse clinical tasks.

To make MIL pretraining feasible, in this work we propose a multi-teacher distillation framework for pretraining MIL networks from pathology slide foundation models. Specifically, we employ TITAN [9] and CARE [51], two state-of-the-art pretrained slide encoders as complementary expert teachers. Guided by a angular dispersion normalized distillation loss, we distill their slide-level representational knowledge into 9 MIL architectures, covering attention-based, Transformer-based, graph-based, and state-space designs. To validate the effectiveness of the pretrained MIL aggregators, we initialize downstream fine-tuning with the distillation-pretrained weights. Across six benchmark datasets and more than fifteen clinical tasks, including cancer subtyping, biomarker prediction, and mutation detection, we compare this paradigm with from-scratch MIL training and direct teacher-model fine-tuning. Our results demonstrate that, compared with from-scratch training, distillation-based pretraining enhances MIL aggregators across fine-tuning, linear-probing, and few-shot evaluation settings, and further enables these lightweight models to outperform their teacher models in many scenarios.

2 Related Work

2.1 Multiple Instance Learning for Weakly Supervised WSI Analysis

Multiple instance learning (MIL) has long been the dominant paradigm for weakly supervised WSI analysis. In this framework, pretrained patch encoders are used to extract local visual features, while an aggregator further integrates instance-level information to generate slide-level predictions [12, 3]. Early MIL methods mainly focused on permutation-invariant aggregation designs, such as the attention-based pooling in ABMIL [17] and the clustering-constrained attention mechanism in CLAM [30]. To capture dependencies among instances, TransMIL [42] introduced self-attention together with pyramidal positional encoding, whereas graph-based methods such as WiKG [26] and DAGMIL [11] construct spatial or feature graphs to propagate contextual information across patches. More recently, state space models have also been introduced into MIL aggregation, leveraging their linear-complexity sequence modeling capability to efficiently handle the thousands of instances typically encountered in WSIs [52]. Nevertheless, most existing MIL methods still treat the aggregator as a task-specific module that must be trained from scratch on each downstream dataset using limited slide-level labels, and its cross-task transferability remains largely underexplored.

2.2 Pretraining from patch-level to slide-level in computational pathology

With the rapid development of self-supervised learning, pathology image pretraining has achieved substantial progress at the patch level. Methods such as SimCLR [8], MoCo [14], and DINO [6] have been applied to millions of tiles cropped from hundreds of thousands of WSIs, leading to a series of pretrained models with different scales and capabilities [20, 25, 44]. Vision Transformer backbones trained under the DINOv2 [34] framework, such as UNI [7] and Virchow [43], have become widely adopted image encoders in digital pathology. Meanwhile, works such as CONCH [29] and MUSK [47] further explored vision-language pretraining, offering pathology models with stronger cross-modal capabilities. Building upon these powerful pathology-specific encoders, recent studies have further investigated slide-level pretraining to learn general-purpose WSI representations. For example, CHIEF [45] combines tile-level representation learning with weakly supervised slide-level pretraining to capture generalizable WSI patterns. MADELEINE [18] introduces a multi-stain pretraining strategy, where matched slides from different staining modalities are aligned through global-local cross-stain objectives. TITAN [9] further extends slide-level pretraining to the multimodal setting by aligning WSI representations with pathology reports and synthetic captions. CARE [51] incorporates molecular guidance by first learning morphological representations from WSIs and then refining adaptive region representations through RNA and protein alignment. However, MIL aggregators are typically lightweight, which limits their capacity to benefit from conventional supervised pretraining in the same way as large-scale encoders. Recent work such as FEATHER [41] has attempted to pretrain MIL models with supervised slide-level classification labels, but this paradigm requires large-scale labeled datasets, making MIL pretraining costly and difficult to scale to broader pathology tasks.

2.3 Knowledge Distillation and Multi-teacher Knowlodege distillation

Knowledge distillation provides an effective mechanism for transferring knowledge from well-trained teacher models to compact student models, commonly through soft-logit supervision or feature-level representation alignment [16, 46, 2, 4, 32]. Multi-teacher distillation further extends this paradigm by allowing a student model to inherit complementary knowledge from multiple heterogeneous teachers [28, 35], which is particularly useful for foundation models with diverse architectures and representation spaces. Recent vision foundation models have validated the effectiveness of this strategy. AM-RADIO [37] distills multiple powerful vision models, including CLIP [36], DINOv2 [34], and SAM [22], into a unified student representation, while C-RADIOv4 [38] further improves representation compatibility and transferability. In computational pathology, GPFM [31] similarly distills multiple expert encoders into a unified pathology foundation model to improve generalization across downstream tasks. Despite these advances, existing pathology distillation studies mainly focus on low-resolution images encoders, while high-resolution whole-slide representation learning remain largely underexplored.

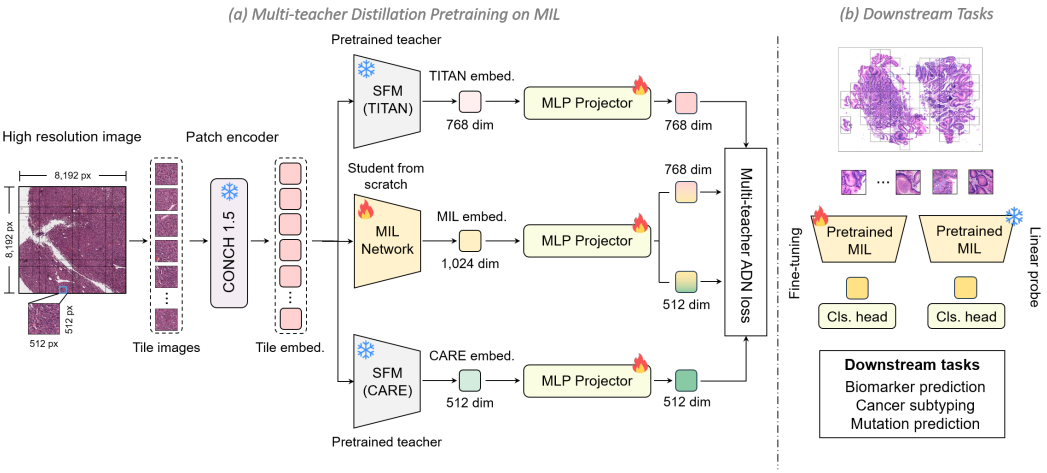

Figure 2: Overview of the proposed framework. Left: ROI-level pretraining of the MIL network via multi-teacher distillation from TITAN and CARE. Right: Transfer of the pretrained MIL aggregator to downstream WSI-level evaluation tasks.

3 Method

We propose a multi-teacher distillation pretraining method on MIL, a general framework for pretraining lightweight MIL aggregators through slide-level representation transfer. Figure 2 provides an overview of our method. We first introduce the multi-teacher distillation pretraining procedure, where MIL aggregators are pretrained on high-resolution pathology regions of interest (ROIs) under the guidance of pretrained whole-slide foundation models. We then describe how the pretrained MIL weights are transferred to downstream WSI analysis through full-parameter fine-tuning or linear probing.

3.1 Multi-teacher Distillation Pretraining on MIL

In WSI analysis, each slide is commonly represented as a bag of patch-level features. Given a WSI or a high-resolution pathology region, we divide it into N non-overlapping patches and use a frozen patch encoder f_{patch} to extract patch embeddings:

$$\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times D}, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the feature vector of the i-th patch. When spatial information is available, we also record the corresponding patch coordinates:

$$\mathbf{C} = \{(r_i, c_i)\}_{i=1}^{N}. \tag{2}$$

A MIL aggregator g_θ maps the bag of patch features, optionally together with spatial coordinates, into a slide-level representation:

$$\mathbf{z} = g_\theta(\mathbf{X}, \mathbf{C}) \in \mathbb{R}^d. \tag{3}$$

For downstream classification, $\mathbf{z}$ is further passed into a task-specific prediction head h_ψ:

$$\hat{\mathbf{y}} = h_\psi(\mathbf{z}). \tag{4}$$

Large-scale data are essential for pretraining MIL aggregators. However, directly using full WSIs for batch-level MIL pretraining is computationally inconvenient because different WSIs contain highly variable numbers of patches, resulting in irregular bag sizes and unstable memory consumption. Inspired by the high-resolution image pretraining strategy used in TITAN [9], we instead construct the pretraining data from extremely high-resolution pathology ROIs. Specifically, we use TCGA-UT-8K [9] as pretraining data. Each ROI is divided into non-overlapping 512×512 patches, yielding a regular 16×16 grid of 256 patches per ROI. For each patch, we extract a 768-dimensional feature vector using CONCH v1.5 [29].

Teacher models. We select TITAN [9] and CARE [51] as teacher models because they provide complementary sources of slide-level knowledge. TITAN [9] offers general-purpose and transferable whole-slide semantics learned from large-scale multimodal pretraining, whereas CARE [51] provides morphology-aware and molecularly informed regional representations. Their combination provides diverse supervisory signals and reduces the bias introduced by relying on a single teacher representation.

Given patch features $\mathbf{X}$ and spatial coordinates $\mathbf{C}$, TITAN [9] first arranges patch features into a 2D spatial grid and produces a pooled slide-level embedding:

$$\mathbf{t}^{\text{titan}} = T_{\text{titan}}(\mathbf{X}, \mathbf{C}) \in \mathbb{R}^{D_{\text{t}}}, \qquad D_{\text{t}} = 768. \tag{5}$$

Similarly, CARE [51] constructs morphologically relevant adaptive regions from the input bag and outputs a slide-level embedding:

$$\mathbf{t}^{\text{care}} = T_{\text{care}}(\mathbf{X}, \mathbf{C}) \in \mathbb{R}^{D_{\text{c}}}, \qquad D_{\text{c}} = 512. \tag{6}$$

During distillation pretraining, both teacher models are kept frozen.

Student models. The student is a MIL aggregator g_θ with its classification head removed, so that it only outputs a slide-level embedding:

$$\mathbf{z} = g_\theta(\mathbf{X}, \mathbf{C}) \in \mathbb{R}^d. \tag{7}$$

Since TITAN [9] and CARE [51] have different representation dimensions, we attach two learnable teacher-specific projection heads to map the student embedding into each teacher space:

$$\hat{\mathbf{t}}^{\text{titan}} = \phi_{\text{titan}}(\mathbf{z}) \in \mathbb{R}^{D_{\text{t}}}, \qquad \hat{\mathbf{t}}^{\text{care}} = \phi_{\text{care}}(\mathbf{z}) \in \mathbb{R}^{D_{\text{c}}}, \tag{8}$$

where $\phi_{\text{titan}} : \mathbb{R}^d \to \mathbb{R}^{D_{\text{t}}}$ and $\phi_{\text{care}} : \mathbb{R}^d \to \mathbb{R}^{D_{\text{c}}}$ are learnable linear projection heads. The projected student embeddings are then aligned with the corresponding frozen teacher embeddings during multi-teacher distillation.

Angular Dispersion Normalized Distillation Loss. A central challenge in multi-teacher distillation is that different teachers may exhibit different concentration levels in their representation distributions on the hypersphere. If one teacher produces more tightly clustered embeddings, the scale of its supervision signal may differ systematically from that of another teacher, leading to imbalanced optimization across teachers. To address this issue, inspired by c-RADIOv4 [38], we propose the Angular Dispersion Normalized Distillation Loss (ADN Loss), whose key idea is to normalize each teacher-specific alignment error by the running angular dispersion of that teacher's representation distribution.

For a batch of B slides, let $\hat{\mathbf{t}}_i$ and $\mathbf{t}_i$ denote the ℓ_2-normalized student prediction and teacher target for teacher k, respectively. The student-teacher discrepancy is measured by angular distance:

$$\Theta_i^{(k)} = \arccos\Big(\text{clip}\Big((\hat{\mathbf{t}}_i^{(k)})^\top \mathbf{t}_i^{(k)}, -1+\epsilon, 1-\epsilon\Big)\Big), \tag{9}$$

where ϵ is a small constant for numerical stability.

Different teachers may exhibit different angular variances in their representation spaces, which may cause one teacher to dominate the optimization. We therefore introduce a simple teacher-wise normalization strategy. For each teacher branch k, we estimate a running angular dispersion σ_k from the standard deviation of teacher embeddings around their update center direction on the unit hypersphere. The distillation loss for teacher k is computed as

$$\mathcal{L}_{\text{distill}}^{(k)} = \frac{1}{B}\sum_{i=1}^{B} \frac{\Theta_i^{(k)}}{\sigma_k + \epsilon}, \tag{10}$$

where $\sigma^{(k)}$ is the running angular dispersion. It is clamped to $[\sigma_{\min}, \infty)$ with $\sigma_{\min} = 0.05$ to prevent numerical instability.

For multi-teacher distillation, the final objective is the weighted sum over all teacher branches:

$$\mathcal{L}_{\text{ADN}} = \sum_{k \in \{\text{titan,care}\}} w_k \cdot \mathcal{L}_{\text{distill}}^{(k)} \tag{11}$$

where w_k is the teacher weight (set to 1 for both teachers) and $\epsilon = 10^{-6}$. By dividing each teacher's error by its angular dispersion, a teacher whose embeddings are more tightly clustered (smaller $\sigma^{(k)}$) will *not* dominate the gradient, since its normalized error is rescaled to a comparable magnitude as that of the more dispersed teacher. Detailed information of algorithm are provided in Appendix A.1.

3.2 Using Pretrained Weights for Downstream Tasks

After multi-teacher distillation pretraining, we discard the teacher-specific projection heads and retain only the pretrained MIL aggregator weights θ^*. These weights serve as reusable initialization for downstream WSI analysis. Given a downstream dataset, we load θ^* into the corresponding MIL architecture and attach a task-specific linear classification head h_ψ. The pretrained MIL model can then be used in two standard adaptation settings.

Full-parameter fine-tuning. In this setting, the pretrained aggregator is used as initialization, and both the MIL aggregator and the task-specific classification head are updated with downstream slide-level labels:

$$\min_{\theta,\psi} \mathcal{L}_{\mathrm{CE}} \left(h_\psi(g_\theta(\mathbf{X}, \mathbf{C})), \mathbf{y} \right). \tag{12}$$

This protocol evaluates whether the pretrained MIL weights provide a better optimization starting point than random initialization.

Linear probing. In this setting, the pretrained MIL aggregator is frozen, and only the task-specific linear head is trained:

$$\min_{\psi} \mathcal{L}_{\mathrm{CE}} \left(h_\psi(g_{\theta^*}(\mathbf{X}, \mathbf{C})), \mathbf{y} \right). \tag{13}$$

This protocol directly evaluates the transferability of the learned slide-level representation.

4 Experiments

4.1 Dataset

Pretraining dataset. During distillation pretraining, we use TCGA-UT-8K [9], a large-scale region-level pan-cancer pathology dataset. It contains 25,495 ROIs of size $8{,}192 \times 8{,}192$ pixels, derived from 9,662 H&E-stained FFPE diagnostic whole-slide images (WSIs) from TCGA.

Evaluation datasets. We evaluate our method on six public WSI cohorts covering 15 downstream tasks, including biomarker prediction, cancer subtyping, pan-cancer classification, and mutation prediction. BCNB [48] is used for ER, PR, and HER2 prediction; BRACS [5] for coarse- and fine-grained breast lesion subtyping; CPTAC [33] for pan-cancer and NSCLC classification; EBRAINS [39] for brain tumor subtyping and IDH mutation prediction; KIDRARE [13] for pediatric tumor classification; and MUT-HET-RCC [1] for BAP1, PBRM1, and SETD2 mutation prediction. For all tasks, we construct label-stratified train/validation/test splits following the same experimental protocol. Detailed dataset information is provided in Appendix A.2.

4.2 Implementation Details

Pretraining setup. Each ROI is partitioned into non-overlapping 512×512 patches, and each patch is encoded into a 768-dimensional feature vector using CONCH v1.5 [29]. Teacher embeddings are pre-computed and cached before training to avoid repeated inference. We jointly optimize the student aggregator and the two projection heads using Adam [21] with a learning rate of 5×10^{-5} and weight decay of 10^{-4} for 500 epochs. Training is conducted on 4 NVIDIA L20 48GB GPUs using PyTorch DistributedDataParallel (DDP), with a per-GPU batch size of 1024.

Downstream fine-tuning setup. For downstream tasks, we initialize the student aggregator with the pretrained weights θ^* and attach a task-specific linear classification head. We fine-tune the whole model end-to-end using Adam [21] with a learning rate of 10^{-4} and weight decay of 10^{-4}, 1 batch size. We apply early stopping with a patience of 5 based on validation balanced accuracy (bacc), and train for at most 50 epochs. Experiments are run on a single NVIDIA L20 48GB GPU and all baseline methods are trained under the same hyperparameters and settings.

4.3 Linear-probing Comparison between Teacher Models and Pretrained MIL Aggregators

Table 1 summarizes the test balanced accuracy of linear-probing trained on frozen slide-level representations produced by pretrained nine MIL aggregators and two teacher slide-embedding models,

Table 1: Merged-dataset linear-probing comparison between teacher models and pretrained MIL aggregators. Results are reported as test balanced accuracy (bacc, %) with standard deviation shown in the lower-right subscript. Δ denotes the difference between each pretrained MIL aggregator and the better teacher model on the same dataset.

Dataset	Teacher Models		Metric	Pretrained Student MIL Aggregators								
	TITAN	CARE		ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
BCNB (3 tasks)	$51.7_{(2.3)}$	$50.0_{(0.0)}$	BACC	$69.7_{(1.3)}$	$69.4_{(2.4)}$	$70.6_{(2.3)}$	$67.9_{(2.3)}$	$68.0_{(1.9)}$	$50.0_{(0.0)}$	$50.2_{(0.4)}$	$66.8_{(1.3)}$	$63.9_{(2.5)}$
			Δ	+18.0	+17.7	+18.9	+16.2	+16.3	−1.7	−1.5	+15.1	+12.2
BRACS (2 tasks)	$40.3_{(5.0)}$	$39.6_{(1.2)}$	BACC	$45.2_{(3.4)}$	$46.0_{(3.5)}$	$46.9_{(4.7)}$	$47.5_{(4.0)}$	$45.0_{(3.1)}$	$33.5_{(3.8)}$	$32.3_{(1.9)}$	$53.1_{(4.8)}$	$38.3_{(2.2)}$
			Δ	+4.9	+5.7	+6.6	+7.2	+4.7	−6.8	−8.0	+12.8	−2.0
CPTAC (2 tasks)	$92.9_{(2.7)}$	$86.7_{(4.0)}$	BACC	$93.3_{(0.9)}$	$94.4_{(1.3)}$	$94.8_{(0.2)}$	$95.5_{(1.2)}$	$94.3_{(0.9)}$	$91.1_{(1.4)}$	$90.1_{(1.2)}$	$94.2_{(0.4)}$	$94.2_{(1.1)}$
			Δ	+0.4	+1.5	+1.9	+2.6	+1.4	−1.8	−2.8	+1.3	+1.3
KIDRARE (2 tasks)	$51.5_{(5.5)}$	$44.9_{(1.2)}$	BACC	$61.0_{(1.8)}$	$62.7_{(1.6)}$	$61.6_{(2.7)}$	$60.6_{(1.6)}$	$53.8_{(2.4)}$	$41.9_{(3.3)}$	$42.3_{(2.5)}$	$61.1_{(2.9)}$	$55.4_{(3.0)}$
			Δ	+9.5	+11.2	+10.1	+9.1	+2.3	−9.6	−9.2	+9.6	+3.9
EBRAINS (3 tasks)	$56.8_{(11.4)}$	$49.0_{(3.7)}$	BACC	$79.9_{(1.0)}$	$80.8_{(1.7)}$	$81.6_{(1.7)}$	$81.5_{(1.9)}$	$79.8_{(1.2)}$	$71.7_{(3.2)}$	$75.8_{(2.0)}$	$82.2_{(0.9)}$	$79.6_{(1.7)}$
			Δ	+23.1	+24.0	+24.8	+24.7	+23.0	+14.9	+19.0	+25.4	+22.8
MUT_HET_RCC (3 tasks)	$52.4_{(1.5)}$	$51.2_{(1.3)}$	BACC	$62.6_{(2.3)}$	$64.0_{(2.3)}$	$62.0_{(1.7)}$	$62.6_{(1.6)}$	$59.7_{(1.6)}$	$52.8_{(0.6)}$	$52.8_{(1.4)}$	$60.9_{(2.8)}$	$56.1_{(1.9)}$
			Δ	+10.2	+11.6	+9.6	+10.2	+7.3	+0.4	+0.4	+8.5	+3.7

CARE [51] and TITAN [9] on six merged dataset groups covering 15 tasks. Detailed results for all 15 individual downstream tasks are provided in the Appendix A.3. Overall, the pretrained MIL aggregators achieve competitive performance under the linear-probing setting and, in many cases, outperform the teacher-level slide encoders. In particular, ABMIL [17], CLAM [30], TransMIL [42], WiKG [26], DAGMIL [11], and AMDMIL [27] outperform the corresponding teacher models across all dataset groups. This suggests that, after knowledge distillation, attention-based, Transformer-based, and graph-based MIL aggregators can effectively preserve transferable slide-level representational knowledge, which can be readily exploited by a simple linear classifier. Meanwhile, GDFMIL [53] and 2DMamba [52] show relatively weaker performance on several dataset, indicating that the transferability of distilled representations is still affected by architecture-specific inductive biases. Under the frozen-feature linear-probing protocol, their representational advantages may not be fully activated.

4.4 Fine-tuning Comparison between From-scratch and Pretrained MIL Aggregators

We evaluate distillation-pretrained initialization for downstream WSI classification using nine MIL aggregators under the same training protocol in Table 2. Detailed results for all 15 individual downstream tasks are provided in the Appendix A.4. Overall, pretrained initialization improves most MIL models over from-scratch training, indicating that teacher-guided pretraining provides useful slide-level priors. At the dataset level, BRACS [5] shows the largest average gain, suggesting that pretrained initialization is especially useful for tasks requiring subtle morphology and cross-patch aggregation. The pretrained MIL aggregators also achieve competitive performance compared with the teacher slide encoders in several cohorts, suggesting that knowledge from large slide foundation models can be effectively transferred into lightweight MIL architectures. Notably, although GDFMIL [53] and 2DMamba [52] show relatively weaker transferability under the frozen linear-probing setting, their performance is substantially improved after full-parameter fine-tuning. These results show that slide-level knowledge from large teachers can be effectively compressed into lightweight MIL aggregators for better downstream adaptation.

4.5 Few-shot transfer to downstream MIL tasks

To evaluate the data efficiency of the proposed initialization, we further assess model performance under few-shot slide classification settings with the number of shots per class $K \in \{1, 2, 4, 8, 16\}$. For each setting, we randomly sample K slides from each class to form the support set. Balanced accuracy is used as the evaluation metric. Figure 3 shows the results of WiKG [26] on BRACS-Coarse, CPTAC-NSCLC, and KidneyRare-Coarse. Complete index results on other models and tasks are provided in Appendix A.5. Across all three datasets, distillation initialization consistently outperforms scratch training at every evaluated shot number. On BRACS-Coarse, the gain ranges from 6.78% to 20.55% percentage points; on CPTAC-NSCLC, from 4.06% to 15.49% percentage points; and on KidneyRare-Coarse, from 6.80% to 15.57% percentage points. The largest improvements are generally observed in lower-shot settings, indicating that distillation is particularly beneficial

Table 2: Merged-dataset fine-tuning comparison between teacher models, from-scratch MIL, and pretrained MIL. Results are reported as test balanced accuracy (bacc, %) with standard deviation shown in the lower-right subscript. Δ denotes pretrained minus scratch and is reported using mean values only.

Dataset	Teacher Models		Setting	Student MIL Aggregators								
	TITAN	CARE		ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
BCNB (3 tasks)	$69.3_{(2.6)}$	$71.3_{(2.7)}$	Scratch	$70.4_{(1.6)}$	$70.0_{(2.6)}$	$65.2_{(3.3)}$	$69.6_{(2.7)}$	$67.4_{(3.1)}$	$70.0_{(5.2)}$	$69.2_{(2.3)}$	$70.4_{(2.1)}$	$72.3_{(1.4)}$
			Pretrain	$70.5_{(1.8)}$	$71.3_{(2.2)}$	$70.0_{(3.3)}$	$69.4_{(3.6)}$	$72.1_{(1.9)}$	$71.5_{(1.2)}$	$70.5_{(2.0)}$	$72.5_{(2.3)}$	$71.1_{(1.8)}$
			Δ	+0.1	+1.3	+4.7	−0.2	+4.7	+1.5	+1.3	+2.1	−1.2
BRACS (2 tasks)	$52.2_{(2.8)}$	$52.8_{(2.3)}$	Scratch	$50.9_{(5.9)}$	$52.9_{(2.3)}$	$47.2_{(2.2)}$	$47.4_{(6.6)}$	$55.1_{(1.5)}$	$50.7_{(3.3)}$	$57.9_{(2.7)}$	$55.5_{(4.7)}$	$54.0_{(4.4)}$
			Pretrain	$54.6_{(2.4)}$	$52.9_{(3.8)}$	$54.0_{(3.6)}$	$53.0_{(4.0)}$	$57.1_{(5.1)}$	$56.8_{(5.7)}$	$59.8_{(2.9)}$	$54.6_{(3.5)}$	$56.9_{(2.1)}$
			Δ	+3.6	0.0	+6.8	+5.6	+1.9	+6.1	+2.0	−0.8	+2.9
CPTAC (2 tasks)	$95.5_{(0.6)}$	$94.9_{(0.9)}$	Scratch	$92.8_{(0.6)}$	$93.5_{(1.2)}$	$94.5_{(1.5)}$	$91.9_{(1.0)}$	$92.4_{(2.3)}$	$93.2_{(1.8)}$	$93.4_{(1.1)}$	$94.9_{(0.9)}$	$94.5_{(1.5)}$
			Pretrain	$94.1_{(1.0)}$	$93.4_{(1.2)}$	$94.0_{(1.1)}$	$94.4_{(2.0)}$	$95.4_{(0.8)}$	$94.2_{(0.5)}$	$92.9_{(1.3)}$	$94.5_{(1.1)}$	$93.2_{(0.9)}$
			Δ	+1.2	−0.1	−0.6	+2.5	+3.0	+1.0	−0.5	−0.4	−1.3
KIDRARE (2 tasks)	$71.0_{(5.4)}$	$69.7_{(4.4)}$	Scratch	$67.9_{(2.3)}$	$68.6_{(1.9)}$	$70.0_{(3.4)}$	$69.7_{(4.2)}$	$70.0_{(3.1)}$	$69.6_{(3.8)}$	$70.8_{(3.4)}$	$75.3_{(3.0)}$	$68.1_{(2.8)}$
			Pretrain	$69.6_{(1.7)}$	$67.9_{(2.4)}$	$72.9_{(2.2)}$	$68.7_{(3.2)}$	$67.4_{(2.8)}$	$71.2_{(3.6)}$	$71.1_{(3.8)}$	$73.5_{(2.7)}$	$70.1_{(2.8)}$
			Δ	+1.7	−0.7	+2.9	−1.0	−2.6	+1.6	+0.3	−1.8	+2.0
EBRAINS (3 tasks)	$79.3_{(3.7)}$	$82.0_{(1.9)}$	Scratch	$81.2_{(1.5)}$	$81.4_{(1.3)}$	$79.6_{(2.1)}$	$79.6_{(2.4)}$	$80.7_{(2.0)}$	$79.9_{(2.6)}$	$82.7_{(1.3)}$	$81.9_{(1.9)}$	$82.4_{(1.4)}$
			Pretrain	$81.1_{(2.0)}$	$81.1_{(1.4)}$	$80.2_{(1.8)}$	$80.3_{(2.2)}$	$81.0_{(2.8)}$	$79.8_{(1.8)}$	$81.6_{(2.0)}$	$81.9_{(2.2)}$	$82.9_{(1.4)}$
			Δ	−0.2	−0.3	+0.6	+0.7	+0.3	−0.1	−1.1	0.0	+0.5
MUT_HET_RCC (3 tasks)	$70.1_{(2.8)}$	$67.2_{(2.5)}$	Scratch	$63.0_{(2.9)}$	$63.6_{(2.9)}$	$59.3_{(4.0)}$	$62.3_{(3.8)}$	$60.6_{(4.9)}$	$59.0_{(2.6)}$	$64.9_{(3.3)}$	$65.3_{(2.9)}$	$67.0_{(3.6)}$
			Pretrain	$63.5_{(3.6)}$	$67.1_{(3.1)}$	$64.9_{(2.0)}$	$60.8_{(6.3)}$	$60.9_{(4.9)}$	$63.2_{(5.3)}$	$67.1_{(3.4)}$	$66.4_{(3.3)}$	$66.4_{(3.1)}$
			Δ	+0.4	+3.5	+5.6	−1.4	+0.2	+4.2	+2.3	+1.1	−0.6

when labeled supervision is scarce. These results suggest that distillation transfers a useful prior for downstream MIL, improving both optimization stability and label efficiency.

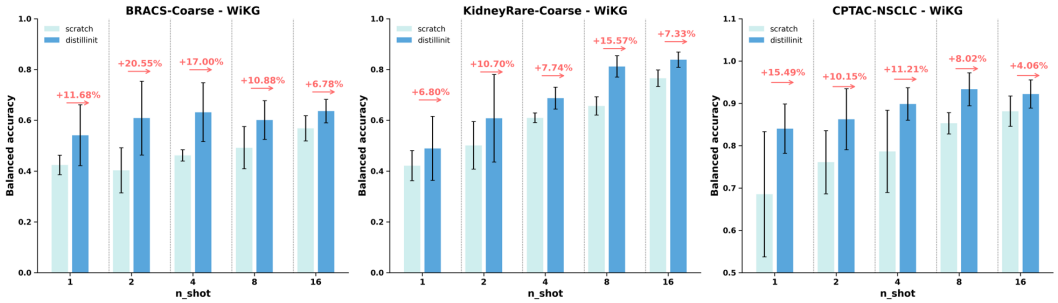

Figure 3: Few-shot comparison of from-scratch and distillation-initialized WiKG across three datasets.

4.6 Ablation

Effectiveness of multi-teacher distillation. We ablate the teacher source used for distillation pretraining by comparing TITAN-only, CARE-only, and joint distillation from both teachers. The results in Figure 4 show that the optimal teacher configuration varies across downstream datasets and MIL architectures, indicating that the transferred supervision from different slide-level teachers is architecture- and task-dependent. Specifically, CARE-only achieves strong results for 2DMamba [52] on BCNB-ER, while TITAN-only performs competitively for WiKG [26] on BCNB-ER. Joint distillation from both teachers also brings advantages in several set-

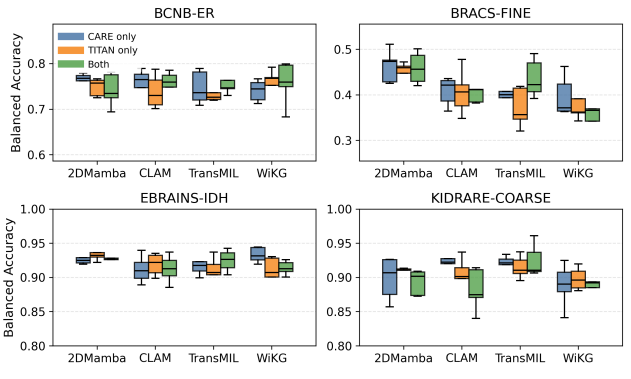

Figure 4: Teacher-source ablation for distillation pretraining. Results compare TITAN-only, CARE-only, and joint distillation across downstream datasets and MIL architectures.

tings, such as Transmil [42] on KIDRARE-COARSE, 2DMamba [52] on BRACS-FINE and EBRAINS-IDH. These results suggest that CARE [51] and TITAN [9] encode complementary but non-uniform supervisory signals, and that combining them can be beneficial when their inductive biases are well aligned with the target task and MIL architecture. Detailed results for all 15 individual downstream tasks are provided in the Appendix A.6.

Effectiveness of distillation loss. We further compare three distillation objectives: ADN loss, mean squared error (MSE) loss, and cosine loss. Since different slide-level teachers may have distinct feature scales and embedding distributions, these objectives provide complementary ways to align student and teacher representations. Table 3 reports the averaged bacc within each dataset group. Overall, the three losses achieve comparable performance, while angle loss shows clear advantages on several dataset groups, including BRACS [5], CPTAC [33], and KIDRARE [13]. This suggests that angular supervision can provide an effective geometric constraint for transferring slide-level teacher knowledge into lightweight MIL aggregators, especially when preserving relative representation structure is important. Complete results on all 15 downstream tasks are provided in Appendix A.7.

Table 3: Averaged balanced accuracy (bacc, %) with standard deviation of different type loss across dataset groups.

Dataset	ADN	MSE	Cosine
BCNB (3)	$70.5_{(1.8)}$	$\mathbf{70.8}_{(1.7)}$	$70.7_{(1.5)}$
BRACS (2)	$\mathbf{54.6}_{(2.4)}$	$53.5_{(2.7)}$	$53.8_{(3.2)}$
CPTAC (2)	$\mathbf{94.1}_{(1.0)}$	$93.4_{(0.8)}$	$93.6_{(1.3)}$
KIDRARE (2)	$\mathbf{69.6}_{(1.7)}$	$69.0_{(2.3)}$	$68.4_{(1.6)}$
EBRAINS (3)	$81.1_{(2.0)}$	$81.2_{(1.8)}$	$\mathbf{81.7}_{(1.5)}$
MUT-HET-RCC (3)	$63.5_{(3.6)}$	$63.1_{(3.7)}$	$\mathbf{64.4}_{(3.4)}$

Memory scalability analysis. We evaluate memory scalability on BRACS-coarse by increasing the number of patches from 5k to 30k for a single WSI. Compared with representative MIL models, CARE [51] and TITAN [9] show substantially faster memory growth. TITAN [9] already exceeds 35 GiB at 10k patches and becomes OOM beyond 15k patches, while CARE [51] reaches over 30 GiB at 30k patches. In contrast, MIL aggregators remain stable at 30k patches with much lower memory usage, including WiKG [26] at about 5.5 GiB, TransMIL [42] at about 3.3 GiB, and ABMIL [17] at about 0.24 GiB. These results highlight the practical scalability of lightweight MIL aggregators for large-patch WSI analysis and efficient downstream fine-tuning.

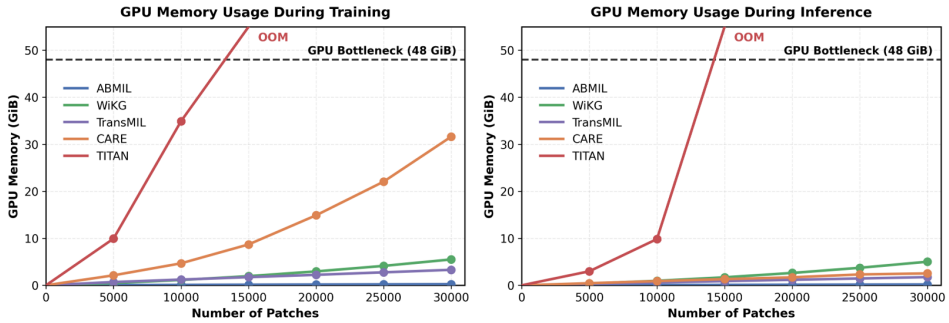

Figure 5: GPU memory scalability during training and testing under increasing patch numbers.

5 Conclusion and Future Work

In this work, we demonstrate that lightweight MIL aggregators can be effectively pretrained through slide-level knowledge distillation. Using CARE and TITAN as teacher models, we transfer their slide-level representations into diverse MIL architectures with an angular dispersion normalized distillation loss. The resulting pretrained aggregators provide reusable initialization for downstream adaptation, offering a practical alternative to conventional from-scratch MIL training. Across multiple computational pathology benchmarks, distilled initialization improves most MIL architectures and brings clear benefits in few-shot settings. Overall, distilled MIL initialization enables transferable slide-level priors while remaining efficient to adapt for WSI tasks with limited labels or computational resources. Future work will scale up the pretraining data and evaluate the proposed framework on a broader range of downstream tasks.

References

[1] P. H. Acosta et al. Intratumoral resolution of driver gene mutation heterogeneity in renal cancer using deep learning. *Cancer Research*, 82(15):2792–2806, 2022.

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9155–9163, 2019.

[3] Khaled Al-Thelaya, Nauman Ullah Gilal, Mahmood Alzubaidi, Fahad Majeed, Marco Agus, Jens Schneider, and Mowafa Househ. Applications of discriminative and deep learning feature extraction methods for whole slide image analysis: A survey. *Journal of Pathology Informatics*, 14:100335, 2023.

[4] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10925–10934, 2022.

[5] Nadia Brancati, Maria Frucci, Daniela Riccio, Giuseppe De Pietro, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, page baac093, 2022.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[7] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, Mane Williams, Anurag Vaidya, Sharifa Sahai, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Walt Williams, Long Phi Le, Georg Gerber, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[9] Tong Ding, Guillaume Jaume, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, 2025.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Mingxi Fu, Xitong Ling, Yuxuan Chen, Jiawen Li, Fanglei Fu, Huaitian Yuan, Tian Guan, Yonghong He, and Lianghui Zhu. Deformable attention graph representation learning for histopathology whole slide image analysis. *arXiv preprint arXiv:2508.05382*, 2025.

[12] Narmin Ghaffari Laleh, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y. Lu, Christian Trautwein, Rupert Langer, Bastian Dislich, Roman D. Buelow, Heike Irmgard Grabsch, Hermann Brenner, Jenny Chang-Claude, Elizabeth Alwers, Titus J. Brinker, Firas Khader, Daniel Truhn, Nadine T. Gaisa, Peter Boor, Volkmar Schulz, Johannes Krisam, and Jakob Nikolas Kather. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Medical Image Analysis*, 79:102474, 2022.

[13] Dexuan He, Xiao Zhou, Wenbin Guan, Liyuan Zhang, Xiaoman Zhang, Sinuo Xu, Ge Wang, Lifeng Wang, Xiaojun Yuan, Jing Ma, Xin Sun, Yanfeng Wang, Kun Sun, Ya Zhang, and Weidi Xie. Boosting pathology foundation models via few-shot prompt-tuning for rare cancer subtyping. *Nature Communications*, 2026.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018.

[18] Guillaume Jaume, Anurag Vaidya, Andrew Zhang, Andrew H. Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Faisal Mahmood. Multistain pretraining for slide representation learning in pathology. In *Computer Vision – ECCV 2024*, pages 19–37. Springer, 2024.

[19] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[20] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, 2023.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[23] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision (ECCV)*, pages 491–507, 2020.

[24] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2661–2671, 2019.

[25] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, 2021.

[26] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11323–11332, 2024.

[27] Xitong Ling, Minxi Ouyang, Yizhi Wang, Xinrui Chen, Renao Yan, Hongbo Chu, Junru Cheng, Tian Guan, Sufang Tian, Xiaoping Liu, et al. Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2795–2803, 2024.

[28] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

[29] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[30] Ming Y. Lu, Drew F. K. 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.

[31] Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering*, 10(3):545–564, 2026.

[32] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020.

[33] National Cancer Institute Genomic Data Commons. Clinical proteomic tumor analysis consortium (cptac). `https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/clinical-proteomic-tumor-analysis-consortium-cptac`, 2026. Accessed: 2026-04-21.

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2024.

[35] SeongUk Park and Nojun Kwak. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI 2020*, pages 1411–1418. IOS Press, 2020.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[37] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024.

[38] Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-radiov4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026.

[39] Theresa Roetzer-Pejrimovsky et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9, 2022.

[40] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

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

[42] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147, 2021.

[43] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Noam Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024.

[44] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[45] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

[46] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

[47] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, et al. A vision-language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

[48] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, 11:759007, 2021.

[49] Hanwen Xu, Naoto Usuyama, Jaume Bagga, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[50] Hongming Xu, Mingkang Wang, Duanbo Shi, Huamin Qin, Yunpeng Zhang, Zaiyi Liu, Anant Madabhushi, Peng Gao, Fengyu Cong, and Cheng Lu. When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. *Medical Image Analysis*, 101:103456, 2025.

[51] Di Zhang, Zhangpeng Gong, Xiaobo Pang, Jiashuai Liu, Junbo Lu, Hao Cui, Jiusong Ge, Zhi Zeng, Kai Yi, Yinghua Li, Si Liu, Tingsong Yu, Haoran Wang, Mireia Crispin-Ortuzar, Weimiao Yu, Chen Li, and Zeyu Gao. Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. *arXiv preprint arXiv:2602.21637*, 2026.

[52] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S Hosseini. 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3583–3592, 2025.

[53] 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.

A Technical appendices and supplementary material

A.1 Algorithm: Dual-teacher Slide-level Distillation Pretraining.

For completeness, we provide the detailed pseudocode of the proposed dual-teacher slide-level distillation pretraining procedure in Algorithm 1. The algorithm summarizes the full training pipeline, including offline teacher cache construction, student forward propagation, teacher-specific projection and normalization, EMA-based angular dispersion estimation, and the final weighted dual-teacher distillation objective. This appendix is intended to clarify the implementation details of the pretraining stage and to improve reproducibility of our framework.

A.2 Detailed description of evaluation datasets.

We evaluate our method on six public whole-slide image (WSI) cohorts covering biomarker prediction, subtype classification, pan-cancer classification, and mutation prediction. Below, we provide detailed descriptions of all evaluation datasets, including task definitions, class compositions, and train/validation/test splits.

Algorithm 1 Dual-teacher slide-level distillation pretraining

Require: Unlabeled slides $\mathcal{D} = \{(\mathbf{X}_s, \mathbf{C}_s)\}_{s=1}^{S}$
Require: Frozen teachers $T_{\text{titan}}, T_{\text{care}}$
Require: Student MIL aggregator g_θ
Require: Projection heads $\phi_{\text{titan}}, \phi_{\text{care}}$
Require: Teacher weights $w_{\text{titan}}, w_{\text{care}}$
Let $\mathcal{K} = \{\text{titan}, \text{care}\}$
Build teacher cache
for all slides $(\mathbf{X}_s, \mathbf{C}_s) \in \mathcal{D}$ **do**
    $\mathbf{t}_s^{\text{titan}} \leftarrow T_{\text{titan}}(\mathbf{X}_s, \mathbf{C}_s)$
    $\mathbf{t}_s^{\text{care}} \leftarrow T_{\text{care}}(\mathbf{X}_s, \mathbf{C}_s)$
    cache $(\mathbf{t}_s^{\text{titan}}, \mathbf{t}_s^{\text{care}})$
end for
Initialize EMA statistics $(\boldsymbol{\mu}^{(k)}, \sigma^{(k)})$ for all $k \in \mathcal{K}$
for epoch $= 1$ to E **do**
    for all mini-batches $\mathcal{B} = \{(\mathbf{X}_i, \mathbf{C}_i, \mathbf{t}_i^{\text{titan}}, \mathbf{t}_i^{\text{care}})\}_{i=1}^{B}$ **do**
        $\mathbf{z}_i \leftarrow g_\theta(\mathbf{X}_i, \mathbf{C}_i)$ for all i
        $\hat{\mathbf{t}}_i^{\text{titan}} \leftarrow \phi_{\text{titan}}(\mathbf{z}_i)$ for all i
        $\hat{\mathbf{t}}_i^{\text{care}} \leftarrow \phi_{\text{care}}(\mathbf{z}_i)$ for all i
        for all $k \in \mathcal{K}$ **do**
            $\hat{\mathbf{t}}_i^{(k)} \leftarrow \ell_2\text{-norm}(\hat{\mathbf{t}}_i^{(k)})$ for all i
            $\mathbf{t}_i^{(k)} \leftarrow \ell_2\text{-norm}(\mathbf{t}_i^{(k)})$ for all i
            $e_i^{(k)} \leftarrow 1 - \cos(\hat{\mathbf{t}}_i^{(k)}, \mathbf{t}_i^{(k)})$ for all i
            update $\boldsymbol{\mu}^{(k)}$ via EMA
            update $\sigma^{(k)}$ via EMA and clamp to $\sigma_{\min}$
            $\mathcal{L}^{(k)} \leftarrow \frac{1}{B} \sum_{i=1}^{B} \frac{e_i^{(k)}}{\sigma^{(k)} + \epsilon}$
        end for
        $\mathcal{L} \leftarrow w_{\text{titan}} \mathcal{L}^{\text{titan}} + w_{\text{care}} \mathcal{L}^{\text{care}}$
        update $\theta, \phi_{\text{titan}}, \phi_{\text{care}}$ with AdamW on $\mathcal{L}$
    end for
end for
return distilled student weights θ^*

BCNB [48]. The BCNB cohort consists of 1,058 WSIs of early breast cancer core needle biopsies. On this cohort, we construct three biomarker prediction tasks, namely ER positivity prediction, PR positivity prediction, and HER2 positivity prediction. For all three tasks, we adopt label-stratified train/validation/test splits with 677, 169, and 212 slides, respectively.

BRACS [5]. The BRACS cohort consists of 546 WSIs of different subtypes of breast cancer. We construct two subtype classification tasks on this cohort. The first is a coarse-grained 3-class classification task with three classes: benign, atypical, and malignant. The second is a fine-grained 7-class classification task with seven classes: Normal, benign, udh, fea, adh, dcis, and ic. For both tasks, we use the same label-stratified split of 394 training slides, 65 validation slides, and 87 test slides.

CPTAC [33]. The CPTAC cohort consists of 2,059 pan-cancer WSIs from the following nine cancer types: Lung adenocarcinoma (LUAD), Head and neck squamous cell carcinoma (HNSC), Clear cell renal cell carcinoma (CCRCC), Colon adenocarcinoma (COAD), Lung squamous cell carcinoma (LSCC), Ovarian cancer (OV), Pancreatic ductal adenocarcinoma (PDA), Glioblastoma (GBM), and Breast cancer (BRCA). We construct two tasks on this cohort. The first is a 9-class pan-cancer classification task over the nine cancer types listed above, with a label-stratified split of 1313/340/406 slides for train/validation/test. The second is a binary NSCLC classification task, with a label-stratified split of 402/108/120 slides.

EBRAINS [39]. The EBRAINS cohort consists of 2,310 WSIs of different brain tumour subtypes. We construct three tasks on this cohort. The first is a coarse-grained 12-class subtype classification task with the following classes: Metastatic tumours, Meningiomas, Embryonal tumours, Circumscribed astrocytic gliomas, Ependymal tumours, Mesenchymal, non-meningothelial tumours involving the CNS, Glioneuronal and neuronal tumours, Adult-type diffuse gliomas, Paediatric-type diffuse low-

grade gliomas, Haematolymphoid tumours involving the CNS, Tumours of the sellar region, and Cranial and paraspinal nerve tumours. The corresponding train/validation/test split is 1470/359/481.

The second is a fine-grained 30-class subtype classification task with the following classes: Schwannoma, Ependymoma, Fibrous meningioma, Ganglioglioma, Anaplastic astrocytoma, IDH-mutant, Transitional meningioma, Meningothelial meningioma, Anaplastic ependymoma, Glioblastoma, IDH-wildtype, Atypical meningioma, Oligodendroglioma, IDH-mutant and 1p-19q codeleted, Lipoma, Diffuse large B-cell lymphoma of the CNS, Metastatic tumours, Haemangiopericytoma, Pilocytic astrocytoma, Adamantinomatous craniopharyngioma, Anaplastic meningioma, Medulloblastoma, non-WNT-non-SHH, Pituitary adenoma, Angiomatous meningioma, Diffuse astrocytoma, IDH-mutant, Glioblastoma, IDH-mutant, Haemangioblastoma, Anaplastic oligodendroglioma, IDH-mutant and 1p-19q codeleted, Haemangioma, Langerhans cell histiocytosis, Secretory meningioma, Gliosarcoma, and Anaplastic astrocytoma, IDH-wildtype. The corresponding split is 1455/363/492.

The third is an IDH biomarker mutation prediction task. We use a label-stratified train/validation/test split of 540/130/173 slides for this task.

KidRare [13]. The KidRare cohort consists of 1,283 WSIs of rare pediatric tumours. We construct two classification tasks on this cohort. The first is a coarse-grained 4-class classification task over the following pediatric cancer types: Nephroblastoma, Neuroblastoma, Hepatoblastoma, and Medulloblastoma. The second is a fine-grained 13-class classification task with the following classes: Mixed epithelial and mesenchymal hepatoblastoma, Epithelial macrotrabecular pattern of hepatoblastoma, Tumor, Ganglioneuroblastoma, intermixed, Large Cell/Anaplastic medulloblastoma, Normal, Desmoplastic nodular medulloblastoma, unknown, Differentiating neuroblastoma, Poorly differentiated neuroblastoma, Epithelial mixed fetal and embryonal hepatoblastoma, Classic medulloblastoma, and Pure fetal hepatoblastoma with low mitotic activity. For both tasks, we use the same label-stratified split of 821/206/256 slides for train/validation/test.

MUT-HET-RCC [1]. The MUT-HET-RCC cohort consists of 1,291 WSIs of clear cell renal cell carcinoma. On this cohort, we construct three mutation prediction tasks targeting BAP1, PBRM1, and SETD2. For all three tasks, we use label-stratified train/validation/test splits with 825, 207, and 259 slides, respectively.

A.3 Results of linear probe

Table 4: Linear-probe performance of teacher models on the 15 downstream datasets. Values are reported as mean with standard deviation in subscript.

Dataset	CARE				TITAN			
	Acc.	bacc	F1	AUC	Acc.	bacc	F1	AUC
bcnb_er	$78.3_{0.0}$	$50.0_{0.0}$	$43.9_{0.0}$	$80.8_{0.8}$	$79.1_{1.1}$	$51.7_{2.5}$	$47.3_{4.8}$	$80.1_{1.8}$
bcnb_pr	$74.5_{0.0}$	$50.0_{0.0}$	$42.7_{0.0}$	$74.6_{0.9}$	$75.5_{1.3}$	$51.8_{2.5}$	$46.3_{5.0}$	$70.1_{1.8}$
bcnb_her2	$74.1_{0.0}$	$50.0_{0.0}$	$42.5_{0.0}$	$63.1_{2.8}$	$74.7_{0.9}$	$51.4_{1.9}$	$45.4_{3.9}$	$63.4_{5.6}$
bracs_coarse	$56.5_{1.9}$	$51.2_{1.7}$	$44.0_{1.5}$	$74.9_{2.3}$	$57.5_{11.0}$	$52.1_{10.0}$	$43.5_{13.2}$	$83.8_{5.3}$
bracs_fine	$40.5_{1.3}$	$28.0_{0.8}$	$18.9_{0.1}$	$70.4_{2.3}$	$41.4_{0.0}$	$28.6_{0.0}$	$18.9_{0.1}$	$79.2_{1.6}$
cptac_nsclc	$87.0_{7.2}$	$86.8_{7.3}$	$86.5_{8.1}$	$99.4_{0.5}$	$94.3_{5.0}$	$94.2_{5.1}$	$94.2_{5.1}$	$99.9_{0.2}$
cptac_pancancer	$89.0_{0.5}$	$86.7_{0.6}$	$87.2_{0.4}$	$98.8_{0.0}$	$92.4_{0.3}$	$91.5_{0.2}$	$91.9_{0.4}$	$99.6_{0.0}$
ebrains_idh	$64.5_{7.2}$	$54.3_{9.6}$	$44.9_{15.4}$	$93.2_{0.9}$	$68.8_{8.0}$	$59.7_{10.3}$	$54.2_{16.4}$	$96.3_{0.4}$
ebrains_coarse	$79.0_{0.2}$	$51.4_{0.4}$	$52.5_{0.6}$	$97.6_{0.0}$	$86.7_{1.2}$	$67.4_{3.4}$	$70.7_{4.0}$	$98.7_{0.1}$
ebrains_fine	$56.9_{0.9}$	$41.3_{1.1}$	$40.0_{1.3}$	$96.8_{0.0}$	$56.3_{18.7}$	$43.3_{20.4}$	$42.8_{20.4}$	$97.4_{1.1}$
kidrare_coarse	$77.3_{1.3}$	$66.9_{2.1}$	$65.9_{2.8}$	$96.0_{0.1}$	$83.9_{2.9}$	$73.5_{6.4}$	$73.3_{8.4}$	$98.1_{0.2}$
kidrare_fine	$53.7_{0.4}$	$22.9_{0.3}$	$18.9_{0.3}$	$91.2_{0.2}$	$60.4_{2.7}$	$29.4_{4.5}$	$26.8_{5.5}$	$93.4_{0.3}$
mut_het_rcc_bap1_mutation	$87.3_{0.0}$	$50.0_{0.0}$	$46.6_{0.0}$	$80.0_{1.3}$	$87.3_{0.0}$	$50.0_{0.0}$	$46.6_{0.0}$	$79.6_{1.0}$
mut_het_rcc_pbrm1_mutation	$55.1_{3.5}$	$53.6_{3.8}$	$42.8_{9.3}$	$69.3_{1.3}$	$58.0_{3.6}$	$56.9_{4.1}$	$51.0_{10.0}$	$70.3_{0.5}$
mut_het_rcc_setd2_mutation	$73.0_{0.0}$	$50.0_{0.0}$	$42.2_{0.0}$	$68.8_{4.9}$	$73.0_{0.2}$	$50.2_{0.5}$	$42.8_{1.2}$	$67.6_{5.8}$

Table 5: Balanced accuracy (bacc) of teacher models and distillation-initialized MIL models under the linear-probe setting. For each dataset, Δ denotes the difference between each MIL model and the best teacher result on the same dataset. Values are reported as mean with standard deviation in subscript.

Dataset	Teacher Models		Metric	Pretrained Student MIL Aggregators								
	TITAN	CARE		ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	$51.7_{(2.5)}$	$50.0_{(0.0)}$	BACC	$74.0_{(1.2)}$	$74.4_{(2.2)}$	$74.6_{(2.8)}$	$71.5_{(1.9)}$	$70.6_{(1.7)}$	$50.0_{(0.0)}$	$50.0_{(0.0)}$	$66.3_{(1.1)}$	$66.2_{(3.2)}$
			Δ	+22.3	+22.7	+22.9	+19.8	+18.9	−1.7	−1.7	+14.6	+14.5
bcnb_pr	$51.8_{(2.5)}$	$50.0_{(0.0)}$	BACC	$71.3_{(1.4)}$	$70.5_{(3.5)}$	$72.5_{(2.0)}$	$69.5_{(3.1)}$	$68.8_{(2.0)}$	$50.0_{(0.0)}$	$50.0_{(0.0)}$	$68.3_{(1.4)}$	$64.8_{(1.7)}$
			Δ	+19.5	+18.7	+20.7	+17.7	+17.0	−1.8	−1.8	+16.5	+13.0
bcnb_her2	$51.4_{(1.9)}$	$50.0_{(0.0)}$	BACC	$64.0_{(1.3)}$	$63.2_{(1.5)}$	$64.7_{(2.0)}$	$62.8_{(1.9)}$	$64.6_{(2.0)}$	$50.0_{(0.0)}$	$50.6_{(1.2)}$	$65.6_{(1.5)}$	$60.8_{(2.7)}$
			Δ	+12.6	+11.8	+13.3	+11.4	+13.2	−1.4	−0.8	+14.2	+9.4
bracs_coarse	$52.1_{(10.0)}$	$51.2_{(1.7)}$	BACC	$58.1_{(3.4)}$	$61.5_{(2.3)}$	$56.2_{(1.1)}$	$62.6_{(1.5)}$	$54.6_{(1.9)}$	$39.4_{(7.0)}$	$43.8_{(3.1)}$	$65.0_{(2.4)}$	$52.1_{(3.4)}$
			Δ	+6.0	+9.4	+4.1	+10.5	+2.5	−12.7	−8.3	+12.9	0.0
bracs_fine	$28.6_{(0.0)}$	$28.0_{(0.8)}$	BACC	$32.3_{(3.3)}$	$30.4_{(4.7)}$	$37.5_{(8.2)}$	$32.4_{(6.6)}$	$35.4_{(4.3)}$	$27.7_{(0.5)}$	$20.9_{(0.7)}$	$41.2_{(7.2)}$	$24.6_{(1.0)}$
			Δ	+3.7	+1.8	+8.9	+3.8	+6.8	−0.9	−7.7	+12.6	−4.0
cptac_nsclc	$94.2_{(5.1)}$	$86.8_{(7.3)}$	BACC	$94.3_{(0.6)}$	$96.7_{(0.5)}$	$95.8_{(0.0)}$	$98.5_{(0.6)}$	$96.5_{(0.6)}$	$93.5_{(2.6)}$	$91.9_{(1.7)}$	$93.8_{(0.4)}$	$97.0_{(1.0)}$
			Δ	+0.1	+2.5	+1.6	+4.3	+2.3	−0.7	−2.3	−0.4	+2.8
cptac_pancancer	$91.5_{(0.2)}$	$86.7_{(0.6)}$	BACC	$92.2_{(1.3)}$	$92.2_{(2.1)}$	$93.8_{(0.4)}$	$92.4_{(1.8)}$	$92.1_{(1.1)}$	$88.6_{(0.3)}$	$88.3_{(0.7)}$	$94.5_{(0.4)}$	$91.3_{(1.2)}$
			Δ	+0.7	+0.7	+2.3	+0.9	+0.6	−2.9	−3.2	+3.0	−0.2
ebrains_idh	$59.7_{(10.3)}$	$54.3_{(9.6)}$	BACC	$91.4_{(0.8)}$	$91.9_{(1.8)}$	$92.0_{(1.8)}$	$92.7_{(1.5)}$	$92.2_{(1.9)}$	$89.1_{(3.0)}$	$90.2_{(0.5)}$	$93.5_{(0.4)}$	$92.2_{(1.8)}$
			Δ	+31.7	+32.2	+32.3	+33.0	+32.5	+29.4	+30.5	+33.8	+32.5
ebrains_coarse	$67.4_{(3.4)}$	$51.4_{(0.4)}$	BACC	$83.8_{(0.8)}$	$85.1_{(2.0)}$	$85.9_{(1.5)}$	$85.1_{(1.5)}$	$84.2_{(1.5)}$	$70.7_{(2.3)}$	$79.8_{(3.6)}$	$86.2_{(1.0)}$	$82.4_{(1.0)}$
			Δ	+16.4	+17.7	+18.5	+17.7	+16.8	+3.3	+12.4	+18.8	+15.0
ebrains_fine	$43.3_{(20.4)}$	$41.3_{(1.1)}$	BACC	$64.6_{(1.5)}$	$65.5_{(1.4)}$	$67.0_{(1.7)}$	$66.8_{(2.7)}$	$63.2_{(0.2)}$	$55.4_{(4.3)}$	$57.3_{(1.8)}$	$66.9_{(1.4)}$	$64.3_{(2.5)}$
			Δ	+21.3	+22.2	+23.7	+23.5	+19.9	+12.1	+14.0	+23.6	+21.0
kidrare_coarse	$73.5_{(6.4)}$	$66.9_{(2.1)}$	BACC	$84.4_{(1.8)}$	$85.8_{(1.2)}$	$83.9_{(1.6)}$	$82.6_{(0.7)}$	$74.1_{(2.2)}$	$61.6_{(3.3)}$	$63.0_{(2.4)}$	$84.2_{(3.5)}$	$77.6_{(3.0)}$
			Δ	+10.9	+12.3	+10.4	+9.1	+0.6	−11.9	−10.5	+10.7	+4.1
kidrare_fine	$29.4_{(4.5)}$	$22.9_{(0.3)}$	BACC	$37.6_{(1.8)}$	$39.5_{(2.1)}$	$39.3_{(3.7)}$	$38.6_{(2.4)}$	$33.5_{(2.5)}$	$22.3_{(3.4)}$	$21.6_{(2.7)}$	$37.9_{(2.2)}$	$33.2_{(3.0)}$
			Δ	+8.2	+10.1	+9.9	+9.2	+4.1	−7.1	−7.8	+8.5	+3.8
mut_het_rcc_bap1_mutation	$50.0_{(0.0)}$	$50.0_{(0.0)}$	BACC	$61.2_{(3.2)}$	$63.2_{(3.1)}$	$60.9_{(1.8)}$	$61.3_{(1.8)}$	$55.2_{(2.8)}$	$50.0_{(0.0)}$	$50.0_{(0.0)}$	$57.6_{(4.0)}$	$53.6_{(2.1)}$
			Δ	+11.2	+13.2	+10.9	+11.3	+5.2	0.0	0.0	+7.6	+3.6
mut_het_rcc_pbrm1_mutation	$56.9_{(4.1)}$	$53.6_{(3.8)}$	BACC	$66.4_{(1.0)}$	$66.8_{(0.5)}$	$62.5_{(1.5)}$	$64.3_{(2.6)}$	$64.6_{(1.0)}$	$58.3_{(1.7)}$	$58.6_{(3.9)}$	$64.7_{(2.4)}$	$63.3_{(2.2)}$
			Δ	+9.5	+9.9	+5.6	+7.4	+7.7	+1.4	+1.7	+7.8	+6.4
mut_het_rcc_setd2_mutation	$50.2_{(0.5)}$	$50.0_{(0.0)}$	BACC	$60.2_{(2.7)}$	$61.9_{(3.3)}$	$62.5_{(1.8)}$	$62.1_{(0.3)}$	$59.4_{(1.0)}$	$50.0_{(0.0)}$	$49.8_{(0.2)}$	$60.4_{(1.9)}$	$51.3_{(1.4)}$
			Δ	+10.0	+11.7	+12.3	+11.9	+9.2	−0.2	−0.4	+10.2	+1.1

Table 6: Accuracy of teacher models and distillation-initialized MIL models under the linear-probe setting. For each dataset, Δ denotes the difference between each MIL model and the best teacher result on the same dataset. Values are reported as mean with standard deviation in subscript.

Dataset	Teacher Models		Metric	Pretrained Student MIL Aggregators								
	TITAN	CARE		ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	$79.1_{(1.1)}$	$78.3_{(0.0)}$	ACC	$81.6_{(2.0)}$	$83.5_{(0.5)}$	$82.5_{(1.0)}$	$83.1_{(1.1)}$	$83.8_{(1.0)}$	$78.3_{(0.0)}$	$78.3_{(0.0)}$	$81.4_{(1.6)}$	$82.0_{(0.8)}$
			Δ	+2.5	+4.4	+3.4	+4.0	+4.7	−0.8	−0.8	+2.3	+2.9
bcnb_pr	$75.5_{(1.3)}$	$74.5_{(0.0)}$	ACC	$79.3_{(1.0)}$	$79.0_{(2.6)}$	$79.1_{(1.9)}$	$78.5_{(1.4)}$	$80.0_{(0.9)}$	$74.5_{(0.0)}$	$74.5_{(0.0)}$	$77.2_{(1.8)}$	$77.4_{(0.3)}$
			Δ	+3.8	+3.5	+3.6	+3.0	+4.5	−1.0	−1.0	+1.7	+1.9
bcnb_her2	$74.7_{(0.9)}$	$74.1_{(0.0)}$	ACC	$68.0_{(3.9)}$	$69.5_{(4.2)}$	$69.2_{(6.9)}$	$67.9_{(7.0)}$	$73.6_{(0.8)}$	$73.9_{(0.0)}$	$74.1_{(0.4)}$	$73.6_{(1.5)}$	$72.8_{(1.6)}$
			Δ	−6.7	−5.2	−5.5	−6.8	−1.1	−0.8	−0.6	−1.1	−1.9
bracs_coarse	$57.5_{(11.0)}$	$56.5_{(1.9)}$	ACC	$63.4_{(3.3)}$	$66.4_{(2.0)}$	$62.1_{(1.3)}$	$67.6_{(1.3)}$	$60.2_{(2.1)}$	$43.5_{(7.7)}$	$48.3_{(3.4)}$	$70.6_{(2.1)}$	$57.5_{(3.8)}$
			Δ	+5.9	+8.9	+4.6	+10.1	+2.7	−14.0	−9.2	+13.1	0.0
bracs_fine	$41.4_{(0.0)}$	$40.5_{(1.3)}$	ACC	$40.5_{(2.0)}$	$39.3_{(3.7)}$	$43.9_{(6.3)}$	$40.9_{(3.9)}$	$41.5_{(3.9)}$	$35.9_{(0.8)}$	$29.9_{(0.7)}$	$48.3_{(6.2)}$	$34.9_{(1.6)}$
			Δ	−0.9	−2.1	+2.5	−0.5	+0.1	−5.5	−11.5	+6.9	−6.5
cptac_nsclc	$94.3_{(5.0)}$	$87.0_{(7.2)}$	ACC	$94.3_{(0.6)}$	$96.6_{(0.5)}$	$95.8_{(0.0)}$	$98.5_{(0.6)}$	$96.5_{(0.6)}$	$93.5_{(2.7)}$	$91.8_{(1.7)}$	$93.8_{(0.4)}$	$97.0_{(1.0)}$
			Δ	0.0	+2.3	+1.5	+4.2	+2.2	−0.8	−2.5	−0.5	+2.7
cptac_pancancer	$92.4_{(0.3)}$	$89.0_{(0.5)}$	ACC	$93.0_{(0.9)}$	$92.8_{(1.8)}$	$94.4_{(0.6)}$	$93.1_{(1.4)}$	$92.6_{(0.9)}$	$89.5_{(0.4)}$	$90.1_{(0.6)}$	$95.3_{(0.3)}$	$92.4_{(1.0)}$
			Δ	+0.6	+0.4	+2.0	+0.7	+0.2	−2.9	−2.3	+2.9	0.0
ebrains_idh	$68.8_{(8.0)}$	$64.5_{(7.2)}$	ACC	$91.4_{(1.1)}$	$91.5_{(2.3)}$	$91.5_{(2.3)}$	$92.3_{(1.3)}$	$92.3_{(1.7)}$	$88.9_{(3.7)}$	$90.3_{(0.5)}$	$93.6_{(0.5)}$	$92.8_{(1.4)}$
			Δ	+22.6	+22.7	+22.7	+23.5	+23.5	+20.1	+21.5	+24.8	+24.0
ebrains_coarse	$86.7_{(1.2)}$	$79.0_{(0.2)}$	ACC	$87.5_{(1.0)}$	$87.4_{(1.0)}$	$88.8_{(1.7)}$	$87.6_{(1.5)}$	$88.4_{(0.6)}$	$84.8_{(0.8)}$	$87.5_{(1.6)}$	$90.2_{(0.6)}$	$88.3_{(0.7)}$
			Δ	+0.8	+0.7	+2.1	+0.9	+1.7	−1.9	+0.8	+3.5	+1.6
ebrains_fine	$56.3_{(18.7)}$	$56.9_{(0.9)}$	ACC	$71.4_{(1.4)}$	$71.8_{(0.4)}$	$73.9_{(1.4)}$	$72.7_{(1.6)}$	$70.6_{(1.0)}$	$66.3_{(2.4)}$	$66.9_{(1.5)}$	$73.9_{(0.7)}$	$72.0_{(1.7)}$
			Δ	+14.5	+14.9	+17.0	+15.8	+13.7	+9.4	+10.0	+17.0	+15.1
kidrare_coarse	$83.9_{(2.9)}$	$77.3_{(1.3)}$	ACC	$83.6_{(3.0)}$	$86.1_{(1.2)}$	$84.9_{(1.2)}$	$83.7_{(0.6)}$	$77.2_{(1.5)}$	$70.3_{(1.7)}$	$72.4_{(1.4)}$	$86.5_{(1.8)}$	$81.2_{(2.4)}$
			Δ	−0.3	+2.2	+1.0	−0.2	−6.7	−13.6	−11.5	+2.6	−2.7
kidrare_fine	$60.4_{(2.7)}$	$53.7_{(0.4)}$	ACC	$57.3_{(1.3)}$	$59.9_{(1.0)}$	$63.5_{(2.1)}$	$60.7_{(3.0)}$	$60.4_{(1.5)}$	$50.8_{(3.4)}$	$49.0_{(3.2)}$	$62.3_{(1.7)}$	$59.9_{(1.8)}$
			Δ	−3.1	−0.5	+3.1	+0.3	0.0	−9.6	−11.4	+1.9	−0.5
mut_het_rcc_bap1_mutation	$87.3_{(0.0)}$	$87.3_{(0.0)}$	ACC	$86.8_{(0.7)}$	$86.6_{(0.4)}$	$87.0_{(0.4)}$	$86.5_{(0.2)}$	$87.4_{(0.6)}$	$87.0_{(0.0)}$	$87.0_{(0.0)}$	$87.2_{(0.4)}$	$87.4_{(0.3)}$
			Δ	−0.5	−0.7	−0.3	−0.8	+0.1	−0.3	−0.3	−0.1	+0.1
mut_het_rcc_pbrm1_mutation	$58.0_{(3.6)}$	$55.1_{(3.5)}$	ACC	$66.3_{(1.1)}$	$66.6_{(0.5)}$	$62.7_{(1.5)}$	$64.3_{(2.6)}$	$64.7_{(0.9)}$	$58.1_{(2.0)}$	$58.8_{(3.7)}$	$64.6_{(2.3)}$	$63.5_{(2.2)}$
			Δ	+8.3	+8.6	+4.7	+6.3	+6.7	+0.1	+0.8	+6.6	+5.5
mut_het_rcc_setd2_mutation	$73.0_{(0.2)}$	$73.0_{(0.0)}$	ACC	$72.5_{(0.6)}$	$71.5_{(3.2)}$	$72.9_{(2.7)}$	$73.8_{(1.8)}$	$74.2_{(0.6)}$	$72.3_{(0.0)}$	$72.1_{(0.3)}$	$73.4_{(0.7)}$	$72.2_{(0.4)}$
			Δ	−0.5	−1.5	−0.1	+0.8	+1.2	−0.7	−0.9	+0.4	−0.8

Table 7: AUC of teacher models and distillation-initialized MIL models under the linear-probe setting. For each dataset, Δ denotes the difference between each MIL model and the best teacher result on the same dataset. Values are reported as mean with standard deviation in subscript.

Dataset	Teacher Models		Metric	Pretrained Student MIL Aggregators								
	TITAN	CARE		ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	$80.1_{(1.8)}$	$80.8_{(0.8)}$	AUC	$83.3_{(1.1)}$	$86.2_{(0.5)}$	$84.4_{(0.7)}$	$83.5_{(0.4)}$	$83.0_{(0.6)}$	$55.9_{(5.3)}$	$51.0_{(3.1)}$	$83.6_{(1.3)}$	$81.8_{(0.9)}$
			Δ	+2.5	+5.4	+3.6	+2.7	+2.2	−24.9	−29.8	+2.8	+1.0
bcnb_pr	$70.1_{(1.8)}$	$74.6_{(0.9)}$	AUC	$76.7_{(1.5)}$	$76.7_{(2.3)}$	$78.4_{(0.9)}$	$77.3_{(1.1)}$	$76.7_{(0.7)}$	$55.3_{(3.4)}$	$52.9_{(3.0)}$	$76.3_{(1.2)}$	$72.9_{(0.9)}$
			Δ	+2.1	+2.1	+3.8	+2.7	+2.1	−19.3	−21.7	+1.7	−1.7
bcnb_her2	$63.4_{(5.6)}$	$63.1_{(2.8)}$	AUC	$69.2_{(1.4)}$	$68.2_{(1.8)}$	$71.6_{(1.4)}$	$69.8_{(0.8)}$	$71.7_{(1.0)}$	$56.5_{(2.9)}$	$61.7_{(4.3)}$	$72.8_{(1.6)}$	$68.9_{(1.1)}$
			Δ	+5.8	+4.8	+8.2	+6.4	+8.3	−6.9	−1.7	+9.4	+5.5
bracs_coarse	$83.8_{(5.3)}$	$74.9_{(2.3)}$	AUC	$83.0_{(1.2)}$	$86.5_{(1.9)}$	$81.9_{(1.7)}$	$86.0_{(1.1)}$	$78.1_{(3.7)}$	$67.6_{(6.5)}$	$64.7_{(4.0)}$	$88.4_{(0.6)}$	$78.6_{(1.6)}$
			Δ	−0.8	+2.7	−1.9	+2.2	−5.7	−16.2	−19.1	+4.6	−5.2
bracs_fine	$79.2_{(1.6)}$	$70.4_{(2.3)}$	AUC	$76.5_{(2.3)}$	$77.1_{(2.9)}$	$72.8_{(6.1)}$	$71.9_{(5.9)}$	$77.5_{(3.5)}$	$69.1_{(1.9)}$	$55.0_{(1.7)}$	$80.6_{(3.1)}$	$67.5_{(1.5)}$
			Δ	−2.7	−2.1	−6.4	−7.3	−1.7	−10.1	−24.2	+1.4	−11.7
cptac_nsclc	$99.9_{(0.2)}$	$99.4_{(0.5)}$	AUC	$99.2_{(0.1)}$	$99.7_{(0.1)}$	$98.7_{(0.1)}$	$99.8_{(0.1)}$	$99.7_{(0.0)}$	$99.1_{(0.6)}$	$97.6_{(1.2)}$	$99.4_{(0.1)}$	$99.6_{(0.0)}$
			Δ	−0.7	−0.2	−1.2	−0.1	−0.2	−0.8	−2.3	−0.5	−0.3
cptac_pancancer	$99.6_{(0.0)}$	$98.8_{(0.0)}$	AUC	$99.7_{(0.1)}$	$99.6_{(0.1)}$	$99.8_{(0.0)}$	$99.6_{(0.2)}$	$99.6_{(0.1)}$	$99.1_{(0.1)}$	$99.3_{(0.1)}$	$99.8_{(0.0)}$	$99.6_{(0.1)}$
			Δ	+0.1	0.0	+0.2	0.0	0.0	−0.5	−0.3	+0.2	0.0
ebrains_idh	$96.3_{(0.4)}$	$93.2_{(0.9)}$	AUC	$97.8_{(0.4)}$	$98.0_{(0.9)}$	$98.4_{(0.3)}$	$98.5_{(0.3)}$	$98.2_{(0.4)}$	$96.6_{(0.5)}$	$96.7_{(0.6)}$	$98.6_{(0.1)}$	$98.2_{(0.2)}$
			Δ	+1.5	+1.7	+2.1	+2.2	+1.9	+0.3	+0.4	+2.3	+1.9
ebrains_coarse	$98.7_{(0.1)}$	$97.6_{(0.0)}$	AUC	$98.6_{(0.2)}$	$98.8_{(0.1)}$	$99.2_{(0.1)}$	$99.1_{(0.1)}$	$98.8_{(0.1)}$	$97.9_{(0.2)}$	$97.8_{(0.3)}$	$99.3_{(0.0)}$	$99.0_{(0.1)}$
			Δ	−0.1	+0.1	+0.5	+0.4	+0.1	−0.8	−0.9	+0.6	+0.3
ebrains_fine	$97.4_{(1.1)}$	$96.8_{(0.0)}$	AUC	$98.2_{(0.1)}$	$98.2_{(0.1)}$	$98.4_{(0.1)}$	$98.2_{(0.0)}$	$98.1_{(0.1)}$	$97.5_{(0.3)}$	$97.2_{(0.3)}$	$98.4_{(0.1)}$	$98.1_{(0.1)}$
			Δ	+0.8	+0.8	+1.0	+0.8	+0.7	+0.1	−0.2	+1.0	+0.7
kidrare_coarse	$98.1_{(0.2)}$	$96.0_{(0.1)}$	AUC	$96.8_{(0.2)}$	$97.1_{(0.1)}$	$97.0_{(0.3)}$	$96.3_{(0.1)}$	$94.3_{(0.7)}$	$93.4_{(0.6)}$	$92.2_{(0.8)}$	$97.5_{(0.2)}$	$96.3_{(0.4)}$
			Δ	−1.3	−1.0	−1.1	−1.8	−3.8	−4.7	−5.9	−0.6	−1.8
kidrare_fine	$93.4_{(0.3)}$	$91.2_{(0.2)}$	AUC	$91.6_{(0.6)}$	$92.0_{(0.4)}$	$92.6_{(0.4)}$	$92.6_{(0.5)}$	$91.0_{(0.4)}$	$84.6_{(3.0)}$	$86.3_{(1.1)}$	$92.7_{(0.5)}$	$90.3_{(0.9)}$
			Δ	−1.8	−1.4	−0.8	−0.8	−2.4	−8.8	−7.1	−0.7	−3.1
mut_het_rcc_bap1_mutation	$79.6_{(1.0)}$	$80.0_{(1.3)}$	AUC	$83.7_{(0.3)}$	$84.0_{(0.3)}$	$83.3_{(0.4)}$	$82.8_{(0.6)}$	$81.8_{(1.2)}$	$50.3_{(14.5)}$	$63.6_{(11.9)}$	$80.8_{(2.8)}$	$82.2_{(1.6)}$
			Δ	+3.7	+4.0	+3.3	+2.8	+1.8	−29.7	−16.4	+0.8	+2.2
mut_het_rcc_pbrm1_mutation	$70.3_{(0.5)}$	$69.3_{(1.3)}$	AUC	$71.8_{(1.3)}$	$70.9_{(1.4)}$	$68.5_{(0.8)}$	$70.8_{(2.6)}$	$70.7_{(0.5)}$	$62.3_{(2.6)}$	$63.4_{(1.7)}$	$69.8_{(2.7)}$	$69.3_{(2.0)}$
			Δ	+1.5	+0.6	−1.8	+0.5	+0.4	−8.0	−6.9	−0.5	−1.0
mut_het_rcc_setd2_mutation	$67.6_{(5.8)}$	$68.8_{(4.9)}$	AUC	$68.3_{(0.6)}$	$68.5_{(0.4)}$	$69.3_{(0.3)}$	$70.3_{(1.9)}$	$70.6_{(0.7)}$	$52.8_{(7.8)}$	$66.1_{(7.9)}$	$69.9_{(1.1)}$	$64.7_{(7.0)}$
			Δ	−0.5	−0.3	+0.5	+1.5	+1.8	−16.0	−2.7	+1.1	−4.1

Table 8: F1 score of teacher models and distillation-initialized MIL models under the linear-probe setting. For each dataset, Δ denotes the difference between each MIL model and the best teacher result on the same dataset. Values are reported as mean with standard deviation in subscript.

Dataset	Teacher Models		Metric	Pretrained Student MIL Aggregators								
	TITAN	CARE		ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	$47.3_{(4.8)}$	$43.9_{(0.0)}$	F1	$81.8_{(1.5)}$	$83.2_{(0.5)}$	$82.5_{(1.0)}$	$82.4_{(1.1)}$	$82.6_{(1.1)}$	$68.8_{(0.0)}$	$68.8_{(0.0)}$	$79.8_{(1.2)}$	$80.1_{(1.6)}$
			Δ	+34.5	+35.9	+35.2	+35.1	+35.3	+21.5	+21.5	+32.5	+32.8
bcnb_pr	$46.3_{(5.0)}$	$42.7_{(0.0)}$	F1	$79.0_{(0.8)}$	$78.5_{(2.7)}$	$79.1_{(1.5)}$	$77.9_{(1.8)}$	$78.7_{(1.1)}$	$63.6_{(0.0)}$	$63.6_{(0.0)}$	$76.7_{(1.1)}$	$75.7_{(0.8)}$
			Δ	+32.7	+32.2	+32.8	+31.6	+32.4	+17.3	+17.3	+30.4	+29.4
bcnb_her2	$45.4_{(3.9)}$	$42.5_{(0.0)}$	F1	$69.1_{(2.9)}$	$69.9_{(3.2)}$	$69.8_{(5.9)}$	$68.5_{(6.0)}$	$73.1_{(0.5)}$	$62.8_{(0.0)}$	$63.6_{(1.5)}$	$73.5_{(1.0)}$	$71.2_{(0.6)}$
			Δ	+23.7	+24.5	+24.4	+23.1	+27.7	+17.4	+18.2	+28.1	+25.8
bracs_coarse	$43.5_{(13.2)}$	$44.0_{(1.5)}$	F1	$56.7_{(4.4)}$	$61.8_{(3.8)}$	$53.8_{(1.1)}$	$63.3_{(2.2)}$	$52.0_{(1.9)}$	$30.2_{(11.3)}$	$40.7_{(2.8)}$	$64.8_{(3.4)}$	$49.4_{(3.8)}$
			Δ	+12.7	+17.8	+9.8	+19.3	+8.0	−13.8	−3.3	+20.8	+5.4
bracs_fine	$18.9_{(0.1)}$	$18.9_{(0.1)}$	F1	$32.4_{(4.7)}$	$30.4_{(5.8)}$	$38.2_{(0.6)}$	$32.8_{(7.0)}$	$35.4_{(6.4)}$	$22.1_{(1.1)}$	$19.7_{(1.0)}$	$42.7_{(7.9)}$	$25.8_{(0.9)}$
			Δ	+13.5	+11.5	+19.3	+13.9	+16.5	+3.2	+0.8	+23.8	+6.9
cptac_nsclc	$94.2_{(5.1)}$	$86.5_{(8.1)}$	F1	$94.3_{(0.6)}$	$96.6_{(0.5)}$	$95.8_{(0.0)}$	$98.5_{(0.6)}$	$96.5_{(0.6)}$	$93.4_{(2.7)}$	$91.8_{(1.7)}$	$93.8_{(0.4)}$	$97.0_{(1.0)}$
			Δ	+0.1	+2.4	+1.6	+4.3	+2.3	−0.8	−2.4	−0.4	+2.8
cptac_pancancer	$91.9_{(0.4)}$	$87.2_{(0.4)}$	F1	$93.1_{(1.0)}$	$92.8_{(1.7)}$	$94.3_{(0.6)}$	$93.1_{(1.4)}$	$92.6_{(0.9)}$	$89.5_{(0.4)}$	$90.1_{(0.6)}$	$95.2_{(0.3)}$	$92.4_{(1.0)}$
			Δ	+1.2	+0.9	+2.4	+1.2	+0.7	−2.4	−1.8	+3.3	+0.5
ebrains_idh	$54.2_{(16.4)}$	$44.9_{(15.4)}$	F1	$91.4_{(1.0)}$	$91.5_{(2.3)}$	$91.5_{(2.3)}$	$92.3_{(1.3)}$	$92.3_{(1.7)}$	$88.9_{(3.6)}$	$90.4_{(0.5)}$	$93.6_{(0.5)}$	$92.8_{(1.4)}$
			Δ	+37.2	+37.3	+37.3	+38.1	+38.1	+34.7	+36.2	+39.4	+38.6
ebrains_coarse	$70.7_{(4.0)}$	$52.5_{(0.6)}$	F1	$87.8_{(0.8)}$	$87.9_{(0.7)}$	$89.1_{(1.2)}$	$88.2_{(1.1)}$	$88.5_{(0.5)}$	$83.1_{(1.1)}$	$87.0_{(2.0)}$	$90.3_{(0.6)}$	$88.4_{(0.5)}$
			Δ	+17.1	+17.2	+18.4	+17.5	+17.8	+12.4	+16.3	+19.6	+17.7
ebrains_fine	$42.8_{(20.4)}$	$40.0_{(1.3)}$	F1	$69.9_{(1.1)}$	$70.5_{(0.8)}$	$72.3_{(1.8)}$	$71.3_{(2.1)}$	$69.0_{(0.8)}$	$62.3_{(3.7)}$	$64.1_{(1.8)}$	$72.8_{(1.1)}$	$69.9_{(2.0)}$
			Δ	+27.1	+27.7	+29.5	+28.5	+26.2	+19.5	+21.3	+30.0	+27.1
kidrare_coarse	$73.3_{(8.4)}$	$65.9_{(2.8)}$	F1	$83.3_{(3.3)}$	$86.0_{(1.2)}$	$84.8_{(1.2)}$	$83.7_{(0.6)}$	$76.8_{(1.5)}$	$67.2_{(2.8)}$	$69.5_{(2.0)}$	$86.3_{(2.0)}$	$80.5_{(2.8)}$
			Δ	+10.0	+12.7	+11.5	+10.4	+3.5	−6.1	−3.8	+13.0	+7.2
kidrare_fine	$26.8_{(5.5)}$	$18.9_{(0.3)}$	F1	$55.0_{(2.7)}$	$58.4_{(1.7)}$	$60.0_{(2.3)}$	$58.4_{(3.1)}$	$56.3_{(1.6)}$	$41.8_{(4.8)}$	$41.2_{(4.0)}$	$60.1_{(1.4)}$	$55.1_{(2.2)}$
			Δ	+28.2	+31.6	+33.2	+31.6	+29.5	+15.0	+14.4	+33.3	+28.3
mut_het_rcc_bap1_mutation	$46.6_{(0.0)}$	$46.6_{(0.0)}$	F1	$85.0_{(0.5)}$	$85.3_{(0.7)}$	$85.1_{(0.7)}$	$84.9_{(0.4)}$	$83.5_{(1.4)}$	$80.9_{(0.0)}$	$80.9_{(0.0)}$	$84.1_{(1.5)}$	$82.8_{(1.1)}$
			Δ	+38.4	+38.7	+38.5	+38.3	+36.9	+34.3	+34.3	+37.5	+36.2
mut_het_rcc_pbrm1_mutation	$51.0_{(10.0)}$	$42.8_{(9.3)}$	F1	$66.2_{(1.1)}$	$66.3_{(0.6)}$	$62.6_{(1.5)}$	$64.0_{(2.9)}$	$64.7_{(1.0)}$	$57.7_{(2.6)}$	$58.3_{(4.1)}$	$64.5_{(2.4)}$	$63.3_{(2.2)}$
			Δ	+15.2	+15.3	+11.6	+13.0	+13.7	+6.7	+7.3	+13.5	+12.3
mut_het_rcc_setd2_mutation	$42.8_{(1.2)}$	$42.2_{(0.0)}$	F1	$70.3_{(1.6)}$	$70.2_{(2.3)}$	$71.4_{(1.5)}$	$71.9_{(1.0)}$	$70.6_{(0.9)}$	$60.7_{(0.0)}$	$60.6_{(0.1)}$	$70.8_{(1.0)}$	$62.6_{(1.8)}$
			Δ	+27.5	+27.4	+28.6	+29.1	+27.8	+17.9	+17.8	+28.0	+19.8

A.4 Results of fine tuning

Table 9: Balanced accuracy (%) comparison between scratch-trained MIL models and pretrain-initialized MIL models. Each dataset contains three rows: Scratch, Pretrain, and Δ (Pretrain–Scratch, percentage points).

Dataset	Setting	ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	Scratch	$74.6_{(1.2)}$	$74.9_{(3.2)}$	$72.4_{(3.9)}$	$74.5_{(3.0)}$	$75.9_{(1.6)}$	$72.6_{(3.8)}$	$74.3_{(2.4)}$	$74.2_{(2.0)}$	$76.2_{(1.2)}$
	Pretrain	$74.8_{(0.9)}$	$75.5_{(2.7)}$	$75.5_{(2.1)}$	$75.8_{(4.2)}$	$76.8_{(2.1)}$	$75.2_{(1.7)}$	$74.2_{(3.2)}$	$77.5_{(4.4)}$	$76.8_{(2.4)}$
	Δ	+0.2	+0.6	+3.2	+1.2	+1.0	+2.6	−0.1	+3.2	+0.6
bcnb_her2	Scratch	$64.5_{(1.9)}$	$62.2_{(2.4)}$	$55.4_{(3.3)}$	$62.7_{(3.3)}$	$53.6_{(4.5)}$	$65.3_{(8.2)}$	$61.4_{(1.9)}$	$64.9_{(1.9)}$	$65.7_{(1.9)}$
	Pretrain	$63.7_{(2.1)}$	$65.3_{(2.3)}$	$63.5_{(4.9)}$	$59.8_{(3.8)}$	$66.1_{(1.6)}$	$67.2_{(1.0)}$	$64.2_{(0.7)}$	$66.4_{(1.1)}$	$63.0_{(1.4)}$
	Δ	−0.8	+3.1	+8.1	−2.9	+12.5	+1.9	+2.8	+1.5	−2.7
bcnb_pr	Scratch	$72.2_{(1.6)}$	$72.9_{(2.1)}$	$67.9_{(2.7)}$	$71.5_{(1.9)}$	$72.9_{(3.1)}$	$72.1_{(3.7)}$	$71.8_{(2.6)}$	$72.0_{(2.3)}$	$74.9_{(1.2)}$
	Pretrain	$73.1_{(2.3)}$	$73.1_{(1.7)}$	$70.8_{(3.0)}$	$72.5_{(2.8)}$	$73.5_{(2.0)}$	$72.1_{(1.0)}$	$73.0_{(1.9)}$	$73.8_{(1.3)}$	$73.4_{(1.7)}$
	Δ	+0.9	+0.2	+2.9	+1.0	+0.6	0.0	+1.1	+1.7	−1.6
bracs_coarse	Scratch	$61.9_{(5.7)}$	$64.7_{(1.5)}$	$62.3_{(1.1)}$	$62.4_{(7.3)}$	$67.2_{(1.9)}$	$64.1_{(5.0)}$	$68.9_{(3.5)}$	$67.5_{(5.7)}$	$66.2_{(5.0)}$
	Pretrain	$66.2_{(2.2)}$	$64.5_{(4.1)}$	$64.3_{(3.5)}$	$70.2_{(4.2)}$	$72.8_{(3.5)}$	$68.4_{(8.3)}$	$73.8_{(2.7)}$	$67.9_{(4.3)}$	$71.1_{(1.5)}$
	Δ	+4.3	−0.2	+2.1	+7.8	+5.6	+4.3	+4.9	+0.4	+5.0
bracs_fine	Scratch	$40.0_{(6.0)}$	$41.2_{(3.2)}$	$32.0_{(3.3)}$	$32.4_{(6.0)}$	$43.0_{(1.2)}$	$37.4_{(1.7)}$	$46.8_{(1.8)}$	$43.5_{(3.7)}$	$41.8_{(3.9)}$
	Pretrain	$42.9_{(2.6)}$	$41.4_{(3.6)}$	$43.6_{(3.8)}$	$35.8_{(3.7)}$	$41.3_{(6.7)}$	$45.2_{(3.2)}$	$45.9_{(3.1)}$	$41.4_{(2.8)}$	$42.7_{(2.7)}$
	Δ	+3.0	+0.2	+11.6	+3.4	−1.7	+7.9	−0.9	−2.1	+0.9
cptac_nsclc	Scratch	$95.0_{(0.7)}$	$95.6_{(1.9)}$	$96.6_{(1.5)}$	$92.0_{(1.4)}$	$93.0_{(1.8)}$	$95.0_{(0.9)}$	$95.0_{(1.2)}$	$95.1_{(1.0)}$	$96.5_{(1.5)}$
	Pretrain	$96.3_{(1.1)}$	$94.6_{(1.1)}$	$94.9_{(1.1)}$	$96.5_{(1.6)}$	$97.0_{(1.1)}$	$94.9_{(0.5)}$	$94.6_{(1.1)}$	$95.0_{(1.3)}$	$95.1_{(0.3)}$
	Δ	+1.3	−1.0	−1.7	+4.5	+4.0	0.0	−0.4	−0.2	−1.4
cptac_pancancer	Scratch	$90.7_{(0.5)}$	$91.3_{(0.4)}$	$92.4_{(1.5)}$	$91.8_{(0.7)}$	$91.7_{(2.9)}$	$91.5_{(2.8)}$	$91.8_{(0.9)}$	$94.7_{(0.9)}$	$92.6_{(1.5)}$
	Pretrain	$91.8_{(0.9)}$	$92.1_{(1.3)}$	$93.0_{(1.0)}$	$92.3_{(2.4)}$	$93.7_{(0.5)}$	$93.5_{(0.4)}$	$91.1_{(1.5)}$	$94.1_{(0.9)}$	$91.3_{(1.5)}$
	Δ	+1.1	+0.8	+0.6	+0.5	+2.0	+2.1	−0.7	−0.6	−1.2
ebrains_coarse	Scratch	$84.5_{(2.4)}$	$85.0_{(1.0)}$	$83.4_{(1.9)}$	$82.9_{(2.9)}$	$85.2_{(2.2)}$	$83.7_{(3.4)}$	$87.2_{(1.1)}$	$84.0_{(2.9)}$	$87.3_{(1.2)}$
	Pretrain	$83.9_{(2.9)}$	$84.9_{(1.1)}$	$82.7_{(2.2)}$	$84.2_{(1.5)}$	$84.8_{(3.2)}$	$82.1_{(2.2)}$	$85.8_{(2.9)}$	$85.4_{(2.5)}$	$87.5_{(1.3)}$
	Δ	−0.5	0.0	−0.7	+1.3	−0.3	−1.6	−1.4	+1.4	+0.2
ebrains_fine	Scratch	$67.0_{(0.8)}$	$66.1_{(0.9)}$	$64.7_{(2.7)}$	$65.1_{(2.0)}$	$66.1_{(1.2)}$	$63.8_{(4.0)}$	$68.3_{(1.4)}$	$70.3_{(1.6)}$	$67.5_{(1.8)}$
	Pretrain	$66.5_{(1.1)}$	$67.0_{(1.4)}$	$65.5_{(1.7)}$	$65.2_{(4.1)}$	$64.7_{(3.9)}$	$64.5_{(2.0)}$	$65.9_{(1.7)}$	$67.5_{(3.2)}$	$68.9_{(1.6)}$
	Δ	−0.5	+0.9	+0.9	+0.1	−1.4	+0.8	−2.4	−2.8	+1.4
ebrains_idh	Scratch	$92.3_{(1.3)}$	$93.0_{(2.0)}$	$90.7_{(1.7)}$	$90.7_{(2.3)}$	$90.8_{(2.6)}$	$92.2_{(0.3)}$	$92.6_{(1.4)}$	$91.3_{(1.2)}$	$92.3_{(1.1)}$
	Pretrain	$92.8_{(1.9)}$	$91.3_{(1.8)}$	$92.5_{(1.4)}$	$91.4_{(0.9)}$	$93.5_{(1.2)}$	$92.7_{(1.3)}$	$93.2_{(1.4)}$	$92.8_{(0.8)}$	$92.2_{(1.4)}$
	Δ	+0.5	−1.8	+1.7	+0.7	+2.7	+0.5	+0.5	+1.4	−0.1
kidrare_coarse	Scratch	$90.1_{(2.2)}$	$90.4_{(1.7)}$	$91.1_{(2.2)}$	$88.0_{(1.2)}$	$88.2_{(1.6)}$	$89.8_{(1.6)}$	$91.2_{(1.4)}$	$93.2_{(0.9)}$	$88.4_{(2.4)}$
	Pretrain	$91.6_{(1.5)}$	$88.2_{(2.8)}$	$92.5_{(2.1)}$	$88.7_{(1.7)}$	$87.9_{(1.1)}$	$91.5_{(1.6)}$	$89.3_{(1.6)}$	$93.0_{(1.2)}$	$91.7_{(2.6)}$
	Δ	+1.6	−2.2	+1.4	+0.7	−0.2	+1.7	−1.9	−0.3	+3.3
kidrare_fine	Scratch	$45.7_{(2.4)}$	$46.8_{(2.0)}$	$48.9_{(4.7)}$	$51.4_{(7.1)}$	$51.8_{(4.5)}$	$49.5_{(5.9)}$	$50.3_{(5.3)}$	$57.3_{(5.1)}$	$47.7_{(3.2)}$
	Pretrain	$47.6_{(1.8)}$	$47.5_{(1.9)}$	$53.3_{(2.2)}$	$48.6_{(4.7)}$	$46.8_{(4.5)}$	$51.0_{(5.5)}$	$52.9_{(5.9)}$	$54.1_{(4.3)}$	$48.5_{(3.0)}$
	Δ	+1.9	+0.7	+4.4	−2.8	−5.0	+1.5	+2.6	−3.2	+0.8
mut_het_rcc_bap1_mutation	Scratch	$63.7_{(4.2)}$	$65.0_{(5.4)}$	$62.7_{(4.6)}$	$64.9_{(3.3)}$	$58.6_{(7.6)}$	$52.9_{(3.5)}$	$66.6_{(5.9)}$	$67.1_{(3.7)}$	$69.5_{(6.3)}$
	Pretrain	$62.3_{(5.7)}$	$69.6_{(4.6)}$	$70.2_{(1.7)}$	$62.6_{(8.2)}$	$62.7_{(6.8)}$	$62.0_{(9.3)}$	$67.1_{(6.5)}$	$66.5_{(5.7)}$	$67.5_{(4.4)}$
	Δ	−1.5	+4.6	+7.5	−2.2	+3.9	+9.1	+0.5	−0.6	−2.0
mut_het_rcc_pbrm1_mutation	Scratch	$62.9_{(2.7)}$	$63.7_{(1.4)}$	$57.8_{(3.2)}$	$64.3_{(1.7)}$	$63.6_{(2.2)}$	$60.7_{(1.8)}$	$64.6_{(2.1)}$	$65.8_{(1.7)}$	$69.2_{(0.8)}$
	Pretrain	$64.5_{(2.6)}$	$67.9_{(2.1)}$	$63.3_{(2.9)}$	$60.5_{(2.8)}$	$62.5_{(1.4)}$	$62.4_{(2.9)}$	$67.9_{(1.6)}$	$69.2_{(1.2)}$	$68.7_{(1.8)}$
	Δ	+1.6	+4.3	+5.5	−3.8	−1.0	+1.7	+3.3	+3.4	−0.5
mut_het_rcc_setd2_mutation	Scratch	$62.5_{(1.9)}$	$62.2_{(2.0)}$	$57.5_{(4.1)}$	$57.6_{(6.3)}$	$59.6_{(4.9)}$	$63.4_{(2.5)}$	$63.5_{(1.8)}$	$62.9_{(3.2)}$	$62.2_{(3.9)}$
	Pretrain	$63.6_{(2.4)}$	$63.9_{(2.6)}$	$61.2_{(1.5)}$	$59.4_{(7.8)}$	$57.4_{(6.4)}$	$65.3_{(3.9)}$	$66.4_{(2.0)}$	$63.5_{(3.0)}$	$62.9_{(3.1)}$
	Δ	+1.2	+1.7	+3.7	+1.7	−2.2	+1.9	+2.9	+0.6	+0.7

Table 10: AUC comparison between scratch-trained MIL models and pretrain-initialized MIL models. Each dataset contains Scratch, Pretrain, and Δ rows. Values are reported as AUC (%) with one decimal place.

Dataset	Setting	ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	Scratch	$84.8_{(2.6)}$	$85.6_{(2.0)}$	$84.0_{(2.0)}$	$83.7_{(2.9)}$	$85.8_{(1.4)}$	$84.4_{(1.1)}$	$85.8_{(0.9)}$	$84.7_{(1.4)}$	$86.0_{(1.2)}$
	Pretrain	$86.1_{(1.8)}$	$84.8_{(2.1)}$	$84.2_{(1.5)}$	$84.4_{(2.5)}$	$85.9_{(0.6)}$	$86.3_{(1.3)}$	$86.2_{(1.5)}$	$86.3_{(1.9)}$	$87.0_{(1.8)}$
	Δ	+1.3	−0.8	+0.2	+0.7	+0.1	+1.9	+0.4	+1.6	+1.0
bcnb_her2	Scratch	$71.0_{(1.2)}$	$70.5_{(3.5)}$	$59.1_{(4.6)}$	$69.4_{(2.8)}$	$65.3_{(3.2)}$	$72.4_{(6.6)}$	$66.8_{(1.6)}$	$72.4_{(0.6)}$	$74.6_{(1.4)}$
	Pretrain	$70.8_{(0.5)}$	$70.9_{(2.3)}$	$68.5_{(4.9)}$	$68.7_{(1.9)}$	$71.8_{(1.1)}$	$73.2_{(0.3)}$	$69.9_{(1.6)}$	$73.7_{(0.7)}$	$71.9_{(2.5)}$
	Δ	−0.3	+0.4	+9.5	−0.7	+6.4	+0.8	+3.1	+1.3	−2.7
bcnb_pr	Scratch	$78.5_{(1.1)}$	$79.5_{(2.8)}$	$75.3_{(2.5)}$	$75.4_{(2.2)}$	$77.7_{(0.6)}$	$77.5_{(2.8)}$	$79.0_{(2.8)}$	$79.3_{(1.0)}$	$79.2_{(0.4)}$
	Pretrain	$79.2_{(0.6)}$	$79.2_{(1.2)}$	$78.2_{(1.8)}$	$77.1_{(2.4)}$	$79.3_{(1.6)}$	$78.4_{(0.7)}$	$80.3_{(1.9)}$	$79.5_{(1.2)}$	$79.7_{(0.5)}$
	Δ	+0.7	−0.3	+2.9	+1.7	+1.6	+0.9	+1.3	+0.1	+0.5
bracs_coarse	Scratch	$84.1_{(5.7)}$	$86.7_{(1.6)}$	$82.7_{(1.5)}$	$85.6_{(3.2)}$	$88.1_{(1.3)}$	$85.8_{(3.0)}$	$89.6_{(0.2)}$	$87.6_{(1.6)}$	$88.6_{(1.2)}$
	Pretrain	$88.6_{(2.5)}$	$87.2_{(1.6)}$	$85.7_{(2.1)}$	$88.6_{(2.4)}$	$89.5_{(1.2)}$	$87.5_{(2.5)}$	$91.5_{(0.9)}$	$89.8_{(1.6)}$	$89.5_{(0.9)}$
	Δ	+4.5	+0.5	+3.1	+3.0	+1.4	+1.7	+1.9	+2.2	+0.9
bracs_fine	Scratch	$78.8_{(5.0)}$	$80.5_{(2.6)}$	$75.8_{(3.0)}$	$75.3_{(4.0)}$	$82.6_{(1.2)}$	$82.7_{(1.7)}$	$83.8_{(0.9)}$	$82.9_{(0.8)}$	$80.4_{(1.2)}$
	Pretrain	$82.6_{(1.4)}$	$80.4_{(1.2)}$	$81.4_{(1.6)}$	$81.9_{(1.6)}$	$81.5_{(2.0)}$	$83.0_{(1.1)}$	$82.7_{(1.6)}$	$82.8_{(0.8)}$	$81.3_{(1.2)}$
	Δ	+3.8	−0.1	+5.7	+6.6	−1.1	+0.4	−1.1	−0.1	+0.8
cptac_nsclc	Scratch	$99.3_{(0.8)}$	$98.3_{(1.6)}$	$99.8_{(0.2)}$	$98.9_{(0.2)}$	$99.1_{(0.5)}$	$99.5_{(0.1)}$	$99.5_{(0.2)}$	$99.6_{(0.2)}$	$99.8_{(0.1)}$
	Pretrain	$99.5_{(0.5)}$	$98.7_{(1.1)}$	$99.7_{(0.3)}$	$99.8_{(0.1)}$	$99.7_{(0.1)}$	$99.9_{(0.0)}$	$99.6_{(0.1)}$	$99.5_{(0.4)}$	$99.8_{(0.1)}$
	Δ	+0.2	+0.4	−0.1	+0.9	+0.6	+0.3	+0.1	−0.1	+0.0
cptac_pancancer	Scratch	$99.6_{(0.1)}$	$99.5_{(0.1)}$	$99.7_{(0.1)}$	$99.6_{(0.1)}$	$99.6_{(0.1)}$	$99.7_{(0.1)}$	$99.6_{(0.1)}$	$99.8_{(0.1)}$	$99.7_{(0.1)}$
	Pretrain	$99.7_{(0.1)}$	$99.5_{(0.1)}$	$99.7_{(0.1)}$	$99.8_{(0.1)}$	$99.7_{(0.1)}$	$99.7_{(0.0)}$	$99.6_{(0.1)}$	$99.8_{(0.0)}$	$99.7_{(0.1)}$
	Δ	+0.1	+0.0	0.0	+0.2	+0.2	+0.0	0.0	−0.1	0.0
ebrains_coarse	Scratch	$99.1_{(0.1)}$	$99.1_{(0.1)}$	$98.8_{(0.1)}$	$98.8_{(0.2)}$	$99.0_{(0.2)}$	$99.0_{(0.3)}$	$99.1_{(0.1)}$	$99.1_{(0.2)}$	$99.3_{(0.1)}$
	Pretrain	$99.1_{(0.1)}$	$99.1_{(0.1)}$	$98.8_{(0.3)}$	$99.1_{(0.1)}$	$99.0_{(0.3)}$	$98.9_{(0.2)}$	$99.1_{(0.2)}$	$99.0_{(0.1)}$	$99.2_{(0.2)}$
	Δ	+0.0	+0.0	+0.0	+0.2	0.0	−0.1	+0.0	−0.1	−0.2
ebrains_fine	Scratch	$98.3_{(0.0)}$	$98.2_{(0.1)}$	$97.6_{(0.3)}$	$98.0_{(0.2)}$	$98.2_{(0.0)}$	$97.8_{(0.3)}$	$98.3_{(0.2)}$	$98.4_{(0.1)}$	$98.3_{(0.1)}$
	Pretrain	$98.4_{(0.0)}$	$98.3_{(0.1)}$	$98.2_{(0.1)}$	$98.2_{(0.2)}$	$98.2_{(0.2)}$	$98.0_{(0.2)}$	$98.2_{(0.1)}$	$98.2_{(0.3)}$	$98.4_{(0.1)}$
	Δ	+0.1	+0.1	+0.6	+0.2	+0.1	+0.1	−0.1	−0.2	+0.0
ebrains_idh	Scratch	$98.1_{(0.5)}$	$98.2_{(0.3)}$	$97.6_{(0.3)}$	$97.5_{(1.0)}$	$97.7_{(1.2)}$	$98.3_{(0.1)}$	$98.3_{(0.5)}$	$98.0_{(0.3)}$	$98.2_{(0.3)}$
	Pretrain	$98.2_{(0.4)}$	$97.5_{(0.4)}$	$98.4_{(0.2)}$	$98.2_{(0.3)}$	$98.5_{(0.4)}$	$98.5_{(0.2)}$	$98.9_{(0.2)}$	$98.4_{(0.1)}$	$98.6_{(0.4)}$
	Δ	+0.1	−0.7	+0.8	+0.7	+0.8	+0.2	+0.6	+0.4	+0.4
kidrare_coarse	Scratch	$98.7_{(0.1)}$	$98.6_{(0.2)}$	$98.8_{(0.3)}$	$98.2_{(0.2)}$	$97.9_{(0.3)}$	$98.4_{(0.4)}$	$98.8_{(0.1)}$	$98.8_{(0.4)}$	$98.5_{(0.1)}$
	Pretrain	$98.9_{(0.3)}$	$98.3_{(0.7)}$	$99.0_{(0.2)}$	$98.3_{(0.3)}$	$97.5_{(0.2)}$	$98.9_{(0.1)}$	$98.9_{(0.2)}$	$98.9_{(0.2)}$	$98.8_{(0.4)}$
	Δ	+0.2	−0.3	+0.2	+0.1	−0.4	+0.5	+0.1	+0.1	+0.3
kidrare_fine	Scratch	$94.4_{(0.4)}$	$94.0_{(0.4)}$	$92.8_{(1.7)}$	$93.6_{(0.7)}$	$94.1_{(0.7)}$	$94.5_{(0.7)}$	$94.2_{(0.6)}$	$95.7_{(0.3)}$	$94.4_{(0.5)}$
	Pretrain	$94.7_{(0.5)}$	$94.2_{(0.6)}$	$95.2_{(1.0)}$	$95.0_{(0.4)}$	$94.0_{(0.4)}$	$94.1_{(0.9)}$	$93.8_{(0.9)}$	$96.0_{(0.4)}$	$94.2_{(0.6)}$
	Δ	+0.3	+0.2	+2.4	+1.4	−0.1	−0.3	−0.3	+0.3	−0.1
mut_het_rcc_bap1_mutation	Scratch	$80.9_{(0.8)}$	$82.1_{(1.7)}$	$80.6_{(4.9)}$	$84.2_{(0.6)}$	$83.8_{(0.7)}$	$83.1_{(0.8)}$	$82.5_{(1.9)}$	$84.2_{(0.2)}$	$82.5_{(3.0)}$
	Pretrain	$81.3_{(1.1)}$	$80.6_{(3.3)}$	$81.9_{(2.0)}$	$82.4_{(1.1)}$	$81.8_{(1.4)}$	$82.8_{(1.7)}$	$83.3_{(1.1)}$	$83.6_{(1.3)}$	$84.4_{(0.6)}$
	Δ	+0.3	−1.5	+1.4	−1.8	−2.0	−0.2	+0.7	−0.6	+1.9
mut_het_rcc_pbrm1_mutation	Scratch	$69.7_{(3.5)}$	$67.9_{(0.9)}$	$61.2_{(1.2)}$	$69.4_{(2.1)}$	$71.1_{(1.2)}$	$66.8_{(3.0)}$	$71.3_{(1.4)}$	$73.2_{(1.7)}$	$75.6_{(0.9)}$
	Pretrain	$69.5_{(3.9)}$	$72.3_{(2.1)}$	$68.7_{(1.7)}$	$67.8_{(3.0)}$	$69.7_{(2.0)}$	$69.5_{(3.3)}$	$74.9_{(1.5)}$	$76.3_{(1.6)}$	$75.1_{(2.1)}$
	Δ	−0.1	+4.4	+7.5	−1.6	−1.4	+2.6	+3.6	+3.1	−0.5
mut_het_rcc_setd2_mutation	Scratch	$70.4_{(1.1)}$	$71.4_{(1.6)}$	$61.7_{(5.2)}$	$69.8_{(0.4)}$	$69.0_{(1.2)}$	$70.7_{(1.0)}$	$69.8_{(1.7)}$	$70.5_{(3.2)}$	$70.2_{(2.0)}$
	Pretrain	$71.0_{(1.2)}$	$71.1_{(1.6)}$	$69.7_{(2.2)}$	$69.7_{(1.9)}$	$69.2_{(3.4)}$	$73.7_{(1.0)}$	$72.4_{(0.4)}$	$72.3_{(2.3)}$	$69.1_{(2.6)}$
	Δ	+0.6	−0.3	+8.0	−0.1	+0.1	+2.9	+2.7	+1.8	−1.1

Table 11: F1 comparison between scratch-trained MIL models and pretrain-initialized MIL models. Each dataset contains Scratch, Pretrain, and Δ rows. Values are reported as F1 (%) with one decimal place.

Dataset	Setting	ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	Scratch	$81.8_{(2.1)}$	$82.5_{(1.8)}$	$77.9_{(3.1)}$	$80.5_{(3.3)}$	$78.8_{(0.8)}$	$80.0_{(0.8)}$	$82.7_{(0.9)}$	$81.0_{(1.2)}$	$81.0_{(2.8)}$
	Pretrain	$81.6_{(3.1)}$	$81.2_{(1.7)}$	$79.8_{(2.1)}$	$79.9_{(2.2)}$	$81.8_{(2.2)}$	$82.1_{(1.7)}$	$82.6_{(1.5)}$	$81.4_{(2.7)}$	$83.2_{(1.6)}$
	Δ	−0.2	−1.3	+1.8	−0.6	+3.0	+2.1	−0.1	+0.3	+2.2
bcnb_her2	Scratch	$72.9_{(2.0)}$	$71.5_{(2.2)}$	$66.8_{(2.1)}$	$70.6_{(2.4)}$	$65.9_{(3.5)}$	$70.9_{(5.6)}$	$65.9_{(5.0)}$	$67.2_{(8.4)}$	$71.3_{(2.5)}$
	Pretrain	$73.5_{(2.1)}$	$74.0_{(1.5)}$	$72.7_{(3.0)}$	$69.0_{(3.8)}$	$69.5_{(4.1)}$	$73.7_{(2.0)}$	$66.5_{(3.4)}$	$69.7_{(3.2)}$	$69.7_{(2.6)}$
	Δ	+0.6	+2.4	+5.9	−1.5	+3.5	+2.8	+0.6	+2.5	−1.6
bcnb_pr	Scratch	$79.0_{(2.0)}$	$79.4_{(1.7)}$	$74.3_{(1.9)}$	$77.5_{(1.6)}$	$74.8_{(6.1)}$	$75.8_{(4.6)}$	$76.2_{(4.9)}$	$77.1_{(3.8)}$	$78.3_{(1.5)}$
	Pretrain	$79.1_{(1.8)}$	$78.6_{(1.5)}$	$75.2_{(3.3)}$	$77.5_{(2.4)}$	$73.9_{(2.9)}$	$77.5_{(2.4)}$	$78.0_{(3.4)}$	$78.0_{(1.8)}$	$78.0_{(1.4)}$
	Δ	+0.2	−0.8	+0.8	+0.1	−0.8	+1.7	+1.8	+0.9	−0.3
bracs_coarse	Scratch	$60.8_{(7.0)}$	$63.7_{(2.6)}$	$64.3_{(1.1)}$	$62.8_{(8.3)}$	$68.7_{(2.0)}$	$65.7_{(5.5)}$	$69.7_{(4.2)}$	$68.7_{(6.3)}$	$67.1_{(5.2)}$
	Pretrain	$66.8_{(2.7)}$	$63.8_{(5.8)}$	$65.9_{(4.4)}$	$71.0_{(5.1)}$	$74.5_{(2.9)}$	$69.5_{(8.8)}$	$74.9_{(3.8)}$	$68.6_{(6.2)}$	$72.9_{(1.5)}$
	Δ	+6.1	+0.0	+1.6	+8.1	+5.8	+3.8	+5.1	0.0	+5.8
bracs_fine	Scratch	$40.7_{(6.6)}$	$42.0_{(3.3)}$	$35.8_{(3.5)}$	$33.4_{(7.0)}$	$42.9_{(1.5)}$	$38.9_{(2.1)}$	$48.3_{(0.5)}$	$45.3_{(4.4)}$	$45.2_{(2.6)}$
	Pretrain	$44.0_{(3.0)}$	$42.3_{(2.7)}$	$45.6_{(2.9)}$	$37.9_{(4.0)}$	$42.8_{(8.0)}$	$46.1_{(3.6)}$	$49.6_{(3.6)}$	$45.7_{(4.0)}$	$47.7_{(3.4)}$
	Δ	+3.3	+0.3	+9.9	+4.5	−0.1	+7.2	+1.3	+0.4	+2.5
cptac_nsclc	Scratch	$95.0_{(0.7)}$	$95.7_{(1.9)}$	$96.7_{(1.5)}$	$92.0_{(1.4)}$	$93.0_{(1.8)}$	$95.0_{(0.9)}$	$95.0_{(1.2)}$	$95.2_{(1.0)}$	$96.5_{(1.4)}$
	Pretrain	$96.3_{(1.1)}$	$94.7_{(1.1)}$	$95.0_{(1.1)}$	$96.5_{(1.6)}$	$97.0_{(1.1)}$	$95.0_{(0.5)}$	$94.7_{(1.1)}$	$95.0_{(1.3)}$	$95.2_{(0.3)}$
	Δ	+1.3	−1.0	−1.7	+4.5	+4.0	0.0	−0.3	−0.2	−1.3
cptac_pancancer	Scratch	$92.2_{(0.5)}$	$92.3_{(0.8)}$	$93.6_{(1.8)}$	$93.4_{(0.9)}$	$92.6_{(2.8)}$	$92.6_{(2.1)}$	$92.2_{(0.7)}$	$94.9_{(0.9)}$	$93.2_{(1.2)}$
	Pretrain	$92.9_{(1.0)}$	$93.3_{(1.0)}$	$93.8_{(0.7)}$	$93.3_{(1.7)}$	$94.3_{(0.5)}$	$94.2_{(0.4)}$	$92.2_{(1.5)}$	$94.5_{(1.3)}$	$92.5_{(1.5)}$
	Δ	+0.7	+1.0	+0.1	−0.1	+1.7	+1.6	0.0	−0.4	−0.6
ebrains_coarse	Scratch	$90.4_{(1.5)}$	$91.0_{(0.8)}$	$88.2_{(1.8)}$	$87.3_{(1.4)}$	$88.1_{(1.7)}$	$88.6_{(1.5)}$	$90.2_{(1.4)}$	$88.9_{(0.8)}$	$89.9_{(1.7)}$
	Pretrain	$90.6_{(1.6)}$	$91.7_{(1.0)}$	$86.8_{(1.1)}$	$88.7_{(2.1)}$	$87.6_{(3.7)}$	$87.6_{(0.6)}$	$88.8_{(2.8)}$	$88.9_{(1.1)}$	$89.2_{(2.4)}$
	Δ	+0.3	+0.7	−1.5	+1.5	−0.5	−0.9	−1.4	+0.1	−0.7
ebrains_fine	Scratch	$72.0_{(1.0)}$	$70.7_{(1.4)}$	$69.7_{(2.2)}$	$69.5_{(1.5)}$	$70.3_{(1.3)}$	$67.8_{(3.7)}$	$71.6_{(1.1)}$	$73.2_{(1.4)}$	$71.9_{(2.2)}$
	Pretrain	$71.9_{(0.6)}$	$71.8_{(0.7)}$	$70.1_{(1.1)}$	$69.0_{(2.8)}$	$68.8_{(3.7)}$	$68.8_{(1.9)}$	$70.6_{(2.0)}$	$71.5_{(2.8)}$	$73.2_{(1.5)}$
	Δ	−0.1	+1.1	+0.4	−0.5	−1.5	+1.0	−1.1	−1.6	+1.3
ebrains_idh	Scratch	$92.5_{(1.2)}$	$93.2_{(1.5)}$	$90.9_{(1.8)}$	$90.6_{(2.8)}$	$90.3_{(2.9)}$	$91.8_{(0.2)}$	$92.4_{(1.8)}$	$90.8_{(1.1)}$	$92.4_{(1.2)}$
	Pretrain	$92.8_{(1.8)}$	$91.5_{(1.6)}$	$91.5_{(1.5)}$	$91.3_{(1.2)}$	$93.4_{(1.7)}$	$92.6_{(0.7)}$	$93.2_{(1.3)}$	$92.8_{(1.1)}$	$92.0_{(1.1)}$
	Δ	+0.2	−1.7	+0.6	+0.7	+3.1	+0.8	+0.8	+2.0	−0.4
kidrare_coarse	Scratch	$92.1_{(1.7)}$	$92.2_{(1.4)}$	$91.8_{(1.8)}$	$89.2_{(0.8)}$	$90.0_{(1.6)}$	$91.7_{(1.4)}$	$91.8_{(0.7)}$	$93.6_{(0.8)}$	$90.1_{(2.3)}$
	Pretrain	$92.7_{(1.2)}$	$90.7_{(2.1)}$	$92.0_{(2.6)}$	$90.2_{(1.0)}$	$89.1_{(1.7)}$	$92.0_{(1.0)}$	$90.6_{(1.5)}$	$93.0_{(0.7)}$	$92.0_{(2.8)}$
	Δ	+0.6	−1.5	+0.2	+1.1	−0.9	+0.4	−1.2	−0.6	+1.9
kidrare_fine	Scratch	$64.7_{(1.7)}$	$64.7_{(1.3)}$	$60.9_{(4.2)}$	$62.8_{(5.4)}$	$61.3_{(2.2)}$	$60.7_{(4.9)}$	$64.4_{(3.1)}$	$69.5_{(1.2)}$	$63.9_{(3.1)}$
	Pretrain	$65.6_{(1.2)}$	$64.8_{(2.2)}$	$67.3_{(1.7)}$	$63.0_{(3.4)}$	$60.7_{(1.1)}$	$64.0_{(2.5)}$	$65.5_{(3.0)}$	$67.2_{(2.1)}$	$64.6_{(3.2)}$
	Δ	+0.9	+0.1	+6.4	+0.1	−0.7	+3.3	+1.1	−2.4	+0.6
mut_het_rcc_bap1_mutation	Scratch	$84.5_{(1.3)}$	$84.1_{(0.2)}$	$84.6_{(2.4)}$	$85.5_{(0.9)}$	$84.0_{(1.9)}$	$82.1_{(0.9)}$	$84.6_{(1.1)}$	$85.7_{(0.7)}$	$84.8_{(1.6)}$
	Pretrain	$84.3_{(1.5)}$	$84.6_{(1.2)}$	$84.6_{(2.1)}$	$84.1_{(1.0)}$	$84.7_{(1.8)}$	$83.8_{(1.5)}$	$83.7_{(1.2)}$	$85.0_{(0.5)}$	$85.0_{(1.5)}$
	Δ	−0.2	+0.5	+0.0	−1.4	+0.7	+1.7	−0.9	−0.6	+0.2
mut_het_rcc_pbrm1_mutation	Scratch	$62.8_{(2.9)}$	$63.5_{(1.3)}$	$57.6_{(3.3)}$	$64.3_{(1.7)}$	$63.5_{(2.1)}$	$60.5_{(1.9)}$	$64.6_{(2.0)}$	$65.6_{(2.2)}$	$69.1_{(0.9)}$
	Pretrain	$64.6_{(2.5)}$	$68.0_{(2.1)}$	$62.6_{(3.2)}$	$60.1_{(3.0)}$	$62.5_{(1.4)}$	$62.3_{(2.8)}$	$67.8_{(1.4)}$	$68.9_{(1.5)}$	$68.7_{(1.9)}$
	Δ	+1.8	+4.4	+5.0	−4.3	−1.0	+1.8	+3.2	+3.3	−0.4
mut_het_rcc_setd2_mutation	Scratch	$68.7_{(6.0)}$	$70.9_{(2.7)}$	$66.1_{(4.6)}$	$62.1_{(6.6)}$	$67.4_{(4.7)}$	$71.5_{(1.6)}$	$68.5_{(5.1)}$	$70.3_{(3.4)}$	$67.1_{(5.0)}$
	Pretrain	$69.7_{(4.3)}$	$71.5_{(3.0)}$	$68.7_{(3.4)}$	$67.8_{(4.6)}$	$64.9_{(3.9)}$	$71.7_{(4.2)}$	$70.3_{(4.5)}$	$70.6_{(4.2)}$	$70.3_{(2.2)}$
	Δ	+1.0	+0.6	+2.6	+5.7	−2.5	+0.2	+1.8	+0.3	+3.2

Table 12: Accuracy comparison between scratch-trained MIL models and pretrain-initialized MIL models. Each dataset contains Scratch, Pretrain, and Δ rows. Values are reported as Accuracy (%) with one decimal place.

Dataset	Setting	ABMIL	CLAM	TransMIL	WiKG	DAGMIL	GDFMIL	2DMamba	AMDMIL	AEMMIL
bcnb_er	Scratch	$81.4_{(2.6)}$	$82.4_{(1.8)}$	$77.2_{(4.3)}$	$79.8_{(4.0)}$	$77.5_{(1.1)}$	$79.7_{(1.7)}$	$82.8_{(0.9)}$	$80.6_{(1.4)}$	$80.3_{(3.7)}$
	Pretrain	$81.2_{(3.9)}$	$80.6_{(2.0)}$	$78.7_{(2.7)}$	$78.8_{(2.3)}$	$81.1_{(3.0)}$	$81.8_{(2.4)}$	$82.7_{(1.4)}$	$80.5_{(2.9)}$	$82.9_{(1.9)}$
	Δ	−0.2	−1.8	+1.5	−1.0	+3.7	+2.1	−0.1	−0.1	+2.6
bcnb_her2	Scratch	$73.1_{(2.9)}$	$71.9_{(2.5)}$	$68.1_{(2.0)}$	$70.9_{(4.1)}$	$73.6_{(2.0)}$	$72.2_{(4.6)}$	$64.5_{(5.9)}$	$66.2_{(8.5)}$	$70.7_{(3.3)}$
	Pretrain	$74.5_{(2.7)}$	$74.3_{(1.6)}$	$73.6_{(2.6)}$	$71.4_{(5.9)}$	$68.1_{(5.1)}$	$73.3_{(2.9)}$	$64.6_{(4.3)}$	$68.2_{(3.8)}$	$69.0_{(3.7)}$
	Δ	+1.4	+2.4	+5.5	+0.5	−5.5	+1.1	+0.1	+2.0	−1.7
bcnb_pr	Scratch	$79.0_{(2.4)}$	$79.4_{(1.9)}$	$74.0_{(2.5)}$	$77.2_{(1.8)}$	$73.6_{(6.7)}$	$74.8_{(5.2)}$	$75.5_{(5.8)}$	$76.5_{(4.5)}$	$77.5_{(1.8)}$
	Pretrain	$79.0_{(2.0)}$	$78.3_{(1.8)}$	$74.5_{(4.2)}$	$77.1_{(2.8)}$	$72.3_{(3.4)}$	$77.2_{(3.3)}$	$77.5_{(4.3)}$	$77.3_{(2.1)}$	$77.5_{(1.9)}$
	Δ	0.0	−1.1	+0.6	−0.1	−1.3	+2.4	+2.1	+0.7	0.0
bracs_coarse	Scratch	$67.1_{(5.6)}$	$70.1_{(1.0)}$	$65.1_{(1.9)}$	$65.8_{(6.8)}$	$70.6_{(2.0)}$	$66.4_{(5.7)}$	$70.3_{(4.1)}$	$70.1_{(4.7)}$	$69.9_{(3.9)}$
	Pretrain	$70.8_{(1.9)}$	$69.4_{(3.0)}$	$67.6_{(2.7)}$	$73.3_{(2.9)}$	$74.6_{(2.5)}$	$70.8_{(7.3)}$	$75.4_{(3.5)}$	$70.1_{(2.4)}$	$73.1_{(1.2)}$
	Δ	+3.7	−0.7	+2.5	+7.6	+4.0	+4.4	+5.1	0.0	+3.2
bracs_fine	Scratch	$46.0_{(5.5)}$	$46.9_{(4.6)}$	$40.0_{(3.9)}$	$39.1_{(7.3)}$	$49.0_{(2.0)}$	$44.1_{(2.8)}$	$51.9_{(1.1)}$	$50.1_{(4.0)}$	$47.6_{(2.5)}$
	Pretrain	$48.5_{(2.5)}$	$47.4_{(3.8)}$	$50.1_{(4.5)}$	$42.1_{(4.6)}$	$46.2_{(6.0)}$	$51.9_{(3.8)}$	$52.2_{(2.8)}$	$48.5_{(5.0)}$	$50.3_{(2.3)}$
	Δ	+2.5	+0.5	+10.1	+3.0	−2.7	+7.8	+0.2	−1.6	+2.7
cptac_nsclc	Scratch	$95.0_{(0.8)}$	$95.7_{(1.9)}$	$96.7_{(1.5)}$	$92.0_{(1.4)}$	$93.0_{(1.8)}$	$95.0_{(0.9)}$	$95.0_{(1.2)}$	$95.2_{(1.0)}$	$96.5_{(1.4)}$
	Pretrain	$96.3_{(1.1)}$	$94.7_{(1.1)}$	$95.0_{(1.1)}$	$96.5_{(1.6)}$	$97.0_{(1.1)}$	$95.0_{(0.5)}$	$94.7_{(1.1)}$	$95.0_{(1.3)}$	$95.2_{(0.3)}$
	Δ	+1.3	−1.0	−1.7	+4.5	+4.0	0.0	−0.3	−0.2	−1.3
cptac_pancancer	Scratch	$92.1_{(0.5)}$	$92.3_{(0.7)}$	$93.6_{(1.9)}$	$93.4_{(0.9)}$	$92.5_{(2.8)}$	$92.6_{(2.1)}$	$92.2_{(0.8)}$	$94.9_{(0.9)}$	$93.1_{(1.3)}$
	Pretrain	$92.9_{(1.0)}$	$93.3_{(1.0)}$	$93.7_{(0.7)}$	$93.3_{(1.7)}$	$94.3_{(0.5)}$	$94.2_{(0.5)}$	$92.2_{(1.5)}$	$94.5_{(1.4)}$	$92.5_{(1.5)}$
	Δ	+0.7	+1.1	+0.1	−0.1	+1.8	+1.6	0.0	−0.4	−0.6
ebrains_coarse	Scratch	$90.3_{(1.6)}$	$91.1_{(0.9)}$	$87.6_{(2.4)}$	$86.7_{(2.0)}$	$87.6_{(1.8)}$	$88.2_{(1.9)}$	$89.7_{(1.6)}$	$88.8_{(0.9)}$	$89.3_{(2.5)}$
	Pretrain	$90.8_{(1.6)}$	$91.9_{(1.1)}$	$86.5_{(1.3)}$	$88.6_{(2.7)}$	$86.6_{(4.5)}$	$87.5_{(0.8)}$	$88.1_{(3.1)}$	$88.4_{(1.1)}$	$88.6_{(3.1)}$
	Δ	+0.4	+0.8	−1.2	+1.9	−1.0	−0.7	−1.6	−0.4	−0.7
ebrains_fine	Scratch	$73.8_{(0.6)}$	$73.0_{(1.1)}$	$71.2_{(2.1)}$	$70.4_{(1.5)}$	$71.2_{(1.5)}$	$69.6_{(2.6)}$	$73.3_{(1.1)}$	$74.2_{(1.5)}$	$73.5_{(1.5)}$
	Pretrain	$73.5_{(0.4)}$	$73.4_{(0.9)}$	$71.9_{(0.8)}$	$71.1_{(2.0)}$	$70.9_{(2.7)}$	$71.0_{(1.4)}$	$72.6_{(1.2)}$	$72.7_{(2.3)}$	$74.5_{(1.0)}$
	Δ	−0.2	+0.3	+0.7	+0.7	−0.4	+1.4	−0.8	−1.4	+1.0
ebrains_idh	Scratch	$92.5_{(1.2)}$	$93.2_{(1.5)}$	$90.9_{(1.9)}$	$90.5_{(2.8)}$	$90.2_{(2.9)}$	$91.8_{(0.2)}$	$92.4_{(1.8)}$	$90.8_{(1.1)}$	$92.4_{(1.2)}$
	Pretrain	$92.7_{(1.8)}$	$91.5_{(1.6)}$	$91.5_{(1.5)}$	$91.2_{(1.3)}$	$93.4_{(1.7)}$	$92.6_{(0.7)}$	$93.2_{(1.3)}$	$92.8_{(1.1)}$	$91.9_{(1.1)}$
	Δ	+0.2	−1.7	+0.6	+0.7	+3.2	+0.8	+0.8	+2.1	−0.5
kidrare_coarse	Scratch	$92.2_{(1.6)}$	$92.2_{(1.4)}$	$91.8_{(1.8)}$	$89.1_{(0.8)}$	$90.0_{(1.6)}$	$91.7_{(1.3)}$	$91.8_{(0.7)}$	$93.6_{(0.8)}$	$90.2_{(2.3)}$
	Pretrain	$92.7_{(1.2)}$	$90.8_{(2.0)}$	$92.0_{(2.5)}$	$90.2_{(1.1)}$	$89.0_{(1.8)}$	$92.0_{(1.1)}$	$90.6_{(1.5)}$	$93.0_{(0.7)}$	$92.0_{(2.7)}$
	Δ	+0.5	−1.4	+0.2	+1.1	−1.0	+0.2	−1.2	−0.5	+1.9
kidrare_fine	Scratch	$66.4_{(1.4)}$	$65.8_{(0.6)}$	$62.1_{(5.0)}$	$63.6_{(5.5)}$	$62.6_{(1.4)}$	$62.1_{(4.4)}$	$65.2_{(2.9)}$	$69.9_{(1.3)}$	$64.8_{(3.0)}$
	Pretrain	$66.3_{(1.2)}$	$65.7_{(2.3)}$	$68.8_{(1.9)}$	$63.7_{(2.3)}$	$62.3_{(0.6)}$	$65.6_{(1.6)}$	$66.6_{(2.9)}$	$68.4_{(1.8)}$	$65.2_{(2.9)}$
	Δ	−0.1	−0.1	+6.7	+0.2	−0.4	+3.5	+1.3	−1.6	+0.3
mut_het_rcc_bap1_mutation	Scratch	$85.0_{(1.7)}$	$84.2_{(1.0)}$	$85.5_{(3.5)}$	$86.1_{(0.7)}$	$86.7_{(0.5)}$	$86.4_{(1.1)}$	$84.6_{(1.7)}$	$86.0_{(1.5)}$	$84.3_{(2.8)}$
	Pretrain	$85.2_{(1.1)}$	$84.0_{(2.3)}$	$83.7_{(3.2)}$	$85.4_{(1.6)}$	$86.3_{(0.9)}$	$85.6_{(1.7)}$	$83.2_{(2.1)}$	$85.3_{(1.4)}$	$84.9_{(2.3)}$
	Δ	+0.2	−0.2	−1.8	−0.7	−0.4	−0.8	−1.4	−0.6	+0.5
mut_het_rcc_pbrm1_mutation	Scratch	$62.9_{(2.8)}$	$63.6_{(1.3)}$	$57.9_{(3.3)}$	$64.5_{(1.8)}$	$63.5_{(2.1)}$	$60.9_{(1.6)}$	$64.6_{(2.0)}$	$66.0_{(1.5)}$	$69.3_{(0.9)}$
	Pretrain	$64.7_{(2.7)}$	$68.0_{(2.0)}$	$63.3_{(2.6)}$	$61.0_{(2.8)}$	$62.6_{(1.4)}$	$62.5_{(2.8)}$	$68.0_{(1.5)}$	$69.0_{(1.2)}$	$68.7_{(1.9)}$
	Δ	+1.8	+4.4	+5.4	−3.5	−1.0	+1.5	+3.3	+3.0	−0.5
mut_het_rcc_setd2_mutation	Scratch	$69.3_{(7.4)}$	$72.4_{(4.5)}$	$66.1_{(5.9)}$	$65.9_{(8.8)}$	$69.7_{(5.3)}$	$72.2_{(2.6)}$	$68.7_{(7.0)}$	$70.6_{(4.6)}$	$67.0_{(6.6)}$
	Pretrain	$70.0_{(5.6)}$	$71.8_{(3.8)}$	$69.3_{(5.1)}$	$72.2_{(1.7)}$	$69.3_{(5.0)}$	$72.3_{(5.3)}$	$69.7_{(5.8)}$	$71.4_{(5.5)}$	$70.8_{(3.3)}$
	Δ	+0.8	−0.6	+3.2	+6.3	−0.4	+0.1	+1.0	+0.9	+3.9

A.5 Results of few shot

Table 13: Few-shot comparison between from-scratch training and distillation initialization measured by bacc. Mean and standard deviation are computed across seeds.

Dataset	MIL	Setting	1-shot	2-shot	4-shot	8-shot	16-shot
BRACS-Coarse	2DMamba	Scratch	$35.1_{(3.3)}$	$33.8_{(4.5)}$	$41.7_{(6.3)}$	$50.4_{(8.2)}$	$58.8_{(5.7)}$
		Pretrain	$\mathbf{47.6}_{(10.7)}$	$\mathbf{41.7}_{(8.4)}$	$\mathbf{58.5}_{(8.9)}$	$\mathbf{64.5}_{(8.2)}$	$\mathbf{72.9}_{(5.4)}$
	ABMIL	Scratch	$38.6_{(6.8)}$	$39.0_{(8.6)}$	$\mathbf{48.8}_{(10.4)}$	$44.9_{(8.2)}$	$55.8_{(5.5)}$
		Pretrain	$\mathbf{39.1}_{(7.7)}$	$\mathbf{40.6}_{(11.3)}$	$47.3_{(12.8)}$	$\mathbf{50.7}_{(10.2)}$	$\mathbf{57.8}_{(5.0)}$
	AEMMIL	Scratch	$40.7_{(8.1)}$	$38.4_{(8.1)}$	$46.3_{(7.8)}$	$52.6_{(12.9)}$	$60.7_{(6.9)}$
		Pretrain	$\mathbf{45.2}_{(10.0)}$	$\mathbf{52.5}_{(16.4)}$	$\mathbf{56.3}_{(12.6)}$	$\mathbf{60.3}_{(10.6)}$	$\mathbf{67.9}_{(5.1)}$
	AMDMIL	Scratch	$41.2_{(5.5)}$	$41.9_{(6.2)}$	$46.0_{(7.1)}$	$55.2_{(6.4)}$	$61.8_{(4.5)}$
		Pretrain	$\mathbf{49.0}_{(8.0)}$	$\mathbf{58.3}_{(3.5)}$	$\mathbf{59.4}_{(5.7)}$	$\mathbf{60.0}_{(3.0)}$	$\mathbf{67.0}_{(4.9)}$
	CLAM	Scratch	$\mathbf{32.9}_{(0.8)}$	$34.2_{(1.0)}$	$42.4_{(10.1)}$	$49.6_{(8.6)}$	$57.6_{(7.3)}$
		Pretrain	$32.5_{(2.1)}$	$\mathbf{49.2}_{(10.0)}$	$\mathbf{47.3}_{(5.9)}$	$\mathbf{57.2}_{(3.8)}$	$\mathbf{57.9}_{(6.1)}$
	GDFMIL	Scratch	$36.7_{(5.7)}$	$38.3_{(4.7)}$	$38.7_{(9.4)}$	$51.3_{(14.1)}$	$64.1_{(5.7)}$
		Pretrain	$\mathbf{46.4}_{(9.1)}$	$\mathbf{45.9}_{(7.4)}$	$\mathbf{57.0}_{(10.6)}$	$\mathbf{65.1}_{(8.3)}$	$\mathbf{68.9}_{(10.2)}$
	TransMIL	Scratch	$42.3_{(4.9)}$	$38.5_{(7.0)}$	$45.1_{(9.3)}$	$51.1_{(11.9)}$	$56.4_{(5.0)}$
		Pretrain	$\mathbf{49.1}_{(4.4)}$	$\mathbf{57.0}_{(5.8)}$	$\mathbf{55.7}_{(7.9)}$	$\mathbf{61.7}_{(7.5)}$	$\mathbf{69.2}_{(2.8)}$
	WiKG	Scratch	$42.4_{(3.5)}$	$40.3_{(8.0)}$	$46.2_{(2.0)}$	$49.2_{(7.4)}$	$56.8_{(4.4)}$
		Pretrain	$\mathbf{54.1}_{(10.7)}$	$\mathbf{60.9}_{(13.0)}$	$\mathbf{63.2}_{(10.4)}$	$\mathbf{60.1}_{(6.8)}$	$\mathbf{63.6}_{(4.1)}$
BRACS-Fine	2DMamba	Scratch	$15.5_{(3.8)}$	$\mathbf{17.7}_{(2.5)}$	$26.9_{(9.2)}$	$38.2_{(6.9)}$	$\mathbf{46.1}_{(4.3)}$
		Pretrain	$\mathbf{22.6}_{(5.4)}$	$17.3_{(3.3)}$	$\mathbf{31.5}_{(4.1)}$	$\mathbf{41.9}_{(3.3)}$	$42.6_{(4.0)}$
	ABMIL	Scratch	$\mathbf{19.2}_{(5.1)}$	$16.7_{(3.9)}$	$27.3_{(9.5)}$	$33.4_{(3.7)}$	$38.8_{(5.8)}$
		Pretrain	$17.6_{(2.2)}$	$\mathbf{24.6}_{(6.6)}$	$\mathbf{31.0}_{(4.1)}$	$\mathbf{36.8}_{(6.2)}$	$\mathbf{43.3}_{(6.0)}$
	AEMMIL	Scratch	$20.0_{(7.8)}$	$21.3_{(4.3)}$	$30.2_{(5.5)}$	$33.2_{(4.9)}$	$\mathbf{43.6}_{(3.9)}$
		Pretrain	$\mathbf{24.5}_{(9.0)}$	$\mathbf{29.6}_{(4.2)}$	$\mathbf{37.8}_{(5.0)}$	$\mathbf{39.6}_{(1.9)}$	$39.8_{(2.3)}$
	AMDMIL	Scratch	$18.7_{(4.2)}$	$17.9_{(4.8)}$	$26.7_{(5.5)}$	$30.4_{(7.4)}$	$\mathbf{42.7}_{(4.6)}$
		Pretrain	$\mathbf{27.0}_{(4.9)}$	$\mathbf{28.4}_{(6.8)}$	$\mathbf{32.7}_{(5.3)}$	$\mathbf{36.3}_{(4.6)}$	$42.0_{(1.6)}$
	CLAM	Scratch	$\mathbf{20.1}_{(7.5)}$	$\mathbf{23.3}_{(7.2)}$	$27.9_{(8.1)}$	$32.2_{(6.7)}$	$38.0_{(4.0)}$
		Pretrain	$19.9_{(5.4)}$	$17.5_{(7.1)}$	$\mathbf{32.4}_{(5.3)}$	$\mathbf{34.9}_{(3.5)}$	$\mathbf{40.3}_{(6.7)}$
	GDFMIL	Scratch	$19.8_{(4.7)}$	$24.2_{(4.0)}$	$29.2_{(11.6)}$	$33.8_{(6.3)}$	$38.7_{(3.9)}$
		Pretrain	$\mathbf{25.1}_{(6.1)}$	$\mathbf{27.3}_{(6.4)}$	$\mathbf{37.7}_{(3.6)}$	$\mathbf{40.2}_{(3.5)}$	$\mathbf{41.2}_{(4.5)}$
	TransMIL	Scratch	$22.9_{(5.4)}$	$20.5_{(5.0)}$	$\mathbf{34.0}_{(7.0)}$	$35.1_{(5.6)}$	$36.8_{(3.4)}$
		Pretrain	$\mathbf{28.8}_{(2.3)}$	$\mathbf{35.6}_{(3.6)}$	$32.5_{(4.8)}$	$\mathbf{36.2}_{(3.6)}$	$\mathbf{41.0}_{(3.7)}$
	WiKG	Scratch	$19.8_{(7.5)}$	$18.5_{(7.1)}$	$26.8_{(5.2)}$	$25.6_{(3.2)}$	$\mathbf{39.4}_{(6.8)}$
		Pretrain	$\mathbf{29.1}_{(4.0)}$	$\mathbf{30.8}_{(2.0)}$	$\mathbf{33.4}_{(5.0)}$	$\mathbf{38.2}_{(5.9)}$	$36.8_{(3.9)}$
CPTAC-NSCLC	2DMamba	Scratch	$57.0_{(10.8)}$	$72.6_{(8.2)}$	$77.7_{(5.0)}$	$86.0_{(6.0)}$	$87.8_{(5.5)}$
		Pretrain	$\mathbf{72.1}_{(10.7)}$	$\mathbf{81.4}_{(5.0)}$	$\mathbf{86.2}_{(3.9)}$	$\mathbf{89.1}_{(2.7)}$	$\mathbf{92.0}_{(2.8)}$
	ABMIL	Scratch	$\mathbf{56.7}_{(9.5)}$	$71.2_{(9.2)}$	$80.5_{(5.3)}$	$81.5_{(4.4)}$	$86.1_{(3.9)}$
		Pretrain	$55.9_{(9.5)}$	$\mathbf{75.4}_{(10.5)}$	$\mathbf{81.5}_{(4.9)}$	$\mathbf{87.3}_{(4.5)}$	$\mathbf{88.8}_{(3.5)}$
	AEMMIL	Scratch	$65.8_{(11.2)}$	$74.4_{(9.0)}$	$82.7_{(10.7)}$	$89.8_{(4.7)}$	$\mathbf{93.0}_{(5.9)}$
		Pretrain	$\mathbf{72.9}_{(14.1)}$	$\mathbf{79.0}_{(6.3)}$	$\mathbf{85.5}_{(5.4)}$	$\mathbf{91.1}_{(3.0)}$	$92.8_{(2.2)}$
	AMDMIL	Scratch	$64.9_{(14.5)}$	$75.1_{(2.4)}$	$78.2_{(8.4)}$	$86.1_{(3.1)}$	$87.8_{(3.4)}$
		Pretrain	$\mathbf{73.6}_{(13.5)}$	$\mathbf{82.2}_{(5.3)}$	$\mathbf{87.5}_{(4.4)}$	$\mathbf{89.4}_{(6.8)}$	$\mathbf{91.2}_{(3.6)}$
	CLAM	Scratch	$\mathbf{70.7}_{(12.7)}$	$\mathbf{79.9}_{(6.9)}$	$80.7_{(4.4)}$	$\mathbf{87.0}_{(7.6)}$	$87.3_{(8.3)}$
		Pretrain	$68.6_{(17.2)}$	$79.2_{(6.9)}$	$\mathbf{83.3}_{(3.4)}$	$86.4_{(4.7)}$	$\mathbf{88.7}_{(2.5)}$
	GDFMIL	Scratch	$61.2_{(9.1)}$	$61.7_{(8.6)}$	$78.3_{(8.3)}$	$83.8_{(5.9)}$	$89.1_{(3.5)}$
		Pretrain	$\mathbf{77.6}_{(9.7)}$	$\mathbf{85.2}_{(5.0)}$	$\mathbf{86.1}_{(6.8)}$	$\mathbf{93.3}_{(4.2)}$	$\mathbf{94.3}_{(3.7)}$
	TransMIL	Scratch	$\mathbf{74.2}_{(13.2)}$	$81.4_{(5.5)}$	$83.1_{(5.3)}$	$86.9_{(2.9)}$	$89.0_{(4.3)}$
		Pretrain	$71.7_{(15.2)}$	$\mathbf{81.8}_{(5.7)}$	$\mathbf{86.1}_{(4.4)}$	$\mathbf{91.3}_{(3.8)}$	$\mathbf{93.3}_{(3.0)}$
	WiKG	Scratch	$68.5_{(13.2)}$	$76.1_{(6.7)}$	$78.6_{(8.7)}$	$85.3_{(2.3)}$	$88.1_{(3.2)}$
		Pretrain	$\mathbf{84.0}_{(5.2)}$	$\mathbf{86.2}_{(6.4)}$	$\mathbf{89.8}_{(3.5)}$	$\mathbf{93.3}_{(3.5)}$	$\mathbf{92.2}_{(3.0)}$
KidRare-Coarse	2DMamba	Scratch	$39.1_{(6.2)}$	$32.0_{(5.0)}$	$58.0_{(11.2)}$	$74.5_{(5.8)}$	$79.0_{(2.3)}$
		Pretrain	$\mathbf{48.8}_{(9.3)}$	$\mathbf{33.5}_{(15.8)}$	$\mathbf{65.2}_{(2.5)}$	$\mathbf{76.7}_{(2.7)}$	$\mathbf{83.0}_{(2.9)}$
	ABMIL	Scratch	$40.0_{(13.2)}$	$52.3_{(11.8)}$	$64.7_{(8.4)}$	$76.7_{(5.6)}$	$81.5_{(5.0)}$
		Pretrain	$\mathbf{44.0}_{(18.8)}$	$\mathbf{59.5}_{(15.0)}$	$\mathbf{69.7}_{(6.1)}$	$\mathbf{78.0}_{(4.3)}$	$\mathbf{83.5}_{(3.8)}$
	AEMMIL	Scratch	$39.9_{(11.2)}$	$49.9_{(11.0)}$	$61.4_{(4.0)}$	$75.2_{(5.1)}$	$81.8_{(2.8)}$
		Pretrain	$\mathbf{45.2}_{(11.3)}$	$\mathbf{56.1}_{(16.7)}$	$\mathbf{69.7}_{(3.7)}$	$\mathbf{79.7}_{(4.1)}$	$\mathbf{83.1}_{(2.7)}$
	AMDMIL	Scratch	$44.0_{(7.8)}$	$56.6_{(9.3)}$	$65.1_{(3.2)}$	$76.3_{(2.5)}$	$83.0_{(2.2)}$
		Pretrain	$\mathbf{48.5}_{(8.6)}$	$\mathbf{63.6}_{(8.8)}$	$\mathbf{69.5}_{(4.1)}$	$\mathbf{79.6}_{(2.1)}$	$\mathbf{86.3}_{(1.3)}$
	CLAM	Scratch	$37.2_{(11.7)}$	$46.4_{(10.7)}$	$57.8_{(15.2)}$	$66.1_{(8.8)}$	$76.6_{(6.8)}$
		Pretrain	$\mathbf{39.1}_{(14.4)}$	$\mathbf{56.4}_{(9.5)}$	$\mathbf{64.2}_{(7.1)}$	$\mathbf{74.8}_{(3.6)}$	$\mathbf{80.2}_{(3.3)}$
	GDFMIL	Scratch	$39.0_{(7.5)}$	$52.1_{(11.0)}$	$59.1_{(8.0)}$	$74.9_{(5.2)}$	$80.5_{(2.8)}$
		Pretrain	$\mathbf{51.5}_{(11.4)}$	$\mathbf{65.2}_{(5.3)}$	$\mathbf{72.8}_{(4.7)}$	$\mathbf{80.3}_{(2.5)}$	$\mathbf{83.0}_{(4.0)}$
	TransMIL	Scratch	$43.0_{(10.7)}$	$50.1_{(9.3)}$	$56.9_{(4.7)}$	$70.6_{(5.1)}$	$80.3_{(3.0)}$
		Pretrain	$\mathbf{45.0}_{(15.1)}$	$\mathbf{57.5}_{(11.5)}$	$\mathbf{70.7}_{(3.4)}$	$\mathbf{81.1}_{(3.3)}$	$\mathbf{86.8}_{(1.9)}$
	WiKG	Scratch	$42.1_{(5.3)}$	$50.1_{(8.4)}$	$61.0_{(1.7)}$	$65.6_{(3.2)}$	$76.6_{(2.9)}$
		Pretrain	$\mathbf{48.9}_{(11.3)}$	$\mathbf{60.8}_{(15.4)}$	$\mathbf{68.7}_{(3.9)}$	$\mathbf{81.2}_{(3.8)}$	$\mathbf{83.9}_{(2.7)}$
KidRare-Fine	2DMamba	Scratch	$20.6_{(5.3)}$	$\mathbf{13.6}_{(5.3)}$	$43.7_{(5.0)}$	$44.8_{(5.3)}$	$\mathbf{53.6}_{(4.2)}$
		Pretrain	$\mathbf{24.7}_{(7.8)}$	$10.2_{(2.6)}$	$\mathbf{44.9}_{(5.9)}$	$\mathbf{47.9}_{(4.5)}$	$52.9_{(4.7)}$
	ABMIL	Scratch	$\mathbf{25.0}_{(10.9)}$	$27.5_{(4.7)}$	$\mathbf{38.7}_{(5.2)}$	$45.7_{(4.2)}$	$46.3_{(7.2)}$
		Pretrain	$23.2_{(8.5)}$	$\mathbf{32.4}_{(2.9)}$	$38.7_{(8.3)}$	$\mathbf{50.7}_{(2.6)}$	$\mathbf{46.8}_{(3.4)}$
	AEMMIL	Scratch	$22.5_{(4.3)}$	$32.4_{(4.6)}$	$\mathbf{44.0}_{(5.1)}$	$\mathbf{49.7}_{(3.8)}$	$\mathbf{50.7}_{(4.2)}$
		Pretrain	$\mathbf{28.9}_{(2.4)}$	$\mathbf{33.2}_{(0.8)}$	$42.1_{(4.7)}$	$48.9_{(5.0)}$	$49.6_{(7.0)}$
	AMDMIL	Scratch	$25.7_{(4.0)}$	$33.6_{(4.3)}$	$\mathbf{44.2}_{(5.8)}$	$51.1_{(3.8)}$	$54.0_{(5.5)}$
		Pretrain	$\mathbf{31.0}_{(7.9)}$	$\mathbf{37.9}_{(2.3)}$	$43.4_{(7.6)}$	$\mathbf{51.3}_{(2.7)}$	$\mathbf{58.1}_{(3.3)}$
	CLAM	Scratch	$\mathbf{22.1}_{(5.1)}$	$30.4_{(2.6)}$	$\mathbf{37.8}_{(1.8)}$	$\mathbf{43.3}_{(4.9)}$	$41.3_{(3.7)}$
		Pretrain	$21.4_{(8.3)}$	$\mathbf{31.1}_{(4.7)}$	$36.5_{(5.4)}$	$42.6_{(6.7)}$	$\mathbf{46.2}_{(6.4)}$
	GDFMIL	Scratch	$21.8_{(7.4)}$	$\mathbf{30.6}_{(4.3)}$	$\mathbf{41.1}_{(5.9)}$	$44.7_{(6.1)}$	$45.9_{(6.0)}$
		Pretrain	$\mathbf{26.7}_{(4.8)}$	$30.2_{(6.3)}$	$40.3_{(6.0)}$	$\mathbf{46.2}_{(5.5)}$	$\mathbf{47.6}_{(5.2)}$
	TransMIL	Scratch	$27.8_{(4.2)}$	$34.7_{(3.4)}$	$39.2_{(3.8)}$	$46.6_{(1.4)}$	$51.8_{(3.3)}$
		Pretrain	$\mathbf{32.4}_{(5.4)}$	$\mathbf{40.2}_{(5.4)}$	$\mathbf{47.3}_{(3.7)}$	$\mathbf{54.1}_{(4.5)}$	$\mathbf{54.8}_{(3.0)}$
	WiKG	Scratch	$\mathbf{28.5}_{(3.8)}$	$28.1_{(3.9)}$	$41.6_{(4.2)}$	$43.9_{(5.7)}$	$52.0_{(2.4)}$
		Pretrain	$28.5_{(6.5)}$	$\mathbf{35.9}_{(1.4)}$	$\mathbf{47.6}_{(4.0)}$	$\mathbf{48.1}_{(6.3)}$	$\mathbf{52.9}_{(4.4)}$

Table 14: Few-shot comparison between from-scratch training and distillation initialization measured by Accuracy. Mean and standard deviation are computed across seeds. DAGMIL is excluded because only scratch results are available, and 32-shot results are omitted.

Dataset	MIL	Setting	1-shot	2-shot	4-shot	8-shot	16-shot
BRACS-Coarse	2DMamba	Scratch	$33.6_{(4.1)}$	$35.6_{(5.2)}$	$40.5_{(6.1)}$	$50.8_{(7.7)}$	$61.8_{(5.6)}$
		Pretrain	$\mathbf{47.8}_{(9.8)}$	$\mathbf{45.8}_{(7.6)}$	$\mathbf{57.2}_{(8.5)}$	$\mathbf{65.1}_{(8.2)}$	$\mathbf{73.6}_{(4.4)}$
	ABMIL	Scratch	$37.0_{(6.1)}$	$35.9_{(10.2)}$	$\mathbf{48.3}_{(11.6)}$	$44.8_{(9.1)}$	$55.4_{(6.6)}$
		Pretrain	$\mathbf{37.2}_{(7.9)}$	$\mathbf{38.4}_{(11.2)}$	$46.4_{(12.8)}$	$\mathbf{50.6}_{(9.6)}$	$\mathbf{58.2}_{(5.3)}$
	AEMMIL	Scratch	$41.1_{(7.1)}$	$39.1_{(7.2)}$	$44.6_{(10.7)}$	$52.9_{(12.2)}$	$62.1_{(8.2)}$
		Pretrain	$\mathbf{43.5}_{(9.0)}$	$\mathbf{52.4}_{(15.9)}$	$\mathbf{54.2}_{(13.4)}$	$\mathbf{59.1}_{(11.2)}$	$\mathbf{66.9}_{(6.3)}$
	AMDMIL	Scratch	$40.0_{(3.4)}$	$40.0_{(6.6)}$	$43.0_{(8.1)}$	$54.9_{(10.0)}$	$60.2_{(4.7)}$
		Pretrain	$\mathbf{46.2}_{(9.2)}$	$\mathbf{58.6}_{(4.6)}$	$\mathbf{60.0}_{(7.2)}$	$\mathbf{57.9}_{(3.5)}$	$\mathbf{67.6}_{(5.4)}$
	CLAM	Scratch	$\mathbf{31.0}_{(4.8)}$	$32.6_{(4.7)}$	$42.5_{(12.1)}$	$48.3_{(9.3)}$	$\mathbf{58.2}_{(9.5)}$
		Pretrain	$30.3_{(5.4)}$	$\mathbf{48.0}_{(12.0)}$	$\mathbf{47.1}_{(8.1)}$	$\mathbf{57.0}_{(5.3)}$	$57.7_{(7.9)}$
	GDFMIL	Scratch	$35.4_{(6.6)}$	$38.4_{(4.3)}$	$36.1_{(9.8)}$	$50.8_{(13.4)}$	$62.5_{(6.2)}$
		Pretrain	$\mathbf{44.1}_{(7.6)}$	$\mathbf{47.4}_{(7.0)}$	$\mathbf{54.9}_{(11.3)}$	$\mathbf{65.1}_{(6.4)}$	$\mathbf{70.8}_{(8.8)}$
	TransMIL	Scratch	$41.8_{(3.4)}$	$35.2_{(6.9)}$	$42.5_{(9.6)}$	$48.7_{(12.6)}$	$56.3_{(6.0)}$
		Pretrain	$\mathbf{48.3}_{(7.7)}$	$\mathbf{57.0}_{(5.9)}$	$\mathbf{54.9}_{(10.2)}$	$\mathbf{61.4}_{(8.2)}$	$\mathbf{69.4}_{(3.0)}$
	WiKG	Scratch	$41.8_{(3.6)}$	$40.7_{(8.0)}$	$44.4_{(5.1)}$	$48.3_{(9.4)}$	$56.3_{(6.4)}$
		Pretrain	$\mathbf{53.6}_{(9.8)}$	$\mathbf{61.8}_{(14.1)}$	$\mathbf{61.2}_{(12.2)}$	$\mathbf{60.5}_{(8.8)}$	$\mathbf{63.0}_{(4.8)}$
BRACS-Fine	2DMamba	Scratch	$14.0_{(3.0)}$	$\mathbf{18.2}_{(4.2)}$	$29.9_{(9.6)}$	$41.8_{(5.6)}$	$\mathbf{48.5}_{(3.8)}$
		Pretrain	$\mathbf{24.1}_{(4.3)}$	$16.0_{(6.9)}$	$\mathbf{33.8}_{(5.6)}$	$\mathbf{44.1}_{(3.5)}$	$45.3_{(4.9)}$
	ABMIL	Scratch	$17.0_{(5.1)}$	$16.3_{(4.7)}$	$28.3_{(9.7)}$	$34.7_{(4.3)}$	$39.5_{(5.2)}$
		Pretrain	$\mathbf{17.5}_{(4.7)}$	$\mathbf{24.8}_{(9.3)}$	$\mathbf{31.7}_{(5.9)}$	$\mathbf{38.2}_{(5.2)}$	$\mathbf{43.2}_{(7.9)}$
	AEMMIL	Scratch	$18.9_{(8.7)}$	$21.1_{(6.7)}$	$33.6_{(6.0)}$	$36.1_{(5.8)}$	$\mathbf{46.7}_{(4.3)}$
		Pretrain	$\mathbf{23.4}_{(10.2)}$	$\mathbf{30.8}_{(5.7)}$	$\mathbf{40.5}_{(4.8)}$	$\mathbf{43.2}_{(2.2)}$	$42.5_{(1.0)}$
	AMDMIL	Scratch	$15.9_{(3.0)}$	$17.7_{(3.8)}$	$29.4_{(4.5)}$	$32.9_{(8.7)}$	$\mathbf{44.1}_{(5.1)}$
		Pretrain	$\mathbf{25.5}_{(6.0)}$	$\mathbf{27.6}_{(6.2)}$	$\mathbf{34.2}_{(6.5)}$	$\mathbf{39.3}_{(4.1)}$	$43.5_{(2.6)}$
	CLAM	Scratch	$20.2_{(7.3)}$	$\mathbf{24.8}_{(9.3)}$	$29.0_{(9.9)}$	$33.8_{(6.2)}$	$39.8_{(3.9)}$
		Pretrain	$\mathbf{21.4}_{(5.9)}$	$19.1_{(8.5)}$	$\mathbf{33.8}_{(7.5)}$	$\mathbf{37.5}_{(3.0)}$	$\mathbf{42.1}_{(7.1)}$
	GDFMIL	Scratch	$17.0_{(2.8)}$	$23.2_{(6.2)}$	$28.7_{(14.6)}$	$36.3_{(6.4)}$	$40.9_{(3.5)}$
		Pretrain	$\mathbf{25.3}_{(8.8)}$	$\mathbf{26.4}_{(5.3)}$	$\mathbf{39.8}_{(3.8)}$	$\mathbf{42.8}_{(2.5)}$	$\mathbf{43.5}_{(3.4)}$
	TransMIL	Scratch	$21.6_{(5.8)}$	$19.5_{(5.1)}$	$35.6_{(7.3)}$	$36.3_{(6.3)}$	$40.5_{(2.9)}$
		Pretrain	$\mathbf{28.5}_{(3.3)}$	$\mathbf{36.8}_{(4.0)}$	$\mathbf{35.9}_{(3.4)}$	$\mathbf{39.8}_{(4.2)}$	$\mathbf{43.5}_{(5.9)}$
	WiKG	Scratch	$17.9_{(7.6)}$	$18.4_{(7.3)}$	$29.4_{(3.1)}$	$24.4_{(6.3)}$	$\mathbf{41.4}_{(6.1)}$
		Pretrain	$\mathbf{28.3}_{(6.0)}$	$\mathbf{32.9}_{(2.0)}$	$\mathbf{37.0}_{(3.9)}$	$\mathbf{41.6}_{(4.3)}$	$38.4_{(4.2)}$
CPTAC-NSCLC	2DMamba	Scratch	$57.0_{(10.8)}$	$72.8_{(8.1)}$	$77.7_{(5.0)}$	$86.0_{(5.9)}$	$87.8_{(5.5)}$
		Pretrain	$\mathbf{72.0}_{(11.0)}$	$\mathbf{81.5}_{(4.9)}$	$\mathbf{86.2}_{(3.9)}$	$\mathbf{89.2}_{(2.8)}$	$\mathbf{92.0}_{(2.8)}$
	ABMIL	Scratch	$\mathbf{56.7}_{(9.6)}$	$71.3_{(9.1)}$	$80.5_{(5.3)}$	$81.5_{(4.4)}$	$86.2_{(3.9)}$
		Pretrain	$55.8_{(9.7)}$	$\mathbf{75.5}_{(10.3)}$	$\mathbf{81.5}_{(4.8)}$	$\mathbf{87.3}_{(4.6)}$	$\mathbf{88.8}_{(3.4)}$
	AEMMIL	Scratch	$65.7_{(11.4)}$	$74.5_{(8.9)}$	$82.7_{(10.6)}$	$89.8_{(4.7)}$	$\mathbf{93.0}_{(5.9)}$
		Pretrain	$\mathbf{72.8}_{(14.3)}$	$\mathbf{79.0}_{(6.3)}$	$\mathbf{85.5}_{(5.4)}$	$\mathbf{91.2}_{(3.0)}$	$92.8_{(2.3)}$
	AMDMIL	Scratch	$65.0_{(14.3)}$	$75.2_{(2.5)}$	$78.3_{(8.4)}$	$86.2_{(3.1)}$	$87.8_{(3.4)}$
		Pretrain	$\mathbf{73.7}_{(13.3)}$	$\mathbf{82.3}_{(5.1)}$	$\mathbf{87.5}_{(4.5)}$	$\mathbf{89.3}_{(6.9)}$	$\mathbf{91.2}_{(3.6)}$
	CLAM	Scratch	$\mathbf{70.7}_{(12.9)}$	$\mathbf{80.0}_{(6.9)}$	$80.7_{(4.5)}$	$\mathbf{87.0}_{(7.6)}$	$87.2_{(8.5)}$
		Pretrain	$68.7_{(17.3)}$	$79.3_{(6.8)}$	$\mathbf{83.3}_{(3.5)}$	$86.5_{(4.7)}$	$\mathbf{88.7}_{(2.5)}$
	GDFMIL	Scratch	$61.0_{(9.2)}$	$61.8_{(8.9)}$	$78.2_{(8.4)}$	$83.8_{(6.0)}$	$89.2_{(3.5)}$
		Pretrain	$\mathbf{77.5}_{(9.9)}$	$\mathbf{85.2}_{(5.0)}$	$\mathbf{86.0}_{(6.8)}$	$\mathbf{93.3}_{(4.2)}$	$\mathbf{94.3}_{(3.7)}$
	TransMIL	Scratch	$\mathbf{74.2}_{(13.5)}$	$81.5_{(5.6)}$	$83.2_{(5.4)}$	$87.0_{(2.9)}$	$89.0_{(4.4)}$
		Pretrain	$71.8_{(15.1)}$	$\mathbf{82.0}_{(5.6)}$	$\mathbf{86.2}_{(4.4)}$	$\mathbf{91.3}_{(3.8)}$	$\mathbf{93.3}_{(3.0)}$
	WiKG	Scratch	$68.7_{(13.2)}$	$76.2_{(6.7)}$	$78.7_{(8.8)}$	$85.3_{(2.3)}$	$88.2_{(3.2)}$
		Pretrain	$\mathbf{84.0}_{(5.1)}$	$\mathbf{86.2}_{(6.5)}$	$\mathbf{89.8}_{(3.4)}$	$\mathbf{93.3}_{(3.5)}$	$\mathbf{92.2}_{(3.0)}$
KidRare-Coarse	2DMamba	Scratch	$38.8_{(5.4)}$	$34.2_{(7.2)}$	$55.8_{(10.4)}$	$74.6_{(6.2)}$	$79.1_{(2.2)}$
		Pretrain	$\mathbf{43.0}_{(9.9)}$	$\mathbf{37.5}_{(13.4)}$	$\mathbf{62.5}_{(1.5)}$	$\mathbf{76.4}_{(4.8)}$	$\mathbf{82.2}_{(2.2)}$
	ABMIL	Scratch	$39.2_{(9.7)}$	$49.9_{(11.4)}$	$63.9_{(5.5)}$	$76.7_{(5.9)}$	$80.5_{(5.9)}$
		Pretrain	$\mathbf{41.9}_{(19.1)}$	$\mathbf{57.7}_{(15.4)}$	$\mathbf{67.0}_{(5.8)}$	$\mathbf{76.9}_{(5.4)}$	$\mathbf{82.5}_{(4.7)}$
	AEMMIL	Scratch	$36.3_{(7.1)}$	$46.3_{(10.8)}$	$59.1_{(2.5)}$	$73.8_{(6.0)}$	$81.8_{(2.8)}$
		Pretrain	$\mathbf{42.3}_{(8.7)}$	$\mathbf{54.9}_{(14.3)}$	$\mathbf{65.5}_{(5.3)}$	$\mathbf{80.2}_{(4.8)}$	$\mathbf{83.2}_{(3.0)}$
	AMDMIL	Scratch	$37.1_{(5.6)}$	$53.1_{(9.2)}$	$62.2_{(3.9)}$	$75.2_{(2.7)}$	$82.0_{(2.4)}$
		Pretrain	$\mathbf{41.3}_{(8.5)}$	$\mathbf{58.4}_{(9.8)}$	$\mathbf{63.6}_{(5.4)}$	$\mathbf{78.8}_{(3.0)}$	$\mathbf{85.2}_{(2.0)}$
	CLAM	Scratch	$38.5_{(7.0)}$	$44.8_{(9.9)}$	$57.4_{(12.8)}$	$67.0_{(9.9)}$	$76.2_{(7.2)}$
		Pretrain	$\mathbf{39.8}_{(8.6)}$	$\mathbf{55.3}_{(8.7)}$	$\mathbf{62.4}_{(8.9)}$	$\mathbf{73.9}_{(4.6)}$	$\mathbf{80.0}_{(3.9)}$
	GDFMIL	Scratch	$35.9_{(9.6)}$	$46.6_{(11.3)}$	$55.9_{(5.8)}$	$72.9_{(6.6)}$	$79.8_{(3.0)}$
		Pretrain	$\mathbf{48.0}_{(12.2)}$	$\mathbf{61.1}_{(8.0)}$	$\mathbf{70.7}_{(4.8)}$	$\mathbf{79.8}_{(3.8)}$	$\mathbf{83.8}_{(4.3)}$
	TransMIL	Scratch	$41.1_{(6.7)}$	$46.9_{(10.1)}$	$55.1_{(3.4)}$	$70.1_{(5.0)}$	$80.9_{(3.4)}$
		Pretrain	$\mathbf{42.7}_{(13.8)}$	$\mathbf{54.1}_{(8.4)}$	$\mathbf{68.3}_{(2.3)}$	$\mathbf{80.5}_{(3.9)}$	$\mathbf{86.4}_{(1.1)}$
	WiKG	Scratch	$40.6_{(5.3)}$	$49.4_{(9.3)}$	$58.8_{(1.9)}$	$67.0_{(2.7)}$	$77.0_{(2.3)}$
		Pretrain	$\mathbf{42.7}_{(10.3)}$	$\mathbf{58.4}_{(14.7)}$	$\mathbf{66.1}_{(6.2)}$	$\mathbf{80.3}_{(3.6)}$	$\mathbf{83.7}_{(3.4)}$
KidRare-Fine	2DMamba	Scratch	$19.7_{(3.3)}$	$\mathbf{11.8}_{(7.8)}$	$\mathbf{37.4}_{(3.9)}$	$42.2_{(3.0)}$	$\mathbf{52.0}_{(2.0)}$
		Pretrain	$\mathbf{20.3}_{(7.4)}$	$10.4_{(5.9)}$	$\mathbf{37.4}_{(6.6)}$	$\mathbf{44.3}_{(6.6)}$	$50.1_{(4.6)}$
	ABMIL	Scratch	$\mathbf{21.4}_{(9.1)}$	$21.9_{(7.1)}$	$30.5_{(5.2)}$	$42.6_{(6.2)}$	$\mathbf{47.4}_{(7.2)}$
		Pretrain	$20.7_{(8.7)}$	$\mathbf{26.3}_{(6.7)}$	$\mathbf{32.9}_{(10.7)}$	$\mathbf{44.0}_{(5.9)}$	$47.3_{(5.3)}$
	AEMMIL	Scratch	$19.2_{(3.8)}$	$26.3_{(7.7)}$	$37.0_{(5.4)}$	$47.7_{(5.8)}$	$\mathbf{48.7}_{(5.8)}$
		Pretrain	$\mathbf{23.4}_{(4.1)}$	$\mathbf{30.2}_{(7.5)}$	$\mathbf{37.4}_{(6.2)}$	$\mathbf{48.2}_{(8.7)}$	$48.1_{(7.8)}$
	AMDMIL	Scratch	$20.3_{(4.3)}$	$31.0_{(3.8)}$	$\mathbf{40.6}_{(5.1)}$	$45.4_{(4.5)}$	$50.6_{(3.9)}$
		Pretrain	$\mathbf{27.6}_{(2.5)}$	$\mathbf{35.0}_{(5.4)}$	$37.9_{(6.1)}$	$\mathbf{48.1}_{(4.9)}$	$\mathbf{54.1}_{(3.9)}$
	CLAM	Scratch	$20.8_{(4.9)}$	$24.4_{(2.1)}$	$32.7_{(2.6)}$	$\mathbf{39.8}_{(4.8)}$	$46.1_{(3.3)}$
		Pretrain	$\mathbf{22.6}_{(6.6)}$	$\mathbf{26.4}_{(5.3)}$	$\mathbf{33.6}_{(3.3)}$	$36.7_{(6.5)}$	$\mathbf{48.0}_{(4.9)}$
	GDFMIL	Scratch	$17.3_{(3.7)}$	$24.1_{(5.5)}$	$34.0_{(3.9)}$	$\mathbf{43.0}_{(3.4)}$	$42.3_{(4.4)}$
		Pretrain	$\mathbf{24.9}_{(4.6)}$	$\mathbf{26.3}_{(3.9)}$	$\mathbf{42.8}_{(6.5)}$	$42.5_{(8.3)}$	$\mathbf{46.8}_{(8.1)}$
	TransMIL	Scratch	$23.4_{(3.2)}$	$28.9_{(4.0)}$	$36.0_{(2.0)}$	$41.9_{(3.2)}$	$49.1_{(2.8)}$
		Pretrain	$\mathbf{28.7}_{(5.9)}$	$\mathbf{31.3}_{(6.4)}$	$\mathbf{39.0}_{(5.8)}$	$\mathbf{49.9}_{(5.1)}$	$\mathbf{52.4}_{(3.0)}$
	WiKG	Scratch	$21.2_{(4.9)}$	$26.3_{(4.5)}$	$32.1_{(3.0)}$	$38.9_{(5.9)}$	$47.6_{(2.4)}$
		Pretrain	$\mathbf{24.7}_{(4.8)}$	$\mathbf{29.6}_{(4.0)}$	$\mathbf{39.7}_{(7.9)}$	$\mathbf{46.4}_{(5.2)}$	$\mathbf{51.6}_{(7.7)}$

Table 15: Few-shot comparison between from-scratch training and distillation initialization measured by AUC. Mean and standard deviation are computed across seeds. DAGMIL is excluded because only scratch results are available, and 32-shot results are omitted.

Dataset	MIL	Setting	1-shot	2-shot	4-shot	8-shot	16-shot
BRACS-Coarse	2DMamba	Scratch	$55.1_{(4.9)}$	$54.5_{(6.0)}$	$60.7_{(5.1)}$	$68.9_{(8.5)}$	$81.3_{(4.8)}$
		Pretrain	$\mathbf{65.0}_{(10.5)}$	$\mathbf{60.3}_{(9.7)}$	$\mathbf{76.4}_{(9.4)}$	$\mathbf{81.6}_{(6.9)}$	$\mathbf{88.8}_{(2.4)}$
	ABMIL	Scratch	$\mathbf{51.3}_{(12.5)}$	$50.2_{(11.2)}$	$63.0_{(14.3)}$	$62.4_{(15.0)}$	$74.0_{(9.3)}$
		Pretrain	$48.9_{(13.4)}$	$\mathbf{54.4}_{(18.6)}$	$\mathbf{65.7}_{(13.3)}$	$\mathbf{65.1}_{(14.3)}$	$\mathbf{77.6}_{(4.2)}$
	AEMMIL	Scratch	$55.3_{(9.2)}$	$59.1_{(9.2)}$	$64.5_{(10.2)}$	$71.2_{(11.9)}$	$80.2_{(5.6)}$
		Pretrain	$\mathbf{64.7}_{(12.6)}$	$\mathbf{71.7}_{(16.4)}$	$\mathbf{75.3}_{(12.7)}$	$\mathbf{79.3}_{(9.6)}$	$\mathbf{86.7}_{(3.1)}$
	AMDMIL	Scratch	$60.9_{(9.5)}$	$59.5_{(9.6)}$	$63.1_{(8.4)}$	$74.7_{(10.4)}$	$83.3_{(3.2)}$
		Pretrain	$\mathbf{70.4}_{(5.7)}$	$\mathbf{78.7}_{(5.9)}$	$\mathbf{78.5}_{(4.7)}$	$\mathbf{81.9}_{(1.8)}$	$\mathbf{87.5}_{(2.2)}$
	CLAM	Scratch	$45.5_{(3.2)}$	$46.1_{(3.5)}$	$58.9_{(11.4)}$	$65.8_{(13.2)}$	$76.4_{(6.1)}$
		Pretrain	$\mathbf{47.9}_{(3.4)}$	$\mathbf{66.5}_{(11.2)}$	$\mathbf{65.6}_{(7.8)}$	$\mathbf{77.2}_{(5.0)}$	$\mathbf{77.0}_{(4.1)}$
	GDFMIL	Scratch	$51.4_{(6.8)}$	$53.9_{(8.8)}$	$53.1_{(12.8)}$	$72.0_{(10.2)}$	$83.4_{(3.9)}$
		Pretrain	$\mathbf{66.3}_{(10.3)}$	$\mathbf{65.8}_{(9.1)}$	$\mathbf{75.4}_{(11.3)}$	$\mathbf{83.4}_{(5.2)}$	$\mathbf{86.1}_{(6.1)}$
	TransMIL	Scratch	$58.3_{(5.5)}$	$58.9_{(8.1)}$	$63.6_{(8.5)}$	$68.5_{(9.8)}$	$78.0_{(4.3)}$
		Pretrain	$\mathbf{68.0}_{(4.4)}$	$\mathbf{76.4}_{(6.0)}$	$\mathbf{76.6}_{(8.7)}$	$\mathbf{82.5}_{(5.1)}$	$\mathbf{86.4}_{(3.4)}$
	WiKG	Scratch	$61.8_{(6.1)}$	$59.9_{(5.5)}$	$61.3_{(3.8)}$	$70.2_{(9.6)}$	$79.9_{(5.5)}$
		Pretrain	$\mathbf{78.1}_{(5.7)}$	$\mathbf{78.2}_{(11.3)}$	$\mathbf{79.2}_{(8.9)}$	$\mathbf{80.1}_{(8.7)}$	$\mathbf{84.6}_{(2.1)}$
BRACS-Fine	2DMamba	Scratch	$52.8_{(7.0)}$	$\mathbf{53.7}_{(4.9)}$	$65.0_{(8.5)}$	$74.6_{(4.7)}$	$\mathbf{82.2}_{(2.6)}$
		Pretrain	$\mathbf{58.8}_{(3.4)}$	$52.3_{(6.4)}$	$\mathbf{66.6}_{(4.5)}$	$\mathbf{79.5}_{(5.9)}$	$81.0_{(3.9)}$
	ABMIL	Scratch	$55.8_{(3.2)}$	$57.9_{(3.7)}$	$65.0_{(3.3)}$	$67.0_{(4.8)}$	$73.9_{(4.4)}$
		Pretrain	$\mathbf{58.9}_{(7.1)}$	$\mathbf{63.9}_{(7.4)}$	$\mathbf{68.1}_{(3.5)}$	$\mathbf{73.2}_{(5.8)}$	$\mathbf{78.5}_{(4.2)}$
	AEMMIL	Scratch	$54.6_{(7.7)}$	$60.1_{(3.4)}$	$67.4_{(6.2)}$	$69.4_{(4.3)}$	$\mathbf{80.5}_{(2.7)}$
		Pretrain	$\mathbf{62.1}_{(6.7)}$	$\mathbf{63.7}_{(4.9)}$	$\mathbf{73.4}_{(4.0)}$	$\mathbf{77.3}_{(3.4)}$	$79.8_{(3.9)}$
	AMDMIL	Scratch	$58.1_{(5.4)}$	$58.5_{(1.8)}$	$64.0_{(4.5)}$	$70.6_{(4.9)}$	$80.2_{(2.0)}$
		Pretrain	$\mathbf{64.0}_{(4.2)}$	$\mathbf{65.5}_{(3.0)}$	$\mathbf{70.8}_{(4.1)}$	$\mathbf{76.4}_{(3.5)}$	$\mathbf{80.7}_{(2.9)}$
	CLAM	Scratch	$57.6_{(4.1)}$	$56.8_{(8.0)}$	$62.9_{(7.2)}$	$68.3_{(6.3)}$	$76.6_{(2.5)}$
		Pretrain	$\mathbf{58.2}_{(5.5)}$	$\mathbf{56.8}_{(9.5)}$	$\mathbf{66.5}_{(7.2)}$	$\mathbf{69.0}_{(3.3)}$	$\mathbf{77.9}_{(5.0)}$
	GDFMIL	Scratch	$55.1_{(10.3)}$	$60.9_{(3.0)}$	$60.4_{(14.5)}$	$72.2_{(6.0)}$	$79.6_{(1.6)}$
		Pretrain	$\mathbf{62.5}_{(2.7)}$	$\mathbf{61.9}_{(5.9)}$	$\mathbf{74.9}_{(4.4)}$	$\mathbf{77.0}_{(5.0)}$	$\mathbf{82.3}_{(2.4)}$
	TransMIL	Scratch	$58.0_{(5.3)}$	$60.7_{(3.2)}$	$68.2_{(4.4)}$	$71.2_{(2.5)}$	$75.6_{(1.8)}$
		Pretrain	$\mathbf{63.2}_{(4.8)}$	$\mathbf{66.3}_{(3.7)}$	$\mathbf{72.3}_{(2.6)}$	$\mathbf{75.0}_{(4.5)}$	$\mathbf{79.1}_{(3.0)}$
	WiKG	Scratch	$57.1_{(6.2)}$	$59.2_{(2.2)}$	$63.9_{(3.3)}$	$63.9_{(4.3)}$	$76.7_{(3.0)}$
		Pretrain	$\mathbf{65.2}_{(4.2)}$	$\mathbf{68.7}_{(2.7)}$	$\mathbf{70.5}_{(7.0)}$	$\mathbf{77.3}_{(4.1)}$	$\mathbf{79.1}_{(2.2)}$
CPTAC-NSCLC	2DMamba	Scratch	$56.7_{(18.0)}$	$78.0_{(6.9)}$	$86.0_{(6.5)}$	$93.1_{(4.0)}$	$95.4_{(4.2)}$
		Pretrain	$\mathbf{77.5}_{(22.4)}$	$\mathbf{91.0}_{(4.0)}$	$\mathbf{93.6}_{(3.8)}$	$\mathbf{96.0}_{(2.4)}$	$\mathbf{97.0}_{(2.9)}$
	ABMIL	Scratch	$62.9_{(12.6)}$	$78.1_{(8.6)}$	$85.8_{(6.6)}$	$89.9_{(3.2)}$	$93.3_{(3.8)}$
		Pretrain	$\mathbf{63.8}_{(17.4)}$	$\mathbf{81.6}_{(10.7)}$	$\mathbf{89.2}_{(4.6)}$	$\mathbf{94.6}_{(3.1)}$	$\mathbf{95.4}_{(1.7)}$
	AEMMIL	Scratch	$76.9_{(9.0)}$	$85.1_{(8.4)}$	$88.7_{(9.0)}$	$96.4_{(2.8)}$	$97.6_{(3.4)}$
		Pretrain	$\mathbf{78.9}_{(18.5)}$	$\mathbf{89.3}_{(5.4)}$	$\mathbf{93.1}_{(4.5)}$	$\mathbf{97.5}_{(1.3)}$	$\mathbf{98.4}_{(1.4)}$
	AMDMIL	Scratch	$69.6_{(24.8)}$	$84.4_{(3.8)}$	$86.7_{(9.5)}$	$92.8_{(2.0)}$	$94.4_{(2.6)}$
		Pretrain	$\mathbf{76.9}_{(20.8)}$	$\mathbf{91.1}_{(4.0)}$	$\mathbf{94.1}_{(3.2)}$	$\mathbf{96.9}_{(2.0)}$	$\mathbf{98.0}_{(1.2)}$
	CLAM	Scratch	$76.0_{(18.3)}$	$\mathbf{88.1}_{(6.5)}$	$90.6_{(3.5)}$	$\mathbf{94.2}_{(4.5)}$	$\mathbf{95.5}_{(3.1)}$
		Pretrain	$\mathbf{77.1}_{(19.6)}$	$87.7_{(8.1)}$	$\mathbf{92.4}_{(3.8)}$	$93.0_{(4.4)}$	$95.3_{(2.4)}$
	GDFMIL	Scratch	$70.9_{(11.8)}$	$68.4_{(12.1)}$	$86.8_{(7.3)}$	$92.0_{(4.4)}$	$95.9_{(1.8)}$
		Pretrain	$\mathbf{87.2}_{(6.1)}$	$\mathbf{93.0}_{(4.1)}$	$\mathbf{94.7}_{(2.2)}$	$\mathbf{98.1}_{(1.4)}$	$\mathbf{98.7}_{(1.6)}$
	TransMIL	Scratch	$\mathbf{79.7}_{(12.4)}$	$88.0_{(3.5)}$	$88.9_{(6.6)}$	$93.8_{(2.2)}$	$95.4_{(2.8)}$
		Pretrain	$78.8_{(15.7)}$	$\mathbf{89.4}_{(4.7)}$	$\mathbf{94.2}_{(3.1)}$	$\mathbf{96.8}_{(2.8)}$	$\mathbf{98.5}_{(1.6)}$
	WiKG	Scratch	$75.3_{(13.2)}$	$83.6_{(6.6)}$	$85.0_{(8.8)}$	$91.7_{(2.1)}$	$93.7_{(2.9)}$
		Pretrain	$\mathbf{93.3}_{(2.5)}$	$\mathbf{94.3}_{(4.3)}$	$\mathbf{96.5}_{(2.5)}$	$\mathbf{98.7}_{(0.6)}$	$\mathbf{98.1}_{(1.5)}$
KidRare-Coarse	2DMamba	Scratch	$62.0_{(8.5)}$	$53.4_{(5.3)}$	$79.8_{(8.8)}$	$91.5_{(2.9)}$	$94.7_{(0.7)}$
		Pretrain	$\mathbf{68.0}_{(11.6)}$	$\mathbf{61.9}_{(14.9)}$	$\mathbf{86.9}_{(1.0)}$	$\mathbf{93.5}_{(1.2)}$	$\mathbf{95.9}_{(0.8)}$
	ABMIL	Scratch	$66.2_{(11.7)}$	$77.0_{(7.6)}$	$85.0_{(4.5)}$	$91.5_{(3.6)}$	$94.2_{(2.1)}$
		Pretrain	$\mathbf{72.2}_{(14.5)}$	$\mathbf{80.5}_{(10.6)}$	$\mathbf{87.7}_{(4.8)}$	$\mathbf{92.5}_{(2.3)}$	$\mathbf{95.8}_{(1.5)}$
	AEMMIL	Scratch	$64.2_{(9.3)}$	$74.3_{(7.2)}$	$83.2_{(1.6)}$	$92.1_{(3.0)}$	$95.7_{(0.9)}$
		Pretrain	$\mathbf{67.8}_{(12.1)}$	$\mathbf{80.7}_{(9.8)}$	$\mathbf{89.5}_{(2.5)}$	$\mathbf{94.3}_{(1.7)}$	$\mathbf{96.6}_{(0.9)}$
	AMDMIL	Scratch	$69.6_{(8.5)}$	$78.8_{(5.9)}$	$86.1_{(1.2)}$	$91.6_{(1.5)}$	$95.7_{(1.3)}$
		Pretrain	$\mathbf{74.9}_{(5.7)}$	$\mathbf{84.3}_{(4.2)}$	$\mathbf{88.3}_{(1.8)}$	$\mathbf{94.0}_{(1.2)}$	$\mathbf{97.0}_{(0.5)}$
	CLAM	Scratch	$63.3_{(14.2)}$	$71.0_{(12.5)}$	$80.2_{(10.7)}$	$86.2_{(6.4)}$	$92.1_{(4.3)}$
		Pretrain	$\mathbf{63.6}_{(15.4)}$	$\mathbf{79.9}_{(10.5)}$	$\mathbf{84.7}_{(6.6)}$	$\mathbf{90.2}_{(2.8)}$	$\mathbf{94.6}_{(1.7)}$
	GDFMIL	Scratch	$64.5_{(8.8)}$	$73.6_{(9.2)}$	$81.0_{(6.2)}$	$92.5_{(2.2)}$	$95.5_{(0.6)}$
		Pretrain	$\mathbf{72.7}_{(13.6)}$	$\mathbf{86.8}_{(3.4)}$	$\mathbf{91.4}_{(2.5)}$	$\mathbf{94.4}_{(1.4)}$	$\mathbf{96.7}_{(1.0)}$
	TransMIL	Scratch	$66.8_{(10.3)}$	$75.4_{(5.9)}$	$80.4_{(3.7)}$	$89.5_{(2.9)}$	$94.6_{(1.2)}$
		Pretrain	$\mathbf{67.4}_{(15.2)}$	$\mathbf{83.0}_{(7.3)}$	$\mathbf{90.2}_{(1.8)}$	$\mathbf{95.2}_{(1.6)}$	$\mathbf{97.4}_{(0.6)}$
	WiKG	Scratch	$66.1_{(7.3)}$	$75.3_{(7.4)}$	$83.1_{(1.3)}$	$87.7_{(2.7)}$	$93.1_{(0.6)}$
		Pretrain	$\mathbf{73.7}_{(9.1)}$	$\mathbf{82.7}_{(8.8)}$	$\mathbf{89.4}_{(2.1)}$	$\mathbf{94.2}_{(2.2)}$	$\mathbf{96.7}_{(0.7)}$
KidRare-Fine	2DMamba	Scratch	$\mathbf{72.7}_{(3.0)}$	$76.1_{(1.7)}$	$85.8_{(1.7)}$	$90.0_{(1.1)}$	$92.3_{(1.0)}$
		Pretrain	$71.0_{(7.3)}$	$\mathbf{78.4}_{(2.0)}$	$\mathbf{86.9}_{(2.2)}$	$\mathbf{90.6}_{(1.0)}$	$\mathbf{92.4}_{(1.1)}$
	ABMIL	Scratch	$73.7_{(5.7)}$	$78.6_{(5.4)}$	$84.2_{(3.8)}$	$90.2_{(1.2)}$	$90.9_{(1.4)}$
		Pretrain	$\mathbf{75.5}_{(4.0)}$	$\mathbf{82.3}_{(2.7)}$	$\mathbf{87.1}_{(3.7)}$	$\mathbf{90.7}_{(0.6)}$	$\mathbf{91.5}_{(1.0)}$
	AEMMIL	Scratch	$71.5_{(1.1)}$	$79.5_{(3.9)}$	$\mathbf{87.6}_{(2.7)}$	$\mathbf{90.9}_{(1.5)}$	$\mathbf{92.7}_{(1.0)}$
		Pretrain	$\mathbf{75.8}_{(0.9)}$	$\mathbf{82.6}_{(1.8)}$	$87.4_{(2.7)}$	$90.8_{(1.7)}$	$92.2_{(1.7)}$
	AMDMIL	Scratch	$75.0_{(3.6)}$	$82.3_{(1.9)}$	$87.2_{(2.0)}$	$90.9_{(1.0)}$	$\mathbf{92.8}_{(0.9)}$
		Pretrain	$\mathbf{78.6}_{(1.1)}$	$\mathbf{83.6}_{(1.3)}$	$\mathbf{87.5}_{(2.4)}$	$\mathbf{91.0}_{(1.1)}$	$92.4_{(1.3)}$
	CLAM	Scratch	$72.2_{(3.1)}$	$77.3_{(1.7)}$	$\mathbf{85.0}_{(2.9)}$	$\mathbf{88.7}_{(0.4)}$	$91.1_{(0.9)}$
		Pretrain	$\mathbf{72.7}_{(6.8)}$	$\mathbf{81.0}_{(2.5)}$	$84.3_{(1.8)}$	$88.5_{(2.5)}$	$\mathbf{91.5}_{(2.1)}$
	GDFMIL	Scratch	$69.0_{(4.3)}$	$78.5_{(3.1)}$	$88.1_{(2.0)}$	$\mathbf{91.5}_{(1.2)}$	$91.5_{(2.4)}$
		Pretrain	$\mathbf{78.1}_{(2.9)}$	$\mathbf{81.9}_{(3.9)}$	$\mathbf{88.8}_{(2.5)}$	$91.4_{(0.8)}$	$\mathbf{93.0}_{(1.2)}$
	TransMIL	Scratch	$74.7_{(1.9)}$	$81.1_{(1.1)}$	$85.6_{(1.2)}$	$89.2_{(0.9)}$	$91.8_{(1.0)}$
		Pretrain	$\mathbf{80.6}_{(1.5)}$	$\mathbf{85.7}_{(2.0)}$	$\mathbf{88.9}_{(2.0)}$	$\mathbf{91.2}_{(1.2)}$	$\mathbf{92.9}_{(1.3)}$
	WiKG	Scratch	$75.8_{(0.9)}$	$80.4_{(0.9)}$	$85.0_{(2.0)}$	$88.5_{(0.8)}$	$91.7_{(1.1)}$
		Pretrain	$\mathbf{78.8}_{(2.0)}$	$\mathbf{83.8}_{(1.5)}$	$\mathbf{87.5}_{(2.0)}$	$\mathbf{90.7}_{(1.2)}$	$\mathbf{93.3}_{(0.6)}$

Table 16: Few-shot comparison between from-scratch training and distillation initialization measured by F1. Mean and standard deviation are computed across seeds. DAGMIL is excluded because only scratch results are available, and 32-shot results are omitted.

Dataset	MIL	Setting	1-shot	2-shot	4-shot	8-shot	16-shot
BRACS-Coarse	2DMamba	Scratch	$25.2_{(5.6)}$	$26.2_{(6.6)}$	$37.8_{(8.1)}$	$49.6_{(10.2)}$	$59.9_{(5.4)}$
		Pretrain	$\mathbf{45.7}_{(13.2)}$	$\mathbf{36.4}_{(13.6)}$	$\mathbf{57.6}_{(9.0)}$	$\mathbf{66.0}_{(8.3)}$	$\mathbf{73.4}_{(5.3)}$
	ABMIL	Scratch	$\mathbf{30.0}_{(8.8)}$	$29.7_{(13.0)}$	$\mathbf{45.8}_{(15.6)}$	$42.2_{(12.7)}$	$54.2_{(8.9)}$
		Pretrain	$27.8_{(10.9)}$	$\mathbf{32.8}_{(15.8)}$	$45.4_{(14.9)}$	$\mathbf{48.0}_{(13.5)}$	$\mathbf{58.2}_{(6.0)}$
	AEMMIL	Scratch	$33.5_{(11.8)}$	$32.6_{(11.0)}$	$41.3_{(16.6)}$	$51.1_{(15.1)}$	$60.7_{(8.9)}$
		Pretrain	$\mathbf{40.0}_{(12.3)}$	$\mathbf{49.9}_{(19.7)}$	$\mathbf{53.7}_{(15.5)}$	$\mathbf{57.9}_{(13.2)}$	$\mathbf{66.3}_{(7.9)}$
	AMDMIL	Scratch	$31.2_{(7.8)}$	$34.2_{(5.8)}$	$36.2_{(11.6)}$	$51.0_{(12.6)}$	$59.1_{(7.3)}$
		Pretrain	$\mathbf{44.1}_{(11.7)}$	$\mathbf{56.3}_{(7.8)}$	$\mathbf{58.7}_{(7.6)}$	$\mathbf{57.0}_{(4.1)}$	$\mathbf{68.4}_{(5.2)}$
	CLAM	Scratch	$\mathbf{17.2}_{(4.3)}$	$21.2_{(7.1)}$	$35.2_{(16.3)}$	$44.8_{(10.0)}$	$56.6_{(9.2)}$
		Pretrain	$16.5_{(3.2)}$	$\mathbf{45.0}_{(15.9)}$	$\mathbf{47.1}_{(7.7)}$	$\mathbf{56.9}_{(6.1)}$	$\mathbf{57.1}_{(8.2)}$
	GDFMIL	Scratch	$29.4_{(8.1)}$	$30.6_{(5.4)}$	$27.0_{(14.5)}$	$46.8_{(16.8)}$	$62.1_{(8.9)}$
		Pretrain	$\mathbf{38.8}_{(12.9)}$	$\mathbf{46.1}_{(9.1)}$	$\mathbf{53.7}_{(14.0)}$	$\mathbf{65.0}_{(8.0)}$	$\mathbf{69.7}_{(9.8)}$
	TransMIL	Scratch	$34.2_{(9.2)}$	$30.1_{(11.7)}$	$39.5_{(12.1)}$	$47.4_{(15.3)}$	$57.7_{(5.5)}$
		Pretrain	$\mathbf{45.1}_{(7.6)}$	$\mathbf{57.2}_{(6.3)}$	$\mathbf{53.0}_{(13.5)}$	$\mathbf{61.5}_{(8.0)}$	$\mathbf{69.9}_{(3.1)}$
	WiKG	Scratch	$33.7_{(6.1)}$	$37.1_{(7.3)}$	$37.1_{(5.8)}$	$46.0_{(10.9)}$	$55.9_{(5.7)}$
		Pretrain	$\mathbf{48.3}_{(15.7)}$	$\mathbf{59.3}_{(16.3)}$	$\mathbf{59.6}_{(15.4)}$	$\mathbf{56.9}_{(11.0)}$	$\mathbf{61.9}_{(4.6)}$
BRACS-Fine	2DMamba	Scratch	$9.0_{(3.3)}$	$\mathbf{11.7}_{(4.3)}$	$27.9_{(11.2)}$	$37.2_{(5.9)}$	$\mathbf{44.6}_{(4.6)}$
		Pretrain	$\mathbf{19.9}_{(6.3)}$	$9.0_{(5.9)}$	$\mathbf{30.2}_{(6.9)}$	$\mathbf{38.9}_{(5.3)}$	$43.0_{(5.0)}$
	ABMIL	Scratch	$12.8_{(6.0)}$	$13.4_{(6.2)}$	$25.6_{(8.3)}$	$30.2_{(3.3)}$	$34.2_{(6.0)}$
		Pretrain	$\mathbf{13.2}_{(7.2)}$	$\mathbf{22.2}_{(12.5)}$	$\mathbf{29.1}_{(7.3)}$	$\mathbf{34.6}_{(4.7)}$	$\mathbf{40.2}_{(7.5)}$
	AEMMIL	Scratch	$14.8_{(8.8)}$	$20.2_{(6.4)}$	$31.3_{(4.1)}$	$33.9_{(6.6)}$	$\mathbf{42.7}_{(5.3)}$
		Pretrain	$\mathbf{19.8}_{(10.1)}$	$\mathbf{26.8}_{(6.1)}$	$\mathbf{34.9}_{(4.7)}$	$\mathbf{40.0}_{(1.7)}$	$40.7_{(1.8)}$
	AMDMIL	Scratch	$10.6_{(3.5)}$	$13.7_{(6.0)}$	$27.4_{(4.3)}$	$30.0_{(9.1)}$	$\mathbf{41.1}_{(3.7)}$
		Pretrain	$\mathbf{22.4}_{(6.6)}$	$\mathbf{23.4}_{(4.9)}$	$\mathbf{29.5}_{(5.0)}$	$\mathbf{36.2}_{(4.7)}$	$37.5_{(3.5)}$
	CLAM	Scratch	$16.2_{(7.1)}$	$\mathbf{21.0}_{(11.8)}$	$24.8_{(10.8)}$	$31.0_{(6.4)}$	$35.5_{(2.9)}$
		Pretrain	$\mathbf{19.5}_{(7.2)}$	$13.6_{(10.2)}$	$\mathbf{30.6}_{(9.1)}$	$\mathbf{34.4}_{(1.9)}$	$\mathbf{38.5}_{(5.2)}$
	GDFMIL	Scratch	$10.9_{(5.1)}$	$20.3_{(6.8)}$	$23.8_{(14.9)}$	$31.0_{(5.7)}$	$36.7_{(5.3)}$
		Pretrain	$\mathbf{21.7}_{(8.9)}$	$\mathbf{23.6}_{(6.1)}$	$\mathbf{35.1}_{(4.2)}$	$\mathbf{39.3}_{(4.6)}$	$\mathbf{38.0}_{(2.7)}$
	TransMIL	Scratch	$18.1_{(5.7)}$	$18.7_{(6.5)}$	$33.2_{(7.5)}$	$34.1_{(5.3)}$	$38.6_{(2.5)}$
		Pretrain	$\mathbf{24.1}_{(4.9)}$	$\mathbf{34.9}_{(4.7)}$	$\mathbf{34.2}_{(4.4)}$	$\mathbf{37.3}_{(3.7)}$	$\mathbf{41.9}_{(6.6)}$
	WiKG	Scratch	$13.1_{(5.6)}$	$16.5_{(6.9)}$	$25.0_{(3.0)}$	$19.0_{(7.6)}$	$\mathbf{34.2}_{(7.7)}$
		Pretrain	$\mathbf{23.4}_{(7.9)}$	$\mathbf{31.0}_{(1.6)}$	$\mathbf{33.1}_{(4.1)}$	$\mathbf{37.4}_{(4.5)}$	$34.1_{(3.0)}$
CPTAC-NSCLC	2DMamba	Scratch	$51.6_{(16.7)}$	$71.3_{(9.9)}$	$77.6_{(5.0)}$	$85.9_{(6.1)}$	$87.8_{(5.6)}$
		Pretrain	$\mathbf{68.9}_{(16.1)}$	$\mathbf{81.2}_{(5.0)}$	$\mathbf{86.1}_{(3.9)}$	$\mathbf{89.1}_{(2.8)}$	$\mathbf{92.0}_{(2.8)}$
	ABMIL	Scratch	$\mathbf{49.2}_{(15.7)}$	$70.5_{(10.2)}$	$80.5_{(5.3)}$	$81.5_{(4.4)}$	$86.1_{(3.9)}$
		Pretrain	$47.8_{(15.0)}$	$\mathbf{74.9}_{(11.2)}$	$\mathbf{81.5}_{(4.8)}$	$\mathbf{87.3}_{(4.6)}$	$\mathbf{88.8}_{(3.5)}$
	AEMMIL	Scratch	$61.2_{(17.0)}$	$73.5_{(9.8)}$	$82.5_{(10.8)}$	$89.8_{(4.7)}$	$\mathbf{93.0}_{(5.9)}$
		Pretrain	$\mathbf{71.7}_{(15.9)}$	$\mathbf{78.7}_{(6.6)}$	$\mathbf{85.5}_{(5.4)}$	$\mathbf{91.2}_{(3.0)}$	$92.8_{(2.3)}$
	AMDMIL	Scratch	$62.7_{(16.9)}$	$75.0_{(2.5)}$	$78.2_{(8.4)}$	$86.1_{(3.1)}$	$87.8_{(3.4)}$
		Pretrain	$\mathbf{72.2}_{(15.8)}$	$\mathbf{81.9}_{(5.8)}$	$\mathbf{87.5}_{(4.5)}$	$\mathbf{89.2}_{(7.0)}$	$\mathbf{91.2}_{(3.6)}$
	CLAM	Scratch	$67.4_{(18.2)}$	$\mathbf{79.8}_{(6.9)}$	$80.5_{(4.5)}$	$\mathbf{87.0}_{(7.6)}$	$86.8_{(9.2)}$
		Pretrain	$\mathbf{68.2}_{(17.4)}$	$79.0_{(6.9)}$	$\mathbf{83.1}_{(3.5)}$	$86.4_{(4.8)}$	$\mathbf{88.6}_{(2.5)}$
	GDFMIL	Scratch	$59.2_{(9.7)}$	$56.9_{(14.4)}$	$77.9_{(8.6)}$	$83.8_{(6.0)}$	$89.1_{(3.5)}$
		Pretrain	$\mathbf{76.3}_{(11.6)}$	$\mathbf{85.1}_{(5.0)}$	$\mathbf{85.9}_{(6.9)}$	$\mathbf{93.3}_{(4.2)}$	$\mathbf{94.3}_{(3.7)}$
	TransMIL	Scratch	$\mathbf{71.7}_{(17.5)}$	$81.4_{(5.6)}$	$83.1_{(5.3)}$	$87.0_{(2.9)}$	$89.0_{(4.4)}$
		Pretrain	$71.4_{(15.5)}$	$\mathbf{81.6}_{(5.9)}$	$\mathbf{86.1}_{(4.4)}$	$\mathbf{91.3}_{(3.8)}$	$\mathbf{93.3}_{(3.0)}$
	WiKG	Scratch	$67.6_{(13.9)}$	$75.9_{(6.7)}$	$78.4_{(9.2)}$	$85.3_{(2.4)}$	$88.2_{(3.2)}$
		Pretrain	$\mathbf{83.9}_{(5.2)}$	$\mathbf{86.1}_{(6.6)}$	$\mathbf{89.8}_{(3.4)}$	$\mathbf{93.3}_{(3.5)}$	$\mathbf{92.2}_{(3.0)}$
KidRare-Coarse	2DMamba	Scratch	$38.8_{(5.3)}$	$24.0_{(11.1)}$	$55.8_{(11.0)}$	$75.6_{(5.8)}$	$79.2_{(2.4)}$
		Pretrain	$\mathbf{42.0}_{(12.5)}$	$\mathbf{25.4}_{(17.6)}$	$\mathbf{61.5}_{(1.5)}$	$\mathbf{76.6}_{(4.3)}$	$\mathbf{82.0}_{(2.4)}$
	ABMIL	Scratch	$39.6_{(10.6)}$	$50.9_{(10.0)}$	$64.9_{(5.6)}$	$77.6_{(5.1)}$	$81.2_{(5.3)}$
		Pretrain	$\mathbf{41.1}_{(19.4)}$	$\mathbf{58.0}_{(14.9)}$	$\mathbf{68.0}_{(6.0)}$	$\mathbf{78.0}_{(4.7)}$	$\mathbf{83.0}_{(4.3)}$
	AEMMIL	Scratch	$35.3_{(6.2)}$	$46.1_{(11.3)}$	$59.6_{(4.0)}$	$75.3_{(5.0)}$	$82.3_{(2.2)}$
		Pretrain	$\mathbf{41.4}_{(9.0)}$	$\mathbf{54.9}_{(14.6)}$	$\mathbf{66.1}_{(5.6)}$	$\mathbf{80.7}_{(4.4)}$	$\mathbf{83.6}_{(3.0)}$
	AMDMIL	Scratch	$36.0_{(7.6)}$	$54.6_{(8.9)}$	$63.7_{(2.2)}$	$76.0_{(1.6)}$	$82.2_{(2.4)}$
		Pretrain	$\mathbf{40.5}_{(9.6)}$	$\mathbf{57.9}_{(11.1)}$	$\mathbf{65.0}_{(4.5)}$	$\mathbf{79.5}_{(2.8)}$	$\mathbf{85.3}_{(2.1)}$
	CLAM	Scratch	$33.9_{(12.5)}$	$45.0_{(9.4)}$	$58.0_{(13.5)}$	$68.1_{(9.6)}$	$77.2_{(6.9)}$
		Pretrain	$\mathbf{34.5}_{(14.9)}$	$\mathbf{53.4}_{(10.0)}$	$\mathbf{63.2}_{(9.1)}$	$\mathbf{74.8}_{(4.4)}$	$\mathbf{80.3}_{(3.9)}$
	GDFMIL	Scratch	$35.0_{(10.2)}$	$45.8_{(10.9)}$	$56.3_{(7.4)}$	$74.0_{(6.4)}$	$80.2_{(3.4)}$
		Pretrain	$\mathbf{47.8}_{(13.5)}$	$\mathbf{61.6}_{(8.4)}$	$\mathbf{71.4}_{(4.4)}$	$\mathbf{80.7}_{(3.5)}$	$\mathbf{84.0}_{(4.3)}$
	TransMIL	Scratch	$\mathbf{42.1}_{(8.8)}$	$46.9_{(10.7)}$	$56.1_{(4.7)}$	$71.2_{(5.0)}$	$81.0_{(3.5)}$
		Pretrain	$41.4_{(15.9)}$	$\mathbf{54.2}_{(9.6)}$	$\mathbf{68.4}_{(2.5)}$	$\mathbf{81.0}_{(3.2)}$	$\mathbf{86.5}_{(1.1)}$
	WiKG	Scratch	$40.0_{(6.2)}$	$50.2_{(9.7)}$	$59.5_{(2.4)}$	$67.9_{(2.5)}$	$77.4_{(2.7)}$
		Pretrain	$\mathbf{40.4}_{(12.4)}$	$\mathbf{58.7}_{(15.4)}$	$\mathbf{65.5}_{(8.8)}$	$\mathbf{81.0}_{(2.7)}$	$\mathbf{83.7}_{(3.5)}$
KidRare-Fine	2DMamba	Scratch	$20.7_{(4.1)}$	$\mathbf{6.1}_{(4.9)}$	$\mathbf{39.1}_{(4.5)}$	$43.0_{(3.3)}$	$\mathbf{52.5}_{(1.6)}$
		Pretrain	$\mathbf{20.8}_{(9.2)}$	$5.1_{(4.1)}$	$38.8_{(6.9)}$	$\mathbf{45.8}_{(6.2)}$	$50.8_{(5.2)}$
	ABMIL	Scratch	$\mathbf{23.2}_{(10.2)}$	$23.0_{(7.4)}$	$32.9_{(6.3)}$	$45.3_{(6.9)}$	$49.0_{(8.2)}$
		Pretrain	$21.1_{(10.9)}$	$\mathbf{27.3}_{(6.9)}$	$\mathbf{34.1}_{(13.4)}$	$\mathbf{46.4}_{(6.3)}$	$\mathbf{49.2}_{(5.6)}$
	AEMMIL	Scratch	$19.6_{(4.5)}$	$27.7_{(8.9)}$	$37.7_{(6.6)}$	$\mathbf{50.0}_{(5.5)}$	$\mathbf{49.7}_{(5.4)}$
		Pretrain	$\mathbf{23.3}_{(4.9)}$	$\mathbf{31.1}_{(8.9)}$	$\mathbf{39.2}_{(7.0)}$	$49.1_{(8.5)}$	$48.8_{(8.8)}$
	AMDMIL	Scratch	$21.0_{(5.6)}$	$31.6_{(5.2)}$	$\mathbf{40.8}_{(5.6)}$	$46.2_{(5.1)}$	$51.3_{(4.3)}$
		Pretrain	$\mathbf{28.4}_{(2.9)}$	$\mathbf{36.4}_{(5.4)}$	$39.4_{(7.3)}$	$\mathbf{49.2}_{(4.3)}$	$\mathbf{55.5}_{(3.5)}$
	CLAM	Scratch	$\mathbf{22.4}_{(5.8)}$	$25.9_{(2.2)}$	$34.8_{(2.9)}$	$\mathbf{42.7}_{(4.9)}$	$48.2_{(3.3)}$
		Pretrain	$21.9_{(9.2)}$	$\mathbf{27.4}_{(6.5)}$	$\mathbf{35.8}_{(4.5)}$	$38.7_{(8.5)}$	$\mathbf{49.8}_{(4.9)}$
	GDFMIL	Scratch	$14.6_{(4.9)}$	$25.2_{(5.7)}$	$33.3_{(4.5)}$	$42.6_{(6.7)}$	$40.6_{(5.3)}$
		Pretrain	$\mathbf{24.6}_{(6.6)}$	$\mathbf{27.0}_{(5.0)}$	$\mathbf{43.6}_{(6.9)}$	$\mathbf{44.6}_{(8.9)}$	$\mathbf{47.0}_{(8.6)}$
	TransMIL	Scratch	$23.9_{(3.3)}$	$31.1_{(5.7)}$	$38.0_{(2.6)}$	$43.2_{(3.7)}$	$50.0_{(2.7)}$
		Pretrain	$\mathbf{29.5}_{(7.6)}$	$\mathbf{32.6}_{(8.3)}$	$\mathbf{40.4}_{(6.3)}$	$\mathbf{51.8}_{(5.5)}$	$\mathbf{52.8}_{(3.9)}$
	WiKG	Scratch	$21.8_{(6.6)}$	$25.2_{(5.6)}$	$31.0_{(3.7)}$	$40.1_{(5.4)}$	$47.4_{(3.4)}$
		Pretrain	$\mathbf{26.1}_{(5.8)}$	$\mathbf{29.8}_{(4.7)}$	$\mathbf{40.0}_{(8.8)}$	$\mathbf{48.6}_{(5.8)}$	$\mathbf{52.6}_{(7.5)}$

A.6 Results of multi-teacher distillation

Table 17: Ablation results of different teacher settings. Mean and standard deviation are computed over five seeds. The best result for each dataset–MIL pair is highlighted in bold.

Dataset	MIL	Teacher	Accuracy	bacc	AUC	F1
BCNB-ER	2DMamba	CARE-only	$82.0_{(2.1)}$	$76.6_{(1.3)}$	$86.4_{(1.1)}$	$82.5_{(1.7)}$
		TITAN-only	$82.3_{(2.2)}$	$74.8_{(1.9)}$	$86.9_{(1.9)}$	$82.4_{(1.6)}$
		Both	$82.7_{(1.6)}$	$74.2_{(3.6)}$	$86.2_{(1.7)}$	$82.6_{(1.7)}$
	CLAM	CARE-only	$82.8_{(2.0)}$	$75.0_{(4.5)}$	$85.2_{(2.5)}$	$82.8_{(2.3)}$
		TITAN-only	$81.7_{(2.5)}$	$73.9_{(3.7)}$	$84.8_{(2.9)}$	$81.8_{(2.4)}$
		Both	$80.6_{(2.2)}$	$75.5_{(3.0)}$	$84.8_{(2.4)}$	$81.2_{(1.9)}$
	TransMIL	CARE-only	$78.9_{(2.8)}$	$74.7_{(3.6)}$	$83.9_{(2.6)}$	$79.8_{(2.3)}$
		TITAN-only	$80.1_{(4.1)}$	$73.8_{(2.7)}$	$83.1_{(2.0)}$	$80.6_{(3.1)}$
		Both	$78.7_{(3.0)}$	$75.5_{(2.4)}$	$84.2_{(1.7)}$	$79.7_{(2.4)}$
	WiKG	CARE-only	$81.0_{(1.0)}$	$74.1_{(2.3)}$	$85.0_{(1.5)}$	$81.3_{(0.5)}$
		TITAN-only	$78.2_{(4.9)}$	$75.6_{(3.6)}$	$84.9_{(1.5)}$	$79.3_{(4.1)}$
		Both	$78.8_{(2.5)}$	$75.8_{(4.7)}$	$84.4_{(2.9)}$	$79.9_{(2.5)}$
BRACS-Fine	2DMamba	CARE-only	$52.4_{(3.9)}$	$46.3_{(3.6)}$	$83.5_{(1.2)}$	$48.1_{(4.4)}$
		TITAN-only	$50.3_{(2.5)}$	$45.2_{(2.1)}$	$83.6_{(1.2)}$	$46.3_{(2.0)}$
		Both	$52.2_{(3.1)}$	$45.9_{(3.5)}$	$82.7_{(1.8)}$	$49.6_{(4.0)}$
	CLAM	CARE-only	$48.3_{(2.9)}$	$40.8_{(3.1)}$	$78.8_{(2.3)}$	$44.4_{(3.0)}$
		TITAN-only	$48.5_{(3.8)}$	$40.6_{(4.9)}$	$80.4_{(1.8)}$	$42.9_{(3.4)}$
		Both	$47.4_{(4.3)}$	$41.4_{(4.0)}$	$80.4_{(1.4)}$	$42.3_{(3.0)}$
	TransMIL	CARE-only	$46.0_{(3.3)}$	$40.0_{(3.2)}$	$80.5_{(1.1)}$	$42.3_{(1.6)}$
		TITAN-only	$44.6_{(3.2)}$	$37.2_{(4.3)}$	$80.1_{(3.5)}$	$41.2_{(3.3)}$
		Both	$50.1_{(5.0)}$	$43.7_{(4.2)}$	$81.4_{(1.8)}$	$45.6_{(3.2)}$
	WiKG	CARE-only	$45.3_{(5.4)}$	$39.7_{(4.4)}$	$82.2_{(1.4)}$	$40.1_{(4.3)}$
		TITAN-only	$43.7_{(6.2)}$	$37.9_{(3.8)}$	$81.9_{(0.7)}$	$38.4_{(6.5)}$
		Both	$42.1_{(5.2)}$	$35.8_{(4.2)}$	$81.9_{(1.8)}$	$37.8_{(4.4)}$
EBrains-IDH	2DMamba	CARE-only	$93.3_{(1.0)}$	$92.9_{(1.2)}$	$98.4_{(0.6)}$	$93.3_{(1.0)}$
		TITAN-only	$93.9_{(1.6)}$	$93.7_{(1.7)}$	$99.0_{(0.2)}$	$93.9_{(1.6)}$
		Both	$93.2_{(1.4)}$	$93.2_{(1.6)}$	$98.9_{(0.2)}$	$93.2_{(1.4)}$
	CLAM	CARE-only	$91.1_{(2.1)}$	$91.2_{(2.0)}$	$97.5_{(1.0)}$	$91.1_{(2.1)}$
		TITAN-only	$91.6_{(1.8)}$	$91.9_{(1.6)}$	$97.8_{(0.5)}$	$91.6_{(1.8)}$
		Both	$91.4_{(1.8)}$	$91.3_{(2.0)}$	$97.5_{(0.5)}$	$91.5_{(1.8)}$
	TransMIL	CARE-only	$91.3_{(2.1)}$	$91.9_{(1.8)}$	$98.3_{(0.6)}$	$91.4_{(2.0)}$
		TITAN-only	$90.8_{(2.0)}$	$90.9_{(2.2)}$	$97.8_{(0.3)}$	$90.8_{(2.0)}$
		Both	$91.4_{(1.7)}$	$92.5_{(1.6)}$	$98.4_{(0.2)}$	$91.5_{(1.6)}$
	WiKG	CARE-only	$93.4_{(1.2)}$	$93.3_{(1.1)}$	$98.6_{(0.3)}$	$93.4_{(1.2)}$
		TITAN-only	$90.2_{(3.0)}$	$90.5_{(2.9)}$	$97.6_{(1.3)}$	$90.2_{(3.0)}$
		Both	$91.2_{(1.4)}$	$91.4_{(1.0)}$	$98.2_{(0.3)}$	$91.3_{(1.4)}$
KidRare-Coarse	2DMamba	CARE-only	$90.9_{(2.1)}$	$89.8_{(3.1)}$	$98.5_{(0.5)}$	$90.8_{(2.1)}$
		TITAN-only	$91.6_{(0.7)}$	$90.8_{(0.9)}$	$98.9_{(0.2)}$	$91.6_{(0.7)}$
		Both	$90.6_{(1.6)}$	$89.3_{(1.8)}$	$98.9_{(0.3)}$	$90.6_{(1.6)}$
	CLAM	CARE-only	$92.9_{(1.6)}$	$91.7_{(2.5)}$	$98.8_{(0.6)}$	$92.8_{(1.6)}$
		TITAN-only	$92.3_{(1.1)}$	$91.0_{(1.6)}$	$98.5_{(0.5)}$	$92.3_{(1.1)}$
		Both	$90.8_{(2.3)}$	$88.2_{(3.1)}$	$98.3_{(0.7)}$	$90.7_{(2.3)}$
	TransMIL	CARE-only	$92.2_{(2.1)}$	$91.6_{(2.0)}$	$99.0_{(0.2)}$	$92.2_{(2.1)}$
		TITAN-only	$92.9_{(1.7)}$	$91.5_{(1.7)}$	$99.0_{(0.1)}$	$92.9_{(1.7)}$
		Both	$92.0_{(2.8)}$	$92.5_{(2.4)}$	$99.0_{(0.2)}$	$92.0_{(2.9)}$
	WiKG	CARE-only	$90.5_{(1.9)}$	$88.9_{(3.2)}$	$98.5_{(0.3)}$	$90.5_{(2.0)}$
		TITAN-only	$90.9_{(1.3)}$	$89.8_{(1.6)}$	$98.5_{(0.2)}$	$91.0_{(1.1)}$
		Both	$90.2_{(1.2)}$	$88.7_{(1.9)}$	$98.3_{(0.3)}$	$90.2_{(1.2)}$

A.7 Results of different type distillation loss

Table 18: Comparison of different distillation loss types across downstream datasets. All values are reported in percentage form, with standard deviations shown in the lower-right subscript.

Dataset	Metric	Angle	Cosine	MSE	Dataset	Metric	Angle	Cosine	MSE
bcnb-er	BACC	$76.0_{(0.9)}$	$75.0_{(2.4)}$	$74.9_{(1.3)}$	kidrare-fine	BACC	$45.4_{(2.0)}$	$45.9_{(1.9)}$	$46.9_{(1.9)}$
	AUC	$86.0_{(1.5)}$	$86.6_{(1.8)}$	$86.3_{(1.0)}$		AUC	$94.4_{(0.4)}$	$94.5_{(0.4)}$	$94.7_{(0.4)}$
	F1	$82.1_{(2.0)}$	$82.0_{(3.3)}$	$81.4_{(2.9)}$		F1	$64.2_{(1.4)}$	$64.8_{(1.3)}$	$64.6_{(1.7)}$
	Acc	$81.6_{(2.6)}$	$81.7_{(4.1)}$	$80.9_{(3.8)}$		Acc	$66.5_{(1.2)}$	$66.7_{(0.5)}$	$66.0_{(1.2)}$
bcnb-pr	BACC	$72.3_{(1.0)}$	$72.4_{(1.0)}$	$71.8_{(2.8)}$	ebrains-coarse	BACC	$84.5_{(2.1)}$	$86.3_{(1.4)}$	$83.5_{(2.8)}$
	AUC	$79.8_{(1.4)}$	$79.2_{(1.3)}$	$78.7_{(1.7)}$		AUC	$99.1_{(0.0)}$	$99.1_{(0.1)}$	$99.1_{(0.1)}$
	F1	$79.7_{(0.4)}$	$78.8_{(0.5)}$	$78.3_{(1.4)}$		F1	$90.3_{(0.9)}$	$90.9_{(1.3)}$	$90.1_{(0.6)}$
	Acc	$80.0_{(0.7)}$	$78.8_{(0.8)}$	$78.3_{(1.2)}$		Acc	$90.1_{(1.0)}$	$90.9_{(1.4)}$	$90.3_{(0.4)}$
bcnb-her2	BACC	$64.3_{(1.3)}$	$64.7_{(1.2)}$	$65.6_{(0.9)}$	ebrains-fine	BACC	$67.4_{(1.0)}$	$66.7_{(1.5)}$	$67.5_{(0.9)}$
	AUC	$71.6_{(0.6)}$	$72.5_{(1.6)}$	$71.7_{(0.3)}$		AUC	$98.4_{(0.0)}$	$98.4_{(0.1)}$	$98.4_{(0.0)}$
	F1	$73.3_{(2.0)}$	$73.0_{(2.5)}$	$74.2_{(2.3)}$		F1	$72.7_{(1.3)}$	$72.2_{(0.6)}$	$72.0_{(0.9)}$
	Acc	$73.9_{(2.8)}$	$73.2_{(3.3)}$	$74.6_{(3.2)}$		Acc	$74.0_{(1.0)}$	$73.6_{(0.8)}$	$73.7_{(0.7)}$
bracs-coarse	BACC	$62.3_{(1.5)}$	$65.5_{(3.5)}$	$66.2_{(2.4)}$	ebrains-idh	BACC	$91.6_{(1.5)}$	$92.1_{(1.6)}$	$92.5_{(1.7)}$
	AUC	$89.0_{(1.8)}$	$88.3_{(1.6)}$	$87.8_{(1.9)}$		AUC	$97.8_{(0.5)}$	$97.9_{(0.6)}$	$98.1_{(0.5)}$
	F1	$61.0_{(1.4)}$	$65.6_{(5.0)}$	$67.0_{(3.3)}$		F1	$91.8_{(1.3)}$	$92.5_{(1.6)}$	$92.7_{(1.5)}$
	Acc	$67.8_{(1.8)}$	$70.3_{(2.7)}$	$70.8_{(1.9)}$		Acc	$91.8_{(1.3)}$	$92.5_{(1.6)}$	$92.7_{(1.5)}$
bracs-fine	BACC	$40.9_{(3.6)}$	$42.2_{(2.9)}$	$41.9_{(3.0)}$	mut-het-rcc-bap1	BACC	$62.5_{(8.1)}$	$64.9_{(4.8)}$	$61.1_{(7.1)}$
	AUC	$81.4_{(2.4)}$	$82.7_{(2.0)}$	$81.8_{(2.0)}$		AUC	$83.3_{(1.3)}$	$82.6_{(1.3)}$	$82.5_{(1.0)}$
	F1	$43.1_{(3.7)}$	$43.2_{(2.2)}$	$43.3_{(3.9)}$		F1	$83.6_{(1.2)}$	$83.6_{(1.1)}$	$83.2_{(1.1)}$
	Acc	$47.4_{(3.0)}$	$48.3_{(2.9)}$	$48.5_{(2.7)}$		Acc	$84.7_{(1.5)}$	$83.4_{(1.7)}$	$84.6_{(1.9)}$
cptac-nsclc	BACC	$96.3_{(1.0)}$	$96.8_{(1.4)}$	$95.6_{(0.6)}$	mut-het-rcc-pbrm1	BACC	$66.1_{(1.9)}$	$65.0_{(2.4)}$	$64.6_{(1.6)}$
	AUC	$99.5_{(0.8)}$	$99.5_{(0.7)}$	$99.2_{(0.9)}$		AUC	$70.6_{(2.4)}$	$69.9_{(3.0)}$	$70.4_{(2.3)}$
	F1	$96.3_{(1.0)}$	$96.8_{(1.3)}$	$95.7_{(0.6)}$		F1	$66.1_{(2.0)}$	$65.0_{(2.4)}$	$64.6_{(1.6)}$
	Acc	$96.3_{(1.0)}$	$96.8_{(1.3)}$	$95.7_{(0.6)}$		Acc	$66.2_{(1.9)}$	$65.1_{(2.3)}$	$64.6_{(1.6)}$
cptac-pancancer	BACC	$92.6_{(0.3)}$	$91.1_{(1.2)}$	$92.2_{(0.9)}$	mut-het-rcc-setd2	BACC	$65.0_{(3.3)}$	$63.4_{(2.9)}$	$64.4_{(2.4)}$
	AUC	$99.7_{(0.0)}$	$99.6_{(0.1)}$	$99.7_{(0.0)}$		AUC	$71.6_{(1.8)}$	$71.0_{(1.0)}$	$71.8_{(0.9)}$
	F1	$93.7_{(0.5)}$	$92.3_{(1.2)}$	$93.4_{(0.9)}$		F1	$70.6_{(3.6)}$	$69.7_{(4.1)}$	$70.1_{(5.6)}$
	Acc	$93.7_{(0.5)}$	$92.3_{(1.3)}$	$93.5_{(0.9)}$		Acc	$70.5_{(4.4)}$	$69.8_{(5.1)}$	$70.2_{(6.6)}$
kidrare-coarse	BACC	$92.2_{(2.3)}$	$91.0_{(1.4)}$	$92.0_{(2.6)}$					
	AUC	$99.0_{(0.2)}$	$98.8_{(0.4)}$	$99.0_{(0.2)}$					
	F1	$93.1_{(1.9)}$	$92.8_{(1.1)}$	$93.1_{(1.9)}$					
	Acc	$93.1_{(1.9)}$	$92.8_{(1.1)}$	$93.1_{(1.8)}$					