RLT: LEARNING FROM RELATIONAL LONG TAILS IN WHOLE SLIDE IMAGE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-tailed whole slide image (WSI) classification is usually studied from the imbalance of slide labels or sparse instance-level evidence. We argue that this view is incomplete. Diagnostic evidence in pathology is also encoded by interactions among heterogeneous tissues, while these tissue relations are themselves highly imbalanced. Common relations dominate graph propagation, whereas rare but class-discriminative relations can be suppressed, especially for tail classes with few training slides. We therefore propose RLT, a relational long-tail framework for WSI classification. First, we construct a morphology-aware heterogeneous graph in which patch nodes are associated with tissue semantics and relational edges encode tissue interactions. Second, we introduce diagnosticity-aware relation debiasing, which combines relation rarity with counterfactual contribution to enhance rare relations only when they provide class-specific evidence. Third, we develop prototype-guided representation enrichment to densify the sparsely sampled representation regions of data-scarce classes and stabilize the decision space. A class-prior-aware Balanced Softmax objective is further used to reduce residual classifier bias. Experiments on Camelyon+-LT and PANDA-LT demonstrate that RLT consistently outperforms representative baselines in both overall and tail-class performance. Code and configurations are available at: `https://anonymous.4open.science/r/RLT_ICLR2027-anonymous/`.

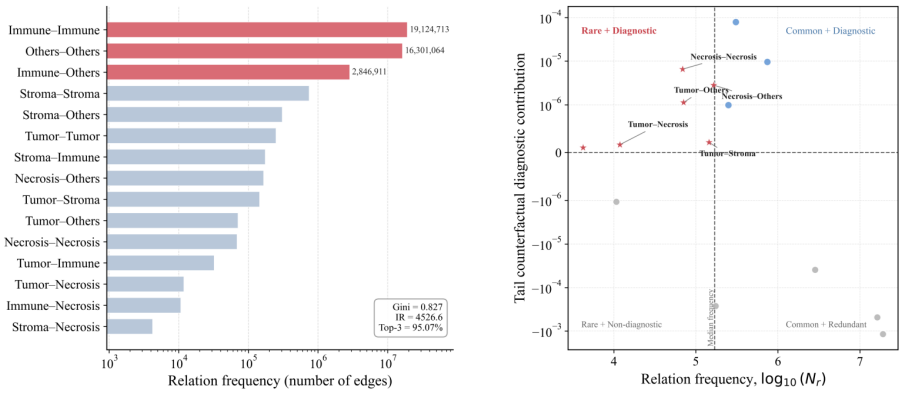
1 INTRODUCTION

Whole slide images (WSIs) contain gigapixel-scale histopathological information and are widely used for computer assisted diagnosis. Because exhaustive region-level annotation is expensive, multiple instance learning (MIL) has become a standard weakly supervised paradigm for WSI analysis. A WSI is divided into patches, patch features are extracted by a pretrained encoder, and an aggregator predicts the slide label from the resulting bag (Ilse et al., 2018; Lu et al., 2021; Shao et al., 2021). Recent pathology foundation models further improve patch representations and make weakly supervised WSI analysis increasingly effective (Lu et al., 2024).

Most WSI methods, however, are developed under a roughly balanced label distribution. Real clinical data are rarely balanced. Common diagnoses may contain hundreds or thousands of slides, whereas rare subtypes can contain only a few dozen. This imbalance across classes biases optimization toward head classes and weakens rare class recognition. Long tail pathology methods therefore introduce supervised contrastive learning, multimodal distillation, expert ensembles, and class aware rebalancing to improve minority classes (Juyal et al., 2024; Ling et al., 2026). More recently, DeCo-MIL (Li et al., 2026) argues that WSI classification contains a nested dual long tail: slide frequencies are imbalanced across classes, while discriminative evidence is also concentrated in a small subset of patches within each slide (Li et al., 2026). This observation explains why class-level rebalancing alone can be insufficient.

We take a different but complementary view. A WSI is not only a set of independent patches. Pathological evidence is often expressed by the coexistence and interaction of heterogeneous tissue components, such as tumor–stroma, tumor–immune, and tumor–necrosis interfaces. Graph methods for WSI analysis have shown that modeling dependencies among patches can provide information that bag aggregation does not explicitly preserve (Hou et al., 2022; Li et al., 2024; Wu et al., 2024). Yet existing graph methods mainly improve relational representation under standard classification

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) Relation Frequency Long Tail (b) Relation Rarity vs. Diagnosticity

Figure 1: Relation-level long-tail analysis on Camelyon+-LT. (a) Tissue-relation frequencies are highly skewed. (b) Some rare relations retain positive tail-class counterfactual contributions, motivating joint modeling of rarity and diagnosticity.

or survival prediction. They do not ask whether the relations themselves become imbalanced in a long tail recognition setting.

Our preliminary analysis reveals a pronounced long tail in tissue relations. When patch nodes are categorized into tumor, stroma, immune, necrosis, and other tissues, a few relation types occupy most graph edges, while many relations occur rarely. As shown in Fig. 1(a), on Camelyon+-LT, the latent graph exhibits a highly skewed relation distribution, with a Gini coefficient of 0.827 and the three most frequent relations accounting for 95.07% of all edges. More importantly, low frequency does not imply low diagnostic value. As illustrated in Fig. 1(b), some rare relations still exhibit positive counterfactual contributions to the tail class. These observations suggest that standard message passing can repeatedly optimize common relations while underusing rare but informative ones.

A complementary challenge arises from the scarcity of tail class slides. Limited training samples provide insufficient coverage of graph representations for tail classes, even when informative relations are preserved. Moreover, highly skewed class priors can still bias the final decision toward head classes. These observations motivate us to enrich tail representations at the graph level and separately calibrate the classifier prior.

Based on these observations, we propose RLT, a relational framework for long tail WSI learning that treats the problem as structured representation learning rather than only as an issue of sample counts. The framework first represents a slide as a heterogeneous graph guided by tissue morphology. It then estimates relation rarity and class-specific diagnosticity, using a frozen judge model and counterfactual relation removal, and injects the resulting relation score into message passing. Finally, prototype based interpolation enriches graph representations, while Balanced Softmax corrects the residual bias from class priors during optimization.

Our contributions are summarized as follows:

- We introduce a relational view of long tail WSI learning. We model WSIs as heterogeneous graphs guided by tissue morphology and empirically identify a long tail in tissue relations, in which a few relations dominate graph propagation while rare relations can remain discriminative for specific classes.
- We propose relation debiasing guided by diagnosticity. Instead of treating all rare relations as important, the method combines relation rarity with counterfactual diagnostic contribution and selectively enhances rare relations that support the target diagnosis.
- We develop prototype based representation enrichment to broaden the representation neighborhood available to classes with scarce data. Together with class prior calibration, this module improves overall class balance and reduces the instability caused by extreme class scarcity.

2 Related Work

2.1 Weakly Supervised WSI Representation Learning

Early WSI learning largely follows the MIL formulation. Ilse et al. (2018) introduced learnable attention pooling to replace fixed bag aggregation, and Lu et al. (2021) further incorporated clustering constraints to improve instance discrimination under supervision from slide labels. To capture dependencies beyond independent instance weighting, Shao et al. (2021) used Transformer correlation modeling, while Xiang & Zhang (2023) exploited low-rank structure to improve bag representation. These methods progressively strengthen the aggregator, but the slide is still primarily treated as a bag of instances.

A complementary line of work models explicit structure between patches. Hou et al. (2022) represented hierarchical dependencies in heterogeneous WSI graphs. From a dynamic interaction perspective, Li et al. (2024) constructed patch relations guided by knowledge to capture dependencies between spatially distant regions. Wu et al. (2024) further incorporated tissue heterogeneity into heterogeneous graphs for survival prediction, showing that tissue identity and interactions across tissues are clinically informative. These graph methods motivate structured WSI modeling, but they do not target imbalanced relation distributions under long-tailed recognition. Our work focuses on this missing interaction between graph structure and long-tailed learning.

2.2 Long-Tailed Recognition in Computational Pathology

Long-tailed recognition has been studied through resampling, margin adjustment, classifier calibration, decoupled learning, and expert ensembles (Buda et al., 2018; Cao et al., 2019; Kang et al., 2019; Ren et al., 2020; Menon et al., 2020; Wang et al., 2020). These methods mainly correct unequal class frequencies or biased decision boundaries. In pathology, Juyal et al. (2024) combined supervised contrastive learning with MIL to improve representation and classifier learning under label imbalance. Ling et al. (2026) introduced multiple expert branches and multimodal distillation to capture complementary distributions for long-tailed WSI analysis.

A more WSI-specific perspective studies imbalance within each slide. Methods that reduce redundancy, such as MHIM-MIL and AMD-MIL, suppress uninformative or easy instances (Tang et al., 2023; Ling et al., 2024). DeCo-MIL explicitly characterizes the coexistence of the class long tail across slides and the long tail of discriminative evidence within each slide, and uses counterfactual contributions corrected by frequency to preserve scarce morphology evidence (Li et al., 2026). Our work is complementary: rather than asking only which instances are informative, we ask which *relations among tissue instances* are informative and whether their frequencies introduce a new bias during message passing.

2.3 Counterfactual Attribution and Prototype Enrichment

Counterfactual attribution evaluates feature importance by intervening on an input and measuring the prediction change. Occlusion and meaningful perturbation methods use this principle for visual explanations (Zeiler & Fergus, 2014; Fong & Vedaldi, 2017), while counterfactual inference has also been used for WSI selection and deconfounding (Lin et al., 2023; 2024). We adopt counterfactual reasoning at the relation level: an intervention removes one tissue relation type while preserving the remaining graph, allowing us to estimate relational diagnosticity for each class.

Prototype learning provides a compact description of recurring data modes. In computational pathology, PANTHER uses morphological prototypes for slide representation (Song et al., 2024). Our prototype module serves a different purpose. It does not replace the classifier or summarize a slide by prototype counts. Instead, it interpolates graph embeddings toward class prototypes during training, increasing representation coverage without duplicating minority samples.

3 Method

Figure 2 summarizes RLT. The framework has three functional stages. First, a morphology-aware heterogeneous graph converts patch sets into typed relational structures. Second, a frozen judge

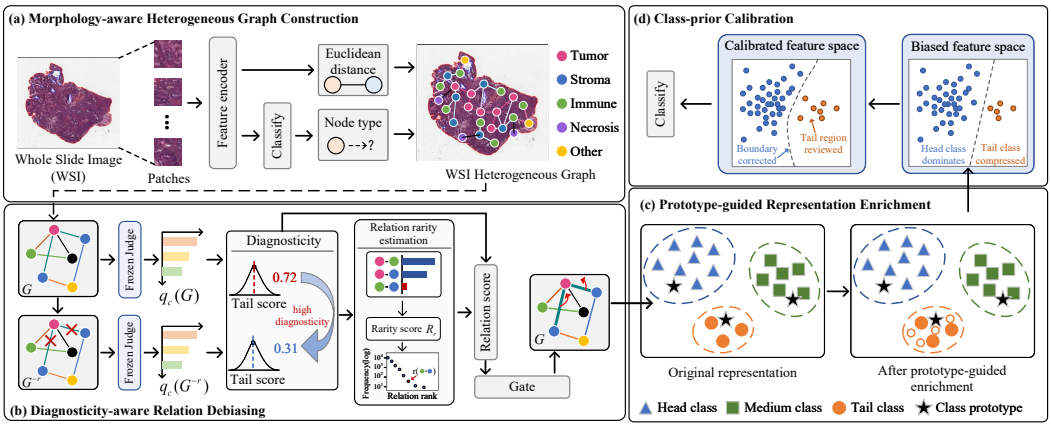


Figure 2: Overview of the proposed RLT framework. (a) Morphology-aware heterogeneous graph construction. (b) Diagnosticity-aware relation debiasing using relation rarity and counterfactual diagnosticity. (c) Prototype-guided representation enrichment. (d) Class-prior calibration for long-tailed classification.

model measures the class-specific effect of relation removal. Relation rarity and diagnostic contribution are combined into a diagnosticity-aware relation score, which modulates graph message passing. Third, graph-level class prototypes guide feature interpolation during final-model training, increasing representation coverage. Balanced Softmax is used in the final classifier to compensate for the long-tailed class prior. Counterfactual interventions and prototype augmentation are used only during training, while the frozen judge provides class posterior estimates for relation gating during both training and inference.

3.1 Problem Formulation

Let the training set be $\mathcal{D} = \{(W_i, y_i)\}_{i=1}^{M}$, where W_i is a WSI and $y_i \in \{1, \dots, C\}$ is its slide label. After tissue segmentation and tiling, slide W_i contains N_i patches. A frozen pathology encoder ϕ maps the n-th patch to a feature $\mathbf{x}_{i,n} \in \mathbb{R}^d$. Each patch is further assigned a tissue type $t_{i,n} \in \mathcal{T}$.

The class distribution is long-tailed. Let n_c denote the number of training slides in class c. We define the class imbalance ratio as

$$\mathrm{IR}_{\mathrm{cls}} = \frac{\max_c n_c}{\min_c n_c}. \tag{1}$$

Unlike MIL, we represent each WSI as a graph $G_i = (V_i, E_i)$. Each node corresponds to a patch, and each edge connects two patches according to their proximity in the pathology feature space. The edge relation type is determined by the tissue identities of its endpoints. Let N_r be the number of training edges assigned to relation r. We analogously define a relation imbalance ratio

$$\mathrm{IR}_{\mathrm{rel}} = \frac{\max_r N_r}{\min_{r:N_r>0} N_r}. \tag{2}$$

Our goal is to learn a classifier that remains sensitive to minority classes while preventing frequent tissue relations from dominating graph propagation.

3.2 Morphology-Aware Heterogeneous Graph Construction

MIL aggregation does not explicitly preserve which tissue types interact. To retain this information, we assign each patch a pathology-aware tissue category and construct a graph over patch features. In our current implementation, CONCH (Lu et al., 2024) provides a patch representation. A zero-shot tissue classifier maps patches into five categories:

$$\mathcal{T} = \{\text{tumor}, \text{stroma}, \text{immune}, \text{necrosis}, \text{other}\}. \tag{3}$$

For node v_n, the stored attributes are the feature $\mathbf{x}_n$, coordinate $\mathbf{p}_n$, and tissue type t_n. We construct a k-nearest-neighbor graph. This graph connects patches by Euclidean distance in the feature space.

For edge (u, v), we define an unordered tissue relation

$$r_{uv} = \big(\min(t_u, t_v), \max(t_u, t_v)\big). \tag{4}$$

Five tissue types therefore produce at most 15 relation types. This representation separates who is connected from what tissues are interacting, which is required for relation-level debiasing.

3.3 Diagnosticity-Aware Relation Debiasing

WSI graphs exhibit highly imbalanced relation frequencies, allowing common relations to dominate message passing. Simply up-weighting rare relations is insufficient because rarity does not necessarily imply diagnostic importance. We therefore distinguish relation rarity from diagnosticity and estimate the class-specific effect of each relation through counterfactual removal. The resulting relation score combines rarity and diagnostic contribution, selectively enhancing rare but informative interactions while avoiding indiscriminate amplification of noisy relations.

Relation rarity. We count relation occurrences on the training set only. Let $\mathcal{R}$ denote the set of tissue-relation types. For each relation $r \in \mathcal{R}$, let $N_r(G)$ denote the number of directed message-passing edges of type r in graph G, and let

$$N_r = \sum_{G \in \mathcal{D}_{\text{train}}} N_r(G) \tag{5}$$

denote its total occurrence count over the training set. Let

$$p_r = \frac{N_r}{\sum_{r' \in \mathcal{R}} N_{r'}}. \tag{6}$$

be the frequency of relation r. Its normalized rarity score is

$$R_r = \mathrm{Norm}\big(-\log(p_r + \epsilon)\big), \tag{7}$$

where $\mathrm{Norm}(\cdot)$ denotes min–max normalization over relation types.

Counterfactual relational diagnosticity. We first train a standard graph classifier f_{judge} and freeze it, where $q_c(G) = \mathrm{softmax}(f_{\text{judge}}(G))_c$ denotes its posterior probability for class c. For each training graph G, we construct G^{-r} by removing all edges of relation r while retaining the nodes and all other relations. For class c, the counterfactual contribution is

$$E_{r,c}(G) = \log q_c(G) - \log q_c(G^{-r}). \tag{8}$$

A positive value means that relation r supports class c. Let

$$\mathcal{D}_c = \{G_i \in \mathcal{D}_{\text{train}} : y_i = c\}$$

denote the set of training graphs belonging to class c. If relation r is absent from a graph G, then $G^{-r} = G$ and $E_{r,c}(G) = 0$. We compute the class-level average contribution as

$$\bar{E}_{r,c} = \frac{1}{|\mathcal{D}_c|} \sum_{G \in \mathcal{D}_c} E_{r,c}(G). \tag{9}$$

The positive diagnosticity is

$$D_{r,c} = \mathrm{Norm}\big(\max(\bar{E}_{r,c}, 0)\big). \tag{10}$$

This design explicitly avoids treating every rare relation as beneficial.

Diagnostic relation score and message gate. We combine rarity and diagnosticity as

$$S_{r,c} = \alpha R_r D_{r,c} + (1 - \alpha) D_{r,c}, \tag{11}$$

where the second term preserves common but diagnostic relations. Because the ground-truth class is unavailable at inference, we use the frozen judge posterior $q_c(G)$ to obtain a soft graph-specific relation score

$$S_r(G) = \sum_{c=1}^{C} q_c(G) S_{r,c}. \tag{12}$$

For a message $\mathbf{m}_{uv}$ carried by relation r_{uv}, we use a residual gate

$$\widetilde{\mathbf{m}}_{uv} = \big(1 + \lambda_{\text{gate}} S_{r_{uv}}(G)\big)\mathbf{m}_{uv}, \tag{13}$$

where λ_{gate} controls the strength of relation reweighting. The residual form leaves the original message path intact and selectively enhances relations supported by both rarity and diagnosticity.

3.4 Prototype-Guided Representation Richness Enhancement

Relation debiasing protects informative graph interactions, but a tail class still contains few slides. Its graph-level representation therefore covers only a narrow part of the class manifold. Repeating the same slides does not increase representation diversity. We instead enrich the training representation by interpolating each graph embedding with a prototype from the same class.

Let $\mathbf{z}_i \in \mathbb{R}^{d_g}$ be the graph representation produced after relation-debiased message passing and pooling. Using only the training split, we construct one feature prototype for each class by averaging its graph representations:

$$\mathbf{p}_c = \frac{1}{|\mathcal{D}_c|} \sum_{i:y_i=c} \mathbf{z}_i. \tag{14}$$

The augmented representation is then obtained by interpolating each graph embedding toward the prototype of its ground-truth class:

$$\widetilde{\mathbf{z}}_i = (1-\beta_i)\mathbf{z}_i + \beta_i \mathbf{p}_{y_i}. \tag{15}$$

where

$$\beta_i = \min(\hat{\beta}_i, \beta_{\max}), \qquad \hat{\beta}_i \sim \mathrm{Beta}(a, b). \tag{16}$$

Here, $\beta_{\max}$ controls the maximum interpolation strength and prevents excessive deviation from the original graph representation.

Both original and augmented embeddings are supervised. Let h be the shared classifier. We optimize

$$\mathcal{L}_{\mathrm{rep}} = \mathcal{L}_{\mathrm{cls}}(h(\mathbf{z}_i), y_i) + \lambda_{\mathrm{aug}} \mathcal{L}_{\mathrm{cls}}(h(\widetilde{\mathbf{z}}_i), y_i) + \lambda_{\mathrm{cons}} \mathcal{L}_{\mathrm{KL}}, \tag{17}$$

where $\mathcal{L}_{\mathrm{KL}}$ encourages prediction consistency between the original and interpolated representations. Prototype interpolation is used only during training.

3.5 Class-Prior Calibration and Training Objective

The previous modules improve relational representation, but severe class imbalance can still induce residual classifier bias, as reflected by the poor tail-class F1 of models without explicit class-prior calibration. We therefore adopt Balanced Softmax (Ren et al., 2020). Given class counts n_c and logits z_c, we introduce a calibration strength τ and define

$$\mathcal{L}_{\mathrm{BS}} = -\log \frac{n_y^{\tau} \exp(z_y)}{\sum_{c=1}^{C} n_c^{\tau} \exp(z_c)}. \tag{18}$$

Here, τ controls the strength of class-prior calibration. When $\tau = 0$, the loss reduces to standard cross-entropy, while larger values impose stronger correction for the long-tailed class prior. At inference, the original classifier logits are used without prototype interpolation or additional prior correction. The final training objective combines original and prototype-augmented Balanced Softmax supervision with the consistency term described above.

4 Experiments

4.1 Datasets and Evaluation Protocol

We evaluate our method on two long-tailed pathology benchmarks, Camelyon+-LT (Ling et al., 2025) and PANDA-LT (Bulten et al., 2022). Camelyon+-LT is used for lymph-node metastasis classification with four categories: Negative, Macro, Micro, and ITC. The dataset contains $871/251/174/54$ slides for the four classes, respectively. We reserve 10 slides per class for validation and 16 slides per class for testing, resulting in $845/225/148/28$ training slides. The resulting training imbalance ratio is approximately 30.2. PANDA-LT is used for prostate cancer grading according to six ISUP grades, ranging from ISUP 0 to ISUP 5. Following the long-tailed construction protocol, the validation and test sets are kept class-balanced, while the training set is downsampled to obtain an imbalance ratio of approximately 50. For both datasets, we report slide-level accuracy (ACC), AUC, F1, and Tail-F1. All final results are reported as mean $\pm$ sample standard deviation over five independent runs using a fixed data split.

Table 1: Comparison with representative methods on two long-tailed WSI benchmarks.

Method	Camelyon+-LT (IR=30)				PANDA-LT (IR=50)			
	ACC	AUC	F1	Tail-F1	ACC	AUC	F1	Tail-F1
AB-MIL	$61.35_{2.80}$	$87.50_{1.90}$	$52.10_{3.10}$	$0.00_{0.00}$	$51.10_{0.85}$	$86.60_{0.38}$	$45.70_{1.35}$	$19.50_{3.42}$
CLAM-SB	$63.50_{1.80}$	$89.20_{0.70}$	$54.20_{1.90}$	$0.00_{0.00}$	$49.30_{0.52}$	$86.65_{0.34}$	$41.50_{0.88}$	$6.50_{3.25}$
DS-MIL	$64.40_{3.90}$	$88.40_{2.50}$	$55.30_{3.80}$	$0.00_{0.00}$	$57.20_{1.25}$	$88.20_{0.38}$	$55.00_{1.55}$	$41.20_{6.20}$
TransMIL	$62.90_{3.30}$	$83.60_{2.70}$	$56.20_{4.60}$	$13.50_{10.20}$	$53.15_{1.52}$	$85.30_{1.15}$	$50.90_{2.35}$	$40.20_{8.30}$
ILRA-MIL	$64.40_{3.50}$	$88.50_{1.40}$	$56.40_{4.50}$	$7.20_{6.10}$	$57.85_{0.88}$	$88.35_{0.35}$	$56.40_{0.72}$	$45.30_{2.85}$
MHIM-MIL	$62.00_{1.80}$	$87.90_{1.20}$	$52.80_{2.20}$	$0.00_{0.00}$	$56.65_{0.80}$	$87.50_{0.35}$	$55.20_{1.42}$	$44.70_{3.35}$
AMD-MIL	$61.30_{3.00}$	$84.50_{3.50}$	$53.80_{2.80}$	$13.80_{12.80}$	$56.85_{1.12}$	$87.60_{0.45}$	$55.00_{1.58}$	$43.50_{3.30}$
WiKG-MIL	$62.90_{3.60}$	$89.10_{2.80}$	$56.60_{4.40}$	$15.10_{5.80}$	$59.05_{1.88}$	$88.40_{0.75}$	$57.60_{2.50}$	$44.50_{5.10}$
MiCo-MIL	$59.80_{1.80}$	$87.00_{1.50}$	$50.90_{1.70}$	$4.50_{6.10}$	$54.60_{2.20}$	$86.90_{0.72}$	$52.30_{2.55}$	$37.50_{5.85}$
MDE-MIL	$60.40_{2.10}$	$85.60_{2.80}$	$52.60_{3.00}$	$16.90_{5.20}$	$55.85_{1.65}$	$86.65_{1.00}$	$54.40_{2.48}$	$45.20_{6.80}$
DeCo-MIL	$68.10_{1.30}$	$90.45_{0.45}$	$64.20_{1.30}$	$35.50_{4.50}$	$59.60_{0.42}$	$87.50_{0.22}$	$55.30_{0.45}$	$44.80_{1.38}$
RLT	$\mathbf{70.63}_{4.61}$	$\mathbf{91.06}_{0.84}$	$\mathbf{70.06}_{4.42}$	$\mathbf{58.15}_{7.79}$	$\mathbf{61.00}_{1.61}$	$\mathbf{89.45}_{2.23}$	$\mathbf{60.94}_{2.44}$	$\mathbf{55.36}_{2.50}$

4.2 Implementation Details

The graph encoder uses a hidden dimension of 128, and the relation gate coefficient λ_{gate} is set to 0.5 for Camelyon+-LT and 0.7 for PANDA-LT. Balanced Softmax uses $\tau = 1.0$. For prototype-guided enrichment, one feature prototype is maintained per class, with augmentation probability 0.3 and interpolation sampled from $\text{Beta}(2, 5)$ and capped at 0.5. Models are trained for up to 50 epochs using Adam with a learning rate of 5×10^{-5}, weight decay of 1×10^{-5}, and batch size 1. Early stopping uses a patience of 10 epochs, and the best model is selected by validation AUC. Experiments are conducted on a workstation equipped with two NVIDIA GeForce RTX 5090 GPUs.

4.3 Comparison with State-of-the-Art Methods

Table 1 compares our method with representative WSI classification approaches, including conventional MIL methods such as AB-MIL (Ilse et al., 2018), CLAM-SB (Lu et al., 2021), DS-MIL (Li et al., 2021), TransMIL (Shao et al., 2021), ILRA-MIL (Xiang & Zhang, 2023), MHIM-MIL (Tang et al., 2023), and AMD-MIL (Ling et al., 2024); relation- and structure-aware methods including WiKG-MIL (Li et al., 2024) and MiCo-MIL (Li et al., 2025); and pathology-specific long-tailed methods including MDE-MIL (Ling et al., 2026) and DeCo-MIL (Li et al., 2026). These comparisons evaluate whether relational long-tail learning improves both overall classification and rare-class recognition. All baseline methods are reimplemented and evaluated under the same experimental protocol for a fair comparison. Subscripts denote standard deviations over five independent runs.

As shown in Table 1, RLT achieves the best overall performance on both benchmarks. On Camelyon+-LT, it obtains 70.63% ACC, 91.06% AUC, 70.06% F1, and 58.15% Tail-F1, outperforming DeCo-MIL by 2.53, 0.61, 5.86, and 22.65 percentage points, respectively. On the more severely imbalanced PANDA-LT benchmark, RLT also achieves the highest ACC, AUC, F1, and Tail-F1, with particularly clear gains in F1 and Tail-F1. These results indicate that explicitly modeling long-tailed tissue relations improves not only overall WSI classification but also recognition of minority classes under severe class imbalance.

4.4 Ablation of Morphology-Aware Heterogeneous Graph

We compare the proposed heterogeneous graph with a homogeneous graph under the same latent k-NN topology, patch features, backbone, and training settings. The homogeneous graph treats all edges equally, while the heterogeneous graph distinguishes tissue interactions by endpoint tissue types.

As shown in Table 2, heterogeneous modeling improves overall performance on both datasets, increasing F1 from 54.30% to 56.88% on Camelyon+-LT and from 48.50% to 51.20% on PANDA-LT. However, its Tail-F1 remains poor, indicating that tissue-aware structure alone cannot resolve long-

Table 2: Ablation study of morphology-aware heterogeneous graph modeling on Camelyon+-LT and PANDA-LT.

Method	Camelyon+-LT				PANDA-LT			
	ACC	AUC	F1	Tail-F1	ACC	AUC	F1	Tail-F1
Homogeneous Graph	$64.20_{2.85}$	$81.50_{2.10}$	$54.30_{3.15}$	$12.40_{6.20}$	$53.10_{1.95}$	$84.20_{1.10}$	$48.50_{2.30}$	$18.20_{4.80}$
Heterogeneous Graph	$65.31_{2.15}$	$83.12_{2.04}$	$56.88_{3.42}$	$0.00_{0.00}$	$55.80_{1.40}$	$86.50_{0.42}$	$51.20_{1.60}$	$0.00_{0.00}$
RLT	$\mathbf{70.63}_{4.61}$	$\mathbf{91.06}_{0.84}$	$\mathbf{70.06}_{4.42}$	$\mathbf{58.15}_{7.79}$	$\mathbf{61.00}_{1.61}$	$\mathbf{89.45}_{2.23}$	$\mathbf{60.94}_{2.44}$	$\mathbf{55.36}_{2.50}$

Table 3: Ablation study of relation-level debiasing on Camelyon+-LT and PANDA-LT.

Method	Camelyon+-LT				PANDA-LT			
	ACC	AUC	F1	Tail-F1	ACC	AUC	F1	Tail-F1
Graph + CE	$65.31_{2.15}$	$83.12_{2.04}$	$56.88_{3.42}$	$0.00_{0.00}$	$55.80_{1.40}$	$86.50_{0.42}$	$51.20_{1.60}$	$0.00_{0.00}$
+ Freq. Debias	$62.19_{2.82}$	$82.75_{2.31}$	$55.45_{4.25}$	$28.60_{12.15}$	$53.50_{1.70}$	$85.40_{0.50}$	$49.80_{1.90}$	$26.80_{4.90}$
+ Diag. Debias	$65.94_{2.45}$	$85.48_{1.88}$	$57.24_{3.86}$	$0.00_{0.00}$	$56.20_{1.25}$	$87.30_{0.38}$	$52.40_{1.50}$	$4.20_{3.10}$
+ Diag. Debias + BS	$69.69_{2.78}$	$88.84_{1.62}$	$69.52_{5.14}$	$30.40_{18.25}$	$59.40_{0.95}$	$88.90_{0.28}$	$59.10_{0.85}$	$48.50_{3.20}$
RLT	$\mathbf{70.63}_{4.61}$	$\mathbf{91.06}_{0.84}$	$\mathbf{70.06}_{4.42}$	$\mathbf{58.15}_{7.79}$	$\mathbf{61.00}_{1.61}$	$\mathbf{89.45}_{2.23}$	$\mathbf{60.94}_{2.44}$	$\mathbf{55.36}_{2.50}$

tail bias. The full RLT further improves both overall and tail-class performance by incorporating relation debiasing, prototype enrichment, and class-prior calibration.

4.5 Ablation of Relation-Level Debiasing

Table 3 compares variants of the proposed relation-learning strategy. Graph + CE is the basic heterogeneous graph model, Freq. Debias uses relation frequency only, Diag. Debias introduces diagnosticity-aware relation weighting, Diag. Debias + BS further applies Balanced Softmax, and RLT additionally includes prototype-guided representation enrichment. Frequency-only debiasing improves Tail-F1 but harms overall performance, indicating that rare relations are not necessarily informative. Diagnosticity-aware debiasing better preserves overall performance, while Balanced Softmax markedly improves tail recognition, especially on PANDA-LT. The full RLT achieves the strongest overall results and further improves Tail-F1, demonstrating the complementary effects of relation diagnosticity, class-prior calibration, and prototype-guided enrichment.

4.6 Effect of Prototype-Guided Representation Enrichment

The results in Table 4 show that prototype design has a clear impact on long-tailed representation learning. Structure-aware and tail-aware multi-prototype variants provide limited or unstable gains and may even reduce the overall F1 compared with the model without prototypes. In contrast, the proposed Feature Prototype achieves the best overall performance on both datasets. On Camelyon+-LT, it improves F1 from 69.52% to 70.06% and Tail-F1 from 30.40% to 58.15%. On PANDA-LT, F1 increases from 59.10% to 60.94% and Tail-F1 from 48.50% to 55.36%. These results indicate that a simple class-level feature prototype provides a more effective and stable reference for representation enrichment than more complex prototype designs.

4.7 Visualization and Interpretability

Figure 3(a) visualizes the learned relational evidence on tail-class slides. The tissue assignment maps (red: tumor, blue: stroma, green: immune, purple: necrosis, gray: others) reveal heterogeneous tissue distributions, while the lesion-focused contribution maps concentrate stronger diagnostic evidence around annotated lesion regions. The relation maps visualize edges associated with high-scoring relation types, showing that diagnostically relevant tissue interactions are frequently observed within lesion-associated regions and tissue interfaces. For clarity, only a subset of these edges is displayed.

Table 4: Prototype ablation on Camelyon+-LT and PANDA-LT.

Method	Camelyon+-LT ACC	Camelyon+-LT AUC	Camelyon+-LT F1	Camelyon+-LT Tail-F1	PANDA-LT ACC	PANDA-LT AUC	PANDA-LT F1	PANDA-LT Tail-F1
w/o Prototype	$69.69_{2.78}$	$88.84_{1.62}$	$69.52_{5.14}$	$30.40_{18.25}$	$59.40_{0.95}$	$88.90_{0.28}$	$59.10_{0.85}$	$48.50_{3.20}$
Structure-aware Prototype	$67.81_{3.95}$	$89.15_{1.80}$	$63.45_{6.80}$	$38.20_{25.10}$	$58.20_{1.20}$	$87.90_{0.35}$	$56.40_{1.40}$	$45.10_{4.50}$
Tail-aware Multi-Prototype	$68.44_{4.12}$	$89.48_{1.72}$	$64.18_{6.55}$	$36.50_{28.30}$	$58.60_{1.30}$	$88.20_{0.38}$	$56.90_{1.50}$	$42.30_{6.10}$
Feature Prototype	$\mathbf{70.63}_{4.61}$	$\mathbf{91.06}_{0.84}$	$\mathbf{70.06}_{4.42}$	$\mathbf{58.15}_{7.79}$	$\mathbf{61.00}_{1.61}$	$\mathbf{89.45}_{2.23}$	$\mathbf{60.94}_{2.44}$	$\mathbf{55.36}_{2.50}$

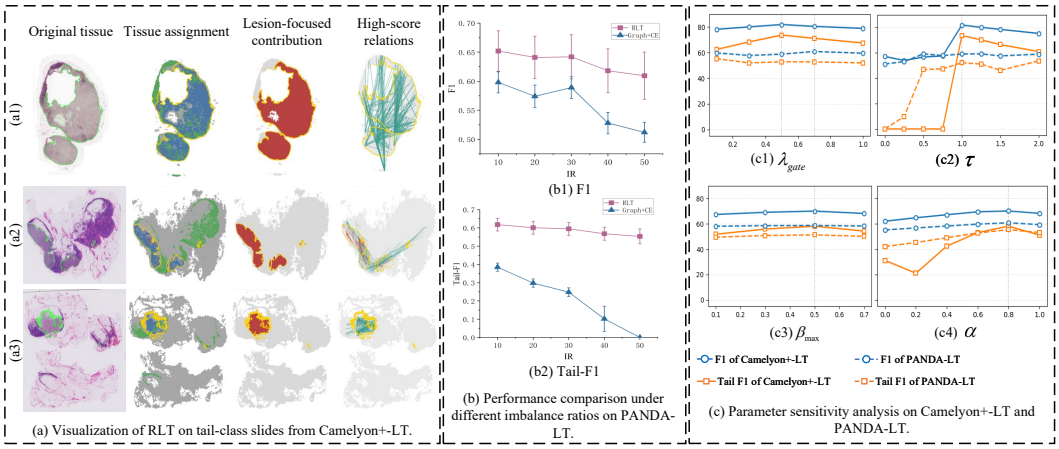


Figure 3: Analysis of the proposed RLT framework. (a) Qualitative visualization on tail-class Camelyon+-LT slides. (b) Robustness under different imbalance ratios on PANDA-LT. (c) Sensitivity analysis of λ_{gate}, τ, $\beta_{\max}$, and α on Camelyon+-LT and PANDA-LT.

4.8 Robustness to Different Imbalance Ratios

To evaluate robustness to class imbalance, we construct PANDA-LT with IR $\in \{10, 20, 30, 40, 50\}$ while keeping the validation and test sets balanced. As shown in Fig. 3(b), both methods degrade as the imbalance increases, but RLT remains more stable than Graph+CE. The difference is especially clear for Tail-F1: Graph+CE drops sharply and approaches zero at IR $= 50$, whereas RLT maintains a Tail-F1 above 50%. This demonstrates stronger robustness to severe long-tailed distributions, particularly for minority classes.

4.9 Parameter Sensitivity.

Figure 3(c) evaluates the sensitivity of λ_{gate}, τ, $\beta_{\max}$, and α. For efficiency, all parameter sensitivity experiments are conducted using one fixed seed. We vary λ_{gate} over $\{0.1, 0.3, 0.5, 0.7, 1.0\}$, τ over $\{0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 2.0\}$, $\beta_{\max}$ over $\{0.1, 0.3, 0.5, 0.7\}$, and α over $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Overall, moderate parameter values provide a better balance between Macro-F1 and Tail-F1. We therefore select the hyperparameters based on their joint performance on overall and tail-class recognition rather than optimizing either metric individually. We set $\lambda_{\text{gate}} = 0.5$ and 0.7 for Camelyon+-LT and PANDA-LT, respectively, while $\tau = 1.0$, $\beta_{\max} = 0.5$, and $\alpha = 0.8$ are used for both datasets.

5 Conclusion

We study long-tailed WSI classification from a relational perspective and propose RLT, which combines morphology-aware heterogeneous graph modeling with diagnosticity-aware relation debiasing. Prototype-guided representation enrichment and class-prior calibration further improve tail-class recognition, with experiments on Camelyon+-LT and PANDA-LT demonstrating the effectiveness and robustness of the proposed framework.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist with the drafting and organization of parts of the manuscript, and support research ideation and execution, including code-related discussion and debugging assistance. We did not use generative AI tools to generate synthetic datasets or to prove mathematical claims.

All AI-assisted content was carefully reviewed and verified by the authors. AI-assisted technical descriptions and research ideas were checked against the implemented methods, experimental results, and relevant literature. Any AI-assisted code suggestions were manually inspected and tested before being used in our experiments, and all reported numerical results were obtained from our own experimental runs. References and factual claims were also manually verified against their original sources. We take full responsibility for the final content of this work, including all text, claims, code, and experimental results produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available pathology datasets and involves no new patient recruitment or clinical intervention. All datasets are used in accordance with their original licenses, privacy requirements, and ethical conditions.

REPRODUCIBILITY STATEMENT

Reproducibility details, including WSI preprocessing, graph construction, model configurations, prototype-guided representation enrichment, frozen-judge training, dataset splits, and additional experimental settings, are provided in Appendix A. The main training objectives and inference procedure are described in Sec. 3, while the evaluation protocol and implementation settings are presented in Sec. 4. Code, configuration files, and data split files are available at the anonymous repository: `https://anonymous.4open.science/r/RLT_ICLR2027-anonymous/`.

REFERENCES

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. volume 106, pp. 249–259. Elsevier, 2018.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. volume 32, 2019.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3449–3457. IEEE, 2017.

Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 933–941, 2022.

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

Dinkar Juyal, Siddhant Shingi, Syed Ashar Javed, Harshith Padigela, Chintan Shah, Anand Sampat, Archit Khosla, John Abel, and Amaro Taylor-Weiner. Sc-mil: Supervised contrastive multiple instance learning for imbalanced classification in pathology. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7931–7940. IEEE, 2024.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. 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 *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14313–14323. IEEE, 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.

Junjian Li, Jin Liu, Hulin Kuang, Hailin Yue, Mengshen He, and Jianxin Wang. Mico: Multiple instance learning with context-aware clustering for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 376–385. Springer, 2025.

Xiaoxiao Li, Xitong Ling, Jiawen Li, Weiming Chen, Zhenyang Cai, Xidong Wang, Tian Guan, Benyou Wang, and Yonghong He. Deco-mil: Debiased counterfactual reasoning for long-tailed whole slide image analysis. *arXiv preprint arXiv:2608.14719*, 2026.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19830–19839, 2023.

Weiping Lin, Zhenfeng Zhuang, Lequan Yu, and Liansheng Wang. Boosting multiple instance learning models for whole slide image classification: A model-agnostic framework based on counterfactual inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3477–3485, 2024.

Xitong Ling, Minxi Ouyang, Yizhi Wang, Xinrui Chen, Renao Yan, Hongbo Chu, Junru Cheng, Tian Guan, Sufang Tian, Xiaoping Liu, et al. Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2795–2803, 2024.

Xitong Ling, Yuanyuan Lei, Jiawen Li, Junru Cheng, Wenting Huang, Tian Guan, Jian Guan, and Yonghong He. Comprehensive benchmark dataset for pathological lymph node metastasis in breast cancer sections. *Scientific Data*, 12(1):1381, 2025.

Xitong Ling, Xiaoxiao Li, Jiawen Li, Mingxi Fu, Yuxuan Chen, Jing Peng, Minxi Ouyang, Yizhi Wang, Yonghong He, Xiaoping Liu, et al. Multimodal distillation-driven ensemble learning for long-tailed histopathology whole slide images analysis. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2026.

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.

Ming Y Lu, Bowen Chen, Drew FK 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.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. 2020.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. volume 33, pp. 4175–4186, 2020.

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. volume 34, pp. 2136–2147, 2021.

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.

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.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. 2020.

Junxian Wu, Xinyi Ke, Xiaoming Jiang, Huanwen Wu, Youyong Kong, and Lizhi Shao. Leveraging tumor heterogeneity: Heterogeneous graph representation learning for cancer survival prediction in whole slide images. volume 37, pp. 64312–64337, 2024.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. volume 2, pp. 7, 2023.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

A Implementation and Reproducibility Details

For WSI preprocessing, each slide is divided into non-overlapping patches of size 256×256 at $20\times$ magnification. Patch features are subsequently used to construct the graph representation of each slide. For the latent k-nearest-neighbor graph, the number of neighboring nodes is set to $k = 4$. During graph construction, edges are retained as directed edges for message passing, whereas the relation identity between two connected nodes is defined as an unordered pair of their tissue types.

Each patch node is assigned to one of five tissue categories, including Tumor, Stroma, Immune, Necrosis, and Others. Accordingly, the heterogeneous graph contains 15 possible tissue-relation types: Tumor–Tumor, Tumor–Stroma, Tumor–Immune, Tumor–Necrosis, Tumor–Others, Stroma–Stroma, Stroma–Immune, Stroma–Necrosis, Stroma–Others, Immune–Immune, Immune–Necrosis, Immune–Others, Necrosis–Necrosis, Necrosis–Others, and Others–Others. Although graph edges are directed, the relation type is determined by the unordered tissue-type pair, such that edges connecting the same two tissue categories share the same relation identity regardless of direction.

The graph encoder uses a hidden dimension of 128. Its multi-level representation consists of the initial hidden representation followed by the outputs of three message-passing blocks. The four representations are concatenated and projected to obtain the graph-context features used for subsequent prototype aggregation and classification. For prototype-guided representation enrichment, we set $\lambda_{\text{aug}} = 0.5$ and $\lambda_{\text{cons}} = 0.1$ for both Camelyon+-LT and PANDA-LT.

A.1 Dataset Split Details

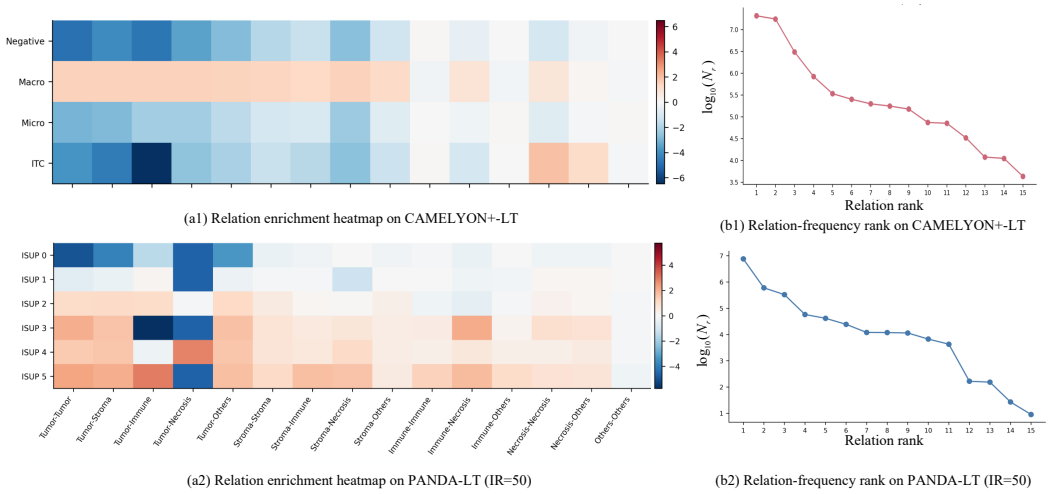
The detailed class-wise splits of Camelyon+-LT and PANDA-LT are summarized in Tables 5 and 6, respectively. For Camelyon+-LT, the validation and test subsets are class-balanced, while the remaining samples form the long-tailed training set. For PANDA-LT, the split is fixed across all experiments and is generated by deterministic class-stratified file-level sampling. The validation and test subsets are class-balanced, whereas only the training set follows the long-tailed distribution.

Table 5: Class-wise data split of Camelyon+-LT.

Split	Negative	Macro	Micro	ITC	Total
Train	845	225	148	28	1246
Validation	10	10	10	10	40
Test	16	16	16	16	64
Total	871	251	174	54	1350

Table 6: Class-wise data split of PANDA-LT.

Split	ISUP 0	ISUP 1	ISUP 2	ISUP 3	ISUP 4	ISUP 5	Total
Train	2303	1053	481	100	220	46	4203
Validation	100	100	100	100	100	100	600
Test	200	200	200	200	200	200	1200
Total	2603	1353	781	400	520	346	6003



(a1) Relation enrichment heatmap on CAMELYON+-LT

(b1) Relation-frequency rank on CAMELYON+-LT

(a2) Relation enrichment heatmap on PANDA-LT (IR=50)

(b2) Relation-frequency rank on PANDA-LT (IR=50)

Figure 4: Additional analysis of relational long tails on Camelyon+-LT and PANDA-LT. (a1–a2) Class-specific relation enrichment maps show that tissue-relation distributions vary across WSI classes. (b1–b2) Relation-frequency rank curves demonstrate highly skewed relation distributions spanning several orders of magnitude on both datasets.

A.2 Frozen Judge Implementation

The frozen judge uses the same graph encoder and classifier architecture as the final model, without relation debiasing and prototype enrichment. It is first trained as a standard graph classifier without relation debiasing, and its parameters are then fixed throughout the subsequent training stages. Given a graph G, the judge produces the class posterior $q_c(G)$, which is combined with the class-conditioned relation scores to generate graph-specific relation gates. The judge is used during training, validation, and test inference of the main model, whereas relation-removal counterfactuals and class-specific relation-score estimation are performed only on the training split.

The judge is trained for at most 50 epochs using Adam with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-5}. Early stopping is applied with a patience of 10 epochs.

B Additional Experiment

B.1 Additional Analysis of Relational Long Tails

Figure 4 provides additional evidence for the relational long-tail phenomenon on both Camelyon+-LT and PANDA-LT. The relation-frequency rank curves in Fig. 4(b1-b2) show that relation counts span several orders of magnitude, with only a few common relations dominating the graph while many relation types occur sparsely. This pattern is consistently observed on both datasets, indicating that relation imbalance is not dataset-specific.

The enrichment heatmaps in Fig. 4(a1-a2) further reveal that relation composition varies substantially across classes. For example, tumor-associated relations are relatively enriched in metastatic classes but suppressed in negative or low-tumor-burden classes, while different tail classes exhibit

Table 7: Ablation study of relation gating strategies on Camelyon+-LT and PANDA-LT.

Gating Strategy	Camelyon+-LT				PANDA-LT			
	ACC	AUC	F1	Tail-F1	ACC	AUC	F1	Tail-F1
No Gate	$66.80_{3.12}$	$87.40_{1.25}$	$64.20_{3.80}$	$42.10_{8.50}$	$57.20_{1.80}$	$87.10_{1.15}$	$55.40_{2.10}$	$46.80_{3.20}$
Additive Weighting	$68.10_{2.85}$	$88.90_{1.10}$	$66.80_{3.20}$	$49.50_{7.90}$	$58.80_{1.55}$	$88.05_{0.92}$	$57.60_{1.85}$	$50.10_{2.80}$
Residual Gate (Ours)	$\mathbf{70.63}_{4.61}$	$\mathbf{91.06}_{0.84}$	$\mathbf{70.06}_{4.42}$	$\mathbf{58.15}_{7.79}$	$\mathbf{61.00}_{1.61}$	$\mathbf{89.45}_{2.23}$	$\mathbf{60.94}_{2.44}$	$\mathbf{55.36}_{2.50}$

distinct relation profiles. These observations show that relation frequency is both highly imbalanced and class dependent. Therefore, global rarity alone is insufficient to determine relation importance, which further motivates the use of class-specific diagnosticity in RLT.

B.2 Effect of Relation Gating Strategy

To evaluate how relation scores should be incorporated into graph propagation, we compare no gating, additive weighting, and the proposed residual gate under the same full RLT configuration. As shown in Table 7, relation-aware gating consistently improves performance over the ungated variant, while the residual formulation achieves the best results on both datasets. Compared with additive weighting, the residual gate further improves both overall F1 and Tail-F1, indicating that preserving the original message path while selectively enhancing informative relations provides more stable and effective relation modulation. These results validate the proposed gating design for injecting diagnostic relation scores into heterogeneous graph message passing.

B.3 Additional Visualizations

Figure 5 provides additional qualitative examples of the learned tissue and relation patterns. The tissue assignment maps show that lesion regions are not composed exclusively of tumor patches. In several cases, a large proportion of lesion-associated patches are classified as stroma. This is expected because the lesion annotation denotes the pathological lesion region rather than a pure tumor-cell mask, while tissue assignment is performed at the patch level. A lesion region may therefore contain substantial stromal, immune, necrotic, or mixed tissue components, especially near tissue interfaces.

The lesion-focused contribution maps further show that stronger model evidence is generally concentrated around lesion-associated regions rather than being uniformly distributed across the slide. This suggests that the learned representation captures spatially meaningful diagnostic evidence even when the local tissue composition is heterogeneous. For relation visualization, the learned score is defined at the relation-type level, such that edges belonging to the same relation type share the same score. We therefore visualize high-scoring relation types and sample only 5% of their associated edges to reduce visual clutter. These displayed edges are frequently observed within lesion-associated regions or along heterogeneous tissue interfaces, although some extend beyond the annotated lesion. This is consistent with the latent graph construction, where morphologically similar patches can be connected even when they are spatially distant. Therefore, the relation maps should be interpreted as relation-level diagnostic evidence rather than direct lesion localization.

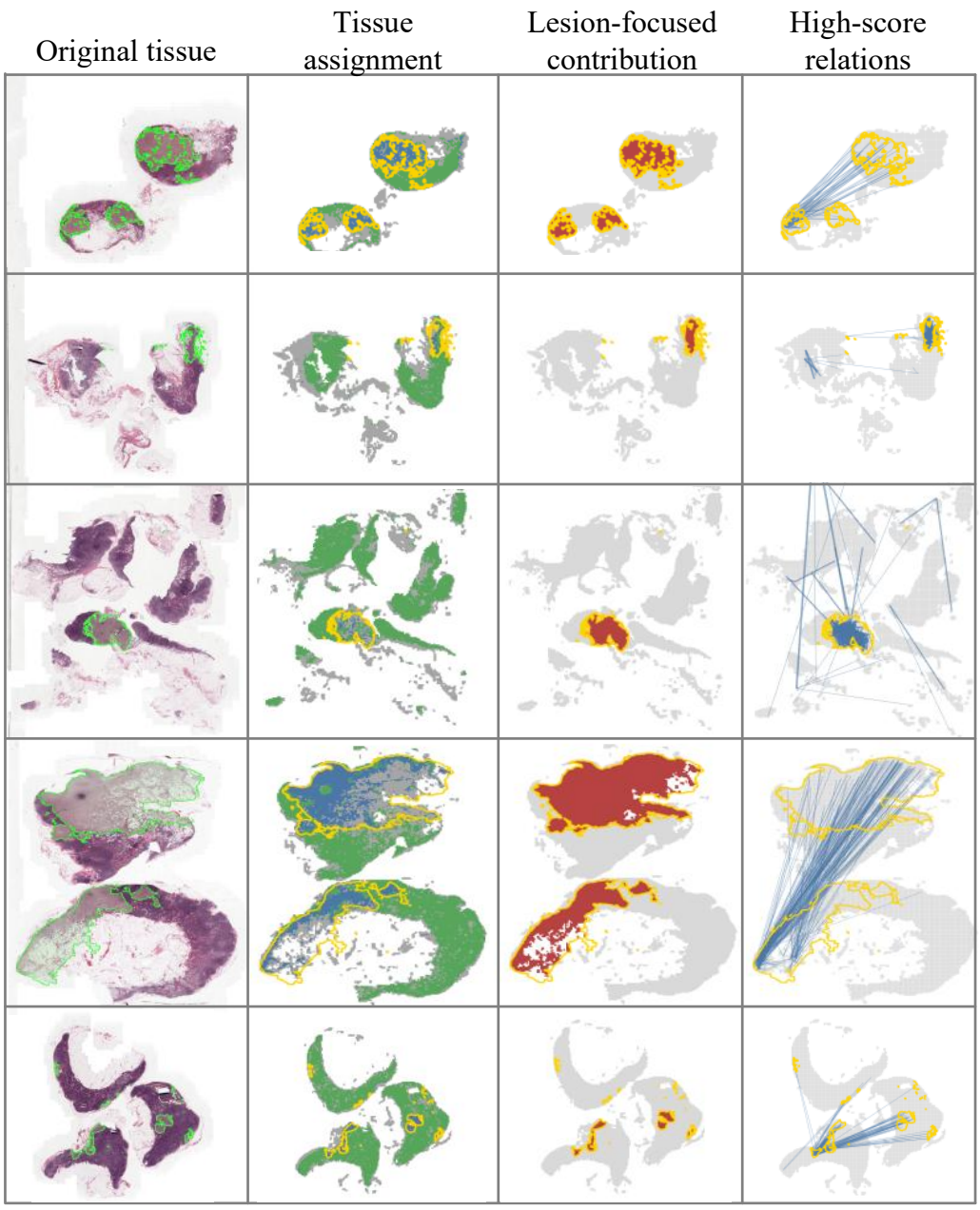


Figure 5: Additional visualization of tissue assignments, lesion-associated evidence, and diagnostic relations. From left to right, each row shows the original WSI with lesion annotation, patch-level tissue assignment, lesion-focused contribution, and edges associated with high-scoring relation types. In the tissue assignment maps, red, blue, green, purple, and gray denote tumor, stroma, immune, necrosis, and other tissues, respectively. Since the learned score is defined at the relation-type level, only 5% of the edges associated with the visualized relation types are sampled and displayed for clarity.