ASMIL: ATTENTION-STABILIZED MULTIPLE INSTANCE LEARNING FOR WHOLE SLIDE IMAGING

Linfeng Ye[1], Shayan Mohajer Hamidi[2], Zhixiang Chi[1], Guang Li[3],
Mert Pilanci[2], Takahiro Ogawa[3], Miki Haseyama[3], Konstantinos N. Plataniotis[1]
[1]University of Toronto, [2]Stanford University, [3]Hokkaido University
[1]{linfeng.ye,zhixiang.chi}@mail.utoronto.ca
[1]kostas@ece.utoronto.ca
[2]{smohajer,pilanci}@stanford.edu
[3]{guang,ogawa,mhaseyama}@lmd.ist.hokudai.ac.jp

ABSTRACT

Attention-based multiple instance learning (MIL) has emerged as a powerful framework for whole slide image (WSI) diagnosis, leveraging attention to aggregate instance-level features into bag-level predictions. Despite this success, we find that such methods exhibit a new failure mode: unstable attention dynamics. Across four representative attention-based MIL methods and two public WSI datasets, we observe that attention distributions oscillate across epochs rather than converging to a consistent pattern, degrading performance. This instability adds to two previously reported challenges: overfitting and over-concentrated attention distribution. To simultaneously overcome these three limitations, we introduce attention-stabilized multiple instance learning (ASMIL), a novel unified framework. ASMIL uses an anchor model to stabilize attention, replaces softmax with a normalized sigmoid function in the anchor to prevent over-concentration, and applies token random dropping to mitigate overfitting. Extensive experiments demonstrate that ASMIL achieves up to a 6.49% F1 score improvement over state-of-the-art methods. Moreover, integrating the anchor model and normalized sigmoid into existing attention-based MIL methods consistently boosts their performance, with F1 score gains up to 10.73%. All code and data are publicly available at https://anonymous.4open.science/r/ASMIL-5018/.

1 INTRODUCTION

Computational pathology, at the intersection of digital imaging, machine learning, and clinical diagnostics, has transformed modern workflows (Verghese et al., 2023). Advances in whole slide imaging (WSI) now allow glass slides to be digitized into gigapixel images (Bacus, 2001), which are central to cancer diagnosis and treatment planning. WSIs preserve rich spatial context and enable large-scale sharing, but their extreme size and sparsity create major challenges: diagnostically relevant regions often occupy only a tiny fraction of the slide, and exhaustive pixel- or tile-level annotations are infeasible in practice. As a result, most datasets provide only weak slide-level labels, making it critical to design methods that learn effectively under weak supervision.

This weakly supervised setting naturally motivates multiple instance learning (MIL) (Keeler et al., 1990; Dietterich et al., 1997; Maron & Lozano-Pérez, 1998). In MIL, a bag of instances is mapped to a single bag-level label. For WSIs, the image is divided into tiles, each treated as an instance, while only the slide-level label is required. This dramatically reduces annotation costs and makes large-scale WSI datasets more practical for research and clinical use.

Early approaches to MIL-based WSI analysis focused on simple aggregation strategies, such as clustering instance features (Xu et al., 2014) or applying global pooling layers (Kraus et al., 2016). A major breakthrough came with the introduction of attention-based MIL (ABMIL) (Ilse et al., 2018), which provided theoretical guidance for neural network-based MIL algorithms and introduced a permutation-invariant attention mechanism to aggregate instance information into bag-level representations. ABMIL established a strong baseline for WSI analysis (Shao et al., 2025) and, impor-

tantly, enhanced interpretability through visualized attention scores, which is an essential property for clinical adoption. Building on this foundation, subsequent works have refined ABMIL to further improve performance, scalability, and robustness (Xiong et al., 2021; Shao et al., 2021; Zhang et al., 2022; Tang et al., 2023b; Zhang et al., 2024). In particular, TransMIL replaces independent instance weighting with a transformer encoder that explicitly models inter-instance relations within a bag (Shao et al., 2021). As a result, attention-based MIL has become the de facto choice for WSI subtyping not only because it aggregates instance features but also because its attention maps are used as clinical evidence of model interpretability.

Despite its success, attention-based MIL still suffers from three major problems, which we denote as **(PI)**, **(PII)**, and **(PIII)**, and elaborate on in the sequel.

A critical yet underexplored aspect of MIL-based WSI analysis is the convergence behavior of attention mechanisms during training. The gigapixel scale of WSIs, coupled with weak supervision, high variability, and sparsity, makes it difficult for models to consistently identify informative tiles among thousands of candidates. Our investigation reveals that existing MIL algorithms often fail to converge stably on WSI datasets. To the best of our knowledge, we are the first to identify and systematically analyze **(PI)** *unstable attention dynamics*, where attention distributions for individual WSIs oscillate substantially across epochs instead of converging into consistent patterns. To quantify this phenomenon, we measure the Jensen-Shannon divergence (Cover, 1999) between consecutive attention distributions of the same WSI, as illustrated for TransMIL (Shao et al., 2021) in Figure 1. Additional experiments across methods and datasets are provided in Appendix P. This persistent oscillation results in unstable training and degraded performance, reflected in higher cross-entropy values compared to our proposed method.

Beyond this new limitation identified in our study, prior work has highlighted two additional challenges. One is **(PII)** *over-concentrated attention distribution* (Zhang et al., 2024; Lu et al., 2021), where models allocate excessive importance to only a few tiles, thereby harming generalization and interpretability. The other is **(PIII)** *overfitting* (Zhang et al., 2022; Lin et al., 2023), a common issue in histopathology WSI classification caused by the limited number of available training samples.

In this paper, we aim to simultaneously address the challenges **(PI)**–**(PIII)**. To stabilize attention distribution and the training process, we introduce an *anchor model*, which has the same architecture as the online model's attention module and receives the same input, but is updated via an exponential moving average (EMA) instead of by backpropagation. Acting as a stable reference, the anchor provides smoother and more consistent attention distributions. To transfer this stability, we encourage the online model to mimic the anchor by minimizing the Kullback–Leibler (KL) divergence between their attention distributions. To mitigate over-concentration, which we attribute to the exponential sensitivity of the softmax function, we replace softmax in the anchor branch with a normalized sigmoid function (NSF), as defined in Equation (5). Finally, we propose a simple yet effective token dropout strategy that regularizes the model and reduces overfitting. Together with the anchor model, these components form a unified framework called attention-stabilized multiple instance learning (ASMIL), which improves both the stability and generalization of MIL-based WSI analysis.

In summary, this paper's contributions are as follows:

• We are the first to identify and systematically analyze the problem of *unstable attention dynamics* in attention-based MIL for WSI analysis. This overlooked issue not only limits predictive performance but also undermines interpretability, since fluctuating attention distributions prevent consistent identification of the tissue regions that drive the model's decisions.

• To overcome this instability, we introduce an anchor model that stabilizes attention distribution throughout training. The anchor model is updated using an exponential moving average of the online model, which ensures stable training dynamics and improves both performance and interpretability.

• We show *mathematically* that replacing softmax with an NSF alleviates attention over-concentration. Since applying the NSF to the online model causes vanishing gradients, we apply it to the anchor model instead, ensuring stable and well-distributed attention.

• To mitigate overfitting, we introduce token dropout, which randomly discards a portion of feature tokens during training while retaining all tokens during inference.

• By integrating these innovations, we present attention-stabilized MIL (ASMIL), a novel MIL-based WSI analysis algorithm. Through comprehensive experiments on multiple public WSI datasets, we demonstrate that ASMIL achieves state-of-the-art performance in subtyping and localization tasks.

Paper Organization. The remainder of this paper is structured as follows: Section 2 reviews related work on MIL and attention mechanisms in WSI analysis; Section 3 presents the preliminaries and motivation of our approach; Section 4 details the ASMIL framework; Section 5 presents the experimental setup and results; and finally Section 6 concludes the paper with future research directions.

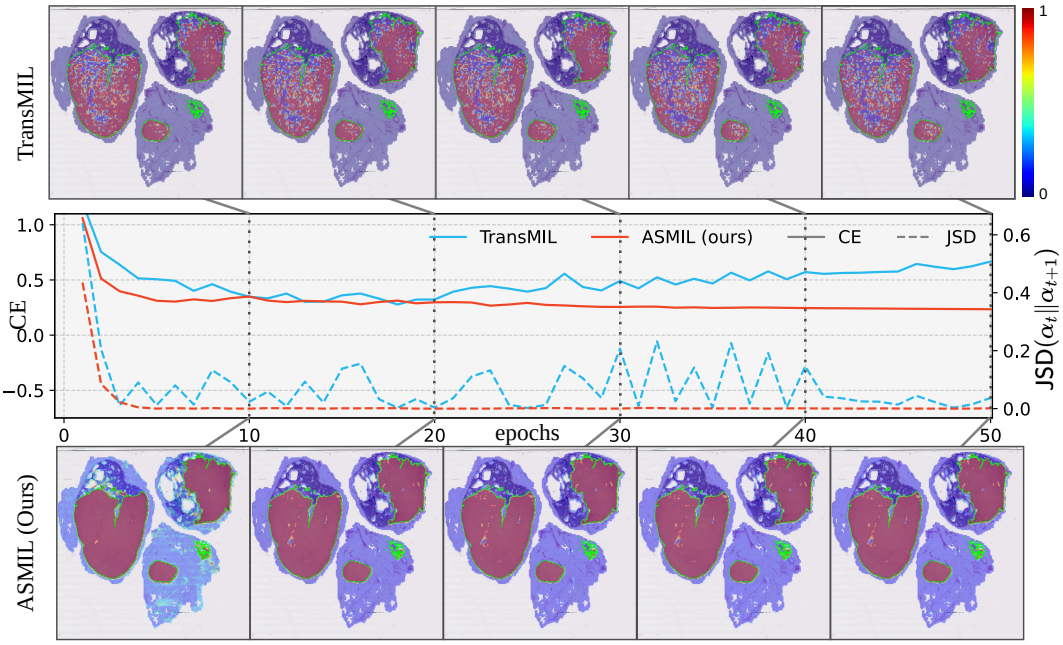

Figure 1: Visualization of attention dynamics on a tumor WSI for TransMIL (Shao et al., 2021) vs. ASMIL (our method). The green contours in the figures indicate the annotated tumor regions. **Top**: TransMIL attention distribution at selected training iterations. **Middle**: Jensen-Shannon divergence (JSD) between attention distributions at successive steps and the cross entropy loss (CE), comparing TransMIL (blue) and ASMIL (red). **Bottom**: Attention distribution from ASMIL over different training iterations. Due to the weakly supervised nature of WSI subtyping datasets, TransMIL's attention patterns never converge during training, further, it focuses on only a subset of cancerous regions. In contrast, our method (i) produces stable attention distributions throughout training and (ii) consistently highlights cancerous regions.

2 Related Work

Early weakly supervised approaches in computational pathology leveraged multi-view convolutional neural network ensembles and basic MIL pooling to transition from patch-level labels to slide-level predictions (Das et al., 2017; 2018). As datasets scaled and slide-level supervision became the norm, methods shifted from fixed pooling to attention mechanisms that make aggregation learnable. Building on this trend, attention-based MIL (Ilse et al., 2018) introduced learnable instance weights and generated heatmaps from slide-level labels, achieving breast and colon cancer classification on par with fully supervised methods at scale. Complementary to weighting instances, subsequent work reduced morphological redundancy in tile representations, Song et al. (2024) used a Gaussian mixture model, and sped up inference by skipping irrelevant patches (Dong et al., 2025). Li et al. (2021b) propose DSMIL, a dual-stream MIL framework that selects a critical instance via max-pooling and then applies a trainable non-local, distance-based attention from this instance to all others to form bag embeddings for WSI classification. Subsequent works extend this line of research by leveraging multi-scale fusion to aggregate information across resolutions (Zhang et al., 2021; Guo et al., 2023; Tran et al., 2025; Buzzard et al., 2024; Li et al., 2019).

Several works further refine training strategies for attention-based MIL. To prevent the attention distribution from collapsing onto a few input patches and to obtain more faithful attention maps, Zhang et al. (2024) stochastically masks the top-K instances, while Zhang et al. (2025b) adds an entropy regularization term that explicitly flattens the attention distribution. In a complementary direction, Fourkioti et al. (2024) introduces neighbor-constrained attention to suppress noise in the feature maps. Because WSI datasets usually contain only a few hundred training samples, many methods focus on mitigating overfitting, for example, by introducing bag splitting to create pseudo-bags (Zhang et al., 2022), designing efficient instance-based classifiers (Qu et al., 2024), and performing hard-negative mining with EMA teachers (Tang et al., 2023b). Lu et al. (2021) introduce clustering-constrained attention multiple-instance learning (CLAM), which replaces max-pooling with class-specific attention pooling and adds instance-level clustering supervision so that weakly supervised slide-level MIL can be both data-efficient and interpretable on WSIs, or using contrastive critical-instance branches (Li et al., 2021a). Recently, Zhu et al. (2025; 2023) systematically studied the effect of random dropping in MIL and proposed to randomly remove the top-K instances with the highest attention weights together with $G \times k$ similar tokens during training, which mitigates overfitting and encourages convergence to flatter regions of the loss landscape, thereby improving generalization. Since our anchor leverages an EMA update, we relate it to EMA/teacher models and provide additional details in Appendix A.

3 Preliminaries and Motivation

3.1 Notation

Scalars are denoted by non-bold letters (e.g., a, β), vectors by bold lowercase letters (e.g., $\boldsymbol{a}$), and matrices by bold uppercase letters (e.g., $\boldsymbol{A}$). The i-th entry of a vector $\boldsymbol{a}$ is written as $\boldsymbol{a}_i$. A C-dimensional probability simplex is denoted by Δ^C. For two distributions $P_1, P_2 \in \Delta^C$, the Kullback–Leibler divergence (KL divergence) is defined as $\mathsf{KL}(P_1 \| P_2) = \sum_{c=1}^{C} P_1[c] \log \frac{P_1[c]}{P_2[c]}$.

3.2 Multiple Instance Learning with Attention

In MIL, supervision is provided only at the bag level. A slide is represented as a bag $X = \{\boldsymbol{x}_i\}_{i=1}^N$ with unknown instance labels. After a pretrained encoder, we obtain instance embeddings $\{\boldsymbol{h}_i\}_{i=1}^N$.

Attention-based MIL assigns a scalar *attention score* to each embedding via a learnable scorer $f_{\boldsymbol{\theta}}$:

$$z_i \;=\; f_{\boldsymbol{\theta}}(\boldsymbol{h}_i), \qquad \boldsymbol{z} = (z_1, \ldots, z_N) \in \mathbb{R}^N. \tag{1}$$

Scores are normalized into an *attention distribution* on the probability simplex Δ^N using a softmax:

$$\alpha_i \;=\; \frac{\exp(z_i)}{\sum_{j=1}^{N} \exp(z_j)}, \qquad \sum_{i=1}^{N} \alpha_i = 1, \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N) \in \Delta^N. \tag{2}$$

The slide-level representation, $\boldsymbol{h}_{\text{bag}} \;=\; \sum_{i=1}^{N} \alpha_i \, \boldsymbol{h}_i$, is a convex combination of instance features weighted by the attention distribution and is passed to a classifier to produce the bag-level prediction.

3.3 Motivation

MIL is effective for WSI analysis, but its weak supervision and small WSI dataset sizes introduce three failure modes: unstable attention dynamics, over-concentrated attention, and overfitting.

• **(PI)** ***Unstable attention dynamics.*** Under bag-level supervision, we empirically observe that attention distribution oscillates across epochs rather than converging to a consistent pattern. To the best of our knowledge, this phenomenon has not been previously identified or explicitly addressed in the literature. To quantify stability, we measure the Jensen-Shannon divergence (JSD) between consecutive attention distributions for the same WSI. Let $\boldsymbol{\alpha}_t \in \Delta^N$ denote the attention over N tiles at epoch t. With $\mathsf{KL}(\cdot\|\cdot)$ denoting the KL divergence and $\bar{\boldsymbol{\alpha}} = \frac{1}{2}(\boldsymbol{\alpha}_t + \boldsymbol{\alpha}_{t+1})$, we define

$$\mathsf{JSD}(\boldsymbol{\alpha}_t \| \boldsymbol{\alpha}_{t+1}) = \tfrac{1}{2}\mathsf{KL}(\boldsymbol{\alpha}_t \| \bar{\boldsymbol{\alpha}}) + \tfrac{1}{2}\mathsf{KL}(\boldsymbol{\alpha}_{t+1} \| \bar{\boldsymbol{\alpha}}). \tag{3}$$

As shown in Figure 1, TransMIL (Shao et al., 2021) exhibits large JSD fluctuations, indicating a lack of stable convergence. Similar behavior appears in other attention-based MIL models; additional results are provided in Appendix P.

• **(PII)** ***Over-concentration of attention.*** Complementary to instability, prior works report that ABMIL often assigns most mass to a few tiles, which harms generalization and interpretability (Zhang et al., 2024; 2025b). Distinct from previous approaches, we attribute these over-concentrated attention distributions to the exponential nature of the softmax function.

• **(PIII)** ***Overfitting.*** WSI datasets typically contain only a few slides per class and highly redundant tiles (Zhang et al., 2022). High-capacity neural-network-based MIL models can memorize spurious tile-level patterns, leading to poor out-of-distribution performance. To alleviate this, we introduce a random token drop mechanism specialized for our method.

In the next section, we present our proposed methodology, which simultaneously addresses the three problems **(PI)**, **(PII)**, and **(PIII)**.

4 METHODOLOGY

To address the limitations of attention-based MIL, we propose a framework illustrated in Figure 2. Our methodology addresses **(PI)** by stabilizing attention through an anchor model, tackles **(PII)** by replacing softmax with an NSF in the anchor, and mitigates **(PIII)** by token random dropping to regularize training. The next subsections detail each component and the overall objective.

4.1 STABILIZING ATTENTION DISTRIBUTIONS VIA AN ANCHOR MODEL

As discussed in Section 3.3, weak supervision in MIL often leads to unstable attention distributions that fluctuate across epochs, preventing convergence. To mitigate this, we introduce an *anchor model* that mirrors the attention block of the online model. The anchor serves as a stable reference by being updated through an EMA of the online model's parameters. Specifically, at training step t, the anchor parameters $\boldsymbol{\theta}'_t$ are updated as

$$\boldsymbol{\theta}'_t \leftarrow m\boldsymbol{\theta}'_{t-1} + (1-m)\boldsymbol{\theta}_t, \quad (4)$$

where $\boldsymbol{\theta}_t$ are the online model's parameters and $m \in [0, 1)$ is the EMA factor. Both the anchor and online models receive the same inputs, but **only** the online model is updated by backpropagation, the anchor is updated via EMA. The goal is to align the online attention distribution to the anchor distribution, which yields a stabilization loss.

In Appendix C, we show that standard attention-based MIL yields poorly separated bag-level feature clusters during training because attention distributions do not converge reliably. Introducing the an-

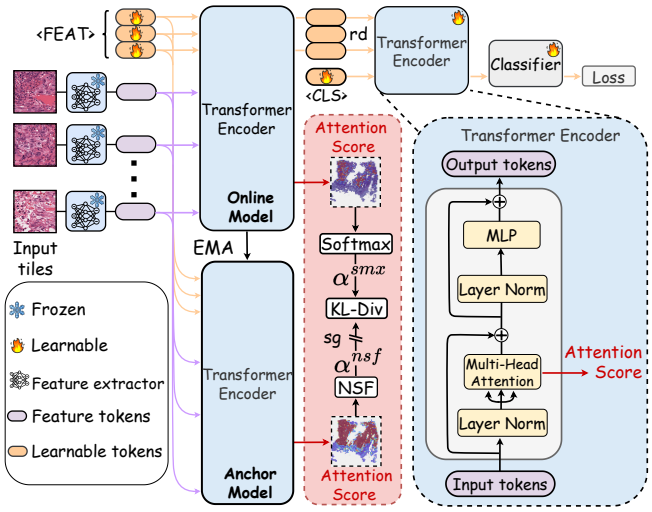

Figure 2: Overview of ASMIL. Each WSI is divided into tiles and embedded into vision tokens using a pretrained encoder. These tokens, along with trainable FEAT tokens, feed into both online and anchor encoders. The anchor encoder's attention scores over the FEAT tokens are transformed into a probability vector using an NSF, while the online encoder applies a softmax. To stabilize training and prevent the online model's attention from becoming overly concentrated, we compute the KL divergence between the two distributions. Gradients are blocked to the anchor encoder using a stop-gradient (sg) operator, and its parameters are updated via EMA from the online encoder. During training, we randomly drop (rd) N FEAT tokens, feed the remaining tokens into a second transformer with a trainable [CLS] token, and train a classifier on its output. 🔥 and ❄ indicate learnable and frozen components, respectively.

chor model stabilizes attention, improves convergence, and produces clearly separated bag-level clusters.

Remark 1. ***Why an anchor model instead of a single regularizer.*** *Scalar penalties on attention, such as entropy, ℓ_2, or temperature, are content-agnostic and act only on the current batch. They cannot encode relational structure among instances. An EMA* anchor model *yields a data-dependent attention distribution conditioned on the bag. Encouraging the online attention to stay close to this target performs functional regularization that captures inter-instance relations and stabilizes training, which a scalar regularizer cannot do.*

The anchor is discarded at inference, adding no extra FLOPs or latency. In the next subsection, we describe how we further improve the anchor's attention using an NSF, which alleviates over-concentration before applying this stabilization loss.

4.2 Preventing Attention Concentration with NSF in the Anchor Model

In conventional transformer architectures, the softmax function maps self-attention scores $\boldsymbol{z} \in \mathbb{R}^N$ to a probability vector. However, softmax often produces over-concentrated attention, in which a few tokens dominate while the weights of the remaining tokens vanish. Temperature scaling is an incomplete remedy: small temperatures preserve concentration, while large temperatures flatten the distribution so aggressively that weak tokens receive undue weight. We therefore seek a mechanism that equalizes attention among genuinely informative tokens while suppressing weak ones.

We compare softmax with normalized sigmoid function (NSF).[1] For $\boldsymbol{z} = (z_1, \ldots, z_N)$, define

$$\alpha_i^{\mathrm{smx}}(\boldsymbol{z}; T) \;=\; \frac{e^{z_i/T}}{\sum_{j=1}^{N} e^{z_j/T}}, \qquad \alpha_i^{\mathrm{nsf}}(\boldsymbol{z}) \;=\; \frac{\sigma(z_i)}{\sum_{j=1}^{N} \sigma(z_j)}, \qquad \sigma(t) = \frac{1}{1+e^{-t}}. \tag{5}$$

For thresholds $\tau > 0$ and bandwidth $\gamma \geq 0$, let $\mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$ be the set of score vectors with "high" indices $\mathcal{H}$ satisfying $z_i \in [\tau, \tau + \gamma]$ for $i \in \mathcal{H}$ and "low" indices $\mathcal{L}$ satisfying $z_j \leq -\tau$ for $j \in \mathcal{L}$. Denote $h \triangleq |\mathcal{H}|$ and $\ell \triangleq |\mathcal{L}|$. The following theorem (proof deferred to Appendix E) formalizes the selective flattening property of NSF and shows that softmax cannot match it with a single temperature.

Theorem 1 (NSF achieves selective flattening; softmax cannot with a single T)**.** *Fix $\tau > 0$, $\gamma \geq 0$, and index sets $\mathcal{H}, \mathcal{L}$ with $h \geq 1$, $\ell \geq 1$. For any $\boldsymbol{z} \in \mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$:*

(A) NSF bounds. *For any $i, h' \in \mathcal{H}$ and any $j \in \mathcal{L}$,*

$$\frac{\alpha_i^{\mathrm{nsf}}(\boldsymbol{z})}{\alpha_{h'}^{\mathrm{nsf}}(\boldsymbol{z})} = \frac{\sigma(z_i)}{\sigma(z_{h'})} \;\leq\; \frac{\sigma(\tau+\gamma)}{\sigma(\tau)} = \frac{1+e^{-\tau}}{1+e^{-(\tau+\gamma)}} \;\leq\; 1+e^{-\tau}, \quad \alpha_j^{\mathrm{nsf}}(\boldsymbol{z}) \leq \frac{\sigma(-\tau)}{h\,\sigma(\tau)} \;=\; \frac{e^{-\tau}}{h}. \tag{6}$$

Hence, NSF equalizes the high tokens up to a factor $1+e^{-\tau}$ and suppresses lows to at most $e^{-\tau}/h$. As $\tau \to \infty$ with fixed γ, ratios among high tokens approach 1 and low-token weights vanish.

(B) Softmax incompatibility with one temperature. *Suppose we desire suppression and equalization targets (ε, κ) on $\mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$:*

$$\textit{(Suppression)} \quad \alpha_j^{\mathrm{smx}}(\boldsymbol{z}; T) \leq \varepsilon \;\; \forall j \in \mathcal{L}, \qquad \textit{(Equalization)} \quad \frac{\max_{i\in\mathcal{H}} \alpha_i^{\mathrm{smx}}(\boldsymbol{z}; T)}{\min_{h'\in\mathcal{H}} \alpha_{h'}^{\mathrm{smx}}(\boldsymbol{z}; T)} \leq \kappa.$$

Then T must satisfy $T \leq \dfrac{2\tau}{\log\left(\frac{h}{\varepsilon}\right)}$ and $T \geq \dfrac{\gamma}{\log \kappa}$ simultaneously, which is impossible whenever $\dfrac{\gamma}{\log \kappa} > \dfrac{2\tau}{\log\left(\frac{h}{\varepsilon}\right)}$. Thus, no single temperature achieves both targets for all $\boldsymbol{z} \in \mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$.

We further illustrate this effect in Figure 3 by comparing attention maps with softmax and NSF using ABMIL (Ilse et al., 2018) on a cancer slide from the CAMELYON-16 dataset (Ehteshami Bejnordi et al., 2017). Softmax yields a highly concentrated map that obscures broader context, whereas NSF produces a less concentrated attention map that highlights most cancerous regions.

[1] We discuss alternatives to NSF, including entmax and softmax with temperature scaling in Appendix F.

A naive option is to apply NSF directly in the online model. In practice, this induces vanishing gradients and degrades performance; see Appendix G. We therefore place NSF in the *anchor* model as a stable prior, guiding the online model without hindering its learning dynamics. As attention distributions lie on the probability simplex, we use the KL divergence to align the online attention distribution with the NSF-based anchor distribution:

$$\mathcal{L}_{\mathrm{AS}} = \mathsf{KL}\big(\boldsymbol{\alpha}^{\mathrm{nsf}} \,\|\, \boldsymbol{\alpha}\big), \quad (7)$$

where $\boldsymbol{\alpha}$ is the online attention (softmax over $\boldsymbol{z}$) and $\boldsymbol{\alpha}^{\mathrm{nsf}}$ is the anchor attention (NSF over the anchor scores). Using $\frac{\partial \alpha_j}{\partial z_i} = \alpha_j(\delta_{ij} - \alpha_i)$ and treating $\boldsymbol{\alpha}^{\mathrm{nsf}}$ as fixed, the gradient with respect to the online attention score z_i is

$$\begin{aligned} \frac{\partial \mathsf{KL}(\boldsymbol{\alpha}^{\mathrm{nsf}} \,\|\, \boldsymbol{\alpha})}{\partial z_i} &= \sum_{j=1}^{N} \alpha_j^{\mathrm{nsf}}(\delta_{ij} - \alpha_i) \\ &= \alpha_i - \alpha_i^{\mathrm{nsf}}. \end{aligned} \quad (8)$$

Thus, gradient descent moves the online attention toward the anchor distribution, promoting stability and discouraging over-concentration.

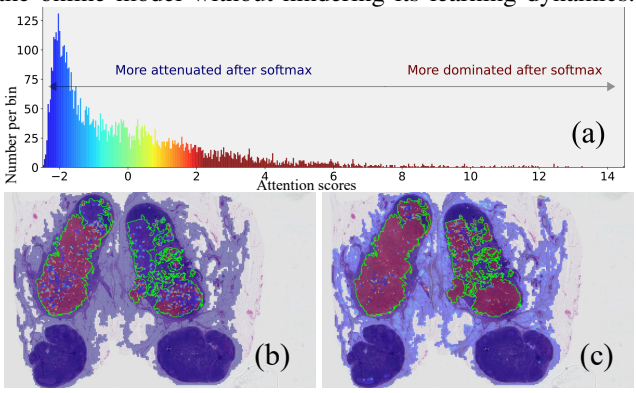

Figure 3: (a) Distribution of attention scores in ABMIL, which exhibits a long-tailed pattern. (b) Attention distribution obtained with the softmax function and (c) with the NSF. Unlike softmax, the normalized sigmoid suppresses large values in the long tail, yielding a less sparse and more interpretable attention distribution.

Remark 2. *The anchor in ASMIL superficially resembles the teacher in MHIM-MIL (Tang et al., 2023b): both are EMA-updated copies of the online model. Their roles, however, differ in two important ways.* ***(i)*** *MHIM-MIL uses the teacher to mine hard instances, whereas ASMIL uses the anchor to stabilize attention and prevent over-concentration.* ***(ii)*** *MHIM-MIL matches softmax bag-level features, while ASMIL directly matches attention distributions. Appendix I discusses why softmax bag-level matching fails to stabilize attention maps.*

4.3 Mitigating Overfitting with Token Random Dropping

To reduce overfitting, we designed a token-level regularizer, specialized for ASMIL, that operates on the trainable tokens used by the online model. Let a WSI $\boldsymbol{x}$ be partitioned into M tiles and embedded by a pretrained encoder into tile tokens $\mathcal{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_M\}$. We augment these with N trainable FEAT tokens $\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\}$ and feed the concatenation $[\mathcal{T}; \mathcal{P}]$ into the online encoder. After the online encoder, only the FEAT tokens are retained. Since the number of FEAT tokens is much smaller than the tile tokens (*i.e.*, $N \ll M$), this design acts as information aggregation via token reduction.

During training, we sample an independent Bernoulli mask over FEAT tokens and drop a fraction $B \in [0, 1)$ of them. Denote the kept set by $\mathcal{P}_{\mathrm{keep}}$ with $|\mathcal{P}_{\mathrm{keep}}| = \tilde{N} \sim \mathrm{Binomial}(N, 1 - B)$ and $\mathbb{E}[\tilde{N}] = (1 - B)N$. The remaining tokens, together with a trainable [CLS] token, are passed to a second transformer to produce a bag representation $\boldsymbol{h}_{\mathrm{bag}}$, which is then classified to obtain $\hat{y}$. At inference time, no tokens are dropped ($B = 0$). Since ASMIL stabilize the attention via aligning the anchor model, which assumes a one-to-one correspondence, as thus general instance dropout method, such as MIL-Dropout Zhu et al. (2025), could not be integrated easily.

This stochastic removal prevents co-adaptation among FEAT tokens and discourages the model from over-relying on a subset of tokens, while preserving image content by keeping all FEAT tokens at inference. Empirically, this acts as an effective regularizer that improves generalization. In Appendix K.4 we study the effect of B and observe a consistent peak in performance around $B \approx 0.5$.

4.4 Overall Training Objective

Based on the discussion thus far, we train with a joint objective that couples standard bag-level classification with attention stabilization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \beta\, \mathcal{L}_{\mathrm{AS}}, \quad (9)$$

Table 1: The F1 score and AUC of different MIL approaches across three WSI datasets. **Bold** and underlined values denote the best and second-best results, respectively.

Backbone	Dataset / Method	CAMELYON-16 F1 score ↑	CAMELYON-16 AUC ↑	CAMELYON-17 F1 score ↑	CAMELYON-17 AUC ↑	BRACS F1 score ↑	BRACS AUC ↑
ResNet-18 ImageNet Pretrained	ABMIL ICML 2018	$0.757_{\pm 0.020}$	$0.790_{\pm 0.027}$	$0.508_{\pm 0.032}$	$0.779_{\pm 0.021}$	$0.523_{\pm 0.028}$	$0.723_{\pm 0.035}$
	Clam-SB Nature 2021	$0.742_{\pm 0.024}$	$0.763_{\pm 0.049}$	$0.504_{\pm 0.012}$	$0.778_{\pm 0.024}$	$0.521_{\pm 0.046}$	$0.750_{\pm 0.039}$
	TransMIL NeurIPS 2021	$0.643_{\pm 0.088}$	$0.706_{\pm 0.076}$	$0.499_{\pm 0.082}$	$0.794_{\pm 0.053}$	$0.444_{\pm 0.040}$	$0.732_{\pm 0.043}$
	DSMIL CVPR 2021b	$0.736_{\pm 0.025}$	$0.773_{\pm 0.034}$	$0.473_{\pm 0.052}$	$0.705_{\pm 0.022}$	$0.511_{\pm 0.052}$	$0.751_{\pm 0.028}$
	DTFD-MIL CVPR 2022	$0.758_{\pm 0.051}$	$0.815_{\pm 0.063}$	$0.546_{\pm 0.010}$	$0.735_{\pm 0.011}$	$0.469_{\pm 0.016}$	$0.717_{\pm 0.032}$
	IBMIL CVPR 2023	$0.777_{\pm 0.009}$	$0.799_{\pm 0.050}$	$0.533_{\pm 0.015}$	$0.813_{\pm 0.092}$	$0.510_{\pm 0.043}$	$0.726_{\pm 0.034}$
	MHIM-MIL ICCV 2023b	$0.752_{\pm 0.034}$	$0.772_{\pm 0.026}$	$0.56_{\pm 0.029}$	$0.815_{\pm 0.019}$	$0.511_{\pm 0.022}$	$0.775_{\pm 0.021}$
	ACMIL ECCV 2024	$0.798_{\pm 0.029}$	$0.841_{\pm 0.030}$	$0.528_{\pm 0.053}$	$0.789_{\pm 0.046}$	$0.552_{\pm 0.048}$	$0.754_{\pm 0.008}$
	CAMIL ICLR 2024	$0.778_{\pm 0.011}$	$0.812_{\pm 0.017}$	$0.503_{\pm 0.007}$	$0.806_{\pm 0.006}$	$0.569_{\pm 0.007}$	$\underline{0.787}_{\pm 0.011}$
	AEM MICCAI 2025b	$\underline{0.804}_{\pm 0.022}$	$\underline{0.859}_{\pm 0.031}$	$0.525_{\pm 0.043}$	$0.828_{\pm 0.054}$	$0.554_{\pm 0.004}$	$0.764_{\pm 0.008}$
	HDMIL CVPR 2025	$0.790_{\pm 0.023}$	$0.856_{\pm 0.027}$	$\underline{0.557}_{\pm 0.007}$	$\mathbf{0.853}_{\pm 0.013}$	$\underline{0.578}_{\pm 0.012}$	$0.761_{\pm 0.011}$
	ASMIL (Ours)	$\mathbf{0.814}_{\pm 0.052}$	$\mathbf{0.870}_{\pm 0.064}$	$\mathbf{0.564}_{\pm 0.020}$	$\underline{0.851}_{\pm 0.061}$	$\mathbf{0.601}_{\pm 0.072}$	$\mathbf{0.810}_{\pm 0.054}$
VIT-S SSL pretrained	ABMIL ICML 2018	$0.914_{\pm 0.031}$	$0.945_{\pm 0.027}$	$0.522_{\pm 0.050}$	$0.853_{\pm 0.016}$	$0.680_{\pm 0.051}$	$0.866_{\pm 0.029}$
	Clam-SB Nature 2021	$0.925_{\pm 0.085}$	$0.969_{\pm 0.024}$	$0.523_{\pm 0.020}$	$0.846_{\pm 0.020}$	$0.631_{\pm 0.034}$	$0.863_{\pm 0.005}$
	TransMIL NeurIPS 2021	$0.922_{\pm 0.019}$	$0.943_{\pm 0.009}$	$0.554_{\pm 0.048}$	$0.792_{\pm 0.029}$	$0.631_{\pm 0.030}$	$0.841_{\pm 0.006}$
	DSMIL CVPR 2021b	$0.943_{\pm 0.007}$	$0.966_{\pm 0.009}$	$0.532_{\pm 0.064}$	$0.804_{\pm 0.032}$	$0.577_{\pm 0.028}$	$0.816_{\pm 0.028}$
	DTFD-MIL CVPR 2022	$0.948_{\pm 0.007}$	$\underline{0.980}_{\pm 0.011}$	$0.627_{\pm 0.015}$	$0.866_{\pm 0.012}$	$0.612_{\pm 0.080}$	$0.870_{\pm 0.022}$
	IBMIL CVPR 2023	$0.912_{\pm 0.034}$	$0.954_{\pm 0.022}$	$0.557_{\pm 0.064}$	$0.850_{\pm 0.024}$	$0.645_{\pm 0.041}$	$0.871_{\pm 0.014}$
	MHIM-MIL ICCV 2023b	$0.932_{\pm 0.024}$	$0.970_{\pm 0.037}$	$0.541_{\pm 0.022}$	$0.845_{\pm 0.026}$	$0.625_{\pm 0.060}$	$0.865_{\pm 0.017}$
	ACMIL ECCV 2024	$0.954_{\pm 0.012}$	$0.974_{\pm 0.012}$	$0.562_{\pm 0.050}$	$0.863_{\pm 0.004}$	$0.722_{\pm 0.030}$	$0.888_{\pm 0.010}$
	CAMIL ICLR 2024	$0.930_{\pm 0.009}$	$0.963_{\pm 0.011}$	$0.633_{\pm 0.022}$	$0.886_{\pm 0.034}$	$0.709_{\pm 0.011}$	$0.836_{\pm 0.014}$
	AEM MICCAI 2025b	$0.947_{\pm 0.003}$	$0.974_{\pm 0.007}$	$\underline{0.647}_{\pm 0.007}$	$\underline{0.887}_{\pm 0.013}$	$\underline{0.742}_{\pm 0.030}$	$\underline{0.905}_{\pm 0.010}$
	HDMIL CVPR 2025	$\underline{0.958}_{\pm 0.013}$	$0.976_{\pm 0.017}$	$0.571_{\pm 0.012}$	$0.796_{\pm 0.022}$	$0.717_{\pm 0.033}$	$0.874_{\pm 0.010}$
	ASMIL (Ours)	$\mathbf{0.965}_{\pm 0.020}$	$\mathbf{0.985}_{\pm 0.017}$	$\mathbf{0.689}_{\pm 0.005}$	$\mathbf{0.898}_{\pm 0.010}$	$\mathbf{0.781}_{\pm 0.042}$	$\mathbf{0.914}_{\pm 0.014}$

where the coefficient $\beta > 0$ balances the stabilization and classification objectives. In practice, to calculate $\mathcal{L}_{\mathrm{AS}}$, $\boldsymbol{\alpha}$ is computed by a softmax over the online scores, $\boldsymbol{\alpha}^{\mathrm{nsf}}$ is computed by applying the NSF to the anchor scores, and the anchor model is treated as *stop-gradient* while its parameters are updated via EMA. The KL divergence is taken over the attention distributions on the FEAT token set used for aggregation. This objective discourages attention concentration through $\mathcal{L}_{\mathrm{AS}}$ and preserves task performance through $\mathcal{L}_{\mathrm{CE}}$. ASMIL can be easily applied to other tasks, including survival prediction by replacing the objective function and the classification head accordingly. During training, the online model is updated by gradient descent

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}, \tag{10}$$

where η is the learning rate. $\mathcal{L}$ is computed as in Equation (9). The anchor model is then updated according to Equation (4). The gradient is only used to update the online model, while the anchor model influences learning through Equation (7). At inference time, ASMIL uses only the online model and discards the anchor model; therefore, the anchor does not increase the computational budget at inference.

5 Experiments

To demonstrate the effectiveness of ASMIL, we evaluate it on three well-known public WSI subtyping datasets: (i) CAMELYON-16 (Ehteshami Bejnordi et al., 2017), (ii) CAMELYON-17 (Bándi et al., 2019), and (iii) BRACS (Brancati et al., 2022). Details of the data splits, preprocessing, training setup, and baselines are provided in the Appendix B. We further evaluate ASMIL on survival prediction and non-WSI datasets in Appendix N and Appendix O, respectively.

5.1 Subtyping Performance

We compare ASMIL against eleven attention-based MIL baselines that are designed for WSIs: CLAM-SB (Lu et al., 2021), TransMIL (Shao et al., 2021), DSMIL (Li et al., 2021b), DTFD-MIL (Zhang et al., 2022), IBMIL (Lin et al., 2023), MHIM-MIL (Tang et al., 2023b), ABMIL (Ilse et al., 2018), ACMIL (Zhang et al., 2024), CAMIL (Fourkioti et al., 2024), AEM (Zhang et al., 2025b) and HDMIL (Dong et al., 2025). Because WSI datasets are class-imbalanced, we report the F1 score and area under the ROC curve (AUC) for each dataset in Table 1[2].

[2]See Appendix D for details on metric computation and interpretation.

Table 2: Applying anchor model and NSF to other attention-based MIL methods.

Dataset			CAMELYON-16		CAMELYON-17		BRACS	
Method	Anchor	NSF	F1 score ↑	AUC ↑	F1 score ↑	AUC ↑	F1 score ↑	AUC ↑
ABMIL ICML 2018	✗	✗	$0.914_{\pm 0.031}$	$0.945_{\pm 0.027}$	$0.522_{\pm 0.050}$	$0.853_{\pm 0.016}$	$0.680_{\pm 0.051}$	$0.866_{\pm 0.029}$
	✓	✗	$0.951_{\pm 0.015}$ +0.037	$0.963_{\pm 0.008}$ +0.018	$0.573_{\pm 0.011}$ +0.051	$0.871_{\pm 0.010}$ +0.018	$0.751_{\pm 0.013}$ +0.071	$0.877_{\pm 0.007}$ +0.011
	✓	✓	$0.953_{\pm 0.009}$ +0.039	$0.967_{\pm 0.006}$ +0.022	$0.574_{\pm 0.010}$ +0.052	$0.883_{\pm 0.014}$ +0.030	$0.753_{\pm 0.009}$ +0.073	$0.887_{\pm 0.014}$ +0.021
CLAM-SB Nature 2021	✗	✗	$0.925_{\pm 0.085}$	$0.969_{\pm 0.024}$	$0.523_{\pm 0.020}$	$0.846_{\pm 0.020}$	$0.631_{\pm 0.034}$	$0.863_{\pm 0.005}$
	✓	✗	$0.937_{\pm 0.004}$ +0.012	$0.979_{\pm 0.015}$ +0.010	$0.547_{\pm 0.006}$ +0.024	$0.887_{\pm 0.0014}$ +0.041	$0.678_{\pm 0.018}$ +0.047	$0.866_{\pm 0.007}$ +0.003
	✓	✓	$0.948_{\pm 0.014}$ +0.023	$0.981_{\pm 0.021}$ +0.012	$0.550_{\pm 0.006}$ +0.027	$0.886_{\pm 0.0015}$ +0.040	$0.679_{\pm 0.013}$ +0.048	$0.887_{\pm 0.002}$ +0.024
TransMIL NeurIPS 2021	✗	✗	$0.922_{\pm 0.019}$	$0.943_{\pm 0.009}$	$0.554_{\pm 0.048}$	$0.792_{\pm 0.029}$	$0.631_{\pm 0.030}$	$0.841_{\pm 0.006}$
	✓	✗	$0.931_{\pm 0.001}$ +0.009	$0.947_{\pm 0.008}$ +0.004	$0.577_{\pm 0.006}$ +0.023	$0.824_{\pm 0.012}$ +0.032	$0.647_{\pm 0.024}$ +0.016	$0.853_{\pm 0.021}$ +0.012
	✓	✓	$0.933_{\pm 0.023}$ +0.011	$0.954_{\pm 0.021}$ +0.011	$0.580_{\pm 0.008}$ +0.026	$0.829_{\pm 0.010}$ +0.037	$0.672_{\pm 0.024}$ +0.041	$0.883_{\pm 0.041}$ +0.045
DSMIL CVPR 2021b	✗	✗	$0.943_{\pm 0.007}$	$0.966_{\pm 0.009}$	$0.532_{\pm 0.064}$	$0.804_{\pm 0.032}$	$0.577_{\pm 0.028}$	$0.816_{\pm 0.028}$
	✓	✗	$0.943_{\pm 0.001}$ ±0.000	$0.974_{\pm 0.007}$ +0.008	$0.544_{\pm 0.038}$ +0.012	$0.819_{\pm 0.031}$ +0.015	$0.609_{\pm 0.012}$ +0.032	$0.837_{\pm 0.013}$ +0.021
	✓	✓	$0.942_{\pm 0.026}$ −0.001	$0.985_{\pm 0.022}$ +0.019	$0.559_{\pm 0.028}$ +0.027	$0.823_{\pm 0.019}$ +0.019	$0.612_{\pm 0.031}$ +0.035	$0.849_{\pm 0.042}$ +0.033

Overall, ASMIL demonstrates superior performance, achieving state-of-the-art performance on all datasets when paired with an in-domain ViT-SSL backbone, and remains competitive with the best baseline on ImageNet-pretrained ResNet-18 features. On the BRACS dataset, our method attains an F1 score of 0.781 and an AUC of 0.914, exceeding the previous best results by 3.9 and 0.9 percentage points, respectively. This shows its effectiveness in capturing subtle histopathological features in heterogeneous subtyping tasks.

For CAMELYON-16 and CAMELYON-17 datasets with sparse tumor regions, where malignant tissue may occupy as little as 5% of a slide (Cheng et al., 2021),the advantages are even more pronounced. on CAMELYON-16, we observe a 3.3% increase in F1 score and a 1.6% uplift in AUC compared to the strongest baseline; similarly, on CAMELYON-17, ASMIL improves the F1 score by 6.49%, which highlights ASMIL's efficacy under an ill-posed, weakly supervised task. We compare the computational cost of ASMIL with that of other benchmarks in Appendix M.1.

5.2 Integrating the Anchor Model and NSF with Other MIL Methods

We regard the anchor model as a general plug-in module for attention-based MIL in WSI analysis. Accordingly, for each baseline we evaluate two variants while keeping all other components and hyperparameters fixed: (i) **+Anchor** (EMA-updated anchor with attention matching), and (ii) **+Anchor+NSF** (anchor updated by EMA and using NSF). The results are summarized in Table 2. As shown, adding the anchor model and the NSF consistently improves performance, with F1 score gains up to 10.73% (for ABMIL on BRACS), except when adding the anchor to DSMIL on the CAMELYON-16 dataset, where the F1 score decreases by 0.001 relative to the original model. The additional computational cost introduced by the anchor model is reported in Appendix M.2.

5.3 Localization

We evaluate tumor localization on CAMELYON-16 both qualitatively and quantitatively. Qualitative heatmaps are shown in Figure 4. Compared with baseline methods, ASMIL consistently highlights all cancerous regions. We attribute these gains to reduced over-concentration by the NSF in the anchor model, which yields more faithful attention distributions.

Table 3: Component-wise ablation of ASMIL on BRACS. We evaluate the contribution of the anchor model, NSF, and random drop (rd).

Anchor	NSF	rd	F1 score ↑	AUC ↑
✓	✓	✓	$\mathbf{0.781_{\pm 0.042}}$	$\mathbf{0.914_{\pm 0.014}}$
✓	✓	✗	$0.765_{\pm 0.030}$	$0.903_{\pm 0.018}$
✓	✗	✓	$0.759_{\pm 0.028}$	$0.895_{\pm 0.012}$
✓	✗	✗	$0.747_{\pm 0.026}$	$0.887_{\pm 0.015}$
✗	✗	✓	$0.728_{\pm 0.019}$	$0.868_{\pm 0.010}$
✗	✗	✗	$0.712_{\pm 0.020}$	$0.860_{\pm 0.012}$

Following the official CAMELYON-16 and Fourkioti et al. (2024), we report lesion-level Free-Response ROC (FROC) (Miller, 1969; Bunch, 1978) the Dice coefficient on cancerous slides, and tile-level specificity on normal slides. To obtain the predicted masks, we use scaled attention distributions for CLAM (Lu et al., 2021), TransMIL (Shao et al., 2021), DSMIL (Li et al., 2021b), and CAMIL (Fourkioti et al., 2024); tile-level logits for DTFD-MIL (Zhang et al., 2022); and for ASMIL, the per-tile average of FEAT-token attentions. Quantitative results for FROC, Dice, and specificity, as well as additional attention-map visualizations, are provided in Appendix L.

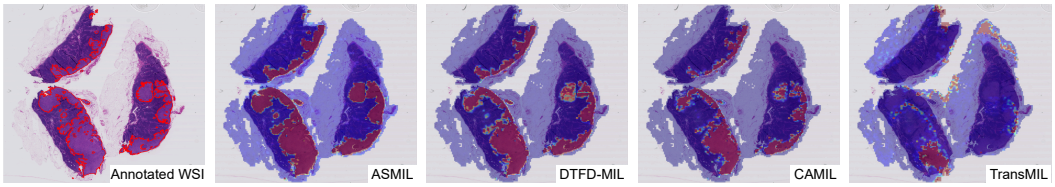

Figure 4: Visual comparison of attention maps on the CAMELYON-16 dataset. The left column shows the original WSI with ground-truth tumor annotations outlined in red; the remaining columns present attention maps for ASMIL (ours), DTFD-MIL, CAMIL, and TransMIL (left to right).

5.4 Ablation Study

Lastly, we evaluate the effect of the anchor model, NSF, and random drop (rd) by enabling or disabling them in all combinations. As shown in Table 3, the full model (all three enabled) achieves the best F1 score and AUC. Removing any component degrades performance, with the anchor model having the largest impact. Without all three, the model drops to the lowest scores, confirming that each component contributes to the overall effectiveness of ASMIL. Additional ablations on the loss weight β, the number of trainable FEAT tokens, the EMA factor m, the anchor update frequency, and the random drop rate are reported in Appendix K.

6 Conclusion

In this work, we identified a previously overlooked failure mode in attention-based MIL for WSI: unstable attention dynamics that hinder convergence. We proposed ASMIL, which stabilizes training via an anchor model, prevents over-concentration by using a normalized sigmoid in the anchor, and mitigates overfitting with token dropout. Across multiple WSI benchmarks, ASMIL improves classification performance and state-of-the-art localization performance. These results underscore the importance of jointly controlling attention stability, concentration, and overfitting in weakly supervised WSI analysis. We anticipate that the proposed anchor model and normalized sigmoid function will serve as building blocks for future MIL-based WSI analysis algorithms, ultimately facilitating more accurate and interpretable analysis of gigapixel pathology images. Due to space constraints, we defer the discussion of future work and limitations to Appendix Q.

Ethics Statement

All WSI datasets used in this work are publicly available and were obtained from open-access websites. The usage of these datasets strictly follows the terms and conditions set by the dataset providers and adheres to established academic and research community standards. No personally identifiable information or sensitive patient data is involved.

Reproducibility Statement

We have taken steps to ensure our results are reproducible. All model and algorithmic details, training procedures, hyperparameters, evaluation protocols, and metrics are specified in the main text. The appendix provides complete proofs, implementation notes, ablations, and additional qualitative results. An anonymized GitHub repository contains the source code and configuration files, and pre-trained checkpoints. All datasets used in our experiments are publicly available; download links, data splits, and preprocessing steps are documented in the repository and referenced in the appendix.

References

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/3e6260b81898beacda3d16db379ed329-Paper.pdf.

James Bacus. Method and apparatus for acquiring and reconstructing magnified specimen images from a computer-controlled microscope, December 2001. URL https://patents.google.com/patent/US20010050999A1/a. US Patent Application.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022.

PC Bunch. Free response approach to measurement and characterization of radiographic observer performance. *AJR Am J Roentgenol*, 130(2):382, 1978.

Zak Buzzard, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Paths: A hierarchical transformer for efficient whole slide image analysis. *arXiv preprint arXiv:2411.18225*, 2024.

Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

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 International Conference on Computer Vision (ICCV)*, 2021.

David Chapman and Ajay Jain. Musk (Version 1). UCI Machine Learning Repository, 1994a. DOI: https://doi.org/10.24432/C5ZK5B.

David Chapman and Ajay Jain. Musk (Version 2). UCI Machine Learning Repository, 1994b. DOI: https://doi.org/10.24432/C51608.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 339–349. Springer, 2021.

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

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

Jun Cheng, Yuting Liu, Wei Huang, Wenhui Hong, Lingling Wang, Xiaohui Zhan, Zhi Han, Dong Ni, Kun Huang, and Jie Zhang. Computational image analysis identifies histopathological image features associated with somatic mutations and patient survival in gastric adenocarcinoma. *Frontiers in Oncology*, 11:623382, 2021.

Zhixiang Chi, Li Gu, Tao Zhong, Huan Liu, YUANHAO YU, Konstantinos N Plataniotis, and Yang Wang. Adapting to distribution shift by visual domain prompt generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Kausik Das, Sri Phani Krishna Karri, Abhijit Guha Roy, Jyotirmoy Chatterjee, and Debdoot Sheet. Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 1024–1027, 2017. doi: 10.1109/ISBI.2017.7950690.

Kausik Das, Sailesh Conjeti, Abhijit Guha Roy, Jyotirmoy Chatterjee, and Debdoot Sheet. Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 578–581, 2018. doi: 10.1109/ISBI.2018.8363642.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1–2):31–71, January 1997. ISSN 0004-3702. doi: 10.1016/S0004-3702(96)00034-3. URL https://doi.org/10.1016/S0004-3702(96)00034-3.

Jiuyang Dong, Junjun Jiang, Kui Jiang, Jiahan Li, and Yongbing Zhang. Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30818–30828, 2025.

Zhaolong Du, Shasha Mao, Yimeng Zhang, Shuiping Gou, Licheng Jiao, and Lin Xiong. RGMIL: Guide your multiple-instance learning model with regressor. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=eGoE9CVRPc.

Zhaolong Du, Shasha Mao, Xuequan Lu, Mengnan Qi, Yimeng Zhang, Jing Gu, and Licheng Jiao. Rethinking multiple-instance learning from feature space to probability space. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=torbeUlslS.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 12 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

Olga Fourkioti, Matt De Vries, and Chris Bakal. CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rzBskAEmoc.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 755–764. Springer, 2023.

Shayan Mohajer Hamidi and Linfeng Ye. Adversarial training via adaptive knowledge amalgamation of an ensemble of teachers, 2024. URL https://arxiv.org/abs/2405.13324.

Shayan Mohajer Hamidi and Linfeng Ye. Distributed quasi-newton method for fair and fast federated learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=KbteA50cni.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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

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

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)*, pp. 3344–3354, June 2023.

James Keeler, David Rumelhart, and Wee Leow. Integrated segmentation and recognition of hand-printed numerals. In R.P. Lippmann, J. Moody, and D. Touretzky (eds.), *Advances in Neural Information Processing Systems*, volume 3. Morgan-Kaufmann, 1990. URL https://proceedings.neurips.cc/paper_files/paper/1990/file/e46de7e1bcaaced9a54f1e9d0d2f800d-Paper.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2021a. doi: 10.1109/CVPR46437.2021.01409.

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

Jiayun Li, Wenyuan Li, Arkadiusz Gertych, Beatrice S Knudsen, William Speier, and Corey W Arnold. An attention-based multi-resolution model for prostate whole slide imageclassification and localization. *arXiv preprint arXiv:1905.13208*, 2019.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023.

Pei Liu, Luping Ji, Jiaxiang Gou, Bo Fu, and Mao Ye. Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=trj2Jq8riA.

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

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19764–19775, June 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, NIPS '97, pp. 570–576, Cambridge, MA, USA, 1998. MIT Press. ISBN 0262100762.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/martins16.html.

Harold Miller. The froc curve: A representation of the observer's performance for the method of free response. *The Journal of the Acoustical Society of America*, 46(6B):1473–1476, 1969.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Linhao Qu, Yingfan Ma, Xiaoyuan Luo, Qinhao Guo, Manning Wang, and Zhijian Song. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9732–9744, 2024. doi: 10.1109/TCSVT.2024.3400876.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hfLqdquVt3.

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 A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=LKUfuWxajHc.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, XINHENG LYU, Tao Lin, and Lin Yang. Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rFpZnn11gj.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Disambiguated attention embedding for multi-instance partial-label learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=NYwbmCrrni.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4078–4087, October 2023b.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11343–11352, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Hervé Delingette. Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images. *Medical Image Analysis*, 85:102763, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102763. URL https://www.sciencedirect.com/science/article/pii/S1361841523000245.

Manuel Tran, Sophia Wagner, Wilko Weichert, Christian Matek, Melanie Boxberg, and Tingying Peng. Navigating through whole slide images with hierarchy, multi-object, and multi-scale data. *IEEE Transactions on Medical Imaging*, 44(5):2002–2015, 2025. doi: 10.1109/TMI.2025.3532728.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

Gregory Verghese, Jochen K Lennerz, Danny Ruta, Wen Ng, Selvam Thavaraj, Kalliopi P Siziopikou, Threnesan Naidoo, Swapnil Rane, Roberto Salgado, Sarah E Pinder, et al. Computational pathology in cancer diagnosis, prognosis, and prediction–present day and prospects. *The Journal of pathology*, 260(5):551–563, 2023.

Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1655–1665, 2023.

Yanan Wu, Zhixiang Chi, Yang Wang, Konstantinos N Plataniotis, and Songhe Feng. Test-time domain adaptation by learning domain-aware batch normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15961–15969, 2024.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=01KmhBsEPFO.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14138–14148, 2021.

Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3): 591–604, 2014.

En-hui Yang and Linfeng Ye. Markov knowledge distillation: Make nasty teachers trained by self-undermining knowledge distillation fully distillable. In *European Conference on Computer Vision*, pp. 154–171. Springer, 2024.

En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning for classification (2023). *URL https://arxiv.org/abs/2309.09123*, 5.

En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning: Framework and preliminary results. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 569–574, 2024. doi: 10.1109/ISIT57864.2024.10619241.

En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning for classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):15436–15448, 2025. doi: 10.1109/TNNLS.2025.3540014.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65:101789, 2020.

Linfeng Ye, Shayan Mohajer Hamidi, Renhao Tan, and En-Hui Yang. Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information, 2024. URL https://arxiv.org/abs/2401.08732.

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

Houston H Zhang, Tao Zhang, Baoze Lin, Yuanqi Xue, Yincheng Zhu, Huan Liu, Li Gu, Linfeng Ye, Ziqiang Wang, Xinxin Zuo, et al. Widget2code: From visual widgets to ui code via multimodal llms. *arXiv preprint arXiv:2512.19918*, 2025a.

Jingwei Zhang, Ke Ma, John Van Arnam, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3771–3779, 2021. doi: 10.1109/CVPRW53098.2021.00418.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 125–143. Springer, 2024.

Yunlong Zhang, Zhongyi Shui, Yunxuan Sun, Honglin Li, Jingxiong Li, Chenglu Zhu, and Lin Yang. Aem: Attention entropy maximization for multiple instance learning based whole slide image classification. *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2025b.

Wenhui Zhu, Peijie Qiu, Xiwen Chen, Oana M. Dumitrascu, and Yalin Wang. Pdl: Regularizing multiple instance learning with progressive dropout layers, 2023.

Wenhui Zhu, Peijie Qiu, Xiwen Chen, Zhangsihao Yang, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. How effective can dropout be in multiple instance learning ? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qsYHqLFCH5.

A Related Work on EMA Models and Anchoring Strategies

EMA-based target networks are central in self-supervised representation learning. Mean Teacher Tarvainen & Valpola (2017) maintains an EMA of student parameters and enforces prediction consistency with this temporal ensemble under limited supervision. BYOL (Grill et al., 2020; Wu et al., 2023) uses an EMA-updated target network to provide representation targets for an online network with an additional predictor, and avoids collapse through architectural asymmetry and the EMA update instead of negatives. DINO-style methods (Caron et al., 2021; Oquab et al., 2023) adapt EMA self-distillation to Vision Transformers, where an EMA teacher produces soft probability targets on multi-crop views; centering, sharpening, and a momentum schedule control the stability–adaptation trade-off of these targets.

DINOv3 (Siméoni et al., 2025) revisits EMA teachers for dense prediction and studies how dense features drift or collapse under long training. It introduces *Gram anchoring*, which aligns Gram matrices of patch–patch similarities between a student and its EMA teacher so that dense features remain close to a temporally smoothed reference. The EMA momentum and the strength of this anchoring loss jointly determine how strongly dense features are tied to the teacher versus how quickly they adapt.

ASMIL also maintains an EMA-updated copy of the model, but uses it in a different regime and on a different target. Training is fully supervised at the bag level, and the EMA branch does not supply pseudo-labels or representation targets. Instead, it defines a temporally smoothed *attention distribution* over tiles, and the anchor enters the loss only through the KL term in Eq. equation 7, while the bag-level cross-entropy in Eq. equation 9 provides all semantic supervision. The shared encoder and classifier parameters are optimized by standard backpropagation; the EMA update acts purely as a temporal regularizer on attention, in contrast to BYOL/DINO, which anchors global embeddings, and DINOv3, which anchors patch–patch similarity structure.

The same EMA hyperparameters induce an analogous stability–adaptation trade-off but at the level of attention rather than features. The EMA momentum in ASMIL sets how rapidly the anchor follows the online model, and the weight β on the KL term controls how strongly attention is pulled toward the temporally smoothed reference. Unlike BYOL and DINO/DINOv3, where EMA model is designed to avoid global representation collapse, ASMIL uses EMA anchoring to reduce unstable and over-concentrated attention patterns observed under purely online MIL training, while the supervised objective already discourages trivial constant-attention solutions.

B EXPERIMENTAL DETAILS

We train all models for 50 epochs with a batch size of 1, using Adam (weight decay 10^{-4}) and a cosine learning rate schedule with an initial learning rate of 10^{-4}. All reported results are averaged over five random seeds.

B.1 WSI PRE-PROCESSING

For all datasets, we used the publicly available CLAM WSI preprocessing toolbox (Lu et al., 2021) to segment tissue regions and divide each slide into non-overlapping 256×256 patches at $20\times$ magnification. Tissue segmentation was performed automatically using Otsu's thresholding. To reduce computational overhead and leverage previously learned representations, we adopted a ResNet-18 model (He et al., 2016) pretrained on ImageNet (Russakovsky et al., 2015) and an open-source self-supervised ViT-small model (Kang et al., 2023) as feature extractors[3]. The ViT-small model was pretrained on 36,666 whole slide images from The Cancer Genome Atlas (TCGA) and the internally collected TULIP dataset. For consistency and fairness in the subtyping task, we used the same feature extractors across all baseline methods.

For the localization experiments, following Tourniaire et al. (2023), we used a ResNet-18 backbone pretrained with SimCLR (Chen et al., 2020)[4]. This feature extractor maps each tile to a 1024-dimensional feature vector.

B.2 DATASETS

CAMELYON-16 (Ehteshami Bejnordi et al., 2017) is a widely used publicly available WSI dataset designed for lymph node metastasis detection. It contains 270 training and 129 test slides collected from two medical centers, with detailed pixel-level annotations provided by expert pathologists. Notably, some slides include only partial annotations, making the dataset particularly challenging due to the presence of small or sparse metastatic regions. CAMELYON-16 has become a standard benchmark for evaluating weakly supervised and fully supervised algorithms in computational pathology.

CAMELYON-17 (Bándi et al., 2019) extends the scope of CAMELYON-16 by including a total of 1,000 WSIs from five medical centers, making it a more diverse and clinically representative dataset. Among these, 500 slides are publicly available and come with slide-level labels, while the remaining 500 are held out for challenge-based evaluations. The inclusion of data from multiple institutions introduces significant variability in staining and scanning conditions, making CAMELYON-17 a suitable benchmark for testing the generalization performance of WSI-based models.

The BRACS dataset (Brancati et al., 2022) is a large-scale WSI dataset curated for the task of breast cancer subtype classification. It comprises 547 WSIs collected from several medical institutions and annotated by expert pathologists into clinically relevant categories: benign tumors, atypical tumors, and malignant tumors. These labels reflect the progression of breast lesions and are critical for diagnostic decision-making and treatment planning. BRACS captures a wide range of histological appearances and staining variations, making it a valuable resource for developing and benchmarking MIL and weakly supervised classification models in real-world clinical settings.

B.3 DATA SPLITS

Following Zhang et al. (2025b; 2024), we partition the datasets as follows. For CAMELYON-16, the WSIs are divided into training, validation, and test sets. The 270 WSIs from Hospital 1 are split, five times, into training (90%) and validation (10%) subsets; the 130 WSIs from Hospital 2 are used as a test set. The official test set of 129 WSIs is used for final evaluation. For CAMELYON-17, we use 500 WSIs in total: 300 WSIs from three hospitals for training/validation (90%, 10%) and 200 WSIs from two other hospitals for testing to assess out-of-distribution (OOD) performance. For BRACS, we follow the official split: 395 slides for training, 65 for validation, and 87 for testing.

[3]The checkpoint is available at `https://github.com/lunit-io/benchmark-ssl-pathology`.

[4]The checkpoint is available at `https://github.com/binli123/dsmil-wsi`.

The task is a three-class WSI classification—benign tumor, atypical tumor, and malignant tumor. All results are averaged over five random seeds, and we report the mean performance on the official competition test set.

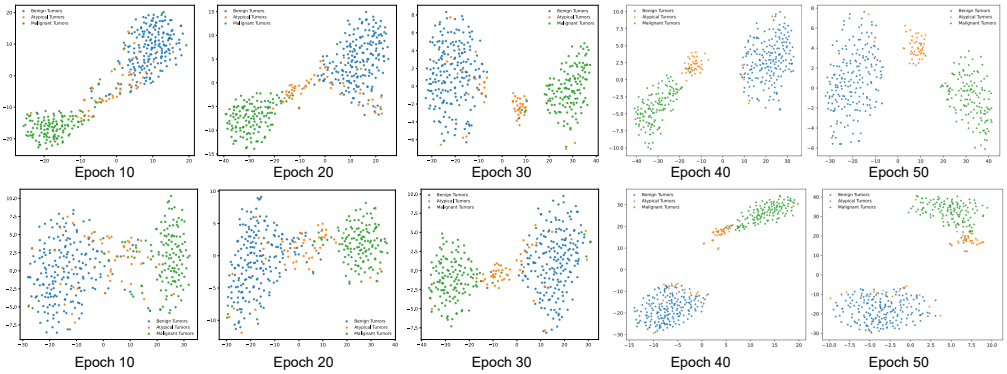

Figure 5: T-SNE embeddings of ASMIL bag-level features on the BRACS training set across training epochs. **Top:** with the anchor model; **Bottom:** without the anchor model.

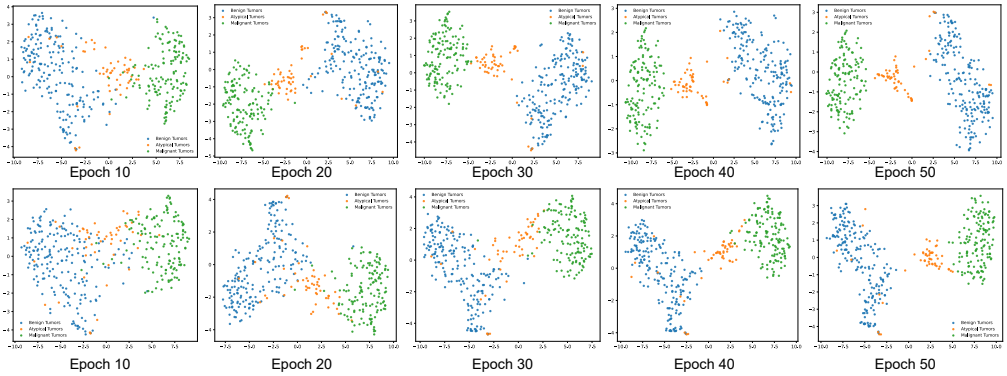

Figure 6: T-SNE embeddings of TransMIL bag-level features on the BRACS training set across training epochs. **Top:** with the anchor model; **Bottom:** without the anchor model.

C T-SNE Visualization of Bag-Level Features

To assess how the anchor model stabilizes attention during training, we visualize the bag-level representations learned by ASMIL using t-SNE Maaten & Hinton (2008); see Figure 5. Compared to ASMIL without the anchor, the model with an anchor forms more distinct clusters and exhibits clearer inter-class boundaries across training epochs, indicating faster convergence and more discriminative features. We observe a similar trend for TransMIL in Figure 6.

D Macro AUC and Macro F1 score under Class Imbalance

Since all datasets considered in this work are class-imbalanced, we report *macro-averaged* variants of the area under the ROC curve (AUC) and the F1 score as our primary summary metrics. Macro-averaging assigns equal weight to each class and therefore prevents majority classes from dominating the overall score.

Setup. Let $\mathcal{Y} = \{1, \ldots, K\}$ denote the set of classes. For a sample x with true label $y \in \mathcal{Y}$, let $s_k(x) \in \mathbb{R}$ be the model score for class k. Define one-vs-rest binary indicators $y_k = \mathbb{1}[y = k]$ for each class k, and the corresponding confusion-matrix counts $(\mathrm{TP}_k, \mathrm{FP}_k, \mathrm{FN}_k, \mathrm{TN}_k)$ computed by treating class k as "positive" and all others as "negative".

D.1 MACRO-$F1$

For class k, precision and recall are

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}, \qquad \text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}. \tag{11}$$

The per-class $F1$ is the harmonic mean of precision and recall:

$$F1_k \;=\; \frac{2\,\text{Precision}_k\,\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{12}$$

The *macro-$F1$* averages the per-class values uniformly:

$$\text{Macro-}F1 \;=\; \frac{1}{K}\sum_{k=1}^{K} F_{1,k}. \tag{13}$$

As a thresholded, decision-level metric, $F1_k$ (and thus macro-$F1$) depends on the classification threshold applied to scores $s_k(x)$. We use a threshold of 0.5 for all experiments. The same definition applies to multilabel settings by averaging over labels.

D.2 MACRO-AUC (ROC)

For class k, the ROC curve plots the true positive rate against the false positive rate as the threshold on $s_k(x)$ varies:

$$\text{TPR}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \qquad \text{FPR}_k = \frac{\text{FP}_k}{\text{FP}_k + \text{TN}_k}. \tag{14}$$

The per-class AUC, $\text{AUC}_k \in [0, 1]$, is the area under this curve; equivalently, it is the probability that a randomly chosen positive example (for class k) receives a higher score than a randomly chosen negative example. The *macro-AUC* is the uniform average across classes:

$$\text{Macro-AUC} \;=\; \frac{1}{K}\sum_{k=1}^{K} \text{AUC}_k. \tag{15}$$

Unlike $F1$, AUC is threshold-agnostic and measures the ranking quality of scores.

E PROOF OF THEOREM 1

Proof. We proceed in two parts.

Part A: NSF bounds. Let $s_i = \sigma(z_i)$ and $S = \sum_{j=1}^{N} \sigma(z_j)$, so $\alpha_i^{\text{nsf}} = s_i/S$.

Equalization among highs. For $i, h' \in \mathcal{H}$,

$$\frac{\alpha_i^{\text{nsf}}}{\alpha_{h'}^{\text{nsf}}} = \frac{s_i}{s_{h'}} = \frac{\sigma(z_i)}{\sigma(z_{h'})}. \tag{16}$$

Since σ is strictly increasing and $z_i, z_{h'} \in [\tau, \tau + \gamma]$,

$$\frac{\sigma(z_i)}{\sigma(z_{h'})} \;\le\; \frac{\sigma(\tau+\gamma)}{\sigma(\tau)} \;=\; \frac{1+e^{-\tau}}{1+e^{-(\tau+\gamma)}} \;\le\; 1 + e^{-\tau}. \tag{17}$$

Suppression of lows. For any $j \in \mathcal{L}$ we have $z_j \le -\tau$. Using monotonicity and the identity

$$\sigma(-t) = e^{-t}\,\sigma(t) \qquad \text{for all } t \in \mathbb{R}, \tag{18}$$

we get $\sigma(z_j) \le \sigma(-\tau) = e^{-\tau}\sigma(\tau)$. Meanwhile

$$S = \sum_{i=1}^{N} \sigma(z_i) \;\ge\; \sum_{i\in\mathcal{H}} \sigma(z_i) \;\ge\; h\,\sigma(\tau), \tag{19}$$

since $z_i \geq \tau$ for $i \in \mathcal{H}$. Hence

$$\alpha_j^{\text{nsf}} = \frac{\sigma(z_j)}{S} \;\leq\; \frac{e^{-\tau}\sigma(\tau)}{h\,\sigma(\tau)} \;=\; \frac{e^{-\tau}}{h}. \tag{20}$$

For completeness, equation 18 follows from $\sigma(-t) = \frac{1}{1+e^t} = \frac{e^{-t}}{1+e^{-t}} = e^{-t}\sigma(t)$.

Part B: Softmax temperature constraints. Fix $T > 0$ and $\boldsymbol{z} \in \mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$.

Equalization among highs. For any $i, h' \in \mathcal{H}$,

$$\frac{\alpha_i^{\text{smx}}}{\alpha_{h'}^{\text{smx}}} = \frac{e^{z_i/T}}{e^{z_{h'}/T}} = e^{(z_i - z_{h'})/T}. \tag{21}$$

Over $\mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$, the worst high to high ratio occurs at $z_i = \tau + \gamma$ and $z_{h'} = \tau$, so

$$\frac{\max_{i\in\mathcal{H}} \alpha_i^{\text{smx}}}{\min_{h'\in\mathcal{H}} \alpha_{h'}^{\text{smx}}} \;\geq\; e^{\gamma/T}. \tag{22}$$

Therefore, the uniform bound $\frac{\max_{i\in\mathcal{H}} \alpha_i^{\text{smx}}}{\min_{h'\in\mathcal{H}} \alpha_{h'}^{\text{smx}}} \;\leq\; \kappa$ for all $\boldsymbol{z} \in \mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$ implies

$$T \;\geq\; \frac{\gamma}{\log \kappa}. \tag{23}$$

Suppression of lows. Fix $j \in \mathcal{L}$. For a given T, the quantity $\alpha_j^{\text{smx}}(\boldsymbol{z}; T)$ is maximized over $\mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$ by taking $z_j = -\tau,\ z_i = \tau\ \forall i \in \mathcal{H},\ z_k \to -\infty$ for $k \notin \mathcal{H} \cup \{j\}$, which minimizes the denominator subject to the constraints. Thus

$$\sup_{\boldsymbol{z}\in\mathcal{S}(\tau,\gamma,\mathcal{H},\mathcal{L})} \alpha_j^{\text{smx}}(\boldsymbol{z}; T) = \frac{e^{-\tau/T}}{h\,e^{\tau/T} + e^{-\tau/T}} = \frac{1}{h\,e^{2\tau/T} + 1}. \tag{24}$$

Consequently, the uniform suppression requirement $\alpha_j^{\text{smx}}(\boldsymbol{z}; T) \leq \varepsilon$ for all $\boldsymbol{z} \in \mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$ forces

$$\frac{1}{h\,e^{2\tau/T} + 1} \;\leq\; \varepsilon \quad\Longleftrightarrow\quad h\,e^{2\tau/T} \;\geq\; \frac{1}{\varepsilon} - 1 \quad\Longleftrightarrow\quad T \;\leq\; \frac{2\tau}{\log\!\left(\frac{1}{\varepsilon} - 1\right) - \log h}. \tag{25}$$

Combining equation 23 and equation 25 yields the simultaneous constraints $T \leq \frac{2\tau}{\log\left(\frac{1}{\varepsilon}-1\right)-\log h}$, $T \geq \frac{\gamma}{\log\kappa}$. If

$$\frac{\gamma}{\log \kappa} \;>\; \frac{2\tau}{\log\!\left(\frac{1}{\varepsilon} - 1\right) - \log h}, \tag{26}$$

no T can satisfy both.

Instantiating NSF targets. Set $\varepsilon = \varepsilon_{\text{nsf}} = e^{-\tau}/h$ and $\kappa = \kappa_{\text{nsf}} = \frac{1+e^{-\tau}}{1+e^{-(\tau+\gamma)}}$. Then

$$\log\!\left(\tfrac{1}{\varepsilon_{\text{nsf}}} - 1\right) - \log h = \log\!\left(\frac{1}{e^{-\tau}/h} - 1\right) - \log h = \log\!\left(he^{\tau} - 1\right) - \log h = \log\!\left(e^{\tau} - h^{-1}\right), \tag{27}$$

so the right side of the incompatibility condition equals

$$\frac{2\tau}{\log\!\left(e^{\tau} - h^{-1}\right)} \;\xrightarrow[\tau\to\infty]{}\; 2. \tag{28}$$

Meanwhile,

$$\log \kappa_{\text{nsf}} = \log(1 + e^{-\tau}) - \log\!\left(1 + e^{-(\tau+\gamma)}\right) \tag{29}$$

$$= \log\!\left(1 + \frac{e^{-\tau}\,(1 - e^{-\gamma})}{1 + e^{-(\tau+\gamma)}}\right) \;\sim\; e^{-\tau}\,(1 - e^{-\gamma}) \quad (\tau \to \infty), \tag{30}$$

hence

$$\frac{\gamma}{\log \kappa_{\text{nsf}}} \;\xrightarrow[\tau\to\infty]{}\; \infty. \tag{31}$$

Therefore, for any fixed $\gamma > 0$, the incompatibility condition holds for all sufficiently large τ, so no single softmax temperature can match NSF uniformly on $\mathcal{S}(\tau, \gamma, \mathcal{H}, \mathcal{L})$. □

Remark 3 (Middle scores)**.** *Allowing additional scores in $(-\tau, \tau)$ only strengthens the NSF suppression bound because the denominator S increases, and it does not weaken the softmax lower bound equation 23 on the high to high ratio since that ratio is independent of other coordinates. The softmax low suppression supremum equation 24 is still attained by driving all non-high and non-j scores to $-\infty$, so the temperature constraints remain necessary.*

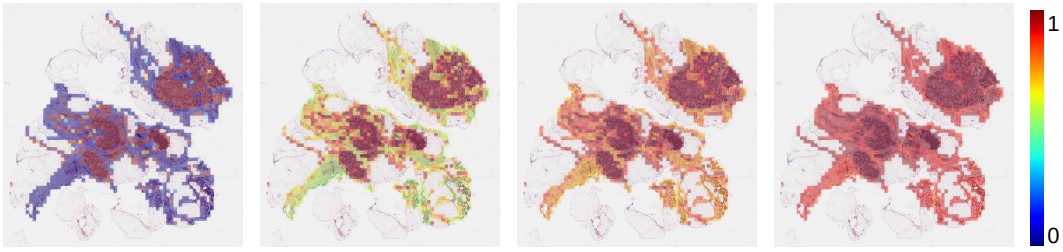

Figure 7: Ablation study of applying the softmax function with temperature scaling to the attention scores: (a) attention distribution of the proposed ASMIL, (b) softmax with $T = 2$ applied to the anchor model, (c) softmax with $T = 4$ applied to the anchor model, (d) softmax with $T = 8$ applied to the anchor model.

F Alternative to NSF in Anchor Model

F.1 Softmax With Temperature Scaling

A straightforward approach to mitigating over-concentration is to apply softmax with temperature scaling (Hinton et al., 2015; Ye et al., 2024; Yang et al.; 2024; 2025). This can indeed yield less concentrated attention distribution; however, as we observe in this section, a large temperature produces an overly smooth distribution, approaching a uniform distribution. This makes all tiles nearly indistinguishable, effectively reducing the operation to mean pooling and compromising interpretability. To illustrate this, we conduct experiments on the BRACS dataset using the same training protocol as in Section 5, summarize the results in Table 4, and visualize the attention maps in Figure 7.

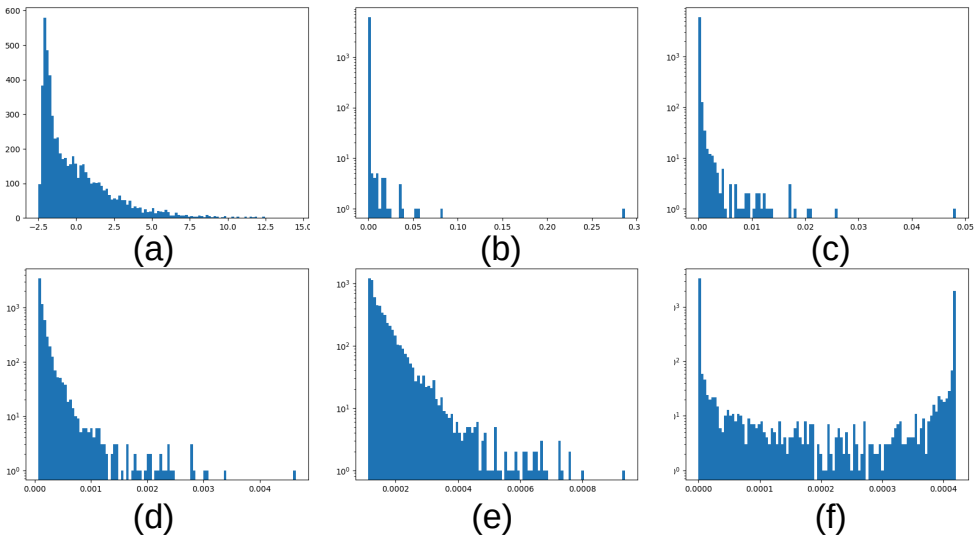

Figure 8: Histograms of (a) raw attention scores, (b) attention distribution obtained by the softmax function with temperature $T = 1$, (c) $T = 2$, (d) $T = 4$, (e) $T = 8$, and **(f) attention distribution computed using an NSF**. The Y-axis is displayed on a logarithmic scale for better visualization.

Furthermore, to clarify the differences between the NSF and softmax, we plot the histograms of the attention scores—(a) outputs from the NSF and softmax with various temperature scalings—in Figure 8. As shown, the saturation property of the NSF suppresses excessively large values.

Table 4: Subtyping performance on BRACS, when we apply softmax with temperature scaling to the anchor model.

BRACS							
Normalized Sigmoid		Softmax T=2		Softmax T=4		Softmax T=8	
F1 score ↑	AUC ↑	F1 score ↑	AUC ↑	F1 score ↑	AUC ↑	F1 score ↑	AUC ↑
$0.781_{\pm 0.042}$	$0.914_{\pm 0.014}$	$0.667_{\pm 0.049}$	$0.860_{\pm 0.027}$	$0.712_{\pm 0.029}$	$0.876_{\pm 0.012}$	$0.688_{\pm 0.037}$	$0.858_{\pm 0.031}$

Table 5: ASMIL performance when replacing NSF with entmax on CAMELYON-16.

CAMELYON-16						
Metric	NSF	entmax$_{\alpha=2}$ (sparsemax)	entmax$_{\alpha=1.75}$	entmax$_{\alpha=1.5}$ (entmax-15)	entmax$_{\alpha=1.25}$	entmax$_{\alpha=1}$ (softmax)
F1 score ↑	$0.965_{\pm 0.020}$	$0.938_{\pm 0.031}$	$0.927_{\pm 0.034}$	$0.937_{\pm 0.014}$	$0.910_{\pm 0.026}$	$0.942_{\pm 0.0147}$
AUC ↑	$0.985_{\pm 0.017}$	$0.964_{\pm 0.012}$	$0.959_{\pm 0.017}$	$0.960_{\pm 0.017}$	$0.925_{\pm 0.031}$	$0.963_{\pm 0.020}$
Time per epoch ↓	6.340s	8.451s	8.452s	8.451s	8.450s	6.336s

F.2 ENTMAX

Entmax is a family of mappings that convert a score vector $z \in \mathbb{R}^d$ into a probability vector $p \in \Delta^d$ by maximizing a linear score plus Tsallis-α entropy Tsallis (1988) H_α^T:

$$\text{entmax}_\alpha(z) = argmax_{p \in \Delta^d} z^T p + H_\alpha^T(p), \tag{32}$$

The solution admits a closed form

$$\boldsymbol{\alpha}_i = \Big[\frac{\alpha - 1}{\alpha}(z_i - \tau)\Big]_+^{\frac{1}{\alpha-1}}, \text{with } \sum_i \boldsymbol{\alpha}_i = 1, \tag{33}$$

where τ is a threshold chosen so that the probabilities sum to one. As limiting cases, $\alpha \to 1$, yields softmax, and while $\alpha = 2$ yields sparsemax (Martins & Astudillo, 2016).

While entmax offers controllable sparsity, two drawbacks are pertinent to MIL-based WSI analysis: (i) **Lack of selective flattening**, entmax is monotone in z on its active support and does not explicitly equalize top-probability entries. (ii) **Higher computational cost**. Computing τ in Equation (33) requires the bisection method, which adds non-trivial overhead vs. NSF's fully closed-form normalization. These differences matter for MIL on WSIs, where multiple correlated tumor foci can be present: we prefer mechanisms that both discourage over-peaky attention and keep computation predictable. We replaced NSF with $entmax_\alpha$ inside ASMIL and swept $\alpha \in \{2, 1.75, 1.5, 1.25, 1\}$. For $\alpha = 1$, we used PyTorch softmax; for $\alpha > 1$, we solved for τ via bisection. The implementation follows the reference code from DeepSPIN[5]. All other hyperparameters, model, and data pipeline were kept fixed. We report results on CAMELYON-16 in Table 5. As seen, across α, entmax underperforms NSF on both F1 and AUC and incurs a 33.5% increase in epoch time vs. NSF.

G APPLYING NORMALIZED SIGMOID TO THE ONLINE MODEL

One might question the rationale behind applying the NSF to the anchor model while using the softmax function for the online model during training. To investigate this design choice, we experiment with applying the NSF to both the online and anchor models and evaluate the model's subtyping performance on the CAMELYON-16 and BRACS datasets. The results, presented in Table 6, reveal a F1 score drop of over 6% on the BRACS dataset. We attribute this degradation to the inherent characteristics of the sigmoid function: when it saturates, its gradients diminish, leading to vanishing gradients in the attention mechanism and thereby impairing the learning process.

To further investigate the potential of applying NSF in the online model, we consider the following mixed attention variant:

$$\alpha_i'(z) = \zeta\, \alpha_i^{\text{SMX}}(z) + (1 - \zeta)\, \alpha_i^{\text{NSF}}(z), \tag{34}$$

where $\zeta = \sigma(\xi)$ and ξ is a trainable scalar that balances the contributions of the softmax and NSF mappings, initialized with $\xi = 0$. We evaluate this variant on CAMELYON-16, CAMELYON-17,

[5] https://github.com/deep-spin/entmax

Table 6: Ablation study on the impact of applying the normalized sigmoid (NS) function to both the online and anchor models. ✓ indicates that NSF is applied to both models, while ✗ denotes the default setting where NSF is applied only to the anchor model. Subtyping performance is evaluated on the CAMELYON-16 and BRACS datasets using F1 score and AUC. A significant performance drop is observed on CAMELYON-16 when NSF is applied to both models.

Dataset	CAMELYON-16	
Online NSF	F1 score ↑	AUC ↑
✓	$0.920_{\pm 0.020}$	$0.936_{\pm 0.021}$
✗	$0.965_{\pm 0.020}$	$0.985_{\pm 0.017}$
Dataset	**BRACS**	
Online NSF	F1 score ↑	AUC ↑
✓	$0.726_{\pm 0.014}$	$0.865_{\pm 0.017}$
✗	$0.781_{\pm 0.042}$	$0.914_{\pm 0.014}$

Table 7: Comparison between ASMIL trained with the standard softmax function in the online model (ASMIL w. Softmax) and with the mixed attention function defined in Equation (34) (ASMIL w. Mixture). The more flexible trainable mapping does not yield improvements over the simpler softmax baseline.

Dataset	CAMELYON-16		CAMELYON-17		BRACS	
Mertic	F1 score	AUC	F1 score	AUC	F1 score	AUC
ASMIL W. SoftMax	$0.965_{\pm 0.020}$	$0.985_{\pm 0.017}$	$0.689_{\pm 0.005}$	$0.898_{\pm 0.010}$	$0.781_{\pm 0.042}$	$0.914_{\pm 0.014}$
ASMIL W. Mixture	$0.953_{\pm 0.023}$	$0.972_{\pm 0.030}$	$0.686_{\pm 0.012}$	$0.889_{\pm 0.009}$	$0.774_{\pm 0.054}$	$0.910_{\pm 0.067}$
ζ in Equation (34)	0.9952		0.9894		0.9963	

and BRACS, and report the results in Table 7. The mixed mapping does not outperform the default softmax, and the learned ζ consistently converges to values close to one, indicating that the online model prefers softmax, which does not suffer from gradient-vanishing issues.

H Alternative Stabilization Methods and Why the Anchor is Preferable

Let $\boldsymbol{\alpha}_t(x) \in \Delta^N$ denote the attention distribution for slide x at epoch t, obtained from scores $\boldsymbol{z}_t(x) \in \mathbb{R}^N$. We diagnose instability by the Jensen-Shannon divergence

$$\mathrm{JSD}_t(x) \;=\; \mathrm{JSD}\big(\boldsymbol{\alpha}_t(x) \| \boldsymbol{\alpha}_{t-1}(x)\big), \tag{35}$$

which we empirically find remains high when training attention-based MIL with only bag-level labels. We present a natural alternative that targets this instability and explain why the anchor model is preferred.

H.1 Alternative: Per-slide temporal ensembling of attention

Maintain a per slide exponential moving average (EMA) of past attentions and penalize deviation from it:

$$\tilde{\boldsymbol{\alpha}}_t(x) \;=\; \rho\, \tilde{\boldsymbol{\alpha}}_{t-1}(x) \;+\; (1-\rho)\, \boldsymbol{\alpha}_t(x),\; \rho \in (0,1); \qquad \mathcal{L}_{\mathrm{AS}}(x) \;= \mathsf{KL}\Big(\boldsymbol{\alpha}_t(x) \| \mathtt{sg}\big(\tilde{\boldsymbol{\alpha}}_t(x)\big)\Big). \tag{36}$$

The EMA target changes slowly when ρ is close to one, which directly shrinks epoch-to-epoch drift of $\boldsymbol{\alpha}_t$ and reduces $\mathsf{JSD}(\boldsymbol{\alpha}_t \| \boldsymbol{\alpha}_{t-1})$. *However,*

(i) It has to maintain a length-N vector per slide. For S slides and average $\bar{N}$ tiles, memory is $O(S\bar{N})$ floats, which can be substantial for gigapixel WSIs and prevent scaling to larger datasets. (ii) The EMA target still uses softmax normalization, which cannot achieve selective flattening across informative tokens; see Theorem 1.

H.2 WHY ASMIL'S ANCHOR IS PREFERABLE

We highlight two main reasons for using an anchor model to stabilize the attention distribution rather than relying on temporal ensembling.

NSF provides selective flattening that softmax cannot match.

Replacing softmax with the normalized sigmoid function (NSF) in the anchor yields $\boldsymbol{\alpha}^{\mathrm{nsf}}(x)$, which equalizes probabilities among truly high-score tiles while suppressing low-score ones. By Theorem 1, no single softmax temperature can realize both behaviors across a broad class of score vectors. Consequently, methods that retain softmax-based targets inherit these limitations.

Memory and implementation simplicity.

The anchor-based approach adds only one extra forward pass and maintains an exponential moving average (EMA) of the anchor parameters during training. It does not require storing per-slide attention distributions, making the approach scalable to large WSI datasets.

Thus, an anchor model is preferable for scalable training on large MIL datasets and for preventing attention over-concentration.

I WHY MATCHING THE TEACHER (ANCHOR) MODEL'S SOFTMAX FEATURE VECTOR CANNOT STABILIZE THE ATTENTION DISTRIBUTION

Table 8: Ratio of affinely dependent feature bags in the CAMELYON-16, CAMELYON-17, and BRACS datasets; most bags are affinely dependent.

Dataset	CAMELYON-16	CAMELYON-17	BRACS
The ratio of affine dependent feature bags	99.24%	99.80%	96.08%

In this section, we show why matching the softmax of the bag-level feature is a suboptimal strategy for stabilizing attention distributions. To this end, we prove that recovering the attention vector $\boldsymbol{\alpha}$ by matching $\mathrm{softmax}(\alpha^T X)$ is, in general, ill-posed: the map $f : \Delta^K \to \Delta^d$, defined by $f(\alpha) = \mathrm{softmax}(\alpha^T X)$ with $X \in \mathbb{R}^{K\times d}$, fails to be injective when the feature matrix X is affinely dependent.

Proof. Assume the rows $x_1, \ldots, x_K \in \mathbb{R}^d$ of X are affinely dependent. By definition there exists a nonzero vector $\psi \in \mathbb{R}^K$ such that

$$\sum_{i=1}^{K} \psi_i = 0 \quad \text{and} \quad \sum_{i=1}^{K} \psi_i x_i = 0.$$

Let $\alpha \in \Delta^K$ be any probability vector and choose $\epsilon > 0$ small enough that $\alpha' = \alpha + \epsilon\psi$ satisfies $\alpha'_i \geq 0$ for every i. Note $\sum_i \alpha'_i = \sum_i \alpha_i + \epsilon \sum_i \psi_i = 1$, so $\alpha' \in \Delta^K$. Since $\sum_{i=1}^K \psi_i x_i = 0$ we have

$$(\alpha')^T X = \alpha^T X + \epsilon \psi^T X = \alpha^T X.$$

Therefore

$$f(\alpha') = \mathrm{softmax}((\alpha')^T X) = \mathrm{softmax}(\alpha^T X) = f(\alpha).$$

Because $\psi \neq 0$ and $\epsilon \neq 0$ we have $\alpha' \neq \alpha$, hence f is not injective. □

Thus, matching the softmax of the bag feature cannot reliably recover or stabilize the attention distributions when the feature bag is affinely dependent. Table 8 confirms that most feature bags extracted by VIT-S (Kang et al., 2023) from WSI datasets are indeed affinely dependent.

Table 9: The F1 score and AUC of different MIL approaches on two WSI subtyping datasets.

PathGen-Clip-VIT-L				
Dataset	CAMELYON-16		CAMELYON-17	
Method	F1 score ↑	AUC ↑	F1 score ↑	AUC ↑
Clam-SB	$0.941_{\pm 0.014}$	$0.960_{\pm 0.015}$	$0.622_{\pm 0.031}$	$0.899_{\pm 0.012}$
TransMIL	$0.951_{\pm 0.024}$	$0.968_{\pm 0.028}$	$0.656_{\pm 0.021}$	$0.892_{\pm 0.014}$
DSMIL	$0.895_{\pm 0.038}$	$0.949_{\pm 0.017}$	$0.582_{\pm 0.062}$	$0.887_{\pm 0.013}$
IBMIL	$0.935_{\pm 0.014}$	$0.953_{\pm 0.009}$	$0.629_{\pm 0.027}$	$0.884_{\pm 0.016}$
MHIM-MIL	$0.946_{\pm 0.33}$	$0.984_{\pm 0.016}$	$0.594_{\pm 0.090}$	$0.912_{\pm 0.009}$
ABMIL	$0.953_{\pm 0.018}$	$0.972_{\pm 0.010}$	$0.610_{\pm 0.025}$	$0.864_{\pm 0.017}$
AEM	$0.967_{\pm 0.025}$	$0.988_{\pm 0.013}$	$0.688_{\pm 0.016}$	$0.905_{\pm 0.005}$
ASMIL	$0.974_{\pm 0.021}$	$0.990_{\pm 0.014}$	$0.699_{\pm 0.020}$	$0.929_{\pm 0.016}$
UNI-VIT-L				
Method	F1 score ↑	AUC ↑	F1 score ↑	AUC ↑
ABMIL	$0.968_{\pm 0.011}$	$0.996_{\pm 0.003}$	$0.605_{\pm 0.047}$	$0.885_{\pm 0.015}$
AEM	$0.975_{\pm 0.003}$	$0.998_{\pm 0.003}$	$0.633_{\pm 0.024}$	$0.863_{\pm 0.017}$
ASMIL	$0.980_{\pm 0.004}$	$0.998_{\pm 0.002}$	$0.672_{\pm 0.035}$	$0.866_{\pm 0.014}$

J Applying ASMIL to Features Extracted by a WSI Foundation Model

In recent years, foundation models have enabled strong open-source feature extractors that markedly improve the performance of computational-pathology systems. To assess the generalizability of our approach, we apply ASMIL to features produced by two such extractors, UNI Chen et al. (2024) and PATHGEN-clip Sun et al. (2025), for the subtyping task on the CAMELYON-16 and CAMELYON-17 datasets. As reported in Table 9, ASMIL consistently outperforms all baseline methods when used with features extracted by foundation models, yielding the best F1 and AUC.

K Ablation Study

K.1 Ablation of the Coefficient β

The coefficient $\beta > 0$ in Equation (9) balances the stabilization and classification objectives. To assess its impact on final performance, we sweep $\beta \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 4, 5\}$ on the CAMELYON-16 and BRACS datasets. Except for β, all experimental settings are identical to those in Section 5.1. We report F1 score and AUC in Figures 9 and 10; results are averaged over five random seeds. Overall, model performance is relatively insensitive to the choice of β: both

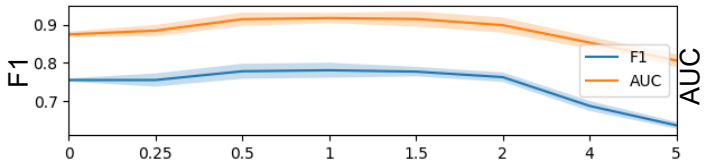

Figure 9: Ablation study on the coefficient β, on CAMELYON-16 dataset.

F1 score and AUC plateau for $\beta \in [0.5, 1.5]$. Accordingly, we set $\beta = 1$ as the default for all experiments.

K.2 Ablation Study on Number of Trainable FEAT Tokens

In this section, we investigate how varying the number of trainable tokens influences model performance. To this end, we sweep a number of trainable tokens in the range of $[2, 4, 8, 16]$, and report the corresponding accuracy on CAMELYON-16, CAMELYON-17, and BRACS in Table 10 In the experiment, we apply 8 trainable tokens for CAMELYON-16 and BRACS, and 16 trainable tokens on the CAMELYON-17 dataset.

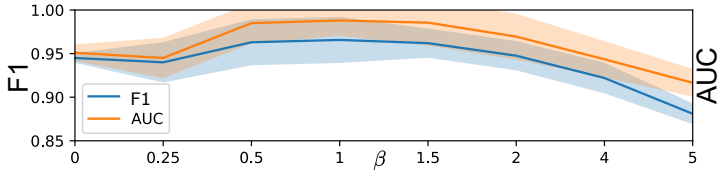

Figure 10: Ablation study on the coefficient β, on BRACS dataset.

Table 10: Ablation results on the number of tokens on different WSI datasets.

CAMELYON-16				
# FEAT tokens	2	4	8	16
F1 score ↑	$0.930_{\pm 0.012}$	$0.946_{\pm 0.009}$	$0.965_{\pm 0.012}$	$0.960_{\pm 0.006}$
AUC ↑	$0.932_{\pm 0.017}$	$0.973_{\pm 0.011}$	$0.985_{\pm 0.013}$	$0.981_{\pm 0.009}$
CAMELYON-17				
F1 score ↑	$0.556_{\pm 0.012}$	$0.610_{\pm 0.009}$	$0.674_{\pm 0.016}$	$0.689_{\pm 0.005}$
AUC ↑	$0.784_{\pm 0.019}$	$0.833_{\pm 0.011}$	$0.879_{\pm 0.024}$	$0.898_{\pm 0.010}$
BRACS				
F1 score ↑	$0.721_{\pm 0.009}$	$0.766_{\pm 0.012}$	$0.781_{\pm 0.004}$	$0.782_{\pm 0.004}$
AUC ↑	$0.871_{\pm 0.004}$	$0.903_{\pm 0.014}$	$0.914_{\pm 0.004}$	$0.912_{\pm 0.026}$

K.3 Ablation on anchor model update

K.3.1 Effect of Anchor Model Update Frequency

Table 11: Ablation study on anchor model update frequency, where batch-wise updates consistently outperform epoch-wise updates in both F1 score and AUC on BRACS and CAMELYON-16.

Dataset	BRACS	
Update	F1 score ↑	AUC ↑
Epoch	$0.742_{\pm 0.015}$	$0.871_{\pm 0.003}$
Batch	$0.781_{\pm 0.042}$	$0.914_{\pm 0.014}$
Dataset	**CAMELYON-16**	
Update	F1 score ↑	AUC ↑
Epoch	$0.920_{\pm 0.020}$	$0.936_{\pm 0.021}$
Batch	$0.965_{\pm 0.020}$	$0.984_{\pm 0.017}$

To assess the impact of anchor update frequency, we compare epoch-wise and batch-wise update strategies on BRACS and CAMELYON-16 (Table 11). The results show that batch-wise updates consistently deliver superior performance. On BRACS, batch-wise updates improve the F1 score by 3.9% and the AUC by 4.9%. On CAMELYON-16, the improvement is even more substantial, with the F1 score increasing by 4.9% and the AUC by 5.1%. These gains confirm that frequent updates enable the anchor model to provide a stable and closely aligned attention reference for the online model, leading to better performance.

K.4 Impact of the Random Drop Rate

We evaluated the effect of random token dropping on model performance using CAMELYON-16 and BRACS, measuring both F1 and AUC across several trainable-token budgets. Results in Figure 11 show a consistent trend: performance rises from low B, peaks around $B = 0.5$, then degrades for larger values. This pattern holds across datasets and capacities, indicating a stable trade-off between regularization and information loss.

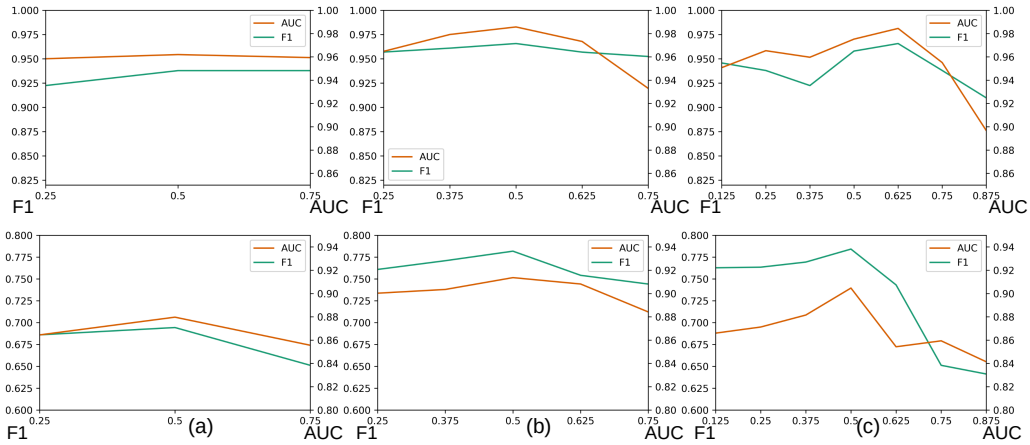

Figure 11: Ablation study of random drop probability (B) vs. model F1 score and AUC on CAMELYON-16 (top row) and BRACS (bottom row). Across both datasets and trainable-token settings (a) 4 tokens, (b) 8 tokens, and (c) 16 tokens, the test F1 score and AUC consistently peak around $B = 0.5$.

Mechanistically, moderate token dropping (0.4–0.7) provides useful regularization, encouraging robustness to missing context and reducing overfitting to redundant or spurious tiles, while excessive dropping increases the chance of discarding diagnostically critical patches and thus harms recall and ranking. We therefore recommend tuning B in the range of $0.4 - 0.6$. In Appendix K.5 we plot test F1 score and AUC across training epochs to demonstrate that random token dropping mitigates overfitting.

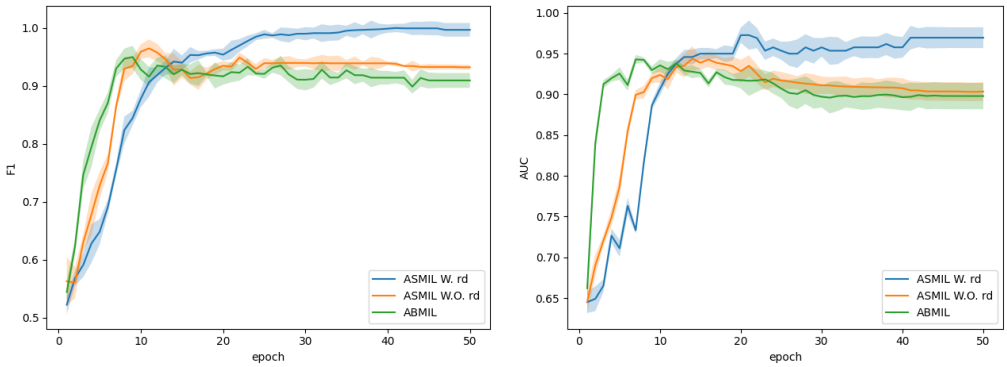

Figure 12: Performance comparison between ABMIL (Ilse et al., 2018), ASMIL with random drop (ASMIL W. rd), and ASMIL without random drop (ASMIL w/o rd). Both ABMIL and ASMIL w/o rd show signs of overfitting, as their F1 score and AUC peak and then decline. In contrast, ASMIL with random drop maintains stable performance across training, demonstrating that random drop effectively mitigates overfitting.

K.5 Random Drop Mitigates Overfitting

To verify that random drop is an efficient regularizer for attention-based MIL on WSIs, we trained three variants on CAMELYON-16: (i) ABMIL, (ii) ASMIL without random drop, and (iii) ASMIL with random drop (ours) with $B = 0.5$. The figure reports validation F1 and AUC over training epochs.

As shown in the Figure 12, both ABMIL and ASMIL without random drop exhibit overfitting: F1 score and AUC rise early, peak, and then decline with continued training. In contrast, ASMIL with random drop maintains high and stable F1/AUC throughout later epochs, with noticeably reduced run-to-run variability (shaded regions). These trajectories empirically validate that random drop curbs the late-epoch degradation that accompanies weak supervision on CAMELYON-16. This

Table 12: Statistical comparison of ASMIL with and without the anchor model. We report the mean performance over 10 random seeds along with p-values from DeLong tests for AUC and permutation tests for F1.

Dataset	Model	AUC	F1	p_{AUC}	p_{F1}
CAMELYON-16	w/o anchor	0.942	0.979	0.013	0.024
	w/ anchor	0.967	0.983		
CAMELYON-17	w/o anchor	0.642	0.879	0.024	0.035
	w/ anchor	0.693	0.899		
BRACS	w/o anchor	0.729	0.866	0.012	0.009
	w/ anchor	0.784	0.916		

observation aligns with our analysis that overfitting is a recurring failure mode for attention-based MIL on WSI datasets.

K.6 Significance test on the effect of online model

To assess whether the performance gains from the anchor model are statistically meaningful, we perform paired significance tests between ASMIL with and without the anchor over multiple 10 seeds. For AUC, we apply DeLong's test, and for F1, we use a non-parametric permutation test. Across CAMELYON-16, CAMELYON-17, and BRACS, the anchor-augmented ASMIL consistently achieves higher AUC and F1 than its non-anchor counterpart, and these improvements are statistically significant ($p < 0.05$) for both metrics on each dataset (see Table 12).

L Quantitative localization results and Additional Visualization

Predicted masks are generated as follows. For attention-based methods (CLAM (Lu et al., 2021), TransMIL (Shao et al., 2021), DTFD-MIL (Zhang et al., 2022), DSMIL (Li et al., 2021b) and CAMIL Fourkioti et al. (2024)), we use the tile-level attention distribution. For ASMIL, the per-tile attention distribution is computed by averaging the attention distributions from all FEAT tokens to that tile. Unless otherwise noted, we rescale all per-tile scores to $[0, 1]$ and threshold at 0.5 to produce binary masks across all methods.

For tumor localization on CAMELYON-16, we follow the official challenge protocol and report the lesion-level Free-Response ROC (FROC) (Miller, 1969; Bunch, 1978; Zhang et al., 2025a). Concretely, model outputs are converted to point detections; a detection is counted as a true positive if it lies within 75 µm of any ground-truth tumor region (implemented in the official script via a distance-transform "evaluation mask"), otherwise it is a false positive. We then sweep the detection score threshold to trace sensitivity versus the average number of false positives per normal WSI, and compute the standard CAMELYON-16 FROC score as the mean sensitivity at 0.25, 0.5, 1, 2, 4, 8 FP/WSI.

Quantitative results for FROC, Dice, and specificity are reported in Table 13, ASMIL achieves the best FROC and Dice on cancerous slides and higher specificity on normal slides, yielding fewer false positives and more contiguous lesion maps compared to baselines.

Figure 13 presents additional visualizations on the CAMELYON-16 dataset. It shows ASMIL attention maps for tumor slides containing both small and large cancerous regions; rows 1 and 3 provide the ground-truth annotations, and rows 2 and 4 show the corresponding attention maps.

M Computational Cost

This section reports the computational cost of ASMIL, as well as the additional cost incurred when integrating the anchor model into the baseline methods.

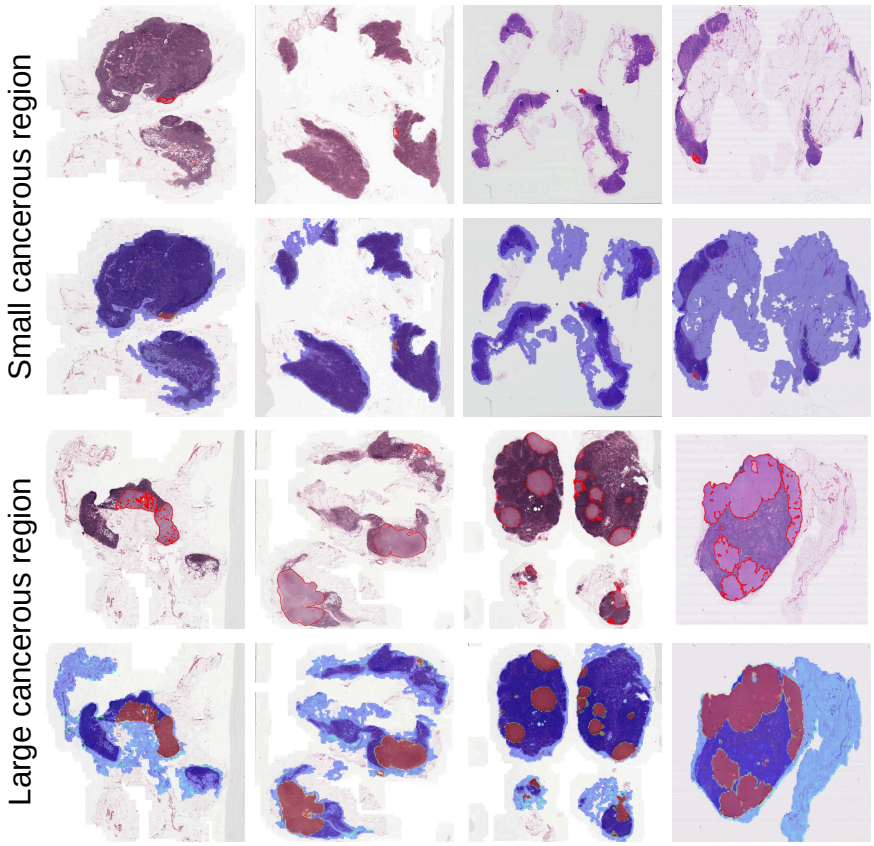

Figure 13: Additional qualitative examples of tumor regions and ASMIL attention maps. Rows 1 and 3 show the ground-truth tumor annotations (cancerous regions outlined in red), and rows 2 and 4 show the corresponding ASMIL attention maps.

Table 13: Localization results on CAMELYON-16.

Method	Dice ↑	Specificity ↑	FROC ↑
CLAM-SB	0.459	0.987	0.4257
TransMIL	0.103	0.999	0.0866
DTFD-MIL	0.525	0.999	0.4712
DSMIL	0.259	0.863	0.4506
CAMIL	0.515	0.980	0.4612
ASMIL	0.586	0.999	0.4941

M.1 Comparison of the computational cost between ASMIL and baseline methods

We conducted a detailed evaluation of the computational overhead introduced by our proposed ASMIL framework, focusing on three primary metrics: floating-point operations (FLOPs), training time per epoch, and peak memory consumption. All experiments were executed under uniform hardware conditions, specifically a single NVIDIA RTX 5000 GPU coupled with an Intel Xeon W-2265 CPU and 64 GB of RAM, ensuring a fair comparison across methods.

During training, ASMIL demonstrates a competitive balance between efficiency and computational demand. On average, ASMIL requires 542M FLOPs per batch, which is lower than MHIM-MIL. The training time per epoch for ASMIL is 7.49s, substantially faster than MHIM-MIL (19.4s) and comparable to TransMIL (5.99s), while remaining higher than ABMIL and CLAM-SB. In terms of peak memory usage, ASMIL consumes 570 MB, markedly lower than MHIM-MIL (2178 MB).

Table 14: Computational cost on BRACS (lower is better). We report training time and peak memory per epoch, and inference FLOPs, latency, and memory. ASMIL (ours) delivers efficient inference, cutting compute by 30.6%, latency by 29.2%, and memory by 20.3% compared with TransMIL, while requiring $4\times$ less training memory than MHIM-MIL.

BRACS	Training				
Method	CLAM-SB	ABMIL	TransMIL	MHIM-MIL	ASMIL
Time	2.26s	0.95s	5.99s	19.4s	7.49s
Memory	94MB	90MB	340MB	2178MB	570MB
BRACS	Inference				
FLOPs	162M	164M	781M	345M	542M
Time	0.45s	0.37s	0.74s	0.40s	0.52s
Memory	69MB	39MB	246MB	61MB	196MB

Table 15: Inference FLOPs, training time per epoch (Time), and memory usage (Memory) for four well-known methods, CLAM-SB, TransMIL, DSMIL, and ABMIL, with and without the anchor model. The anchor model incurs only minor computational overhead. FLOPs are measured using a fixed bag size of 2000 instances.

BRACS	Training		Inference		
Method	Time	Memory	FLOPs	Time	Memory
CLAM-SB w/o anchor	2.26s	94MB	162M	0.45s	69MB
CLAM-SB W. anchor	2.69s	120MB	162M	0.45s	69MB
TransMIL w/o anchor	5.99s	340MB	781M	0.74s	246MB
TransMIL W. anchor	7.27s	443MB	781M	0.74s	246MB
DSMIL w/o anchor	0.57s	60 MB	103M	1.09s	113 MB
DSMIL W. anchor	0.58s	145MB	103M	1.09s	113 MB
ABMIL w/o anchor	0.95s	90MB	164M	0.37s	39MB
ABMIL W. anchor	1.17s	162MB	164M	0.37s	39MB

These results indicate that ASMIL maintains a favorable computational profile, offering a scalable alternative to more resource-intensive methods.

In inference, ASMIL continues to show strong efficiency. It requires 542M FLOPs, substantially fewer than TransMIL and comparable to MHIM-MIL. Inference time for ASMIL is 0.52s per epoch, slightly slower than CLAM-SB (0.45s) but faster than TransMIL. Peak memory usage during inference is 196 MB, markedly lower than TransMIL, highlighting ASMIL's efficient memory footprint relative to its computational performance. Overall, ASMIL delivers high-performance multiple-instance learning while keeping computational cost affordable.

M.2 Additional Computational Cost Introduced by Anchor Model

We conducted a detailed evaluation of the computational overhead introduced by integrating the anchor model into four widely used MIL methods, namely CLAM-SB, TransMIL, DSMIL, and ABMIL, all measured on the BRACS dataset. The results are summarized in Table 15.

Because no gradients are computed through the anchor model, and only the attention layer is updated during training, the computational overhead is small. As shown in Table 15, integrating the anchor model into CLAM-SB, TransMIL, DSMIL, and ABMIL introduces only a modest increase in training time and memory usage, while the FLOPs remain unchanged. For example, training time for CLAM-SB increases from 2.26s to 2.69s and memory usage from 94 MB to 120 MB, with larger models like TransMIL showing slightly higher overhead. Importantly, during inference, the anchor model is discarded, resulting in identical FLOPs, execution time, and memory consumption compared to the baseline methods. These results demonstrate that the anchor model provides performance benefits during training with minimal computational cost and does not affect deployment efficiency, making it an effective and practical addition to existing MIL frameworks.

Table 16: C-index for WSI-based survival prediction using vision-only MIL models.

Method	BLCA	BRCA	GBMLGG	LUAD	UCEC
ABMIL ICML 2018	$0.5581_{\pm 0.031}$	$0.5825_{\pm 0.035}$	$0.7935_{\pm 0.032}$	$0.6121_{\pm 0.050}$	$0.6667_{\pm 0.033}$
TransMIL NeurIPS 2021	$0.5885_{\pm 0.055}$	$0.6140_{\pm 0.060}$	$0.7956_{\pm 0.015}$	$0.5708_{\pm 0.050}$	$0.6380_{\pm 0.067}$
ILRA ICLR 2023	$0.5549_{\pm 0.053}$	$0.5705_{\pm 0.067}$	$0.7742_{\pm 0.014}$	$0.5179_{\pm 0.081}$	$0.6503_{\pm 0.064}$
R^2T-MIL CVPR 2024	$0.5775_{\pm 0.024}$	$0.5476_{\pm 0.095}$	$0.7757_{\pm 0.024}$	$0.5711_{\pm 0.076}$	$0.6510_{\pm 0.087}$
DeepAttnMISL MIA 2020	$0.5646_{\pm 0.035}$	$0.5346_{\pm 0.036}$	$0.6750_{\pm 0.048}$	$0.4678_{\pm 0.039}$	$0.6259_{\pm 0.086}$
Patch-GCN MICCAI 2021	$0.6124_{\pm 0.031}$	$0.6375_{\pm 0.033}$	$0.7999_{\pm 0.021}$	$0.5922_{\pm 0.053}$	$0.7212_{\pm 0.025}$
ASMIL (Ours)	$0.6133_{\pm 0.047}$	$0.6396_{\pm 0.044}$	$0.8036_{\pm 0.018}$	$0.6001_{\pm 0.093}$	$0.7243_{\pm 0.0488}$

N Survival Prediction

To assess whether ASMIL is also beneficial for prognosis, we extend ASMIL from slide-level classification to discrete-time overall survival prediction on histopathology WSIs. Following (Liu et al., 2025), we apply an incidence-based discrete survival formulation, *i.e.*, the survival times are mapped to C non-overlapping time intervals, and the model outputs a discrete distribution over first-event times.

We follow the experimental setup of (Liu et al., 2025), and evaluate on five TCGA datasets, namely BLCA, BRCA, LUAD, and UCEC. We use the concordance index (C-index) to evaluate the model's performance; specifically, it measures how often the model assigns a higher risk score to a patient who experiences the event earlier. Formally, with a little abuse of notations, let $t_i, \delta_i, \hat{R}_i$ denote the observed time, event indicator, and predicted risk for patient i, the C-index is defined as

$$\text{CI} = \frac{\sum_{i,j} \mathbf{1}[t_i < t_j]\mathbf{1}[\hat{R}_i > \hat{R}_j]\delta_i}{\sum_{i,j} \mathbf{1}[t_i < t_j]\delta_i}, \tag{37}$$

where $\mathbf{1}[\cdot]$ is the indicator function. A value of $\text{CI} = 0.5$ corresponds to a random ranking, and larger values indicate better risk discrimination Yang & Ye (2024); Hamidi & Ye (2024; 2025). Following Liu et al. (2025), we compare ASMIL against six vision-only WSI survival prediction methods, namely ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), ILRA Xiang & Zhang (2023), R^2T-MIL (Tang et al., 2024), DeepAttnMISL (Yao et al., 2020), and Patch-GCN (Chen et al., 2021), all implemented on top of the same CONCH-derived patch features (Lu et al., 2023).

Table 16 reports the C-index on each TCGA dataset. ASMIL achieves the highest mean C-index among all vision-only baselines. These results indicate that stabilizing slide-level attention not only improves weakly supervised classification but also yields stronger prognostic discrimination in survival analysis.

O Evaluate ASMIL Over Non-WSI Dataset

Table 17: MIL dataset statistics.

Dataset	Domain	Bags (pos/neg)	Total instances	Dim./inst.
MUSK1	Drug activity	92 (47/45)	476	166
MUSK2	Drug activity	102 (39/63)	6598	166
TIGER	Images (Blobworld segments)	200 (100/100)	1220	230
FOX	Images (Blobworld segments)	200 (100/100)	1320	230
ELEPHANT	Images (Blobworld segments)	200 (100/100)	1391	230

To demonstrate ASMIL's applicability beyond WSI, we evaluate it on five classic multiple-instance learning (MIL) benchmarks: *MUSK1* Chapman & Jain (1994a) and *MUSK2* Chapman & Jain (1994b), where each bag is a molecule and instances are its low-energy 3D conformations described by 166 attributes (a bag is positive if at least one conformation is active); and the image MIL datasets *TIGER*, *FOX*, and *ELEPHANT* Andrews et al. (2002), where each bag is a Corel image segmented

into "Blobworld" regions (instances) with 230-D color/texture/shape features (a bag is positive if at least one segment contains the named animal). Standard size statistics are reported in Table 17.

Table 18: Results on the small MIL benchmark datasets.

Methods	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
ABMIL ICML 2018	$0.916_{\pm 0.118}$	$0.928_{\pm 0.109}$	$0.952_{\pm 0.051}$	$0.953_{\pm 0.042}$	$0.969_{\pm 0.036}$
DSMIL CVPR 2021b	$0.959_{\pm 0.053}$	$0.952_{\pm 0.066}$	$0.939_{\pm 0.060}$	$0.951_{\pm 0.053}$	$\mathbf{0.989_{\pm 0.023}}$
TransMIL NeurIPS 2021	$0.927_{\pm 0.093}$	$0.877_{\pm 0.127}$	$0.944_{\pm 0.050}$	$0.963_{\pm 0.042}$	$0.979_{\pm 0.030}$
DEMIL NeurIPS 2023a	$0.963_{\pm 0.073}$	$0.961_{\pm 0.057}$	$0.941_{\pm 0.047}$	$0.965_{\pm 0.035}$	$0.969_{\pm 0.034}$
RGMIL Neurips2023	$0.968_{\pm 0.060}$	$0.963_{\pm 0.048}$	$0.954_{\pm 0.048}$	$0.949_{\pm 0.047}$	$0.965_{\pm 0.032}$
PSMIL ICLR2025	$0.968_{\pm 0.053}$	$\mathbf{0.968_{\pm 0.052}}$	$0.942_{\pm 0.054}$	$0.947_{\pm 0.047}$	$0.985_{\pm 0.030}$
ASMIL (Ours)	$\mathbf{0.971_{\pm 0.060}}$	$0.968_{\pm 0.058}$	$\mathbf{0.961_{\pm 0.025}}$	$\mathbf{0.969_{\pm 0.037}}$	$0.985_{\pm 0.025}$

Since these datasets are relatively balanced, following Du et al. (2025), we report accuracy as the primary metric. We train for 40 epochs with the Adam optimizer (Kingma & Ba, 2014) and a learning rate of 0.0005.

We compare ASMIL against six MIL methods—ABMIL (Ilse et al., 2018), DSMIL (Li et al., 2021b), TransMIL (Shao et al., 2021), DEMIL (Tang et al., 2023a), RGMIL (Du et al., 2023), and PSMIL (Du et al., 2025)—and report accuracies in Table 18. As shown, ASMIL outperforms all baselines on 4 of 5 datasets, demonstrating strong performance on non-WSI benchmarks.

P Attention Dynamics of different MIL methods on Various Datasets

In this section, we illustrate that the issue of attention convergence on the WSI dataset is not unique to the ABMIL and CAMELYON-16 datasets. To this end, similar to the method we describe in Figure 1, we plot the JSD of two attention distributions between two consecutive epochs.

P.1 CAMELYON-16 DATASET

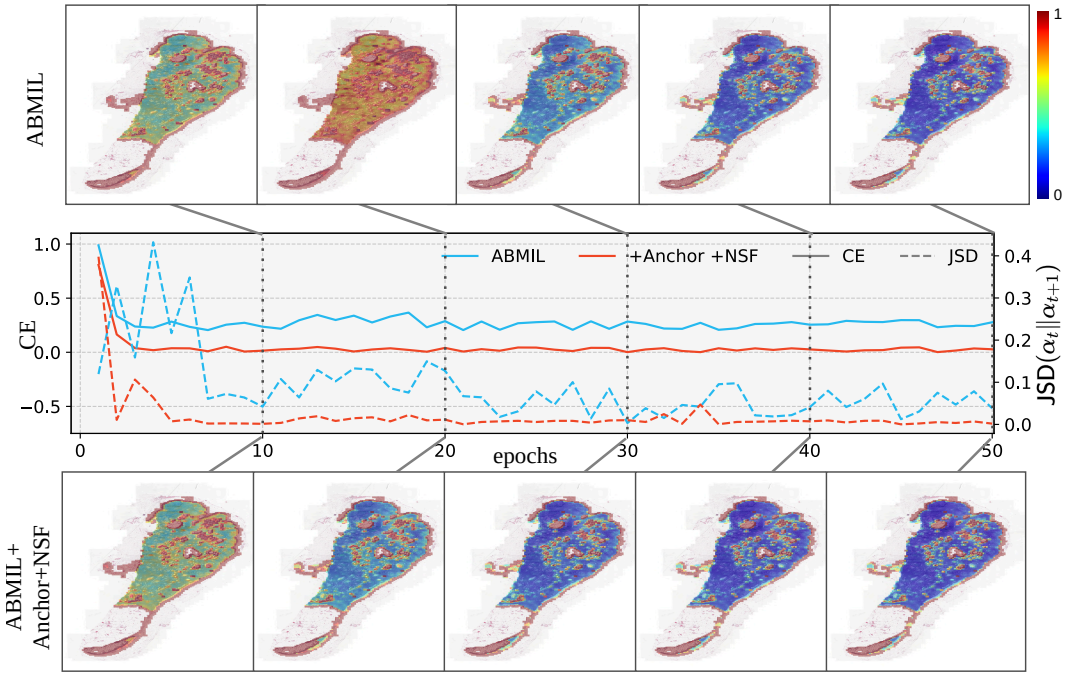

Figure 14: Visualization of attention dynamics on a normal WSI for ABMIL vs. ABMIL + anchor + NSF.

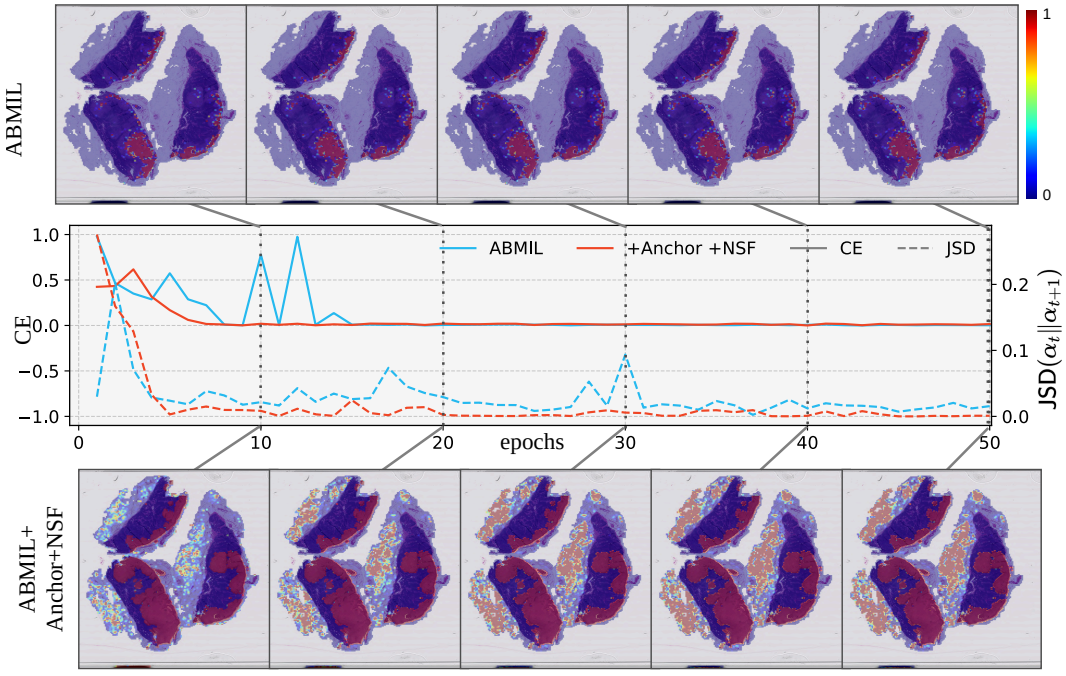

Figure 15: Visualization of attention dynamics on a tumor WSI for ABMIL vs. ABMIL + anchor + NSF.

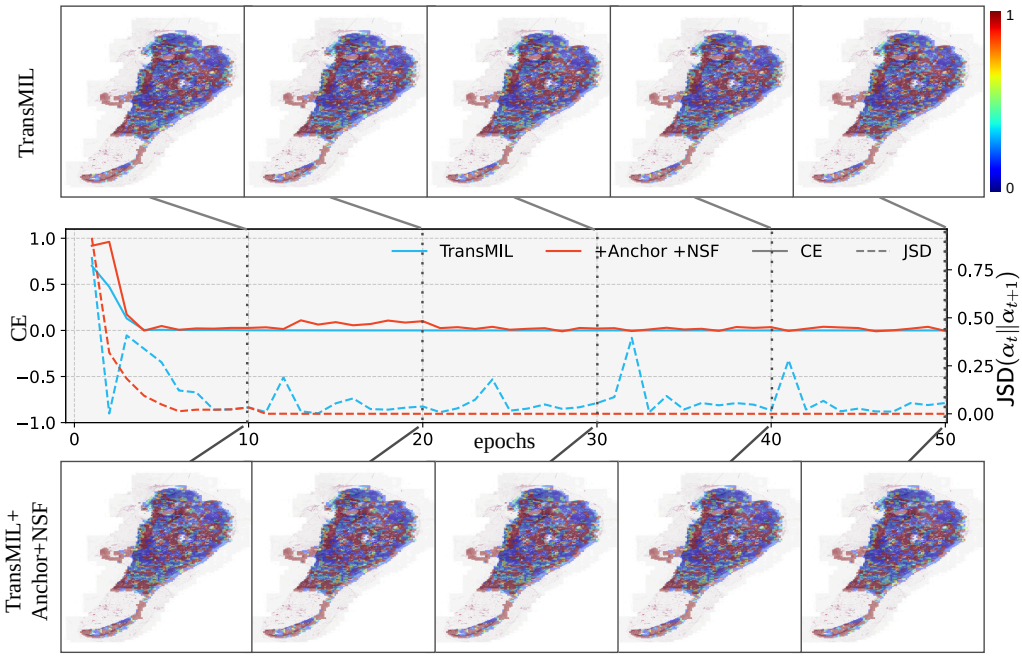

Figure 16: Visualization of attention dynamics on a normal WSI for TransMIL vs. TransMIL + anchor + NSF.

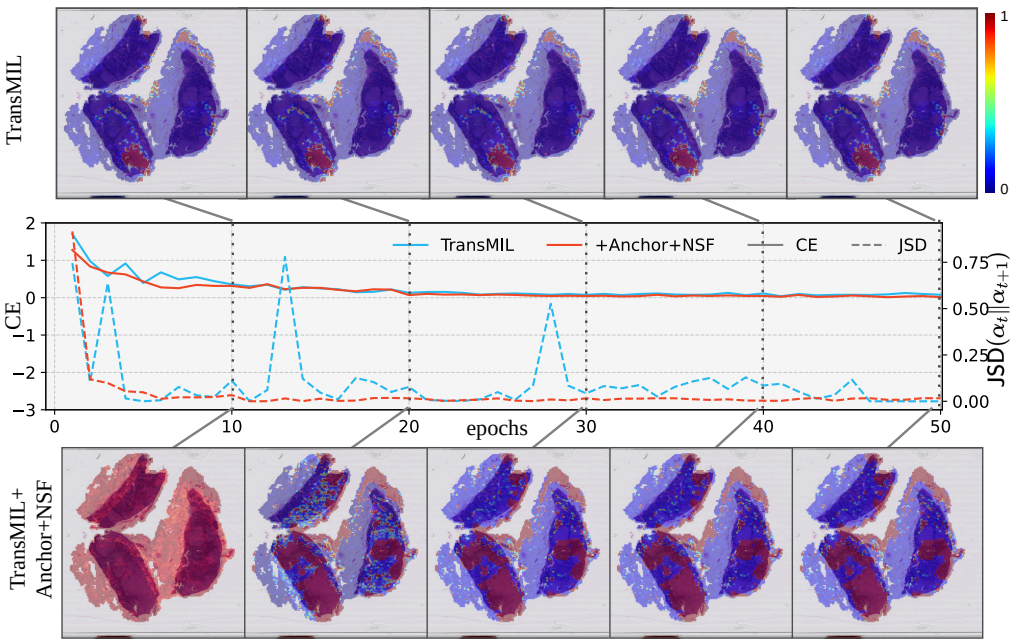

Figure 17: Visualization of attention dynamics on a tumor WSI for TransMIL vs. TransMIL + anchor + NSF.

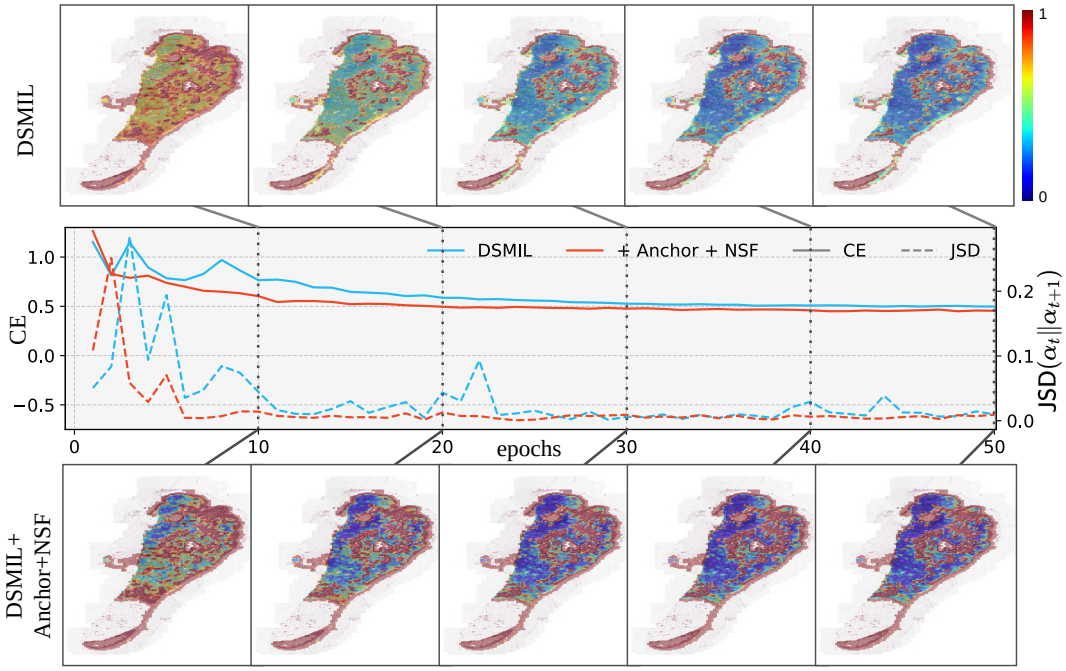

Figure 18: Visualization of attention dynamics on a normal WSI for DSMIL vs. DSMIL + anchor + NSF.

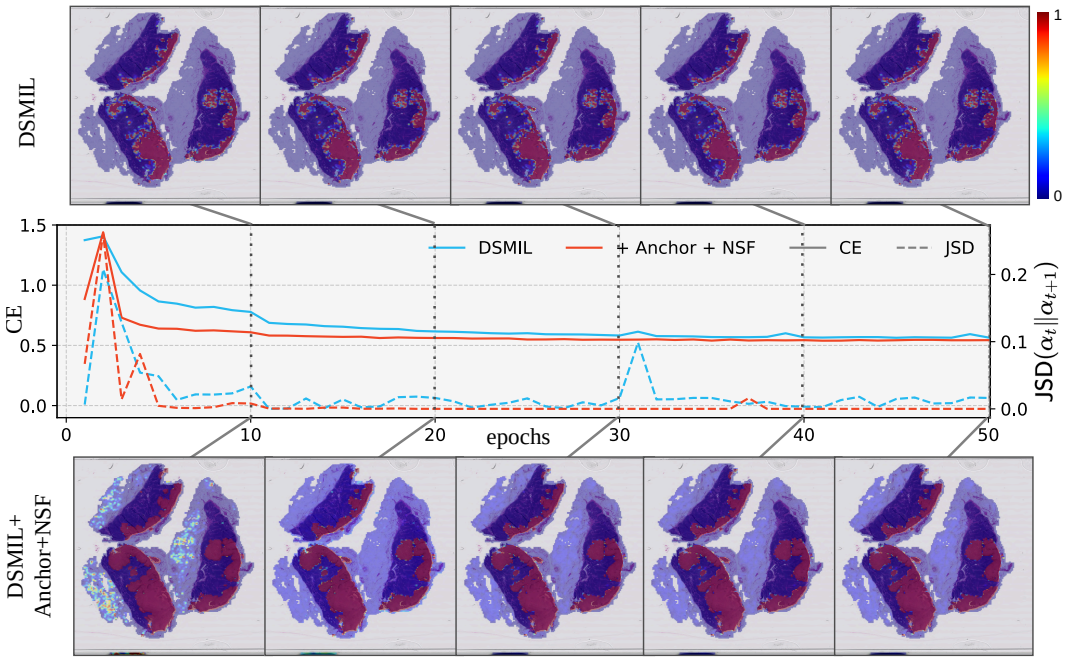

Figure 19: Visualization of attention dynamics on a tumor WSI for DSMIL vs. DSMIL + anchor + NSF.

P.2 BRACS DATASET

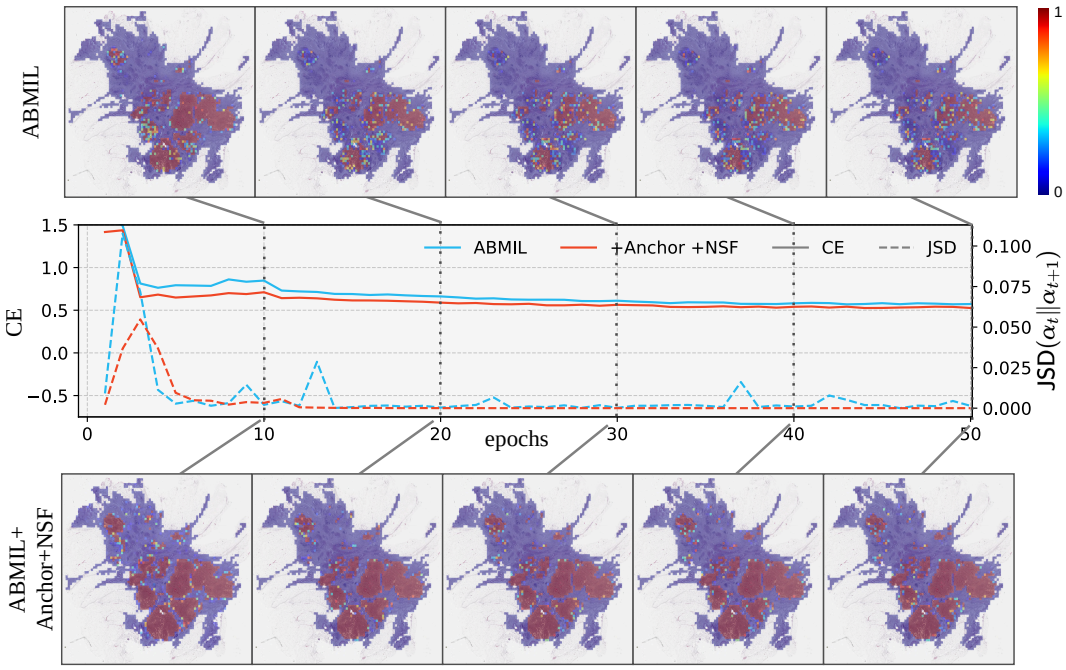

Figure 20: Visualization of attention dynamics on a normal WSI for ABMIL vs. ABMIL + anchor + NSF.

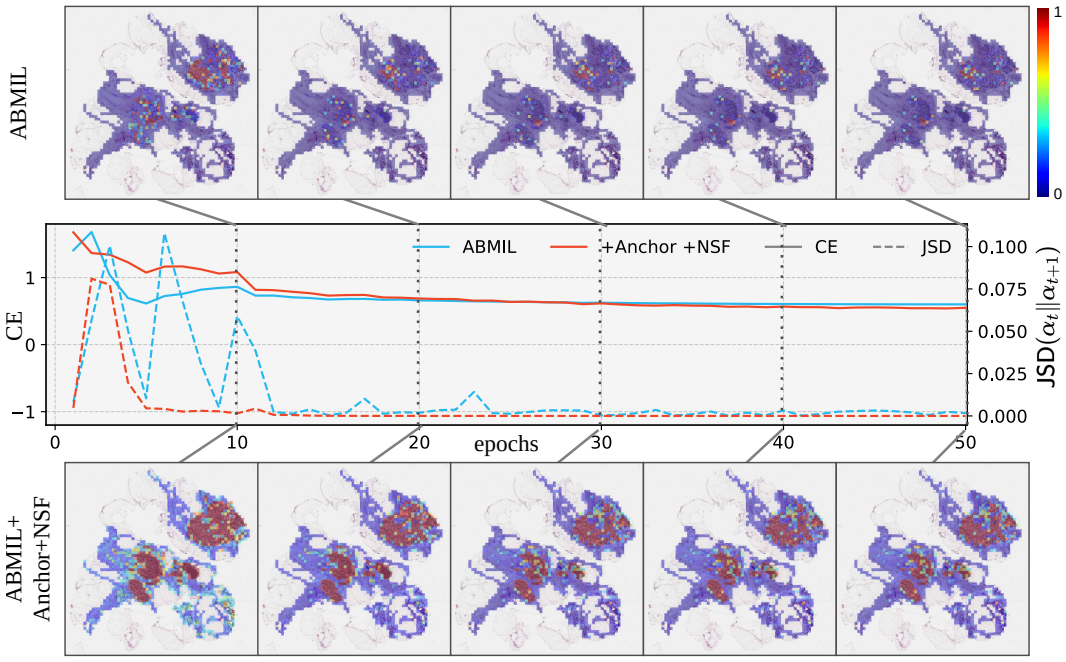

Figure 21: Visualization of attention dynamics on a tumor WSI for ABMIL vs. ABMIL + anchor + NSF.

Despite these advances, several avenues remain open for future investigation:

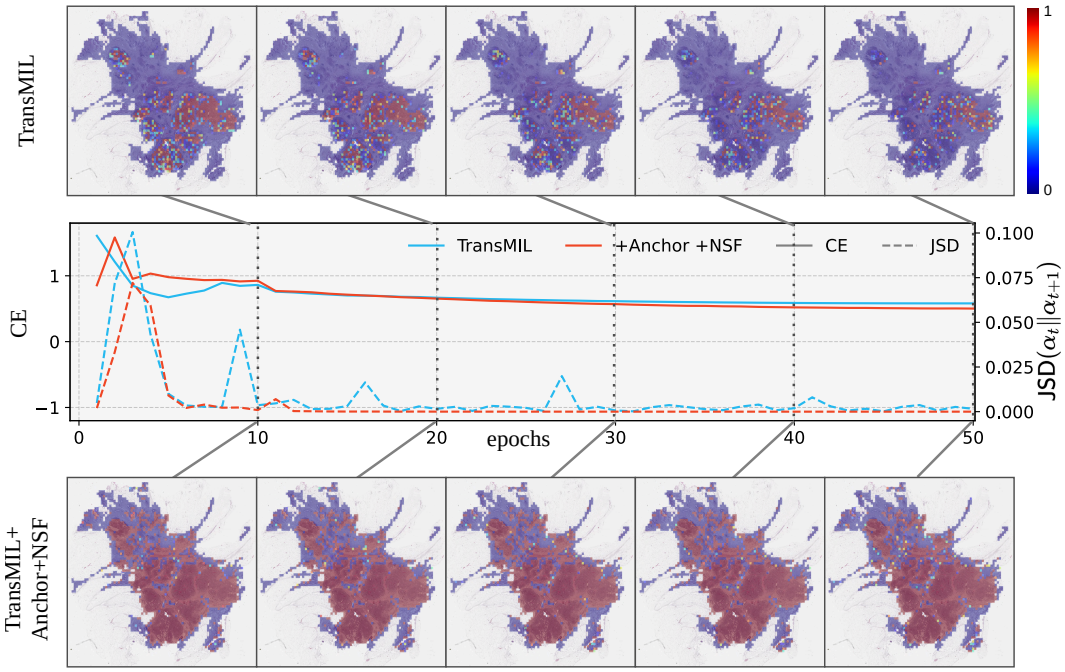

Figure 22: Visualization of attention dynamics on a tumor WSI for TransMIL vs. TransMIL + anchor + NSF.

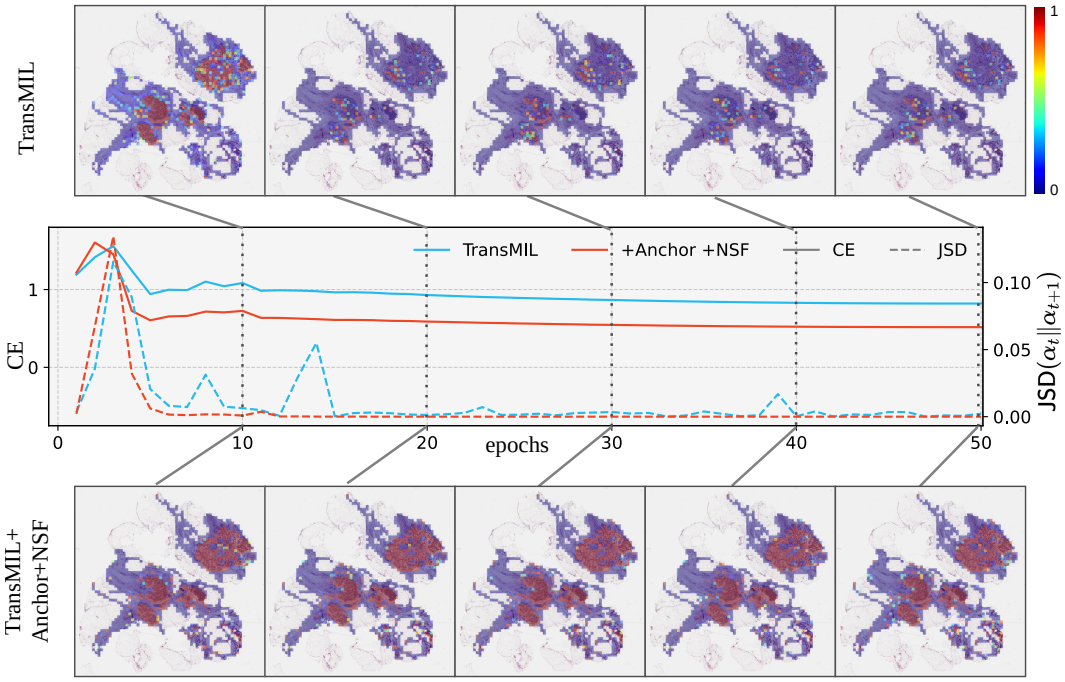

Figure 23: Visualization of attention dynamics on a tumor WSI for TransMIL vs. TransMIL + anchor + NSF.

ASMIL employs an EMA-updated anchor model to stabilize attention dynamics, but this introduces additional computational overhead. An important direction is the development of intrinsic train-

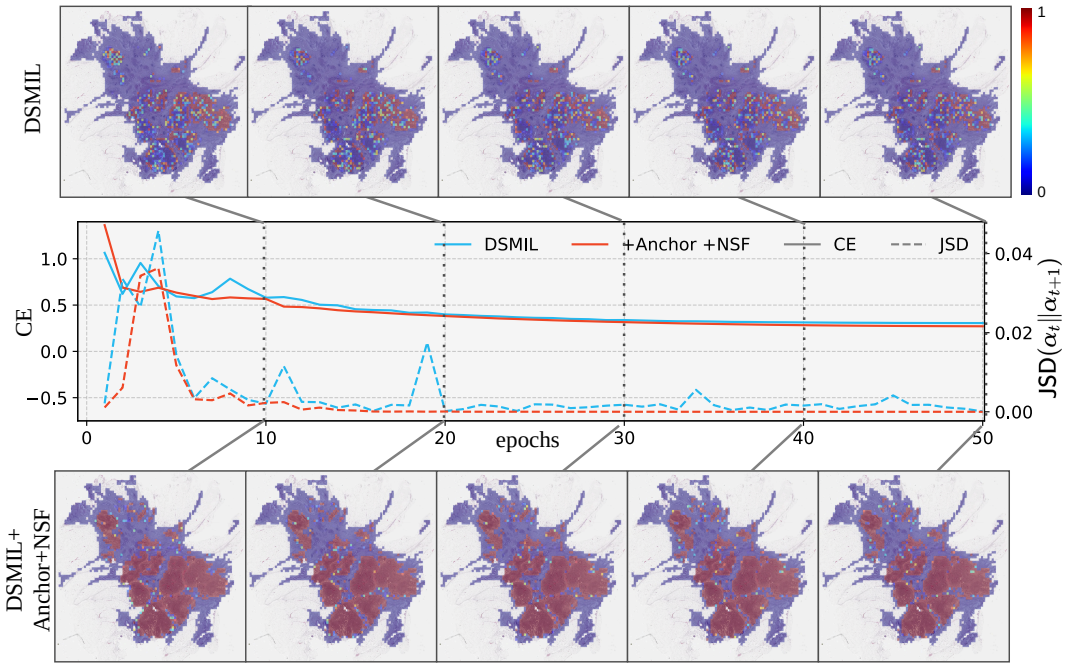

Figure 24: Visualization of attention dynamics on a tumor WSI for DSMIL vs. DSMIL + anchor + NSF.

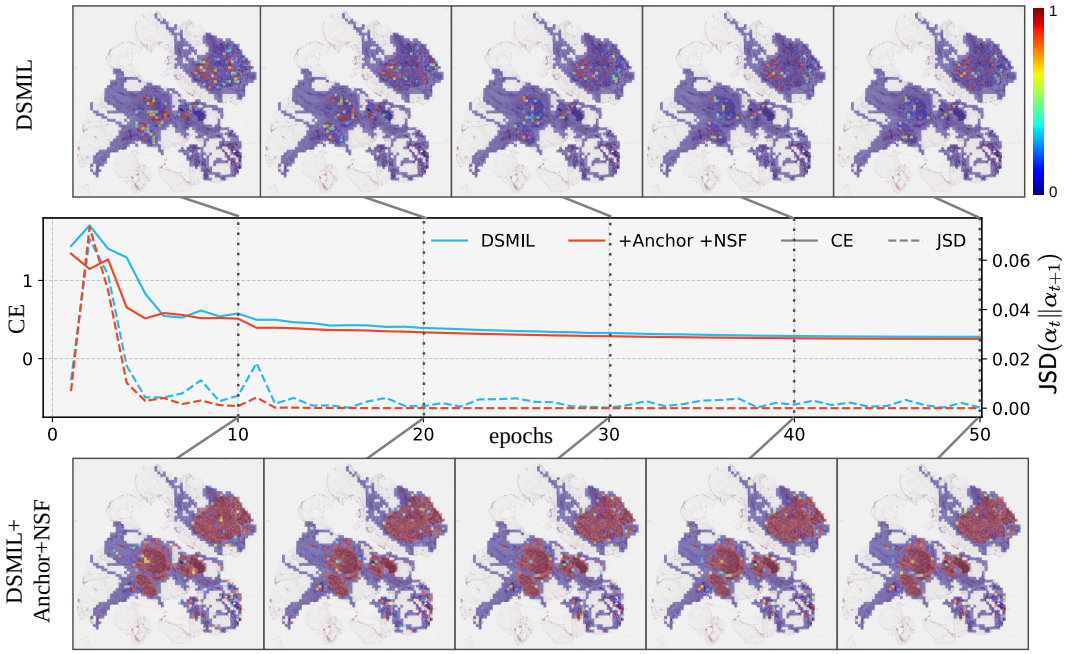

Figure 25: Visualization of attention dynamics on a tumor WSI for DSMIL vs. DSMIL + anchor + NSF.

ing strategies, such as regularization, that achieve comparable stability without auxiliary modules, thereby improving efficiency in large-scale WSI applications.

Table 19: Rate of cancerous WSIs without missed regions on the CAMELYON-16 dataset.

Method	Clam	TransMIL	DTFD-MIL	DSMIL	CAMIL	ASMIL
Rate	46.93%	3.246%	50.34%	48.22%	49.78%	54.63%

Q Limitations and Future Work

Meanwhile, as more advanced regularization techniques such as MIL-dropout (Zhu et al., 2025) continue to emerge, integrating them into the ASMIL framework represents a highly promising direction for future work. Such enhancements could further improve the model's generalization ability while yielding more faithful and stable attention distributions.

Furthermore, a limitation of our approach is that ASMIL can fail by assigning low attention to tiny foci and small tumor regions (see Figure 26), particularly when large and small cancerous regions coexist within a single WSI. We fix the attention threshold at 0.5 and count a cancerous WSI as successfully localized if all regions inside the tumor annotation exceed this threshold; the success rates are reported in Table 19. ASMIL achieves the highest success rate, which we attribute to the NSF in the anchor model that mitigates over-concentrated attention. This indicates room for improvement. Nevertheless, compared with published baselines, ASMIL's attention maps consistently achieve higher Dice and FROC scores. One avenue to further enhance localization performance is to bootstrap training with a mixture of synthetic data and real WSI data. These directions are beyond the scope of this work and will be investigated in future research Wu et al. (2024); Chi et al. (2024).

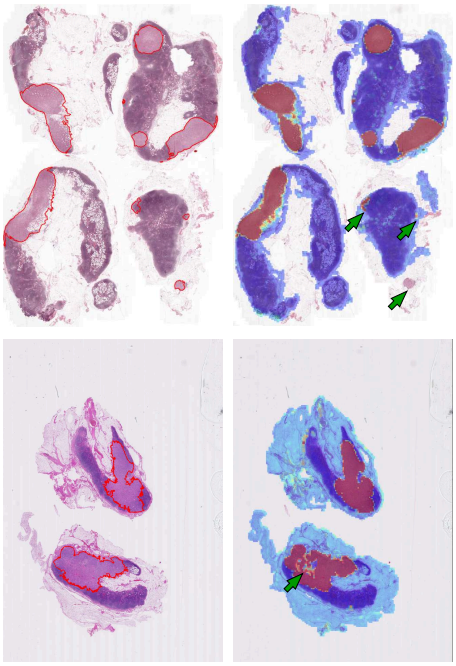

Figure 26: Left: annotated WSI. Right: attention map generated by ASMIL, which fails to assign high attention to all tumor regions, as highlighted by the green arrow.

R LLM Usage Statement

LLM used only for grammar and wording edits; no generation of ideas, methods, analyses, results, or citations. The authors reviewed all edits and accept full responsibility.