STATISTICAL TESTING FOR MULTIPLE INSTANCE LEARNING VIA SELECTIVE INFERENCE WITH APPLICATIONS TO COMPUTATIONAL PATHOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple instance learning (MIL) is widely used in computational pathology for weakly supervised analysis of whole-slide images (WSIs) without patch-level annotations. In attention-based MIL, high-attention instances are often interpreted as diagnostically important regions, but attention scores alone do not establish whether these instances significantly differ from normal tissue. We formulate this problem as a statistical hypothesis test that compares a selected high-attention instance with a representative normal reference selected based on feature similarity. Because both the target and reference instances are selected through data-dependent procedures, standard hypothesis testing is invalid. We therefore develop a selective inference (SI) framework that accounts for attention-based instance selection and adaptive reference selection, enabling valid selective p-values. Experiments demonstrate Type-I error control on synthetic and MNIST-based data and practical applicability to pathological WSIs, with higher statistical power than the conventional over-conditioning approach.

1 INTRODUCTION

Multiple instance learning (MIL) (Dietterich et al., 1997; Maron & Lozano-Pérez, 1997) is a weakly supervised framework in which labels are assigned to bags while individual instances remain unlabeled. It has become a standard paradigm for whole-slide image (WSI) analysis in computational pathology (Das et al., 2018; Yamamoto et al., 2019; Campanella et al., 2019; Hashimoto et al., 2020). Because WSIs are extremely large and dense patch-level annotations are prohibitively expensive, MIL treats image patches as instances and slide-level diagnoses as bag labels, enabling WSI classification without exhaustive patch-level annotation.

Among MIL approaches, attention-based MIL (ABMIL) (Ilse et al., 2018) has become influential because it provides both competitive predictive performance and instance-level attention scores. High-attention patches are often interpreted as diagnostically important regions and are widely used as visual explanations in computational pathology (Sudharshan et al., 2019; Campanella et al., 2019; Hashimoto et al., 2020; Li et al., 2021; Lu et al., 2021; Zhang et al., 2022; Takagi et al., 2023; Hashimoto et al., 2024). However, attention scores do not necessarily correspond to causal or clinically meaningful evidence (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019). In pathology, high-attention patches may also arise from dataset-specific biases, staining artifacts, or optimization effects rather than genuine pathological abnormalities (Wang et al., 2022; Cai et al., 2024; Sun et al., 2025; Zehnder et al., 2025). Thus, attention scores alone do not provide statistical evidence that a selected high-attention instance significantly differs from normal tissue.

We address this limitation by formulating the evaluation of high-attention instances as a statistical hypothesis testing problem. Specifically, we assess whether a selected high-attention instance significantly deviates from a representative normal reference instance. A key issue is that normal tissue can exhibit substantial heterogeneity due to variations in tissue appearance, staining, slide preparation, and other dataset-specific factors. Therefore, comparison with an arbitrary or globally representative normal instance may not provide an appropriate reference. Instead, for each selected target instance, we retrieve similar normal instances and adaptively select a representative local reference, thereby tailoring the comparison to the target instance.

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.

This adaptive procedure introduces an important statistical challenge. The target instance is selected according to its attention score, while the representative normal reference is selected adaptively from similar normal instances. Because standard hypothesis tests assume that the hypothesis being tested is specified independently of the data used for testing, ignoring these adaptive selections leads to selection bias and invalid statistical inference (Kriegeskorte et al., 2009). Properly accounting for both target selection and reference selection is essential for obtaining valid significance measures.

To address this issue, we employ selective inference (SI) (Lee et al., 2016; Tian & Taylor, 2018; Taylor & Tibshirani, 2018), which performs statistical inference conditional on the selection event induced by the data-dependent procedure. In our setting, we jointly account for the attention-based target-selection outcome and the adaptively selected representative reference, and derive valid selective p-values for high-attention instances.

Figure 1 illustrates the overall framework. We evaluate the proposed method on synthetic data, MNIST-based data, and real-world pathological WSIs. The results demonstrate Type-I error control on synthetic and MNIST-based data and practical applicability to pathological WSIs, with higher statistical power than the conventional over-conditioning approach. To the best of our knowledge, this is the first study to formulate statistical significance testing for instances selected by MIL models within an SI framework. We use ABMIL as a proof-of-concept because its relatively simple attention mechanism allows the data-dependent selection procedure to be explicitly characterized. The proposed framework is not conceptually restricted to ABMIL and can potentially be extended to other instance-selection mechanisms whose selection events can be characterized.

Related work. MIL was originally introduced for learning from labeled bags of unlabeled instances (Dietterich et al., 1997; Maron & Lozano-Pérez, 1997; Zhou & Zhang, 2002; Andrews et al., 2002) and has since been widely adopted in computational pathology. Early deep MIL approaches used instance-level or embedding-based aggregation strategies (Wu et al., 2015; Feng & Zhou, 2017; Wang et al., 2018; Kraus et al., 2016), while ABMIL (Ilse et al., 2018) introduced learnable attention-based aggregation and has been widely applied to WSI classification (Hashimoto et al., 2020; Li et al., 2021; Lu et al., 2021; Zhang et al., 2022). More recent transformer-based MIL architectures model interactions among instances and have further improved slide-level predictive performance (Lee et al., 2019; Chen et al., 2021; Shao et al., 2021; Chen et al., 2022; Takagi et al., 2023). The reliability of attention as an explanation has also been questioned (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019), including in computational pathology, where attention maps may reflect spurious or dataset-specific effects rather than clinically meaningful abnormalities (Wang et al., 2022; Cai et al., 2024; Sun et al., 2025; Zehnder et al., 2025). Our goal is therefore not to develop a new MIL architecture or attention mechanism, but to statistically evaluate whether MIL-selected instances differ from appropriate normal references.

SI provides a general framework for valid inference after data-dependent selection by conditioning on the selection event generated by the underlying algorithm (Taylor & Tibshirani, 2015; Lee et al., 2016; Tibshirani et al., 2016). Early work primarily focused on feature and model selection in regression, including sparse models such as the Lasso (Lee et al., 2016; Hyun et al., 2018; Suzumura et al., 2017; Yamada et al., 2018). The framework has subsequently been extended to a broad range of data-dependent inference problems, including clustering, changepoint detection, outlier detection, anomaly detection, and principal component analysis (Chen & Witten, 2023; Gao et al., 2024; Duy et al., 2020; Chen & Bien, 2020; Perry et al., 2026). Methodological developments such as randomized SI and parametric-programming-based approaches have also been proposed to reduce unnecessary conditioning and improve statistical power (Tian & Taylor, 2018; Nguyen Le Duy & Takeuchi, 2022).

More recently, SI has been extended to selection mechanisms involving deep neural networks. Representative examples include statistical inference for neural-network-based image segmentation and salient-region detection (Duy et al., 2022; Miwa et al., 2023), deep nearest-neighbor anomaly detection (Niihori et al., 2025), saliency analysis for graph neural networks (Nishino et al., 2025), and attention-based hypotheses generated by transformer models (Shiraishi et al., 2024; 2026). Computational frameworks for automatically characterizing neural-network-induced selection events have also been developed (Katsuoka et al., 2025). These advances make it possible to consider SI for MIL, where the instances subjected to statistical testing are themselves selected by the MIL model. Our setting introduces an additional challenge because both the high-attention target instance and

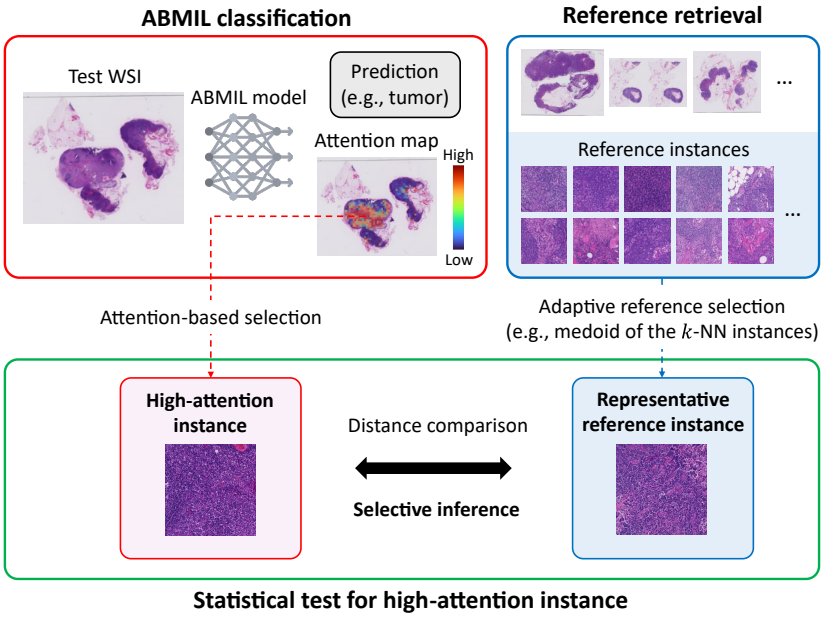


Figure 1: Overview of the proposed framework. An ABMIL model selects a high-attention instance from a test WSI, while a representative normal reference is adaptively selected from reference instances. The proposed SI framework tests whether the selected target significantly differs from the reference while accounting for both data-dependent selection procedures.

its representative normal reference are selected adaptively. We account for both selection processes within a unified SI framework.

Contributions. The main contributions of this paper are:

- We formulate the evaluation of high-attention instances in ABMIL as a statistical hypothesis testing problem, enabling quantitative assessment of instance-level significance.
- We develop an SI framework that accounts for both attention-based target selection and adaptive reference selection, enabling valid selective p-values.
- We demonstrate Type-I error control on synthetic and MNIST-based data and practical applicability to pathological WSI data, while achieving higher statistical power than the conventional over-conditioning approach.

2 Problem Setup: Multiple Instance Learning

In MIL, each sample is represented as a bag of instances. For any positive integer N, we write $[N] := \{1, \dots, N\}$. The n-th bag is denoted by

$$X^n = \{\boldsymbol{x}^{n,m}\}_{m \in [M^n]},$$

where $\boldsymbol{x}^{n,m} \in \mathbb{R}^d$ is the m-th instance, d is the instance dimension, and M^n is the number of instances. Each bag is associated with a bag-level label Y^n, whereas instance-level labels are unavailable. In computational pathology, a bag typically corresponds to a WSI, and its instances correspond to image patches extracted from the WSI.

We employ ABMIL (Ilse et al., 2018), which assigns an attention score to each instance and aggregates instance features into a bag-level representation. For each instance, the feature representation and the pre-sigmoid attention logit are computed as

$$\boldsymbol{h}^{n,m} = f_{\mathrm{enc}}(\boldsymbol{x}^{n,m}) \in \mathbb{R}^{d'}, \qquad a^{n,m} = f_{\mathrm{att}}(\boldsymbol{h}^{n,m}),$$

where $f_{\mathrm{enc}} : \mathbb{R}^d \to \mathbb{R}^{d'}$ denotes the feature extractor, d' is the feature dimension, and f_{att} denotes the attention network. The attention logits are subsequently transformed into normalized attention weights and used to construct the bag-level representation for classification.

In computational pathology, high-attention instances are often interpreted as diagnostically important regions and used as visual explanations (Hashimoto et al., 2020; Li et al., 2021; Lu et al., 2021). However, attention scores do not necessarily provide faithful explanations (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019) and, by themselves, do not provide statistical evidence that the corresponding instances differ from normal tissue. Statistical hypothesis testing therefore provides a complementary way to quantify such deviations.

Motivated by this observation, we use the pre-sigmoid attention logits to select candidate instances for statistical testing. For a test bag $X = \{\boldsymbol{x}^m\}_{m\in[M]}$, we define the set of selected high-attention instances as

$$\mathcal{C}(X) = \{m \in [M] \mid a^m > \tau\},$$

where τ is a fixed attention threshold. Depending on the experimental setting, τ is either specified a priori or estimated beforehand using an independent dataset. For each $m \in \mathcal{C}(X)$, the corresponding instance $\boldsymbol{x}^m$ is treated as a target instance. Our objective is to test whether each selected high-attention instance significantly differs from a representative normal reference instance. The corresponding hypothesis-testing problem is formulated in the next section.

3 Statistical testing formulation for high-attention instances

Let

$$X_{\mathrm{test}} = \{\boldsymbol{x}^i_{\mathrm{test}}\}_{i\in[M_{\mathrm{test}}]} \quad \text{and} \quad X_{\mathrm{ref}} = \{\boldsymbol{x}^j_{\mathrm{ref}}\}_{j\in[M_{\mathrm{ref}}]}$$

denote a test bag and an independent reference set of normal instances, respectively. For each test instance, the trained ABMIL model computes $\boldsymbol{h}^i_{\mathrm{test}} = f_{\mathrm{enc}}(\boldsymbol{x}^i_{\mathrm{test}})$ and $a^i_{\mathrm{test}} = f_{\mathrm{att}}(\boldsymbol{h}^i_{\mathrm{test}})$. As defined in Section 2, we focus on high-attention instances

$$\mathcal{C}(X_{\mathrm{test}}) = \{i \in [M_{\mathrm{test}}] \mid a^i_{\mathrm{test}} > \tau\}.$$

For each $i \in \mathcal{C}(X_{\mathrm{test}})$, we test whether $\boldsymbol{x}^i_{\mathrm{test}}$ significantly differs from a representative normal reference instance. Because normal tissue can be heterogeneous, we adapt the reference to each target instance rather than using a fixed global reference. Specifically, we first identify the k nearest reference instances in the feature space, where $\boldsymbol{h}^j_{\mathrm{ref}} = f_{\mathrm{enc}}(\boldsymbol{x}^j_{\mathrm{ref}})$, and then select the medoid of these instances as the representative reference. Let $j_{\mathrm{med}} \in [M_{\mathrm{ref}}]$ denote the index of this selected representative reference. Other adaptive reference-selection rules can in principle be accommodated within the same SI framework, provided that their selection events can be characterized.

For notational simplicity, we hereafter omit the index of the selected test instance and denote it by $\boldsymbol{x}_{\mathrm{test}}$. We assume the signal-plus-noise model

$$\boldsymbol{X}_{\mathrm{test}} = \boldsymbol{s}_{\mathrm{test}} + \boldsymbol{\epsilon}_{\mathrm{test}}, \qquad \boldsymbol{X}^j_{\mathrm{ref}} = \boldsymbol{s}^j_{\mathrm{ref}} + \boldsymbol{\epsilon}^j_{\mathrm{ref}},$$

where $\boldsymbol{\epsilon}_{\mathrm{test}}, \boldsymbol{\epsilon}^j_{\mathrm{ref}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$ are mutually independent. The signal vectors are arbitrary unknown fixed vectors, and σ^2 is assumed known for the theoretical analysis. We test

$$\mathrm{H}_0 : \boldsymbol{s}_{\mathrm{test}} = \boldsymbol{s}^{j_{\mathrm{med}}}_{\mathrm{ref}} \quad \text{vs.} \quad \mathrm{H}_1 : \boldsymbol{s}_{\mathrm{test}} \neq \boldsymbol{s}^{j_{\mathrm{med}}}_{\mathrm{ref}}.$$

We use the Euclidean distance between the selected target and reference instances as the test statistic:

$$T = \frac{1}{\sqrt{2}} \left\| \boldsymbol{X}_{\mathrm{test}} - \boldsymbol{X}^{j_{\mathrm{med}}}_{\mathrm{ref}} \right\|_2, \qquad \widetilde{T} = \frac{T}{\sigma}. \tag{1}$$

If one ignores the data-dependent selection and treats the selected target and reference indices as fixed, then under H_0,

$$\widetilde{T} \sim \chi_d.$$

A naive analysis therefore computes

$$p_{\text{naive}} = \mathbb{P}\Big(\chi_d \geq \widetilde{T}_{\text{obs}}\Big), \tag{2}$$

where $\widetilde{T}_{\text{obs}}$ is the observed standardized test statistic. However, both the target instance and its representative reference are selected from the observed data. Consequently, the fixed-selection χ_d distribution does not represent the post-selection distribution of $\widetilde{T}$, and p_{naive} does not in general control the Type-I error rate. In the next section, we account for these data-dependent selections using SI.

4 Statistical inference framework

We develop an SI framework that accounts for the data-dependent selection of both the target and reference instances. For the derivation, we first rewrite the statistical test introduced in Section 3 in vector form.

4.1 Vector-form formulation

Let

$$\boldsymbol{Y} = \text{vec}\left(\boldsymbol{X}_{\text{test}}, \boldsymbol{X}_{\text{ref}}^{1}, \ldots, \boldsymbol{X}_{\text{ref}}^{M_{\text{ref}}}\right) \in \mathbb{R}^{(M_{\text{ref}}+1)d},$$

and let $\boldsymbol{y}$ denote its observed value. For the selected representative reference j_{med}, the test statistic in Section 3 can be written as

$$T(\boldsymbol{Y}) = \|P\boldsymbol{Y}\|_2, \qquad \widetilde{T}(\boldsymbol{Y}) = \frac{\|P\boldsymbol{Y}\|_2}{\sigma},$$

where P is an orthogonal projection matrix satisfying

$$P\boldsymbol{Y} = \frac{1}{2}\,\text{vec}\left(\boldsymbol{X}_{\text{test}} - \boldsymbol{X}_{\text{ref}}^{j_{\text{med}}}, 0, \ldots, \boldsymbol{X}_{\text{ref}}^{j_{\text{med}}} - \boldsymbol{X}_{\text{test}}, \ldots, 0\right).$$

Thus,

$$\|P\boldsymbol{Y}\|_2 = \frac{1}{\sqrt{2}}\left\|\boldsymbol{X}_{\text{test}} - \boldsymbol{X}_{\text{ref}}^{j_{\text{med}}}\right\|_2,$$

which recovers the test statistic in Eq. (1). As discussed in Section 3, treating the adaptively determined projection matrix P as fixed leads to the invalid naive p-value in Eq. (2). We therefore account for the selection determining P using SI.

4.2 Selective inference

We represent the data-dependent selection procedure by the selection map

$$\mathcal{E}(\boldsymbol{Y}) = (\mathcal{A}(\boldsymbol{Y}), \mathcal{M}(\boldsymbol{Y})),$$

where $\mathcal{A}$ represents attention-based target selection and $\mathcal{M}$ represents adaptive reference selection. The reference-selection map $\mathcal{M}$ is defined by first retrieving the k nearest reference instances and then selecting their medoid. The corresponding selection event is

$$\mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}), \tag{3}$$

which requires the attention-selection outcome and the selected representative reference to remain identical to those observed from $\boldsymbol{y}$. The explicit definitions of the selection maps and the resulting constraints are provided in Appendix C.

Following the standard SI framework (Lee et al., 2016; Yang et al., 2016), we additionally condition on the nuisance statistic

$$\mathcal{Q}(\boldsymbol{Y}) = \left(\frac{P\boldsymbol{Y}}{\|P\boldsymbol{Y}\|_2}, (I-P)\boldsymbol{Y}\right).$$

Conditioning on $\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})$ fixes the direction of the projected component and the component orthogonal to the range of P, leaving only $\widetilde{T}(\boldsymbol{Y})$ random. Consequently, the conditional sample space reduces to a one-dimensional path, and the selection event in Eq. (3) restricts this path to a truncation region. We characterize this truncation region using parametric programming (Nguyen Le Duy & Takeuchi, 2021; 2022). Specifically, parametric programming explores the one-dimensional path of the test statistic and identifies all regions that preserve the observed selection outcome. Within the selection event defined by $\mathcal{E}(\boldsymbol{Y}) = (\mathcal{A}(\boldsymbol{Y}), \mathcal{M}(\boldsymbol{Y}))$, this avoids additional over-conditioning on a single local truncation region. The one-dimensional parameterization and the explicit characterization of the selection constraints are provided in Appendix C.

The selective p-value is defined as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}\left(\widetilde{T}(\boldsymbol{Y}) \geq \widetilde{T}(\boldsymbol{y}) \;\middle|\; \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}), \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\right).$$

Conditioning on the selection event fixes the selected representative reference instance and hence the projection matrix P. The conditional null distribution of the test statistic is characterized by the following theorem.

Theorem 4.1. *Under the null hypothesis* H_0*,*

$$\widetilde{T}(\boldsymbol{Y}) \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}), \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})$$

follows a χ_d *distribution truncated to the region determined by the conditioning events.*

Therefore, $p_{\text{selective}}$ can be computed as the upper-tail probability of the truncated χ_d distribution. The proof of Theorem 4.1 is provided in Appendix B.1.

The following theorem establishes the validity of the proposed selective p-value.

Theorem 4.2. *For any significance level* $\alpha \in (0, 1)$*,*

$$\mathbb{P}_{\text{H}_0}\left(p_{\text{selective}} \leq \alpha \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}), \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\right) = \alpha.$$

Hence, the proposed selective p*-value exactly controls the conditional Type-I error rate at level* α*.*

The proof of Theorem 4.2 is provided in Appendix B.2.

5 Experiments

5.1 Experimental settings

All experiments were conducted at the significance level $\alpha = 0.05$. We evaluated Type-I error rates and statistical power on synthetic and MNIST-based data, and rejection proportions on CAMELYON16, using the following methods. **Proposed** conditions on attention-based target selection and adaptive reference selection and uses parametric programming to characterize the entire truncation region. **Ablation 1** ignores attention thresholding, whereas **Ablation 2** ignores adaptive reference selection. **Over-conditioning (OC)** uses only the local truncation region containing the observed test statistic. **Bonferroni** applies a multiplicity correction for attention-based target selection and adaptive reference selection. **Naive** ignores all selection events and is reported only in the synthetic experiments. Detailed definitions of the methods are provided in Appendix D.1.

5.2 Synthetic experiments

We first evaluated the proposed method using synthetic data, for which Type-I error and statistical power can be directly assessed. Positive and negative instances were generated from $\mathcal{N}(\mathbf{1}, \sigma^2 I_d)$ and $\mathcal{N}(\mathbf{0}, \sigma^2 I_d)$, respectively, and an ABMIL model with a multilayer perceptron (MLP) backbone was trained on bags of 10 instances. High-attention instances were selected using the pre-sigmoid attention logit with threshold $\tau = 0$. For the main experiment, we fixed $d = 32$, $\sigma^2 = 1.0$, and $M_{\text{ref}} = 100$, and varied $k \in \{1, 3, 5, 10\}$. Negative and positive instances were used to evaluate the Type-I error rate and statistical power, respectively. Each experiment was repeated until 10,000 high-attention instances were collected. Additional experimental settings are provided in Appendix D.2.

Figure 2 shows the Type-I error rate and statistical power. The proposed method controlled the Type-I error rate near the nominal level across all values of k. Ablation 1 also remained close

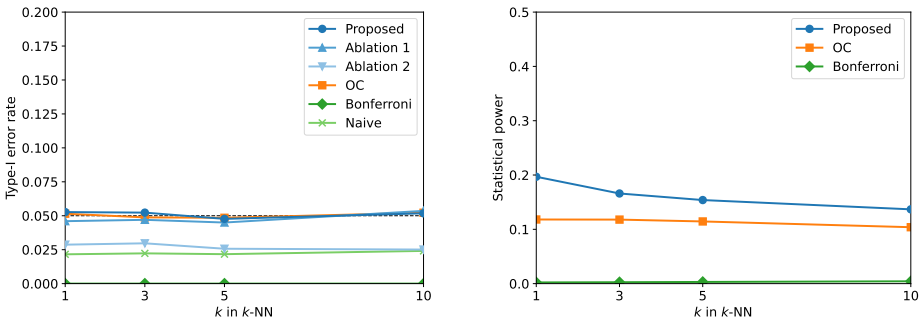


Figure 2: Type-I error rate (left) and statistical power (right) on synthetic data for varying numbers of nearest neighbors k. The dashed line indicates the nominal significance level of 0.05.

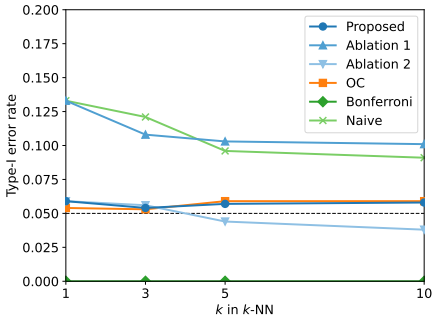


Figure 3: Type-I error rate under strong attention selection. The attention threshold was set to the 99.99th percentile of attention logits computed from independent negative instances, with $M_{\mathrm{ref}} = k \in \{1, 3, 5, 10\}$. The dashed line indicates the nominal significance level of 0.05.

to the nominal level in this setting, whereas Ablation 2 and Naive were conservative, indicating a substantial effect of adaptive reference selection. The proposed method consistently achieved higher statistical power than OC, while Bonferroni was highly conservative. These results demonstrate that the proposed method accounts for adaptive reference selection without the substantial loss of power caused by excessive conditioning or multiplicity correction.

Stress test under strong attention selection. The preceding experiment uses a relatively mild attention threshold, and the effect of attention-based selection is therefore limited. To examine a setting with strong selection bias, we set τ to the 99.99th percentile of attention logits obtained from independent negative instances. We further set $k = M_{\mathrm{ref}} \in \{1, 3, 5, 10\}$ so that k-NN retrieval becomes trivial while medoid selection remains adaptive for $k > 1$. For each setting, 1,000 negative instances satisfying the attention threshold were collected.

As shown in Figure 3, the proposed method maintained the Type-I error rate near the nominal level. In contrast, Naive and Ablation 1, both of which ignore attention-based selection, exhibited substantial Type-I error inflation, whereas Bonferroni was highly conservative. This result confirms that explicitly accounting for attention thresholding is necessary when its selection effect is strong.

5.3 MNIST-BASED EXPERIMENTS

We next evaluated the proposed method under a heterogeneous distribution constructed from MNIST (Deng, 2012). Digit "0" was used as the negative class, and each digit in $\{1, \ldots, 9\}$ was independently treated as the positive class. Instances were generated by adding isotropic Gaussian noise to MNIST images used as cluster centers. Negative test and reference instances were generated from three shared cluster centers. Type-I error was evaluated only for selected test–reference pairs from the same cluster. Images were downsampled to 14×14 and represented as 196-dimensional vectors. We fixed $\sigma^2 = 0.25$, $k = 5$, and $M_{\mathrm{ref}} = 100$, and set τ to the top-5% attention threshold

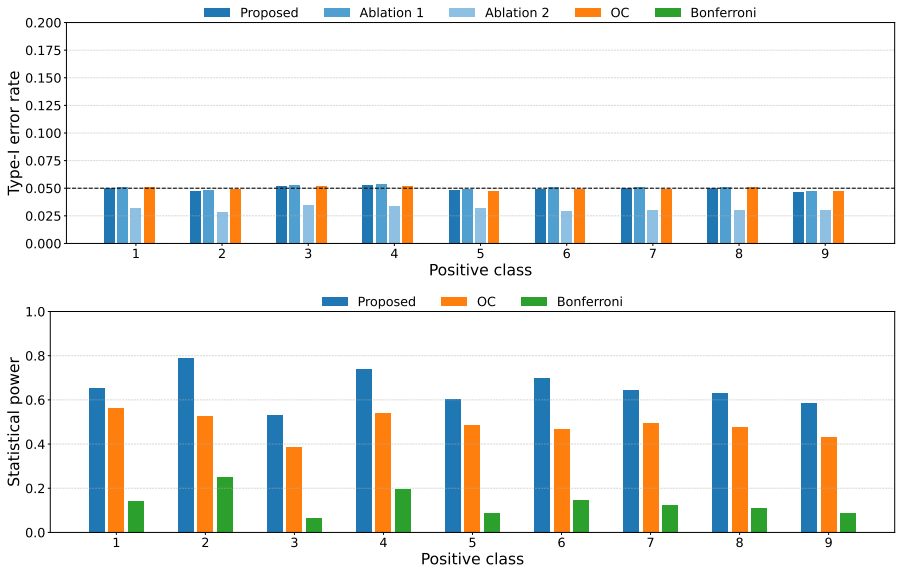


Figure 4: Type-I error rate (top) and statistical power (bottom) in the MNIST-based experiment with three Gaussian clusters. The horizontal axis denotes the positive digit class.

estimated from an independent set of negative instances. Further details and results for other settings are provided in Appendix D.3.

Figure 4 shows the results across the nine positive classes. The proposed method controlled the Type-I error rate near the nominal level across all classes. Ablation 1 showed similar Type-I error control, whereas Ablation 2 was conservative. The proposed method also consistently achieved higher statistical power than OC, while Bonferroni was overly conservative. These results demonstrate that the proposed framework remains valid and powerful under heterogeneous, image-derived data distributions.

5.4 Real-world experiments on CAMELYON16

Finally, we evaluated the proposed method on CAMELYON16 (Bejnordi et al., 2017) to demonstrate its applicability to real-world WSI analysis. We used WSIs from a single medical center and restricted the tumor cases to macro-metastases. The training data consisted of 59 normal and 26 tumor WSIs. Of the 59 normal WSIs, 37 were used for ABMIL training, 11 for reference construction, and 11 for independent noise-variance estimation and attention-threshold determination. The test set contained 30 normal and 9 tumor WSIs.

Each WSI was divided into image patches, from which 1,536-dimensional UNI2 features (Chen et al., 2024) were extracted and reduced to 192 dimensions by average pooling. These 192-dimensional vectors were used as the input instances for hypothesis testing and reference selection. The reference set contained 1,009 instances sampled from the 11 reference WSIs. The noise variance and attention threshold were estimated using the independent 11 normal WSIs and treated as fixed during inference. We report the results for $k = 3$ and a threshold corresponding to the top 5% of attention logits; additional implementation details are provided in Appendix D.4.

Table 1 summarizes the quantitative results. For normal WSIs, we report the rejection proportion among attention-selected patches as an empirical proxy for the Type-I error rate. For tumor WSIs, we report the corresponding rejection proportion as an empirical proxy for statistical power. Because the equality or inequality of the underlying signal vectors is not directly observed for individual patch pairs, these quantities are operational evaluation metrics. The proposed method yielded a rejection proportion close to 0.05 on normal WSIs while achieving a substantially higher rejection proportion than OC on tumor WSIs. Bonferroni yielded substantially lower rejection proportions.

Table 1: Quantitative results on CAMELYON16 with $k = 3$ and the top-5% attention threshold. Numbers in parentheses indicate the numbers of tested instances.

	Proposed	OC	Bonferroni
Normal rejection rate (297)	0.037	0.037	0.000
Tumor rejection rate (1962)	<u>0.570</u>	0.468	0.117

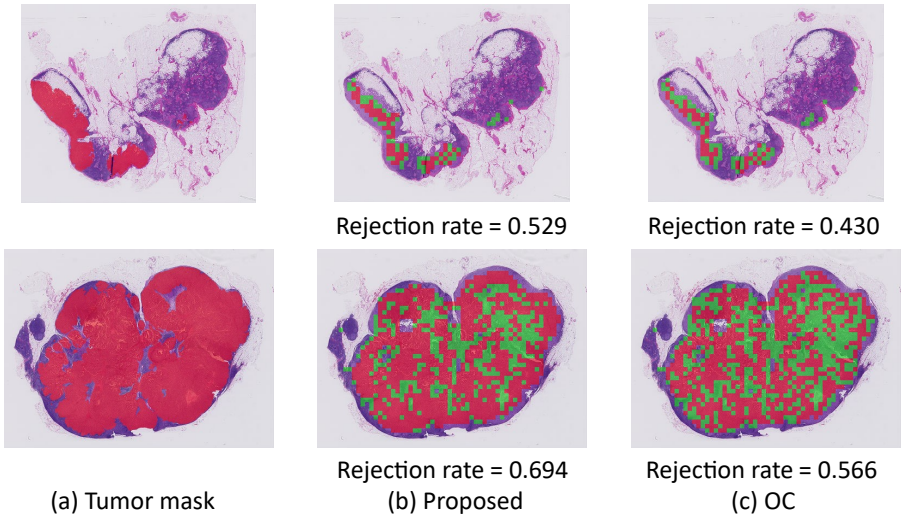


Figure 5: Visualization of two representative CAMELYON16 tumor WSIs. (a) Ground-truth tumor masks; (b) Proposed; and (c) OC. Red and green rectangles indicate attention-selected patches with $p < 0.05$ and $p \geq 0.05$, respectively. Tumor masks are used only for visualization.

Figure 5 visualizes the results for two representative tumor WSIs. Red and green rectangles indicate attention-selected patches with $p < 0.05$ and $p \geq 0.05$, respectively, while the tumor masks are shown only for visualization and were not used for inference. Compared with OC, the proposed method identifies more attention-selected patches as statistically significant, resulting in a higher rejection proportion. Importantly, some high-attention patches are not rejected by the proposed test, illustrating that high attention alone does not imply a statistically significant deviation from normal tissue. Results for all tumor WSIs, including those shown in Figure 5, together with the Bonferroni results, are provided in Appendix D.4.

6 Conclusion

In this paper, we proposed a selective inference framework for statistically evaluating high-attention instances in ABMIL with adaptive reference selection. The proposed method accounts for attention-based target selection and adaptive reference selection using parametric programming to compute valid selective p-values. Experiments demonstrated Type-I error control on synthetic and MNIST-based data and practical applicability to CAMELYON16, with higher statistical power than the conventional over-conditioning approach.

A limitation of this study is that we focused on ABMIL as a proof-of-concept architecture, whose relatively simple attention mechanism allows its selection event to be explicitly characterized. More generally, the proposed framework is applicable when the data-dependent selection procedure can be characterized along the one-dimensional conditional path. In particular, neural networks composed of affine operations, such as fully connected and convolutional layers, and piecewise-linear activations such as ReLU are piecewise affine with respect to the input, allowing the corresponding selection constraints and truncation region to be characterized by parametric programming. Extending the framework to more complex MIL architectures and foundation-model-based pathology models (Chen et al., 2024; Vorontsov et al., 2024; Ding et al., 2025), as well as broader retrieval and explanation mechanisms, is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research execution, particularly the implementation and debugging of experimental code, and to draft and refine portions of the manuscript. This writing assistance included drafting proof descriptions based on established arguments in the selective inference literature. We did not use generative AI tools to generate the synthetic datasets used in our experiments. We additionally used generative AI tools for English-language editing and polishing. All AI-assisted text, including the proof descriptions, was reviewed and revised by the authors, and AI-assisted code was verified and tested for correctness before being used in the experiments. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a statistical inference methodology and does not involve the collection of new human-subject data. The real-world experiments use the publicly available CAMELYON16 dataset. The proposed method is intended to provide statistically valid inference for data-dependent instance selection in MIL and is not intended to provide clinical diagnoses on its own. No patient-identifiable information was used in this study.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, complete proofs of the theoretical results are provided in Appendix B, and detailed characterizations of the selection events are provided in Appendix C. Detailed experimental settings, implementation and inference details, computational environment, and additional results are provided in Appendix D. The datasets used in the experiments are publicly available or generated according to the procedures described in the paper and appendices. Anonymous source code for reproducing the experiments is provided in the supplementary material.

REFERENCES

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002.

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

Linghan Cai, Shenjin Huang, Ye Zhang, Jinpeng Lu, and Yongbing Zhang. Rethinking attention-based multiple instance learning for whole-slide pathological image classification: An instance attribute viewpoint. *arXiv preprint arXiv:2404.00351*, 2024.

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

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155, 2022.

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

Shuxiao Chen and Jacob Bien. Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, 29(2):323–334, 2020. doi: 10.1080/10618600.2019.1660180.

Yiqun T. Chen and Daniela M. Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

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

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.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 31(11):3749–3761, 2025.

Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11356–11367. Curran Associates, Inc., 2020.

Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi. Quantifying statistical significance of neural network-based image segmentation by selective inference. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31627–31639. Curran Associates, Inc., 2022. doi: 10.52202/068431-2293.

Ji Feng and Zhi-Hua Zhou. Deep miml network. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

Lucy L. Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024. doi: 10.1080/01621459.2022.2116331.

Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3852–3861, 2020.

Noriaki Hashimoto, Hiroyuki Hanada, Hiroaki Miyoshi, Miharu Nagaishi, Kensaku Sato, Hidekata Hontani, Koichi Ohshima, and Ichiro Takeuchi. Multimodal gated mixture of experts using whole slide image and flow cytometry for multiple instance learning classification of lymphoma. *Journal of Pathology Informatics*, 15:100359, 2024.

Sangwon Hyun, Max G'Sell, and Ryan J Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

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.

Teruyuki Katsuoka, Tomohiro Shiraishi, Daiki Miwa, Shuichi Nishino, and Ichiro Takeuchi. si4onnx: A python package for selective inference in deep learning models. *arXiv preprint arXiv:2501.17415*, 2025. doi: 10.48550/arXiv.2501.17415.

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

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, 2009. doi: 10.1038/nn.2303.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

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.

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.

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

Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Valid p-value for deep learning-driven salient region. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=qihMOPw4Sf_`.

Vo Nguyen Le Duy and Ichiro Takeuchi. Parametric programming approach for more powerful and general lasso selective inference. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 901–909. PMLR, 2021. URL https://proceedings.mlr.press/v130/nguyen-le-duy21a.html.

Vo Nguyen Le Duy and Ichiro Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *Journal of Machine Learning Research*, 23(300):1–37, 2022.

Mizuki Niihori, Shuichi Nishino, Teruyuki Katsuoka, Tomohiro Shiraishi, Kouichi Taji, and Ichiro Takeuchi. Quantifying statistical significance of deep nearest neighbor anomaly detection via selective inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 173174–173203. Curran Associates, Inc., 2025. doi: 10.52202/085713-5762.

Shuichi Nishino, Tomohiro Shiraishi, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for saliency maps of graph neural networks via selective inference. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=5NkXTCVa7F.

Ronan Perry, Snigdha Panigrahi, Jacob Bien, and Daniela Witten. Inference on the proportion of variance explained in principal component analysis. *Journal of the American Statistical Association*, 121(553):667–677, 2026. doi: 10.1080/01621459.2025.2538895.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2931–2951, 2019.

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.

Tomohiro Shiraishi, Daiki Miwa, Teruyuki Katsuoka, Vo Nguyen Le Duy, Kouichi Taji, and Ichiro Takeuchi. Statistical test for attention maps in vision transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45079–45096. PMLR, 2024. URL https://proceedings.mlr.press/v235/shiraishi24a.html.

Tomohiro Shiraishi, Daiki Miwa, Teruyuki Katsuoka, Vo Nguyen Le Duy, Shuichi Nishino, Kouichi Taji, and Ichiro Takeuchi. Statistical test for attention in transformers for images and time series. *Journal of Machine Learning Research*, 27(119):1–43, 2026.

PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.

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.

Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *International Conference on Machine Learning*, pp. 3338–3347. PMLR, 2017.

Yusuke Takagi, Noriaki Hashimoto, Hiroki Masuda, Hiroaki Miyoshi, Koichi Ohshima, Hidekata Hontani, and Ichiro Takeuchi. Transformer-based personalized attention mechanism for medical images with clinical records. *Journal of Pathology Informatics*, 14:100185, 2023.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. doi: 10.1073/pnas.1507583112.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018. doi: 10.1214/17-AOS1564.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. doi: 10.1080/01621459.2015.1108848.

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

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

Zhenzhen Wang, Carla Saoud, Sintawat Wangsiricharoen, Aaron W James, Aleksander S Popel, and Jeremias Sulam. Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. *IEEE transactions on medical imaging*, 41(12):3952–3968, 2022.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3460–3469, 2015.

Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post selection inference with kernels. In *International Conference on Artificial Intelligence and Statistics*, pp. 152–160. PMLR, 2018.

Yoichiro Yamamoto, Toyonori Tsuzuki, Jun Akatsuka, Masao Ueki, Hiromu Morikawa, Yasushi Numata, Taishi Takahara, Takuji Tsuyuki, Kotaro Tsutsumi, Ryuto Nakazawa, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature communications*, 10(1):5642, 2019.

Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2469–2477, 2016.

Philip Zehnder, Jeffrey Feng, Trung Nguyen, Philip Shen, Ruth Sullivan, Reina N Fuji, and Fangyao Hu. Diagnostic classification in toxicologic pathology using attention-guided weak supervision and whole slide image features: a pilot study in rat livers. *Scientific Reports*, 15(1):4202, 2025.

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.

Zhi-Hua Zhou and Min-Ling Zhang. Neural networks for multi-instance learning. In *Proceedings of the International Conference on Intelligent Information Technology, Beijing, China*, pp. 455–459. Citeseer, 2002.

A SUMMARY OF NOTATION

This appendix summarizes the principal notation used throughout the paper. Bold lowercase and uppercase letters denote vectors and random vectors, respectively, while calligraphic letters denote sets, selection maps, or selection outcomes. For a positive integer N, $[N] := \{1, \ldots, N\}$.

A.1 MULTIPLE INSTANCE LEARNING

Symbol	Description
N, M^n	Number of bags and number of instances in the n-th bag.
$X^n = \{\boldsymbol{x}^{n,m}\}_{m\in[M^n]}$	The n-th bag and its constituent input instances.
Y^n	Bag-level label of the n-th bag.
f_{enc}, f_{att}	Feature extractor and attention network.
$\boldsymbol{h}^{n,m}$	Feature representation of the m-th instance in the n-th bag.
$a^{n,m}$	Pre-sigmoid attention logit of the m-th instance in the n-th bag.
τ, $\mathcal{C}(X)$	Attention threshold and the set of indices of high-attention instances in bag X.
d, d'	Input dimension and encoded feature dimension.

A.2 TEST AND REFERENCE INSTANCES

Symbol	Description
$X_{\mathrm{test}} = \{\boldsymbol{x}^i_{\mathrm{test}}\}_{i\in[M_{\mathrm{test}}]}$	Test bag containing M_{test} observed instances.
$X_{\mathrm{ref}} = \{\boldsymbol{x}^j_{\mathrm{ref}}\}_{j\in[M_{\mathrm{ref}}]}$	Reference set containing M_{ref} observed normal instances.
$\boldsymbol{h}^i_{\mathrm{test}}$, $\boldsymbol{h}^j_{\mathrm{ref}}$	Encoded features of test instance i and reference instance j.
a^i_{test}	Pre-sigmoid attention logit of test instance i.
k, j_{med}	Number of nearest neighbors and index of the selected medoid reference instance.
$\boldsymbol{X}_{\mathrm{test}}$, $\boldsymbol{X}^j_{\mathrm{ref}}$	Random vector corresponding to the selected test instance and random vectors corresponding to the reference instances.
$\boldsymbol{s}_{\mathrm{test}}$, $\boldsymbol{s}^j_{\mathrm{ref}}$	Unknown signal vectors of the selected test instance and reference instance j.
$\boldsymbol{\epsilon}_{\mathrm{test}}$, $\boldsymbol{\epsilon}^j_{\mathrm{ref}}$	Independent Gaussian noise vectors with covariance $\sigma^2 I_d$.
σ^2, I_d	Noise variance and the d-dimensional identity matrix.
H_0, H_1	Null and alternative hypotheses comparing the signal vectors of the selected test instance and selected medoid reference instance.

A.3 TEST STATISTICS AND SELECTIVE INFERENCE

Symbol	Description
$\boldsymbol{Y}$, $\boldsymbol{y}$	Concatenated random vector formed from the selected test instance and all reference instances, and its observed value.
$\mathrm{vec}(\cdot)$	Vector concatenation operator used to construct $\boldsymbol{Y}$.
P	Orthogonal projection matrix representing the contrast between the test instance and the selected medoid reference instance.
$T(\boldsymbol{Y})$, $\widetilde{T}(\boldsymbol{Y})$	Test statistic $\|P\boldsymbol{Y}\|_2$ and its standardized version $\|P\boldsymbol{Y}\|_2/\sigma$.
p_{naive}, $p_{\mathrm{selective}}$	Naive p-value and selective p-value.
$p_{\mathrm{Bonferroni}}$	Bonferroni-corrected p-value.
$\mathcal{A}(\boldsymbol{Y})$, $\mathcal{A}(\boldsymbol{y})$	Attention selection map and its observed outcome.
$\mathcal{K}(\boldsymbol{Y})$, $\mathcal{K}(\boldsymbol{y})$	k-nearest-neighbor retrieval map and its observed set.
$\mathcal{M}(\boldsymbol{Y})$, $\mathcal{M}(\boldsymbol{y})$	Adaptive reference-selection map, defined by nearest-neighbor retrieval followed by medoid selection, and its observed outcome.
$\mathcal{E}(\boldsymbol{Y})$, $\mathcal{E}(\boldsymbol{y})$	Overall selection map $(\mathcal{A}, \mathcal{M})$ and its observed outcome.
$\mathcal{Q}(\boldsymbol{Y})$, $\mathcal{Q}(\boldsymbol{y})$	Nuisance statistic and its observed value, consisting of the direction of $P\boldsymbol{Y}$ and the component orthogonal to the range of P.
α	Significance level.
λ_{att}	Multiplicity factor accounting for attention-based target selection in the Bonferroni correction.

A.4 ONE-DIMENSIONAL CONDITIONAL PATH AND SELECTION EVENTS

Symbol	Description
$\boldsymbol{a}$, $\boldsymbol{b}$, z, $\boldsymbol{Y}(z)$	Quantities defining the conditional path $\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z$, where $z = \widetilde{T}(\boldsymbol{Y}) > 0$.
$\boldsymbol{X}_{\mathrm{test}}(z)$, $\boldsymbol{X}^j_{\mathrm{ref}}(z)$	Conditional-path representations of the test instance and the j-th reference instance.
$j^{\mathrm{obs}}_{\mathrm{med}}$	Index of the observed representative reference, $j^{\mathrm{obs}}_{\mathrm{med}} = \mathcal{M}(\boldsymbol{y})$.
$\mathcal{Z}_{\mathcal{A}}$	Feasible region preserving the observed attention-selection outcome.
$\mathcal{Z}_{\mathcal{K}}(K)$	Local region on the conditional path over which the k-nearest-neighbor set is fixed to K.
$\mathcal{Z}_{\mathcal{M}\mid\mathcal{K}}(K)$	Local region within $\mathcal{Z}_{\mathcal{K}}(K)$ in which $j^{\mathrm{obs}}_{\mathrm{med}}$ is selected as the medoid.
$\mathcal{Z}_{\mathcal{M}}$	Feasible region preserving the observed selected representative reference.
$\mathcal{Z}$	Overall truncation region $\mathcal{Z}_{\mathcal{A}} \cap \mathcal{Z}_{\mathcal{M}}$.
$\boldsymbol{\alpha}_{\mathrm{test}}$, $\boldsymbol{\beta}_{\mathrm{test}}$	Affine coefficients of the encoded test feature on the conditional path.
$\boldsymbol{\alpha}^j_{\mathrm{ref}}$, $\boldsymbol{\beta}^j_{\mathrm{ref}}$	Affine coefficients of the encoded feature of reference instance j.
$D^j(z)$, D^j_r $(r = 0, 1, 2)$	Squared feature-space distance from the test instance to reference instance j and its constant, linear, and quadratic coefficients.
$S^j_K(z)$, $S^j_{K,r}$ $(r = 0, 1, 2)$	Medoid score of candidate j within a fixed k-nearest-neighbor set K and its constant, linear, and quadratic coefficients.
ℓ_r, u_r	Lower and upper endpoints of the r-th interval constituting the truncation region $\mathcal{Z}$.
R	Number of disjoint intervals constituting the truncation region.

B PROOFS OF THEOREMS

Noise variance. The theoretical guarantees in Section 4.2 assume that the true noise variance σ^2 is known. In the synthetic and MNIST-based experiments, the true variance used for data generation is also used for inference. For CAMELYON16, an estimate obtained from independent normal

instances is substituted for the true variance. The proofs below do not account for uncertainty in this estimate and do not establish exact finite-sample validity for this plug-in implementation.

B.1 Proof of Theorem 4.1

Fix the observed selection outcome and let P denote the corresponding deterministic projection matrix throughout this proof. The unconditional distributional calculations below use this fixed matrix before conditioning on the selection event.

We next condition on $\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})$, which gives

$$\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y}) \quad \Longleftrightarrow \quad \left(\frac{P\boldsymbol{Y}}{\|P\boldsymbol{Y}\|_2}, (I-P)\boldsymbol{Y}\right) = \left(\frac{P\boldsymbol{y}}{\|P\boldsymbol{y}\|_2}, (I-P)\boldsymbol{y}\right) \quad \Longleftrightarrow \quad \boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z,$$

where

$$\begin{aligned} \boldsymbol{a} &= (I-P)\boldsymbol{y}, \\ \boldsymbol{b} &= \sigma \frac{P\boldsymbol{y}}{\|P\boldsymbol{y}\|_2}, \\ z &= \widetilde{T}(\boldsymbol{Y}) \in \mathbb{R}_{>0}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\left\{\boldsymbol{Y} \in \mathbb{R}^{(M_{\text{ref}}+1)d} \;\middle|\; \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}),\ \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\right\} \\ &= \{\boldsymbol{a} + \boldsymbol{b}z \mid z \in \mathcal{Z}\}, \end{aligned}$$

where

$$\mathcal{Z} = \{z \in \mathbb{R}_{>0} \mid \mathcal{E}(\boldsymbol{a} + \boldsymbol{b}z) = \mathcal{E}(\boldsymbol{y})\},$$

is the truncation region determined by the observed selection event.

Under the Gaussian model with known σ^2 and the null hypothesis associated with the fixed selection outcome, $P\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{0}$. Before conditioning on the selection event, $P\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 P)$. Since P is an orthogonal projection matrix of rank d, $\widetilde{T}(\boldsymbol{Y}) = \|P\boldsymbol{Y}\|_2/\sigma$ follows a χ_d distribution. Moreover, $P\boldsymbol{Y}$ and $(I-P)\boldsymbol{Y}$ are independent, and the norm of $P\boldsymbol{Y}$ is independent of its direction. Therefore, $\widetilde{T}(\boldsymbol{Y})$ is independent of $\mathcal{Q}(\boldsymbol{Y})$ before conditioning on the selection event. Conditioning additionally on $\mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y})$ restricts the support of $\widetilde{T}(\boldsymbol{Y})$ to $\mathcal{Z}$, yielding the stated truncated χ_d distribution.

B.2 Proof of Theorem 4.2

By Theorem 4.1, under the null hypothesis,

$$\widetilde{T}(\boldsymbol{Y}) \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}),\ \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y}),$$

follows a χ distribution with d degrees of freedom truncated to $\mathcal{Z}$. Since the selective p-value is defined as the upper-tail probability of this conditional distribution, the probability integral transform gives

$$\mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}),\ \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\right) = \alpha,$$

for any $\alpha \in (0,1)$.

Since this equality holds for all possible values of $\mathcal{Q}(\boldsymbol{Y})$, the law of iterated expectations gives

$$\begin{aligned} &\mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y})\right) \\ &= \mathbb{E}_{\mathrm{H}_0}\left[\mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}),\ \mathcal{Q}(\boldsymbol{Y})\right) \mid \mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y})\right] \\ &= \alpha. \end{aligned}$$

Therefore, the selective p-value exactly controls the conditional Type-I error rate given the observed selection event. This guarantee holds for each selection outcome for which the corresponding null hypothesis is true.

C SELECTION EVENT CHARACTERIZATION

In this appendix, we provide the explicit definitions of the selection maps introduced in Section 4.2 and characterize the corresponding selection events along the one-dimensional conditional path used for selective inference. Recall that the overall selection map is

$$\mathcal{E}(\boldsymbol{Y}) = (\mathcal{A}(\boldsymbol{Y}), \mathcal{M}(\boldsymbol{Y}))\,,$$

where $\mathcal{A}$ represents attention thresholding and $\mathcal{M}$ represents adaptive reference selection. The reference-selection map $\mathcal{M}$ is determined by first retrieving the k nearest reference instances and then selecting their medoid. We first define these maps explicitly and then derive the corresponding constraints that determine the truncation region.

C.1 SELECTION MAPS

Throughout selective inference, the trained feature extractor f_{enc} and attention network f_{att} are treated as fixed.

Attention thresholding. The attention selection map determines whether the test instance is selected according to its pre-sigmoid attention logit:

$$\mathcal{A}(\boldsymbol{Y}) = \mathbb{I}\left\{f_{\mathrm{att}}\left(f_{\mathrm{enc}}(\boldsymbol{X}_{\mathrm{test}})\right) > \tau\right\},$$

where τ is fixed before selective inference. Depending on the experimental setting, τ is either specified a priori or estimated from an independent dataset.

k-nearest-neighbor retrieval. The k-nearest-neighbor map is

$$\mathcal{K}(\boldsymbol{Y}) = \mathrm{Top}_k\left(\left\{\left\|f_{\mathrm{enc}}(\boldsymbol{X}_{\mathrm{test}}) - f_{\mathrm{enc}}(\boldsymbol{X}_{\mathrm{ref}}^{j})\right\|_2^2\right\}_{j\in[M_{\mathrm{ref}}]}\right),$$

where Top_k returns the indices corresponding to the k smallest squared Euclidean distances. The map $\mathcal{K}(\boldsymbol{Y})$ is used internally to determine the representative reference but is not itself included in the conditioning event.

Medoid selection. Among the retrieved nearest neighbors, the representative reference instance is selected as

$$\mathcal{M}(\boldsymbol{Y}) = \arg\min_{j\in\mathcal{K}(\boldsymbol{Y})} \sum_{j'\in\mathcal{K}(\boldsymbol{Y})} \left\|f_{\mathrm{enc}}(\boldsymbol{X}_{\mathrm{ref}}^{j}) - f_{\mathrm{enc}}(\boldsymbol{X}_{\mathrm{ref}}^{j'})\right\|_2^2.$$

Any ties in k-nearest-neighbor retrieval or medoid selection are resolved using a prespecified deterministic rule. The selection inequalities in the following subsections are interpreted consistently with this rule. A strict inequality is used whenever a tie would favor a competing index over the corresponding selected index.

The definitions above use distances in the encoded feature space. The k-nearest-neighbor retrieval and medoid selection can also be performed in the input space by replacing $f_{\mathrm{enc}}(\boldsymbol{X})$ with $\boldsymbol{X}$ in the corresponding maps. The choice of space is fixed before selective inference. The attention-selection map remains unchanged.

The observed overall selection event is therefore

$$\mathcal{E}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{y}),$$

which requires the attention-thresholding outcome and the selected representative reference to remain identical to those observed from $\boldsymbol{y}$. The k-nearest-neighbor set is allowed to change as long as the final selected representative reference remains unchanged.

C.2 ONE-DIMENSIONAL CONDITIONAL PATH

As described in Section 4.2, selective inference additionally conditions on

$$\mathcal{Q}(\boldsymbol{Y}) = \left(\frac{P\boldsymbol{Y}}{\|P\boldsymbol{Y}\|_2}, (I-P)\boldsymbol{Y} \right).$$

Conditioning on $\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})$ reduces the conditional sample space to the one-dimensional path

$$\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z, \qquad z \in \mathbb{R}_{>0}, \tag{4}$$

where

$$\boldsymbol{a} = (I-P)\boldsymbol{y}, \qquad \boldsymbol{b} = \sigma \frac{P\boldsymbol{y}}{\|P\boldsymbol{y}\|_2}, \qquad z = \widetilde{T}(\boldsymbol{Y}).$$

Thus, z is the only remaining random quantity after conditioning on $\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})$.

Let $\boldsymbol{X}_{\mathrm{test}}(z)$ and $\boldsymbol{X}^j_{\mathrm{ref}}(z)$ denote the test instance and the j-th reference instance contained in $\boldsymbol{Y}(z)$, respectively. Along the path in Eq. (4), the observed selection event is preserved when the attention-thresholding outcome and the selected representative reference remain unchanged. The k-nearest-neighbor set may change along the path. We therefore use regions with a fixed k-nearest-neighbor set as local partitions for characterizing the adaptive reference-selection event.

C.3 ATTENTION SELECTION EVENT

Along the conditional path $\boldsymbol{Y}(z)$, the attention logit of the test instance is

$$f_{\mathrm{att}}\left(f_{\mathrm{enc}}\left(\boldsymbol{X}_{\mathrm{test}}(z)\right)\right).$$

The observed attention-selection outcome is preserved when

$$\mathbb{I}\left\{f_{\mathrm{att}}\left(f_{\mathrm{enc}}\left(\boldsymbol{X}_{\mathrm{test}}(z)\right)\right) > \tau\right\} = \mathcal{A}(\boldsymbol{y}).$$

Since both f_{enc} and f_{att} are piecewise affine, the attention logit is affine in z within each activation region. Therefore, the feasible region associated with the attention selection event is

$$\mathcal{Z}_{\mathcal{A}} = \left\{z \in \mathbb{R}_{>0} \mid \mathcal{A}(\boldsymbol{a} + \boldsymbol{b}z) = \mathcal{A}(\boldsymbol{y})\right\}.$$

C.4 k-NEAREST-NEIGHBOR PARTITION

For each reference instance, define its squared Euclidean distance from the test instance along the conditional path by

$$D^j(z) = \left\| f_{\mathrm{enc}}\left(\boldsymbol{X}_{\mathrm{test}}(z)\right) - f_{\mathrm{enc}}\left(\boldsymbol{X}^j_{\mathrm{ref}}(z)\right) \right\|_2^2. \tag{5}$$

Using squared Euclidean distances does not change the nearest-neighbor ordering compared with ordinary Euclidean distances.

For any index set $K \subseteq [M_{\mathrm{ref}}]$ with $|K| = k$, the k-nearest-neighbor set is equal to K when every reference instance in K is no farther from the test instance than every candidate outside K:

$$D^j(z) \le D^{j'}(z), \qquad j \in K, \quad j' \in [M_{\mathrm{ref}}] \setminus K. \tag{6}$$

Within each activation region, the encoded features are affine functions of z:

$$f_{\mathrm{enc}}\left(\boldsymbol{X}_{\mathrm{test}}(z)\right) = \boldsymbol{\alpha}_{\mathrm{test}} + \boldsymbol{\beta}_{\mathrm{test}} z, \tag{7}$$

$$f_{\mathrm{enc}}\left(\boldsymbol{X}^j_{\mathrm{ref}}(z)\right) = \boldsymbol{\alpha}^j_{\mathrm{ref}} + \boldsymbol{\beta}^j_{\mathrm{ref}} z. \tag{8}$$

Here, $\boldsymbol{\alpha}_{\mathrm{test}}$, $\boldsymbol{\beta}_{\mathrm{test}}$, $\boldsymbol{\alpha}^j_{\mathrm{ref}}$, and $\boldsymbol{\beta}^j_{\mathrm{ref}}$ are fixed within the corresponding activation region and are obtained by propagating $\boldsymbol{a}$ and $\boldsymbol{b}$ through the encoder.

Substituting Eqs. (7)–(8) into Eq. (5) gives

$$\begin{aligned} D^j(z) &= \left\| \left(\boldsymbol{\alpha}_{\mathrm{test}} - \boldsymbol{\alpha}^j_{\mathrm{ref}}\right) + \left(\boldsymbol{\beta}_{\mathrm{test}} - \boldsymbol{\beta}^j_{\mathrm{ref}}\right) z \right\|_2^2 \\ &= D^j_0 + D^j_1 z + D^j_2 z^2, \end{aligned}$$

where

$$D_0^j = \left\| \boldsymbol{\alpha}_{\text{test}} - \boldsymbol{\alpha}_{\text{ref}}^j \right\|_2^2,$$
$$D_1^j = 2\left(\boldsymbol{\alpha}_{\text{test}} - \boldsymbol{\alpha}_{\text{ref}}^j \right)^\top \left(\boldsymbol{\beta}_{\text{test}} - \boldsymbol{\beta}_{\text{ref}}^j \right),$$
$$D_2^j = \left\| \boldsymbol{\beta}_{\text{test}} - \boldsymbol{\beta}_{\text{ref}}^j \right\|_2^2.$$

Therefore, each ordering condition in Eq. (6) is equivalent to the quadratic inequality

$$\left(D_0^j - D_0^{j'} \right) + \left(D_1^j - D_1^{j'} \right) z + \left(D_2^j - D_2^{j'} \right) z^2 \leq 0.$$

For a fixed k-nearest-neighbor set K, define the corresponding local region by

$$\mathcal{Z}_{\mathcal{K}}(K) = \bigcap_{\substack{j \in K \\ j' \in [M_{\text{ref}}] \setminus K}} \left\{ z \in \mathbb{R}_{>0} \;\middle|\; D^j(z) - D^{j'}(z) \leq 0 \right\}.$$

These regions partition the conditional path according to the k-nearest-neighbor set. Importantly, $\mathcal{Z}_{\mathcal{K}}(K)$ is used only as a local partition for characterizing the reference-selection map; the k-nearest-neighbor set itself is not conditioned upon.

If k-nearest-neighbor retrieval is performed in the input space, the same derivation applies by replacing the encoded test and reference features with $\boldsymbol{X}_{\text{test}}(z)$ and $\boldsymbol{X}_{\text{ref}}^j(z)$, respectively.

C.5 Medoid selection event

Within a region where the k-nearest-neighbor set is fixed to K, the medoid is the reference instance that minimizes the sum of squared Euclidean distances to the other instances in K.

For each candidate $j \in K$, define its medoid score by

$$S_K^j(z) = \sum_{j' \in K} \left\| f_{\text{enc}}\left(\boldsymbol{X}_{\text{ref}}^j(z) \right) - f_{\text{enc}}\left(\boldsymbol{X}_{\text{ref}}^{j'}(z) \right) \right\|_2^2.$$

Let

$$j_{\text{med}}^{\text{obs}} = \mathcal{M}(\boldsymbol{y})$$

denote the observed representative reference. For a fixed K containing $j_{\text{med}}^{\text{obs}}$, the observed reference-selection outcome is preserved when

$$S_K^{j_{\text{med}}^{\text{obs}}}(z) \leq S_K^j(z), \qquad j \in K \setminus \{j_{\text{med}}^{\text{obs}}\}. \tag{9}$$

If $j_{\text{med}}^{\text{obs}} \notin K$, the observed reference-selection outcome cannot be preserved within that k-nearest-neighbor region.

Using the affine representation in Eq. (8), each medoid score can be written as

$$\begin{aligned} S_K^j(z) &= \sum_{j' \in K} \left\| \left(\boldsymbol{\alpha}_{\text{ref}}^j - \boldsymbol{\alpha}_{\text{ref}}^{j'} \right) + \left(\boldsymbol{\beta}_{\text{ref}}^j - \boldsymbol{\beta}_{\text{ref}}^{j'} \right) z \right\|_2^2 \\ &= S_{K,0}^j + S_{K,1}^j z + S_{K,2}^j z^2, \end{aligned}$$

where

$$S_{K,0}^j = \sum_{j' \in K} \left\| \boldsymbol{\alpha}_{\text{ref}}^j - \boldsymbol{\alpha}_{\text{ref}}^{j'} \right\|_2^2,$$
$$S_{K,1}^j = 2 \sum_{j' \in K} \left(\boldsymbol{\alpha}_{\text{ref}}^j - \boldsymbol{\alpha}_{\text{ref}}^{j'} \right)^\top \left(\boldsymbol{\beta}_{\text{ref}}^j - \boldsymbol{\beta}_{\text{ref}}^{j'} \right),$$
$$S_{K,2}^j = \sum_{j' \in K} \left\| \boldsymbol{\beta}_{\text{ref}}^j - \boldsymbol{\beta}_{\text{ref}}^{j'} \right\|_2^2.$$

Thus, Eq. (9) reduces to

$$\left(S_{K,0}^{j_{\mathrm{med}}^{\mathrm{obs}}} - S_{K,0}^{j}\right) + \left(S_{K,1}^{j_{\mathrm{med}}^{\mathrm{obs}}} - S_{K,1}^{j}\right) z + \left(S_{K,2}^{j_{\mathrm{med}}^{\mathrm{obs}}} - S_{K,2}^{j}\right) z^2 \leq 0.$$

For a fixed K containing $j_{\mathrm{med}}^{\mathrm{obs}}$, define the corresponding local medoid region by

$$\mathcal{Z}_{\mathcal{M}|\mathcal{K}}(K) = \bigcap_{j \in K \setminus \{j_{\mathrm{med}}^{\mathrm{obs}}\}} \left\{ z \in \mathbb{R}_{>0} \;\middle|\; S_K^{j_{\mathrm{med}}^{\mathrm{obs}}}(z) - S_K^{j}(z) \leq 0 \right\}.$$

Since the k-nearest-neighbor set is not conditioned upon, the overall region preserving the observed representative reference is obtained by combining all k-nearest-neighbor regions for which the same medoid is selected:

$$\mathcal{Z}_{\mathcal{M}} = \bigcup_{\substack{K \subseteq [M_{\mathrm{ref}}],\, |K|=k \\ j_{\mathrm{med}}^{\mathrm{obs}} \in K}} \left[\mathcal{Z}_{\mathcal{K}}(K) \cap \mathcal{Z}_{\mathcal{M}|\mathcal{K}}(K) \right].$$

Thus, the k-nearest-neighbor set may change across different parts of $\mathcal{Z}_{\mathcal{M}}$ as long as the final selected medoid remains $j_{\mathrm{med}}^{\mathrm{obs}}$.

If medoid selection is performed in the input space, the same derivation applies by replacing the encoded reference features $f_{\mathrm{enc}}(\boldsymbol{X}_{\mathrm{ref}}^{j}(z))$ with $\boldsymbol{X}_{\mathrm{ref}}^{j}(z)$.

Throughout this appendix, an intersection over an empty index set is interpreted as $\mathbb{R}_{>0}$. In particular, this convention applies to $\mathcal{Z}_{\mathcal{K}}(K)$ when $k = M_{\mathrm{ref}}$ and to $\mathcal{Z}_{\mathcal{M}|\mathcal{K}}(K)$ when $k = 1$.

C.6 Overall selection event

Preserving the observed overall selection outcome $\mathcal{E}(\boldsymbol{y})$ requires simultaneous preservation of the attention selection and the selected representative reference. The resulting truncation region is

$$\begin{aligned} \mathcal{Z} &= \{ z \in \mathbb{R}_{>0} \mid \mathcal{E}(\boldsymbol{a} + \boldsymbol{b}z) = \mathcal{E}(\boldsymbol{y}) \} \\ &= \mathcal{Z}_{\mathcal{A}} \cap \mathcal{Z}_{\mathcal{M}}. \end{aligned}$$

Within each activation region, the attention constraint is characterized by linear inequalities, whereas the boundaries associated with k-nearest-neighbor retrieval and medoid selection are characterized by quadratic inequalities. Because these inequalities can generate disconnected feasible subsets, the overall truncation region generally takes the form

$$\mathcal{Z} = \bigcup_{r=1}^{R} [\ell_r, u_r], \tag{10}$$

where $R \in \mathbb{N}$ is the number of disjoint feasible intervals and ℓ_r and u_r are the lower and upper endpoints of the r-th interval. Whether individual endpoints are included does not affect the selective p-value because the null distribution is continuous.

C.7 Computation of the truncation region

We characterize the entire selection-dependent truncation region using parametric programming. Along the one-dimensional path $\boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z$, the affine representations of the encoded features and attention logits remain fixed within each activation region of the feature extractor and attention network. Consequently, the attention constraint and the boundaries associated with k-nearest-neighbor retrieval and medoid selection reduce to linear or quadratic inequalities in z within each activation region.

The k-nearest-neighbor boundaries partition the conditional path into local regions over which $\mathcal{K}(\boldsymbol{Y}(z))$ is fixed. Within each such region, the medoid constraints determine whether the selected representative reference is equal to $j_{\mathrm{med}}^{\mathrm{obs}}$. Parametric programming explores these local regions along the one-dimensional path and retains those portions satisfying the observed attention-selection outcome and $\mathcal{M}(\boldsymbol{Y}(z)) = \mathcal{M}(\boldsymbol{y})$. The retained intervals are then combined over all local regions

and merged when adjacent. Importantly, a change in the k-nearest-neighbor set does not by itself exclude an interval from the truncation region if the final selected medoid remains unchanged. In contrast to over-conditioning, which uses only the local feasible interval containing the observed statistic, this procedure identifies the entire truncation region consistent with the observed selection outcome.

Given the decomposition in Eq. (10), the selective p-value can be evaluated explicitly as

$$p_{\text{selective}} = \frac{\displaystyle\sum_{r:\, u_r \geq \widetilde{T}(\boldsymbol{y})} \int_{\max\{\ell_r, \widetilde{T}(\boldsymbol{y})\}}^{u_r} f_{\chi_d}(z)\, dz}{\displaystyle\sum_{r=1}^{R} \int_{\ell_r}^{u_r} f_{\chi_d}(z)\, dz}, \tag{11}$$

where f_{χ_d} denotes the probability density function of the χ distribution with d degrees of freedom. Equation (11) is the upper-tail probability of the χ_d distribution truncated to the selection-dependent region $\mathcal{Z}$.

D Experimental details and additional results

This appendix provides additional experimental details omitted from the main paper due to space limitations, together with supplementary results under additional parameter settings. For each experimental setting, we first describe the implementation and inference settings and then present additional results that complement those in Section 5.

D.1 Common experimental details

Computational environment. All experiments, except for UNI2 feature extraction, were implemented in Python and executed on an x86_64 Linux computer equipped with two Intel Xeon Gold 6230 CPUs (20 cores each, 40 CPU cores in total). UNI2 feature extraction for CAMELYON16 was performed on a separate Linux computer equipped with an NVIDIA RTX A6000 GPU.

Attention parameterization for fixed-threshold selection. We introduce a sigmoid transformation of the instance-wise attention logits before softmax normalization to support fixed-threshold instance selection. Direct softmax normalization is invariant to a common additive shift of the logits, so their absolute offset is not determined by the normalized attention weights. This does not affect quantile-based selection, which depends on instance ranking, but motivates our modified parameterization when using a fixed cutoff. Applying sigmoid before softmax breaks this shift invariance and provides a bounded instance-wise score suitable for fixed-threshold selection. For selective inference, we express the selection event using the pre-sigmoid logit a_m. Since sigmoid is strictly increasing, $\mathrm{sigmoid}(a^m) > c$ is equivalent to $a^m > \log(c/(1-c))$ for $c \in (0,1)$. In particular, the fixed logit threshold $\tau = 0$ used in the synthetic experiments corresponds to a sigmoid-score threshold of 0.5. Using the pre-sigmoid logits also preserves the piecewise-affine structure required to characterize the selection event.

All experiments were conducted at the significance level $\alpha = 0.05$. We compared the following methods.

Proposed. The proposed method conditions on the complete observed selection event

$$\mathcal{E}(\boldsymbol{Y}) = (\mathcal{A}(\boldsymbol{Y}), \mathcal{M}(\boldsymbol{Y}))\,,$$

and uses parametric programming to characterize the entire truncation region consistent with this event.

Ablation 1. Ablation 1 ignores the attention-thresholding event $\mathcal{A}(\boldsymbol{Y})$ and conditions only on the adaptive reference-selection event $\mathcal{M}(\boldsymbol{Y})$. This ablation evaluates the effect of ignoring the data-dependent selection of the target instance.

Ablation 2. Ablation 2 conditions on the attention-thresholding event $\mathcal{A}(\boldsymbol{Y})$ but ignores the adaptive reference-selection event $\mathcal{M}(\boldsymbol{Y})$. This ablation evaluates the effect of ignoring adaptive reference selection.

Over-conditioning (OC). OC performs valid selective inference using only the local feasible interval containing the observed test statistic. Unlike the proposed method, OC does not characterize all disconnected regions that preserve the observed selection outcome. Consequently, OC introduces additional conditioning and is generally more conservative, which can reduce statistical power.

Bonferroni. Bonferroni accounts for the data-dependent selection procedure using a conventional multiple-testing correction. To account for attention-based selection, we define

$$\lambda_{\mathrm{att}} = \frac{N_{\mathrm{sample}}}{N_{\mathrm{att}}},$$

where N_{sample} is the number of sampled instances and N_{att} is the number of high-attention instances selected by attention thresholding. For each selected target instance, the adaptive reference-selection procedure selects one representative reference from M_{ref} reference candidates through k-nearest-neighbor retrieval followed by medoid selection. We therefore use the multiplicity factor

$$\lambda_{\mathrm{att}} M_{\mathrm{ref}}.$$

Using the naive p-value in Eq. (2), the Bonferroni-corrected p-value is defined as

$$p_{\mathrm{Bonferroni}} = \min\left\{1,\, \lambda_{\mathrm{att}} M_{\mathrm{ref}} p_{\mathrm{naive}}\right\}.$$

Here, λ_{att} accounts for attention-based target selection, while M_{ref} accounts for the possible choices of the final representative reference. We use this Bonferroni-based correction as a comparison baseline. For the synthetic and MNIST-based experiments, we evaluate its empirical Type-I error rate and statistical power, whereas for CAMELYON16 we report rejection proportions as described in Section 5.4. The correction can be highly conservative when the multiplicity factor is large.

Naive. The naive method ignores all selection events and computes the p-value from the untruncated χ_d distribution. This method is included in the synthetic experiments to illustrate the effect of selection bias and is not expected to provide valid Type-I error control in general.

D.2 Synthetic experiments

Experimental settings. For each combination of the input dimension d and Gaussian noise variance σ^2, we independently trained an ABMIL model using synthetically generated bags. Positive and negative instances were generated from $\mathcal{N}(\mathbf{1}, \sigma^2 I_d)$ and $\mathcal{N}(\mathbf{0}, \sigma^2 I_d)$, respectively. Each training epoch consisted of 1,000 positive bags and 1,000 negative bags, each containing 10 instances. The number of positive instances in a positive bag was sampled uniformly from one to ten, whereas negative bags consisted entirely of negative instances. New training bags were generated at every epoch.

The ABMIL model consisted of an MLP-based feature extractor, an attention network, and a bag-level classifier. The feature extractor had hidden dimensions $d/2$ and $d/4$, while both the attention network and the classifier used a hidden dimension of $d/8$. Thus, for $d \in \{16, 32, 48, 64\}$, the corresponding encoded feature dimensions were $d' \in \{4, 8, 12, 16\}$. ReLU activations were used in all hidden layers. The pre-sigmoid attention logits were transformed by a sigmoid function and subsequently normalized across the instances in each bag using softmax.

The model was trained for 10 epochs using binary cross-entropy loss and the Adam optimizer with an initial learning rate of 10^{-3}. The batch size was one, and the learning rate was halved every five epochs. After training, only the feature extractor and attention network were used for selective inference.

We considered $d \in \{16, 32, 48, 64\}$ and $\sigma^2 \in \{0.6, 0.8, 1.0, 1.2, 1.4\}$. For each hypothesis test, $M_{\mathrm{ref}} = 100$ negative reference instances were independently generated. Since the noise variance is known in this synthetic setting, the true variance used for data generation was also used in the statistical test. Negative and positive instances were used to evaluate the Type-I error rate and statistical power, respectively. Because the attention logit is computed independently for each instance,

hypothesis testing was performed at the instance level. Sampling was continued until 10,000 high-attention test instances were obtained for each configuration.

For the main experiment in Section 5.2, we fixed $d = 32$ and $\sigma^2 = 1.0$ and varied $k \in \{1, 3, 5, 10\}$, with the attention threshold set to $\tau = 0$. For the additional experiments reported below, we used $k = 5$ as a representative nontrivial medoid-selection setting and independently varied the input dimension and Gaussian noise variance.

For the stress test, we fixed $d = 32$ and $\sigma^2 = 0.6$ and set the attention threshold to the 99.99th percentile of pre-sigmoid attention logits computed from an independent set of negative instances. We set $M_{\mathrm{ref}} = k \in \{1, 3, 5, 10\}$, so that the k-nearest-neighbor retrieval became trivial while medoid selection remained adaptive for $k > 1$. For each setting, 1,000 negative instances satisfying the attention threshold were collected.

Additional results. We further evaluated the robustness of the proposed method to the input dimension and Gaussian noise variance. In both experiments, we fixed $k = 5$ as a representative setting in which medoid selection is nontrivial.

The top row of Figure 6 shows the results obtained by varying the input dimension while fixing $\sigma^2 = 1.0$. The proposed method consistently controlled the Type-I error rate near the nominal significance level across all dimensions. Ablation 1 and OC also remained close to the nominal level, whereas Ablation 2 and Naive were conservative and Bonferroni was highly conservative. The statistical power of both the proposed method and OC increased with the input dimension, while the proposed method consistently achieved higher power than OC. Moreover, the difference in power between the two methods became more pronounced as the input dimension increased.

The bottom row shows the results obtained by varying the Gaussian noise variance while fixing $d = 32$. The proposed method again maintained the Type-I error rate near the nominal level across all noise levels. Its statistical power decreased as the noise variance increased, as expected because the positive and negative distributions became more difficult to distinguish. Nevertheless, the proposed method consistently achieved higher statistical power than OC over the entire range of noise variances, whereas Bonferroni remained highly conservative. These results demonstrate that the proposed method retains valid Type-I error control and favorable statistical power across different input dimensions and noise levels.

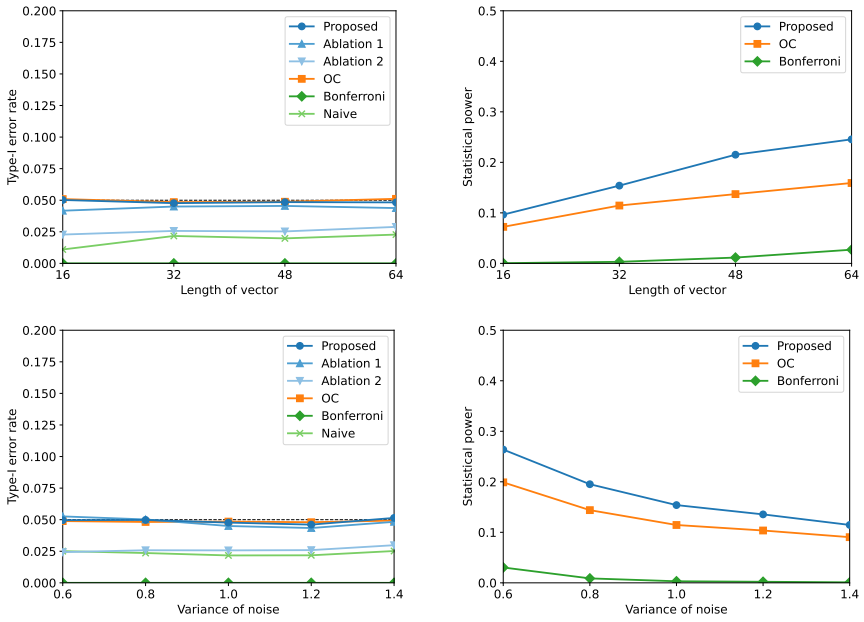


Figure 6: Type-I error rate (left column) and statistical power (right column) under additional synthetic-data settings. The dashed horizontal line indicates the nominal significance level of 0.05. The top row varies the input dimension with $\sigma^2 = 1.0$ and $k = 5$. The bottom row varies the Gaussian noise variance with $d = 32$ and $k = 5$.

D.3 MNIST-BASED EXPERIMENTS

Experimental settings. For each experiment, digit "0" was used as the negative class, and one of the digits in $\{1, \ldots, 9\}$ was independently treated as the positive class. The images were downsampled from 28×28 to 14×14 by average pooling, flattened into 196-dimensional vectors, and normalized to $[0, 1]$.

For model training, 10 positive and 10 negative images were randomly selected from the MNIST training set as cluster centers. Training instances were generated by randomly selecting one of the corresponding cluster centers and adding independent isotropic Gaussian noise. Thus, the training distribution consisted of 10 Gaussian clusters for each class.

For statistical inference, separate cluster centers were selected from the MNIST test set. We considered three- and five-cluster negative distributions, and the same negative cluster centers were used to generate both reference and null test instances. For positive test instances, the same number of positive cluster centers was independently selected from the corresponding positive digit class. The selected cluster centers were fixed throughout each experiment, whereas Gaussian noise was independently regenerated for every hypothesis test. The reference set contained $M_{\text{ref}} = 100$ instances, with cluster centers sampled uniformly when generating the reference instances.

Each training epoch consisted of 1,000 positive bags and 1,000 negative bags, with 10 instances per bag. The number of positive-class instances in each positive bag was sampled uniformly from one to ten, and the remaining instances were generated from the negative class. Negative bags contained only negative-class instances. New Gaussian noise was generated at every epoch.

The ABMIL model consisted of a linear feature extractor mapping the 196-dimensional input to a 32-dimensional feature vector, followed by an attention network and a bag-level classifier. Both the attention network and classifier used a 16-dimensional hidden layer with ReLU activation. The pre-sigmoid attention logits were transformed by sigmoid and normalized across the bag using softmax.

The model was trained for 10 epochs using binary cross-entropy loss and the Adam optimizer with an initial learning rate of 10^{-3}. The batch size was one, and the learning rate was halved every

five epochs. After training, only the feature extractor and attention network were used for selective inference.

The noise variance was fixed to $\sigma^2 = 0.25$ for the experiments reported in this paper. As in the synthetic experiments, the true noise variance was assumed to be known and was used in the statistical test. Negative and positive instances were used for the Type-I error rate and statistical power evaluations, respectively. For Type-I error evaluation, only high-attention test instances whose selected representative reference originated from the same negative cluster as the test instance were counted. Sampling continued until 10,000 such instances were obtained for each setting.

The attention threshold τ was set to the 95th percentile of pre-sigmoid attention logits computed from an independent set of negative instances. These instances were independent of the test and reference instances, and the resulting threshold was treated as fixed throughout selective inference.

The main experiment in Section 5.3 used three negative clusters and $k = 5$. The additional experiments below vary k while retaining the three-cluster setting and increase the number of negative clusters to five while retaining $k = 5$.

Additional results. We conducted two additional experiments to evaluate robustness to the reference-selection parameter and the heterogeneity of the negative distribution. First, while retaining the three-cluster setting used in the main experiment, we varied the number of nearest neighbors to $k = 3$ and $k = 1$. Second, we increased the number of negative clusters from three to five while retaining $k = 5$. All other experimental settings were unchanged.

Figure 7 shows the results for the three-cluster setting with $k = 3$ and $k = 1$. Across both values of k and all positive classes, the proposed method controlled the Type-I error rate near the nominal significance level. Ablation 1 and OC also remained close to the nominal level, whereas Ablation 2 tended to be conservative. The proposed method consistently achieved higher statistical power than OC. Bonferroni remained substantially conservative and markedly less powerful than the proposed method and OC.

Figure 8 shows the results obtained by increasing the number of negative clusters to five while retaining $k = 5$. The proposed method again controlled the Type-I error rate near the nominal level across all positive classes and consistently achieved higher statistical power than OC. The same qualitative behavior as in the three-cluster main experiment was therefore preserved under the more heterogeneous negative distribution. Together, these results demonstrate robustness to both the choice of k and increased distributional heterogeneity.

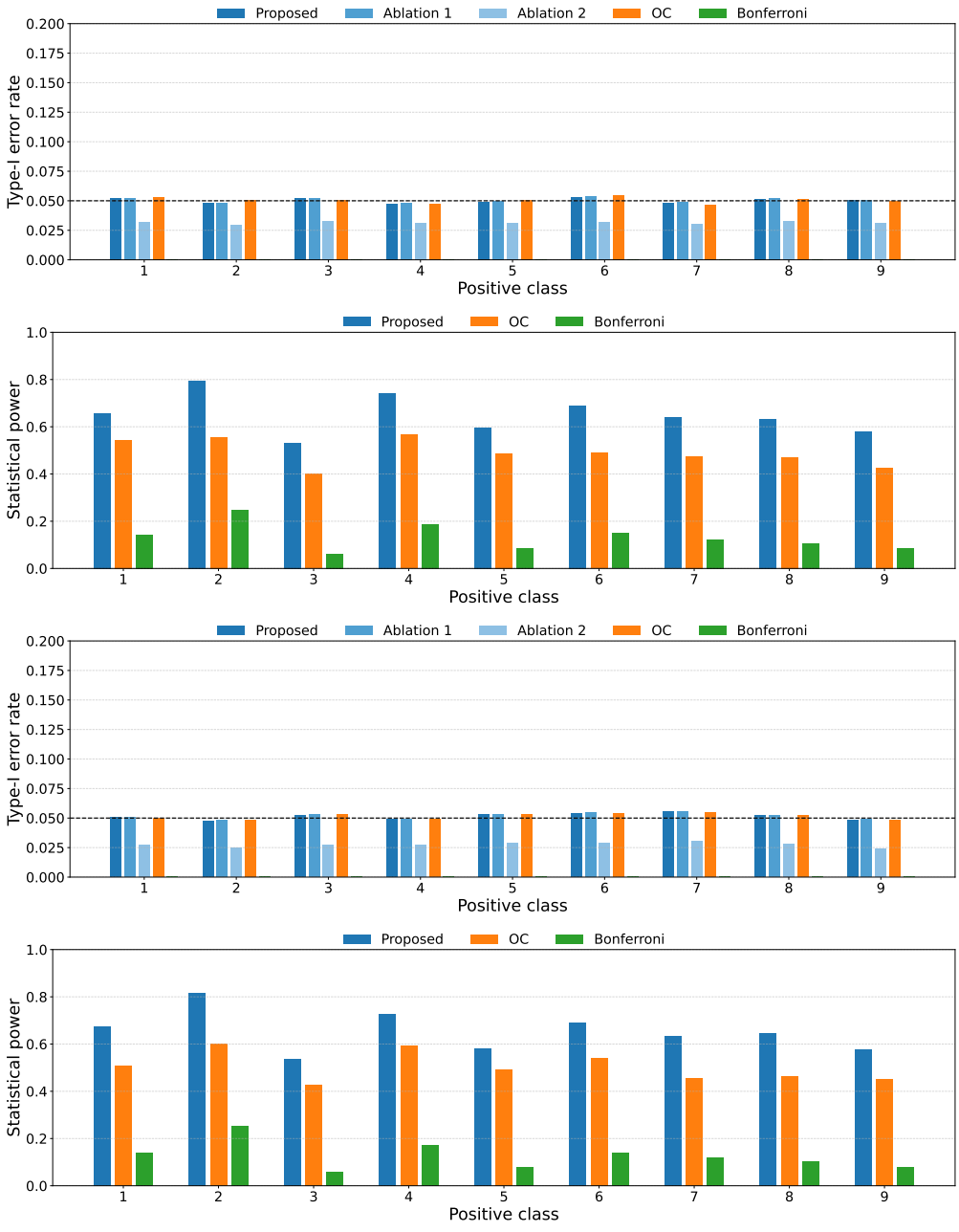


Figure 7: Type-I error rate and statistical power in the MNIST-based experiment with three Gaussian clusters. From top to bottom, the plots show the Type-I error rate and statistical power for $k = 3$, followed by those for $k = 1$. The horizontal axis denotes the positive digit class.

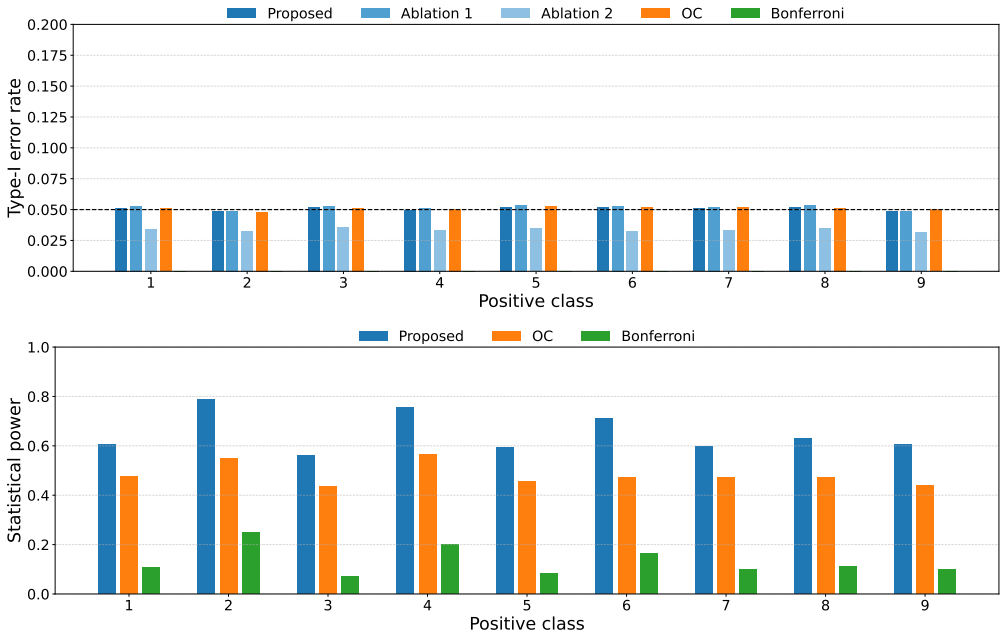


Figure 8: Type-I error rate and statistical power in the MNIST-based experiment with five Gaussian clusters. From top to bottom, the plots show the Type-I error rate and statistical power for $k = 5$. The horizontal axis denotes the positive digit class.

D.4 REAL-WORLD EXPERIMENTS ON CAMELYON16

Experimental settings. CAMELYON16 contains WSIs collected at two medical centers and includes both micro- and macro-metastatic tumor cases. We used WSIs from a single center to reduce center-specific variation arising from differences in staining and slide preparation, and restricted the tumor cases to macro-metastases to facilitate interpretation of the visualization results.

The training split from the selected center contained 59 normal and 26 macro-metastatic tumor WSIs. Among the 59 normal WSIs, 37 were used for ABMIL training, 11 for reference construction, and the remaining 11 exclusively for noise-variance estimation and attention-threshold determination. All 26 tumor WSIs were used for ABMIL training. These three subsets of normal WSIs were mutually disjoint. The test set consisted of 30 normal and 9 macro-metastatic tumor WSIs.

Each WSI was divided into non-overlapping 2048×2048-pixel patches over its tissue regions. Each patch was resized to the input resolution of UNI2, from which a 1,536-dimensional feature vector was extracted. To reduce the degrees of freedom of the χ distribution and improve numerical stability in the computation of selective p-values, average pooling was applied to reduce the UNI2 feature dimension from 1,536 to 192. The resulting 192-dimensional vectors were treated as the input instances for both hypothesis testing and reference selection (k-nearest-neighbor retrieval and medoid selection). The learned 32-dimensional ABMIL representation was therefore used for attention computation but not for reference selection.

The reference instances were constructed from the 11 normal WSIs reserved for reference construction. Up to 100 instances were randomly sampled from each WSI. Because several WSIs contained fewer than 100 eligible patches, the reference set comprised 1,009 instances in total. Each reference WSI was treated as a distinct population with its own mean, allowing heterogeneity among normal tissue samples to be represented.

The remaining 11 normal WSIs were used independently to estimate the noise variance and attention thresholds. Up to 100 instances were sampled from each WSI. Each WSI was modeled as an isotropic Gaussian distribution with its own mean vector, and the isotropic variance was estimated independently for each WSI. The final noise variance was obtained by averaging these variance estimates over the 11 WSIs. The empirical distribution of the pre-sigmoid attention logits from these independently sampled patches was used to determine the attention threshold. The main-paper experiment used the threshold corresponding to the top 5% of attention logits. Both the estimated noise variance and the attention threshold were treated as fixed during selective inference.

The ABMIL model consisted of a linear feature extractor mapping the 192-dimensional pooled UNI2 features to a 32-dimensional representation, followed by an attention network and a bag-level classifier with a 16-dimensional hidden layer. ReLU activations were used in the hidden layers, and the pre-sigmoid attention logits were transformed by sigmoid and normalized using softmax across instances within each bag.

For model training, each WSI was randomly divided into bags containing 50 image patches. Up to 50 bags were generated from each WSI depending on the number of available patches, and the bag assignment was regenerated at each training epoch. The model was trained for 10 epochs using binary cross-entropy loss and the Adam optimizer with an initial learning rate of 10^{-4}. The batch size was one, and the learning rate was halved every five epochs.

For slide-level evaluation, all patches from each WSI were treated as a single bag. The trained ABMIL model achieved a test accuracy of 0.897. After training, only the feature extractor and attention network were used for selective inference.

Because the number of instances differed among test WSIs, the attention-selection multiplicity factor used for Bonferroni correction was computed separately for each WSI. For the n-th WSI, we defined

$$\lambda_{\mathrm{att}}^{n} = \frac{M_{\mathrm{test}}^{n}}{|\mathcal{C}(X_{\mathrm{test}}^{n})|},$$

where M_{test}^{n} and $|\mathcal{C}(X_{\mathrm{test}}^{n})|$ denote the total number of instances and the number of attention-selected instances, respectively. The corresponding Bonferroni multiplicity factor was

$$\lambda_{\mathrm{att}}^{n} M_{\mathrm{ref}}.$$

The main-paper results use $k = 3$ and the top-5% attention threshold.

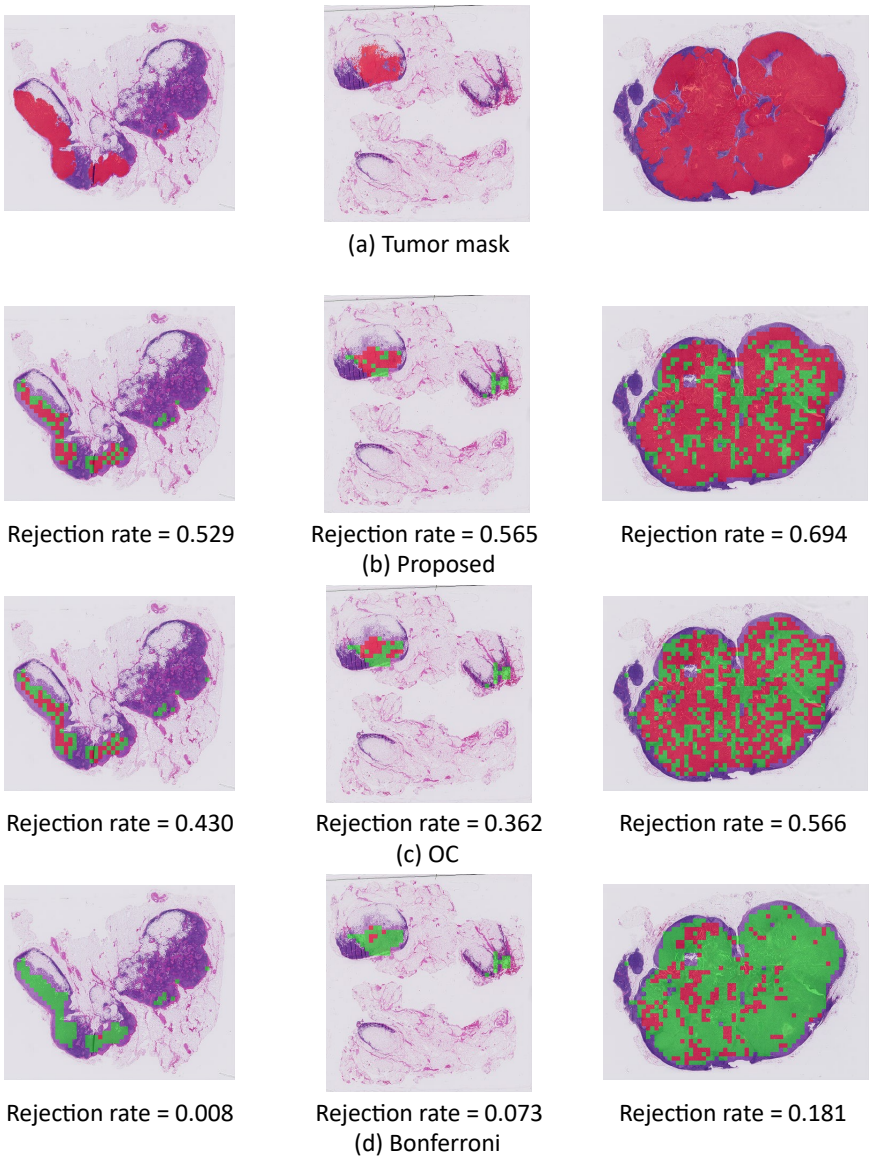


Figure 9: Visualization of p-value maps for CAMELYON16 tumor WSIs (Part I). (a) Ground-truth tumor masks; (b) Proposed; (c) OC; and (d) Bonferroni. Red and green rectangles indicate attention-selected instances with $p < 0.05$ and $p \geq 0.05$, respectively. Tumor masks are used only for visualization.

Additional qualitative results. Figures 9–11 present the visualization results for all nine tumor WSIs in the test set. Unlike the compact visualization in the main paper, the appendix includes Proposed, OC, and Bonferroni for every tumor WSI. Red and green rectangles indicate attention-selected instances with $p < 0.05$ and $p \geq 0.05$, respectively. The tumor annotations are shown only for visualization and were not used for statistical inference. Across the nine WSIs, the proposed method generally identifies more statistically significant regions than OC, whereas Bonferroni yields substantially fewer rejections.

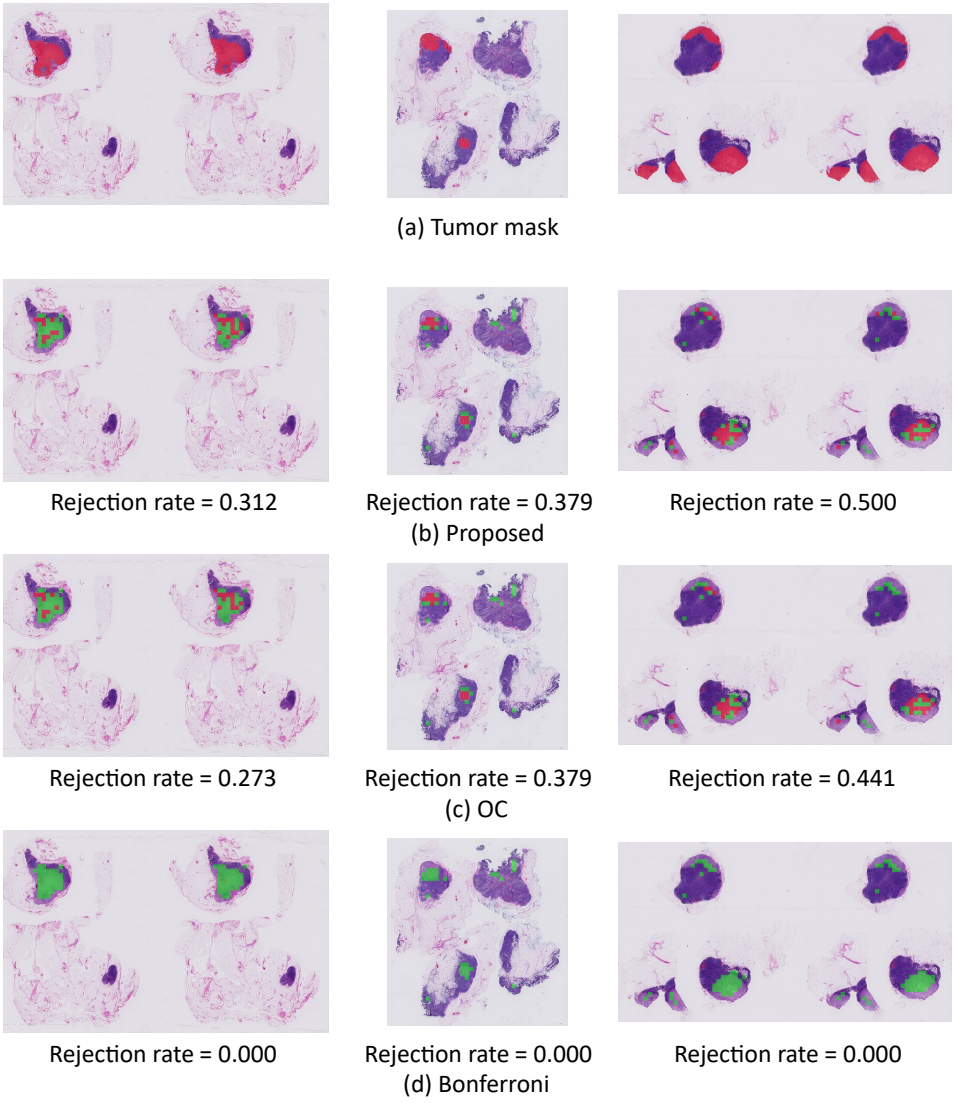


Figure 10: Visualization of p-value maps for CAMELYON16 tumor WSIs (Part II). (a) Ground-truth tumor masks; (b) Proposed; (c) OC; and (d) Bonferroni. Red and green rectangles indicate attention-selected instances with $p < 0.05$ and $p \geq 0.05$, respectively. Tumor masks are used only for visualization.

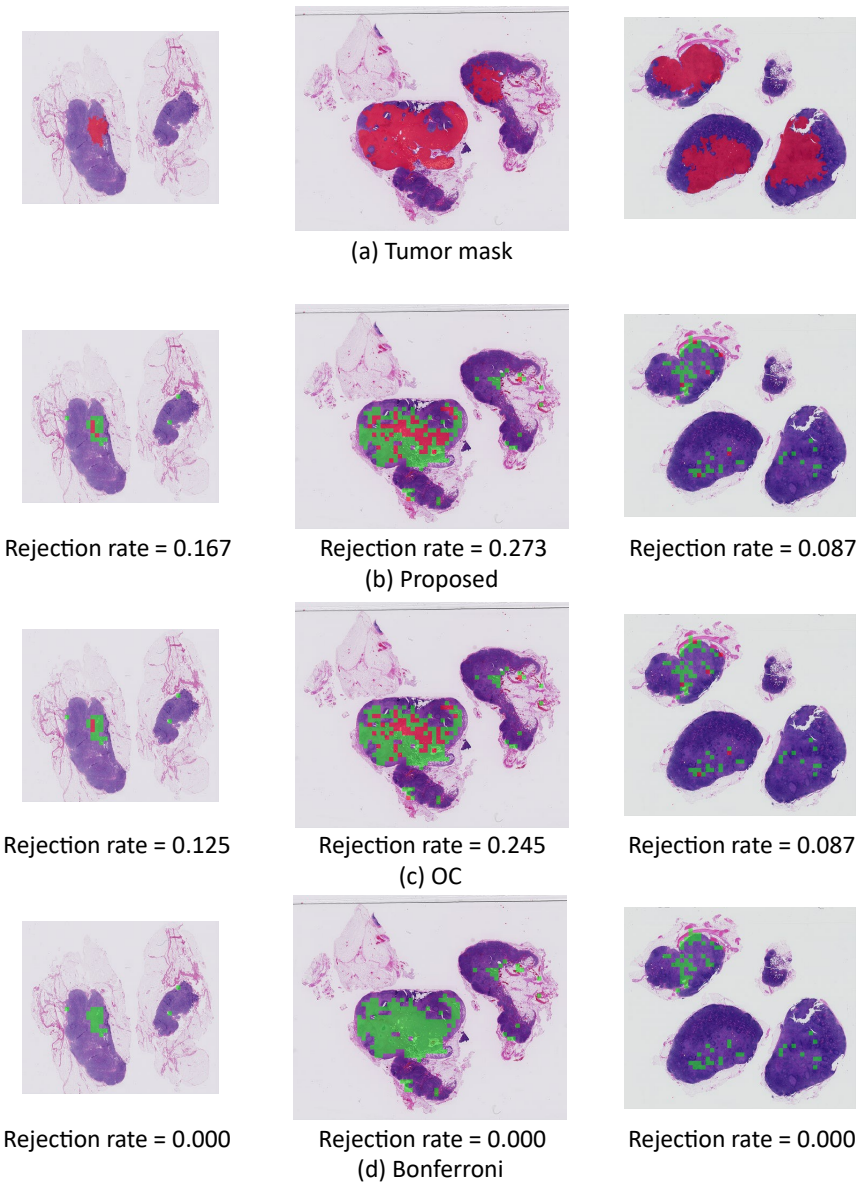


Figure 11: Visualization of p-value maps for CAMELYON16 tumor WSIs (Part III). (a) Ground-truth tumor masks; (b) Proposed; (c) OC; and (d) Bonferroni. Red and green rectangles indicate attention-selected instances with $p < 0.05$ and $p \geq 0.05$, respectively. Tumor masks are used only for visualization.