GEOMETRY-AWARE MULTIPLE INSTANCE LEARNING VIA A SHARED RIEMANNIAN METRIC FOR FEW-SHOT WHOLE-SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The diagnostic paradigm of whole slide images (WSIs) involves a frozen pathomorphological model, providing patch features, while multiple instance learning (MIL) accomplishes aggregation under weak slide-level labels. Nevertheless, this paradigm has two issues in practice: feature manifold collapse and unstable attention. By analyzing them, we realize that both belong to the two ends of the same causal chain, where the disruption of the feature manifold causes oscillatory changes in the attention distribution that depends on the tangent-space component. Based on this finding, we propose geometry-aware multiple instance learning (GeoMIL), which simultaneously drives feature transformation and attention generation with a single learnable Riemannian metric. We design a tangent space low-rank projection (TSLP) module to perform task adaptation while maintaining feature geometry, and use geometric query attention (GQA) to anchor the attention distribution to the tangent space stabilized by TSLP. Through extensive experiments, we demonstrate that our method maintains superior few-shot WSI classification performance while achieving a more stable attention distribution.

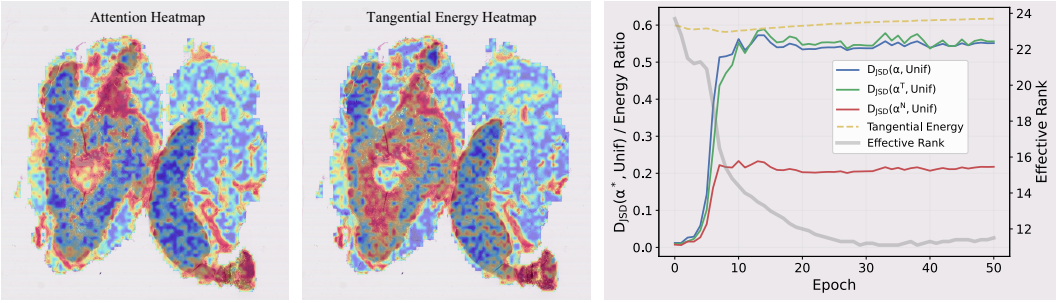
1 INTRODUCTION

Multiple instance learning (MIL) has become the standard paradigm for whole slide image (WSI) analysis: a slide is partitioned into patches that serve as instances, and an aggregator is trained under slide-level labels only (Ilse et al., 2018; Lu et al., 2021). In recent years, this paradigm has converged on a compact recipe, namely that a pathology foundation model supplies highly discriminative patch features (Lu et al., 2024; Chen et al., 2024; Ma et al., 2024; Xu et al., 2025; Wang et al., 2022), and attention-based MIL (Ilse et al., 2018) performs the aggregation and produces interpretable heatmaps. Because the pathology foundation model is kept frozen in most cases (Shao et al., 2025; Ye et al., 2026), the design of the aggregation module becomes the key to improve the performance.

Recently, two lines of work have separately challenged the prevailing aggregation framework (Ye et al., 2026; Xiong et al., 2026). The first targets feature geometry: spectral and tangent-space analyses show that, although the features of pathology foundation models are highly discriminative, they are embedded in a low-dimensional nonlinear manifold whose effective rank is only about 30 (Xiong et al., 2026; Fefferman et al., 2016; Xiang & Zhang, 2023); the linear projection layers commonly used in aggregators lack the inductive bias to preserve this manifold, distort the structure, and amplify overfitting in the few-shot regime (Xiong et al., 2026; Baraniuk & Wakin, 2009; Zhang & Zha, 2004). The second targets attention (Zhang et al., 2025; 2024): attention distributions exhibit unstable, hard-to-converge dynamics across epochs (Ye et al., 2026); existing work mostly treats this as a training problem, stabilizing it with statistical devices such as EMA (exponential moving average) anchors and normalized sigmoid, or suppressing over-concentration from the sampling and loss sides (Ye et al., 2026; Tang et al., 2023; Zhu et al., 2025).

To analyze these two problems in depth, we design a controlled diagnostic experiment (Figure 1; the experimental protocol is given in Appendix B). The experiment uses a standard attention framework (Ilse et al., 2018), and the tangent and normal spaces are estimated independently by data-driven local PCA (Zhang & Zha, 2004). Figure 1 shows that the attention distribution remains markedly concentrated at locations where no diagnostic evidence exists (Figure 1(a)) (Zhang et al.,

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.



(a) (b) (c)

Figure 1: Illustration that attention instability has a geometric origin. (a) Attention heatmap of a normal WSI; (b) Per-patch tangential energy heatmap of the same WSI; its high-value regions coincide closely with the focus regions in (a); (c) The attention concentrations $D_{\mathrm{JSD}}(\alpha, \mathrm{Unif})$, $D_{\mathrm{JSD}}(\alpha^T, \mathrm{Unif})$, and $D_{\mathrm{JSD}}(\alpha^N, \mathrm{Unif})$ computed from the original features, the tangential projection features, and the normal projection features, together with the tangential energy proportion and the effective rank, over epochs. Note that $D_{\mathrm{JSD}}(\alpha, \mathrm{Unif})$ denotes the Jensen–Shannon divergence from the attention distribution α to the uniform distribution. See Appendix B for the specific calculation method and detailed analysis of the values in the plot.

2025; 2024). The attention focus regions coincide closely with the high-value regions of the per-patch tangential energy (Figure 1(b)). Attention is then computed separately from the original features, the tangentially projected features, and the normally projected features, and the attention concentration is evaluated on each of the three resulting distributions: the D_{JSD} trajectories of the original features and the tangentially projected features agree closely, and the attention concentration rises as the effective rank falls (Figure 1(c)). Here $D_{\mathrm{JSD}}(\alpha^*, \mathrm{Unif})$ denotes the distance from the attention distribution α^* to the uniform distribution, measured by the Jensen–Shannon divergence.

It follows that feature manifold collapse and attention instability are not independent of each other, but they are the two ends of the same causal chain. A geometry-agnostic linear map corrupts the manifold structure of the features (Xiong et al., 2026; Baraniuk & Wakin, 2009); the tangent-space component of the feature manifold therefore varies violently during training, which in turn makes the attention distribution that depends on the tangent-space component oscillate. At the same time, the unstable tangent-space component keeps attention markedly concentrated in non-diagnostic regions. This phenomenon also shows that merely stabilizing the attention distribution statistically (Ye et al., 2026; Zhang et al., 2025) is not enough to solve the problem at its root: even after the distribution has converged, attention may still keep pointing along the corrupted tangent-space directions.

To address manifold collapse and attention instability uniformly, starting from their geometric root, we propose **GeoMIL** (geometry-aware multiple instance learning), a framework that uses a single learnable Riemannian metric as a shared geometric foundation to drive both feature transformation and attention generation, organized into two complementary modules. Tangent-space low-rank projection (TSLP) uses the low-rank factor to suppress the normal-space artifacts of each patch while performing dimensionality reduction, and additionally achieves task adaptation through a teacher–student structure, so that adaptation is not paid for by forgetting the data geometry; Geometry query attention (GQA) allows the attention query to be determined by the global position of the slide in the tangent space, so that the attention distribution can be anchored to the tangent space stabilized by TSLP.

Our contributions are summarized as follows.

1. We lay out the chained relationship among feature geometry degradation, variation of the manifold tangent space, and unstable attention distributions, and verify it experimentally.

2. We propose GeoMIL, which uses a single learnable Riemannian metric to achieve a geometry-preserving low-rank feature transformation and a geometrically grounded attention query, obtaining a more stable attention distribution without sacrificing classification performance.

3. We conduct extensive experiments, including comparative experiments, ablation studies, and localization analyses, to validate the effectiveness of our method.

2 Related Work

2.1 Attention-Based Multiple Instance Learning

MIL encodes a WSI as a set of instance features, and an aggregator produces a prediction under slide-level weak supervision (Lu et al., 2021; Zhang et al., 2022). Among these methods, the gated attention of ABMIL (Ilse et al., 2018) has become the de facto standard because its weights can be visualized directly as heatmaps. Subsequent work has broadly followed two directions. The first models inter-instance dependencies, e.g., self-attention (Shao et al., 2021), key-instance guidance (Li et al., 2021), graph structures (Fourkioti et al., 2024; Chen et al., 2021; Guo et al., 2023; Li et al., 2024), and hierarchical selection (Xiong et al., 2023). The second mitigates overfitting and spurious correlations under weak supervision, e.g., clustering constraints (Lu et al., 2021), pseudo-bag distillation (Zhang et al., 2022), causal debiasing (Lin et al., 2023), and low-rank compression (Tang et al., 2024; Song et al., 2024). The attention distribution itself has also been studied extensively: ASMIL (Ye et al., 2026) diagnoses its epoch-to-epoch oscillation and stabilizes it with an EMA anchor, while Zhang et al. (2025; 2024) suppress over-concentration from the sampling and loss sides. These works, however, only optimize the statistical shape of the attention distribution; they do not investigate the intrinsic cause and geometric nature of its abnormality.

2.2 Manifold Preservation and Few-Shot WSI Classification

The core assumption of manifold learning states that high-dimensional observations typically lie near a low-dimensional smooth manifold embedded in the ambient space (Pope et al., 2021; Facco et al., 2017). Through spectral analysis and local tangent-space analysis, MR-Block (Xiong et al., 2026) further confirms that the features output by pathology foundation models lie on a low-dimensional nonlinear manifold whose effective rank is only about 30, far below the original dimension of the feature space, and that the linear layers in aggregators, which lack geometric inductive bias, distort this structure (Brahma et al., 2016). In addition, through a large-scale comparative study, MAMMOTH (Shao et al., 2026) shows that the performance gain brought by replacing the linear transformation of patch features with a low-rank mixture of experts exceeds the gain from changing the aggregator architecture itself. This shows clearly that shaping the geometry of features before they enter the aggregation module matters as much as the aggregation method itself.

Few-shot weakly supervised whole-slide image classification (FSWC) is formally introduced as a research setting by TOP (Qu et al., 2023), whose core assumption is that only a few annotated slides are available per class and that the supervision signal is confined to the slide level (Song et al., 2023; Snell et al., 2017). Existing methods mainly follow two technical routes. The first injects external semantic priors to supplement the supervision signal, e.g., PEMP (Qu et al., 2024) introduces pathology knowledge prompts, ViLa-MIL (Shi et al., 2024) adopts prototype-guided aggregation, and FOCUS (Guo et al., 2025) compresses visual features under the guidance of language priors (Lu et al., 2024; Huang et al., 2023). The second optimizes the feature transformation operator before aggregation, or reuses a pretrained aggregator for cross-task transfer, improving few-shot adaptation at the architectural level (Shao et al., 2026; Xiong et al., 2024; Fu et al., 2024).

2.3 Riemannian Metric Learning and Representation Geometry

General metric learning is originally concerned with learning task-adaptive distances for classification and clustering (Xing et al., 2002; Bellet et al., 2015; Weinberger & Saul, 2009). Closer to this work is a family of methods that estimate geometry directly from data, e.g., graph-spectral and diffusion methods that characterize the manifold through local neighborhoods (Coifman & Lafon, 2006); deep generative models that define the curvature of the latent space through a pullback metric (Arvanitidis et al., 2018; 2016; 2021); and diffusion-model analyses showing that, under small noise, the score points along the normal direction of the manifold and that $I+\sigma^2\nabla^2 \log p_\sigma$ converges to the tangent-space projection operator (Stanczuk et al., 2024; Kharitenko et al., 2026; Diepeveen et al., 2024). Riemannian metric matching (Bamberger et al., 2026) belongs to this line as well: it

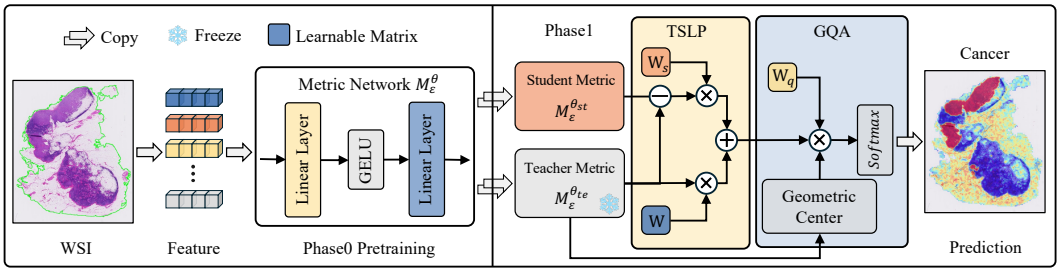


Figure 2: The overall framework of GeoMIL. A WSI is partitioned into N patches, and a pretrained feature extractor produces the instance features. The features are first fed into the shared metric network M_ε^θ, which completes the Phase 0 pretraining of the metric network. The teacher network is copied and frozen after Phase 0 to provide a stable geometric anchor, while the student residual branch remains trainable to handle task adaptation. Tangent-Space Low-rank Projection (TSLP) uses the low-rank factor M_ε^θ to suppress the normal-space artifacts of each patch while performing dimensionality reduction, and additionally achieves task adaptation through a teacher–student structure. Geometry-Query Attention (GQA) lets the attention query be determined by the global position of the slide in the tangent space, so that the attention distribution can be anchored to the tangent space stabilized by TSLP.

recasts the carré du champ operator in the form of a conditional expectation, so that a pointwise Riemannian metric can be regressed directly by a denoising objective, and it proves convergence to the tangent-space projection operator as noise variance $\varepsilon \to 0$.

3 Method

3.1 Problem Formulation

In the standard multiple instance learning framework, a WSI is partitioned into N patches, and a pretrained feature extractor produces instance features $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ ($d = 512$), with supervision consisting only of the slide-level label y. Standard attention-based MIL computes an attention weight for each instance and normalizes it into a distribution,

$$\alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{x}_i/\tau)}{\sum_{j=1}^N \exp(\mathbf{w}^\top \mathbf{x}_j/\tau)}, \qquad \mathbf{z}_{\text{slide}} = \sum_{i=1}^N \alpha_i \mathbf{x}_i, \tag{1}$$

where τ is a temperature, and $\mathbf{z}_{\text{slide}}$ is the aggregated slide representation that is passed through a classifier to give the slide-level prediction. The query $\mathbf{w}$ is learnable and **input-independent**, while in ABMIL (Ilse et al., 2018), $\mathbf{w}$ may also be generated by a gated MLP (multilayer perceptron), but it likewise does not depend on the slide-level context.

3.2 Riemannian Metric Network

The Riemannian metric network $M_\varepsilon^\theta : \mathbb{R}^d \to \mathbb{R}^{r\times d}$ follows the low-rank factor parameterization (Bamberger et al., 2026): the network takes the bandwidth ε as an additional conditioning input (FiLM modulation (Perez et al., 2018)), and outputs an $r \times d$ low-rank factor matrix. The pointwise Riemannian metric is evaluated on a noised sample $\mathbf{y} = \mathbf{x} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \varepsilon\mathbf{I})$, defined as

$$\Gamma_\varepsilon^\theta(\mathbf{y}) = M_\varepsilon^\theta(\mathbf{y})^\top M_\varepsilon^\theta(\mathbf{y}) + \lambda \mathbf{I}_d, \qquad \lambda = 10^{-3}, \tag{2}$$

where the Tikhonov term $\lambda\mathbf{I}_d$ guarantees that $\Gamma_\varepsilon^\theta$ is strictly positive definite. More importantly, since $\mathbf{y} \to \mathbf{x}$ as $\varepsilon \to 0$, the row vectors of M_ε^θ span the column space of $\Gamma_\varepsilon^\theta$; that is, the tangent space of the data manifold at $\mathbf{x}$; hence $M_\varepsilon^\theta(\mathbf{x})\mathbf{x} \in \mathbb{R}^r$ projects $\mathbf{x}$ into tangent-space coefficients. The metric is pretrained by conditional carré du champ matching (CDC) on unlabeled patch features,

$$\mathcal{L}_{\text{metric}} = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{N}(\mathbf{x},\varepsilon\mathbf{I})}\left[\left\|M_\varepsilon^\theta(\mathbf{y})M_\varepsilon^\theta(\mathbf{y})^\top\right\|_F^2 + 2\lambda\left\|M_\varepsilon^\theta(\mathbf{y})\right\|_F^2 - \frac{1}{\varepsilon}\left\|M_\varepsilon^\theta(\mathbf{y})(\mathbf{x}-\mathbf{y})\right\|^2\right]. \tag{3}$$

The three terms inside the metric loss $\mathcal{L}_{\text{metric}}$ are, in order, a low-rank matching term, a Tikhonov regularizer corresponding to $\lambda \mathbf{I}_d$ in Equation (2), and a denoising objective term, whose form follows from the work of Bamberger et al. (2026).

3.3 Tangent-Space Low-Rank Projection

As shown in Figure 2, the metric network must be pretrained with unlabeled patches in Phase 0 and then adapted to the task in Phase 1, which puts its two roles in conflict: task adaptation may destroy the general geometry learned in Phase 0. We therefore decouple the two with a frozen teacher network and a trainable student network. The teacher $M_\varepsilon^{\theta_{\text{te}}}$ is copied after Phase 0 and frozen to lock in the general geometry, while the student $M_\varepsilon^{\theta_{\text{st}}}$ is copied from the same weights and kept trainable to carry out task adaptation.

TSLP aims to suppress normal-space artifacts while preserving the tangent-space semantics of the features, and to compress the d-dimensional features into the task space. Reweighted according to the learned metric, the most direct formulation is left multiplication by the matrix square root of the metric, $\mathbf{x} \mapsto (\Gamma_\varepsilon^\theta)^{1/2}\mathbf{x}$; however, $\Gamma_\varepsilon^\theta$ is of dimension $d \times d$, and its square root requires an eigendecomposition or a Newton–Schulz iteration, which is both expensive to compute per patch and prone to numerical error. The low-rank factorization of Equation (2) provides an equivalent yet efficient alternative: since $\Gamma_\varepsilon^\theta = M_\varepsilon^{\theta\top} M_\varepsilon^\theta + \lambda \mathbf{I}$ already carries a rank-r factor, the scaling can be split into two steps. First, M_ε^θ is used to project the patch into tangent-space coefficients (preserving semantics, discarding normal noise, and completing the $d \to r$ reduction); then a learnable projection $\mathbf{W}^\top$ maps them into the d'-dimensional task space. Combining the two gives

$$\mathbf{x}' = \mathbf{W}^\top M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\,\mathbf{x} + \alpha\,\mathbf{W}_s^\top \big(M_\varepsilon^{\theta_{\text{st}}}(\mathbf{x}) - M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\big)\mathbf{x} \in \mathbb{R}^{d'}, \tag{4}$$

where $\mathbf{W}, \mathbf{W}_s \in \mathbb{R}^{r \times d'}$ are two learnable projections and α is a learnable scalar. Let $\mathbf{c} = M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\mathbf{x} \in \mathbb{R}^r$ be the teacher tangent-space coefficient and $\mathbf{d} = \big(M_\varepsilon^{\theta_{\text{st}}}(\mathbf{x}) - M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\big)\mathbf{x} \in \mathbb{R}^r$ is the student residual coefficient; together they determine the manifold direction.

To prevent the student from forgetting the teacher geometry, we impose a projection consistency constraint at the factor level,

$$\mathcal{L}_{\text{tangent}} = \frac{1}{N}\sum_{i=1}^{N} \big\| M_\varepsilon^{\theta_{\text{st}}}(\mathbf{x}_i)\mathbf{x}_i - M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x}_i)\mathbf{x}_i \big\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{d}_i\|^2, \tag{5}$$

which penalizes the extent to which the student projection deviates from the teacher projection; the deviation required for task adaptation is carried by the residual.

Whether the teacher branch of Equation (4) preserves semantics and suppresses artifacts hinges on one premise: that M_ε^θ really projects patches into the tangent space while leaving normal-space variation such as staining and scanning outside. Theorem 1 below shows that preserving tangent-space semantics and suppressing normal-space artifacts is a provable limiting property rather than a design preference. We take the intrinsic dimension $d_{\mathcal{M}} \approx 30$ and $r = 64$; the sensitivity of the model to the Phase 0 data amount and to the metric dimension r is analysed in Appendix D.

Theorem 1 (Tangent-space preservation)**.** *Assume that the data locally lie on a $d_{\mathcal{M}}$-dimensional manifold and that $r \geq 2d_{\mathcal{M}} - 1$. Then, in the limit $\varepsilon \to 0$, for any $\mathbf{x} \in \mathcal{M}$, $\mathbf{v}, \mathbf{w} \in T_{\mathbf{x}}\mathcal{M}$ and $\mathbf{n} \perp T_{\mathbf{x}}\mathcal{M}$: (i) (tangential isometry) $\|M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\mathbf{v}\| \to \|\mathbf{v}\|$; (ii) (normal annihilation) $\|M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\mathbf{n}\| \to 0$; (iii) (inner-product preservation) $\langle M_\varepsilon^{\theta_{\text{te}}}\mathbf{v}, M_\varepsilon^{\theta_{\text{te}}}\mathbf{w}\rangle \to \langle \mathbf{v}, \mathbf{w}\rangle$; (iv) (tangent–normal separability) writing $\mathbf{x} = \mathbf{x}_T + \mathbf{x}_N$ for the tangent–normal decomposition with respect to $T_{\mathbf{x}}\mathcal{M}$, we have $M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\mathbf{x} \to M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\mathbf{x}_T$ and $\|M_\varepsilon^{\theta_{\text{te}}}(\mathbf{x})\mathbf{x}_N\| \to 0$.*

The proof is given in Appendix A.1. Parts (i) and (iii) of Theorem 1 show that $M_\varepsilon^{\theta_{\text{te}}}$ is asymptotically distance-preserving on the tangent space, while (ii) and (iv) show that it asymptotically annihilates the normal space. Taken together, they mean that $\mathbf{c} = M_\varepsilon^{\theta_{\text{te}}}\mathbf{x}$ is an isometric encoding of the tangential component, and that normal-space variation such as staining and scanning has already been removed before entering the task space.

3.4 Geometry-Query Attention

The residual branch in Equation (4) serves task adaptation and therefore inevitably injects normal-space deviation beyond the teacher projection. To keep attention free of this deviation, we design

the GQA module so that the attention distribution is anchored to a stable tangent space. Concretely, we let the attention query be determined by the geometric position of the slide on the manifold. We first use $M_{\varepsilon}^{\theta_{\mathrm{te}}}$ to aggregate the geometric information of the whole WSI, obtaining the tangent-space geometric centroid,

$$\mathbf{g} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{c}_i = \frac{1}{N}\sum_{i=1}^{N} M_{\varepsilon}^{\theta_{\mathrm{te}}}(\mathbf{x}_i)\mathbf{x}_i \in \mathbb{R}^r. \tag{6}$$

We then condition the query on this centroid and let the query drive the aggregation,

$$\mathbf{q} = \mathbf{W}_q\,\mathbf{g} \in \mathbb{R}^{d'}, \qquad \alpha_i = \frac{\exp(\mathbf{q}^\top \mathbf{x}_i'/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{q}^\top \mathbf{x}_j'/\tau)}, \qquad \mathbf{z}_{\text{slide}} = \sum_{i=1}^{N}\alpha_i\,\mathbf{x}_i', \tag{7}$$

where τ is a temperature parameter and $\mathbf{W}_q \in \mathbb{R}^{d'\times r}$ is a trainable matrix. Since $\mathbf{c}_i$ has already been computed in the TSLP forward pass, this step is computationally efficient.

3.5 Loss Functions and Two-Stage Training

The training of GeoMIL is driven by three types of losses, split across two training stages:

$$\mathcal{L}_{\text{phase0}} = \mathcal{L}_{\text{metric}}, \qquad \mathcal{L}_{\text{phase1}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{tangent}}\,\mathcal{L}_{\text{tangent}}, \tag{8}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss. Phase 0 is metric pretraining: at each step a batch of unlabeled patches and a bandwidth ε are sampled, ε is used as a conditioning input, and $\mathcal{L}_{\text{metric}}$ is minimized on the noised samples. After Phase 0, $M_{\varepsilon}^{\theta_{\mathrm{te}}}$ is produced by copying and is frozen, and a trainable $M_{\varepsilon}^{\theta_{\mathrm{st}}}$ is then copied from it; the two are used jointly to construct the GeoMIL model in Phase 1. Further, λ_{tangent} is a balancing coefficient that is cosine-annealed from 0.1 to 0 over epochs; its role is to strengthen the geometric anchoring and stabilize the student network early on, and to release the residual for task adaptation later.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We conduct extensive experiments on CAMELYON-16 (Ehteshami Bejnordi et al., 2017), TCGA-NSCLC, and TCGA-BRCA. Details of the individual datasets are given in Appendix C.

Evaluation metrics. The primary classification metric is AUC, supplemented by Accuracy (ACC) and Macro-F1. In the few-shot experiments, $K \in \{4, 8, 16\}$ annotated slides are randomly sampled per class; each setting is run with five random seeds, and the mean and standard deviation are reported.

Implementation details. All models use the AdamW optimizer with a weight decay of 1×10^{-5} (Loshchilov & Hutter, 2019). Phase 0 is trained for 100 epochs with a learning rate of 1×10^{-4}. Phase 1 is trained for 100 epochs with an initial learning rate of 1×10^{-4} and a cosine-annealing schedule. The early-stopping patience is 10 epochs and the minimum training length is 20 epochs. The batch size is 1, the tangent-space dimension is $r = 64$, and the temperature is $\tau = 0.1$. All experiments are conducted on NVIDIA A100 GPUs.

4.2 Few-Shot Classification Results

Table 1 reports the AUC, ACC, and F1 of GeoMIL and ten baseline methods on the three datasets under the three few-shot settings (4/8/16-shot), where each cell is the mean$\pm$standard deviation over five random seeds. All methods use features extracted by CONCH at $20\times$ magnification, so that the comparison is fair.

As shown in Table 1, GeoMIL achieves a consistent lead on CAMELYON-16, TCGA-NSCLC, and TCGA-BRCA. The lead is largest on the sparsely positive CAMELYON-16. In a sparsely positive setting, normal regions account for the vast majority of a slide, so a geometry-agnostic linear projection is most likely to divert attention away from diagnostic evidence; this is where geometric protection is worth the most. If the gains came from a general improvement in classification ability,

Table 1: Few-shot classification results. The best performance is shown in **boldface**, the second-best is underlined, and the mean and standard deviation over five random seeds are reported. MR-ABMIL augments ABMIL with MR-Block (Xiong et al., 2026).

	Method	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
4-shot	ABMIL	57.0$_{\pm 8.6}$	58.9$_{\pm 9.0}$	48.0$_{\pm 9.7}$	88.8$_{\pm 3.9}$	81.5$_{\pm 3.3}$	81.7$_{\pm 3.3}$	85.2$_{\pm 8.3}$	72.4$_{\pm 9.3}$	65.9$_{\pm 10.6}$
	CLAM	56.4$_{\pm 8.1}$	57.5$_{\pm 8.1}$	46.9$_{\pm 9.1}$	90.4$_{\pm 2.9}$	82.6$_{\pm 2.6}$	82.9$_{\pm 2.6}$	86.2$_{\pm 8.1}$	73.8$_{\pm 10.4}$	68.2$_{\pm 9.4}$
	DSMIL	54.5$_{\pm 7.0}$	58.8$_{\pm 8.6}$	44.6$_{\pm 8.6}$	88.5$_{\pm 4.2}$	80.6$_{\pm 3.7}$	81.1$_{\pm 3.2}$	83.8$_{\pm 10.3}$	73.7$_{\pm 12.1}$	65.8$_{\pm 12.2}$
	DTFD-MIL	57.4$_{\pm 8.8}$	57.4$_{\pm 9.5}$	52.1$_{\pm 12.8}$	87.4$_{\pm 7.2}$	79.9$_{\pm 5.4}$	80.0$_{\pm 5.4}$	84.8$_{\pm 9.5}$	74.4$_{\pm 7.7}$	66.0$_{\pm 11.8}$
	TransMIL	53.3$_{\pm 5.4}$	56.1$_{\pm 5.0}$	41.2$_{\pm 8.2}$	88.2$_{\pm 2.9}$	79.7$_{\pm 2.5}$	79.3$_{\pm 3.1}$	82.5$_{\pm 9.0}$	69.2$_{\pm 9.4}$	62.2$_{\pm 11.2}$
	CAMIL	54.9$_{\pm 8.8}$	57.8$_{\pm 8.7}$	45.8$_{\pm 9.2}$	88.7$_{\pm 3.3}$	81.3$_{\pm 3.4}$	81.4$_{\pm 3.7}$	85.0$_{\pm 8.1}$	71.0$_{\pm 11.2}$	64.7$_{\pm 12.4}$
	ASMIL	54.2$_{\pm 8.4}$	56.3$_{\pm 9.2}$	42.6$_{\pm 11.4}$	85.4$_{\pm 4.5}$	74.2$_{\pm 5.5}$	75.8$_{\pm 3.7}$	79.3$_{\pm 9.3}$	67.5$_{\pm 9.6}$	59.7$_{\pm 8.7}$
	MR-ABMIL	56.0$_{\pm 7.0}$	57.2$_{\pm 7.2}$	47.8$_{\pm 7.9}$	87.2$_{\pm 4.0}$	78.5$_{\pm 7.6}$	79.0$_{\pm 4.6}$	85.8$_{\pm 7.4}$	72.8$_{\pm 9.6}$	65.5$_{\pm 8.3}$
	ViLa-MIL	56.6$_{\pm 8.3}$	55.8$_{\pm 8.6}$	51.8$_{\pm 9.3}$	88.5$_{\pm 4.1}$	80.7$_{\pm 4.3}$	80.7$_{\pm 4.4}$	83.0$_{\pm 9.2}$	74.5$_{\pm 8.5}$	66.4$_{\pm 11.0}$
	FOCUS	58.0$_{\pm 7.1}$	58.1$_{\pm 8.2}$	53.1$_{\pm 8.6}$	90.5$_{\pm 6.7}$	81.4$_{\pm 4.4}$	81.9$_{\pm 6.7}$	86.6$_{\pm 6.8}$	74.8$_{\pm 9.4}$	67.4$_{\pm 8.4}$
	GeoMIL	**62.8**$_{\pm 8.1}$	**64.3**$_{\pm 8.2}$	**56.0**$_{\pm 9.2}$	**91.9**$_{\pm 4.3}$	**83.7**$_{\pm 6.0}$	**84.7**$_{\pm 5.1}$	**88.9**$_{\pm 7.5}$	**75.3**$_{\pm 8.5}$	**70.0**$_{\pm 8.1}$
8-shot	ABMIL	72.3$_{\pm 8.3}$	72.4$_{\pm 8.3}$	62.7$_{\pm 8.5}$	91.0$_{\pm 2.0}$	82.9$_{\pm 1.0}$	82.1$_{\pm 1.6}$	89.5$_{\pm 8.8}$	79.5$_{\pm 9.4}$	67.1$_{\pm 9.3}$
	CLAM	74.6$_{\pm 8.0}$	72.6$_{\pm 8.1}$	64.5$_{\pm 8.7}$	92.3$_{\pm 1.3}$	84.4$_{\pm 1.0}$	83.9$_{\pm 1.2}$	90.7$_{\pm 7.1}$	81.6$_{\pm 8.1}$	70.0$_{\pm 8.1}$
	DSMIL	66.7$_{\pm 6.5}$	63.3$_{\pm 7.2}$	49.0$_{\pm 7.8}$	89.6$_{\pm 5.5}$	81.0$_{\pm 4.2}$	80.6$_{\pm 4.5}$	89.0$_{\pm 7.4}$	77.3$_{\pm 9.5}$	67.7$_{\pm 9.2}$
	DTFD-MIL	73.3$_{\pm 8.7}$	69.0$_{\pm 8.9}$	60.4$_{\pm 10.7}$	91.5$_{\pm 1.8}$	83.1$_{\pm 1.3}$	82.5$_{\pm 1.9}$	89.6$_{\pm 7.5}$	80.3$_{\pm 9.2}$	68.2$_{\pm 9.3}$
	TransMIL	66.0$_{\pm 5.6}$	65.8$_{\pm 5.3}$	57.5$_{\pm 6.5}$	88.8$_{\pm 4.3}$	80.4$_{\pm 3.0}$	79.9$_{\pm 3.0}$	89.4$_{\pm 7.2}$	78.6$_{\pm 8.6}$	65.8$_{\pm 9.0}$
	CAMIL	69.2$_{\pm 7.9}$	65.9$_{\pm 8.5}$	59.0$_{\pm 8.7}$	87.8$_{\pm 5.9}$	79.5$_{\pm 5.5}$	79.0$_{\pm 5.3}$	89.2$_{\pm 6.9}$	79.0$_{\pm 8.8}$	66.8$_{\pm 9.2}$
	ASMIL	65.0$_{\pm 8.0}$	60.2$_{\pm 8.4}$	56.4$_{\pm 9.6}$	85.5$_{\pm 7.8}$	76.0$_{\pm 7.4}$	75.3$_{\pm 8.0}$	86.6$_{\pm 5.3}$	70.7$_{\pm 6.9}$	60.1$_{\pm 7.0}$
	MR-ABMIL	74.0$_{\pm 7.8}$	72.9$_{\pm 8.5}$	60.7$_{\pm 9.1}$	89.9$_{\pm 4.7}$	81.7$_{\pm 4.8}$	81.3$_{\pm 5.0}$	89.9$_{\pm 6.4}$	79.8$_{\pm 6.9}$	68.7$_{\pm 7.1}$
	ViLa-MIL	71.5$_{\pm 7.4}$	70.0$_{\pm 8.1}$	65.5$_{\pm 9.2}$	89.5$_{\pm 5.4}$	82.2$_{\pm 6.0}$	82.0$_{\pm 6.2}$	88.2$_{\pm 8.8}$	78.7$_{\pm 5.5}$	68.4$_{\pm 6.1}$
	FOCUS	75.6$_{\pm 7.5}$	74.3$_{\pm 7.4}$	66.7$_{\pm 8.8}$	91.8$_{\pm 3.7}$	84.5$_{\pm 5.8}$	83.3$_{\pm 7.7}$	90.2$_{\pm 6.1}$	80.5$_{\pm 7.8}$	69.4$_{\pm 7.1}$
	GeoMIL	**79.8**$_{\pm 7.6}$	**77.4**$_{\pm 7.0}$	**72.8**$_{\pm 8.6}$	**93.5**$_{\pm 0.9}$	**86.2**$_{\pm 1.4}$	**85.8**$_{\pm 1.6}$	**93.6**$_{\pm 5.1}$	**83.1**$_{\pm 7.4}$	**74.5**$_{\pm 7.1}$
16-shot	ABMIL	85.8$_{\pm 7.8}$	84.2$_{\pm 7.1}$	75.1$_{\pm 8.1}$	93.7$_{\pm 1.1}$	87.5$_{\pm 1.4}$	87.1$_{\pm 1.6}$	92.3$_{\pm 4.7}$	83.4$_{\pm 5.6}$	72.4$_{\pm 6.7}$
	CLAM	87.9$_{\pm 7.5}$	85.7$_{\pm 7.7}$	79.2$_{\pm 8.2}$	**96.3**$_{\pm 2.8}$	88.0$_{\pm 1.6}$	87.7$_{\pm 1.6}$	93.2$_{\pm 4.2}$	85.3$_{\pm 4.8}$	73.7$_{\pm 7.6}$
	DSMIL	80.4$_{\pm 6.4}$	74.6$_{\pm 6.1}$	70.8$_{\pm 7.7}$	93.8$_{\pm 1.2}$	87.8$_{\pm 2.6}$	87.4$_{\pm 2.6}$	92.7$_{\pm 4.2}$	84.8$_{\pm 5.1}$	74.3$_{\pm 5.5}$
	DTFD-MIL	87.6$_{\pm 6.9}$	86.1$_{\pm 7.3}$	77.0$_{\pm 9.2}$	95.2$_{\pm 1.0}$	86.6$_{\pm 1.3}$	86.1$_{\pm 1.1}$	92.5$_{\pm 4.9}$	83.7$_{\pm 6.5}$	72.2$_{\pm 8.1}$
	TransMIL	86.9$_{\pm 5.2}$	85.9$_{\pm 6.0}$	82.0$_{\pm 6.8}$	93.8$_{\pm 0.7}$	86.9$_{\pm 1.0}$	86.3$_{\pm 0.9}$	92.8$_{\pm 3.6}$	83.7$_{\pm 3.9}$	72.8$_{\pm 4.0}$
	CAMIL	88.2$_{\pm 6.5}$	85.1$_{\pm 7.6}$	83.3$_{\pm 8.4}$	93.6$_{\pm 0.9}$	87.1$_{\pm 1.2}$	86.6$_{\pm 0.9}$	92.5$_{\pm 4.2}$	84.4$_{\pm 4.4}$	72.0$_{\pm 4.3}$
	ASMIL	79.7$_{\pm 6.8}$	72.7$_{\pm 7.4}$	68.6$_{\pm 7.2}$	92.5$_{\pm 1.1}$	85.6$_{\pm 1.1}$	85.0$_{\pm 1.5}$	88.9$_{\pm 7.7}$	80.0$_{\pm 8.3}$	69.2$_{\pm 5.6}$
	MR-ABMIL	84.3$_{\pm 5.6}$	83.8$_{\pm 6.0}$	75.3$_{\pm 7.3}$	94.8$_{\pm 1.1}$	88.9$_{\pm 1.4}$	**88.6**$_{\pm 1.6}$	92.9$_{\pm 5.5}$	84.9$_{\pm 5.9}$	73.8$_{\pm 5.0}$
	ViLa-MIL	85.3$_{\pm 7.1}$	84.4$_{\pm 7.3}$	80.5$_{\pm 7.4}$	93.2$_{\pm 2.6}$	86.2$_{\pm 2.7}$	86.1$_{\pm 2.6}$	91.3$_{\pm 5.5}$	81.4$_{\pm 5.5}$	70.6$_{\pm 6.8}$
	FOCUS	87.3$_{\pm 6.2}$	85.0$_{\pm 7.3}$	82.0$_{\pm 7.5}$	94.0$_{\pm 1.7}$	87.4$_{\pm 1.6}$	87.3$_{\pm 1.4}$	92.5$_{\pm 4.8}$	83.3$_{\pm 4.9}$	72.1$_{\pm 4.9}$
	GeoMIL	**92.5**$_{\pm 5.3}$	**89.0**$_{\pm 6.8}$	**86.2**$_{\pm 8.1}$	95.9$_{\pm 0.6}$	**89.0**$_{\pm 1.2}$	88.4$_{\pm 1.5}$	**94.7**$_{\pm 1.2}$	**86.5**$_{\pm 4.8}$	**78.0**$_{\pm 5.1}$

the margins on the three datasets should have been roughly comparable, so this ordering itself points to the suppression of spurious correlations. The gains in Table 1 do not fade as the number of annotations grows, which indicates that what our model compensates for is not primarily a shortage of samples but a bias that is independent of sample size. MR-ABMIL and ASMIL each constrain one end of the problem, feature geometry and attention distribution respectively, yet neither produces a stable gain, which indirectly shows that handling the two separately makes it hard to improve classification performance and attention quality at the same time.

4.3 Ablation Studies

Ablation Study of GeoMIL Components. Table 2 removes each component one at a time to quantify its contribution, and all variants share the same features, the same classification head, and the same training procedure. Removing the student branch and the frozen teacher incurs the largest loss, which confirms the necessity of performing task adaptation and feature manifold protection at the same time, since task adaptation or manifold protection alone both degrade performance. The ablations above correspond to the manifold-protection side and the attention-generation side respectively, and neither side alone yields a stable gain, consistent with our claim that a single learned metric drives both.

Ablation Study of TSLP Projection Sources. Keeping the rest of TSLP and the downstream GQA unchanged and replacing only the projection source isolates the effect of the geometric source itself (Table 3). The four projection sources have identical tensor interfaces and dimensions. Metric ranks first in every combination, while the three alternative projections lag consistently and by a clear margin. This shows that the key is not whether a low-rank projection is applied, but whether the projection is learned adaptively from data: Random and PCA contain no trainable parameters, and although Linear is trainable it lacks geometric pretraining, so none of the three can provide data-

Table 2: Ablation studies of our proposed method. Each row removes a single component from the full model: - student removes the entire student branch, - teacher uses only a trainable student for projection, and - GQA replaces the query mechanism with gated attention.

	Methods	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
4-shot	**GeoMIL**	62.8$_{\pm 8.1}$	**64.3**$_{\pm 8.2}$	**56.0**$_{\pm 9.2}$	**91.9**$_{\pm 4.3}$	**83.7**$_{\pm 6.0}$	**84.7**$_{\pm 5.1}$	**88.9**$_{\pm 7.5}$	**75.3**$_{\pm 8.5}$	**70.0**$_{\pm 8.1}$
	- TSLP	57.4$_{\pm 9.4}$	59.7$_{\pm 8.6}$	51.8$_{\pm 10.2}$	88.4$_{\pm 4.4}$	81.6$_{\pm 3.2}$	81.3$_{\pm 3.5}$	86.5$_{\pm 6.1}$	73.1$_{\pm 8.3}$	67.4$_{\pm 7.8}$
	- student	55.5$_{\pm 7.3}$	57.2$_{\pm 8.3}$	52.0$_{\pm 8.6}$	83.2$_{\pm 3.7}$	76.1$_{\pm 8.9}$	79.2$_{\pm 3.7}$	82.1$_{\pm 8.6}$	66.2$_{\pm 9.2}$	59.5$_{\pm 11.0}$
	- teacher	54.1$_{\pm 9.0}$	55.6$_{\pm 9.2}$	49.0$_{\pm 10.2}$	85.7$_{\pm 4.5}$	79.7$_{\pm 3.8}$	79.2$_{\pm 5.8}$	85.4$_{\pm 8.9}$	68.1$_{\pm 12.6}$	66.6$_{\pm 9.8}$
	- GQA	62.1$_{\pm 11.0}$	63.9$_{\pm 9.3}$	54.1$_{\pm 14.1}$	90.5$_{\pm 3.7}$	82.7$_{\pm 2.6}$	82.8$_{\pm 3.0}$	86.7$_{\pm 7.7}$	67.7$_{\pm 8.2}$	64.8$_{\pm 8.0}$
	- $\mathcal{L}_{\text{tangent}}$	**63.5**$_{\pm 8.2}$	59.5$_{\pm 9.2}$	55.6$_{\pm 10.8}$	89.7$_{\pm 4.4}$	80.2$_{\pm 3.9}$	80.3$_{\pm 4.2}$	87.1$_{\pm 7.1}$	72.1$_{\pm 9.5}$	67.7$_{\pm 9.5}$
8-shot	**GeoMIL**	**79.8**$_{\pm 7.6}$	**77.4**$_{\pm 7.0}$	**72.8**$_{\pm 8.6}$	**93.5**$_{\pm 0.9}$	**86.2**$_{\pm 1.4}$	**85.8**$_{\pm 1.6}$	**93.6**$_{\pm 5.1}$	**83.1**$_{\pm 7.4}$	**74.5**$_{\pm 7.1}$
	- TSLP	74.1$_{\pm 8.0}$	71.6$_{\pm 8.5}$	66.4$_{\pm 8.1}$	91.6$_{\pm 2.0}$	82.7$_{\pm 2.3}$	82.3$_{\pm 1.9}$	89.4$_{\pm 7.1}$	79.3$_{\pm 9.0}$	70.3$_{\pm 9.7}$
	- student	69.6$_{\pm 8.1}$	64.5$_{\pm 7.6}$	58.5$_{\pm 9.6}$	85.8$_{\pm 4.0}$	78.4$_{\pm 4.4}$	80.1$_{\pm 5.0}$	83.3$_{\pm 6.5}$	72.9$_{\pm 7.5}$	66.1$_{\pm 8.9}$
	- teacher	69.2$_{\pm 9.0}$	65.8$_{\pm 9.8}$	59.0$_{\pm 11.6}$	88.3$_{\pm 5.6}$	81.1$_{\pm 5.4}$	82.2$_{\pm 6.2}$	87.4$_{\pm 6.2}$	75.9$_{\pm 7.0}$	69.5$_{\pm 6.8}$
	- GQA	77.8$_{\pm 9.7}$	72.1$_{\pm 10.3}$	71.4$_{\pm 9.3}$	92.4$_{\pm 4.7}$	84.1$_{\pm 3.6}$	83.9$_{\pm 2.5}$	92.8$_{\pm 3.2}$	82.3$_{\pm 2.8}$	67.3$_{\pm 6.8}$
	- $\mathcal{L}_{\text{tangent}}$	77.3$_{\pm 8.2}$	69.2$_{\pm 9.3}$	72.1$_{\pm 7.2}$	92.3$_{\pm 1.4}$	84.3$_{\pm 3.1}$	83.8$_{\pm 2.9}$	90.7$_{\pm 5.2}$	80.9$_{\pm 4.5}$	69.5$_{\pm 5.9}$

Table 3: Ablation studies of projection sources. Metric is the learned metric M_{ε}^{θ} of this work; Random is a row-orthogonal frozen random matrix; PCA is the principal components estimated on the training set; Linear is a linear projection without CDC pretraining.

	Methods	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
4-shot	**Metric**	**62.8**$_{\pm 8.1}$	**64.3**$_{\pm 8.2}$	**56.0**$_{\pm 9.2}$	**91.9**$_{\pm 4.3}$	**83.7**$_{\pm 6.0}$	**84.7**$_{\pm 5.1}$	**88.9**$_{\pm 7.5}$	**75.3**$_{\pm 8.5}$	**70.0**$_{\pm 8.1}$
	Random	53.9$_{\pm 4.9}$	55.2$_{\pm 8.0}$	48.1$_{\pm 9.4}$	85.1$_{\pm 9.6}$	75.5$_{\pm 7.9}$	77.8$_{\pm 6.6}$	76.4$_{\pm 7.1}$	70.1$_{\pm 4.8}$	67.5$_{\pm 2.0}$
	PCA	56.5$_{\pm 7.6}$	56.7$_{\pm 8.2}$	55.0$_{\pm 8.4}$	86.7$_{\pm 8.6}$	78.4$_{\pm 9.8}$	79.3$_{\pm 9.8}$	83.1$_{\pm 9.4}$	72.9$_{\pm 8.4}$	68.3$_{\pm 8.4}$
	Linear	52.7$_{\pm 7.1}$	55.5$_{\pm 8.0}$	55.5$_{\pm 8.3}$	89.8$_{\pm 3.8}$	81.4$_{\pm 3.2}$	82.2$_{\pm 3.4}$	75.8$_{\pm 8.2}$	70.8$_{\pm 9.2}$	65.5$_{\pm 10.3}$
8-shot	**Metric**	**79.8**$_{\pm 7.6}$	**77.4**$_{\pm 7.0}$	**72.8**$_{\pm 8.6}$	**93.5**$_{\pm 0.9}$	**86.2**$_{\pm 1.4}$	**85.8**$_{\pm 1.6}$	**93.6**$_{\pm 5.1}$	**83.1**$_{\pm 7.4}$	**74.5**$_{\pm 7.1}$
	Random	72.1$_{\pm 7.3}$	71.6$_{\pm 5.9}$	67.5$_{\pm 9.9}$	88.8$_{\pm 7.4}$	81.4$_{\pm 6.5}$	80.7$_{\pm 6.2}$	83.1$_{\pm 7.3}$	74.5$_{\pm 6.2}$	64.0$_{\pm 8.2}$
	PCA	76.1$_{\pm 8.5}$	74.9$_{\pm 8.1}$	69.0$_{\pm 9.5}$	92.1$_{\pm 4.8}$	83.6$_{\pm 4.5}$	83.1$_{\pm 5.2}$	86.9$_{\pm 4.4}$	77.0$_{\pm 6.5}$	67.9$_{\pm 9.1}$
	Linear	74.4$_{\pm 8.8}$	72.3$_{\pm 7.9}$	67.9$_{\pm 8.0}$	89.1$_{\pm 6.9}$	81.9$_{\pm 5.9}$	81.8$_{\pm 5.5}$	84.2$_{\pm 8.2}$	71.4$_{\pm 8.8}$	63.1$_{\pm 9.5}$

driven geometric protection. The result indicates that the geometry has to be learned from data, and that neither a preset nor a geometry-agnostic projection can substitute for it.

Ablation Study of Query Anchor. Keeping the other components unchanged and replacing only the source of the query $\mathbf{q}$ isolates the effect of the query design itself (Table 4). GQA ranks first in every combination, which shows that the gain comes from anchoring the query to the manifold structure of the slide rather than from the trainable capacity of the query itself. Instance computes the similarity between each instance and a key instance as the attention weight (Li et al., 2021), and its poor performance shows that under few-shot conditions selecting the query from instances within the bag injects instance-level noise into the query and fails to anchor correctly to the manifold structure of the slide.

4.4 Attention Quality and Localization

Figure 3 compares the attention distributions of GeoMIL and three representative baselines side by side on the same batch of CAMELYON-16 test slides. In Figure 3, the leftmost column is the original WSI, in which the yellow contour marks the ground-truth tumor region, and the slide itself consists of two parts, purple stained tissue and a white transparent background. The remaining four columns show, from left to right, the attention heatmaps of GeoMIL (ours), MR-ABMIL (Xiong et al., 2026), ABMIL (Ilse et al., 2018), and CLAM (Lu et al., 2021), each annotated above with the Dice coefficient and the localization AUROC$_{\text{L}}$ of that slide. All four models share the same features, the same test slides, and the same color scale, so that the heatmaps are directly comparable. As Figure 3 shows, the attention of all four models is mainly concentrated within the yellow contour, and GeoMIL attains an average Dice of 52.3 and an average localization AUROC$_{\text{L}}$ of 98.2 over the test slides, outperforming the other three models. Moreover, the attention of ABMIL and CLAM clearly spills over onto the white transparent region outside the tissue. That region contains neither tissue nor any diagnostic evidence, so any non-zero weight there can only arise from non-semantic

Table 4: Ablation studies of query anchor. GQA is our global geometric centroid, Fixed is a fixed query $\mathbf{w}$, Instance is generated from key instances within the bag, and Inducing is a learnable inducing point.

	Methods	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
4-shot	**GQA**	**62.8**$_{\pm 8.1}$	**64.3**$_{\pm 8.2}$	**56.0**$_{\pm 9.2}$	**91.9**$_{\pm 4.3}$	**83.7**$_{\pm 6.0}$	**84.7**$_{\pm 5.1}$	**88.9**$_{\pm 7.5}$	**75.3**$_{\pm 8.5}$	**70.0**$_{\pm 8.1}$
	Fixed	60.1$_{\pm 8.1}$	62.4$_{\pm 9.3}$	53.1$_{\pm 10.4}$	89.8$_{\pm 3.0}$	81.3$_{\pm 4.6}$	82.5$_{\pm 3.0}$	87.4$_{\pm 6.3}$	72.5$_{\pm 9.4}$	69.3$_{\pm 8.7}$
	Instance	52.9$_{\pm 5.2}$	54.7$_{\pm 5.9}$	49.1$_{\pm 9.0}$	82.2$_{\pm 9.3}$	67.6$_{\pm 10.5}$	72.7$_{\pm 8.4}$	81.6$_{\pm 7.9}$	71.4$_{\pm 11.1}$	61.0$_{\pm 10.3}$
	Inducing	61.3$_{\pm 8.9}$	63.1$_{\pm 9.1}$	54.5$_{\pm 11.8}$	89.4$_{\pm 3.8}$	81.9$_{\pm 3.6}$	82.5$_{\pm 3.3}$	86.8$_{\pm 5.5}$	71.7$_{\pm 9.9}$	68.5$_{\pm 7.8}$
8-shot	**GQA**	**79.8**$_{\pm 7.6}$	**77.4**$_{\pm 7.0}$	**72.8**$_{\pm 8.6}$	**93.5**$_{\pm 0.9}$	**86.2**$_{\pm 1.4}$	**85.8**$_{\pm 1.6}$	**93.6**$_{\pm 5.1}$	**83.1**$_{\pm 7.4}$	**74.5**$_{\pm 7.1}$
	Fixed	77.3$_{\pm 8.6}$	71.3$_{\pm 9.8}$	68.6$_{\pm 8.3}$	91.4$_{\pm 1.6}$	83.9$_{\pm 3.2}$	83.4$_{\pm 3.3}$	92.2$_{\pm 2.4}$	80.3$_{\pm 8.2}$	72.0$_{\pm 9.6}$
	Instance	71.6$_{\pm 6.6}$	61.2$_{\pm 8.7}$	62.2$_{\pm 7.3}$	88.1$_{\pm 2.5}$	79.9$_{\pm 3.1}$	78.8$_{\pm 4.8}$	88.9$_{\pm 3.6}$	74.5$_{\pm 9.5}$	68.3$_{\pm 8.1}$
	Inducing	77.1$_{\pm 9.0}$	72.6$_{\pm 9.6}$	66.3$_{\pm 11.0}$	92.9$_{\pm 0.2}$	85.7$_{\pm 1.3}$	84.9$_{\pm 1.5}$	91.4$_{\pm 4.7}$	82.1$_{\pm 6.0}$	72.7$_{\pm 7.3}$

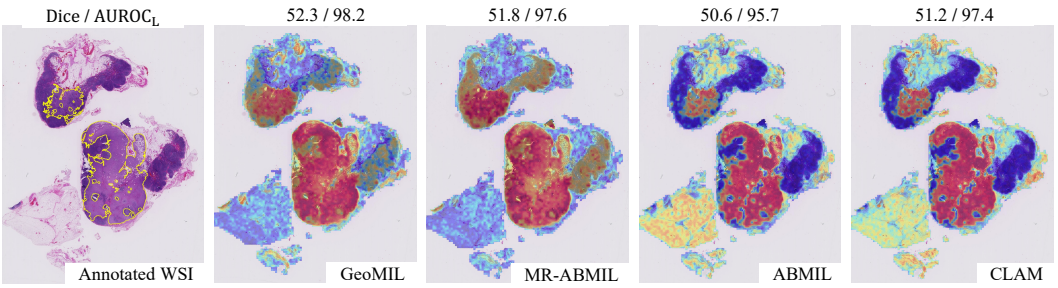


Figure 3: **Visual comparison of attention maps on CAMELYON-16.** The left column shows the original WSI, with the ground-truth tumor annotation outlined in yellow; the slide consists of purple stained tissue and a white transparent background. The remaining columns show the attention maps of GeoMIL (ours), MR-ABMIL (Xiong et al., 2026), ABMIL (Ilse et al., 2018), and CLAM (Lu et al., 2021), from left to right. The corresponding Dice coefficient and localization AUROC$_L$ are displayed above each column.

factors. By contrast, the attention of GeoMIL and MR-ABMIL remains within the tissue region, and the Dice and AUROC$_L$ of MR-ABMIL are also higher than those of ABMIL and CLAM. Preserving feature geometry therefore improves the attention distribution.

5 Conclusion

We propose **GeoMIL**, a geometry-aware multiple instance learning framework driven by a single learned Riemannian metric. Its starting point is a fact that the previously treated two separate problems are in fact interrelated: the degradation of feature geometry and the instability of the attention distribution are two ends of the same causal chain. We therefore do not patch the two separately, but let one learnable Riemannian metric carry both feature transformation and attention generation at once. Along this line, we design TSLP to perform dimensionality reduction and task adaptation through a tangent-space projection together with an independent residual projection, preserving the manifold structure; and GQA to determine the query from the global geometric position of the features in the tangent space, so that attention is anchored to a stable geometry. In few-shot experiments on CAMELYON-16, TCGA-NSCLC, and TCGA-BRCA, GeoMIL achieves consistently superior classification performance.

Ethics Statement

The datasets used in this work, CAMELYON-16, TCGA-NSCLC, and TCGA-BRCA, are all publicly available and can be obtained through open channels. Their use strictly follows the terms of the respective dataset providers and the norms of the academic research community, and involves no personally identifiable information or sensitive patient data. This work studies weakly supervised classification and attention analysis of pathology images; all conclusions are intended for research purposes only and do not constitute a basis for clinical diagnosis.

Reproducibility statement

All model and algorithmic details, the training pipeline, hyperparameters, evaluation protocols, and metrics are given in the main text: the problem formulation and the metric network in Sections 3.1 and 3.2, TSLP and GQA in Sections 3.3 and 3.4, and the two-stage training and the losses in Section 3.5. The experimental protocol, dataset splits, and implementation details are given in Section 4.1 and Appendix C. The complete proof of Theorem 1 is given in Appendix A, and the controlled diagnostic protocol of Figure 1 is given in Appendix B. The three datasets used in this work are all public, and their splits and preprocessing steps are documented in Appendix C.

AI use statement

Generative AI tools were used during the writing of this paper for literature organization, language polishing, and the preparation of the initial draft. All method design, theoretical derivations, experiments, and conclusions are the responsibility of the authors and have been verified item by item. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Georgios Arvanitidis, Lars K Hansen, and Søren Hauberg. A locally adaptive normal distribution. *Advances in Neural Information Processing Systems*, 29, 2016.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Georgios Arvanitidis, Søren Hauberg, and Bernhard Schölkopf. Geometrically enriched latent spaces. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 631–639, 2021.

Jacob Bamberger, Adam Gosztolai, Pierre Vandergheynst, Michael Bronstein, and Iolo Jones. Riemannian metric matching for scalable geometric modeling of distributions. In *International Conference on Machine Learning*, 2026.

Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. *Metric learning*. Morgan & Claypool Publishers, 2015.

Pratik Prabhanjan Brahma, Dehao Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 27 (10):1997–2008, 2016.

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

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. 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.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Willem Diepeveen, Carlos Esteve-Yagüe, Jan Lellmann, Ozan Öktem, and Carola-Bibiane Schönlieb. Riemannian geometry for efficient analysis of protein dynamics data. *Proceedings of the National Academy of Sciences*, 121(33):e2318951121, 2024.

Babak Ehteshami Bejnordi 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.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Olga Fourkioti, Matt De Vries, and Chris Bakal. Camil: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *International Conference on Learning Representations*, 2024.

Kexue Fu, Xiaoyuan Luo, Linhao Qu, Shuo Wang, Ying Xiong, Ilias Maglogiannis, Longxiang Gao, and Manning Wang. Fast: A dual-tier few-shot learning paradigm for whole slide image classification. *Advances in Neural Information Processing Systems*, 37:105090–105113, 2024.

Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15590–15600, 2025.

Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In *International conference on medical image computing and computer-assisted intervention*, pp. 755–764. Springer, 2023.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316, 2023.

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

Andrey Kharitenko, Zebang Shen, Riccardo De Santi, Niao He, and Florian Dörfler. Landing with the score: Riemannian optimization through denoising. In *International Conference on Learning Representations*, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021.

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

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, et al. Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv preprint arXiv:2407.18449*, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Linhao Qu, Xinyi Luo, Kexue Fu, Manning Wang, and Zhijian Song. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67551–67564, 2023.

Linhao Qu, Dingkang Yang, Ding Huang, Qian Guo, Rui Luo, Shaoqing Zhang, and Xiaoying Wang. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In *European Conference on Computer Vision*, volume 15069 of *Lecture Notes in Computer Science*, pp. 196–212, 2024.

Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? In *Forty-second International Conference on Machine Learning*, 2025.

Daniel Shao, Josephine Runevic, Richard J. Chen, Drew F. K. Williamson, Andrew Kim, Andrew H. Song, and Faisal Mahmood. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In *International Conference on Learning Representations*, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021.

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11248–11258, 2024.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11566–11578. IEEE, 2024.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4055–4064. IEEE, 2023.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11343–11352. IEEE, 2024.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *International Conference on Learning Representations*, 2023.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15, 2002.

Conghao Xiong, Hao Chen, Joseph J. Y. Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In *International Joint Conference on Artificial Intelligence*, pp. 1587–1595, 2023.

Conghao Xiong, Yi Lin, Hao Chen, Hao Zheng, Dong Wei, Yefeng Zheng, Joseph JY Sung, and Irwin King. Takt: Target-aware knowledge transfer for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 503–513. Springer, 2024.

Conghao Xiong, Zhengrui Guo, Zhe Xu, Yifei Zhang, Raymond Kai-yu Tong, Si Yong Yeo, Hao Chen, Joseph JY Sung, and Irwin King. Exploiting low-dimensional manifold of features for few-shot whole slide image classification. In *International Conference on Learning Representations*, volume 2026, pp. 86203–86232, 2026.

Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Cheng Jin, Shu Yang, Jinbang Li, Zhengyu Zhang, Chenglong Zhao, Huajun Zhou, et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature communications*, 16(1):11406, 2025.

Linfeng Ye, Shayan Mohajer Hamidi, Zhixiang Chi, Guang Li, Mert Pilanci, Takahiro Ogawa, Miki Haseyama, and Konstantinos N. Plataniotis. Asmil: Attention-stabilized multiple instance learning for whole slide imaging. In *International Conference on Learning Representations*, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022.

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

Yunlong Zhang, Honglin Li, Yuxuan Sun, Zhongyi Shui, Jingxiong Li, Chenglu Zhu, and Lin Yang. Aem: attention entropy maximization for multiple instance learning based whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 45–55. Springer, 2025.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004.

Wenhui Zhu, Peijie Qiu, Xiwen Chen, Zhangsihao Yang, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. How effective can dropout be in multiple instance learning? *arXiv preprint arXiv:2504.14783*, 2025.

A Properties of the Learned Riemannian Metric

This appendix provides the theoretical support for Section 3.3 of the main text. We first state the tangent-space property of the learned Riemannian metric (Property 1) together with the scope within which it holds, then give the proof of Theorem 1 (Appendix A.1), and finally explain the geometric behavior of TSLP under this property (Appendix A.2).

Property 1 (The learned Riemannian metric converges to the tangent-space projection operator)**.** *(Bamberger et al., 2026) Let μ be a probability measure on $\mathbb{R}^d$ and let $\mathbf{x} \in \mathrm{supp}(\mu)$. Suppose there exists $\delta > 0$ such that $B(\mathbf{x}, \delta) \cap \mathrm{supp}(\mu)$ is a $d_{\mathcal{M}}$-dimensional manifold (whose intrinsic dimension $d_{\mathcal{M}}$ may vary with $\mathbf{x}$) carrying an induced Riemannian metric g, and that μ has a smooth density on this manifold. Then the learned Riemannian metric $\Gamma_\varepsilon^\theta$ and its low-rank factor M_ε^θ satisfy the following convergences:*

1. ***Convergence of the CDC.*** *As $\varepsilon \to 0$, for any smooth functions f, h, we have $\Gamma_\varepsilon(f,h)(\mathbf{x}) \to g_{\mathbf{x}}(\nabla f, \nabla h)$; equivalently, $\partial f(\mathbf{x})^\top \Gamma_\varepsilon(\mathbf{x})\, \partial h(\mathbf{x}) \to g_{\mathbf{x}}(\nabla f, \nabla h)$. Here, Γ_ε is the target quantity that the learned metric $\Gamma_\varepsilon^\theta$ approximates.*
2. ***Low-rank factorization.*** *If $r \geq 2d_{\mathcal{M}} - 1$ and M_ε^θ minimizes the low-rank loss, then as $\varepsilon \to 0$ we have $M_\varepsilon^{\theta\top} M_\varepsilon^\theta(\mathbf{x}) \to G(\mathbf{x})$, where $G(\mathbf{x}) = (\Gamma(x_k, x_\ell)(\mathbf{x}))_{k\ell} = \Gamma_\varepsilon(\mathbf{x})$ is the CDC matrix in ambient coordinates. Moreover, as $\varepsilon \to 0$, $G(\mathbf{x})$ converges to the tangent-space projection operator P_T, whose eigenvalues tend to 1 along tangent directions and to 0 along normal directions.*

By Property 1, we have $M_\varepsilon^{\theta\top} M_\varepsilon^\theta \to P_T$; correspondingly, $\Gamma_\varepsilon^\theta = M_\varepsilon^{\theta\top} M_\varepsilon^\theta + \lambda I \to P_T + \lambda I$. The Tikhonov term guarantees that $\Gamma_\varepsilon^\theta$ is strictly positive definite and leaves only an $O(\lambda)$ offset in the limit, without changing the separation between the tangent and the normal components.

A.1 Theorem 1 and Its Proof

For notational brevity, let M denote $M_\varepsilon^{\theta_{\mathrm{te}}}$ (the teacher factor).

Theorem (Tangent-space preservation, restated)**.** *Assume that the data locally lie on a $d_{\mathcal{M}}$-dimensional manifold and that $r \geq 2d_{\mathcal{M}} - 1$. Then, in the limit $\varepsilon \to 0$, for any $\mathbf{x} \in \mathcal{M}$, $\mathbf{v}, \mathbf{w} \in T_{\mathbf{x}}\mathcal{M}$ and $\mathbf{n} \perp T_{\mathbf{x}}\mathcal{M}$: (i) (tangential isometry) $\|M(\mathbf{x})\mathbf{v}\| \to \|\mathbf{v}\|$; (ii) (normal annihilation) $\|M(\mathbf{x})\mathbf{n}\| \to 0$; (iii) (inner-product preservation) $\langle M\mathbf{v}, M\mathbf{w}\rangle \to \langle \mathbf{v}, \mathbf{w}\rangle$; (iv) (tangent–normal separability) writing $\mathbf{x} = \mathbf{x}_T + \mathbf{x}_N$ for the tangent–normal decomposition with respect to $T_{\mathbf{x}}\mathcal{M}$, we have $M(\mathbf{x})\mathbf{x} \to M(\mathbf{x})\mathbf{x}_T$ and $\|M(\mathbf{x})\mathbf{x}_N\| \to 0$.*

Proof of Theorem 1. For any $\mathbf{u} \in \mathbb{R}^d$ we have $\|M\mathbf{u}\|^2 = \mathbf{u}^\top (M^\top M)\mathbf{u}$; since $M^\top M \to P_T$, it follows that $\|M\mathbf{u}\|^2 \to \mathbf{u}^\top P_T \mathbf{u}$. Because P_T is the orthogonal projection onto $T_{\mathbf{x}}\mathcal{M}$, it satisfies $P_T\mathbf{v} = \mathbf{v}$ (for $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$) and $P_T\mathbf{n} = 0$ (for $\mathbf{n} \perp T_{\mathbf{x}}\mathcal{M}$), which yields (i) and (ii). For (iii), the polarization identity

$$\langle M\mathbf{v}, M\mathbf{w}\rangle = \tfrac{1}{4}\big(\|M(\mathbf{v}+\mathbf{w})\|^2 - \|M(\mathbf{v}-\mathbf{w})\|^2\big) \tag{A1}$$

together with $\mathbf{v} \pm \mathbf{w} \in T_{\mathbf{x}}\mathcal{M}$ allows (i) to be applied to both terms on the right-hand side, giving $\langle M\mathbf{v}, M\mathbf{w}\rangle \to \frac{1}{4}(\|\mathbf{v}+\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2) = \langle \mathbf{v}, \mathbf{w}\rangle$. For (iv), linearity gives $M\mathbf{x} = M\mathbf{x}_T + M\mathbf{x}_N$ with $\mathbf{x}_N \perp T_{\mathbf{x}}\mathcal{M}$, so (ii) gives $\|M\mathbf{x}_N\| \to 0$ and hence $M\mathbf{x} \to M\mathbf{x}_T$. □

Theorem 1 holds in the same scope as Property 1, namely in the limit $\varepsilon \to 0$, and the rank condition $r \geq 2d_{\mathcal{M}} - 1$ is already contained in that property; the concrete meaning of this result for the TSLP forward pass is given in Appendix A.2.

A.2 Geometric Behavior of TSLP

Under Property 1, the geometric behavior of TSLP can be described term by term.

The tangential component is approximately distance-preserving. By (i) and (iii) of Theorem 1, for any tangent directions $\mathbf{v}, \mathbf{w} \in T_{\mathbf{x}}\mathcal{M}$ we have $\|M_\varepsilon^{\theta_{\mathrm{te}}}(\mathbf{x})\mathbf{v}\| \approx \|\mathbf{v}\|$ and $\langle M_\varepsilon^{\theta_{\mathrm{te}}}\mathbf{v}, M_\varepsilon^{\theta_{\mathrm{te}}}\mathbf{w}\rangle \approx \langle \mathbf{v}, \mathbf{w}\rangle$. That is, the teacher branch acts as an approximate isometry on the tangent space, so that $\mathbf{c} = M_\varepsilon^{\theta_{\mathrm{te}}}(\mathbf{x})\mathbf{x}$ is an isometric encoding of the tangential component: the reduction from d to r does not compress the scale of the semantic directions, and the metric relations of histological morphology are preserved.

The normal component is removed. By (ii) and (iv) of Theorem 1, for any normal direction $\mathbf{n} \perp T_{\mathbf{x}}\mathcal{M}$ we have $\|M_\varepsilon^{\theta_{\mathrm{te}}}(\mathbf{x})\mathbf{n}\| \approx 0$, and $\|M_\varepsilon^{\theta_{\mathrm{te}}}(\mathbf{x})\mathbf{x}_N\| \approx 0$. That is, normal-space variation such as staining intensity, scanning stripes, and tissue folding has already been removed before entering the task space, and $\mathbf{c}$ carries only the tangential component.

The geometry of the representation is determined by the tangent space. Combining (iv), the output of the teacher branch satisfies $M_{\varepsilon}^{\theta_{\mathrm{te}}}(\mathbf{x})\mathbf{x} \approx M_{\varepsilon}^{\theta_{\mathrm{te}}}(\mathbf{x})\mathbf{x}_T$, so that its value is determined solely by the tangential component. The residual branch injects the deviation required for task adaptation on top of this, and its magnitude is constrained sample-wise by $\mathcal{L}_{\text{tangent}}$.

B Controlled Diagnostic Experiment of Figure 1

B.1 Detailed Analysis of the Two Symptoms

The introduction keeps only the conclusion and the pointer; here we give the concrete analysis of this mechanism at the two ends of the pipeline.

(a) Feature side: the discriminative structure is drowned by technical variation. A geometry-agnostic linear map treats tangential and normal directions alike. Once the manifold structure is corrupted, both the energy fraction of the tangential component and its degree of oscillation rise, so that technical factors such as staining, scanning, and folding become the leading directions of the representation. These directions have large amplitude and vary from slide to slide, and therefore account for most of the variance of the representation; the directions needed for discrimination are suppressed and can only fall into a few low-amplitude normal coordinates. The result is that sample-specific technical differences are amplified as if they were signal, while the histological evidence shared across slides is drowned out. This is precisely the source of overfitting in the few-shot regime: the model learns the technical fingerprint of each individual slide rather than a morphological regularity that transfers across slides.

(b) Attention side: the distribution is driven by the tangential component, which is both high-variance and unstable. The two properties each produce one symptom.

- **Sharpness**: the large variance raises the scale of the scores, so that the softmax tends to saturate and attention concentrates on a few instances.
- **Oscillation**: the corruption of the manifold structure makes the tangential component vary both across slides and over the course of training, so that the attention distribution is hard to converge. Even after statistically stabilizing the attention distribution (Ye et al., 2026; Zhang et al., 2025), the problem is not removed at its root: even after the distribution has converged, attention may still keep pointing along the corrupted tangent-space directions.

(c) The common elimination condition. The symptoms at the two ends share a single elimination condition, namely constraining the operator to the tangent space ($P_N\boldsymbol{\varphi} = 0$). This also motivates the designs of TSLP and GQA: the former reduces the feature transformation to the tangent space, and the latter lets the query be generated from a geometric quantity inside the tangent space rather than from a fixed vector in ambient coordinates.

B.2 How the Panels Are Drawn

All three panels are built on the same standard fixed-query attention model.

(a) Attention heatmap. The model is trained till convergence, and the attention weight of each patch of one slide is upsampled into a heatmap according to the spatial coordinates of the patch within the slide.

(b) Tangential energy heatmap. For the same slide, the tangent-space projection P_T^{loc} is estimated at each patch by data-driven local principal component analysis, and the fraction of the patch feature's energy that falls in the tangent space is computed as

$$\tau_i = \frac{\|P_T^{\mathrm{loc}}\Delta_i\|^2}{\|\Delta_i\|^2}, \qquad \Delta_i = \mathbf{x}_i - \bar{\mathbf{x}}, \tag{B1}$$

where $\bar{\mathbf{x}}$ is the mean of the patch features within the slide. The per-patch τ_i is likewise upsampled into a heatmap according to spatial coordinates and shares the same color scale as (a), so that one can judge directly whether the two high-value regions coincide.

(c) Per-epoch curves. Keeping the model and the training setup unchanged and replacing only the input features, we compute the attention distribution of the model on the three types of features:

$$C_0 : \mathbf{x}_i, \qquad C_1 : \bar{\mathbf{x}} + P_T^{\mathrm{loc}} \Delta_i, \qquad C_2 : \bar{\mathbf{x}} + P_N^{\mathrm{loc}} \Delta_i. \tag{B2}$$

Three quantities are recorded at every epoch during training.

- **Attention concentration**: the distance from the attention distribution α^* to the uniform distribution, measured by the Jensen–Shannon divergence,

$$D_{\mathrm{JSD}}(\alpha^*, \mathrm{Unif}) = \tfrac{1}{2} D_{\mathrm{KL}}\big(\alpha^* \,\|\, m\big) + \tfrac{1}{2} D_{\mathrm{KL}}\big(\mathrm{Unif} \,\|\, m\big), \quad m = \tfrac{1}{2}\big(\alpha^* + \mathrm{Unif}\big). \tag{B3}$$

- **Tangential-energy fraction in the representation** τ_h: the τ_i of (b) is recomputed on the representation h after the linear dimensionality-reduction layer with the same formula, that is, per patch we compute how much of the representation's energy falls in its tangent space and then average over the slide,

$$\tau_h = \frac{1}{N} \sum_{i=1}^{N} \frac{\| P_T^{\mathrm{loc},h} \Delta_i^h \|^2}{\| \Delta_i^h \|^2}, \qquad \Delta_i^h = h_i - \bar{h}, \tag{B4}$$

  where h_i is the value of patch i in that representation, $\bar{h}$ is the mean of h_i over the slide, and $P_T^{\mathrm{loc},h}$ is the tangent-space projection estimated on that representation by the same k-NN local PCA as in (b).

- **Effective rank of the representation** $R_{\mathrm{eff}}(h)$: stacking the per-patch values of the representation h into a matrix and taking its singular value decomposition, we form the normalized spectrum $p_i = \lambda_i / \sum_j \lambda_j$ from the squared singular values λ_i and take the exponential of the spectral entropy,

$$R_{\mathrm{eff}}(h) = \exp\Big(-\sum_i p_i \log p_i \Big), \tag{B5}$$

  that is, the effective dimensionality of the representation is estimated epoch by epoch. Unlike τ_h, R_{eff} is a purely spectral quantity that does not rely on the tangent-space estimate, and it can therefore serve as an independent check on the behavior of τ_h (see Appendix B.3).

B.3 Manifold Structure Collapse

Panel (c) of Figure 1 records two quantities of the representation epoch by epoch: the tangential-energy fraction τ_h and the effective rank $R_{\mathrm{eff}}(h)$. During training τ_h rises with the epoch while R_{eff} falls gradually, and together the two indicate that the feature manifold is collapsing.

(a) Two readings of the same geometric quantity. Let d_{eff} denote the effective dimensionality of the representation and t the rank of the tangent space of the local PCA. If the displacement Δ^h is roughly isotropic across the d_{eff} directions, and the t-dimensional local tangent plane sits in general position among them, then the expected fraction of the displacement falling inside the tangent plane is

$$\mathbb{E}\left[\tau_h\right] \approx \frac{t}{d_{\mathrm{eff}}}. \tag{B6}$$

Since t is fixed during training, a rise in τ_h indicates a fall in d_{eff} under this approximation, and R_{eff} is precisely the spectral estimate of d_{eff}. The two quantities are therefore not two unrelated phenomena but two aspects of the same geometric quantity: τ_h describes the change in the direction of the displacement, and R_{eff} the structural dimensionality of the representation.

(b) The two observations jointly indicate manifold structure collapse. R_{eff} is a purely spectral quantity that relies on no geometric estimate, and it therefore gives the structural dimensionality of the representation reliably; but it does not say where the displacement goes once that dimensionality has contracted. τ_h is a geometric quantity that says where the displacement lies, but its definition of the tangent plane comes from the local PCA on that representation, so whether the tangential structure is preserved or the tangent plane is hijacked by the dominant directions, τ_h is high either way. The fall in R_{eff} tells us that the effective dimensionality of the representation has already contracted. Given that, a high τ_h means that the displacement is concentrated in the few directions of largest

variance that currently remain in the representation. By construction the local PCA always returns the top t directions of local variance; what has changed is the nature of those directions themselves, which have turned from the tangential directions of semantic variation into directions dominated by technical factors. The local PCA therefore no longer estimates a tangent space. This combination also rules out the other geometrically opposite failure path: if energy leaked into directions irrelevant to discrimination, R_{eff} would rise while τ_h fell, the opposite of what Figure 1(c) shows.

(c) Mechanism. The only training signal for this linear layer is the slide-level cross-entropy, and under the few-shot setting the number of samples is very small. The gradient only needs to separate the training slides and carries no geometric constraint; contracting directions irrelevant to discrimination lowers the effective dimensionality without increasing the loss. Training therefore keeps compressing the effective rank of the representation, d_{eff} falls, and τ_h approaches its degenerate upper bound $t/d_{\text{eff}} \to 1$.

C Datasets, Preprocessing, and Splits

C.1 Datasets and tasks. All experiments in this work are conducted on three public datasets: CAMELYON-16, TCGA-NSCLC, and TCGA-BRCA. The three datasets cover both the sparse-positive and the dense-positive bag regime, as well as both detection and subtyping tasks, so that the conclusions are not tied to a single tissue origin or a single data regime.

CAMELYON-16 (Ehteshami Bejnordi et al., 2017) is used for breast cancer lymph node metastasis detection. It contains roughly 400 H&E-stained lymph node whole-slide images in two classes, normal and cancer, with pixel-level annotations of the tumor regions available on the validation data. At $20\times$ magnification, the average number of patches extracted per slide is 3697. In the positive slides the tumor region is extremely sparse (average area fraction $< 10\%$), so the dataset is a **sparse-positive** regime in which normal regions account for the large majority of the instances in the bag. This makes it a stringent setting for attention quality: under sparse positivity, any shift of attention caused by non-semantic factors falls directly on instances that carry no diagnostic evidence, so its effect on the classification result is amplified. The pixel-level annotations also make the dataset suitable for localization evaluation; both the Dice coefficient and the localization AUROC_{L} reported in Section 4.4 are computed on it.

TCGA-NSCLC is used for subtyping of non-small cell lung cancer. It contains H&E whole-slide images of two histological subtypes, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), and the task is to distinguish the two at the slide level, with 1143 diagnostic WSIs in total (541 LUAD / 512 LUSC). In contrast to CAMELYON-16, the diagnostic evidence for both subtypes is distributed over the entire slide and the large majority of the instances in the bag carry it, so this is a dense-positive regime. The task emphasizes the discrimination of morphological differences and is a bag-balanced subtype classification task. After preprocessing, at $20\times$ magnification, the average number of patches is 3168.

TCGA-BRCA is a curated binary classification dataset from The Cancer Genome Atlas (TCGA), used to distinguish invasive ductal carcinoma (IDC) from invasive lobular carcinoma (ILC) on diagnostic WSIs, with 1008 diagnostic WSIs in total (807 IDC / 201 ILC). The dataset is large in scale and highly heterogeneous in morphology, and the tumor cells are often interwoven with dense stroma and lymphocytic infiltrate; it is therefore an effective benchmark for testing the generalization ability and the robustness to interference of a model on large-scale, highly heterogeneous data. After preprocessing, at $20\times$ magnification, the average number of patches is 3680.

C.2 Preprocessing and feature extraction. All WSIs are tiled at $20\times$ magnification into non-overlapping 256×256 patches, and the instance features are extracted by a frozen pretrained model. The main experiments uniformly use the 512-dimensional patch features of CONCH (Lu et al., 2024). The data augmentation and preprocessing pipeline is fixed and identical across all methods, and is not adjusted per method.

C.3 Few-shot split protocol. For each dataset, each $k \in \{4, 8, 16\}$, and each random seed $r = 1, \ldots, 5$: (1) k slides are randomly sampled per class to form the support set, which is the only labeled training data; (2) all remaining slides form the test set, from which 20% are held out as a validation set for early stopping; (3) the metric pretraining of Phase 0 uses only the unlabeled patches of the support-set slides, so as to avoid information leakage from the test set; (4) Phase 1 is

Table 5: 2-shot results. Best performance in bold, second best underlined; mean±standard deviation over 5 random seeds.

Method	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
ABMIL	49.8$_{\pm 5.7}$	49.8$_{\pm 5.7}$	44.3$_{\pm 9.7}$	85.0$_{\pm 5.6}$	72.3$_{\pm 4.3}$	73.7$_{\pm 5.5}$	75.5$_{\pm 11.1}$	71.6$_{\pm 9.9}$	51.9$_{\pm 12.7}$
CLAM	49.6$_{\pm 7.5}$	50.4$_{\pm 6.2}$	44.0$_{\pm 12.3}$	86.3$_{\pm 7.7}$	74.0$_{\pm 7.7}$	75.8$_{\pm 9.6}$	77.0$_{\pm 10.0}$	70.0$_{\pm 10.1}$	54.6$_{\pm 9.1}$
DSMIL	49.4$_{\pm 4.8}$	46.1$_{\pm 7.1}$	46.1$_{\pm 7.4}$	85.2$_{\pm 5.4}$	73.6$_{\pm 6.7}$	73.2$_{\pm 6.4}$	76.6$_{\pm 13.1}$	65.9$_{\pm 12.9}$	57.2$_{\pm 13.2}$
DTFD-MIL	50.4$_{\pm 6.0}$	48.1$_{\pm 7.2}$	44.4$_{\pm 10.0}$	86.3$_{\pm 9.1}$	72.4$_{\pm 9.4}$	74.1$_{\pm 7.7}$	76.0$_{\pm 13.4}$	69.6$_{\pm 11.3}$	51.7$_{\pm 13.0}$
TransMIL	48.3$_{\pm 5.1}$	47.4$_{\pm 4.2}$	40.5$_{\pm 11.9}$	82.9$_{\pm 7.0}$	69.1$_{\pm 9.5}$	70.5$_{\pm 8.4}$	74.1$_{\pm 9.3}$	64.5$_{\pm 12.1}$	56.0$_{\pm 11.5}$
CAMIL	47.8$_{\pm 5.6}$	45.4$_{\pm 4.6}$	47.7$_{\pm 6.0}$	84.5$_{\pm 7.1}$	72.0$_{\pm 6.5}$	73.2$_{\pm 9.3}$	73.7$_{\pm 11.1}$	65.2$_{\pm 11.3}$	56.5$_{\pm 12.9}$
ASMIL	47.7$_{\pm 4.0}$	45.6$_{\pm 5.4}$	46.8$_{\pm 7.3}$	81.4$_{\pm 7.2}$	67.4$_{\pm 8.4}$	68.0$_{\pm 8.5}$	72.1$_{\pm 9.5}$	63.0$_{\pm 10.4}$	52.7$_{\pm 9.8}$
MR-ABMIL	51.3$_{\pm 4.9}$	50.5$_{\pm 8.9}$	45.3$_{\pm 9.2}$	84.3$_{\pm 7.7}$	69.6$_{\pm 9.9}$	71.3$_{\pm 6.2}$	76.2$_{\pm 10.0}$	67.5$_{\pm 9.4}$	55.5$_{\pm 9.7}$
ViLa-MIL	50.4$_{\pm 5.3}$	49.8$_{\pm 6.7}$	45.6$_{\pm 8.2}$	85.6$_{\pm 6.4}$	71.8$_{\pm 8.7}$	72.6$_{\pm 7.4}$	76.8$_{\pm 10.2}$	69.6$_{\pm 11.7}$	55.3$_{\pm 11.2}$
FOCUS	51.1$_{\pm 6.4}$	50.2$_{\pm 7.3}$	46.9$_{\pm 9.2}$	86.4$_{\pm 5.2}$	73.5$_{\pm 7.4}$	75.3$_{\pm 6.8}$	77.3$_{\pm 9.5}$	71.4$_{\pm 10.8}$	57.6$_{\pm 12.2}$
GeoMIL	**53.0**$_{\pm 5.6}$	**53.5**$_{\pm 4.5}$	**48.4**$_{\pm 8.5}$	**88.4**$_{\pm 4.7}$	**76.8**$_{\pm 6.7}$	**78.9**$_{\pm 6.0}$	**79.8**$_{\pm 9.1}$	**74.8**$_{\pm 8.6}$	**61.4**$_{\pm 11.6}$

Table 6: Sensitivity to the Phase 0 data amount n_0. The results reported in Table 1 correspond to $n_0 = 4$ (4-shot) and $n_0 = 8$ (8-shot).

	n_0	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
4-shot	1	61.4$_{\pm 7.8}$	62.1$_{\pm 5.5}$	53.8$_{\pm 10.1}$	90.3$_{\pm 5.9}$	82.5$_{\pm 4.1}$	83.6$_{\pm 3.5}$	87.6$_{\pm 7.3}$	73.9$_{\pm 9.2}$	68.0$_{\pm 9.3}$
	2	62.0$_{\pm 7.1}$	63.2$_{\pm 7.7}$	54.2$_{\pm 11.7}$	91.6$_{\pm 4.2}$	83.5$_{\pm 5.3}$	84.0$_{\pm 4.0}$	87.4$_{\pm 7.4}$	74.9$_{\pm 7.6}$	69.8$_{\pm 7.8}$
	4	**62.8**$_{\pm 8.1}$	64.3$_{\pm 8.2}$	**56.0**$_{\pm 9.2}$	91.9$_{\pm 4.3}$	83.7$_{\pm 6.0}$	84.7$_{\pm 5.1}$	**88.9**$_{\pm 7.5}$	75.3$_{\pm 8.5}$	70.0$_{\pm 8.1}$
	8	62.0$_{\pm 7.4}$	63.1$_{\pm 6.0}$	55.8$_{\pm 7.4}$	**92.1**$_{\pm 3.5}$	**84.5**$_{\pm 3.6}$	**85.1**$_{\pm 3.5}$	88.8$_{\pm 6.4}$	**76.9**$_{\pm 7.8}$	**70.3**$_{\pm 8.7}$
	16	61.9$_{\pm 9.8}$	**64.6**$_{\pm 7.2}$	55.2$_{\pm 8.5}$	91.6$_{\pm 4.0}$	82.8$_{\pm 3.9}$	83.5$_{\pm 3.5}$	88.1$_{\pm 7.6}$	73.9$_{\pm 8.8}$	66.5$_{\pm 9.1}$
	32	60.1$_{\pm 5.2}$	62.7$_{\pm 6.3}$	53.5$_{\pm 7.5}$	91.4$_{\pm 4.1}$	83.6$_{\pm 4.7}$	84.4$_{\pm 4.0}$	85.9$_{\pm 10.1}$	76.5$_{\pm 8.1}$	67.0$_{\pm 10.2}$
8-shot	1	79.1$_{\pm 8.7}$	74.0$_{\pm 9.4}$	70.9$_{\pm 7.4}$	91.7$_{\pm 2.4}$	84.5$_{\pm 3.7}$	84.3$_{\pm 3.5}$	91.2$_{\pm 6.5}$	81.0$_{\pm 7.4}$	72.7$_{\pm 8.1}$
	2	79.3$_{\pm 9.8}$	76.1$_{\pm 9.9}$	**73.1**$_{\pm 8.5}$	93.0$_{\pm 1.5}$	85.9$_{\pm 3.2}$	84.6$_{\pm 2.5}$	92.2$_{\pm 3.1}$	82.8$_{\pm 7.9}$	73.0$_{\pm 6.7}$
	4	**80.1**$_{\pm 9.1}$	**79.1**$_{\pm 7.2}$	71.8$_{\pm 9.8}$	93.3$_{\pm 1.7}$	**86.9**$_{\pm 2.9}$	85.5$_{\pm 2.5}$	92.4$_{\pm 4.7}$	82.8$_{\pm 8.2}$	73.4$_{\pm 8.8}$
	8	79.8$_{\pm 7.6}$	77.4$_{\pm 7.0}$	72.8$_{\pm 8.6}$	**93.5**$_{\pm 0.9}$	86.2$_{\pm 1.4}$	**85.8**$_{\pm 1.6}$	**93.6**$_{\pm 5.1}$	**83.1**$_{\pm 7.4}$	**74.5**$_{\pm 7.1}$
	16	79.2$_{\pm 8.3}$	77.4$_{\pm 7.5}$	69.7$_{\pm 8.9}$	92.8$_{\pm 0.8}$	85.5$_{\pm 1.2}$	84.9$_{\pm 1.3}$	91.3$_{\pm 5.9}$	81.2$_{\pm 9.2}$	72.2$_{\pm 9.7}$
	32	79.8$_{\pm 7.5}$	75.3$_{\pm 9.9}$	72.4$_{\pm 7.0}$	92.2$_{\pm 2.5}$	84.5$_{\pm 2.6}$	84.1$_{\pm 3.0}$	90.3$_{\pm 6.8}$	79.7$_{\pm 9.5}$	71.7$_{\pm 10.2}$

trained on the support set with early stopping on the validation set; (5) one model is trained for each (k, r) combination, and the mean and standard deviation over the 5 repetitions are reported. All baselines are run on the same set of (k, r) splits, the same features, and the same evaluation scripts.

D Additional Experimental Results

D.1 2-shot results. Table 5 reports the complete results under the 2-shot setting, that is, only 2 annotated slides per class, which is the most extreme few-shot setting considered in this work.

GeoMIL is best on all three datasets under 2-shot, with a margin of $+0.7$ to $+3.8$ over the second-best method. The most informative case is the sparsely positive CAMELYON-16, where 6 of the 10 baselines have an AUC no higher than chance (TransMIL 48.3, CAMIL 47.8, ASMIL 47.7, DSMIL 49.4, CLAM 49.6, ABMIL 49.8), whereas GeoMIL still reaches 53.0. This is consistent with the observation in Section 4.2: under sparse positivity normal regions account for the large majority of instances, a geometry-agnostic linear projection diverts attention toward technical variation, and 2 positive slides per class are no longer enough to correct that bias. GeoMIL suppresses normal-space artifacts through the tangent-space constraint and therefore retains above-chance discriminability in this setting.

D.2 Sensitivity to the Phase 0 data amount n_0. Here n_0 is the number of slides per class in the support set used for the Phase 0 metric pretraining. Table 6 sweeps $n_0 \in \{1, 2, 4, 8, 16, 32\}$ under both the 4-shot and the 8-shot setting.

The trend is non-monotone but remains stable overall: GeoMIL is optimal at $n_0 \in \{4, 8\}$ in most cases, and raising n_0 to 16 or 32 degrades the performance instead. For example, the AUC of TCGA-NSCLC at 8-shot is 93.5 at $n_0 = 8$ and drops to 92.2 at $n_0 = 32$; the F1 of TCGA-BRCA at 8-shot is 74.5 at $n_0 = 8$ and drops to 71.7 at $n_0 = 32$. This direction runs counter to the intuition that more unlabeled data should yield a better geometric estimate, and indicates that a small number

Table 7: Sensitivity to the metric dimension r. The results reported in Table 1 correspond to $r = 64$.

	r	CAMELYON-16			TCGA-NSCLC			TCGA-BRCA		
		AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
4-shot	16	$60.6_{\pm 8.5}$	$60.5_{\pm 8.1}$	$54.9_{\pm 9.2}$	$\underline{91.7}_{\pm 3.9}$	$\underline{83.3}_{\pm 3.2}$	$\underline{83.5}_{\pm 3.1}$	$\underline{88.3}_{\pm 7.2}$	$\underline{74.8}_{\pm 9.4}$	$\underline{69.1}_{\pm 8.3}$
	32	$\underline{61.5}_{\pm 9.2}$	$\underline{63.5}_{\pm 10.3}$	$\underline{55.3}_{\pm 8.2}$	$91.6_{\pm 4.0}$	$82.3_{\pm 2.8}$	$83.4_{\pm 2.5}$	$86.2_{\pm 9.3}$	$71.1_{\pm 11.5}$	$67.1_{\pm 8.6}$
	64	$\mathbf{62.8}_{\pm \mathbf{8.1}}$	$\mathbf{64.3}_{\pm \mathbf{8.2}}$	$\mathbf{56.0}_{\pm \mathbf{9.2}}$	$\mathbf{91.9}_{\pm \mathbf{4.3}}$	$\mathbf{83.7}_{\pm \mathbf{6.0}}$	$\mathbf{84.7}_{\pm \mathbf{5.1}}$	$\mathbf{88.9}_{\pm \mathbf{7.5}}$	$\mathbf{75.3}_{\pm \mathbf{8.5}}$	$\mathbf{70.0}_{\pm \mathbf{8.1}}$
	128	$60.9_{\pm 8.6}$	$62.9_{\pm 7.6}$	$52.1_{\pm 7.9}$	$88.6_{\pm 5.6}$	$79.9_{\pm 4.3}$	$80.9_{\pm 4.5}$	$87.6_{\pm 8.2}$	$72.8_{\pm 8.1}$	$67.9_{\pm 10.6}$
8-shot	16	$76.7_{\pm 7.6}$	$73.4_{\pm 8.2}$	$72.2_{\pm 8.7}$	$92.4_{\pm 0.8}$	$84.3_{\pm 3.6}$	$\underline{84.8}_{\pm 2.4}$	$\underline{93.2}_{\pm 2.3}$	$80.8_{\pm 9.3}$	$\underline{73.5}_{\pm 7.9}$
	32	$77.7_{\pm 9.0}$	$\underline{76.0}_{\pm 6.7}$	$71.1_{\pm 9.3}$	$\underline{93.4}_{\pm 1.2}$	$\underline{84.9}_{\pm 2.7}$	$84.6_{\pm 2.1}$	$91.9_{\pm 3.7}$	$77.5_{\pm 8.5}$	$69.9_{\pm 9.2}$
	64	$\mathbf{79.8}_{\pm \mathbf{7.6}}$	$\mathbf{77.4}_{\pm \mathbf{7.0}}$	$\mathbf{72.8}_{\pm \mathbf{8.6}}$	$\mathbf{93.5}_{\pm \mathbf{0.9}}$	$\mathbf{86.2}_{\pm \mathbf{1.4}}$	$\mathbf{85.8}_{\pm \mathbf{1.6}}$	$\mathbf{93.6}_{\pm \mathbf{5.1}}$	$\mathbf{83.1}_{\pm \mathbf{7.4}}$	$\mathbf{74.5}_{\pm \mathbf{7.1}}$
	128	$\underline{79.0}_{\pm 8.3}$	$75.7_{\pm 7.1}$	$\underline{72.6}_{\pm 7.7}$	$92.7_{\pm 1.3}$	$84.9_{\pm 2.0}$	$84.4_{\pm 1.4}$	$92.1_{\pm 5.3}$	$\underline{81.2}_{\pm 5.6}$	$73.0_{\pm 7.1}$

of slides already suffices for Phase 0 to estimate the tangent space stably; adding further slides only folds more slide-specific technical variation into the CDC objective without contributing additional geometric information. The default configuration of this work takes n_0 on the same order as k, which also keeps the additional cost of Phase 0 relatively low.

D.3 Sensitivity to the metric dimension r. Here r is the number of rows of the low-rank factor M_ε^θ, that is, the dimension of the tangent-space coefficients. Table 7 sweeps $r \in \{16, 32, 64, 128\}$.

$r = 64$ retains the best performance on all three datasets under both shot settings, and over the range $r \in \{16, 32, 64, 128\}$ the AUC varies by at most 3.3 points on all three datasets. When r is too small the performance drops somewhat (the AUC of CAMELYON-16 at 4-shot falls from 62.8 at $r = 64$ to 60.6 at $r = 16$), which indicates that when the tangent-space rank is below the intrinsic dimension of the manifold, part of the semantic directions may be discarded together with the normal-space artifacts. When r is too large the performance also drops, possibly because beyond the intrinsic dimension of the data the normal-space artifacts re-enter the representation, and because a large r inflates the parameter count and tends to overfit. The degradation at either end is nonetheless limited, which indicates that the model remains stable overall with respect to the metric dimension r.

E Limitations and Future Work

Limitations. Property 1 and Theorem 1 are both statements in the limit $\varepsilon \to 0$: the convergence of the learned metric to the tangent-space projection operator, and the tangential isometry and normal annihilation derived from it, presuppose that limit together with the rank condition $r \geq 2d_{\mathcal{M}} - 1$. Training in practice uses a finite ε, so the results hold only as approximations. In addition, the reading of τ_h and R_{eff} in Appendix B.3 rests on two premises: that the local PCA returns, by construction, the top t local directions, and that the k-NN adjacency is built in the original feature space and then held fixed. The neighborhood scale k does not adapt to the local density of the data, and changes in the adjacency topology fall outside the scope of the present diagnostics. Moreover, the absolute level of R_{eff} depends on the dimensionality of the backbone representation and is not comparable across backbones, so we use it only as a relative criterion within a single backbone.

Future work. A natural direction is to extend the $\varepsilon \to 0$ result to non-asymptotic bounds at finite ε, and to let the tangent-space rank and the neighborhood scale adapt to the local geometry of the data. On the application side, we plan to move beyond the binary detection and subtype classification considered here to multi-class, grading, and survival-prediction targets, so as to test the role of geometric protection on fine-grained and ordinal objectives. Generalizing the tangent-space constraint to graph-based and hierarchical aggregation frameworks, and to multi-stain and multimodal inputs, is likewise a direction we intend to pursue.