TILTMIL: LEARNING FINITE STATISTICS FROM FUNCTIONAL BAG REPRESENTATIONS FOR MULTIPLE INSTANCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple instance learning (MIL) predicts bag-level labels from multiple instances using only bag-level supervision and is widely applied to whole-slide image analysis in computational pathology. However, existing MIL methods typically compress an instance set directly into a finite-dimensional representation whose retained information form is predetermined by the representation mechanism, although different tasks may require different forms of compositional information. We propose TiltMIL, a Bag composition modeling framework based on finite cumulant-generating function (CGF) readouts. TiltMIL represents a bag as an empirical probability measure in a task-specific feature space, uses the CGF as a functional distributional representation, and learns task-relevant tilt locations to extract finite CGF function values and cumulant statistics for prediction. A lightweight hierarchical fusion mechanism combines these readouts, while empirical perturbation consistency improves stability to finite composition changes. Experiments on three WSI-MIL benchmarks with two feature encoders demonstrate consistent performance improvements across multiple pathology classification tasks.

1 INTRODUCTION

Multiple instance learning (MIL) aims to predict a label for a *bag*, defined as a collection of multiple *instances*, when supervision is available only for the bag as a whole rather than for individual instances. (Dietterich et al., 1997; Carbonneau et al., 2018). Digital pathology is a prominent application of MIL. In this setting, a whole-slide image (WSI) is treated as a bag and partitioned into thousands or even tens of thousands of patches, each serving as an instance, while supervision is typically provided only at the slide level. Because disease-related regions may occupy only a small fraction of the tissue, bag-level prediction requires identifying and integrating diagnostically relevant information across a large number of patches. (Campanella et al., 2019; Lu et al., 2021)

Existing MIL methods largely focus on how to combine a large set of instances into a bag representation for prediction . Classical approaches typically use fixed pooling operators such as max or mean pooling, or attention-based weighted aggregation (Ilse et al., 2018). More recent methods retain more information about the distribution of instances within a bag, for example by modeling marginal feature distributions (Oner et al., 2023), characterizing patch-score distributions at multiple locations (Li et al., 2022), or describing a bag using a probability distribution parameterized by its mean and variance (Zhao et al., 2025). However, different pathology tasks may rely on different aspects of a bag. For example, tumor classification can be driven primarily by whether a localized lesion is present (Campanella et al., 2019); Gleason grading of prostate cancer additionally depends on the proportions of different tissue patterns (Epstein et al., 2016; Bulten et al., 2022) and survival prediction may require integrating evidence distributed across multiple tissue regions (Yao et al., 2020). These examples illustrate that different tasks place different demands on a bag representation. Yet in existing approaches, the form in which a bag is ultimately summarized is largely determined when the aggregation or representation mechanism is chosen.

This motivates a different modeling strategy: rather than compressing a bag into a fixed statistical form at the outset, we first construct a richer representation of the bag and then decide which information should be used for prediction. Specifically, we represent a bag by the empirical probability

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.

measure of its instances in a learned feature space (Doran & Ray, 2016), and use the cumulant-generating function (CGF) to represent this empirical distribution. The CGF is a function of a continuous parameter; changing this parameter emphasizes different parts of the same bag by exponentially reweighting the instance masses. For a finite empirical measure, the full CGF uniquely determines the empirical distribution, so no particular finite statistical summary needs to be chosen at this stage. A practical predictor, however, must ultimately operate on a finite-dimensional representation. This leads to the central question: *which finite statistics should be read from this functional representation for a given prediction task?*

To address this question, we propose **TiltMIL**, which learns a finite, task-relevant statistical representation from the CGF of a bag. Its key design separates *what to read* from *where to read*: the model extracts a small set of explicit statistics from the CGF, including function values and first- and second-order cumulants, while the corresponding tilt locations are learned end-to-end from bag-level supervision. Different tilt locations reweight the same empirical distribution in different ways, allowing the model to extract different information from the same bag. TiltMIL also retains an untilted base statistic and combines the resulting finite readouts through a lightweight fusion mechanism for bag-level prediction. During training, we further encourage consistent predictions under mild perturbations of the bag to improve stability to finite instance changes. We study TiltMIL through both controlled experiments and real WSI-MIL benchmarks. The controlled tasks use explicit bag-level decision rules to examine what information different tasks require from a bag and how predictions respond to changes in bag content. We then evaluate TiltMIL across multiple pathology datasets and feature encoders, and analyze the contribution of its different finite statistical readouts. Our main contributions are as follows:

1. We revisit MIL from the perspective of finite-statistics learning. Instead of directly compressing an instance set into a fixed-dimensional bag representation, we formulate MIL as first constructing a functional representation of the empirical instance distribution and then learning task-relevant finite statistics for bag-level prediction.
2. Based on this principle, we propose **TiltMIL**, which separates *what statistics to read* from *where to read them*: the statistical readout types are explicitly defined, while the corresponding tilt locations are learned from bag-level supervision.
3. We evaluate TiltMIL through controlled MIL tasks and real WSI-MIL benchmarks. TiltMIL effectively models different bag-level decisions and achieves consistent performance improvements across pathology tasks and feature encoders.

2 Understanding Composition-Dependent Decisions in MIL

Different MIL tasks may rely on different information from a bag, while the aggregation mechanism determines how instance-level information is combined into the final bag representation. To examine this process under controlled conditions, we construct MIL tasks with explicit bag-level decision rules, such that the information required by each task is known by design. We then systematically modify bag composition and analyze how the predictions of a direct aggregation model change in response.

2.1 Experimental Design

Controlled MIL tasks. To compare the information requirements induced by different bag-level decision rules, we construct three controlled tasks with explicit generative mechanisms: **Presence**, **Co-occurrence** and **Competitive** tasks. The **Presence** task follows the classical MIL assumption (Carbonneau et al., 2018), where the bag label depends only on whether a target instance type $\mathcal{A}$ is present. Let $n_{\mathcal{A}}$ denote the number of $\mathcal{A}$ instances in the bag. The bag label is 1 when $n_{\mathcal{A}} > 0$ and 0 otherwise, i.e., $y = \mathbb{I}[n_{\mathcal{A}} > 0]$, and therefore requires only instance-presence information. Building on this setting, the **Co-occurrence** task introduces a second decision-relevant instance type $\mathcal{B}$ and assigns each bag to one of four classes, neither target type, $\mathcal{A}$-only, $\mathcal{B}$-only, or $\mathcal{A} + \mathcal{B}$, such that prediction depends on the joint presence of different evidence types. The **Competitive** task further defines the label according to the relative composition of the two types, $y = \mathbb{I}[n_{\mathcal{A}} > n_{\mathcal{B}}]$. These tasks progress from instance presence to co-occurrence and relative composition,

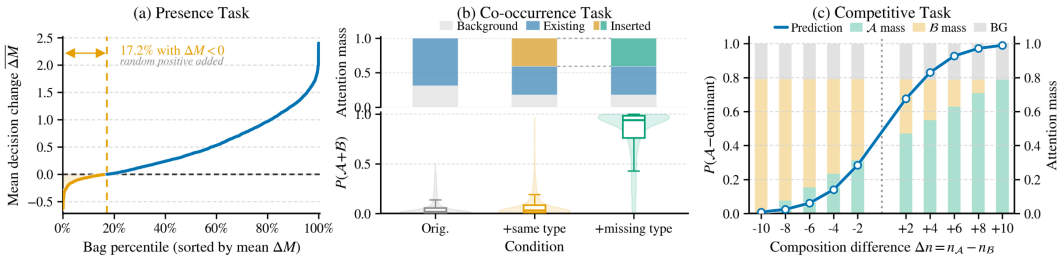


Figure 1: Controlled composition analysis across three MIL tasks. (a) Positive-instance addition can decrease the bag-level margin. (b) Similar attention mass can correspond to different prediction responses. (c) Prediction tracks relative composition as attention mass redistributes across $\mathcal{A}$, $\mathcal{B}$, and background.

providing a controlled spectrum of bag-level information requirements. Additional data construction and training details are provided in the Appendix A.1.

Analysis baseline. For a bag $B = \{\mathbf{x}_i\}_{i=1}^{N}$, a frozen feature encoder g maps each instance to $\mathbf{h}_i = g(\mathbf{x}_i)$. We use ABMIL (Ilse et al., 2018) as a representative direct aggregation model because its explicitly normalized weights allow the effect of composition edits on both the aggregation and the final decision to be analyzed in closed form. Specifically, a scoring network $s(\cdot)$ produces normalized instance weights $a_i = \frac{\exp s(\mathbf{h}_i)}{\sum_{j=1}^{N} \exp s(\mathbf{h}_j)}$ and computes the bag representation $\mathbf{z} = \sum_{i=1}^{N} a_i \mathbf{h}_i$, after which a bag-level classifier maps $\mathbf{z}$ to the final prediction.

Controlled composition edits. After training the ABMIL, we design simple task-specific edits and examine how the resulting predictions change. For the Presence task, we examine how repeated additions of target instances affect correctly classified positive bags. We summarize each bag's response by the mean change in its positive-versus-negative logit margin, $\overline{\Delta M}(B)$, across separate target-instance edits. The detailed protocol is provided in Appendix A.2. For an $\mathcal{A}$-only bag in the Co-occurrence task, we compare two equal-sized edits: adding another five $\mathcal{A}$ instances or adding the missing $\mathcal{B}$ instances; the construction is symmetric for $\mathcal{B}$-only bags. For the Competitive task, we sweep the relative composition $\Delta n = n_{\mathcal{A}} - n_{\mathcal{B}}$ and track the corresponding prediction response.

2.2 Results and Observations

Figure 1 summarizes the composition-editing results for the three controlled tasks. ABMIL reliably learns the corresponding bag-level decision rules, with full predictive results reported in Appendix A.1. We further examine how local and global changes in bag content are reflected in the final decision through the aggregation process.

The effect of a local composition change depends on the current bag state. As shown in Figure 1(a), although every added instance belongs to the target type $\mathcal{A}$, 17.2% of the analyzed positive bags satisfy $\overline{\Delta M} < 0$ after editing. The same target type can therefore strengthen or weaken the current decision when added to different bags. Its effect on the prediction is determined jointly by the added instance and the bag state already formed through aggregation.

Equal-sized composition changes can induce different decision responses. The Co-occurrence task provides a further example of this dependence. As shown in Figure 1(b), for a bag containing only one target type, added instances of the existing type and the missing type can receive similar aggregation weights. However, their decision effects differ substantially: adding the existing type leaves $P(\mathcal{A} + \mathcal{B})$ nearly unchanged, whereas adding the missing type markedly increases it. This shows that aggregation weight alone does not determine the effect of a local composition change. The effect also depends on the information already present in the bag.

Relative counts can affect the decision through cumulative aggregation. In the Competitive task, as $\Delta n = n_{\mathcal{A}} - n_{\mathcal{B}}$ moves from the $\mathcal{B}$-dominant to the $\mathcal{A}$-dominant regime, $P(\mathcal{A}\text{-dominant})$ changes systematically, as shown in Figure 1(c). Meanwhile, $\mathcal{A}$ and $\mathcal{B}$ instances receive similar average per-instance aggregation weights throughout the sweep. Their distinction in the bag representation is therefore driven largely by their counts and the resulting cumulative contribution. This further

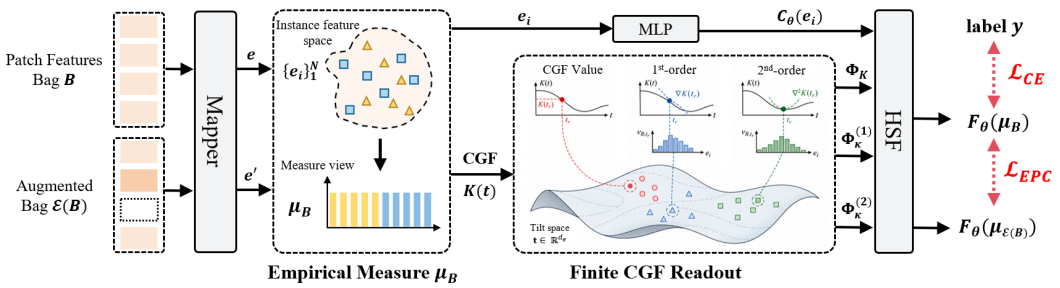


Figure 2: Overview of TiltMIL. The empirical Bag measure is represented through the CGF and queried at learned tilt locations to obtain function-value, first-order, and second-order statistical readouts, which are fused for Bag-level prediction with empirical perturbation consistency. HSF denotes Hierarchical Statistical Fusion, described in Sec. 3.2.

shows that an individual aggregation weight does not by itself determine the role of an instance in the current bag-level decision.

Analysis. These observations admit a unified explanation under the normalized aggregation form of ABMIL (Ilse et al., 2018). For a pair of classes c and k, let $M_{c,k}(B) = \boldsymbol{\beta}_{c,k}^\top \mathbf{z}_B + \Delta b_{c,k}$ denote the pairwise decision margin, where $\boldsymbol{\beta}_{c,k} = \mathbf{w}_c - \mathbf{w}_k$, and define the corresponding instance-level decision quantity as $p_i^{c,k} = \boldsymbol{\beta}_{c,k}^\top \mathbf{h}_i + \Delta b_{c,k}$. Consider inserting one additional copy of instance $\mathbf{x}_i$ into the bag. The resulting margin change is

$$M_{c,k}(B \cup \{\mathbf{x}_i\}) - M_{c,k}(B) = \frac{a_i}{1 + a_i} \left[p_i^{c,k} - M_{c,k}(B) \right]. \tag{1}$$

The aggregation weight a_i controls the magnitude of the change, while its direction depends on the added instance relative to the current bag margin. This relation explains both the negative edit responses in the Presence task and the different effects of similarly weighted edits in the Co-occurrence task. More generally, under this form of direct aggregation, the effect of local instance evidence depends jointly on the instance itself and the current bag state. The full derivation is provided in Appendix A.3.

3 Method: Bag Composition Modeling via Finite CGF Readouts

The controlled study shows that the aggregation process determines both how instances are weighted and how their information is combined with the evidence already present in the bag. Consequently, the effect of a local change depends on the bag state formed through aggregation. Together with the different information requirements of bag-level decision rules, this suggests that the chosen aggregation mechanism directly shapes how bag information is organized before prediction. Motivated by this observation, we separate the representation of a bag from the finite statistical readout used for prediction. As shown in Fig. 2, TiltMIL first represents a bag as an empirical probability measure in a learned feature space and uses its cumulant-generating function (CGF) as a functional representation of the empirical distribution. It then learns a finite set of task-relevant statistical readouts from this functional representation for bag-level prediction.

3.1 Bag Composition and Cumulant-Generating Representation

Empirical bag measure. Given a bag containing N instances, a frozen pretrained encoder first extracts instance features $\{\mathbf{h}_i\}_{i=1}^N$. A shared mapping $E_{\boldsymbol{\theta}}$ then maps each $\mathbf{h}_i$ to a task-relevant representation $\mathbf{e}_i = E_\theta(\mathbf{h}_i) \in \mathbb{R}^{d_e}$. We represent the composition of the current bag in this representation space as an empirical probability measure Muandet et al. (2017); Peyré & Cuturi (2019):

$$\mu_{B,\theta} = \sum_{i=1}^N m_i \delta_{\mathbf{e}_i}, \qquad m_i \geq 0, \qquad \sum_{i=1}^N m_i = 1, \tag{2}$$

where $\delta_{\mathbf{e}_i}$ denotes the Dirac measure at $\mathbf{e}_i$, and m_i is the probability mass associated with instance $\mathbf{e}_i$ in the current bag. By default, $m_i = 1/N$. When there is no ambiguity, we write μ_B as shorthand for $\mu_{B,\theta}$.

Cumulant-generating representation. To avoid compressing the bag into a predefined finite set of statistics at the outset, we further use the cumulant-generating function (CGF) of the empirical measure to construct a functional representation of the current bag composition. For the finite empirical measure considered here, the full CGF uniquely determines the corresponding distribution. We therefore regard the full CGF as a functional representation of the empirical composition of the current bag. For any tilt parameter $\mathbf{t}$, the CGF is defined as

$$K(\mathbf{t}; \mu_B) = \log \mathbb{E}_{\mathbf{e}\sim\mu_B}\left[\exp\left(\mathbf{t}^\top \mathbf{e}\right)\right] = \log \sum_{i=1}^{N} m_i \exp\left(\mathbf{t}^\top \mathbf{e}_i\right), \qquad \mathbf{t} \in \mathbb{R}^{d_e}. \tag{3}$$

Here, each $\mathbf{t}$ defines a reweighting of the instance masses within the same bag, which we refer to as exponential tilting. For an instance $\mathbf{e}_i$, the inner product $\mathbf{t}^\top \mathbf{e}_i$ gives its tilt score: a larger score assigns the instance greater probability mass after reweighting. Let $m_i(\mathbf{t})$ denote the reweighted probability mass of instance $\mathbf{e}_i$ under tilt $\mathbf{t}$. Specifically,

$$m_i(\mathbf{t}) = m_i \exp\left(\mathbf{t}^\top \mathbf{e}_i - K(\mathbf{t}; \mu_B)\right), \quad \nu_{B,\mathbf{t}} = \sum_{i=1}^{N} m_i(\mathbf{t})\, \delta_{\mathbf{e}_i}, \quad \text{hence } \nu_{B,\mathbf{0}} = \mu_B. \tag{4}$$

Thus, different values of $\mathbf{t}$ induce different exponentially tilted measures $\nu_{B,\mathbf{t}}$ of the same empirical measure μ_B by changing the allocation of instance masses while preserving the same support.

Another key property of the CGF is that its derivatives directly correspond to statistics under the associated tilted measure. Its first-order derivative gives the mean of the tilted distribution, while its second-order derivative gives the corresponding covariance:

$$\nabla_{\mathbf{t}} K(\mathbf{t}; \mu_B) = \mathbb{E}_{\nu_{B,\mathbf{t}}}[\mathbf{e}], \qquad \nabla^2_{\mathbf{t}} K(\mathbf{t}; \mu_B) = \mathrm{Cov}_{\nu_{B,\mathbf{t}}}(\mathbf{e}). \tag{5}$$

Therefore, at any tilt location $\mathbf{t}$, the CGF provides its function value together with the first- and second-order statistics of the corresponding tilted measure.

3.2 Learnable Tilting and Finite Statistical Representation

Finite CGF readout. The full CGF $K(\cdot; \mu_B)$ is defined over a continuous space of tilt parameters, whereas a practical bag-level predictor must operate on a finite-dimensional representation. We therefore construct a *finite CGF readout* by observing the CGF at a finite number of tilt locations and collecting the statistics obtained at these locations. This construction involves two related choices: which statistics to retain, and at which tilt locations to observe them. We separate these two choices in our model. The statistical readout types are specified by model design, whereas the corresponding tilt locations are learnable parameters optimized end-to-end with bag-level supervision. In this way, each readout retains a predefined statistical meaning while its observation location can adapt to the prediction task. Concretely, we use two types of finite readouts. The first directly reads the CGF function value at a selected tilt location. The second retains cumulant statistics of the corresponding tilted measure, from the first through order Q, following the differential relations in Eq. (5). Let $\mathcal{R}_K$ denote the index set of function-value readouts and $\mathcal{R}_Q$ the index set of cumulant readouts retained through order Q. We define

$$\begin{aligned} \Phi_K(\mu_B) &= \mathrm{cat}_{r\in\mathcal{R}_K}\left[\beta_r + \tau_K K(\mathbf{t}_r; \mu_B)\right], \\ \Phi^{(Q)}_{\kappa}(\mu_B) &= \mathrm{cat}_{r\in\mathcal{R}_Q}\left[\boldsymbol{\kappa}_{1:Q}\left(P_Q(\mathbf{e}); \nu_{B,\mathbf{t}_r}\right)\right]. \end{aligned} \tag{6}$$

The tilt parameters $\mathbf{t}_r$ associated with different readout types are parameterized and learned separately, allowing different statistic types and orders to be extracted at different tilt locations. Here, τ_K is a scale coefficient, β_r is an affine bias for the corresponding channel, and P_Q is a lightweight affine value map shared within the same order. The vector $\boldsymbol{\kappa}_{1:Q}$ contains coordinate-wise cumulants from first through order Q under the same tilted measure and value space. We use $Q \in \{1, 2\}$, with

$$\begin{aligned} \boldsymbol{\kappa}_{1:1}\left(P_1(\mathbf{e}); \nu_{B,\mathbf{t}_r}\right) &= \mathbb{E}_{\nu_{B,\mathbf{t}_r}}[P_1(\mathbf{e})], \\ \boldsymbol{\kappa}_{1:2}\left(P_2(\mathbf{e}); \nu_{B,\mathbf{t}_r}\right) &= \mathrm{cat}\left[\mathbb{E}_{\nu_{B,\mathbf{t}_r}}[P_2(\mathbf{e})]\,; \mathrm{Var}_{\nu_{B,\mathbf{t}_r}}[P_2(\mathbf{e})]\right]. \end{aligned} \tag{7}$$

where $\mathrm{Var}_{\nu_{B,\mathbf{t}_r}}$ denotes coordinate-wise variance. Accordingly, Eq. (6) yields a first-order cumulant readout $\Phi_\kappa^{(1)}$ for $Q = 1$ and a cumulant readout retained through second order $\Phi_\kappa^{(2)}$ for $Q = 2$. The former consists of projected means under different tilted views, whereas the latter retains both projected means and coordinate-wise variances in its corresponding value space. We explicitly retain cumulant statistics only through second order to capture distributional variation while controlling representation size and estimation cost for bags. Together, Φ_K, $\Phi_\kappa^{(1)}$, and $\Phi_\kappa^{(2)}$ form the finite task-relevant representation extracted from the full CGF.

Untilted composition statistic. The finite readouts above extract task-relevant information from learnable tilt locations. To preserve the original empirical composition as a reference, we additionally construct a base composition statistic under the untilted measure $\nu_{B,\mathbf{0}} = \mu_B$. Specifically, we apply a lightweight nonlinear mapping C_θ to the task-relevant instance representation and take its expectation under the original empirical measure:

$$c_0(\mu_B) = \mathbb{E}_{\mu_B}[C_\theta(\mathbf{e})] = \sum_{i=1}^{N} m_i C_\theta(\mathbf{e}_i). \quad (8)$$

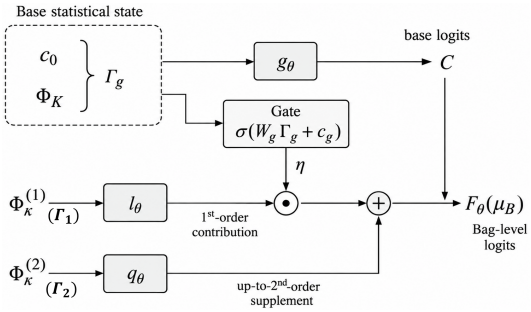


Figure 3: Detailed architecture of the Hierarchical Statistical Fusion (HSF) module.

Here, C_θ maps instance representations into the base composition-statistic space. The statistic $c_0(\mu_B)$ retains the original bag composition before exponential tilting and complements the finite tilted readouts in the subsequent decision function.

Hierarchical Statistical Fusion. The readouts above capture complementary levels of compositional information. As shown in Fig. 3 the untilted statistic c_0 describes the base composition of the original bag, whereas Φ_K aggregates CGF responses at finite tilt locations; we treat their concatenation as the base statistical state of the current bag. First-order tilted statistics further characterize distributional centers under different tilted views, while statistics retained through second order additionally capture distributional variation. We denote

$$\Gamma_g = \mathrm{cat}[c_0(\mu_B); \Phi_K(\mu_B)], \qquad \Gamma_1 = \Phi_\kappa^{(1)}(\mu_B), \qquad \Gamma_2 = \Phi_\kappa^{(2)}(\mu_B). \quad (9)$$

Based on this hierarchy, we use a lightweight hierarchical fusion mechanism: the base state produces the primary prediction and a class-wise gate that modulates the contribution of the first-order readout, while the statistics retained through second order provide a supplementary residual contribution to the final logits:

$$F_\theta(\mu_B) = g_\theta(\Gamma_g) + [\eta \odot \sigma(W_g\Gamma_g + c_g)] \odot l_\theta(\Gamma_1) + \lambda_m q_\theta(\Gamma_2), \quad F_\theta(\mu_B) \in \mathbb{R}^C. \quad (10)$$

Here, g_θ produces class logits from the base statistical state. The sigmoid-transformed term $W_g\Gamma_g + c_g$ yields a class-wise gate, which, together with the learnable scale η, modulates the contribution of $l_\theta(\Gamma_1)$. The mapping $q_\theta(\Gamma_2)$ converts the statistics retained through second order into supplementary logits. To keep the additional model complexity small, both g_θ and l_θ are shallow MLPs, the gate consists of a single linear layer followed by a sigmoid, and q_θ consists only of LayerNorm and a single linear layer.

3.3 Bag-Level Supervised Optimization

Given a bag B and its label y, the model outputs bag-level logits $F_\theta(\mu_B)$ from the empirical measure μ_B. Because the tilt locations in the finite CGF readouts are learned jointly from bag-level supervision, we further regularize the resulting predictor against small perturbations of the finite empirical measure. Specifically, we apply random instance dropout and mild feature noise during training and enforce prediction consistency before and after perturbation. Let $\mathcal{E}(B)$ denote the randomly augmented bag and $\mu_{\mathcal{E}(B)}$ its corresponding empirical measure. The final training objective is

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{EPC}}\mathcal{L}_{\mathrm{EPC}} \\ &= \mathrm{CE}(F_\theta(\mu_B), y) + \lambda_{\mathrm{EPC}}\mathbb{E}_{\mathcal{E}}\left[\frac{1}{C}\left\|F_\theta(\mu_B) - F_\theta\big(\mu_{\mathcal{E}(B)}\big)\right\|_2^2\right],\end{aligned} \quad (11)$$

where $\mathcal{L}_{\text{cls}}$ is the bag-level classification loss, $\mathcal{L}_{\text{EPC}}$ is the empirical perturbation consistency (EPC) loss, C is the number of classes, and λ_{EPC} controls the strength of the consistency term.

4 Related Work

MIL aggregation for computational pathology. Classical MIL summarizes a variable-size bag using fixed pooling or learned instance weighting. ABMIL (Ilse et al., 2018) introduced trainable permutation-invariant attention pooling, followed by methods that enrich bag aggregation through instance clustering (Lu et al., 2021), dual-stream instance–bag modeling (Li et al., 2021), Transformer-based contextualization (Shao et al., 2021), and other expressive aggregation mechanisms (Zhang et al., 2022; 2024; Li et al., 2024). Despite their architectural differences, these methods primarily learn direct mappings from an instance set to a finite bag representation.

Distribution-aware bag representations. Recent work explicitly retains distributional information beyond a single pooled embedding (Doran & Ray, 2016). DeepDisMISL (Li et al., 2022) samples multiple locations of the patch-score distribution, while distribution-based MIL pooling (Oner et al., 2023) models marginal feature distributions. Such methods preserve richer bag statistics, but the retained statistical form is still determined by the representation design. TiltMIL instead represents the empirical bag composition through its CGF and performs finite task-adaptive readout: the statistical quantities are specified, while their tilt locations are learned from bag-level supervision.

5 Experiments

We evaluate TiltMIL on three whole-slide image benchmarks: PANDA (Bulten et al., 2022) for six-class prostate cancer grading, TCGA-NSCLC (Weinstein et al., 2013) for binary lung cancer subtype classification, and TCGA-RCC (Weinstein et al., 2013) for three-class renal carcinoma subtype classification. For each dataset, we use two frozen feature encoders, ResNet-50 and UNI (Chen et al., 2024). For controlled comparison, all MIL models use identical protocol, and we report the mean and standard deviation across folds for accuracy, macro-AUC, and macro-F1. We compare TiltMIL against representative and recent MIL baselines, including ABMIL (Ilse et al., 2018),CLAM-SB and CLAM-MB (Lu et al., 2021), DSMIL (Li et al., 2021), TransMIL (Shao et al., 2021), DTFD (Zhang et al., 2022), ACMIL (Zhang et al., 2024), WiKG (Li et al., 2024), 2DMamba (Zhang et al., 2025), and PTCMIL (Zhao et al., 2026). Implementation details and hyperparameters are provided in Appendix B.1.

5.1 Main Results

Overall performance. Table 1 compares TiltMIL with ten representative MIL baselines across three pathology tasks and two feature encoders. TiltMIL performs consistently across all settings: with ResNet-50 features, it achieves the best results on all three metrics for PANDA and TCGA-NSCLC and the best AUC and macro-F1 on TCGA-RCC; with UNI features, it achieves the best accuracy and macro-F1 on all three datasets while remaining competitive in AUC. The gains are most pronounced on PANDA, where TiltMIL improves over the strongest competing results by up to 1.91 points with ResNet-50 and 5.03 points with UNI. The substantial improvement under UNI further indicates that the proposed bag representation remains complementary to stronger instance-level representations rather than being tied to a specific feature encoder.

Ablation study. Table 2 evaluates the contributions of the finite statistical readouts (Eq. 10) and EPC regularization (Eq. 11). Starting from the base state Γ_g, the rows $+\Gamma_1$ and $+\Gamma_2$ progressively introduce the first-order and up-to-second-order tilted statistics defined in Eq. 7. The first-order readout alone brings limited changes, whereas incorporating the up-to-second-order statistics yields a clear gain on PANDA but slightly lowers performance on TCGA-NSCLC. Adding EPC then gives the *Full* model and further improves accuracy and macro-F1 on both datasets.

The removal studies further clarify the roles of the two readout components. *Full w/o* Φ_K removes the CGF function-value readout in Eq. 6 and causes a marked performance drop, especially on TCGA-NSCLC, while *Full w/o Var* removes the variance term in Eq. 7 and reduces accuracy and macro-F1 on both tasks. These results support the contributions of CGF function values and variance

Table 1: **Overall results.** Performance comparison on PANDA, TCGA-NSCLC, and TCGA-RCC under 5-fold cross-validation with seed 2026. **Bold** numbers indicate the best results, and underlined numbers indicate the second-best results within each feature setting.

	Methods	PANDA			TCGA-NSCLC			TCGA-RCC		
		Acc	AUC	F1-score	Acc	AUC	F1-score	Acc	AUC	F1-score
ResNet-50	ABMIL'18	61.62±1.00	88.56±0.36	55.83±1.15	85.25±2.78	93.56±2.04	85.17±2.84	89.39±3.81	97.44±1.37	88.12±5.19
	CLAM-SB'21	62.33±0.26	88.65±0.32	56.17±1.08	85.71±3.67	93.49±2.41	85.69±3.70	89.34±2.90	97.72±0.88	87.91±3.82
	CLAM-MB'21	63.24±0.65	89.00±0.44	57.60±1.00	84.91±5.09	93.09±2.50	84.72±5.41	90.43±1.93	97.79±0.89	89.03±3.25
	DSMIL'21	59.44±1.24	87.15±0.44	52.75±1.82	81.86±3.09	89.04±3.34	81.81±3.10	86.54±3.09	96.37±1.15	85.67±2.90
	TransMIL'21	60.69±1.66	87.90±0.87	54.67±2.34	85.56±2.03	92.44±2.00	85.52±2.05	90.17±1.85	97.68±1.05	88.58±4.09
	DTFD'22	61.85±0.74	88.63±0.32	55.59±0.73	85.61±2.77	93.73±1.73	85.55±2.79	89.79±2.45	97.77±0.70	87.64±3.35
	ACMIL'24	61.99±1.90	88.92±0.57	56.17±2.98	86.28±2.58	93.08±2.15	86.22±2.58	89.32±5.01	97.86±0.97	89.09±4.67
	WiKG'24	62.36±1.36	89.22±0.36	57.56±1.43	80.59±4.78	91.34±2.18	80.29±5.09	88.57±3.22	97.84±1.16	86.78±4.30
	2DMamba'25	59.29±1.71	86.78±0.43	54.05±1.79	85.94±2.21	92.85±2.94	85.85±2.17	**91.57±1.32**	98.10±1.16	89.88±2.05
	PTCMIL'26	62.52±1.08	88.55±0.69	55.74±2.19	86.06±3.13	93.11±1.86	86.00±3.18	90.32±2.00	97.97±0.75	89.11±2.84
	TiltMIL(Ours)	**64.47±0.79**	**89.99±0.35**	**59.51±1.25**	**86.90±3.46**	**94.12±2.10**	**86.89±3.46**	90.88±2.92	**98.30±0.48**	**90.03±3.12**
UNI	ABMIL'18	74.62±0.50	93.77±0.22	69.96±1.12	92.52±2.05	97.59±1.06	92.50±2.04	94.82±1.98	**99.44±0.24**	94.36±2.38
	CLAM-SB'21	75.69±0.76	94.31±0.30	71.53±1.48	92.13±1.97	97.03±1.96	92.13±1.97	94.62±1.70	99.35±0.31	94.15±1.65
	CLAM-MB'21	76.31±0.56	94.29±0.31	72.32±0.80	93.31±2.29	97.79±1.21	93.30±2.28	94.31±2.20	99.44±0.26	93.92±2.55
	DSMIL'21	72.46±0.62	93.02±0.28	67.05±0.84	93.10±1.70	97.57±1.52	93.09±1.69	94.28±1.62	99.14±0.51	93.84±2.59
	TransMIL'21	66.05±2.05	91.08±0.75	59.98±1.80	92.32±2.84	97.32±1.26	92.29±2.84	94.59±1.31	99.24±0.38	93.93±1.97
	DTFD'22	69.02±2.06	91.67±0.53	62.96±2.36	92.34±2.69	97.37±1.38	92.31±2.72	95.12±1.33	99.41±0.26	94.71±1.57
	ACMIL'24	73.40±2.53	93.20±0.63	68.69±3.32	93.34±2.51	**97.87±1.56**	93.34±2.51	94.81±1.57	98.97±0.09	94.12±1.92
	WiKG'24	74.87±2.58	93.84±0.86	70.76±3.58	90.77±3.46	97.21±1.80	90.73±3.49	94.69±0.91	99.04±0.45	94.18±2.08
	2DMamba'25	69.02±1.18	91.14±0.67	63.09±1.78	91.53±3.63	97.07±1.84	91.51±3.65	94.17±0.95	99.16±0.64	93.66±1.53
	PTCMIL'26	71.82±1.05	92.97±0.36	65.43±1.76	93.25±2.32	97.55±1.87	93.25±2.32	95.59±1.32	99.30±0.40	95.29±1.75
	TiltMIL(Ours)	**80.73±0.46**	**96.04±0.17**	**77.35±0.65**	**93.38±1.97**	96.94±1.87	**93.36±1.97**	**95.99±1.30**	99.14±0.59	**95.88±1.70**

Table 2: **Ablation results on PANDA and TCGA-NSCLC (UNI).** † further extends the full model with third-order cumulant statistics.

	Variant	Acc	AUC	F1-score		Variant	Acc	AUC	F1-score
PANDA	Γ_g	75.00±1.57	93.87±0.54	70.01±1.97	TCGA-NSCLC	Γ_g	92.60±2.42	97.45±1.62	92.59±2.41
	$+\Gamma_1$	74.74±1.49	93.84±0.32	69.73±1.75		$+\Gamma_1$	92.77±3.23	97.44±2.08	92.76±3.23
	$+\Gamma_2$	79.63±1.81	95.11±0.43	75.48±2.15		$+\Gamma_2$	92.71±3.40	96.91±2.37	92.68±3.41
	Full	**80.73±0.46**	**96.04±0.17**	**77.35±0.65**		Full	**93.38±1.97**	96.94±1.87	**93.36±1.97**
	Full $+\Gamma_3^\dagger$	80.49±0.77	95.84±0.17	77.07±1.04		Full $+\Gamma_3^\dagger$	92.79±3.14	97.27±1.75	92.78±3.14
	Full w/o Φ_K	78.54±0.89	95.95±0.25	76.19±0.99		Full w/o Φ_K	89.31±5.04	96.59±1.79	89.19±5.30
	Full w/o Var	79.44±0.23	95.50±0.19	75.69±0.41		Full w/o Var	91.61±4.37	**97.49±2.01**	91.58±4.40

statistics to accuracy and macro-F1 in the full model. Finally, extending the model with third-order cumulant statistics (*Full*+$\Gamma_3^\dagger$) does not provide consistent gains across datasets and metrics.

Robustness and sensitivity analyses. We first evaluate robustness to incomplete bag observation by keeping all trained models fixed and randomly subsampling $n \in \{64, 128, 256, 512, 1024\}$ instances from each PANDA–UNI test bag. As shown in Fig. 4, TiltMIL maintains the highest macro-F1 across all observation budgets and exhibits fewer subsample-induced prediction flips, while also showing low within-WSI prediction variability under repeated sampling. We further examine sensitivity to the number and allocation of finite CGF observation locations on TCGA-NSCLC with ResNet-50 features. As shown in Fig. 5, performance varies only moderately across different budgets and representative allocations, with no systematic gain from increasing the total number of locations. The default configuration $(R_K, R_1, R_2) = (8, 12, 4)$ lies within this broad stable range.

5.2 Task-conditioned and Reusable Finite CGF Views

Task-conditioned finite CGF views remain transferable. To examine whether learned tilt locations capture task-dependent information rather than acting as arbitrary queries, we train observation locations from different source objectives while keeping all shared feature and value mappings fixed. After source training, the temporary heads are discarded and the frozen representations are evaluated using identical linear probes. As shown in Table 3, the task-supervised sources (Rows 2–4) consis-

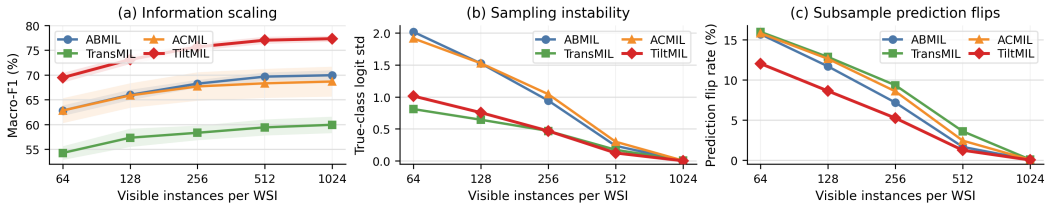


Figure 4: **Bag degradation analysis.** (a) Macro-F1 under different observation budgets. (b) Mean within-WSI standard deviation of the true-class logit across repeated subsamples. (c) Mean prediction flip rate relative to the modal prediction across repeated subsamples.

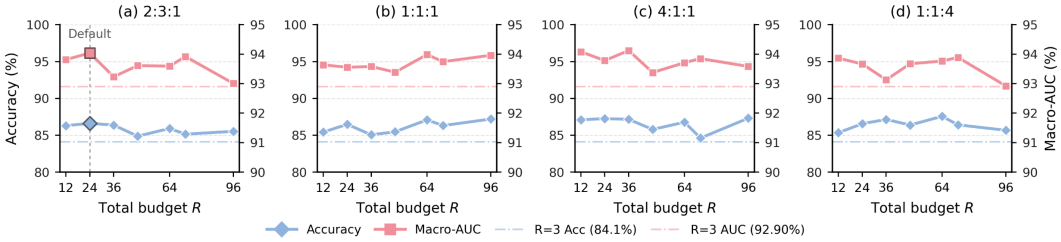


Figure 5: Accuracy and macro-AUC on TCGA-NSCLC with ResNet-50 features under four allocations of $R = R_K + R_1 + R_2$. Horizontal lines indicate the $R = 3$ reference and the marked point in (a) is the default configuration. Panels (a)–(d) use different ratios $R_K : R_1 : R_2$ in Eq. 6.

tently outperform the random baseline (Row 1), confirming that finite CGF observation geometry is shaped by Bag-level supervision. At the same time, transfer is not restricted to the source task: Co-occurrence- and Competitive-trained locations remain effective across multiple decision rules. Thus, learned finite CGF views are task-conditioned but not task-private, and can retain composition information reusable by related downstream decisions.

Reusable observation geometry can also emerge without task labels. Motivated by this cross-task transfer, we further ask whether reusable finite CGF views must be induced by an explicit downstream objective. Using only unlabeled Bag views for pretraining and discarding all self-supervised heads before evaluation, the resulting representation achieves balanced accuracies closely matching the strongest supervised source. These results suggest that reusable ways of observing a Bag distribution need not be tied to a specific decision rule: finite CGF views can also be shaped through task-agnostic self-supervision.

Table 3: **Transferability of finite CGF views.** Composition uses the 10-class Bag histogram as privileged supervision. Self-supervised pretraining uses no task or instance-class labels.

Source	Label	Presence	Co-occurrence	Competitive
Random	×	81.90±0.61	71.98±0.35	80.20±0.48
Presence	✓	86.67±0.59	77.52±0.49	85.19±0.32
Co-occurrence	✓	**87.54±0.81**	79.35±0.99	85.31±0.95
Competitive	✓	86.83±0.34	79.03±0.40	**85.67±0.24**
Composition	✓	86.33±0.32	77.49±0.35	83.89±0.26
Self-supervised	×	87.22±0.32	**79.48±0.17**	84.79±0.67

6 Conclusion

This work addresses the construction of bag representations in MIL and proposes TiltMIL. The proposed framework separates the statistical readout form from the observation locations, enabling the model to extract task-relevant finite statistics from a unified functional compositional representation and combine them through a lightweight fusion mechanism for bag-level prediction. In controlled MIL tasks, we demonstrate that different decision rules rely on different forms of information, including instance presence, evidence co-occurrence, and relative composition. On real WSI-MIL benchmarks, TiltMIL achieves stable predictive performance across different datasets and feature encoders. These results validate the effectiveness of finite task-relevant statistical readouts for modeling complex bag-level decisions.

REFERENCES

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, 28: 154–163, 2022. doi: 10.1038/s41591-021-01620-2.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

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. doi: 10.1016/j.patcog.2017.10.009.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024. doi: 10.1038/s41591-024-02857-3.

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. doi: 10.1016/S0004-3702(96)00034-3.

Gary Doran and Soumya Ray. Multiple-instance learning from distributions. *Journal of Machine Learning Research*, 17(128):1–50, 2016. URL https://www.jmlr.org/papers/v17/15-171.html.

Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American Journal of Surgical Pathology*, 40(2):244–252, 2016. doi: 10.1097/PAS.0000000000000530.

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. doi: 10.1109/CVPR.2016.90.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021. doi: 10.1109/CVPR46437.2021.01409.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11323–11332, 2024. doi: 10.1109/CVPR52733.2024.01076.

Xingyu Li, Jitendra Jonnagaddala, Min Cen, Hong Zhang, and Steven Xu. Colorectal cancer survival prediction using deep distribution-based multiple-instance learning. *Entropy*, 24(11):1669, 2022. doi: 10.3390/e24111669.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017. doi: 10.1561/2200000060.

Mustafa Umit Oner, Jared Marc Song Kye-Jet, Hwee Kuan Lee, and Wing-Kin Sung. Distribution-based MIL pooling filters: Experiments on a lymph node metastases dataset. *Medical Image Analysis*, 87:102813, 2023. doi: 10.1016/j.media.2023.102813.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

John N. Weinstein, Eric A. Collisson, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013. doi: 10.1038/ng.2764.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020. doi: 10.1016/j.media.2020.101789.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022. doi: 10.1109/CVPR52688.2022.01824.

Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S. Hosseini. 2DMamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3583–3592, 2025. doi: 10.1109/CVPR52734.2025.00339.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 125–143. Springer, 2024. doi: 10.1007/978-3-031-73668-1_8.

Beidi Zhao, Sangmook Kim, Hao Chen, Chen Zhou, Zu-hua Gao, Gang Wang, and Xiaoxiao Li. PTCMIL: Multiple instance learning via prompt token clustering for whole slide image analysis. *Medical Image Analysis*, 111:104013, 2026. doi: 10.1016/j.media.2026.104013.

Zhikang Zhao, Kaitao Chen, and Jing Zhao. RPMIL: Rethinking uncertainty-aware probabilistic multiple instance learning for whole slide pathology diagnosis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2467–2475, 2025. doi: 10.24963/ijcai.2025/275.

AI USE STATEMENT

We used large language models to assist with literature discovery, manuscript organization, language polishing, and drafting or revising portions of the text and figures. The research questions, methodological decisions, experimental design, implementation, analyses, and scientific conclusions were determined and validated by the authors. All AI-assisted content, including technical statements and citations, was reviewed and verified by the authors, who take full responsibility for the final manuscript.

A Additional Details for the Controlled Analysis

A.1 Implementation Details and Predictive Performance

We use CIFAR-10 (Krizhevsky, 2009) for all controlled experiments. The official training split is divided class-wise into 4,000 images for training and 1,000 images for validation, while the official test split is used exclusively for testing. Thus, the underlying images are disjoint across training, validation, and test sets. Images are sampled without replacement within each bag, but may be reused across different bags in the same split. The target type $\mathcal{A}$ is the CIFAR-10 *cat* class (class ID 3). For the Co-occurrence and Competitive tasks, $\mathcal{B}$ is *dog* (class ID 5). Presence treats all non-cat classes as background; Co-occurrence uses the remaining eight classes as background; Competitive uses seven background classes, with *truck* excluded from the clean benchmark. Each split contains 30,000/10,000/10,000 train/validation/test bags. Presence and Competitive are balanced between the two labels, while Co-occurrence is balanced across its four bag classes.

Each original bag contains 64 instances. For Presence, positive bags contain $5\mathcal{A} + 59\mathrm{BG}$, while negative bags contain $64\mathrm{BG}$. For Co-occurrence, the four bag classes have compositions $64\mathrm{BG}$, $10\mathcal{A} + 54\mathrm{BG}$, $10\mathcal{B} + 54\mathrm{BG}$, and $10\mathcal{A} + 10\mathcal{B} + 44\mathrm{BG}$, respectively. For Competitive, each bag contains ten target instances and 54 background instances, with $n_{\mathcal{A}} + n_{\mathcal{B}} = 10$ and $\Delta n = n_{\mathcal{A}} - n_{\mathcal{B}} \in \{-10, -8, -6, -4, -2, 2, 4, 6, 8, 10\}$. Tied compositions are excluded from these standard Competitive task datasets. The separately constructed bags and instance additions used for composition-edit analysis are described in Appendix A.2.

Images are resized to 256 pixels, center-cropped to 224×224, and normalized using ImageNet statistics. A frozen ImageNet-pretrained ResNet-18 (He et al., 2016) produces 512-dimensional instance features. ABMIL applies a $512 \to 512$ linear projection followed by ReLU and dropout ($p = 0.1$), and uses a $512 \to 128 \to 1$ Tanh attention network with bag-wise softmax normalization. The final classifier is linear, with two outputs for Presence and Competitive and four outputs for Co-occurrence. All models are trained with Adam, a learning rate of 2×10^{-4}, weight decay of 10^{-5}, and a batch size of 256 bags. Results are reported over model seeds 2024, 2025, and 2026. Checkpoints are selected by validation binary AUC for Presence and Competitive and by validation macro one-vs-rest AUC for Co-occurrence. Table 8 confirms that ABMIL reliably learns all three controlled decision rules before the composition-edit analysis.

A.2 Composition-Edit Protocols

Presence. Figure 1(a) uses the 4,578 positive test bags that are correctly classified by all three model seeds. Starting from $5\mathcal{A} + 59\mathrm{BG}$, each of the five $\mathcal{A}$ instances is considered separately. For a target instance $\mathcal{A}_j$, five additional copies of the same instance are appended to the original bag, increasing its size from 64 to 69. Each edit starts from the original bag, and the five resulting margin changes are averaged to obtain $\overline{\Delta M}$ for that bag. Let $M(B)$ denote the positive-versus-negative logit margin, and let $B_j^{(+5)}$ denote the bag obtained by appending five copies of the j-th target instance to the original bag. We compute

$$\Delta M_j(B) = M(B_j^{(+5)}) - M(B), \qquad \overline{\Delta M}(B) = \frac{1}{5}\sum_{j=1}^{5} \Delta M_j(B).$$

We use logit margins because predicted probabilities are typically near saturation on this task.

Co-occurrence. Figure 1(b) uses 1,000 shared controlled groups constructed from the CIFAR-10 test image pool. Each group contains an $\mathcal{A}$-only bag with $10\mathcal{A} + 54\mathrm{BG}$ and a symmetric $\mathcal{B}$-only bag, together with five unused candidate instances from each target type. Each edit appends five instances of either the type already present or the missing type, increasing the bag size from 64 to 69. The same groups are evaluated by all three model seeds. After retaining source bags that are correctly classified before editing and averaging available seed predictions for each unique bag, the analysis contains 1,887 unique $(\mathrm{group}, \mathrm{source})$ bags. The mean $P(\mathcal{A} + \mathcal{B})$ is 0.057 for the original bags, 0.083 after adding the same type, and 0.820 after adding the missing type; the missing-type edit produces the larger response in 98.09% of the analyzed bags. The inserted instances receive nearly identical total aggregation mass under the two edits, consistent with the comparison shown

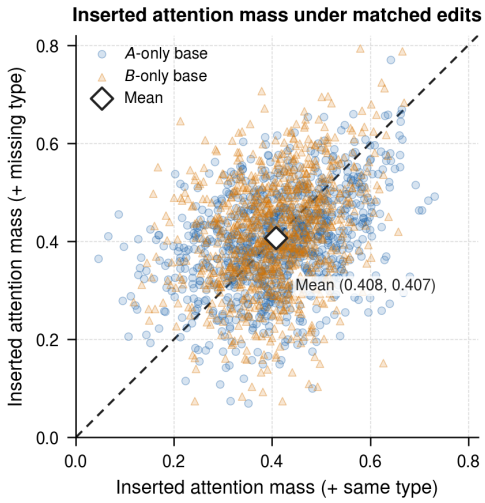


Figure 6: Matched comparison of the aggregation mass assigned to inserted instances in the Co-occurrence task. Each point corresponds to the same source bag under the same-type and missing-type edits, with aggregation mass summed over the five inserted instances. Circles and triangles denote $\mathcal{A}$-only and $\mathcal{B}$-only source bags, respectively. The dashed line indicates equal inserted mass under the two edits, and the diamond marks the overall mean. Despite their markedly different effects on $P(\mathcal{A}+\mathcal{B})$, the two edits receive closely matched aggregation mass.

Table 4: Co-occurrence edit statistics. Results are averaged over the 1,887 unique source bags after averaging available model seeds within each bag.

Condition	$P(\mathcal{A}+\mathcal{B})$	BG mass	Existing mass	Inserted mass
Original	0.057	0.311	0.689	–
Same type	0.083	0.181	0.412	0.408
Missing type	0.820	0.181	0.413	0.407

in Figure 1(b). The matched comparison in Fig. 6 further shows that this similarity is not limited to the aggregate mean: the paired masses remain broadly distributed around the equality line across individual source bags.

Competitive. Figure 1(c) is based on 1,000 paired composition trajectories. Within each trajectory, the same 54 background instances are retained while $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ vary subject to $n_{\mathcal{A}}+n_{\mathcal{B}}=10$. The candidate sequences of $\mathcal{A}$ and $\mathcal{B}$ instances are also fixed within a trajectory, so each point differs only in relative composition. All three model seeds are evaluated on the same trajectories, with no filtering by prediction correctness or model response. The prediction curve reports $P(\mathcal{A}\text{-dominant})$ averaged first across model seeds and then across trajectories. The stacked bars are computed analogously after summing the normalized aggregation weights over all $\mathcal{A}$, $\mathcal{B}$, and background instances within each bag; they therefore represent total aggregation mass for each instance type.

A.3 Closed-Form Response of Normalized Aggregation

Let

$$z_B = \sum_j a_j h_j, \qquad a_j = \frac{\exp s(h_j)}{\sum_\ell \exp s(h_\ell)}.$$

If k additional copies of instance i are inserted into the bag, direct renormalization gives

$$z_{B\cup k\{x_i\}} = \frac{z_B + k a_i h_i}{1 + k a_i}.$$

Table 5: Default TiltMIL configuration used in the WSI experiments.

Quantity	Setting
Task feature dimension p	256
Observation locations (R_K, R_1, R_2)	$(8, 12, 4)$
Base statistic dimension $\dim(c_0)$	128
First-order value dimension $\dim(P_1(e))$	32
Second-order value dimension $\dim(P_2(e))$	128
Tilt temperatures $(\Phi_K, \Phi_\kappa^{(1)}, \Phi_\kappa^{(2)})$	$(0.25, 0.25, 1.0)$
Fusion hidden dimension	128
Second-order residual weight λ_m	0.1

For a pair of classes c and r, define $M_{c,r}(B) = \beta_{c,r}^\top z_B + \Delta b_{c,r}$ and $p_i^{c,r} = \beta_{c,r}^\top h_i + \Delta b_{c,r}$, where $\beta_{c,r} = w_c - w_r$. The corresponding margin change is

$$M_{c,r}(B \cup k\{x_i\}) - M_{c,r}(B) = \frac{ka_i}{1 + ka_i} \left[p_i^{c,r} - M_{c,r}(B)\right].$$

Hence, repeated duplication changes the response magnitude through $ka_i/(1+ka_i)$, while its sign is determined by the position of the inserted instance relative to the current bag margin. The Presence analysis uses $k = 5$.

B Additional Details for Experiments

B.1 Implementation Details

Model instantiation. Unless otherwise specified, all WSI experiments use the same TiltMIL configuration. The shared instance mapping E_θ maps the input patch features to a $p = 256$ dimensional task-specific feature space using a linear layer followed by GELU and dropout (0.1). The untilted composition mapping C_θ is implemented as a 256–256–128 MLP with GELU activation, yielding a 128-dimensional base statistic c_0. The default numbers of observation locations for the three finite CGF readouts in Eq. (6) are $R_K = 8$, $R_1 = 12$, and $R_2 = 4$. For Φ_K and $\Phi_\kappa^{(1)}$, the tilt locations are parameterized by linear scores on the 256-dimensional task feature, with temperature 0.25 for both readouts. In Eq. (7), the first-order value map P_1 projects the task feature to 32 dimensions. For $\Phi_\kappa^{(2)}$, we use four learnable tokens to parameterize tilted views in a 64-dimensional key space, while the value map P_2 projects the task feature to 128 dimensions.

Lightweight statistical fusion. The prediction mappings g_θ and l_θ are both single-hidden-layer MLPs with hidden dimension 128, GELU activation, and dropout (0.1). Their input dimensions are $128 + R_K$ and $32R_1$, respectively. The class-wise gate consists of a single linear mapping from Γ_g to the C output classes followed by a sigmoid. The learnable class-wise scale η is initialized to 0.5, and the gate bias is initialized to -0.5. The mapping q_θ contains no additional hidden layer; it applies LayerNorm to the flattened up-to-second-order statistics followed by a single linear mapping to the class logits. Its residual contribution is weighted by $\lambda_m = 0.1$.

Optimization. All WSI models are trained with AdamW for at most 30 epochs, using an initial learning rate of 2×10^{-4} and weight decay of 10^{-5}. We use a two-epoch warm-up followed by cosine learning-rate decay. In each fold, model selection is performed exclusively on the validation split using macro-AUC; training is stopped if the validation criterion does not improve for eight consecutive epochs, and the checkpoint with the best validation macro-AUC is used for test evaluation. During training, each WSI is represented by at most 4,096 randomly sampled instances, while validation and test bags use deterministic uniform sampling. For the empirical perturbation consistency term in Eq. (11), we set $\lambda_{\mathrm{EPC}} = 0.1$. The perturbed view is generated by independently dropping instances with probability 0.1 and adding Gaussian noise with standard deviation 0.01 to the retained instance features.

Table 6: Summary of the WSI datasets used in the experiments.

Dataset	Task	Classes	WSIs
PANDA	ISUP grading	6	10,616
TCGA-NSCLC	Lung cancer subtype	2	993
TCGA-RCC	Renal cancer subtype	3	884

Table 7: Quadratic weighted kappa (QWK, %) on PANDA. Bold and underlined values denote the best and second-best results, respectively.

Method	ResNet-50 QWK ↑	UNI QWK ↑
ABMIL	77.51 ± 1.59	87.71 ± 0.88
CLAM-SB	77.75 ± 1.81	87.97 ± 0.81
CLAM-MB	78.58 ± 1.87	$\underline{88.05 \pm 1.17}$
DSMIL	75.59 ± 1.53	86.21 ± 0.76
TransMIL	77.33 ± 1.62	84.26 ± 1.67
DTFD	78.14 ± 1.54	83.92 ± 1.26
ACMIL	$\underline{78.88 \pm 1.57}$	87.08 ± 1.62
WiKG	77.64 ± 1.25	87.34 ± 0.73
2DMamba	75.15 ± 2.10	85.48 ± 1.03
PTCMIL	78.58 ± 1.67	86.72 ± 1.17
TiltMIL (ours)	$\mathbf{79.37 \pm 1.49}$	$\mathbf{89.83 \pm 0.86}$

B.2 Datasets

We evaluate TiltMIL on three public WSI benchmarks covering prostate, lung, and renal cancer. Table 6 summarizes the prediction tasks and cohort statistics.

PANDA. The Prostate cANcer graDe Assessment (PANDA) dataset (Bulten et al., 2022) contains 10,616 H&E-stained prostate core-needle biopsy WSIs collected from Radboud University Medical Center and Karolinska Institutet. Each biopsy is assigned an International Society of Urological Pathology (ISUP) grade. We formulate PANDA as a six-class grading task with labels from grade 0 (benign) to grade 5.

TCGA-NSCLC. TCGA-NSCLC combines diagnostic WSIs from the TCGA lung adenocarcinoma (TCGA-LUAD) and lung squamous cell carcinoma (TCGA-LUSC) projects (Weinstein et al., 2013). Following the commonly used WSI-MIL cohort, the dataset contains 993 diagnostic slides, including 507 LUAD and 486 LUSC slides. We use it for binary histological subtype classification.

TCGA-RCC. TCGA-RCC combines diagnostic WSIs from three renal cell carcinoma projects: kidney chromophobe (TCGA-KICH), kidney renal clear cell carcinoma (TCGA-KIRC), and kidney renal papillary cell carcinoma (TCGA-KIRP) (Weinstein et al., 2013). The cohort contains 884 diagnostic slides, including 111 KICH, 489 KIRC, and 284 KIRP slides, and is used for three-class renal carcinoma subtype classification.

B.3 Quadratic Weighted Kappa on PANDA

Metric definition. PANDA grading is an ordinal classification problem with six ordered ISUP grades, $\mathcal{Y} = \{0, 1, \ldots, 5\}$. Accuracy and macro-F1 do not account for the ordinal distance between the predicted and ground-truth grades, whereas quadratic weighted kappa (QWK) penalizes larger grade discrepancies more strongly. Let O_{ij} denote the normalized confusion matrix, i.e., the empirical joint probability of ground-truth grade i and predicted grade j. Its row and column marginals are $p_i = \sum_j O_{ij}$ and $q_j = \sum_i O_{ij}$, respectively, and $E_{ij} = p_i q_j$ denotes the expected matrix under independent ground-truth and prediction marginals. For $K = 6$ ordered classes, the quadratic

disagreement weight is

$$w_{ij} = \frac{(i-j)^2}{(K-1)^2}. \tag{12}$$

QWK is then defined as

$$\kappa_{\mathrm{QW}} = 1 - \frac{\sum_{i=0}^{K-1}\sum_{j=0}^{K-1} w_{ij}O_{ij}}{\sum_{i=0}^{K-1}\sum_{j=0}^{K-1} w_{ij}E_{ij}}. \tag{13}$$

A value of 1 indicates perfect ordinal agreement, 0 corresponds to the agreement expected under independent marginals, and negative values indicate agreement below this reference level.

Discussion. TiltMIL achieves the highest mean QWK with both feature encoders. With ResNet-50 features, it reaches 79.37%, exceeding the strongest baseline. With UNI features, it reaches 89.83%, improving over CLAM-MB by 1.78 percentage points. These results complement the accuracy, AUC, and macro-F1 results in Table 1, showing that the improvement on PANDA also extends to agreement with the ordered ISUP grading structure.

B.4 Sensitivity Analysis

We examine the sensitivity of TiltMIL to the number and relative allocation of finite CGF observation locations. Let R_K, R_1, and R_2 denote the numbers of locations used for the CGF-value, first-order cumulant, and cumulant readouts retained through second order, respectively, with $R = R_K + R_1 + R_2$. We vary the total number of locations over $R \in \{12, 24, 36, 48, 64, 72, 96\}$ and consider the ten allocation ratios shown in Fig. 7. The default configuration $(R_K, R_1, R_2) = (8, 12, 4)$ corresponds to the 2:3:1 allocation at $R = 24$. We additionally include $\mathcal{R} = 3$, with one location for each readout family, as a minimal reference. Across the evaluated range, both accuracy and macro-AUC vary moderately with the total number and allocation of observation locations, and no allocation consistently dominates across different values of R. Increasing R beyond the default configuration also does not lead to a systematic improvement. These results indicate that TiltMIL is not strongly dependent on a narrowly chosen number or allocation of finite CGF observation locations.

B.5 Finite CGF View Pretraining and Evaluation

End-to-end performance on the controlled tasks. Before analyzing the transferability of finite CGF views, we first evaluate the end-to-end performance of TiltMIL on the three controlled tasks. TiltMIL reliably learns the prescribed Bag-level decision rules for Presence, Co-occurrence, and Competitive, and achieves better overall predictive performance than the corresponding ABMIL results reported in Table 8. These results further show that the finite CGF readouts can adapt to different forms of Bag-level information requirements.

Frozen-view evaluation. Table 3 follows a unified frozen-view linear-probe protocol to evaluate the transferability of finite CGF views learned from different source objectives. After source training, the original task classifier is discarded, the finite CGF representation is frozen, and a new linear probe is trained for each target task. The results therefore reflect the linear recoverability of task-relevant information retained in the frozen finite CGF representation. Because a common linear classification head is retrained under this protocol, the diagonal entries in Table 3 differ from the end-to-end classification results reported above.

Table 8: Predictive performance of ABMIL on the controlled tasks.

Task	Acc (%)	AUC (%)	Macro-F1 (%)
Presence	$89.35_{\pm 1.06}$	$95.51_{\pm 0.03}$	$89.34_{\pm 1.07}$
Co-occurrence	$84.64_{\pm 0.19}$	$96.20_{\pm 0.18}$	$84.54_{\pm 0.22}$
Competitive	$90.79_{\pm 0.29}$	$96.54_{\pm 0.13}$	$90.72_{\pm 0.31}$

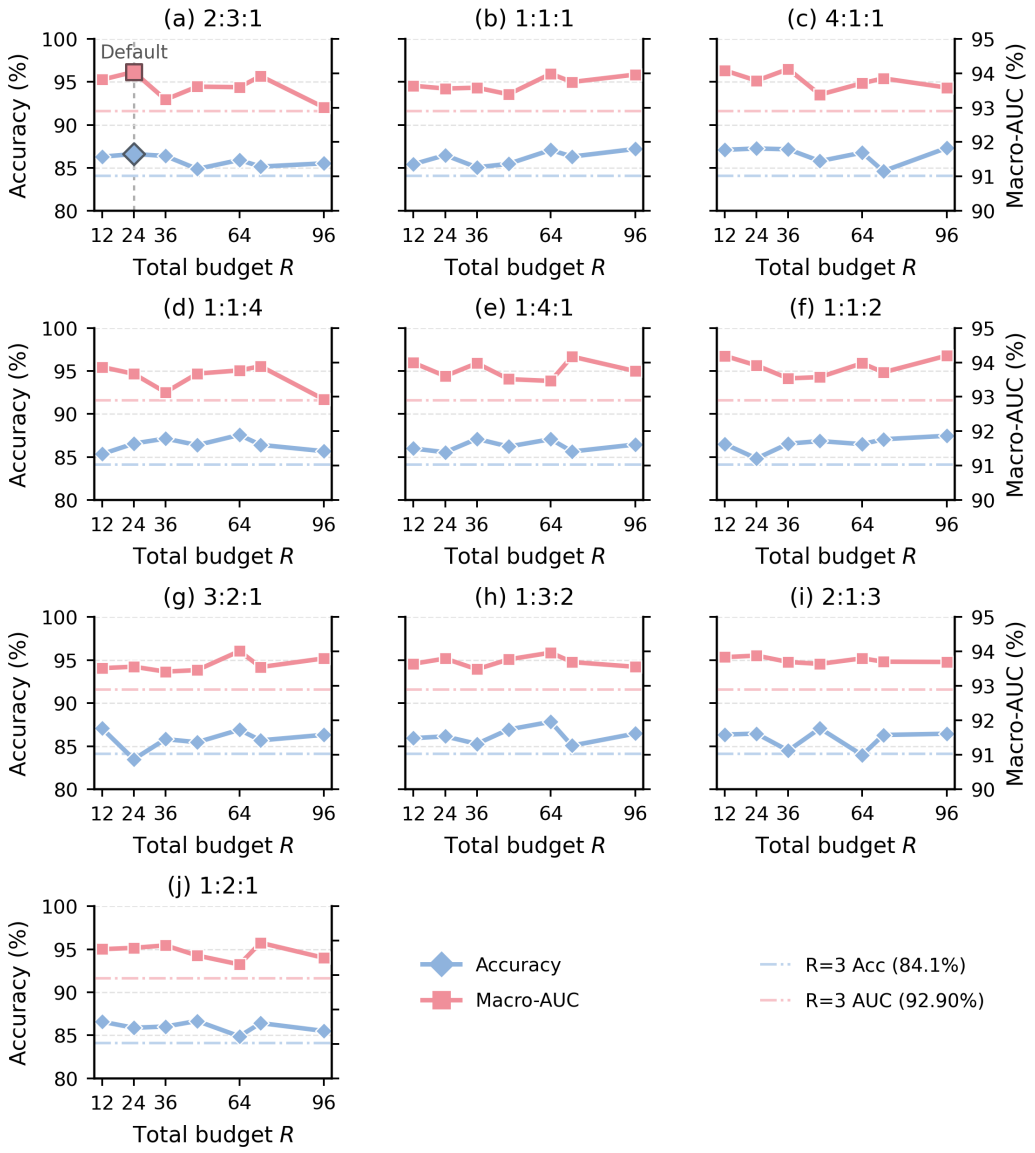


Figure 7: Accuracy and macro-AUC on TCGA-NSCLC with ResNet-50 features under ten allocations of $R = R_K + R_1 + R_2$. Horizontal lines indicate the $R = 3$ reference and the marked point in (a) is the default configuration.

Self-supervised view pretraining. For the self-supervised source in Table 3, we use the same 30,000/10,000/10,000 controlled Bag splits. The ResNet-18 instance features, the 512-to-256 instance mapping, and the branch-specific key/value mappings are kept fixed, while only the finite CGF observation locations are optimized. After pretraining, all temporary self-supervised predic-

Table 9: Predictive performance of TiltMIL on the controlled tasks.

Task	Acc (%)	AUC (%)	Macro-F1 (%)
Presence	$91.18_{\pm 0.41}$	$96.33_{\pm 0.15}$	$91.17_{\pm 0.41}$
Co-occurrence	$86.28_{\pm 0.72}$	$96.99_{\pm 0.25}$	$86.22_{\pm 0.74}$
Competitive	$92.04_{\pm 0.16}$	$97.31_{\pm 0.04}$	$92.01_{\pm 0.16}$

tion heads are removed. The resulting finite CGF representation is then frozen, and separate linear probes are trained for Presence, Co-occurrence, and Competitive.