EC-MIL: DISENTANGLING EVIDENCE AND CONTEXT FOR INTRINSIC INSTANCE-LEVEL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern multiple instance learning (MIL) models enhance bag-level prediction via instance interactions, but the same interactions entangle individual contributions and hinder instance-level attribution. This tension is particularly important in biomedicine, where instance-level supervision is scarce yet identifying supporting or opposing instances is essential. To address this, we introduce Evidence-Context Disentangled MIL (EC-MIL), a framework separating instance-local, class-specific evidence from bag-dependent context weighting. EC-MIL expresses each class logit as a weighted sum of instance evidence, preserving backbone context modeling while making individual contributions directly accessible. The product of each instance's evidence and context weight naturally defines Weighted Evidence (WE), an intrinsic, signed, class-specific attribution whose sum exactly recovers the class logit in a single forward pass. Under explicit assumptions, we justify this decomposition and show that excess population bag-level cross-entropy risk controls instance-contribution decomposition error. Across synthetic, whole-slide image, and mutation-bag tasks, EC-MIL maintains competitive predictive performance, while WE performs well in evidence recovery and instance-removal faithfulness at lower attribution cost than post-hoc alternatives. On mutation bags, WE also identifies gene-level signals consistent with established tissue–gene associations. Together, these results demonstrate that expressive context modeling and efficient intrinsic attribution can be achieved within a unified MIL framework. The code can be obtained from: `https://anonymous.4open.science/r/EC-MIL-6449`.

1 INTRODUCTION

Multiple Instance Learning (MIL) is primarily designed to address weakly supervised tasks, where a bag composed of multiple instances is observed but only bag-level labels are accessible (Maron & Lozano-Pérez, 1997). Originally applied to drug discovery (Dietterich et al., 1997), this learning paradigm has since been widely adopted in biomedical problems (Carbonneau et al., 2018). In whole-slide pathology and cancer genomics, for example, slide- or sample-level labels are readily available, whereas exhaustive patch- or mutation-level annotation is difficult to obtain. Beyond accurate bag-level prediction, physicians and researchers also need to understand which instances support or oppose a decision. These applications therefore call for MIL models that combine strong predictive performance with informative instance-level explanations.

To improve predictive performance, modern MIL models increasingly make use of the context of a bag (Bilal et al., 2023). Transformer and low-rank interaction models, for example, capture relationships among instances that cannot be represented by scoring each instance independently (Shao et al., 2021; Xiang & Zhang, 2023). Yet the same interactions make individual contributions difficult to identify directly: as an instance representation incorporates information from its companions, its role in the prediction becomes entangled with the rest of the bag. The challenge is therefore to retain expressive context modeling while keeping instance contributions directly readable.

Post-hoc methods offer a way to explain these models, but their limitations complicate instance-level interpretation. Attention weights indicate which instances are emphasized, yet do not directly provide signed contributions and may not faithfully reflect the model's decision (Jain & Wallace, 2019). Gradient- and perturbation-based methods can provide signed, class-specific attributions, but their results can depend on baseline or perturbation choices, and repeated model evaluations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

become costly for large bags (Sundararajan et al., 2017; Early et al., 2022). Relevance-propagation methods further require architecture-specific rules (Hense et al., 2024). In high-stakes biomedical applications, where explanations must help users understand the basis of a prediction, these limitations strengthen the case for inherently interpretable models (Rudin, 2019). This motivates a MIL architecture in which instance contributions are directly accessible from the prediction computation while contextual interactions are retained.

We therefore adopt a simple design principle: class evidence should be determined locally by each instance, while the bag context should determine how strongly that evidence is used. Based on this, we introduce **Evidence-Context Disentangled MIL (EC-MIL)**, which separates instance-local evidence from context-dependent weighting through two streams. The Evidence Stream independently computes signed, class-specific evidence for each instance, while the Context Stream uses an existing MIL backbone to model interactions and produce the corresponding weights. EC-MIL thus preserves context modeling within the Context Stream while keeping each instance's evidence unchanged across different bag contexts. Combining the two streams expresses each class logit as a weighted sum of instance evidence. Their product naturally defines Weighted Evidence (WE), which is exactly the contribution of an instance to that logit. WE is therefore signed, class-specific, and complete with respect to the class logit, and is directly available during inference.

We further examine why this decomposition is appropriate and how it can be constrained by bag-level supervision. Under explicit assumptions, we justify the multiplicative evidence-context form and show that excess population bag-level cross-entropy risk bounds instance-contribution decomposition error. The key insight is that constraints on contextual compensation limit cancellation among instance-level errors, linking the quality of the decomposition to bag-level prediction. We complement this analysis with experiments on synthetic MIL, whole-slide images, and a newly constructed mutation-bag dataset, testing evidence recovery, instance-removal faithfulness, and biological plausibility alongside predictive performance and attribution cost. Our main contributions are as follows:

- We introduce the EC-MIL framework, which separates instance-local signed evidence from context-dependent weighting while retaining contextual interactions within MIL backbones. This structure yields WE, an intrinsic attribution that is signed, class-specific, complete with respect to each class logit, and directly accessible during inference.
- Under explicit MIL assumptions, we establish sufficient conditions for the multiplicative evidence-context decomposition and show that excess population bag-level cross-entropy risk bounds instance-contribution decomposition error, providing a theoretical link between bag-level supervision and instance-level attribution.
- We introduce MUT[1], a mutation-bag dataset for tissue-of-origin prediction, and evaluate EC-MIL across synthetic MIL, whole-slide images, and MUT. EC-MIL maintains competitive predictive performance while WE supports evidence recovery and instance-removal faithfulness at lower attribution cost than post-hoc alternatives. On MUT, WE identifies gene-level signals consistent with established tissue-gene associations.

2 Related Work

2.1 MIL Architecture

MIL architectures have evolved from fixed pooling to learned aggregation and contextual modeling. Early models use mean or max pooling (Oquab et al., 2015), while ABMIL learns attention weights to emphasize informative instances (Ilse et al., 2018). In computational pathology, CLAM refines representations through instance-level clustering (Lu et al., 2021), and DSMIL combines instance predictions with a bag branch modeling relationships to a critical instance (Li et al., 2021). TransMIL captures global patch correlations with Transformers (Vaswani et al., 2017; Shao et al., 2021), while DTFDMIL distills pseudo-bag features for a second aggregation stage (Zhang et al., 2022). ILRAMIL uses low-rank approximations of global interactions (Xiang & Zhang, 2023), and structured state space models efficiently process long patch sequences (Fillioux et al., 2023). CAMIL models neighboring patches to learn context-aware representations (Fourkioti et al., 2024), while

[1]We construct MUT from cBioPortal (Cerami et al., 2012; Gao et al., 2013) with OncoTree tissue labels (Kundra et al., 2021) (Construction details in Appendix B.6). The dataset will be released after acceptance.

ASMIL improves training stability (Ye et al., 2026). However, as contextual models mix information across instances, individual contributions become harder to read from their representations. EC-MIL addresses this tension by retaining contextual interactions in the weighting stream while computing class evidence locally.

2.2 Instance Attribution in MIL

To interpret MIL predictions, attention weights are often visualized as instance-level heatmaps. Attention rollout aggregates weights across Transformer layers (Abnar & Zuidema, 2020). Neither attention weights nor rollout scores directly specify signed class contributions (Jain & Wallace, 2019). Gradient-based methods provide signed, class-specific attributions. Gradient $\times$ Input (GI) is inexpensive (Pirovano et al., 2020) but susceptible to gradient saturation (Ancona et al., 2017); Integrated Gradients (IG) integrates gradients along a reference-to-input path (Sundararajan et al., 2017), requiring multiple evaluations and a baseline choice (Sturmfels et al., 2020). Perturbation methods measure predictions on modified bags. Single evaluates instances individually, while leave-one-out (LOO) measures the effect of removing each instance (Early et al., 2022). SHAP estimates contributions across instance coalitions (Shapley, 1953; Lundberg & Lee, 2017); exact enumeration is exponential in bag size. Repeated evaluations are costly for large bags, and perturbations may produce out-of-distribution inputs. LRP and DeepLIFT instead propagate signed scores backward through a network using operation-specific rules (Bach et al., 2015; Shrikumar et al., 2017); xMIL adapts LRP to ABMIL and TransMIL (Hense et al., 2024). These trade-offs motivate EC-MIL's intrinsic attribution, which makes signed, class-specific contributions directly available from the prediction computation in a single forward pass.

2.3 Intrinsically Interpretable MIL

These challenges have motivated architectures that expose instance contributions directly. Additive MIL sums signed class contributions computed from backbone representations, including those of TransMIL (Javed et al., 2022). In its attention-based formulation, a nonlinear classifier acts on attention-weighted instance features. MILLET introduces conjunctive pooling for time series, multiplying predictions and attention weights from the same backbone representations (Early et al., 2024). Inspired by this design, EC-MIL computes evidence from each instance before context weighting, with a separate stream modeling context. This keeps class evidence unchanged across bags while context modulates its contribution. Other approaches explain predictions through named features or concepts. SI-MIL uses handcrafted pathological descriptors in an interpretable branch guided by deep MIL (Kapse et al., 2024). Concept MIL maps patches to predefined pathology concepts using a vision-language model and predicts from their contributions (Sun et al., 2025). These approaches ground explanations in domain semantics; EC-MIL instead targets signed instance contributions from existing representations, without requiring a predefined semantic feature or concept set.

3 Evidence-Context Disentangled MIL

3.1 Problem Formulation

Let $B = \{x_i\}_{i=1}^{n}$ be a bag containing n instances, with a bag label $y \in \{1, \ldots, C\}$. In EC-MIL (Figure 1), a shared projection maps each instance independently to a latent representation $h_i = \phi(x_i)$, which serves as input to both streams.

Our design separates the class evidence carried by an instance from the strength with which the bag context uses it. We therefore represent each instance's contribution as the product of its local evidence and a nonnegative context weight, allowing context to modulate its magnitude without reversing its sign. EC-MIL then aggregates these contributions into the pre-softmax class logit:

$$z_c(B) = \sum_{i=1}^{n} \alpha_i\,(B)\, E_c\,(x_i)\,, \qquad \alpha_i(B) \geq 0, \quad \sum_{i=1}^{n} \alpha_i(B) = 1. \tag{1}$$

Here, $E_c\,(x_i)$ is the instance-local, class-specific signed evidence of instance x_i, while the context weights $\alpha_i\,(B)$ depend on the entire bag. These weights are shared across classes, so class specificity is carried by the evidence scores. Applying softmax to the resulting logits gives the bag-level class probabilities. Both streams and the shared projection are trained jointly using bag-level cross-

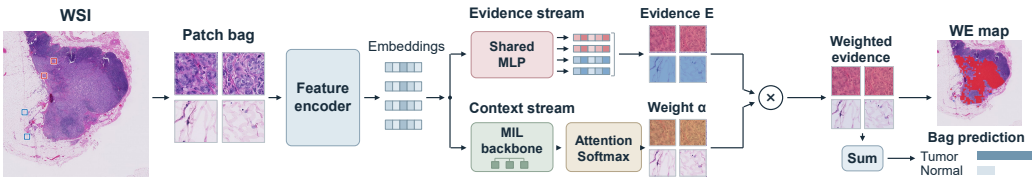


Figure 1: Overview of EC-MIL. Pre-extracted instance embeddings feed two streams: the Evidence Stream produces instance-local, signed class evidence, while the Context Stream computes normalized, class-agnostic weights using a MIL backbone. Their product gives WE, which provides signed, class-specific instance contributions in the same forward pass.

entropy loss. With a linear classification head, standard attention pooling also admits a weighted sum of instance scores (Ilse et al., 2018). MILLET's conjunctive pooling similarly weights instance predictions, with both heads reading the same backbone representations (Early et al., 2024). EC-MIL computes class evidence independently of cross-instance interactions and uses contextual representations only to determine the weights.

3.2 Evidence Stream

The Evidence Stream applies a shared MLP-style network followed by a linear classification head independently to each projected instance h_i, producing a vector of class-specific evidence scores:

$$\mathbf{E}\left(x_i\right) = \left(E_1\left(x_i\right), \ldots, E_C\left(x_i\right)\right) \in \mathbb{R}^C. \tag{2}$$

The linear head produces real-valued scores, allowing each instance to provide supportive or opposing evidence for a class. Because the projection and evidence network process instances independently with shared parameters, a fixed instance produces the same evidence regardless of its bag companions:

$$E_c(x_i; B) = E_c(x_i). \tag{3}$$

3.3 Context Stream

The Context Stream determines how strongly each instance's evidence is used in the current bag. Given the projected instances, a MIL backbone produces a representation for each instance, allowing information from its companions to inform the weighting. An attention MLP head maps these representations to scalar context scores $C_i(B)$, which are normalized across the bag:

$$\alpha_i\left(B\right) = \frac{\exp(C_i\left(B\right))}{\sum_{j=1}^{n} \exp(C_j\left(B\right))}. \tag{4}$$

The resulting weights are positive, sum to one, and are shared across classes. This interface accommodates MIL backbones that expose instance-wise representations before aggregation, supporting different context-modeling mechanisms.

3.4 Weighted Evidence Attribution

Combining the local evidence with its context weight gives the instance-wise contribution in Eq. (1), which we define as Weighted Evidence (WE):

$$\mathrm{WE}_i^c = \alpha_i\left(B\right) E_c\left(x_i\right). \tag{5}$$

WE inherits class specificity from the class evidence $E_c\left(x_i\right)$, while the positive context weights scale its magnitude without reversing its sign. These contributions sum exactly to the class logit, giving WE its completeness property:

$$\sum_{i=1}^{n} \mathrm{WE}_i^c = z_c(B). \tag{6}$$

Both $\alpha_i\left(B\right)$ and $E_c\left(x_i\right)$ are already calculated during prediction, so WE is available in the same forward pass without additional backward passes or model evaluations.

4 THEORETICAL ANALYSIS OF EC-MIL

The EC-MIL design raises two questions: why should instance contributions take the multiplicative form used by WE, and how can bag-level supervision constrain this evidence-context decomposition? We first relate the WE form to bounded contextual effects, then establish conditions under which bag-level prediction risk controls instance-contribution error.

4.1 JUSTIFYING THE WEIGHTED EVIDENCE FORM

To examine the multiplicative form used by WE, consider an instance contribution consisting of a local term and a contextual adjustment, following Deng et al. (2023):

$$c(x, B) = f(x) + g(x, B), \tag{7}$$

where $f(x)$ is the instance-local contribution and $g(x, B)$ is its contextual adjustment.

Assumption A (bounded contextual effects). We assume that the magnitude of the contextual adjustment is bounded by a fraction of the local contribution:

$$|g(x, B)| \le \lambda |f(x)|, \qquad 0 < \lambda \le 1. \tag{8}$$

In this regime, context can strengthen or weaken the contribution without reversing its direction. For $f(x) \neq 0$, setting $s(x) = (1 + \lambda) f(x)$ gives the multiplicative representation

$$c(x, B) = s(x)\beta(x, B), \qquad \beta(x, B) \in \left[\frac{1 - \lambda}{1 + \lambda}, 1\right]. \tag{9}$$

If $f(x) = 0$, the bound gives $g(x, B) = c(x, B) = 0$. This representation motivates the WE form: local evidence determines the direction of contribution, while a nonnegative contextual factor modulates its strength. EC-MIL implements this separation with the additional design choice of normalizing context weights across instances.

4.2 CONSTRAINING THE DECOMPOSITION THROUGH BAG-LEVEL SUPERVISION

Both streams are trained using bag labels. The central question is how this supervision constrains their instance-level products, since an overestimated contribution can be offset by an underestimated one within the same bag.

For the analysis, label the two classes 0 and 1, and let $z_\theta(B)$ be their logit difference for model parameters θ. Write a random bag as $B = (X_1, \ldots, X_n)$ and assume that the true log-odds admit a latent decomposition $z^*(B) = \sum_i c_i^*(B)$. Each model contribution to this margin is the difference between the two class-specific WE scores:

$$\begin{aligned} c_{\theta,i}(B) &= \mathrm{WE}_i^1 - \mathrm{WE}_i^0 = \alpha_i(B)\big(E_1(X_i) - E_0(X_i)\big), \\ \delta_{\theta,i}(B) &= c_{\theta,i}(B) - c_i^*(B), \qquad \sum_i \delta_{\theta,i}(B) = z_\theta(B) - z^*(B). \end{aligned} \tag{10}$$

Here, $\delta_{\theta,i}$ is the error relative to the target margin contribution c_i^*. Summing these errors over instances gives the bag-level margin error, so we need conditions that limit cancellation among instances.

Assumption B (varying instance combinations). Given a data-generating environment T, bags of fixed size $n \ge 2$ contain conditionally independent instances drawn from the same distribution P_T. Both model and target contributions are permutation equivariant, and their errors have finite second moments. This sampling model allows the same local features to occur with different companions.

For a fixed model, we separate each instance's error into its conditional mean across companions and a contextual residual:

$$u_{\theta,T}(x) = \mathbb{E}[\delta_{\theta,i}(B) \mid X_i = x, T], \qquad r_{\theta,i} = \delta_{\theta,i}(B) - u_{\theta,T}(X_i). \tag{11}$$

Assumption C (bounded sensitivity to companions). Let $B^{(j)}$ replace companion X_j with an independent draw from P_T. For every $j \neq i$, assume

$$\frac{1}{2}\mathbb{E}\Big[\big(\delta_{\theta,i}(B) - \delta_{\theta,i}(B^{(j)})\big)^2 \mid X_i = x, T\Big] \le \frac{\kappa}{(n-1)^2} u_{\theta,T}(x)^2, \qquad 0 \le \kappa < \frac{n-1}{n}. \tag{12}$$

This condition limits how much the error varies with the companions relative to its persistent local component. Under B and C, the conditional Efron–Stein inequality yields

$$n\,\mathbb{E}\sum_i r_{\theta,i}^2 \le \eta\,\mathbb{E}\sum_i u_{\theta,T}(X_i)^2, \qquad \eta = \frac{n\kappa}{n-1} < 1. \tag{13}$$

Conditional independence gives $\mathbb{E}[(\sum_i u_{\theta,T}(X_i))^2] \ge \mathbb{E}\sum_i u_{\theta,T}(X_i)^2$. The residual bound limits cancellation by contextual errors, linking aggregate error to instance-level error. When the model and true margins lie in $[-M, M]$, the curvature of binary cross-entropy bounds aggregate margin error by excess risk. Combining these results gives

$$\mathbb{E}\sum_i \delta_{\theta,i}(B)^2 \le \frac{2}{\mu_M \kappa_{\eta,n}}\big(\mathcal{R}_{\mathrm{CE}}(\theta) - \mathcal{R}^*_{\mathrm{CE}}\big), \tag{14}$$

where $\mathcal{R}_{\mathrm{CE}}$ is population cross-entropy risk, $\mathcal{R}^*_{\mathrm{CE}}$ is its Bayes-optimal value, $\mu_M = \sigma(M)(1-\sigma(M))$, and $\kappa_{\eta,n} = (1-\sqrt{\eta})^2/(1+\eta/n) > 0$. Here, σ denotes the sigmoid function.

Implications for EC-MIL. Bounded contextual effects motivate WE's multiplicative form, in which local evidence determines contribution direction and context modulates its strength. Under the stated conditions, excess bag-level cross-entropy risk controls the error of the instance contributions to the binary margin. Detailed derivations are provided in Appendix A.

5 Experiments

5.1 Experimental Setup

We instantiate EC-MIL with ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), and IL-RAMIL (Xiang & Zhang, 2023), and compare each variant with its original backbone. We also evaluate the ABMIL implementation of Additive MIL (Javed et al., 2022) as an intrinsic attribution baseline. Unless otherwise specified, quantitative results report the mean and standard deviation over five random seeds, with error bars indicating standard deviations.

(1) Synthetic MIL. The synthetic experiments follow the signed ground-truth setup of xMIL. Each bag consists of 30 MNIST instances (Deng, 2012). We use a ResNet18 (He et al., 2016) pre-trained on ImageNet to extract 512-dimensional features. For each task and seed, we use 2,000 training bags, 500 validation bags, and 1,000 test bags.

(2) WSI. We use CAMELYON16, a lymph node metastasis dataset (Ehteshami Bejnordi et al., 2017), and BRACS, a breast pathology dataset with diagnostic categories (Brancati et al., 2022). Each slide is a bag of patch embeddings. We compare ResNet18 features with self-supervised features (SSL) from DINO pre-trained on 36,666 WSIs (Caron et al., 2021; Kang et al., 2023).

(3) MUT. We construct MUT for tissue-of-origin prediction using mutation data from cBioPortal (Cerami et al., 2012; Gao et al., 2013). Each bag is a tumor sample, and each instance is a gene mutation event represented by a gene embedding and an 18-dimensional mutation-description vector. Tissue labels follow OncoTree (Kundra et al., 2021). We evaluate scGPT (Cui et al., 2024) and text2vec (Chen & Zou, 2024; Zhang et al., 2025) embeddings (details in Appendix B.6.2).

Attribution evaluation. We compare WE with random rankings, GI (Ancona et al., 2017), IG (Sundararajan et al., 2017), and LOO (Early et al., 2022) on the same EC predictors. We also report original-backbone results, including xMIL-LRP (Hense et al., 2024) on supported architectures. Synthetic evidence recovery and CAMELYON16 localization use AUPRC-2 against signed instance annotations and lesion annotations, respectively. Deletion AUPC (Samek et al., 2016) summarizes the target-class probability as high-scoring instances are removed; lower values indicate faster probability reduction. Implementation and metric definitions are in Appendix B.7.

5.2 Predictive Performance and Context Modeling

Figure 2 compares the AUC of each backbone before and after applying EC-MIL. Predictive performance remains competitive across the three settings. In synthetic MIL, eight of nine paired AUC differences are within 0.005. On WSI, AUC changes range from -0.0342 to $+0.0331$, while all six MUT comparisons favor the EC variant. Full numerical results are provided in Tables A8–A10.

To examine context modeling more directly, we use Adjacent-Pairs, in which a bag is positive if it contains at least one numerically adjacent digit pair among 0–4, such as 1 and 2, regardless of instance order. On this task, ABMIL and EC-ABMIL achieve similar AUCs of 0.8125 and 0.8149. With contextual interactions, EC-TransMIL and EC-ILRAMIL achieve substantially higher AUCs

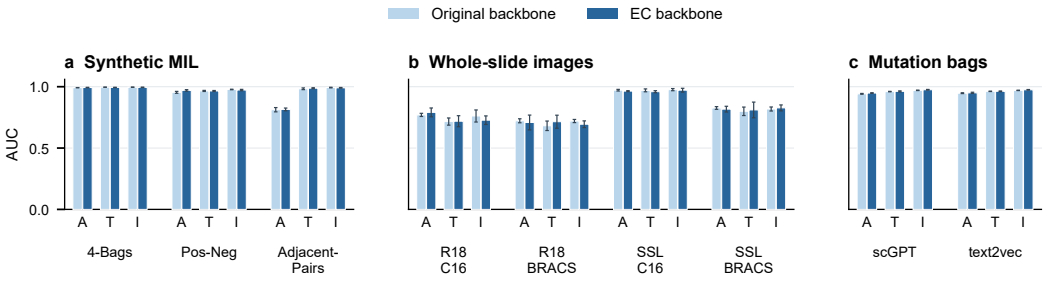


Figure 2: Predictive AUC of original and EC backbones. A / T / I denote ABMIL / TransMIL / ILRAMIL; R18 and C16 denote ResNet18 and CAMELYON16. (Tables A8–A10.)

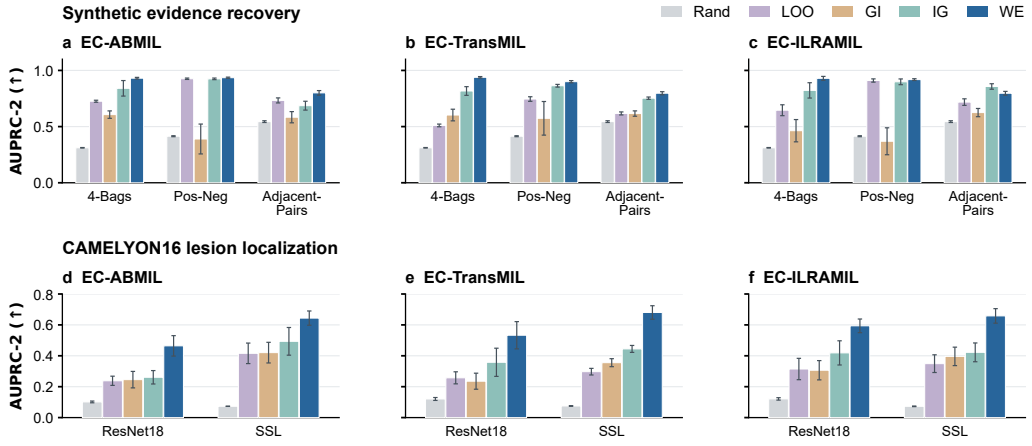


Figure 3: Evidence recovery on EC-MIL predictors: synthetic AUPRC-2 (a–c) and CAMELYON16 localization AUPRC-2 (d–f). Full numerical results are in Tables A11 and A13.

of 0.9878 and 0.9910, respectively, while remaining close to their original backbones. Together, these comparisons show that EC-MIL retains the benefits of backbone context modeling while keeping class evidence local to each instance.

5.3 Signed and Class-Specific Evidence Recovery

The synthetic tasks test complementary aspects of evidence recovery. In 4-Bags, digits 8 and 9 define four classes, with each digit supporting some classes and opposing others. Pos-Neg compares the counts of supportive digits (4, 6, 8) and opposing digits (5, 7, 9). Adjacent-Pairs tests whether attribution identifies digits participating in valid pairs. AUPRC-2 averages retrieval of supportive instances using attribution scores and opposing instances using their negation. Against GI, IG, LOO, and random rankings on the same EC predictor, WE achieves the highest mean AUPRC-2 in eight of nine settings (Figure 3a–c). Tables A11 and A12 provide complete comparisons and synthetic deletion results. To illustrate how context modulates an instance's contribution, Figure 4a follows a fixed digit-1 instance. Replacing a companion digit 7 with digit 2 creates the pair $(1, 2)$, turning a negative bag into a positive one. For EC-ILRAMIL, its class-1 evidence stays at 3.3943, while its weight rises from 0.0753 to 0.9079. Context thus amplifies the instance's contribution when it participates in a class-defining relation.

Compared with Additive MIL's ABMIL implementation, EC-ABMIL achieves higher mean predictive AUC and evidence-recovery AUPRC-2 on all three synthetic tasks (Table 1). Full Additive MIL results, including deletion evaluation, are in Appendix H.

5.4 Attribution on Biomedical Data

WSI localization and deletion. Among methods evaluated on the same EC predictor, WE achieves the highest mean CAMELYON16 AUPRC-2 under both feature settings (Figure 3d–f). Original ABMIL with xMIL-LRP attains higher AUPRC-2 in both settings (Table A13). Figure 4b illustrates

Table 1: Synthetic prediction and intrinsic evidence recovery (means). AdditiveMIL uses ABMIL and Patch-Logits; EC models use WE. Standard deviations are in Tables A8, A11, A19, and A21.

	4-Bags		Pos-Neg		Adjacent-Pairs	
Model	AUC ↑	AUPRC-2 ↑	AUC ↑	AUPRC-2 ↑	AUC ↑	AUPRC-2 ↑
AdditiveMIL (ABMIL)	0.9916	0.9053	0.9492	0.8891	0.7803	0.6027
EC-ABMIL	0.9932	0.9307	0.9702	0.9351	0.8149	0.8000
EC-TransMIL	0.9942	0.9386	0.9652	0.8998	0.9878	0.7970
EC-ILRAMIL	0.9950	0.9295	0.9729	0.9183	0.9910	0.7967

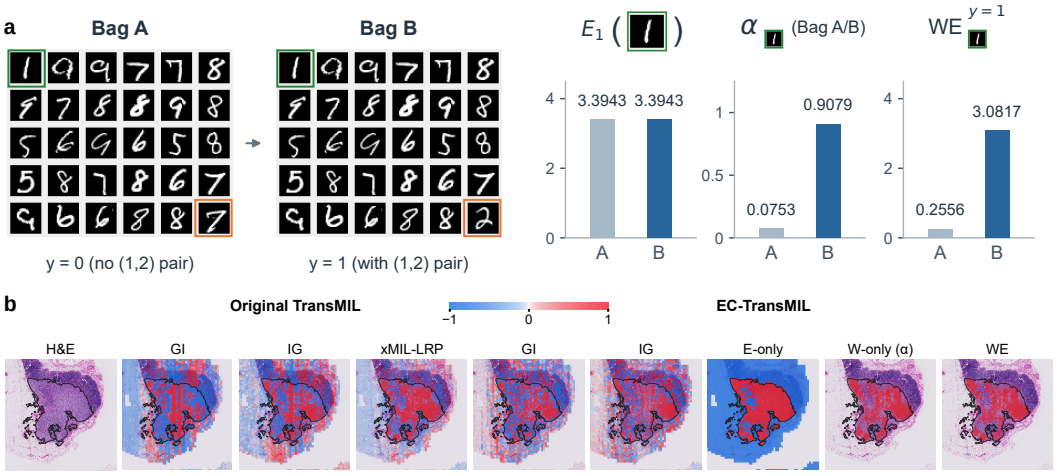


Figure 4: Instance-level explanations. (a) Context modulation of a fixed digit-1 instance in Adjacent-Pairs with EC-ILRAMIL. (b) Tumor-class attribution on CAMELYON16 with TransMIL and EC-TransMIL; black contours mark lesions. More visualization results are in Appendix E.1.1.

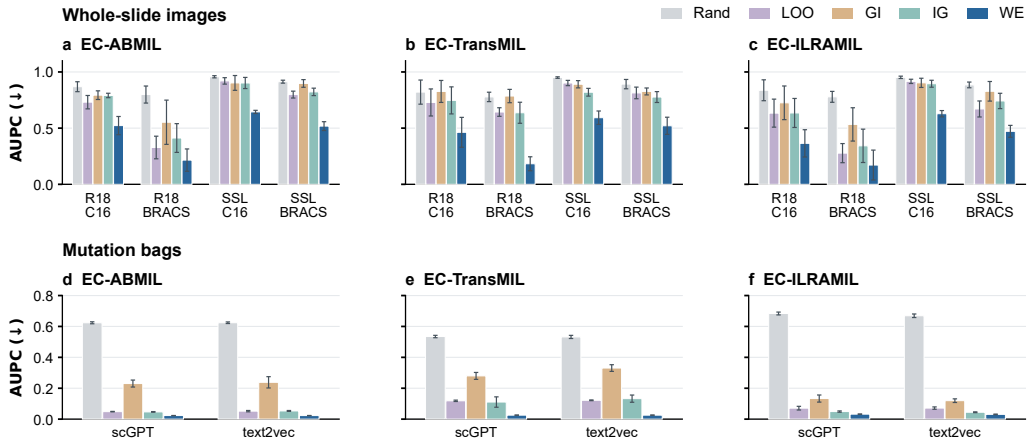


Figure 5: Deletion-based faithfulness on EC-MIL predictors for WSI and MUT. (Full numerical results are in Tables A14 and A15.)

the decomposition on a WSI region: local evidence is positive within much of the lesion and negative in surrounding tissue, while context weights concentrate on the lesion, yielding spatially focused WE. GI and IG exhibit more dispersed positive and negative responses in this TransMIL example. In deletion evaluation, WE gives the lowest mean AUPC in all twelve EC task–feature–backbone settings (Figure 5a–c).

Mutation-bag attribution. The same evaluation extends to mutation events. With both scGPT and text2vec features, WE gives the lowest mean deletion AUPC among methods evaluated on each EC predictor (Figure 5d–f). WE-based gene rankings also reveal tissue-associated patterns, including Bowel–APC, Lung–EGFR, CNS/Brain–IDH1, Pancreas–KRAS, and Kidney–VHL (Figure 6; Table A16), consistent with established cancer genomics (Weinstein et al., 2013; Zehir et al., 2017). The cross-tissue scores also capture shared signals: KRAS receives its highest WE in Pancreas, with

additional positive scores in Bowel and Lung. This pattern is consistent with the established role of KRAS as a recurrent oncogenic driver across multiple cancer types (Simanshu et al., 2017). At the sample level, the Bowel example highlights APC, SOX9, and KRAS, while the Liver example highlights TERT, NPAS3, and CACNA2D1 (Figure A3). Signed WE separates contributions supporting the target tissue from those opposing it, connecting tissue-level patterns with the contributions underlying individual predictions (Table A18).

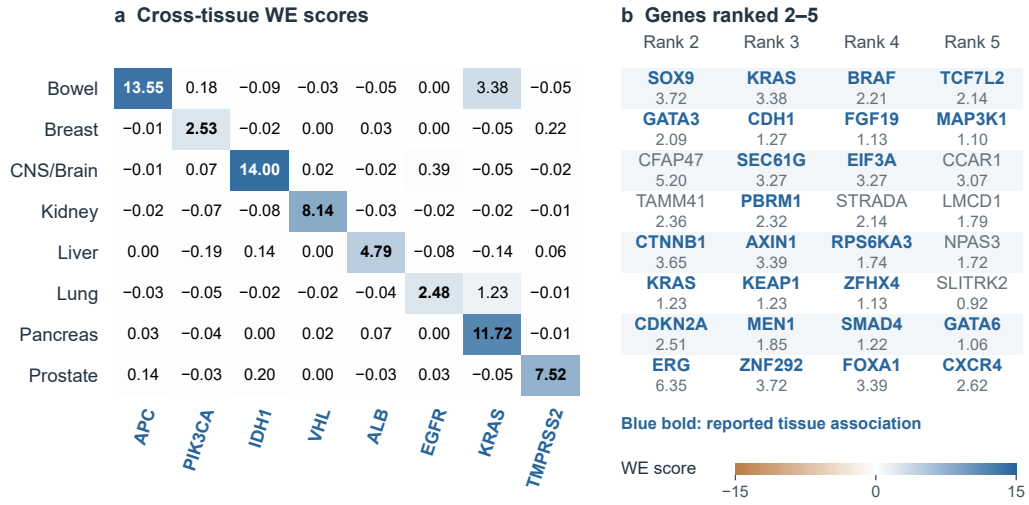


Figure 6: Tissue-level gene attribution on MUT: (a) cross-tissue WE for the top-ranked gene of each selected tissue and (b) the remaining top-five genes and their WE scores.

5.5 Attribution Efficiency

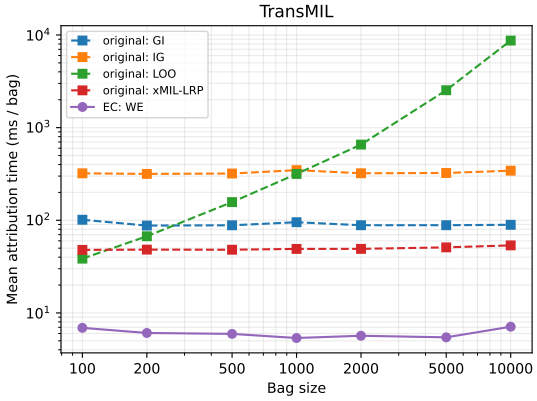


Figure 7: Runtime scaling on TransMIL.

We measure mean all-class attribution time per bag on Adjacent-Pairs with TransMIL for bag sizes of 100–10,000 (Figure 7). WE has the lowest measured runtime and remains nearly constant over this range. LOO exhibits the steepest increase; GI, IG, and xMIL-LRP are also slower than WE.

This advantage also holds on MUT. EC-ILRAMIL takes 29.9 seconds for end-to-end attribution, compared with 7064.3 seconds for LOO on the original backbone, a roughly 236× speedup. Appendix Table A17 further reports times and computational costs across all three backbones on the MUT dataset.

6 Conclusion and Limitations

We propose EC-MIL, an evidence-context disentanglement framework for instance-level interpretation in MIL. By separating instance-local class evidence from bag-dependent context weighting, EC-MIL yields WE: signed, class-specific contributions that sum to the class logit and are obtained in a single forward pass. Our analysis motivates the multiplicative form and relates excess bag-level cross-entropy risk to instance-contribution error. Experiments on synthetic MIL, WSI, and our MUT dataset demonstrate competitive predictive performance, evidence recovery, and deletion-based faithfulness at lower attribution cost. These results show that context modeling and intrinsic attribution can coexist under bag-level supervision. EC-MIL requires architectural adaptation and joint training of both streams, limiting its use for explaining fixed pretrained MIL classifiers.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We have not used generative AI tools for data analysis, idea generation, or code generation, and these disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for improving the grammar, flow, and clarity of the manuscript. We have reviewed all AI-assisted work. Specifically, all LLM-polished paragraphs were carefully inspected and verified by the authors to ensure that the scientific meaning, tone, and technical accuracy remained completely faithful to our original drafts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments in this study are computational and use publicly accessible datasets. No new human participant data were collected, and no participant recruitment or intervention was involved.

REPRODUCIBILITY STATEMENT

Section 3 describes the EC-MIL architecture and the computation of WE. Appendix A provides the theoretical assumptions and complete proofs. Appendix B documents the computing infrastructure, hyperparameters, data preparation, feature extraction, and attribution evaluation, including the construction of MUT in Appendix B.6. The anonymous code repository is linked in the abstract, and the MUT dataset will be released.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

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.

Mohsin Bilal, Robert Jewsbury, Ruoyu Wang, Hammam M AlGhamdi, Amina Asif, Mark Eastwood, and Nasir Rajpoot. An aggregation of aggregation methods in computational pathology. *Medical image analysis*, 88:102885, 2023.

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.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern recognition*, 77: 329–353, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.

Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *BioRxiv*, pp. 2023–10, 2024.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Understanding and unifying fourteen attribution methods with taylor interactions. *arXiv preprint arXiv:2303.01506*, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

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, Christine Evers, and Sarvapali Ramchurn. Model agnostic interpretability for multiple instance learning. *arXiv preprint arXiv:2201.11701*, 2022.

Joseph Early, Gavin KC Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xriGRsoAza.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Camelyon16 Consortium, Meyke Hermsen, Quirine F Manson, 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.

Leo Fillioux, Joseph Boyd, Maria Vakalopoulou, Paul-Henry Cournède, and Stergios Christodoulidis. Structured state space models for multiple instance learning in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 594–604. Springer, 2023.

Olga Fourkioti, Matt De Vries, and Chris Bakal. Camil: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *International conference on learning representations*, volume 2024, pp. 36205–36220, 2024.

Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269): pl1–pl1, 2013.

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*, pp. 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*, pp. 2127–2136. Pmlr, 2018.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20689–20702, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82764461a05e933cc2fd9d312e107d12-Abstract-Conference.html.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3344–3354. IEEE, 2023.

Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R. Gupta, and Prateek Prasanna. SI-MIL: Taming deep MIL for self-interpretability in gigapixel histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11226–11237, 2024.

Ritika Kundra, Hongxin Zhang, Robert Sheridan, Sahussapont Joseph Sirintrapun, Avery Wang, Angelica Ochoa, Manda Wilson, Benjamin Gross, Yichao Sun, Ramyasree Madupuri, et al. Oncotree: a cancer classification system for precision oncology. *JCO clinical cancer informatics*, 5: 221–230, 2021.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14313–14323. IEEE, 2021.

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.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

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

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 685–694, 2015.

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*, pp. 43–53. Springer, 2020.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

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.

Lloyd S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, Princeton, NJ, 1953.

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

Dhirendra K Simanshu, Dwight V Nissley, and Frank McCormick. Ras proteins and their regulators in human disease. *Cell*, 170(1):17–33, 2017.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

Susu Sun, Leslie Tessier, Frédérique Meeuwsen, Clément Grisi, Dominique van Midden, Geert Litjens, and Christian F. Baumgartner. Label-free concept based multiple instance learning for gigapixel histopathology. *arXiv preprint arXiv:2501.02922*, 2025. URL https://arxiv.org/abs/2501.02922.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. *ICLR*, 2(6):7, 2023.

Linfeng Ye, Shayan Mohajer Hamidi, Zhixiang Chi, Guang Li, Mert Pilanci, Takahiro Ogawa, Miki Haseyama, and Konstantinos N Plataniotis. Asmil: Attention-stabilized multiple instance learning for whole slide imaging. *arXiv preprint arXiv:2603.06658*, 2026.

Ahmet Zehir, Ryma Benayed, Ronak H Shah, Aijazuddin Syed, Sumit Middha, Hyunjae R Kim, Preethi Srinivasan, Jianjiong Gao, Debyani Chakravarty, Sean M Devlin, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*, 23(6):703–713, 2017.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18780–18790. IEEE, 2022.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

APPENDIX CONTENTS

A Theoretical Analysis of EC-MIL **15**

A.1 Notation . . . 15
A.2 Local Evidence and the Multiplicative Contribution Form . . . 15
A.3 Contribution Decomposition and the Predictive Distribution . . . 16
A.4 Cross-Entropy Constraints on the Aggregate Contribution Error . . . 16
A.5 Assumptions on the Decomposition Error . . . 17
A.6 Upper Bound on the Decomposition Error . . . 18

B Experimental Setup **19**

B.1 Computing Infrastructure . . . 19
B.2 Training Configurations . . . 19
B.3 Data Availability . . . 20
B.4 Synthetic MIL Experiments . . . 20
B.5 WSI Experiments . . . 21
B.6 MUT Experiments . . . 21
B.7 Attribution . . . 23

C Predictive Performance and Context Modeling **25**

D Signed and Class-Specific Evidence Recovery **27**

E Attribution on Biomedical Data **28**

E.1 WSI Localization and Deletion . . . 28
E.2 Mutation-Bag Attribution . . . 30
E.3 Additional Attribution Plots . . . 30

F Attribution Efficiency **32**

G Biomedical Plausibility of MUT **32**

H Additive MIL Results **33**

H.1 Predictive Performance . . . 33
H.2 Synthetic Evidence Recovery and Deletion . . . 33
H.3 MUT Deletion . . . 34
H.4 WSI Localization and Deletion . . . 34

A THEORETICAL ANALYSIS OF EC-MIL

A.1 NOTATION

Table A1: Notation used in the theoretical analysis.

Symbol	Meaning
$B = (X_1, \ldots, X_n)$	A random bag containing n instances
x and X_i	A realized instance and a random instance, respectively
y and Y	A realized label and a random label, respectively
T	Data-generating environment
θ	Model parameters
$f(x)$ and $g(x, B)$	Instance-local contribution and contextual adjustment, respectively
$s(x)$ and $\beta(x, B)$	Reparameterized local evidence and multiplicative coefficient, respectively
$s_\theta(x)$ and $\alpha_{\theta,i}(B)$	Model local evidence and nonnegative normalized weight, respectively
$c_{\theta,i}(B)$ and $c_i^*(B)$	Model contribution and true contribution, respectively
$z_\theta(B)$ and $z^*(B)$	Model and true binary log-odds, i.e., binary margins
$\ell(z, y)$	Binary cross-entropy for one sample
$\mathcal{Q}_B(z)$	Conditional risk after fixing the bag and taking expectation over the label
$\mathcal{R}_{\mathrm{CE}}(\theta)$ and $\mathcal{R}_{\mathrm{CE}}^*$	Overall model cross-entropy risk and Bayes-optimal risk, respectively
$\delta_{\theta,i}(B)$	Contribution error for the i-th instance
$u_{\theta,T}(x)$ and $r_{\theta,i}$	Conditional mean contribution error and context-dependent residual, respectively
U_θ, V_θ, and D_θ	Conditional mean error energy, residual energy, and total decomposition error energy, respectively
$m_\theta(T)$	Conditional mean error within an environment
λ	Upper-bound coefficient on the magnitude of the contextual adjustment relative to the local contribution
M and μ_M	Absolute binary-margin bound and the corresponding lower bound on the curvature of the cross-entropy risk
κ, η, and $\kappa_{\eta,n}$	Companion-replacement stability coefficient, residual-energy constraint coefficient, and coefficient controlling decomposition error through aggregate error, respectively

Bold symbols $\mathbf{s}$, $\mathbf{c}$, and $\mathbf{z}$ denote multiclass vectors. The scalar symbols s, c, and z denote local evidence, instance contribution, and the bag-level margin in the binary setting, respectively.

A.2 LOCAL EVIDENCE AND THE MULTIPLICATIVE CONTRIBUTION FORM

Let the scalar contribution of an instance be

$$c(x, B) = f(x) + g(x, B). \tag{A1}$$

Here, $f(x)$ is the instance-local contribution and $g(x, B)$ is its contextual adjustment.

Assumption A: Bounded contextual effects In practical MIL tasks, an instance affects the decision through structural features intrinsic to itself. For example, in the toy experiments, particular digits provide the basis for the rule-based decision; in WSI data, a patch contains normal or cancerous tissue; and in MUT data, each gene has its own biological function.

We assume that the magnitude of the contextual adjustment is bounded by a fraction of the local contribution: there exists a constant $0 < \lambda \leq 1$ such that

$$|g(x, B)| \leq \lambda |f(x)|. \tag{A2}$$

By (A2), $f(x) = 0$ implies $g(x, B) = c(x, B) = 0$.

Equation (A2) also constrains the range of $c(x, B)$ according to

$$f(x) - \lambda |f(x)| \leq c(x, B) \leq f(x) + \lambda |f(x)|. \tag{A3}$$

Under Assumption A, define the latent local evidence as

$$s(x) := (1 + \lambda) f(x). \tag{A4}$$

We then define the multiplicative coefficient by

$$\beta(x, B) := \begin{cases} \dfrac{1 + g(x, B)/f(x)}{1 + \lambda}, & f(x) \neq 0, \\ 0, & f(x) = 0. \end{cases} \tag{A5}$$

Equations (A1)–(A5) yield

$$c(x, B) = f(x) + g(x, B) = s(x)\beta(x, B), \qquad \beta(x, B) \in \left[\frac{1 - \lambda}{1 + \lambda}, 1\right] \quad \text{when } f(x) \neq 0. \tag{A6}$$

Since $0 < \lambda \leq 1$, β is nonnegative, so context modulates the contribution's strength without reversing its direction. This motivates WE's multiplicative form; EC-MIL additionally normalizes its context weights across instances.

A.3 Contribution Decomposition and the Predictive Distribution

For a multiclass model, define the local evidence, instance contribution, bag output, and label distribution by

$$\mathbf{c}_{\theta,i}(B) = \alpha_{\theta,i}(B)\mathbf{s}_\theta(X_i), \qquad \mathbf{z}_\theta(B) = \sum_{i=1}^{n} \mathbf{c}_{\theta,i}(B), \qquad p_\theta(y \mid B) = \big[\mathrm{softmax}(\mathbf{z}_\theta(B))\big]_y. \tag{A7}$$

We use the binary case below for clarity. If the model outputs logits for two classes, we take the difference between class 1 and class 0 and define the local evidence and contribution as

$$\begin{aligned} s_\theta(x) &:= [\mathbf{s}_\theta(x)]_1 - [\mathbf{s}_\theta(x)]_0, & c_{\theta,i}(B) &:= \alpha_{\theta,i}(B) s_\theta(X_i), \\ z_\theta(B) &:= [\mathbf{z}_\theta(B)]_1 - [\mathbf{z}_\theta(B)]_0 = \sum_{i=1}^{n} c_{\theta,i}(B). \end{aligned} \tag{A8}$$

Assume that the true log-odds admit an additive latent contribution decomposition. The true and model conditional probabilities are then

$$\begin{aligned} z^*(B) &= \sum_{i=1}^{n} c_i^*(B), \\ p^*(B) &:= \mathbb{P}(Y = 1 \mid B) = \sigma(z^*(B)), \\ p_\theta(B) &= \sigma(z_\theta(B)). \end{aligned} \tag{A9}$$

Here, $\sigma(t) = (1 + e^{-t})^{-1}$. The contribution error and its aggregation satisfy

$$\delta_{\theta,i}(B) := c_{\theta,i}(B) - c_i^*(B), \qquad \sum_{i=1}^{n} \delta_{\theta,i}(B) = z_\theta(B) - z^*(B). \tag{A10}$$

A.4 Cross-Entropy Constraints on the Aggregate Contribution Error

For $y \in \{0, 1\}$, the binary cross-entropy can be written as

$$\ell(z, y) := -y \log \sigma(z) - (1 - y) \log(1 - \sigma(z)) = \log(1 + e^z) - yz. \tag{A11}$$

After fixing a bag and taking the conditional expectation over its possible labels, define the conditional risk

$$\mathcal{Q}_B(z) := \mathbb{E}_{Y|B}[\ell(z, Y)] = \log(1 + e^z) - p^*(B)z. \tag{A12}$$

Its first and second derivatives are

$$\mathcal{Q}'_B(z) = \sigma(z) - p^*(B), \qquad \mathcal{Q}''_B(z) = \sigma(z)(1-\sigma(z)). \tag{A13}$$

By (A9), the first derivative vanishes at the true logit:

$$\mathcal{Q}'_B(z^*(B)) = 0. \tag{A14}$$

Assume that there exists a finite constant $M > 0$ such that, almost surely, $z^*(B), z_\theta(B) \in [-M, M]$. The conditional risk therefore has the uniform curvature lower bound

$$\mathcal{Q}''_B(z) \geq \mu_M, \qquad \mu_M := \sigma(M)(1-\sigma(M)) > 0, \tag{A15}$$

throughout the interval between the two logits.

By the second-order Taylor expansion, together with (A10) and (A14),

$$\begin{aligned}\mathcal{Q}_B(z_\theta(B)) - \mathcal{Q}_B(z^*(B)) &\geq \mathcal{Q}'_B(z^*(B))\big(z_\theta(B) - z^*(B)\big) + \frac{\mu_M}{2}\big(z_\theta(B) - z^*(B)\big)^2 \\ &= \frac{\mu_M}{2}\left(\sum_{i=1}^n \delta_{\theta,i}(B)\right)^2.\end{aligned} \tag{A16}$$

Define the overall cross-entropy risk and its Bayes-optimal value by

$$\mathcal{R}_{\mathrm{CE}}(\theta) := \mathbb{E}_{B,Y}[\ell(z_\theta(B), Y)] = \mathbb{E}_B[\mathcal{Q}_B(z_\theta(B))], \qquad \mathcal{R}^*_{\mathrm{CE}} := \mathbb{E}_B[\mathcal{Q}_B(z^*(B))]. \tag{A17}$$

Taking the expectation of (A16) over the bag distribution gives

$$\mathcal{R}_{\mathrm{CE}}(\theta) - \mathcal{R}^*_{\mathrm{CE}} \geq \frac{\mu_M}{2}\,\mathbb{E}_B\left(\sum_{i=1}^n \delta_{\theta,i}(B)\right)^2. \tag{A18}$$

Expanding the square on the right-hand side of (A18) gives

$$\mathbb{E}_B\left(\sum_{i=1}^n \delta_{\theta,i}(B)\right)^2 = \mathbb{E}_B \sum_{i=1}^n \delta_{\theta,i}(B)^2 + 2\mathbb{E}_B \sum_{1\leq i<j\leq n} \delta_{\theta,i}(B)\delta_{\theta,j}(B). \tag{A19}$$

Equation (A19) reveals the key difficulty. We seek an upper bound on the sum of squared errors that is controlled by the conditional risk. However, the interaction term can contain large negative values that cancel large positive squared-error terms. We therefore focus on controlling the squared-error term in (A19).

A.5 Assumptions on the Decomposition Error

Assumption A motivates the multiplicative contribution form. To relate bag-level prediction risk to instance-contribution error, we introduce conditions that limit cancellation in (A19).

We model varying instance combinations through conditional independent sampling and bound the sensitivity of contribution errors to companion replacement. Together, these conditions control the contextual residual relative to the persistent local component of the error.

Assumption B: Varying instance combinations This assumption models how the same local features can occur with different companions through conditional independent sampling.

Specifically, each bag contains a fixed number $n \geq 2$ of instances. Given the data-generating environment T, assume that

$$X_1, \ldots, X_n \mid T \overset{\text{i.i.d.}}{\sim} P_T. \tag{A20}$$

The true and model contributions are both permutation equivariant, and $\delta_{\theta,i}(B)$ has a finite second moment.

For a fixed model θ, define the conditional mean contribution error of the same instance across different combinations and its residual by

$$u_{\theta,T}(x) := \mathbb{E}[\delta_{\theta,i}(B) \mid X_i = x, T], \qquad r_{\theta,i} := \delta_{\theta,i}(B) - u_{\theta,T}(X_i). \tag{A21}$$

Equation (A21) implies

$$\delta_{\theta,i}(B) = u_{\theta,T}(X_i) + r_{\theta,i}, \qquad \mathbb{E}[r_{\theta,i} \mid X_i, T] = 0. \tag{A22}$$

Assumption C: Bounded sensitivity to companions Fix an instance $X_i = x$ and an environment T. Let $B^{(j)}$ denote the bag obtained by replacing companion X_j with an independent draw from P_T.

We assume that there exists a constant $0 \le \kappa < (n-1)/n$ such that, for every $j \neq i$,

$$\frac{1}{2}\mathbb{E}\left[\left(\delta_{\theta,i}(B) - \delta_{\theta,i}(B^{(j)})\right)^2 \middle| X_i = x, T\right] \le \frac{\kappa}{(n-1)^2} u_{\theta,T}(x)^2. \tag{A23}$$

The term $u_{\theta,T}(x)^2$ is the squared conditional mean error across companions. The condition bounds the change in contribution error under a single-companion replacement relative to this persistent local component.

By the conditional independence in Assumption B and the conditional Efron–Stein inequality, for fixed $X_i = x$ and T, the conditional variance is bounded by one half of the sum of the squared effects of all single-companion replacements. Combining this inequality with (A23) gives

$$\mathrm{Var}(\delta_{\theta,i}(B)|X_i = x, T) \le \frac{\kappa}{n-1} u_{\theta,T}(x)^2. \tag{A24}$$

LEMMA (RESIDUAL-ENERGY CONSTRAINT). Under Assumption C, define $\eta := n\kappa/(n-1) < 1$. Then (A24), together with permutation symmetry, implies

$$n\,\mathbb{E}\sum_{i=1}^{n} r_{\theta,i}^2 \le \eta\,\mathbb{E}\sum_{i=1}^{n} u_{\theta,T}(X_i)^2, \qquad 0 \le \eta < 1. \tag{A25}$$

A.6 UPPER BOUND ON THE DECOMPOSITION ERROR

All expectations below are taken with respect to the data distribution of a new bag and its environment for a fixed model. Define

$$U_\theta := \mathbb{E}\sum_{i=1}^{n} u_{\theta,T}(X_i)^2, \qquad V_\theta := \mathbb{E}\sum_{i=1}^{n} r_{\theta,i}^2, \qquad D_\theta := \mathbb{E}\sum_{i=1}^{n} \delta_{\theta,i}(B)^2. \tag{A26}$$

By (A25), $nV_\theta \le \eta U_\theta$. Define the conditional mean within an environment as

$$m_\theta(T) := \mathbb{E}[u_{\theta,T}(X_i) \mid T]. \tag{A27}$$

For $i \neq j$, Equation (A20) gives

$$\mathbb{E}[u_{\theta,T}(X_i)u_{\theta,T}(X_j) \mid T] = m_\theta(T)^2. \tag{A28}$$

Taking the expectation of the squared sum of the local errors gives

$$\mathbb{E}\left(\sum_{i=1}^{n} u_{\theta,T}(X_i)\right)^2 = U_\theta + n(n-1)\mathbb{E}_T[m_\theta(T)^2] \ge U_\theta. \tag{A29}$$

On the other hand, by the Cauchy–Schwarz inequality and (A25),

$$\mathbb{E}\left(\sum_{i=1}^{n} r_{\theta,i}\right)^2 \le n\,\mathbb{E}\sum_{i=1}^{n} r_{\theta,i}^2 = nV_\theta \le \eta U_\theta. \tag{A30}$$

Let $u_{\theta,i} := u_{\theta,T}(X_i)$ and use the L^2 norm of a random scalar, $\|Z\|_{L^2} := (\mathbb{E}[Z^2])^{1/2}$. By the reverse triangle inequality, together with (A29) and (A30),

$$\begin{aligned}
\sqrt{\mathbb{E}\left(\sum_i \delta_{\theta,i}\right)^2} &= \left\|\sum_i u_{\theta,i} + \sum_i r_{\theta,i}\right\|_{L^2} \\
&\ge \left\|\sum_i u_{\theta,i}\right\|_{L^2} - \left\|\sum_i r_{\theta,i}\right\|_{L^2} \\
&\ge (1 - \sqrt{\eta})\sqrt{U_\theta}.
\end{aligned} \tag{A31}$$

Because $0 \le \eta < 1$, the right-hand side of (A31) is nonnegative. Squaring both sides yields

$$\mathbb{E}\left(\sum_i \delta_{\theta,i}\right)^2 \ge (1-\sqrt{\eta})^2 U_\theta. \tag{A32}$$

The conditional zero-mean property in (A22) makes the local mean error orthogonal to the residual of the corresponding instance:

$$\mathbb{E}[u_{\theta,i} r_{\theta,i}] = \mathbb{E}[u_{\theta,i}\,\mathbb{E}[r_{\theta,i} \mid X_i, T]] = 0. \tag{A33}$$

Therefore, the total decomposition error admits the exact energy decomposition

$$D_\theta = \mathbb{E}\sum_i (u_{\theta,i} + r_{\theta,i})^2 = U_\theta + V_\theta \le \left(1 + \frac{\eta}{n}\right) U_\theta. \tag{A34}$$

Combining (A32) and (A34) and eliminating U_θ gives

$$\mathbb{E}\left(\sum_i \delta_{\theta,i}\right)^2 \ge \kappa_{\eta,n} D_\theta, \qquad \kappa_{\eta,n} := \frac{(1-\sqrt{\eta})^2}{1+\eta/n} > 0. \tag{A35}$$

When $U_\theta = 0$, Equation (A25) also gives $V_\theta = D_\theta = 0$, so (A35) remains valid.

Substituting (A35) into (A19) directly bounds the negative interaction term:

$$2\mathbb{E}\sum_{i<j} \delta_{\theta,i}(B)\delta_{\theta,j}(B) \ge -(1-\kappa_{\eta,n}) D_\theta. \tag{A36}$$

Combining (A18) and (A35) yields the following bound on the decomposition error in terms of the overall cross-entropy risk:

$$\mathcal{R}_{\mathrm{CE}}(\theta) - \mathcal{R}^*_{\mathrm{CE}} \ge \frac{\mu_M \kappa_{\eta,n}}{2} D_\theta. \tag{A37}$$

Finally,

$$\mathbb{E}\sum_{i=1}^n \delta_{\theta,i}(B)^2 \le \frac{2}{\mu_M \kappa_{\eta,n}} \left(\mathcal{R}_{\mathrm{CE}}(\theta) - \mathcal{R}^*_{\mathrm{CE}}\right) = \frac{2(1+\eta/n)}{\mu_M (1-\sqrt{\eta})^2} \left(\mathcal{R}_{\mathrm{CE}}(\theta) - \mathcal{R}^*_{\mathrm{CE}}\right). \tag{A38}$$

B Experimental Setup

B.1 Computing Infrastructure

All training, testing, and attribution experiments were conducted on a single NVIDIA A100 SXM4 80 GB GPU. The main results were obtained using five different random seeds: 42, 3407, 99, 5, and 256. The experiments for MUT attribution efficiency and biological plausibility utilized seed 42.

B.2 Training Configurations

All models are optimized using the AdamW optimizer and the cross-entropy (CE) loss function. Automatic Mixed Precision (AMP) is enabled on CUDA. Early stopping selects the checkpoint with the best validation score under the metrics listed in Table A2.

In all settings, ABMIL uses the hidden dimension specified in its respective configuration as its attention or embedding dimension. EC models use a softmax-weighted sum and fully connected layers. EC-ABMIL applies its three fully connected layers independently to each instance. Detailed configurations can be found in Table A3.

Table A2: Training hyperparameters for the three experimental settings.

Setting	Epochs	Batch	Accm	LR	Weight Decay	Patience	Monitor
Toy	100	32	1	0.0001	0.0000	20	Val AUC
WSI	100	1	1	0.0001	0.0001	20	Val AUC
MUT	100	8	2	0.0001	0.0001	20	Val AUC

Table A3: Architecture hyperparameters for the three experimental settings.

Setting	Hidden dimension	Heads × depth	Dropout	PPEG	Landmarks	ILRA top
Toy	128	2×2	0.0	No	64	4
WSI	256	8×2	0.2	Yes	256	64
MUT	512	8×2	0.2	No	256	64

B.3 Data Availability

Data used in this paper can be accessed via the links in Table A4.

Table A4: Access information for the pre-trained model and datasets.

Resource	Access
ResNet18	`https://download.pytorch.org/models/resnet18-f37072fd.pth`
MNIST	`https://ossci-datasets.s3.amazonaws.com/mnist/`
CAMELYON16	`https://camelyon16.grand-challenge.org/`
BRACS	`https://www.bracs.icar.cnr.it/`
WSI-Feature	`https://github.com/dazhangyu123/ACMIL`
MUT source data	`https://www.cbioportal.org/`
OncoTree	`https://oncotree.mskcc.org/`
NCG	`http://network-cancer-genes.org/`
OncoKB	`https://www.oncokb.org/`

B.4 Synthetic MIL Experiments

The synthetic task definitions below use zero-based class labels, whereas the general multiclass notation above uses labels $1, \ldots, C$.

B.4.1 Bag Generation

The toy experiments use three tasks: 4-Bags, Pos-Neg, and Adjacent-Pairs. Each bag contains 30 MNIST instances; 2,000 training bags, 500 validation bags, and 1,000 test bags are generated for each seed. The MNIST images are passed through a pre-trained ResNet18 to produce 512-dimensional features.

Given a running seed s, the training, validation, and test datasets utilize seeds s, $s+101$, and $s+202$, respectively. Each bag is generated according to the following procedure:

1. For each digit $d \in \{0, \ldots, 9\}$, sample independently $m_d \sim \text{Bernoulli}\,(0.5)$ to obtain the set of active digits for the current bag, $S = \{d : m_d = 1\}$; if S is empty, re-sample.

2. Uniformly sample 30 digits with replacement from S to form the sequence of digits of the bag.

3. For each digit in the sequence, uniformly sample a 512-dimensional feature vector with replacement from the corresponding digit's feature library.

4. Determine the bag label according to the task rules described in Table A5, and generate instance-level ground-truth evidence using the class-conditional evidence matrix.

Table A5: Definitions of the synthetic MIL classification tasks.

Task	Number of classes	Bag rule
4-Bags	4	Determined by whether digits 8 and 9 appear at least once, forming four categories: neither appears, only 8 appears, only 9 appears, or both appear.
Pos-Neg	2	If the number of positive digits is greater than the number of negative digits, the bag is positive; otherwise, it is negative.
Adjacent-Pairs	2	Considering only digits 0–4, the label is 1 if the bag contains both digits of at least one numerically adjacent pair, $(0, 1)$, $(1, 2)$, $(2, 3)$, or $(3, 4)$, regardless of their positions; otherwise, the label is 0.

B.4.2 CLASS-CONDITIONAL EVIDENCE

Let $E_{cd} \in \{-1,\ 0,\ 1\}$ denote the ground-truth evidence of digit d for class c, where 1, -1, and 0 represent supportive, opposing, and neutral (or irrelevant), respectively.

Class-specific evidence for 4-Bags is shown in Table A6:

Table A6: Class-specific instance contributions for the 4-Bags task.

Class c	0–7	8	9
0: neither 8 nor 9	0	-1	-1
1: 8 only	0	1	-1
2: 9 only	0	-1	1
3: both 8 and 9	0	1	1

Class-specific evidence for Pos-Neg is shown in Table A7:

Table A7: Class-specific instance contributions for the Pos-Neg task.

Class c	0–3	4	5	6	7	8	9
0: negative majority	0	-1	1	-1	1	-1	1
1: positive majority	0	1	-1	1	-1	1	-1

The evidence for Adjacent-Pairs depends on the current bag and therefore cannot be represented by a single fixed matrix. For a bag B, let $P(B)$ be the set of digits involved in all numerically adjacent pairs whose two digits are both present in the bag. Then,

$$E_{1d}(B) = \mathbf{1}[d \in P(B)],\ E_{0d}(B) = -\mathbf{1}[d \in P(B)] \tag{A39}$$

Only instance digits belonging to at least one numerically adjacent pair present in the bag have non-zero evidence; the evidence for digits 5–9 and digits not participating in any valid adjacent pair is 0, where $\mathbf{1}[\cdot]$ denotes the indicator function.

B.5 WSI EXPERIMENTS

The WSI experiments use CAMELYON16 and BRACS. We compare ResNet18 features (He et al., 2016) with 384-dimensional SSL features extracted using DINO pre-trained on 36,666 WSIs (Caron et al., 2021; Kang et al., 2023). Each slide is a bag stored as an HDF5 group containing a two-dimensional array `feat`, with one patch embedding per row, and an integer `label`. All patches of each slide are used for training, validation, testing, and attribution.

CAMELYON16 uses the official split: 129 slides are fixed as the test set, while the remaining 269 slides are split into stratified train/validation sets under each seed, with the validation set fixed at 40 slides. The official test set of CAMELYON16 remains invariant in all seeds.

BRACS performs two stratified splits on all 537 slides: first, split them into a training set and a temporary set at a 60% / 40% ratio, and then split the temporary set into validation and test sets at a 37.5%/62.5% ratio, resulting in final proportions of 60%/15%/25%. All stratification operations use the current random seed.

B.6 MUT EXPERIMENTS

B.6.1 COHORT CONSTRUCTION

Figure A1 summarizes the data preprocessing workflow used in the MUT experiments. We used a list of genes obtained from the NCG dataset to pre-screen the genes to be studied. The samples are assigned to OncoTree tissue types; samples with rare tissues or low tumor mutational burden are removed before being used in the MUT experiments. The current MUT implementation performs a

65%/15%/20% train/validation/test split for each random seed, while performing stratified sampling across samples with varying numbers of classes.

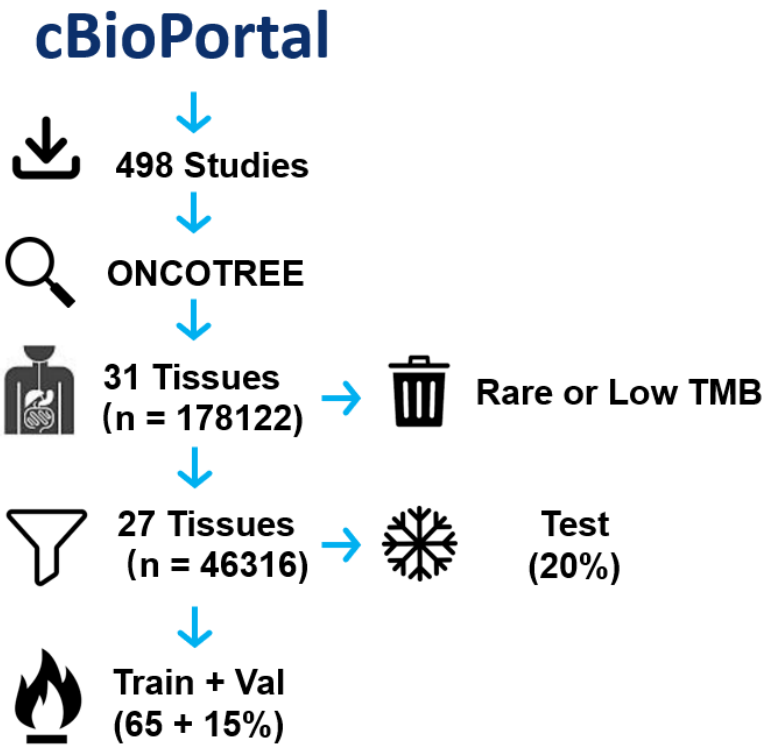


Figure A1: Building MUT cohort.

B.6.2 MUT Instance Features

Each mutation instance is formed by concatenating a 512-dimensional gene vector with an 18-dimensional mutation description vector, resulting in a model input dimension of 530.

In our study, we used two approaches with different focuses to generate 512-dimensional gene embeddings. The text2vec method focuses on existing knowledge about genes, whereas scGPT focuses on the expression patterns of a given gene and other genes in cells.

To obtain text2vec gene embeddings, Qwen3-max is used to generate text descriptions of the gene. Prompts are used to constrain the output into five fields: Molecular Function, Signal Pathways, Role in Cancer, Clinical Relevance, and Summary. The text from these five fields is then fed into a Qwen text-embedding-v4 text2vec model to produce the averaged vector.

To obtain scGPT gene embeddings, we extract gene vectors from the embedding layer of the pre-trained scGPT model (Cui et al., 2024).

The pre-extracted gene embeddings are processed using BERT whitening to facilitate subsequent training.

Identifiable mutation descriptions are encoded in the same 18-dimensional vector. Dimensions 1–15 and 17–18 use 0/1 encoding, while the 16th dimension is numerical, representing the CNV value.

The first 15 dimensions indicate whether the mutation belongs to one of the following SNV categories, respectively: mutations in the $3'$ flanking region, $3'$ untranslated region (UTR), $5'$ flanking region, or $5'$ UTR; frameshift mutations; in-frame insertions or deletions; intronic mutations; missense mutations; mutations subject to nonsense-mediated decay (NMD); nonsense mutations; non-

stop mutations; RNA mutations; silent mutations; splice-site mutations; and translation-start-site mutations.

The 17th dimension indicates whether an SV involving copy-number changes—such as an amplification or deletion—has occurred. The 18th dimension indicates whether another type of SV, such as an inversion, has occurred.

B.7 Attribution

B.7.1 Attribution Methods

Random

Independently sample instance scores from a standard normal distribution.

Leave-One-Out (LOO)

$$a_i = z_c\left(B\right) - z_c\left(B \setminus \{x_i\}\right), \tag{A40}$$

where z_c is the target class score.

Gradient $\times$ Input

$$\sum_j x_{ij} \frac{\partial z_c}{\partial x_{ij}}, \tag{A41}$$

summed over the input feature dimensions.

Integrated Gradients

Uses an all-zero feature vector as the baseline, takes K equidistant points along the straight-line path from 0 to the input, averages the gradients, multiplies the result by the input difference and sums over the feature dimensions. $K = 50$ is used for the synthetic MIL experiments and $K = 20$ for WSI and MUT.

xMIL-LRP

Supports only original ABMIL and original TransMIL. The implementation initializes relevance with the logit of the target class, redistributes relevance through each layer using the ϵ-rule with default $\gamma = 0$ and $\epsilon = 10^{-8}$, does not propagate bias, and finally sums the feature dimensions to obtain the relevance of the instance. The current implementation does not support original ILRAMIL or any EC-MIL models.

WE

Applicable only to EC models:

$$a_i = \mathrm{WE}_i^c = \alpha_i\left(B\right) E_c\left(x_i\right), \tag{A42}$$

which is directly returned from a single inference.

B.7.2 AUPRC-2

Synthetic MIL tasks possess class-conditional signs. For a bag B and class c, let the ground-truth evidence be $e_{ic}(B) \in \{-1,\ 0,\ 1\}$, and the attribution score be $a_{ic}(B)$. If the set of positive evidence is neither empty nor the universal set, define:

$$\mathrm{AP}_{Bc}^{+} = \mathrm{AP}\left(\left(\mathbf{1}[e_{ic}(B) = 1]\right)_{i=1}^{n},\ \left(a_{ic}(B)\right)_{i=1}^{n}\right). \tag{A43}$$

If the set of negative evidence is neither empty nor the universal set, define:

$$\mathrm{AP}_{Bc}^{-} = \mathrm{AP}\left((\mathbf{1}[e_{ic}(B) = -1])_{i=1}^{n},\ (-a_{ic}(B))_{i=1}^{n}\right). \tag{A44}$$

For each bag-class pair, AUPRC-2 first averages the positive and negative average precisions (APs) wherever they are well-defined; the final metric is then computed as the arithmetic mean over all bag-class pairs that have at least one well-defined AP. Bags containing no non-zero instance evidence are excluded from the evaluation. This definition rewards supportive evidence with higher scores and opposing evidence with lower scores. Higher AUPRC-2 values indicate better performance in synthetic MIL experiments.

B.7.3 AUPC

Since WSI and MUT lack complete instance-level ground truth, the removal curve is employed to evaluate faithfulness. Only test bags correctly classified by the model are retained. For the true class c, instances are sorted in descending order of their attribution scores, and at $k = 0, 1, \ldots, 100$, the top

$$m_k = \left\lfloor \frac{k}{100} n \right\rfloor \tag{A45}$$

highest-scoring instances are removed. At each step, the true class output $p_c^{(k)}$ of the remaining bag is recorded; when all instances are removed, an all-zero bag with the same shape as the original bag is used. The AUPC for a single bag is defined as the arithmetic mean of 101 discrete points:

$$\mathrm{AUPC}\,(B) = \frac{1}{101} \sum_{k=0}^{100} p_c^{(k)} \tag{A46}$$

Dataset-level results are first averaged over all correctly classified bags within each seed, and then the mean and standard deviation across five seeds are reported. In the WSI and MUT experiments, lower AUPC values are better because they indicate a more rapid drop in the true class probability after high-scoring instances are removed, which in turn indicates greater consistency between the attribution ranking and the model's reliance.

C Predictive Performance and Context Modeling

The following sections provide all original numerical results, including standard deviations and comparison markings. Tables A8–A10 report ACC, F1, and AUC for the original and EC backbones.

Table A8: Predictive performance in synthetic MIL tasks.

Model	ACC	F1	AUC
4-Bags			
ABMIL	0.9464 ± 0.0059	0.9462 ± 0.0059	0.9923 ± 0.0007
EC-ABMIL	0.9322 ± 0.0106	0.9310 ± 0.0113	0.9932 ± 0.0018
TransMIL	0.9720 ± 0.0028	0.9720 ± 0.0028	0.9960 ± 0.0002
EC-TransMIL	0.9652 ± 0.0049	0.9651 ± 0.0049	0.9942 ± 0.0011
ILRAMIL	0.9700 ± 0.0057	0.9701 ± 0.0055	0.9960 ± 0.0010
EC-ILRAMIL	0.9684 ± 0.0043	0.9684 ± 0.0043	0.9950 ± 0.0014
Pos-Neg			
ABMIL	0.8696 ± 0.0057	0.8691 ± 0.0058	0.9541 ± 0.0077
EC-ABMIL	0.8940 ± 0.0130	0.8938 ± 0.0131	0.9702 ± 0.0059
TransMIL	0.8894 ± 0.0160	0.8893 ± 0.0160	0.9655 ± 0.0044
EC-TransMIL	0.8802 ± 0.0146	0.8792 ± 0.0153	0.9652 ± 0.0039
ILRAMIL	0.9120 ± 0.0092	0.9119 ± 0.0093	0.9774 ± 0.0021
EC-ILRAMIL	0.8992 ± 0.0147	0.8988 ± 0.0148	0.9729 ± 0.0054
Adjacent-Pairs			
ABMIL	0.8174 ± 0.0080	0.7981 ± 0.0070	0.8125 ± 0.0174
EC-ABMIL	0.8202 ± 0.0201	0.8013 ± 0.0161	0.8149 ± 0.0111
TransMIL	0.9398 ± 0.0126	0.9369 ± 0.0143	0.9831 ± 0.0070
EC-TransMIL	0.9504 ± 0.0107	0.9482 ± 0.0114	0.9878 ± 0.0039
ILRAMIL	0.9684 ± 0.0051	0.9670 ± 0.0056	0.9930 ± 0.0024
EC-ILRAMIL	0.9620 ± 0.0064	0.9605 ± 0.0068	0.9910 ± 0.0028

Table A9: Predictive performance in WSI tasks.

Model	ACC	F1	AUC
ResNet18-CAMELYON16			
ABMIL	0.7721 ± 0.0253	0.7446 ± 0.0217	0.7703 ± 0.0128
EC-ABMIL	0.7860 ± 0.0248	0.7669 ± 0.0223	0.7906 ± 0.0360
TransMIL	0.7070 ± 0.0348	0.6735 ± 0.0351	0.7159 ± 0.0292
EC-TransMIL	0.7550 ± 0.0412	0.7211 ± 0.0350	0.7187 ± 0.0450
ILRAMIL	0.7612 ± 0.0499	0.7363 ± 0.0334	0.7613 ± 0.0492
EC-ILRAMIL	0.7659 ± 0.0603	0.7340 ± 0.0539	0.7271 ± 0.0349
ResNet18-BRACS			
ABMIL	0.5930 ± 0.0265	0.4664 ± 0.0190	0.7223 ± 0.0163
EC-ABMIL	0.5535 ± 0.0911	0.4902 ± 0.0731	0.7084 ± 0.0603
TransMIL	0.5326 ± 0.0225	0.4252 ± 0.0339	0.6815 ± 0.0386
EC-TransMIL	0.5721 ± 0.0567	0.4752 ± 0.0314	0.7146 ± 0.0531
ILRAMIL	0.5628 ± 0.0676	0.4706 ± 0.0529	0.7203 ± 0.0129
EC-ILRAMIL	0.5186 ± 0.0379	0.4702 ± 0.0368	0.6941 ± 0.0274
SSL-CAMELYON16			
ABMIL	0.9457 ± 0.0190	0.9421 ± 0.0197	0.9717 ± 0.0066
EC-ABMIL	0.9519 ± 0.0192	0.9488 ± 0.0198	0.9641 ± 0.0036
TransMIL	0.9473 ± 0.0076	0.9438 ± 0.0075	0.9703 ± 0.0117
EC-TransMIL	0.9457 ± 0.0230	0.9418 ± 0.0241	0.9594 ± 0.0082
ILRAMIL	0.9597 ± 0.0058	0.9569 ± 0.0062	0.9761 ± 0.0088
EC-ILRAMIL	0.9442 ± 0.0150	0.9408 ± 0.0155	0.9717 ± 0.0153
SSL-BRACS			
ABMIL	0.6349 ± 0.0261	0.5607 ± 0.0456	0.8267 ± 0.0110
EC-ABMIL	0.6605 ± 0.0171	0.5839 ± 0.0338	0.8172 ± 0.0233
TransMIL	0.6512 ± 0.0382	0.5981 ± 0.0411	0.7995 ± 0.0345
EC-TransMIL	0.6349 ± 0.0474	0.5706 ± 0.0709	0.8107 ± 0.0638
ILRAMIL	0.6186 ± 0.0248	0.5252 ± 0.0484	0.8183 ± 0.0164
EC-ILRAMIL	0.6605 ± 0.0186	0.5793 ± 0.0453	0.8273 ± 0.0242

Table A10: Predictive performance in MUT tasks.

Model	ACC	F1	AUC
scGPT			
ABMIL	0.7181 ± 0.0044	0.5926 ± 0.0122	0.9418 ± 0.0033
EC-ABMIL	0.7656 ± 0.0045	0.6462 ± 0.0152	0.9482 ± 0.0026
TransMIL	0.7300 ± 0.0079	0.5729 ± 0.0156	0.9602 ± 0.0019
EC-TransMIL	0.7525 ± 0.0108	0.6307 ± 0.0172	0.9620 ± 0.0040
ILRAMIL	0.7734 ± 0.0090	0.6489 ± 0.0132	0.9704 ± 0.0018
EC-ILRAMIL	0.8225 ± 0.0089	0.7241 ± 0.0122	0.9752 ± 0.0025
text2vec			
ABMIL	0.7317 ± 0.0035	0.6155 ± 0.0125	0.9473 ± 0.0034
EC-ABMIL	0.7679 ± 0.0036	0.6546 ± 0.0119	0.9501 ± 0.0048
TransMIL	0.7334 ± 0.0098	0.5850 ± 0.0165	0.9622 ± 0.0015
EC-TransMIL	0.7511 ± 0.0067	0.6299 ± 0.0113	0.9626 ± 0.0026
ILRAMIL	0.7852 ± 0.0119	0.6599 ± 0.0208	0.9700 ± 0.0015
EC-ILRAMIL	0.8136 ± 0.0134	0.7087 ± 0.0241	0.9748 ± 0.0022

D SIGNED AND CLASS-SPECIFIC EVIDENCE RECOVERY

Table A11 reports synthetic AUPRC-2 scores; Table A12 provides the corresponding deletion AUPC results.

Table A11: AUPRC-2 attribution results for the 4-Bags, Pos-Neg, and Adjacent-Pairs settings.

	ABMIL	EC-ABMIL	TransMIL	EC-TransMIL	ILRAMIL	EC-ILRAMIL
			4-Bags			
Rand	0.3115 ± 0.0022	0.3115 ± 0.0022	0.3115 ± 0.0022	0.3115 ± 0.0022	0.3115 ± 0.0022	0.3115 ± 0.0022
LOO	0.6366 ± 0.0106	0.7257 ± 0.0089	0.4592 ± 0.0136	0.5104 ± 0.0119	0.5798 ± 0.0624	0.6457 ± 0.0485
G × I	0.8623 ± 0.0061	0.6066 ± 0.0331	0.4121 ± 0.0234	0.6029 ± 0.0518	0.6159 ± 0.0415	0.4635 ± 0.0991
IG	0.8740 ± 0.0053	0.8407 ± 0.0691	0.7597 ± 0.0744	0.8168 ± 0.0387	0.8049 ± 0.0594	0.8223 ± 0.0682
xMIL-LRP	0.9179 ± 0.0098	–	0.7826 ± 0.0565	–	–	–
WE	–	**0.9307 ± 0.0074**	–	**0.9386 ± 0.0066**	–	**0.9295 ± 0.0172**
			Pos-Neg			
Rand	0.4144 ± 0.0043	0.4144 ± 0.0043	0.4144 ± 0.0043	0.4144 ± 0.0043	0.4144 ± 0.0043	0.4144 ± 0.0043
LOO	0.8870 ± 0.0087	0.9259 ± 0.0061	0.7233 ± 0.0304	0.7453 ± 0.0204	0.8736 ± 0.0274	0.9102 ± 0.0145
G × I	0.9076 ± 0.0075	0.3896 ± 0.1329	0.4584 ± 0.0460	0.5737 ± 0.1496	0.8526 ± 0.0335	0.3694 ± 0.1199
IG	0.9144 ± 0.0090	0.9249 ± 0.0071	0.8814 ± 0.0282	0.8636 ± 0.0120	**0.9205 ± 0.0059**	0.8986 ± 0.0256
xMIL-LRP	0.9142 ± 0.0084	–	**0.9067 ± 0.0090**	–	–	–
WE	–	**0.9351 ± 0.0052**	–	0.8998 ± 0.0096	–	0.9183 ± 0.0079
			Adjacent-Pairs			
Rand	0.5443 ± 0.0078	0.5443 ± 0.0078	0.5443 ± 0.0078	0.5443 ± 0.0078	0.5443 ± 0.0078	0.5443 ± 0.0078
LOO	0.6278 ± 0.0181	0.7330 ± 0.0224	0.6188 ± 0.0062	0.6172 ± 0.0134	0.7327 ± 0.0280	0.7181 ± 0.0292
G × I	0.6483 ± 0.0073	0.5835 ± 0.0499	0.6034 ± 0.0125	0.6160 ± 0.0239	0.7342 ± 0.0125	0.6247 ± 0.0366
IG	0.7229 ± 0.0149	0.6862 ± 0.0394	0.7779 ± 0.0319	0.7530 ± 0.0102	0.8298 ± 0.0193	**0.8571 ± 0.0244**
xMIL-LRP	0.7968 ± 0.0113	–	0.7516 ± 0.0408	–	–	–
WE	–	**0.8000 ± 0.0205**	–	**0.7970 ± 0.0137**	–	0.7967 ± 0.0167

Table A12: AUPC results for the 4-Bags, Pos-Neg, and Adjacent-Pairs settings.

	ABMIL	EC-ABMIL	TransMIL	EC-TransMIL	ILRAMIL	EC-ILRAMIL
			4-Bags			
Rand	0.7881 ± 0.0108	0.7729 ± 0.0163	0.8240 ± 0.0095	0.7981 ± 0.0245	0.8268 ± 0.0069	0.8098 ± 0.0168
LOO	0.6809 ± 0.0289	0.6341 ± 0.0435	0.7171 ± 0.0339	0.7393 ± 0.0448	0.5664 ± 0.0428	0.7021 ± 0.0425
G × I	0.4467 ± 0.0142	0.7384 ± 0.0263	0.7834 ± 0.0474	0.7115 ± 0.1017	0.6329 ± 0.0719	0.8329 ± 0.0547
IG	0.4235 ± 0.0127	0.4583 ± 0.0663	0.4522 ± 0.0458	0.4535 ± 0.0455	0.3790 ± 0.0415	0.4355 ± 0.0444
xMIL-LRP	**0.3594 ± 0.0118**	–	0.4119 ± 0.0248	–	–	–
WE	–	0.3618 ± 0.0164	–	**0.3635 ± 0.0058**	–	**0.3654 ± 0.0040**
			Pos-Neg			
Rand	0.8151 ± 0.0073	0.8172 ± 0.0087	0.8473 ± 0.0075	0.8189 ± 0.0133	0.8462 ± 0.0101	0.8425 ± 0.0085
LOO	0.3021 ± 0.0072	0.2734 ± 0.0053	0.4063 ± 0.0291	0.3740 ± 0.0254	0.3202 ± 0.0371	0.3026 ± 0.0165
G × I	0.2721 ± 0.0061	0.8520 ± 0.1520	0.7572 ± 0.0973	0.6306 ± 0.1885	0.3283 ± 0.0444	0.8983 ± 0.1137
IG	0.2515 ± 0.0099	0.2664 ± 0.0045	0.2759 ± 0.0059	0.2912 ± 0.0164	0.2564 ± 0.0066	0.2876 ± 0.0052
xMIL-LRP	0.2408 ± 0.0102	–	0.2645 ± 0.0079	–	–	–
WE	–	**0.2396 ± 0.0067**	–	**0.2641 ± 0.0093**	–	**0.2533 ± 0.0107**
			Adjacent-Pairs			
Rand	0.7621 ± 0.0249	0.7644 ± 0.0138	0.8154 ± 0.0128	0.8222 ± 0.0289	0.8379 ± 0.0077	0.8370 ± 0.0084
LOO	0.7028 ± 0.0272	0.7012 ± 0.0385	0.6726 ± 0.0244	0.7238 ± 0.0276	0.5520 ± 0.0554	0.6728 ± 0.0405
G × I	0.6548 ± 0.0374	0.7707 ± 0.0312	0.7533 ± 0.0290	0.8001 ± 0.0228	0.6095 ± 0.0755	0.8711 ± 0.0375
IG	0.5562 ± 0.0433	0.5852 ± 0.0829	0.5318 ± 0.0183	0.5841 ± 0.0309	0.4761 ± 0.0344	0.5087 ± 0.0088
xMIL-LRP	**0.4097 ± 0.0251**	–	0.5671 ± 0.0322	–	–	–
WE	–	0.4689 ± 0.0588	–	**0.4470 ± 0.0186**	–	**0.4360 ± 0.0108**

E ATTRIBUTION ON BIOMEDICAL DATA

E.1 WSI LOCALIZATION AND DELETION

Tables A13 and A14 report CAMELYON16 localization AUPRC-2 and WSI deletion AUPC, respectively.

Table A13: Lesion localization AUPRC-2 on CAMELYON16 with ResNet18 and SSL features.

	ABMIL	EC-ABMIL	TransMIL	EC-TransMIL	ILRAMIL	EC-ILRAMIL
			ResNet18-CAMELYON16			
Rand	0.1146 ± 0.0126	0.1014 ± 0.0056	0.1205 ± 0.0097	0.1205 ± 0.0094	0.1099 ± 0.0180	0.1210 ± 0.0078
LOO	0.1829 ± 0.0189	0.2382 ± 0.0298	0.1733 ± 0.0156	0.2574 ± 0.0390	0.1734 ± 0.0120	0.3144 ± 0.0691
G × I	0.3405 ± 0.0494	0.2456 ± 0.0533	0.1553 ± 0.0108	0.2355 ± 0.0519	0.3643 ± 0.0605	0.3065 ± 0.0624
IG	0.4553 ± 0.0394	0.2609 ± 0.0434	0.3149 ± 0.0389	0.3579 ± 0.0912	0.3212 ± 0.0609	0.4188 ± 0.0783
xMIL-LRP	**0.6407 ± 0.0158**	–	0.3571 ± 0.0653	–	–	–
WE	–	0.4643 ± 0.0661	–	**0.5330 ± 0.0882**	–	**0.5939 ± 0.0442**
			SSL-CAMELYON16			
Rand	0.0744 ± 0.0017	0.0731 ± 0.0008	0.0734 ± 0.0021	0.0751 ± 0.0022	0.0724 ± 0.0023	0.0729 ± 0.0019
LOO	0.2626 ± 0.0115	0.4159 ± 0.0664	0.2248 ± 0.0215	0.2972 ± 0.0214	0.2264 ± 0.0426	0.3491 ± 0.0573
G × I	0.5821 ± 0.0521	0.4204 ± 0.0668	0.2694 ± 0.0846	0.3554 ± 0.0259	0.3742 ± 0.0709	0.3962 ± 0.0597
IG	0.6938 ± 0.0355	0.4936 ± 0.0896	0.3874 ± 0.0324	0.4445 ± 0.0225	0.3764 ± 0.0469	0.4223 ± 0.0608
xMIL-LRP	**0.7655 ± 0.0279**	–	0.5633 ± 0.0274	–	–	–
WE	–	0.6439 ± 0.0462	–	**0.6804 ± 0.0437**	–	**0.6582 ± 0.0470**

Table A14: Attribution results on the ResNet18 and SSL feature settings.

	ABMIL	EC-ABMIL	TransMIL	EC-TransMIL	ILRAMIL	EC-ILRAMIL
			ResNet18-CAMELYON16			
Rand	0.8426 ± 0.0675	0.8688 ± 0.0442	0.8508 ± 0.0722	0.8207 ± 0.1073	0.8213 ± 0.0480	0.8371 ± 0.0930
LOO	0.6807 ± 0.0506	0.7311 ± 0.0596	0.6976 ± 0.0271	0.7286 ± 0.1204	0.7153 ± 0.0677	0.6335 ± 0.1247
G × I	0.7463 ± 0.0143	0.7935 ± 0.0386	0.7587 ± 0.0898	0.8267 ± 0.0978	0.6856 ± 0.0582	0.7254 ± 0.1498
IG	0.6228 ± 0.0598	0.7908 ± 0.0197	0.5578 ± 0.0250	0.7470 ± 0.1204	0.6684 ± 0.0682	0.6351 ± 0.1291
xMIL-LRP	**0.2714 ± 0.1027**	–	0.4677 ± 0.0540	–	–	–
WE	–	0.5226 ± 0.0811	–	**0.4628 ± 0.1331**	–	**0.3646 ± 0.1214**
			ResNet18-BRACS			
Rand	0.7208 ± 0.0270	0.7990 ± 0.0758	0.6987 ± 0.0940	0.7779 ± 0.0419	0.8569 ± 0.0081	0.7786 ± 0.0485
LOO	0.4829 ± 0.0324	0.3277 ± 0.1006	0.5638 ± 0.0665	0.6434 ± 0.0382	0.5599 ± 0.0357	0.2775 ± 0.0851
G × I	0.4998 ± 0.0737	0.5524 ± 0.1969	0.5560 ± 0.1177	0.7857 ± 0.0593	0.5879 ± 0.0720	0.5332 ± 0.1483
IG	0.4572 ± 0.0757	0.4121 ± 0.1280	0.4214 ± 0.0356	0.6369 ± 0.0936	0.4709 ± 0.0601	0.3425 ± 0.1484
xMIL-LRP	0.2896 ± 0.0269	–	0.3630 ± 0.0377	–	–	–
WE	–	**0.2155 ± 0.0993**	–	**0.1829 ± 0.0622**	–	**0.1714 ± 0.1331**
			SSL-CAMELYON16			
Rand	0.9489 ± 0.0228	0.9573 ± 0.0097	0.9436 ± 0.0224	0.9502 ± 0.0068	0.9539 ± 0.0143	0.9522 ± 0.0105
LOO	0.8990 ± 0.0349	0.9205 ± 0.0287	0.8685 ± 0.0316	0.9016 ± 0.0235	0.9135 ± 0.0267	0.9163 ± 0.0189
G × I	0.8341 ± 0.0755	0.9025 ± 0.0659	0.8721 ± 0.0339	0.8907 ± 0.0323	0.9081 ± 0.0361	0.9032 ± 0.0409
IG	0.7592 ± 0.0604	0.9019 ± 0.0496	0.8193 ± 0.0467	0.8195 ± 0.0345	0.9014 ± 0.0291	0.8952 ± 0.0309
xMIL-LRP	**0.4852 ± 0.1428**	–	0.6421 ± 0.0583	–	–	–
WE	–	0.6444 ± 0.0141	–	**0.5928 ± 0.0598**	–	**0.6284 ± 0.0280**
			SSL-BRACS			
Rand	0.8335 ± 0.0599	0.9136 ± 0.0131	0.8200 ± 0.0402	0.8918 ± 0.0415	0.8832 ± 0.0388	0.8847 ± 0.0254
LOO	0.7390 ± 0.0635	0.7986 ± 0.0306	0.7257 ± 0.0458	0.8134 ± 0.0517	0.7230 ± 0.0388	0.6706 ± 0.0712
G × I	0.6979 ± 0.0483	0.8978 ± 0.0341	0.7197 ± 0.0354	0.8266 ± 0.0308	0.7071 ± 0.0805	0.8279 ± 0.0873
IG	0.6477 ± 0.0477	0.8227 ± 0.0331	0.6129 ± 0.0571	0.7767 ± 0.0484	0.5954 ± 0.0432	0.7420 ± 0.0682
xMIL-LRP	0.5267 ± 0.0480	–	**0.4650 ± 0.0724**	–	–	–
WE	–	**0.5181 ± 0.0393**	–	0.5209 ± 0.0763	–	**0.4716 ± 0.0530**

E.1.1 WSI ATTRIBUTION MAPS

Figure A2 compares attribution maps for one region of CAMELYON16 slide `test_071` and two regions of `test_061` across ABMIL, TransMIL, ILRAMIL, and their EC variants. Each score corresponds to a 512×512 patch at its original slide coordinates; black contours show the supplied lesion annotations. Models share the same patch coordinates within each slide: 27,149 for `test_071` and 9,135 for `test_061`.

To display signed scores across their dynamic range, let q be the 99.5th percentile of absolute scores over the full slide for each map. A raw score s is transformed as

$$t(s) = \frac{\operatorname{asinh}(100 \ \operatorname{clip}(s/q, -1, 1))}{\operatorname{asinh}(100)}. \tag{A47}$$

Red and blue encode positive and negative scores, with overlay opacity $0.78\sqrt{|t(s)|}$; zero is transparent. This common display rule preserves signs and reveals smaller-magnitude responses while retaining the tissue image. Color strength is normalized within each map, and patches are rendered without smoothing.

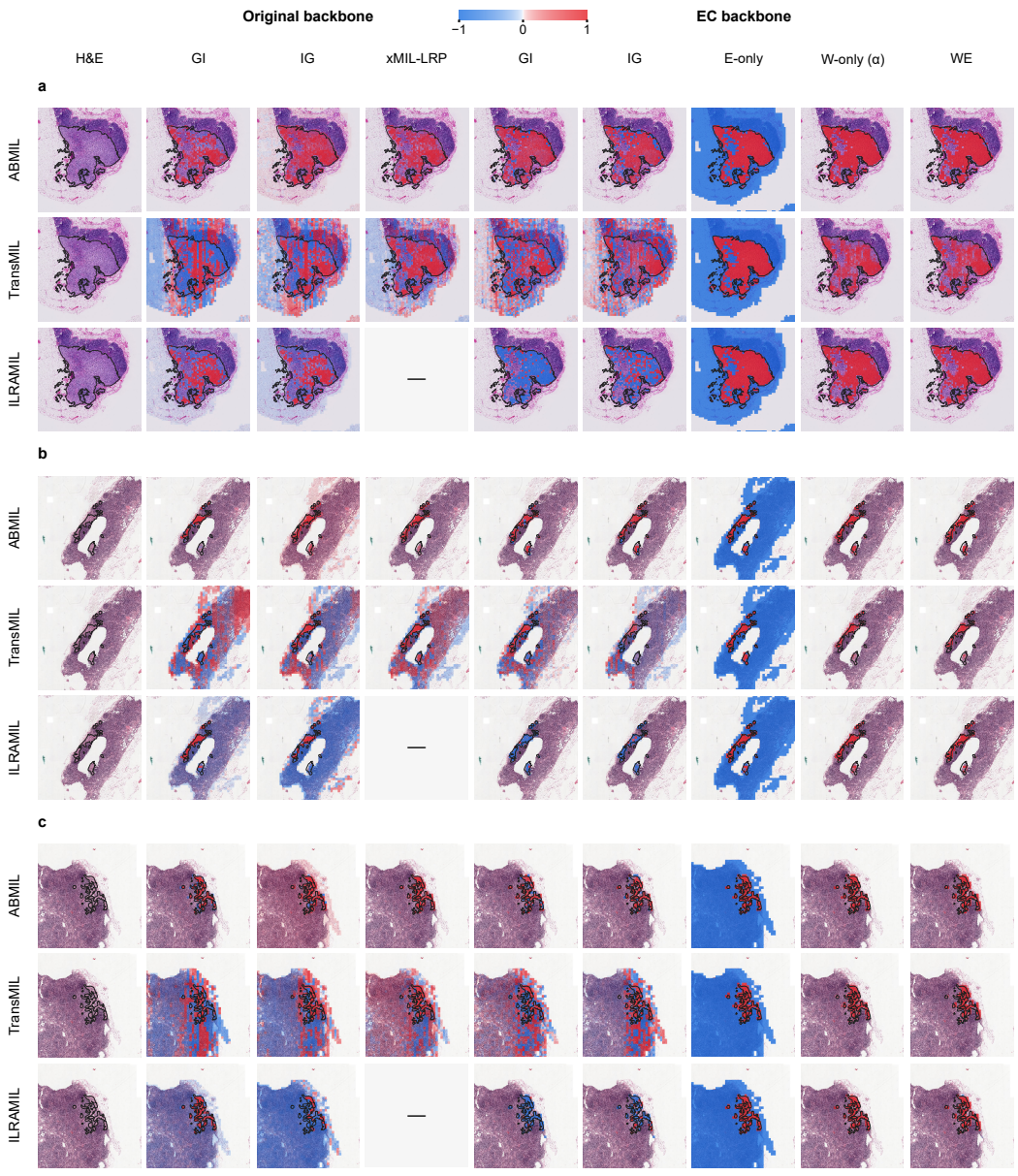


Figure A2: WSI attribution maps of 3 cases: (a)-(c). Within each region, rows show ABMIL, TransMIL, and ILRAMIL; columns compare original-backbone attributions with EC attributions and their evidence and weight components. Black contours mark lesions. All maps use Equation (A47); the dash denotes unsupported xMIL-LRP for ILRAMIL.

E.2 MUTATION-BAG ATTRIBUTION

Tables A15 and A16 report deletion AUPC and tissue-level gene rankings, respectively.

Table A15: Attribution results with scGPT and text2vec features.

	ABMIL	EC-ABMIL	TransMIL	EC-TransMIL	ILRAMIL	EC-ILRAMIL
			scGPT			
Rand	0.5531 ± 0.0046	0.6237 ± 0.0064	0.5523 ± 0.0095	0.5347 ± 0.0077	0.6372 ± 0.0127	0.6838 ± 0.0092
LOO	0.0700 ± 0.0041	0.0485 ± 0.0016	0.1412 ± 0.0175	0.1186 ± 0.0044	0.0833 ± 0.0098	0.0710 ± 0.0119
G × I	0.0240 ± 0.0006	0.2303 ± 0.0226	0.3485 ± 0.0177	0.2798 ± 0.0224	0.1419 ± 0.0374	0.1332 ± 0.0230
IG	0.0336 ± 0.0014	0.0468 ± 0.0021	0.0704 ± 0.0074	0.1106 ± 0.0336	0.0339 ± 0.0025	0.0490 ± 0.0042
xMIL-LRP	**0.0226 ± 0.0013**	–	0.0423 ± 0.0026	–	–	–
WE	–	0.0232 ± 0.0004	–	**0.0257 ± 0.0006**	–	**0.0322 ± 0.0010**
			text2vec			
Rand	0.5570 ± 0.0041	0.6240 ± 0.0052	0.5535 ± 0.0139	0.5321 ± 0.0102	0.6575 ± 0.0088	0.6691 ± 0.0115
LOO	0.0647 ± 0.0040	0.0517 ± 0.0043	0.1374 ± 0.0210	0.1219 ± 0.0024	0.1007 ± 0.0181	0.0715 ± 0.0076
G × I	0.0235 ± 0.0009	0.2383 ± 0.0366	0.3464 ± 0.0139	0.3307 ± 0.0216	0.1795 ± 0.0745	0.1195 ± 0.0118
IG	0.0361 ± 0.0028	0.0530 ± 0.0029	0.0693 ± 0.0075	0.1326 ± 0.0233	0.0425 ± 0.0078	0.0445 ± 0.0026
xMIL-LRP	**0.0225 ± 0.0010**	–	0.0418 ± 0.0033	–	–	–
WE	–	0.0230 ± 0.0007	–	**0.0257 ± 0.0005**	–	**0.0308 ± 0.0019**

Table A16: Top five genes associated with each tissue.

Tissue	Top-5 Genes
Bowel	APC, SOX9, KRAS, BRAF, TCF7L2
Breast	PIK3CA, GATA3, CDH1, FGF19, MAP3K1
CNS/Brain	IDH1, CFAP47, SEC61G, EIF3A, CCAR1
Kidney	VHL, TAMM41, PBRM1, STRADA, LMCD1
Liver	ALB, CTNNB1, AXIN1, RPS6KA3, NPAS3
Lung	EGFR, KRAS, KEAP1, ZFHX4, SLITRK2
Pancreas	KRAS, CDKN2A, MEN1, SMAD4, GATA6
Prostate	TMPRSS2, ERG, ZNF292, FOXA1, CXCR4

E.3 ADDITIONAL ATTRIBUTION PLOTS

Figure A3 shows the original signed-WE examples. Figure A4 shows the MUT deletion curves and per-bag AUPC distributions.

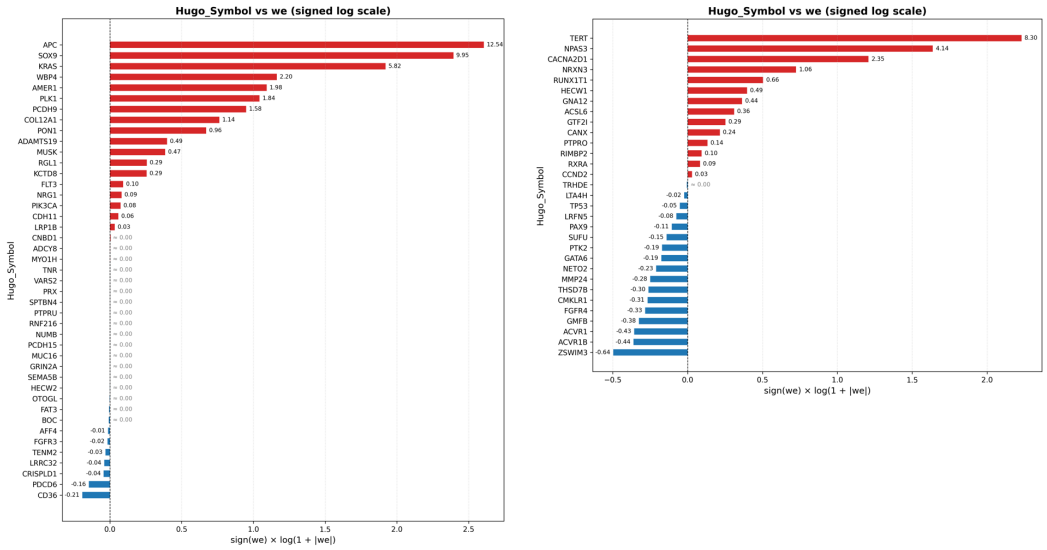


Figure A3: Original Bowel (left) and Liver (right) case plots. Red and blue bars show positive and negative WE, respectively. Bar lengths use the signed-log transform shown on the axes; numerical labels give the original WE values.

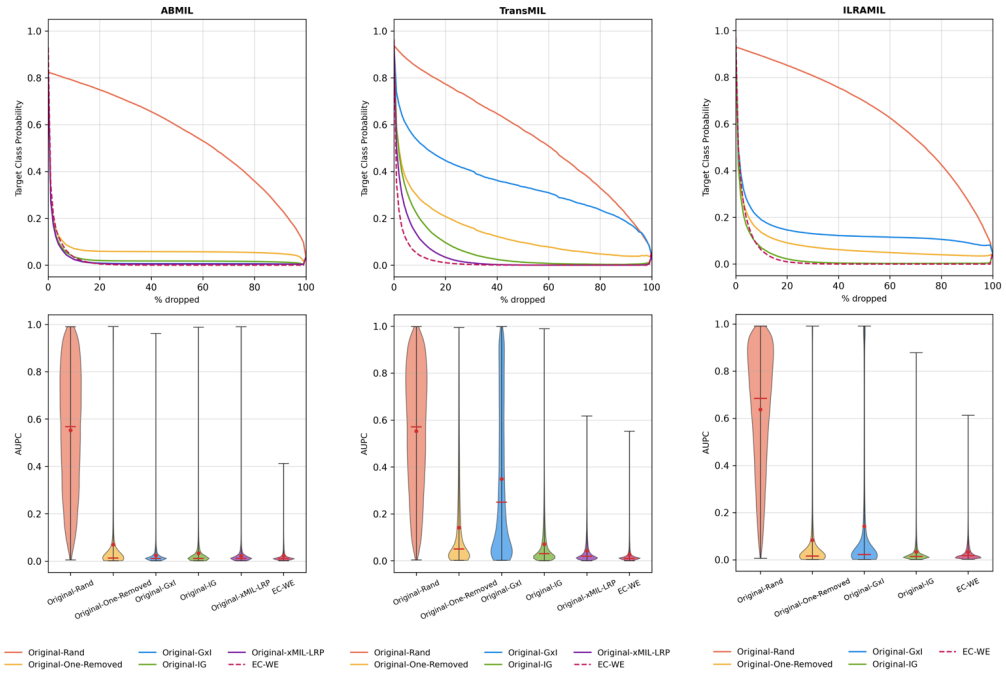


Figure A4: Original deletion curves and AUPC distributions on MUT with scGPT features for EC-ABMIL, EC-TransMIL, and EC-ILRAMIL.

F ATTRIBUTION EFFICIENCY

Table A17 reports computational costs and end-to-end attribution times on the MUT test set (seed 42; average bag size: 952.9 instances). WE uses EC backbones; post-hoc methods use the original backbones.

Table A17: Computational costs and end-to-end attribution times on MUT (seconds; seed 42).

Method	Cost	ABMIL	TransMIL	ILRAMIL
EC-WE	F	**11.3**	**47.5**	**29.9**
$G \times I$	F + B	58.3	490.3	<u>257.0</u>
IG	K(F + B)	146.6	1251.7	639.9
xMIL-LRP	F + B	<u>47.7</u>	<u>294.9</u>	–
LOO	NF	1411.0	12085.1	7064.3

G BIOMEDICAL PLAUSIBILITY OF MUT

For most of the genes mentioned, detailed information on gene–cancer associations can be obtained from the OncoKB database. For the small number of genes that are not included in OncoKB or for which OncoKB does not describe their associations with the corresponding tissues, the references listed in the table provide supporting information.

Table A18: Supporting information sources for the mentioned genes.

FGF19	https://doi.org/10.2147/ijgm.s360177
CFAP47	https://doi.org/10.3389/fimmu.2024.1324010
SEC61G	https://doi.org/10.1073/pnas.2303400120
EIF3A	https://doi.org/10.1186/s12935-019-0867-1
CCAR1	https://doi.org/10.1016/j.neo.2020.10.010
TAMM41	https://doi.org/10.1186/s13062-026-00832-1
STRADA	https://doi.org/10.1016/j.lungcan.2025.108841
LMCD1	https://doi.org/10.1038/onc.2011.440
ALB	https://doi.org/10.1016/j.cell.2017.05.046
RPS6KA3	https://doi.org/10.1016/j.jhep.2023.05.004
NPAS3	https://doi.org/10.3389/fphar.2025.1604473
MEN1	https://doi.org/10.1126/science.1200609
ZNF292	https://doi.org/10.1038/s41586-020-2135-x
ZFHX4	https://doi.org/10.1007/s10637-024-01477-5
SLITRK2	https://doi.org/10.2147/ijgm.s357659

H ADDITIVE MIL RESULTS

We evaluate the ABMIL implementation of Additive MIL (Javed et al., 2022); Patch-Logits denotes its intrinsic instance contributions.

H.1 PREDICTIVE PERFORMANCE

The following tables report the AdditiveMIL predictive results on the synthetic and MUT tasks.

Table A19: Predictive performance of AdditiveMIL in synthetic MIL tasks.

Task	ACC	F1	AUC
4-Bags	0.9424 ± 0.0053	0.9422 ± 0.0054	0.9916 ± 0.0015
Pos-Neg	0.8642 ± 0.0173	0.8631 ± 0.0182	0.9492 ± 0.0063
Adjacent-Pairs	0.7596 ± 0.0707	0.7458 ± 0.0670	0.7803 ± 0.0592

Table A20: Predictive performance of AdditiveMIL on MUT tasks.

Features	ACC	F1	AUC
scGPT	0.7035 ± 0.0090	0.5674 ± 0.0084	0.9442 ± 0.0033
text2vec	0.7236 ± 0.0044	0.5942 ± 0.0097	0.9500 ± 0.0022

H.2 SYNTHETIC EVIDENCE RECOVERY AND DELETION

The following tables report AdditiveMIL attribution and deletion results on the synthetic MIL tasks.

Table A21: AUPRC-2 attribution results for AdditiveMIL in the 4-Bags, Pos-Neg, and Adjacent-Pairs settings.

Method	4-Bags	Pos-Neg	Adjacent-Pairs
Rand	0.3115 ± 0.0022	0.4144 ± 0.0043	0.5443 ± 0.0078
LOO	0.6297 ± 0.0176	0.8788 ± 0.0069	$\mathbf{0.6552 \pm 0.0308}$
Patch-Logits	$\mathbf{0.9053 \pm 0.0370}$	0.8891 ± 0.0145	0.6027 ± 0.0593
$G \times I$	0.8460 ± 0.0352	0.8939 ± 0.0082	0.5431 ± 0.0758
IG	0.8599 ± 0.0370	$\mathbf{0.9008 \pm 0.0085}$	0.6244 ± 0.0310

Table A22: AUROC-2 attribution results for AdditiveMIL in the 4-Bags, Pos-Neg, and Adjacent-Pairs settings.

Method	4-Bags	Pos-Neg	Adjacent-Pairs
Rand	0.4996 ± 0.0038	0.4997 ± 0.0034	0.5016 ± 0.0031
LOO	0.5636 ± 0.0172	0.8787 ± 0.0049	$\mathbf{0.6087 \pm 0.0399}$
Patch-Logits	$\mathbf{0.8997 \pm 0.0453}$	0.9088 ± 0.0131	0.5035 ± 0.0528
$G \times I$	0.8244 ± 0.0433	0.9022 ± 0.0078	0.3999 ± 0.0820
IG	0.8418 ± 0.0473	$\mathbf{0.9138 \pm 0.0075}$	0.5153 ± 0.0275

Table A23: AUPC results for AdditiveMIL in the 4-Bags, Pos-Neg, and Adjacent-Pairs settings.

Method	4-Bags	Pos-Neg	Adjacent-Pairs
Rand	0.7745 ± 0.0112	0.7911 ± 0.0145	0.6510 ± 0.0338
LOO	0.6412 ± 0.0434	0.3216 ± 0.0143	0.5773 ± 0.0265
Patch-Logits	$\mathbf{0.3464 \pm 0.0169}$	$\mathbf{0.2599 \pm 0.0190}$	$\mathbf{0.5767 \pm 0.0956}$
$G \times I$	0.4506 ± 0.0317	0.2914 ± 0.0160	0.6102 ± 0.0483
IG	0.4191 ± 0.0322	0.2744 ± 0.0200	0.5810 ± 0.0728

H.3 MUT DELETION

The following table reports AdditiveMIL deletion results on the MUT tasks.

Table A24: AUPC results for AdditiveMIL on MUT tasks with scGPT and text2vec features.

Method	scGPT	text2vec
Rand	0.5495 ± 0.0136	0.5577 ± 0.0066
LOO	0.1168 ± 0.0204	0.1148 ± 0.0202
Patch-Logits	$\mathbf{0.0190 \pm 0.0010}$	$\mathbf{0.0193 \pm 0.0007}$
$G \times I$	$\underline{0.0226 \pm 0.0014}$	$\underline{0.0229 \pm 0.0008}$
IG	0.0297 ± 0.0022	0.0314 ± 0.0022

H.4 WSI LOCALIZATION AND DELETION

The following table reports AdditiveMIL attribution and deletion results on CAMELYON16 with ResNet18 features.

Table A25: Attribution and deletion results for AdditiveMIL on CAMELYON16 with ResNet18 features.

Method	AUPRC-2 ↑	AUROC-2 ↑	AUPC ↓
Rand	0.1108 ± 0.0232	0.4988 ± 0.0099	0.8270 ± 0.0685
LOO	0.1203 ± 0.0217	0.5069 ± 0.0069	0.7971 ± 0.0816
$G \times I$	0.3390 ± 0.0445	$\underline{0.5753 \pm 0.0257}$	0.7651 ± 0.0985
IG	$\mathbf{0.4540 \pm 0.0599}$	0.5727 ± 0.0509	$\underline{0.7373 \pm 0.1054}$
Patch-Logits	$\underline{0.4063 \pm 0.0716}$	$\mathbf{0.7087 \pm 0.0524}$	$\mathbf{0.5120 \pm 0.2233}$