Counterfactual Intervention in Attention Multiple Instance Learning For Digital Pathology

Imane Chraki[1,2]   IMANE.CHRAKI@CENTRALESUPELEC.FR
Pierre Marza[1,2]   PIERRE.MARZA@CENTRALESUPELEC.FR
Stergios Christodoulidis[1,2]   STERGIOS.CHRISTODOULIDIS@CENTRALESUPELEC.FR
Maria Vakalopoulou[1,2]   MARIA.VAKALOPOULOU@CENTRALESUPELEC.FR

[1] *MICS Laboratory, CentraleSupélec, Université Paris-Saclay*

[2] *IHU PRISM, National Center for Precision Medicine in Oncology, Gustave Roussy*

Editors: Under Review for MIDL 2026

Abstract

Attention-based Multiple Instance Learning (MIL) has become a prominent framework for analysing whole-slide images (WSI). These models have been shown to achieve good performance on classification tasks, while also offering an inherent proxy for interpretability through attention weights. In this work, we first question the validity of using attention for the interpretability of MIL models. Subsequently, we propose Counterfactual Intervention in Attention for MIL (CIA-MIL), a causal extension of attention-based MIL that explicitly measures and optimizes the contribution of attention to slide-level predictions. Across four histopathology classification benchmarks (BRCA, NSCLC, LUAD, Camelyon16) and two feature encoders (Resnet50, UNI), we investigate how the interpretability of attention relates to the representation space, and the downstream performance. We then show that CIA-MIL achieves performance comparable to strong MIL baselines while providing a more causally meaningful attention vector. Qualitative perturbation experiments show that dropping the top-attended patches leads to a larger confidence degradation in CIA-MIL compared to baseline ABMIL, highlighting the potential of causal supervision for reliable and interpretable WSI-based prediction.

Keywords: Multiple Instance Learning, Attention, Interpretability, Whole Slide Images, Digital Pathology, Causal Intervention.

1. Introduction

Multiple Instance Learning (MIL) is a weakly supervised learning framework designed for scenarios in which data are organized into sets of instances, referred to as bags, while supervision labels are only provided at the bag level (Dietterich et al., 1997; Maron and Lozano-Pérez, 1997). MIL is a relevant framework in many domains where fine-grained annotations are scarce or expensive, such as digital histopathology (Amores, 2013; Wang et al., 2016; Diao et al., 2021; Campanella et al., 2019), where labels are typically provided at the level of Whole Slide Images (WSIs), which can reach extremely large resolutions (often as large as 100k pixels per side)(Lu et al., 2021b).

In the standard MIL assumption, a bag is labeled positive if at least one of its instances is positive, and negative otherwise. While this assumption is appropriate in certain detection tasks, it is often restrictive for real-world biomedical applications. In practice, slide-level

labels usually arise from complex interactions between multiple tissue regions rather than from a single discriminative patch. To address this limitation, numerous attention-based MIL approaches have been proposed, enabling models to learn instance-level importance directly from data (Ilse et al., 2018; Shao et al., 2021; Li et al., 2021). These methods replace strategies such as max or mean pooling with learnable attention mechanisms that assign relevance scores to each instance, referred to as attention.

Despite the remarkable success of attention-based MIL in digital pathology, important questions remain regarding the reliability of attention as an interpretability tool. Attention weights are often visualized as heatmaps and interpreted as indicators of model reasoning (Lu et al., 2021a; Wagner et al., 2023). However, recent work on interpretability has shown that attention may not reliably reflect the true importance of instances for the final prediction (Hense et al., 2024; Zhang et al., 2022; Early et al., 2023; Javed et al., 2022). In some cases, models can achieve similar predictions while relying on substantially different attention distributions, revealing a potential disparity between attention and decision-making. This limitation is particularly critical in medical applications, where interpretability is not merely a convenience but a prerequisite for trust, clinical adoption, and regulatory approval.

In this work, we investigate the reliability of raw attention scores as a proxy for interpretability in MIL models. We study a broad family of attention-based MIL architectures within a unified experimental framework and quantitatively and qualitatively report how different design choices—such as attention type, multihead formulations, clustering strategies, and entropy regularization affect the attention effect on predictions. In a second part, we explore the use of causal counterfactual intervention to guide the learning of attention toward representations that are more causally aligned with the model's predictions. By enforcing counterfactual consistency during training, our goal is to promote attention patterns that reflect more accurately the underlying causal mechanisms. This introduces an explicit trade-off between predictive performance and interpretability, which we characterize empirically.

The main contributions of this work can be summarized as follows: *(i) attention reliability analysis*, we perform an extensive evaluation of attention reliability across a wide range of MIL models, including standard attention MIL and its variants such as clustering-based MIL, multi-head attention MIL, and DSMIL. Our analysis reveals that the attention byproduct of current state-of-the-art MILs does not fully align causally with downstream prediction. *(ii) novel counterfactual attention intervention framework (CIA-MIL)*, we propose, according to our knowledge, for the first time a novel counterfactual-guided attention learning strategy designed to improve the causal alignment and stability of attention mechanisms in MIL models. *(iii) analysis of the interplay between downstream performance and attention interpretability*, we conduct extensive experiments on real-world digital pathology datasets, demonstrating the complex trade-offs between predictive performance and interpretability of attention.

2. Related Work

Attention-Based MIL — Embedding-level MIL models operate directly in the instance embedding space to compute a compact bag-level representation that is subsequently passed to a classifier. In the standard attention-based MIL formulation (Ilse et al., 2018), the bag

representation is expressed as a weighted sum of instance embeddings, where the attention weights are learned functions of the instances themselves. Building upon this framework, numerous variants have been proposed to improve representational capacity and performance. TransMIL (Shao et al., 2021) introduces self-attention (Wagner et al., 2023; Xiong et al., 2021), while CLAM (Lu et al., 2021b) integrates clustering-based attention to capture multiple discriminative regions within a bag. DSMIL (Li et al., 2021) adopts a dual-stream architecture to explicitly model instance-level and bag-level interactions. Other extensions leverage multi-head attention and multi-branch aggregation strategies. More recent models aim to regularize attention, reducing over-reliance on a few highly activated instances. For instance, MHIM-MIL (Tang et al., 2023) adopts a Siamese framework with masked attention to mine hard-to-classify instances, while ACMIL (Zhang et al., 2024) introduces multi-branch attention together with stochastic top-K instance masking to promote diversity in discriminative patterns. To further improve learning under limited supervision, data distillation and pseudo-bag generation strategies such as DTFD-MIL (Zhang et al., 2022) have also been proposed.

Interpretability of MIL Models — In digital histopathology, attention-based MIL models typically rely on attention scores to generate patch-level relevance maps that highlight regions of interest. Despite their intuitive appeal, several studies have shown that raw attention maps do not necessarily provide faithful explanations of model behavior (Hense et al., 2024; Javed et al., 2022; Zhang et al., 2022). To address this limitation, alternative explainability strategies have been proposed. DTFD (Zhang et al., 2022) reframes MIL as an equivalent image classification problem and derives patch-level importance scores using Grad-CAM (Selvaraju et al., 2017). Other post-hoc explainability strategies have also been proposed (Adebayo et al., 2018; Kindermans et al., 2019; Pirovano et al., 2020; Wang et al., 2019; Bach et al., 2015; Baehrens et al., 2010; Shrikumar et al., 2017; Montavon et al., 2019; Hense et al., 2024), including perturbation-based approaches (Early et al., 2023), while fully additive models explicitly decompose the bag-level prediction into a sum of instance contributions (Javed et al., 2022). Despite these advances, many existing interpretability methods suffer from high computational cost, limited scalability to large bags, or simplifying assumptions that neglect complex inter-instance dependencies. Developing MIL models that are simultaneously accurate, scalable, and faithfully interpretable remains an open challenge.

Causal Inference in MIL — Recently, causality (Pearl and Mackenzie, 2018; Pearl et al., 2016) has been introduced into the MIL framework to account for confounding factors that may compromise model training. Models such as CAMIL (Chen et al., 2024a) and CATTMIL (Wu et al., 2024) formulate the bag-level representation as a mediator between patch embeddings and the final prediction by applying a front-door adjustment (Pearl et al., 2016). In contrast, (Lin et al., 2023) adopts a back-door adjustment (Pearl et al., 2016) strategy via a two-stage training procedure to explicitly control for co-founders. These approaches highlight the growing interest in causal reasoning within MIL pipelines, particularly for medical imaging applications, where biased signals affecting images might degrade prediction reliability. In our work, instead of reasoning in the image space, we operate directly at the attention level. We climb the causality ladder higher than the adjustment level and apply counterfactual intervention in attention.

Counterfactual Intervention for Attention Learning — Counterfactual analysis provides a principled framework to measure the causal influence of input features on model

predictions. In computer vision, counterfactual attention learning has been introduced to guide attention mechanisms through causal supervision rather than relying solely on conventional likelihood maximization. Notably, (Rao et al., 2021) proposes a counterfactual attention learning framework that explicitly maximizes the prediction difference between factual and counterfactual attentions to encourage the discovery of causally effective visual regions. While these ideas have demonstrated strong performance in fine-grained recognition and re-identification tasks, their integration into multiple instance learning for digital pathology remains largely unexplored. Our work bridges this gap by formulating counterfactual causal supervision directly at the MIL aggregation level.

3. Methods

3.1. Attention-Based Multiple Instance Learning

We adopt the attention-based MIL formulation of (Ilse et al., 2018). A bag $X = \{x_i\}_{i=1}^{N}$ consists of N instances, each mapped to a feature embedding by a shared frozen encoder:

$$\mathbf{z}_i = f(x_i), \qquad i = 1, \ldots, N. \tag{1}$$

The instance embeddings are aggregated into a bag-level representation using a learnable attention-weighted pooling, $\mathbf{Z} = \sum_{i=1}^{N} a_i \mathbf{z}_i$ which is passed to a classifier to obtain the final bag prediction $Y = \varphi(\mathbf{Z})$.

3.2. Counterfactual Attention Intervention

To explicitly model the causal contribution of attention to prediction, we model an attention-based MIL framework through a structural causal model (SCM) graph (Pearl et al., 2016) as shown in Fig.1, where X: WSI (bag), Y: bag label, A: attention distribution. $X \to A$ indicates that attention is generated from X, $A \to Y$ indicates that attention leads to a bag level prediction, and $X \to Y$ indicates that bag instances lead to a bag level prediction. Let $X = \{z_i\}_{i=1}^{N}$ denote the instance embeddings and $A = \{a_i\}_{i=1}^{N}$ the learned attention distribution. The standard prediction is:

$$Y(A, X) = \varphi\left(\sum_{i=1}^{N} a_i z_i\right) = \varphi(\mathbf{Z}) \tag{2}$$

We introduce a counterfactual intervention in attention by cutting the path from X to A and measuring the effect of this intervention on the prediction as shown in Fig.1(b).

$$Y(\text{do}(A = \bar{A}), X) = \varphi\left(\sum_{i=1}^{N} \bar{a}_i z_i\right) = \bar{Y} \tag{3}$$

where the do-operation $\text{do}(\cdot)$ forcibly assigns a specific non-informative value to A, referred to as $\bar{A}$, while keeping X fixed. $\bar{A} \sim \gamma$ may be sampled as random or uniform attention. This way, the attention effect is defined as:

$$Y_{\text{effect}} = \mathbb{E}_{\bar{A} \sim \gamma}\left[Y(A, X) - Y(\text{do}(A = \bar{A}), X)\right] \tag{4}$$
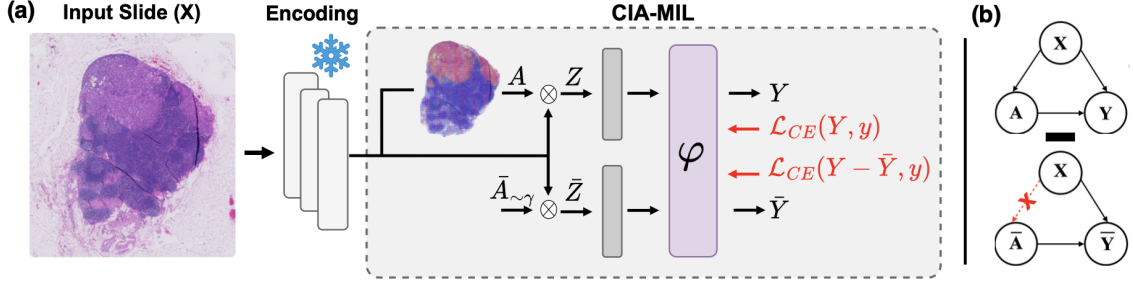
Figure 1: **Overview of CIA-MIL and its causal intervention module.** (a) CIA-MIL: Proposed model architecture. (b) Structural causal graph of the counterfactual attention intervention.

In practice, according to the results in (Baldi and Sadowski, 2014), the expectation is approximated by sampling a single counterfactual attention per bag, resulting in negligible training overhead and no additional inference cost. To guide the learning of causally effective attention, we add a cross-entropy loss between the effect caused by the counterfactual intervention and the ground-truth bag label y. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(Y(A, X), y) + \lambda \, \mathcal{L}_{\text{CE}}(Y_{\text{effect}}, y) \qquad (5)$$

where λ is a hyperparameter, and it controls the influence of the counterfactual supervision. The additional causal intervention loss term to the standard classification cross-entropy loss guides the model to not only learn the classification task, but also to ensure attention patterns that are effect causing for the model: a well-learned attention A should lead to a correct prediction, while a random non-informative attention $\bar{A}$ should fail in leading to the same level of prediction accuracy.

3.3. Attention-Based Perturbation Analysis

We assess the faithfulness of MIL attention as an interpretability proxy using a region perturbation strategy (Hense et al., 2024; Alber et al., 2019). Patches are ranked by decreasing attention scores and partitioned into 100 disjoint subsets $\{r_1, \ldots, r_{100}\}$ of equal size. The perturbed slide at step $k < 100$ is defined as:

$$X^{(k)} = \bigcup_{i=k+1}^{100} r_i, \qquad k = 0, \ldots, 99, \qquad (6)$$

The model's prediction is evaluated at each perturbation step to obtain a perturbation curve. For every slide X, we start first with prediction from the original bag with all patches $X^{(0)} = X$, and start removing subsequently at each following step k, subset r_k, to get the model's prediction $s(k)$ for the perturbed slide $X^{(k)}$. To evaluate the attention influence on prediction, we calculate the Area Under the Perturbation Curve (AUPC). A lower AUPC indicates a faster degradation of predictive confidence when highly attended regions are removed, reflecting an attention mechanism valid for interpreting the model's decision.

4. Experimental Setup

Datasets — We conduct our experiments on whole-slide images (WSIs) from multiple publicly available datasets from TCGA (Tomczak et al., 2015): TCGA-NSCLC (Non–Small Cell Lung Cancer) and TCGA-BRCA (Breast Invasive Carcinoma) for the binary subtyping task, TCGA-LUAD (Lung Adenocarcinoma) for TP53 mutation prediction, Camelyon16 (Bejnordi et al., 2017) for binary metastasis detection in breast lymph node.

Implementation Details — Slides were processed using the same approach as in (Lu et al., 2021b) to obtain patches of size 256×256 at $20\times$ magnification. We use pre-extracted features with two different image encoders: **(i)** ResNet50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) and **(ii)** UNI-V1 (Chen et al., 2024b) foundation model, to test performance in both in and out of domain pre-training. Please note that any other encoder could be integrated into our method. Experiments were conducted under five cross-validation settings for TCGA-NSCLC, TCGA-BRCA, and TCGA-LUAD using a learning rate of 0.0002. We used the originally published train-test split for Camelyon16 in three runs with a learning rate of 0.0002. All experiments were conducted using the Adam optimizer (Kingma, 2014) with a maximum total of 150 epochs and an early stopping criterion of 50 epochs.

5. Results and Discussion

We evaluate and challenge our method against standard ABMIL (Ilse et al., 2018) and its variants, where additional modules are added to mitigate problems encountered during training relative to attention, through clustering (CLAM-SB (Lu et al., 2021b)), multi-head attention and attention masking (ACMIL (Zhang et al., 2024)), attention stage (AddMIL (Javed et al., 2022)), hard instances mining in a 2-stage training framework (MHIM (Tang et al., 2023))). As well as DSMIL (Li et al., 2021), where attention is calculated in a dual stream manner, but the model still satisfies equations 1 and 2. We compare downstream performance also against instance-based baselines, namely average and maximum pooling strategies (MeanMIL/MaxMIL) that do not contain attention modules. When needed, and if not explained differently, raw attention of models is used to assess their reliability. We report performance metrics in terms of Area Under the Curve (AUC) and F1 score. For Camelyon 16, for which we have access to fine-grained annotations at the level of patches, we report the AUPRC, the area under the precision-recall curve for attention as a prediction of instance labels, using both a sigmoid operator and a min-max normalisation per slide.

TCGA Benchmark — We evaluate all models on BRCA and NSCLC tumor subtyping and on the more challenging LUAD TP53 mutation prediction task (Table 1). Across BRCA and NSCLC, most attention-based MIL models, particularly ABMIL-derived methods, achieve very high AUC, often exceeding 0.95 with in-domain foundation models such as the UNI features. However, this strong predictive performance is accompanied by large variability in attention faithfulness. ABMIL-based models frequently exhibit elevated AUPC, whereas DSMIL shows a more favorable performance-interpretability trade-off, although with sensitivity to the feature encoder. In contrast, CIA-MIL consistently achieves among the lowest AUPC while maintaining competitive or superior AUC across both feature extractors, reflecting an explicit and stable performance–interpretability balance. On LUAD, where overall performance drops across all methods, CIA-MIL still achieves low AUPC with

competitive AUC, indicating that counterfactual supervision remains effective even in low-signal data. An important insight is that when attention-based MIL models already achieve strong downstream performance, introducing the causal intervention primarily serves to guide attention toward more interpretable and causally meaningful patterns. Conversely, in settings where predictive performance is weaker, the intervention can also stabilize and improve the training process itself, leading to more reliable attention and better overall performance trade-off.

Table 1: **TCGA Benchmark: Performance vs Attention Interpretability.** Comparison of classification performance (AUC ↑) and attention faithfulness (AUPC ↓) across three tasks (BRCA subtyping, NSCLC subtyping, LUAD TP53 mutation prediction) and two feature extractors (ResNet50, UNI).

| | BRCA - ResNet50 | | BRCA - UNI | | NSCLC - ResNet50 | | NSCLC - UNI | | LUAD - ResNet50 | | LUAD - UNI | |
	AUC (↑)	AUPC (↓)	AUC (↑)	AUPC (↓)	AUC (↑)	AUPC (↓)	AUC (↑)	AUPC (↓)	AUC (↑)	AUPC (↓)	AUC (↑)	AUPC (↓)
MeanMIL	89.0±3.8	N/A	93.2±2.4	N/A	91.1±3.0	N/A	96.9±1.3	N/A	66.9±6.5	N/A	74.5±6.3	N/A
MaxMIL	86.9±2.6	N/A	**95.4±1.5**	N/A	**94.4±1.6**	N/A	97.5±1.0	N/A	61.4±9.2	N/A	76.0±5.4	N/A
ABMIL	89.2±2.6	73.3±25.3	95.3±1.7	70.7±26.5	93.4±1.7	57.0±41.4	**97.6±1.0**	63.3±40.4	68.2±5.6	**49.9±18.9**	74.0±4.3	59.8±33.7
CLAM	89.4±1.8	72.2±24.3	94.5±2.3	60.9±30.3	94.1±1.9	56.5±37.2	97.9±0.8	70.9±28.8	65.5±6.2	55.9±29.7	74.0±3.0	56.6±36.0
DSMIL	88.4±3.0	69.6±23.0	94.1±1.6	**48.4±24.8**	93.6±2.5	55.5 ± 29.2	97.4±1.1	**23.7±28.2**	67.8±6.3	51.9±14.5	66.9±4.7	**46.6±22.7**
ACMIL	88.6±2.8	76.6±26.3	94.4±2.9	74.0±18.2	93.8±1.8	54.9±46.7	97.9±0.8	73.2±22.3	67.5±5.0	50.8±23.2	75.5±7.4	59.9±35.7
AddMIL	88.0±2.0	78.3±30.9	93.7±2.6	78.7±26.8	92.0±2.3	55.4±46.4	94.6±3.0	62.4±42.2	63.4±3.8	54.2±39.3	73.3±3.9	57.1±38.6
MHIM	90.0±1.9	74.7±24.8	94.2±1.1	71.3±28.1	94.8±1.5	57.8±18.1	97.9±0.9	75.2±25.7	68.0±3.1	52.0±12.6	73.8±3.7	57.0±35.9
CIA-MIL	**90.1±2.8**	**61.2±16.6**	94.5±1.2	56.7±25.2	91.6±2.4	49.4±21.2	96.5±1.6	51.5±23.1	**69.1±5.1**	52.3±11.2	**77.5±2.5**	55.6±27.0
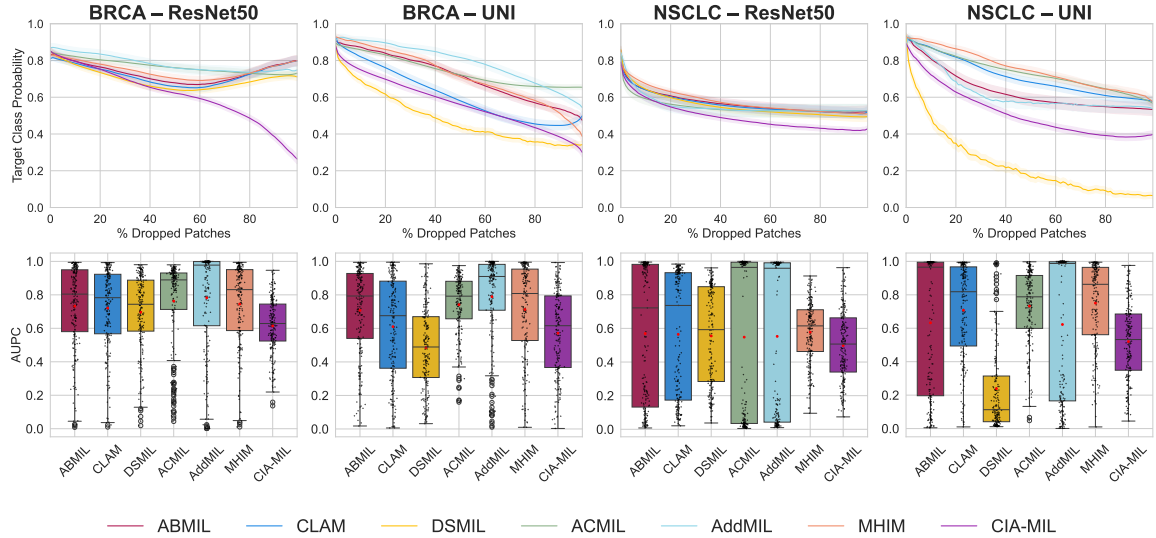
Figure 2: Attention-based perturbation analysis on the BRCA and NSCLC subtyping tasks. The curves show the evolution of the target class probability as increasingly larger fractions of highly attended patches are removed. Boxplots report statistics over slides.

To further analyze the sensitivity of each model to attention-guided perturbations, Fig.2 presents the perturbation curves on BRCA and NSCLC. Models with causally faithful attention exhibit a rapid degradation of prediction confidence when highly attended patches

are removed, whereas models with diffuse attention remain comparatively insensitive. CIA-MIL consistently shows steeper decay profiles, confirming that counterfactual supervision effectively reshapes attention toward causally meaningful tissue regions for the model's prediction.

Camelyon16 Benchmark — We evaluate models on Camelyon16 with UNI features to explicitly disentangle bag-level performance from instance-level interpretability. Table 2 reports bag-level classification performance (AUC, F1) together with instance-level attention interpretability (AUPC) and localization accuracy (AUPRC). AUPRC is computed either by applying a sigmoid to raw patch attentions (Sig), or by min-max normalizing attentions within each slide (MM). Comparing these two variants allows us to distinguish between the global behavior of attention across all slides (Sig) and the relative spatial selectivity within each slide (MM). At the bag level, nearly all attention-based MIL models achieve near-saturated performance, with AUC values consistently above 98%. However, clear differences emerge at the instance level. In particular, DSMIL shows an interesting instance-level behavior, with severely degraded AUPRC under the sigmoid setting, reflecting an attention weighting of instances that tightly depend on the slide. In contrast, CIA-MIL achieves the lowest AUPC while maintaining high AUPRC under both Sig and MM schemes, suggesting that counterfactual intervention in attention yields attention that is simultaneously globally consistent, spatially selective, and causally aligned with outcome without compromising slide-level performance.

Table 2: **Camelyon16: Bag-Level Performance vs Instance-Level Interpretability.**
Comparison of bag-level classification performance (AUC, F1) and instance-level interpretability measured by AUPC and AUPRC. AUPRC (Sig) is computed by applying a sigmoid to raw patch-level attentions, while AUPRC (MM) is obtained by min-max normalizing attentions within each slide.

| | Bag | | Instance | | |
	AUC ($\uparrow$)	**F1** ($\uparrow$)	**AUPC** ($\downarrow$)	$\mathbf{AUPRC}_{sig}(\uparrow)$	$\mathbf{AUPRC}_{MM}(\uparrow)$
Meanmil	62.5 ± 4.8	46.7 ± 9.6	N/A	N/A	N/A
MaxMIL	98.3 ± 0.4	94.2 ± 1.9	N/A	N/A	N/A
ABMIL	98.7 ± 0.3	95.5 ± 2.2	64.8 ± 45.3	93.2 ± 1.2	79.2 ± 5.6
CLAM	$\mathbf{99.7 \pm 0.3}$	97.6 ± 2.1	65.6 ± 44.9	$\underline{94.4 \pm 0.3}$	$\mathbf{92.4 \pm 0.6}$
DSMIL	98.9 ± 1.1	$\mathbf{97.9 \pm 1.1}$	66.2 ± 43.7	23.9 ± 9.8	82.0 ± 3.6
ACMIL	$\underline{99.4 \pm 0.5}$	96.9 ± 2.7	65.5 ± 45.2	$\mathbf{95.4 \pm 0.5}$	$\underline{91.3 \pm 1.6}$
AddMIL	98.2 ± 1.4	95.0 ± 3.4	65.2 ± 45.0	93.4 ± 1.2	82.6 ± 4.5
MHIM	98.2 ± 1.1	94.2 ± 2.8	$\underline{58.8 \pm 32.5}$	93.2 ± 0.5	87.9 ± 3.0
CIA-MIL	99.2 ± 0.5	96.3 ± 2.0	$\mathbf{53.3 \pm 18.6}$	92.9 ± 1.5	82.4 ± 5.4

Qualitative Attention Analysis. Fig.3 provides a qualitative comparison of attention maps on representative crops from a metastatic Camelyon16 slide. The first row shows a metastatic region (red annotations), followed by the corresponding ABMIL, ACMIL and CIA-MIL attention overlays, while the second row displays a non-metastatic region. ABMIL exhibits more diffuse attention and assigns non-negligible attention to non-metastatic tissue. In contrast, ACMIL and CIA-MIL concentrate attention sharply within metastatic regions and suppress responses in non-metastatic areas.

Why Still Use ResNet in the Era of Foundation Models? — Although foundation models such as UNI (Chen et al., 2024b) provide in-domain representations, we include
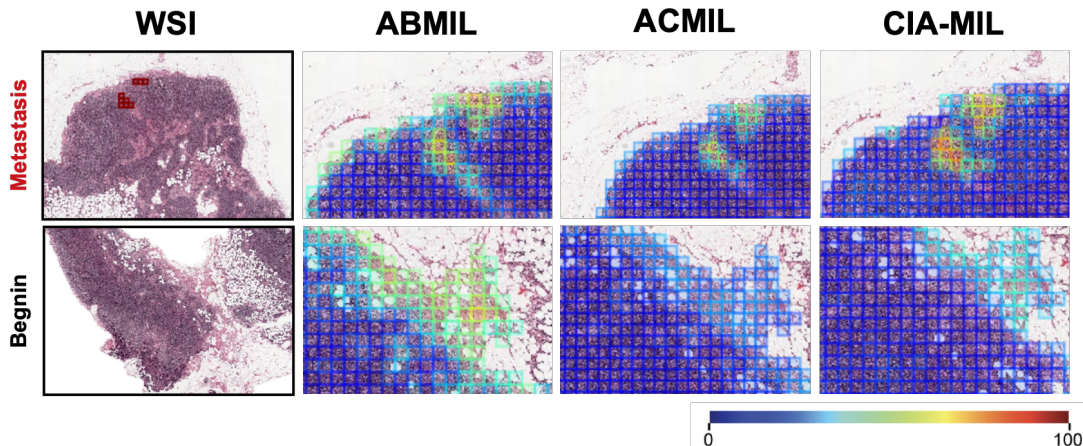
Figure 3: **Qualitative comparison of attention maps on a Camelyon16 WSI.** Attention is normalised with min-max slide within the slide for compared models: ABMIL, ACMIL, and CIA-MIL.

ResNet50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) as a feature extractor. In fact, evaluating both out-of-domain and in-domain features allows us to disentangle the effect of representation learning from that of attention supervision. Also, the fact that CIA-MIL exhibits consistent interpretability improvements across both ResNet50 and UNI demonstrates that the proposed counterfactual intervention is not tied to a specific representation space. Notably, while UNI generally yields higher absolute AUC values, attention faithfulness as measured by AUPC does not automatically improve with stronger foundation features. In some cases, certain configurations of MIL models can achieve the same performance or even slightly degraded with foundation models compared to generic representations, as in the example of DSMIL on the LUAD TP53 mutation prediction as reported in Table 1. This further supports our central claim: improvements in representation power alone do not guarantee causally meaningful attention, and explicit counterfactual supervision of attention can help mitigate this issue.

6. Ablation Study

Random vs Uniform Counterfactual Intervention — We study the impact of the counterfactual attention distribution used for intervention. Table 3 compares random and uniform counterfactual attentions across BRCA and NSCLC. Both strategies lead to consistent improvements in attention faithfulness compared to ABMIL, validating that the proposed causal mechanism is not sensitive to a specific choice of counterfactual distribution. However, subtle differences emerge between the two variants. Random counterfactual attention generally yields slightly lower AUPC, especially on NSCLC with UNI features, indicating stronger selective pressure on attention. Uniform counterfactual attention, by contrast, occasionally preserves marginally higher AUC at the cost of slightly degraded AUPC. This suggests that random intervention introduces stronger stochastic perturbations that better suppress spurious correlations, whereas uniform intervention acts as a weaker regularizer.

Table 3: **Intervention Choice: Random vs Uniform Counterfactual Attention Distribution.** Comparison of CIA-MIL using random and uniform counterfactual attentions across BRCA and NSCLC datasets with both ResNet50 and UNI feature extractors.

| | BRCA - ResNet50 | | BRCA - UNI | | NSCLC - ResNet50 | | NSCLC - UNI | |
	AUC ($\uparrow$)	AUPC ($\downarrow$)	AUC ($\uparrow$)	AUPC ($\downarrow$)	AUC ($\uparrow$)	AUPC ($\downarrow$)	AUC ($\uparrow$)	AUPC ($\downarrow$)
ABMIL (Ilse et al., 2018)	89.2±2.6	73.3±25.3	**95.3±1.7**	70.7±26.5	**93.4±1.7**	57.0±41.4	**97.6±1.0**	63.3±40.4
CIA-MIL (Random)	90.1±2.8	61.2±16.6	94.5±1.2	**56.7±25.2**	91.6±2.4	49.4±21.2	96.5±1.6	**51.5±23.1**
CIA-MIL (Uniform)	**90.2±3.1**	**60.0±17.8**	94.6±2.0	67.5±23.9	91.5±2.4	**48.9±21.3**	97.2±1.9	67.4±28.9

Causal Effect Weighting: Performance-Interpretability Trade-off — We analyse in Fig.4 the influence of the causal effect weighting coefficient λ on the balance between predictive performance and attention interpretability. When $\lambda = 0$, the model reduces to a standard attention-based MIL baseline without causal supervision, resulting in weaker perturbation sensitivity. As λ increases, the drop in AUPC increases on both BRCA and NSCLC, indicating progressively more selective and causally aligned attention. Importantly, AUC remains stable within a narrow range, showing that the improvement in interpretability comes at a moderate and controlled cost in predictive performance.
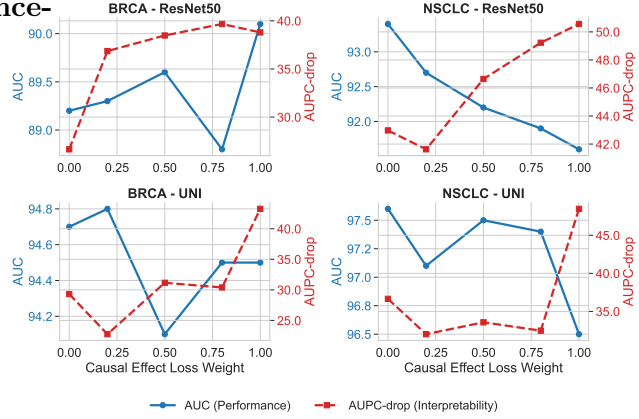

Figure 4: Performance-interpretability trade-off controlled by the causal effect weighting coefficient λ.

7. Conclusion

In this work, we questioned the reliability of attention as an interpretability proxy for attention MIL models, and presented CIA-MIL, a causal attention learning framework for MIL that explicitly enforces the causal contribution of attention to model prediction through counterfactual supervision. Through extensive experiments on tumor subtyping, mutation prediction, and metastasis detection benchmarks, we demonstrated that high predictive performance does not necessarily imply reliable interpretability via attention. CIA-MIL consistently improves attention reliability while maintaining competitive predictive performance across both out-of-domain and in-domain feature spaces. This study establishes counterfactual causal intervention supervision of attention as a promising mechanism to improve the reliability of attention in MIL, taking a step further towards deploying trustworthy and clinically actionable AI systems in digital pathology.

Acknowledgments

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks! *Journal of machine learning research*, 20(93):1–8, 2019.

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210, 2017.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

Kaitao Chen, Shiliang Sun, and Jing Zhao. Camil: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1120–1128, 2024a.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862, 2024b.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

James A Diao, Jason K Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N Mitchell, Benjamin Glass, Sara Hoffman, Sudha K Rao, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature Communications*, 12(1):1–15, 2021.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Joseph Early, Gavin KC Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. *arXiv preprint arXiv:2311.10049*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Julius Hense, Mina Jamshidi Idaji, Oliver Eberle, Thomas Schnake, Jonas Dippel, Laure Ciernik, Oliver Buchstab, Andreas Mock, Frederick Klauschen, and Klaus-Robert Müller. Xmil: Insightful explanations for multiple instance learning in histopathology. *Advances in Neural Information Processing Systems*, 37:8300–8328, 2024.

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

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280. Springer, 2019.

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

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*, pages 14318–14328, 2021.

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*, pages 19830–19839, 2023.

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

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, 2021b.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect.* Basic books, 2018.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer.* John Wiley & Sons, 2016.

Antoine Pirovano, Hippolyte Heuberger, Sylvain Berlemont, Saïd Ladjal, and Isabelle Bloch. Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations. In *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*, pages 43–53. Springer, 2020.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1025–1034, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

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*, pages 4078–4087, 2023.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, Piet A van den Brandt, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer cell*, 41(9):1650–1661, 2023.

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

Xiuying Wang, Dingqian Wang, Zhigang Yao, Bowen Xin, Bao Wang, Chuanjin Lan, Yejun Qin, Shangchen Xu, Dazhong He, and Yingchao Liu. Machine learning models for multiparametric glioma grading with quantitative result interpretations. *Frontiers in neuroscience*, 12:1046, 2019.

Xiaochun Wu, Haitao Wang, and Hejun Wu. Causal attention multiple instance learning for whole slide image classification. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024.

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, pages 14138–14148, 2021.

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

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

Appendix A. Additional details for the WSI Datasets

Table 4 presents in more detail the description of the different datasets used in our study. Details about the number of samples and classes are included, together with a small description of each of the digital pathology whole slide image data. More details can be identified in the original papers.

Table 4: **Overview of Whole-Slide Image (WSI) datasets used in this study.** For each cohort, we report the data source, clinical or diagnostic task, number of slides, and the distribution of labels.

Cohort	Description	# Slides	# Labels
TCGA (Tomczak et al., 2015)			
BRCA	WSIs from TCGA-BRCA. Used for histological subtype classification between invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC).	977	IDC: 779 ILC: 198
NSCLC	WSIs from TCGA-LUAD and TCGA-LUSC. Used for distinguishing between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).	956	LUAD: 478 LUSC: 478
LUAD (TP53)	TCGA-LUAD dataset for TP53 mutation prediction directly from H&E WSIs. Labels correspond to mutation status (mutated vs wild-type).	427	TP53 WT: 199 TP53 Mut: 228
Camelyon 16 Challenge (Bejnordi et al., 2017)			
Camelyon16 (Train)	Lymph node metastasis detection dataset. Training subset of whole-slide images (H&E) annotated for the presence of tumor metastasis.	270	Normal: 159 Tumor: 111
Camelyon16 (Test)	Official test subset from the Camelyon16 challenge.	129	Normal: 80 Tumor: 49

Appendix B. Implementation Details about Attention MIL

We provide additional details here on the attention mechanism used in Attention-Based Multiple Instance Learning (Ilse et al., 2018) used in our work. A bag $B = \{x_i\}_{i=1}^{N}$ consists of N instances, each of which is first transformed into a low-dimensional embedding:

$$\mathbf{z}_i = f(x_i), \qquad i = 1, \ldots, N. \tag{7}$$

These embeddings are aggregated using an attention operator to obtain a bag-level representation:

$$\hat{\mathbf{Z}} = \sum_{i=1}^{N} a_i \mathbf{z}_i. \tag{8}$$

The bag prediction is then obtained by applying a classifier to the aggregated representation:

$$\hat{Y} = \varphi(\hat{\mathbf{Z}}). \tag{9}$$

We adopt a gated attention (GA) mechanism (Dauphin et al., 2017) to produce more expressive attention scores. The unnormalized attention score for instance i is computed as:

$$u_i = \mathbf{w}^\top \left(\tanh(\mathbf{V}_1 \mathbf{z}_i) \odot \sigma(\mathbf{V}_2 \mathbf{z}_i) \right), \tag{10}$$

where

- $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^{L \times M}$ are learnable projection matrices,

- $\mathbf{w} \in \mathbb{R}^L$ is a learnable attention vector,

- $\odot$ denotes element-wise multiplication,

- $\sigma(\cdot)$ denotes the sigmoid nonlinearity.

The normalized attention value is obtained using a softmax over all instances:

$$a_i = \frac{\exp(u_i)}{\sum_{j=1}^{N} \exp(u_j)}. \tag{11}$$

Substituting Eq. 11 into Eq. 8, we obtain the gated attention aggregation:

$$\hat{\mathbf{Z}} = \sum_{i=1}^{N} \frac{\exp(u_i)}{\sum_{j=1}^{N} \exp(u_j)} \mathbf{z}_i. \tag{12}$$