# MM-HAGT: HIERARCHICAL GRAPH TRANSFORMERS CONDITIONED ON INCOMPLETE CLINICAL DATA FOR WHOLE-SLIDE IMAGES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Pathological staging from gigapixel WSIs requires modeling spatially organized tissue patterns together with complementary patient context. Existing WSI models typically ignore structured clinical evidence, whereas conventional multimodal fusion is vulnerable to incomplete clinical records. We introduce MM-HAGT, a hierarchical graph Transformer that accounts for missing information and incorporates clinical context. MM-HAGT represents each WSI as a graph combining topological and semantic relationships, adaptively reweights uninformative patches, and hierarchically aggregates patch features into representations at the region and slide levels. A missing aware clinical encoder transforms variables with a fixed schema and their observation mask into a token controlled by a reliability gate, which interpolates between observed clinical evidence and a learned prior for missing information. Self attention subsequently enables the clinical token to condition localized WSI graph representations. Clinical modality dropout and auxiliary supervision further promote robustness to incomplete records and discourage excessive reliance on structured variables. This framework provides a unified approach to hierarchical WSI representation learning with partially observed clinical context for pathological staging at the patient level.

## 1 INTRODUCTION

Whole-slide images (WSIs) provide detailed morphological evidence for cancer diagnosis and pathological staging, but their gigapixel resolution makes direct processing impractical (Hijazi et al., 2024; El Nahhas et al., 2025). Most existing approaches therefore divide a WSI into patches and aggregate their features using multiple-instance learning (MIL) (Shi et al., 2024). Treating patches as unordered instances, however, can obscure tissue topology, retain many weakly informative regions, and fail to capture the progression from local morphology to region- and slide-level semantics (Ma et al., 2024; Jin et al., 2024).

Graphs provide a natural representation of tissue organization by treating patches as nodes and their spatial or semantic relationships as edges (Wang et al., 2024; Chan et al., 2023b). Combining graph neural networks with Transformers further supports local message passing and long-range interaction (Li et al., 2024; Yu et al., 2023; Chen et al., 2022). Nevertheless, fixed graph structures may not reflect sample-specific tissue relationships, while global attention over thousands of patch nodes remains computationally expensive (Li et al., 2024; Wang et al., 2022; An et al., 2024). These limitations call for adaptive graph construction and hierarchical aggregation before global reasoning.

WSI morphology also provides only part of the evidence required for patient-level staging. Demographic attributes, exposure history, and routine laboratory findings offer complementary context, but clinical records frequently contain mixed variable types and missing fields. Conventional imputation can obscure whether a value was actually observed, whereas late fusion limits interactions between clinical factors and localized pathological patterns. Although multimodal medical learning has shown the value of combining imaging with structured patient information (Jing et al., 2025), conditioning hierarchical WSI representations on incomplete clinical evidence remains underexplored.



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.

To address these challenges, we propose MM-HAGT, a hierarchical graph Transformer conditioned on incomplete clinical data. Dynamic semantic selection suppresses uninformative patches, while joint topological and semantic edge weighting adapts graph relationships to each WSI. Multi-query hierarchical attention then aggregates patch representations into region- and slide-level features, which are integrated through gated cross-level fusion. In parallel, a missing-aware clinical encoder combines fixed-schema variables with their observation mask to construct a reliability-gated clinical token. This token interpolates between observed clinical evidence and a learnable missing-information prior and interacts with localized WSI graph tokens through self-attention. Clinical modality dropout and auxiliary supervision further encourage stable prediction under incomplete clinical observations.

The main contributions are:

(1) An adaptive WSI graph encoder that combines dynamic semantic selection with topological and semantic edge weighting to suppress redundant patches and refine tissue relationships.

(2) A multi-query hierarchical attention mechanism with diversity regularization and gated cross-level fusion for aggregating complementary pathology patterns across patch, region, and slide levels.

(3) A missing-aware clinical conditioning mechanism that uses observation masks, reliability gating, and a learnable missing-information before integrate incomplete clinical evidence with localized WSI representations.

## 2 Related Works

### 2.1 Multiple-Instance Learning for WSIs

Multiple-instance learning (MIL) represents a WSI as a bag of patch instances and has become a standard framework for weakly supervised pathological image analysis (Zheng et al., 2024; Shi et al., 2025; Wu et al., 2025). Recent studies have improved MIL through more selective or robust aggregation. HDMIL filters irrelevant patches using attention-guided hierarchical distillation (Dong et al., 2025a), IBMIL reduces bias induced by bag-level contextual priors through interventional learning (Lin et al., 2023), and CS-MIL models relationships between pathological features at different scales (Deng et al., 2024). Although these methods improve patch selection, debiasing, or multiscale aggregation, they do not jointly model sample-specific tissue graphs, hierarchical pathological structure, and complementary patient context.

### 2.2 Graph-Based WSI Modeling

Graph neural networks provide a structured representation of WSIs by encoding patches as nodes and tissue relationships as edges (Ding et al., 2023). Graph Transformers further combine local graph propagation with global attention for relational representation learning (Wu et al., 2022; Chen et al., 2025). In computational pathology, GTP integrates graph representations with vision Transformers for WSI classification and interpretation (Zheng et al., 2022), while HIGT models interactions between tissue representations across multiple image resolutions (Guo et al., 2023). These approaches demonstrate the benefit of combining tissue topology with long-range dependencies, but they generally rely on predefined graph relationships or WSI morphology alone. Adaptive suppression of redundant patches, refinement of topological and semantic relationships, and clinical conditioning under a unified hierarchy remain insufficiently explored.

### 2.3 Multimodal Learning with Incomplete Clinical Data

Multimodal graph learning has been used to capture relationships between modalities in a shared representation space (D'Souza et al., 2024), while self-supervised hypergraph learning has been explored for aligning multiple medical imaging modalities (Jing et al., 2025). These methods establish the value of cross-modal interaction but primarily focus on multiplex graph data or alignment between imaging modalities. The combination of gigapixel WSI graphs and structured clinical records introduces a different challenge because clinical variables contain mixed data types and patient-specific missing fields. Conventional imputation and late fusion model data availability

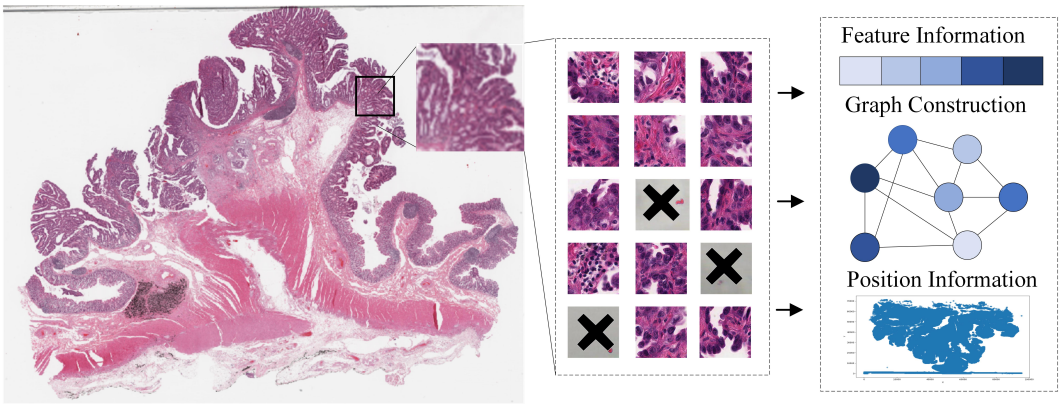



Figure 1: Construction of the adaptive WSI graph. Patch features and spatial coordinates form the initial graph representation. Dynamic semantic selection suppresses weakly informative patches, and topology–semantic edge weighting combines spatial proximity with feature similarity to refine graph connectivity.

only implicitly and provide limited interaction between clinical factors and localized tissue patterns. MM-HAGT instead encodes clinical observations and their missingness into a reliability-controlled token, which directly conditions hierarchical WSI graph representations through self-attention.

# 3 Methodology

## 3.1 Overview

Given a whole-slide image (WSI) and a partially observed clinical record, MM-HAGT constructs a patient-conditioned representation through four stages. As illustrated in Fig. 1, the WSI is first partitioned into patches, from which semantic features and spatial coordinates are extracted to construct the initial graph. A learnable semantic gate suppresses weakly informative patches, after which topology–semantic edge weighting combines spatial proximity with feature similarity to refine the relationships between retained nodes.

The complete architecture is shown in Fig. 2. The refined patch representations are processed by an edge-weighted graph encoder and progressively aggregated into region- and slide-level representations using multi-query hierarchical attention. These representations are combined through gated cross-level fusion. In parallel, a missing-aware clinical encoder transforms the partially observed clinical record and its observation mask into a reliability-gated token. The clinical token then interacts with localized pathology tokens through self-attention to produce the final stage prediction. By applying global reasoning to the resulting hierarchical token set, MM-HAGT retains local tissue information while reducing the computational cost of attention over all WSI patches.

## 3.2 Adaptive WSI Graph Encoding

Each WSI is partitioned into $256 \times 256$ tissue patches. A pretrained feature extractor provides a descriptor $\mathbf{x}_n \in \mathbb{R}^{d_x}$ for patch $n$, and its normalized slide coordinate is denoted by $\mathbf{p}_n \in \mathbb{R}^2$. Patches form the nodes of a radius-neighborhood graph $G = (V, E)$.

**Dynamic semantic selection.** For each patch, we collect its background proportion $b_n$, normalized texture entropy $e_n$, and semantic confidence $q_n$ in $\mathbf{m}_n = [1 - b_n, e_n, q_n]$. A lightweight two-branch network combines this metadata with the patch descriptor to obtain a soft retention gate:

$$\begin{aligned} s_n &= \psi(\phi_x(\mathbf{x}_n) \,\|\, \phi_m(\mathbf{m}_n)), \\ a_n &= \mathrm{sigmoid}((s_n - \tau)/T), \qquad \widehat{\mathbf{x}}_n = a_n \mathbf{x}_n, \end{aligned} \tag{1}$$

where $\tau$ and the parameters of $\phi_x$, $\phi_m$, and $\psi$ are learned, and $T$ controls gate sharpness. Unlike hard top-$k$ pruning, this differentiable gate retains weak evidence when it is useful while reducing

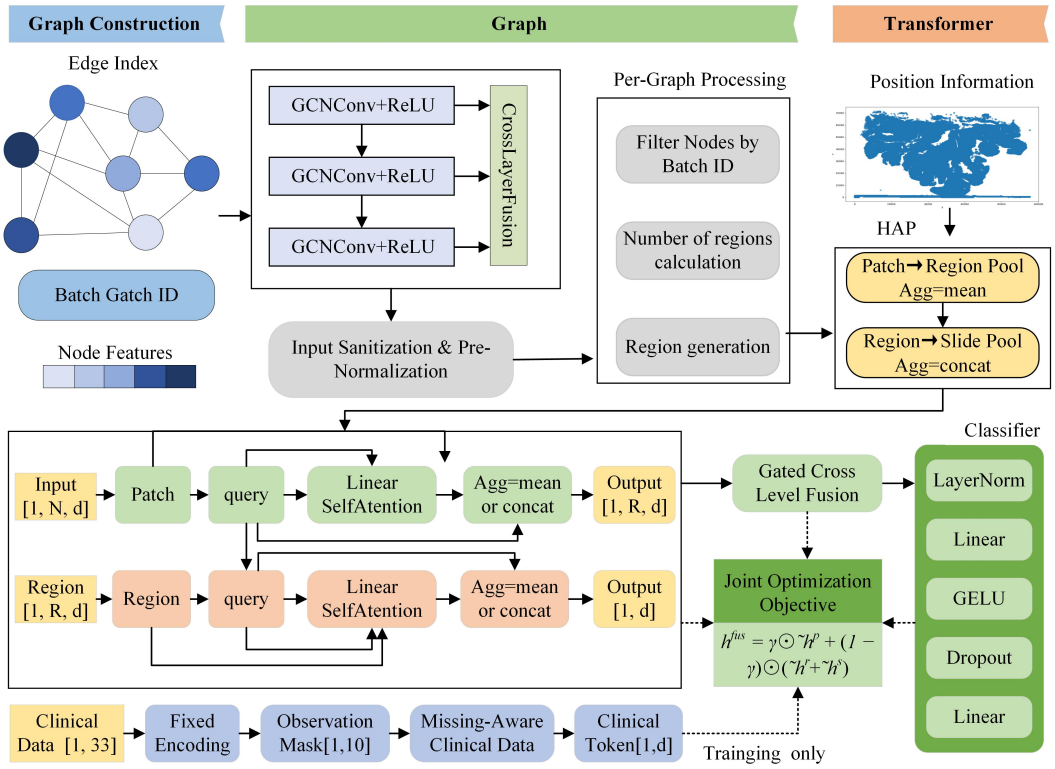



Figure 2: Overview of MM-HAGT. The adaptive graph encoder refines patch nodes and edges, hierarchical query attention produces region- and slide-level pathology representations, and a reliability-gated clinical token conditions the pathology tokens through self-attention for stage prediction.

the influence of background and redundant tissue. We regularize the average retained mass with $\mathcal{L}_{\text{sel}} = |V|^{-1} \sum_n a_n$.

**Topology–semantic edge weighting.** For each existing spatial edge $(u, v) \in E$, MM-HAGT combines a Gaussian spatial kernel with cosine feature similarity:

$$w_{uv} = \alpha \exp\left(-\frac{\|\mathbf{p}_u - \mathbf{p}_v\|_2^2}{2\sigma_p^2 + \epsilon}\right) + (1-\alpha)\frac{1+\cos(\mathbf{x}_u, \mathbf{x}_v)}{2}, \qquad \alpha = \text{sigmoid}(\theta_\alpha). \tag{2}$$

Thus, the spatial graph topology is preserved, but the contribution of each edge adapts to both tissue geometry and appearance. An edge-weighted GCN then produces contextualized patch embeddings $\mathbf{U} = \text{GCN}(\widehat{\mathbf{X}}, E, \mathbf{W}) \in \mathbb{R}^{N \times d}$.

### 3.3 Hierarchical Multi-Query Aggregation

Global attention over all WSI patches is costly and can dilute localized signals. We therefore partition the graph into $R$ spatially coherent regions and aggregate it in two stages. For a learnable query bank $\mathbf{Q}$ and token matrix $\mathbf{H}$, we use

$$\text{MQAttn}(\mathbf{Q}, \mathbf{H}) = \text{softmax}\left(\frac{(\mathbf{Q}\mathbf{W}_Q)(\mathbf{H}\mathbf{W}_K)^\top}{\sqrt{d}}\right)\mathbf{H}\mathbf{W}_V. \tag{3}$$

Let $\mathbf{U}_{\mathcal{R}_r}$ contain the patch embeddings in region $r$. A shared bank of $K$ region queries yields $\mathbf{Z}_r = \text{MQAttn}(\mathbf{Q}^{\text{reg}}, \mathbf{U}_{\mathcal{R}_r})$. The resulting region summaries are passed to $M$ slide queries:

$$\mathbf{h}_r = \text{mean}(\mathbf{Z}_r), \qquad \mathbf{S} = \text{MQAttn}(\mathbf{Q}^s, [\mathbf{h}_1; \ldots; \mathbf{h}_R]). \tag{4}$$

Multiple queries allow each level to capture complementary pathological patterns. To discourage query collapse, we apply

$$\mathcal{L}_{\text{div}} = \frac{1}{R}\sum_{r=1}^{R}\|\overline{\mathbf{A}}_r\overline{\mathbf{A}}_r^{\top} - \mathbf{I}\|_F^2 + \|\overline{\mathbf{B}}\overline{\mathbf{B}}^{\top} - \mathbf{I}\|_F^2, \tag{5}$$

where $\mathbf{A}_r$ and $\mathbf{B}$ are the region- and slide-level attention matrices from Eq. 3, and the bars denote row-wise $\ell_2$ normalization.

**Gated cross-level fusion.** We summarize the three hierarchy levels as $\mathbf{h}^p = \sum_n a_n \mathbf{u}_n / (\sum_n a_n + \epsilon)$, $\mathbf{h}^r = R^{-1}\sum_r \mathbf{h}_r$, and $\mathbf{h}^s = \text{mean}(\mathbf{S})$. After projecting them into a common space, a sample-dependent gate combines fine- and coarse-grained evidence:

$$\begin{aligned} \bar{\mathbf{h}}^{\ell} &= \mathbf{W}_{\ell}\mathbf{h}^{\ell}, \quad \ell \in \{p, r, s\}, \\ \boldsymbol{\gamma} &= \text{sigmoid}\big(\mathbf{W}_g[\bar{\mathbf{h}}^p \parallel \bar{\mathbf{h}}^r \parallel \bar{\mathbf{h}}^s] + \mathbf{b}_g\big), \\ \mathbf{h}^w &= \text{MLP}_f\big(\boldsymbol{\gamma} \odot \bar{\mathbf{h}}^p + (1-\boldsymbol{\gamma}) \odot (\bar{\mathbf{h}}^r + \bar{\mathbf{h}}^s)\big). \end{aligned} \tag{6}$$

### 3.4 Missing-Aware Clinical Conditioning

**Fixed-schema clinical input.** We use ten patient variables: age, sex, tumor laterality, prior malignancy, smoking status, pack-years, and the initial-diagnosis calcium, hemoglobin, leukocyte, and platelet results. Pack-years are log-transformed, and continuous variables are standardized using training-fold statistics only. Categorical fields use a fixed one-hot schema with an explicit missing category. Unobserved continuous values are set to zero after standardization, with missingness preserved by indicators. This produces a 33-dimensional vector $\mathbf{x}_i^c$ and a field-level observation mask $\mathbf{m}_i^c \in \{0,1\}^{10}$. Clinical records, stage labels, and WSIs are linked by patient identifier, but identifiers are never supplied to the model; all slides from a patient remain in the same data split.

**Reliability-gated clinical token.** A two-layer MLP first maps the fixed-schema vector into a candidate embedding. A channel-wise gate uses both the values and their observation mask and is scaled by record coverage:

$$\begin{aligned} \mathbf{h}_i^c &= \text{MLP}_c(\text{LN}(\mathbf{x}_i^c)), & \rho_i &= \frac{1}{10}\sum_{j=1}^{10} m_{ij}^c, \\ \mathbf{r}_i^c &= \rho_i \,\text{sigmoid}(\mathbf{W}_r[\text{LN}(\mathbf{x}_i^c) \parallel \mathbf{m}_i^c] + \mathbf{b}_r), & \mathbf{z}_i^c &= \mathbf{r}_i^c \odot \mathbf{h}_i^c + (1-\mathbf{r}_i^c) \odot \mathbf{e}_{\text{miss}}^c. \end{aligned} \tag{7}$$

Here, $\mathbf{e}_{\text{miss}}^c$ is a learned prior used when clinical evidence is incomplete. During training, clinical modality dropout replaces the entire clinical token with this prior:

$$\widetilde{\mathbf{z}}_i^c = \delta_i \mathbf{z}_i^c + (1-\delta_i)\mathbf{e}_{\text{miss}}^c, \qquad \delta_i \sim \text{Bernoulli}(1-p_c). \tag{8}$$

This exposes the classifier to both complete and unavailable clinical context and limits over-reliance on structured variables.

### 3.5 Multimodal Transformer and Training Objective

The region-query outputs remain localized, whereas $\mathbf{S}$ and $\mathbf{h}^w$ encode progressively broader context. We retain these complementary tokens as $\mathbf{U}_i^w = [\mathbf{Z}_{i1}; \ldots; \mathbf{Z}_{iR}; \mathbf{S}_i; \mathbf{h}_i^w]$ and prepend the clinical and classification tokens:

$$\mathbf{T}_i^0 = [\mathbf{e}_{\text{CLS}}; \widetilde{\mathbf{z}}_i^c; \mathbf{U}_i^w], \qquad \mathbf{T}_i^L = \mathcal{T}_\theta(\mathbf{T}_i^0), \qquad \widehat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}_o\mathbf{T}_{i,\text{CLS}}^L + \mathbf{b}_o). \tag{9}$$

Self-attention lets the clinical token modulate individual region representations rather than interacting only with a single global WSI vector.

The final objective retains only losses tied directly to the proposed components:

$$\mathcal{L} = \mathcal{L}_{\text{stage}} + \lambda_{\text{sel}}\mathcal{L}_{\text{sel}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_c\mathcal{L}_{\text{clin}}, \qquad \mathcal{L}_{\text{clin}} = \text{CE}(\mathbf{W}_c\mathbf{z}_i^c + \mathbf{b}_c, y_i), \tag{10}$$

where $\mathcal{L}_{\text{stage}}$ denotes the class-weighted cross-entropy loss for multimodal stage prediction. The clinical auxiliary head is used only during training and is removed at inference. Implementation details, including the architecture dimensions, loss weights, region construction procedure, and optimization settings, are provided in the experimental setup.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets.** We evaluate MM-HAGT on two cohorts from The Cancer Genome Atlas (TCGA). TCGA-COAD contains 461 WSIs from patients with colon adenocarcinoma, whereas TCGA-KICH contains 113 WSIs from patients with kidney chromophobe carcinoma. For each cohort, the model uses the WSIs, pathological-stage labels, and structured clinical variables described in Section 3; genomic measurements are not used as model inputs.

**Evaluation protocol.** We use stratified five-fold cross-validation with an internal validation split. All WSIs from the same patient are assigned to a single fold to prevent patient-level leakage. In each run, the validation split is used for checkpoint selection, and the held-out fold is used once for evaluation. We report the mean and standard deviation across the five test folds for accuracy, F1 score, and AUC.

**Implementation details.** Models are trained for 50 epochs with a batch size of 16. MM-HAGT uses a hidden dimension of 128, three graph-convolution layers, and a one-layer Transformer with four attention heads. We optimize the cross-entropy objective using Adam with a learning rate of $10^{-4}$ and weight decay of $5 \times 10^{-5}$. Checkpoint selection and hyperparameter tuning use validation data only.

## 4.2 COMPARISON WITH EXISTING METHODS

We compare MM-HAGT with SSM, DTFD-MIL, MambaMIL, HDMIL, HEAT, and CLAM. Tables 1 and 2 report the fold-level mean and standard deviation; all entries are percentages.

Table 1: Comparison of Model Performance on Accuracy (Acc), F1 Score, and AUC (%) with Mean ± Standard Deviation on the TCGA-COAD Dataset

| Method | Acc (%) | F1 (%) | AUC (%) |
|---|---|---|---|
| SSM(Fillioux et al., 2023) | 68.38 ± 6.8 | 67.26 ± 5.9 | 67.55 ± 7.1 |
| DTFD-MIL(Zhang et al., 2022) | 64.45 ± 6.5 | 64.56 ± 6.8 | 67.26 ± 6.7 |
| MambaMIL(Yang et al., 2024) | 67.35 ± 7.6 | 63.26 ± 7.4 | 68.42 ± 6.2 |
| HDMIL(Dong et al., 2025b) | 70.56 ± 6.6 | 66.34 ± 7.2 | 70.75 ± 5.4 |
| HEAT(Chan et al., 2023a) | 65.06 ± 4.3 | 64.42 ± 5.7 | 68.44 ± 6.5 |
| CLAM(Lu et al., 2021) | 69.46 ± 6.4 | 69.86 ± 6.1 | 69.48 ± 6.3 |
| **MM-HAGT** | **73.19 ± 5.1** | **71.37 ± 5.8** | **82.28 ± 0.5** |

On TCGA-COAD (Table 1), the highest baseline results are 70.56% accuracy and 70.75% AUC from HDMIL and 69.86% F1 from CLAM. MM-HAGT reports 73.19% accuracy, 71.37% F1, and 82.28% AUC, corresponding to differences of 2.63, 1.51, and 11.53 percentage points, respectively. Its reported AUC standard deviation is 0.5 percentage points, compared with 5.4–7.1 for the baselines.

Table 2: Comparison of Model Performance on Accuracy (Acc), F1 Score, and AUC (%) with Mean ± Standard Deviation on the TCGA-KICH Dataset

| Method | Acc (%) | F1 (%) | AUC (%) |
|---|---|---|---|
| SSM(Fillioux et al., 2023) | 68.48 ± 7.8 | 68.32 ± 6.4 | 68.95 ± 7.2 |
| DTFD-MIL(Zhang et al., 2022) | 64.35 ± 7.5 | 64.93 ± 5.8 | 64.46 ± 7.1 |
| MambaMIL(Yang et al., 2024) | 64.56 ± 7.7 | 64.28 ± 6.9 | 66.45 ± 4.9 |
| HDMIL(Dong et al., 2025b) | 69.55 ± 6.8 | 65.42 ± 7.7 | 66.65 ± 6.4 |
| HEAT(Chan et al., 2023a) | 66.00 ± 5.6 | 66.40 ± 6.7 | 70.40 ± 5.5 |
| CLAM(Lu et al., 2021) | 68.84 ± 6.6 | 68.56 ± 6.5 | 68.20 ± 6.7 |
| **MM-HAGT** | **72.73 ± 3.2** | **71.88 ± 3.3** | **81.59 ± 4.6** |

On TCGA-KICH (Table 2), the strongest baseline values are 69.55% accuracy from HDMIL, 68.56% F1 from CLAM, and 70.40% AUC from HEAT. MM-HAGT reports 72.73% accuracy, 71.88% F1, and 81.59% AUC; the corresponding differences are 3.18, 3.32, and 11.19 percentage points. The standard deviations of its three metrics are 3.2, 3.3, and 4.6 percentage points. These comparisons are descriptive because statistical significance tests are not reported.

### 4.3 Ablation Study

We remove Hierarchical Attention Pooling (HAP), Gated Cross-Level Fusion (GCLF), and Graph-Aware Positional Encoding (GAPE) individually while retaining the remaining training protocol.

Table 3: Ablation study results evaluating the impact of key modules on Accuracy (Acc), F1 Score, and AUC (%) with Mean ± Standard Deviation on the TCGA-COAD Dataset

| Model Variant | Acc | F1 Score | AUC |
|---|---|---|---|
| w/o HAP | 70.36 ± 5.4 | 66.26 ± 5.9 | 80.50 ± 5.1 |
| w/o GCLF | 71.95 ± 5.5 | 69.96 ± 5.8 | 80.65 ± 5.6 |
| w/o GAPE | 72.25 ± 5.6 | 71.36 ± 5.4 | 81.35 ± 1.2 |
| Full Model (MM-HAGT) | **73.19 ± 5.1** | **71.37 ± 5.8** | **82.28 ± 0.5** |

Relative to the full model in Table 3, removing HAP changes accuracy, F1, and AUC by $-2.83$, $-5.11$, and $-1.78$ percentage points. Removing GCLF changes the same metrics by $-1.24$, $-1.41$, and $-1.63$ points, whereas removing GAPE changes them by $-0.94$, $-0.01$, and $-0.93$ points. Among the reported variants, the largest reductions in mean accuracy and F1 occur when HAP is removed. The table isolates one component at a time and does not by itself identify the mechanism responsible for each difference.

### 4.4 Qualitative Visualization and Optimization Curves

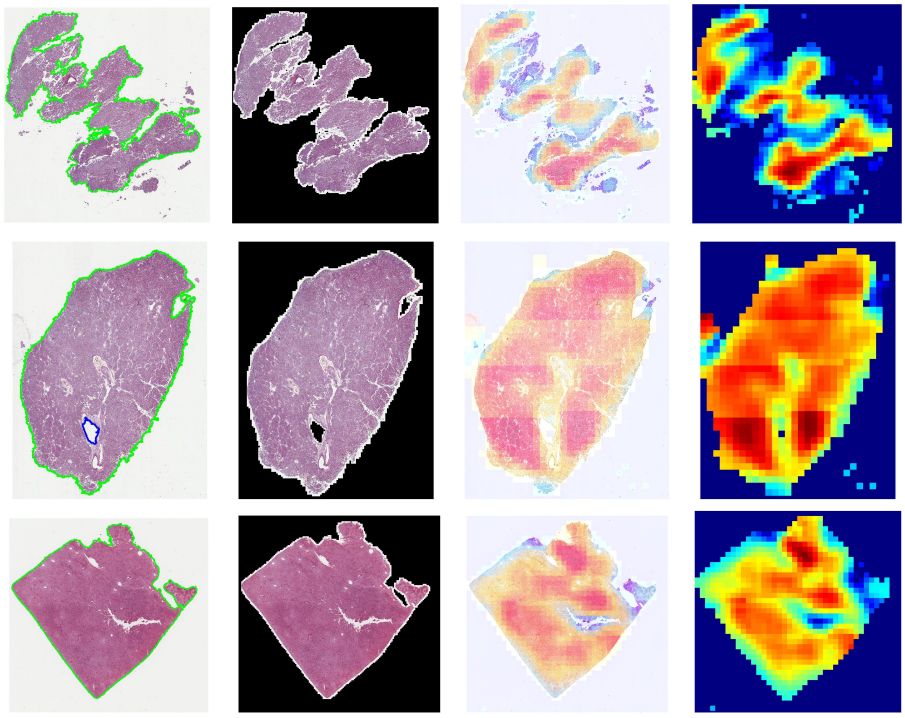



Figure 3: Visualization of whole-slide images and corresponding attention heatmaps.

Figure 3 displays the original WSIs, binary tissue masks, attention overlays, and standalone attention maps. Warmer colors denote higher attention weights and cooler colors denote lower weights.

The heatmaps provide a qualitative view of the regions contributing to the model representation. They do not establish agreement with pathologist-defined regions because no quantitative overlap measure or blinded reader assessment is reported.

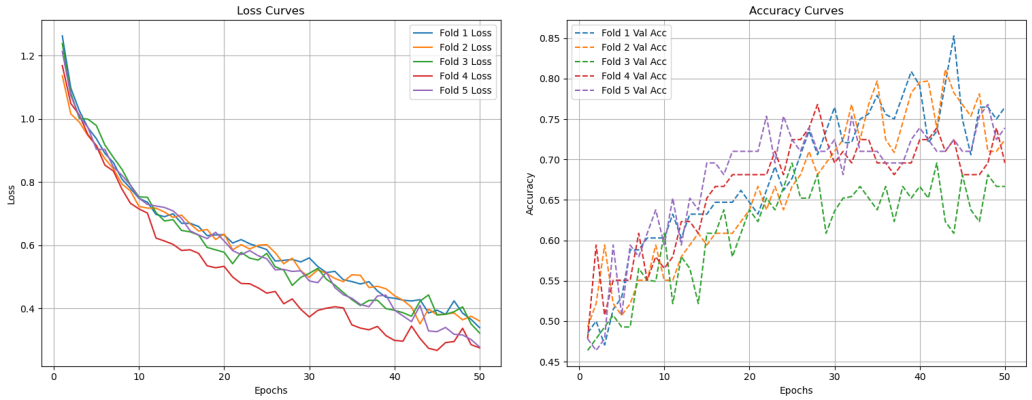



Figure 4: Loss Curve and Accuracy Curve on the TCGA-COAD Dataset.

Figure 4 shows the training loss and validation accuracy for the five TCGA-COAD folds over 50 epochs. Training loss decreases across the folds, while validation accuracy increases and subsequently approaches a plateau. Differences among the fold trajectories indicate sensitivity to the data partition. These curves describe optimization and checkpoint selection; test-set performance is reported separately in Table 1.

Figure 5 reports the corresponding trajectories for the TCGA-KICH kidney chromophobe cohort.

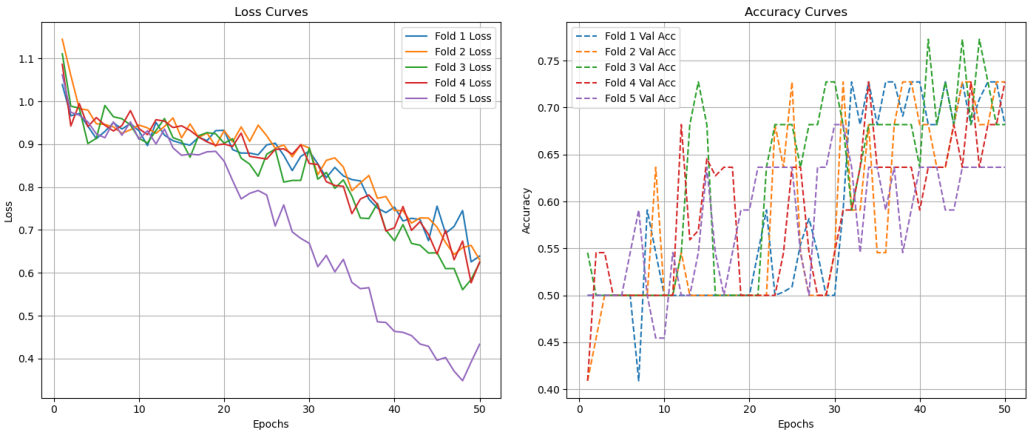
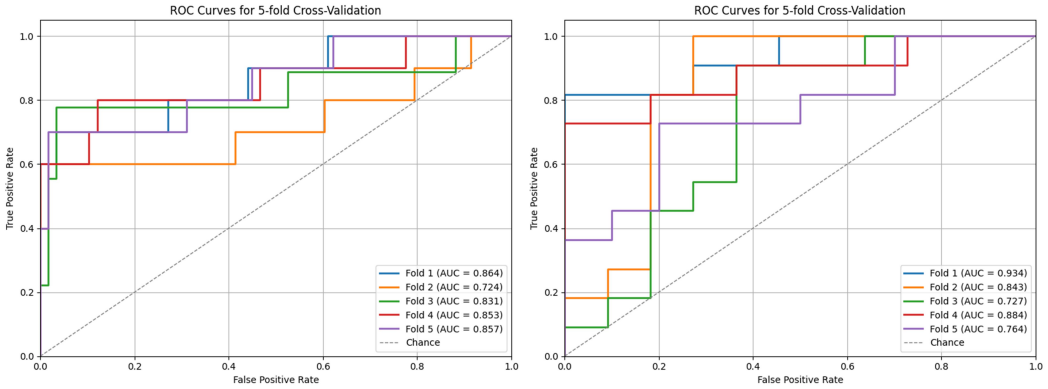



Figure 5: Loss Curves and Accuracy Curves on the TCGA-KICH Dataset.

Across the five folds, training loss decreases from approximately 1.1–1.15 to 0.6–0.7 by epoch 50, and validation accuracy rises from approximately 0.50 to 0.70–0.75 before plateauing in later epochs. As with Fig. 4, these validation trajectories are not used as substitutes for held-out test metrics.

Figure 6 presents the ROC curves, which summarize the trade-off between true-positive and false-positive rates across decision thresholds. The corresponding fold-averaged AUC values are reported in Tables 1 and 2; the ROC curves are not used to make claims about probability calibration.




(A) (B)

Figure 6: ROC Curves.

## 5 Conclusion

We introduced MM-HAGT, a hierarchical graph Transformer for pathological staging from WSIs and partially observed clinical data. MM-HAGT combines dynamic patch selection and topology–semantic edge weighting with multi-query hierarchical aggregation and gated cross-level fusion. A missing-aware clinical encoder further uses observation masks and reliability gating to condition localized WSI representations while reducing dependence on incomplete clinical records. On TCGA-COAD and TCGA-KICH, MM-HAGT achieved the highest mean accuracy, F1 score, and AUC among the evaluated methods. The ablation results quantified the contributions of the examined components, while the attention maps provided a qualitative view of the regions emphasized by the model. The evaluation is currently limited to two public TCGA cohorts and does not include external multi-center validation. The effects of scanner and staining variations, alternative clinical missingness patterns, and the computational cost of constructing ultra-large WSI graphs also require further study. Future work will therefore evaluate MM-HAGT on independent clinical cohorts, investigate more scalable sparse graph representations, and assess model calibration and clinically grounded interpretability.

## REFERENCES


Jianpeng An, Yong Wang, Qing Cai, Gang Zhao, Stephan Dooper, Geert Litjens, and Zhongke Gao. Transformer-based weakly supervised learning for whole slide lung cancer image classification. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15661–15670, 2023a.

Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15661–15670, 2023b.

Chengkuan Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Andrew J Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022.

Dong Chen, Shuai Zheng, Muhao Xu, Zhenfeng Zhu, and Yao Zhao. Signn: A spike-induced graph neural network for dynamic graph representation learning. *Pattern Recognition*, 158:111026, 2025. ISSN 0031-3203.

Ruining Deng, Can Cui, Lucas W Remedios, Shunxing Bao, R Michael Womick, Sophie Chiron, Jia Li, Joseph T Roland, Ken S Lau, Qi Liu, et al. Cross-scale multi-instance learning for pathological image diagnosis. *Medical image analysis*, 94:103124, 2024.

Saisai Ding, Juncheng Li, Jun Wang, Shihui Ying, and Jun Shi. Multi-scale efficient graph-transformer for whole slide image classification. *IEEE Journal of Biomedical and Health Informatics*, 27(12):5926–5936, 2023.

Jiuyang Dong, Junjun Jiang, Kui Jiang, Jiahan Li, and Yongbing Zhang. Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30818–30828, 2025a.

Jiuyang Dong, Junjun Jiang, Kui Jiang, Jiahan Li, and Yongbing Zhang. Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30818–30828, 2025b.

Niharika S D'Souza, Hongzhi Wang, Andrea Giovannini, Antonio Foncubierta-Rodriguez, Kristen L Beck, Orest Boyko, and Tanveer F Syeda-Mahmood. Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond. *Medical Image Analysis*, 93:103064, 2024.

Omar SM El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J Wagner, Katherine J Hewitt, Firas Khader, Sebastian Foersch, et al. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nature Protocols*, 20(1):293–316, 2025.

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

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.

Assia Hijazi, Carlo Bifulco, Pamela Baldin, and Jérôme Galon. Digital pathology for better clinical practice. *Cancers*, 16(9):1686, 2024.

Cheng Jin, Luyang Luo, Huangjing Lin, Jun Hou, and Hao Chen. Hmil: Hierarchical multi-instance learning for fine-grained whole slide image classification. *IEEE Transactions on Medical Imaging*, 2024.

Weipeng Jing, Junze Wang, Donglin Di, Dandan Li, Yang Song, and Lei Fan. Multi-modal hypergraph contrastive learning for medical image segmentation. *Pattern Recognition*, 165:111544, 2025.

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

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.

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.

Yingfan Ma, Xiaoyuan Luo, Kexue Fu, and Manning Wang. Transformer-based video-structure multi-instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14263–14271, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11248–11258, 2024.

Jiangbo Shi, Chen Li, Tieliang Gong, Chunbao Wang, and Huazhu Fu. Cod-mil: Chain-of-diagnosis prompting multiple instance learning for whole slide image classification. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 44(3):1218–1229, MAR 2025. ISSN 0278-0062.

Xiyue Wang, Jinxi Xiang, Jun Zhang, Sen Yang, Zhongyi Yang, Ming-Hui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. *Advances in neural information processing systems*, 35:18009–18021, 2022.

Zhikang Wang, Jiani Ma, Qian Gao, Chris Bain, Seiya Imoto, Pietro Liò, Hongmin Cai, Hao Chen, and Jiangning Song. Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. *Medical Image Analysis*, 97:103252, 2024.

Qitian Wu, Wentao Zhao, Zenan Li, David P Wipf, and Junchi Yan. Nodeformer: A scalable graph structure learning transformer for node classification. *Advances in Neural Information Processing Systems*, 35:27387–27401, 2022.

Zixuan Wu, Haiyong He, Xiushun Zhao, Zhenghui Lin, Yanyan Ye, Jing Guo, Wanming Hu, and Xiaobing Jiang. Reimagining cancer tissue classification: a multi-scale framework based on multi-instance learning for whole slide image classification. *Medical & Biological Engineering & Computing*, pp. 1–19, 2025.

Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 296–306. Springer, 2024.

Jiahui Yu, Tianyu Ma, Yu Fu, Hang Chen, Maode Lai, Cheng Zhuo, and Yingke Xu. Local-to-global spatial learning for whole-slide image representation and classification. *Computerized medical imaging and graphics*, 107:102230, 2023.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18802–18812, 2022.

Tingting Zheng, Kui Jiang, and Hongxun Yao. Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8028–8037, 2024.

Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.