HiPath: Hierarchical Vision-Language Alignment for Structured Pathology Report Prediction

Ruicheng Yuan[1,2], Zhenxuan Zhang[2], Anbang Wang[2], Liwei Hu[2], Xiangqian Hua[3], Yaya Peng[4], Jiawei Luo[1(✉)], and Guang Yang[2(✉)]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, China
{raytion,luojiawei}@hnu.edu.cn
[2] Department of Bioengineering and Imperial-X, Imperial College London, London, UK
g.yang@imperial.ac.uk
[3] Department of Pathology, Xiangtan Maternal and Child Health Hospital, Xiangtan, Hunan, China
[4] Department of Pathology, The First People's Hospital of Xiangtan City, Xiangtan, Hunan, China

Abstract. Pathology reports are structured, multi-granular documents encoding diagnostic conclusions, histological grades, and ancillary test results across one or more anatomical sites; yet existing pathology vision-language models (VLMs) reduce this output to a flat label or free-form text. We present HiPath, a lightweight VLM framework built on frozen UNI2 and Qwen3 backbones that treats structured report prediction as its primary training objective. Three trainable modules totalling 15 M parameters address complementary aspects of the problem: a Hierarchical Patch Aggregator (HiPA) for multi-image visual encoding, Hierarchical Contrastive Learning (HiCL) for cross-modal alignment via optimal transport, and Slot-based Masked Diagnosis Prediction (Slot-MDP) for structured diagnosis generation. Trained on 749 K real-world Chinese pathology cases from three hospitals, HiPath achieves 68.9% strict and 74.7% clinically acceptable accuracy with a 97.3% safety rate, outperforming all baselines under the same frozen backbone. Cross-hospital evaluation confirms generalisation with only a 3.4 pp drop in strict accuracy while maintaining 97.1% safety. Code and evaluation protocol: https://github.com/raytions/hipath.

Keywords: Computational Pathology · Vision-Language Model · Structured Diagnosis · Contrastive Learning

1 Introduction

Recent progress in computational pathology has been driven by large-scale self-supervised foundation models [3,18,10] and vision-language models (VLMs) [7,8] that connect histopathology images with clinical text. These models have become strong generic feature extractors and support a wide range of downstream

tasks, including tissue classification, biomarker prediction, and zero-shot recognition [12,7].

However, the clinical output of pathology is rarely a single label. A routine report is a structured record that jointly specifies (i) a primary diagnosis, (ii) a histological grade, and (iii) immunohistochemistry (IHC) results, often repeated across multiple anatomical sites for the same patient [13]. This structure is central to decision-making: different fields have different clinical consequences, so an evaluation that collapses all errors into flat accuracy can obscure safety-critical failures. At the same time, open-ended text generation offers no guarantee that the output is slot-consistent or reliably parseable.

This leaves an open question for pathology VLMs: can we train and evaluate a model whose primary objective is structured report prediction, rather than flat classification or unconstrained generation? To our knowledge, existing pathology VLMs are not trained or benchmarked for structured, slot-level diagnosis prediction in this setting. A related gap concerns language and reporting conventions. All major pathology VLMs to date are trained exclusively on English corpora [7,12,8], despite substantial clinical demand in non-English contexts. China accounts for approximately 24% of global cancer cases [1,6] and faces an estimated shortage of 90,000 pathologists [20]. Chinese reports also follow distinct conventions, including site-prefixed segment identifiers, standardised grading terminology, and structured IHC formatting. While Chinese groups have introduced notable MIL methods [16], pathology foundation models [19], and VLMs [17], none target structured Chinese diagnosis prediction at scale.

In this work, we propose HiPath (Fig. 1), a lightweight framework that treats structured report prediction as the central learning objective. HiPath is built on frozen UNI2 [3] and Qwen3 [21] backbones and adds 15 M trainable parameters. The model (1) aggregates multi-image evidence selected by pathologists via hierarchical cross-attention (HiPA), (2) uses segment and case-level image-text alignment via optimal transport as an auxiliary training signal (HiCL) [5], and (3) predicts typed diagnostic slots by matching against a closed vocabulary in the frozen LLM embedding space (Slot-MDP). Free-text reports are used only during training; at inference, only frozen vocabulary embeddings and visual features are required. Trained on 749 K cases from three hospitals, HiPath achieves 68.9% strict accuracy, 74.7% clinically acceptable accuracy, and 97.3% safety. Vision-only classifiers on the same frozen visual features plateau at 30–32%, and a non-hierarchical variant with identical text access trails by 12.9 pp. Cross-hospital evaluation (training on two hospitals, testing on the third) shows a 3.4 pp drop while safety remains above 97%.

Our contributions are as follows. (1) We formalise structured slot-level report prediction as a primary objective for pathology VLMs and show that it substantially outperforms flat classification. (2) We introduce a lightweight architecture on frozen backbones that combines hierarchical multi-image aggregation, OT-based local alignment, and vocabulary-grounded slot prediction for structured diagnosis. (3) We propose a four-level clinical evaluation protocol with safety-aware metrics and validate it on 749 K cases from three hospitals. The evaluation

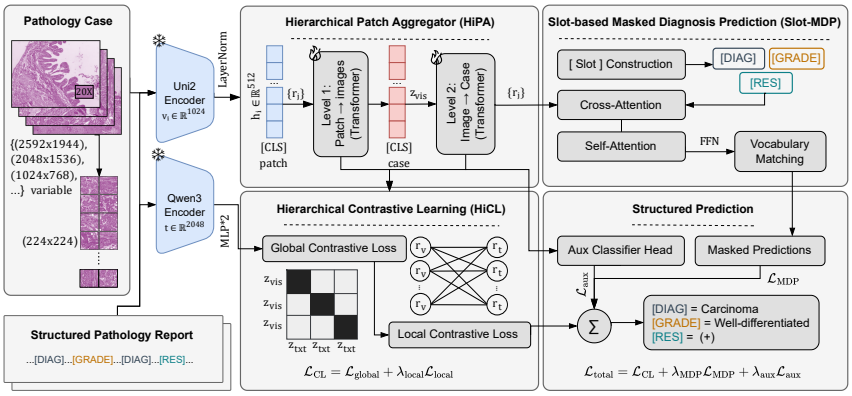

Fig. 1. Overview of HiPath. Frozen UNI2 patch features and Qwen3 text embeddings are processed by three trainable modules: HiPA aggregates patches hierarchically into case representations; HiCL aligns modalities at case and segment level via optimal transport; Slot-MDP fills typed diagnostic slots from visual features alone. Free-text reports are used only during training.

toolkit, including the 304-term vocabulary, synonym mappings, coarse taxonomy, and safety boundary definitions, is publicly released to facilitate reproducible research in structured pathology diagnosis.

2 Methods

Problem formulation. Each case c consists of J physician-selected ROI images and a structured report parsed into anatomical-site segments. In routine Chinese pathology practice, reports follow a conventional structure organised by anatomical site, with each site recording a primary diagnosis, an optional histological grade, and any ancillary findings such as IHC results. We exploit this regularity by parsing historical reports into segments and replacing diagnostic terms with typed masks `[DIAG]`, `[GRADE]`, `[RES]`, yielding K slots $\{(m_k, \tau_k)\}_{k=1}^{K}$, where $m_k \in \mathcal{V}_{\tau_k}$ is the ground-truth term and τ_k its type.

At inference, the model requires a report template specifying the number of segments, slot positions, and slot types. This structure is derivable from information available before pathologist examination: the clinical requisition records the specimen source and number of tissue blocks, while the IHC panel is ordered at accessioning. For example, a breast biopsy with ER/PR testing yields a predictable template of `[DIAG]`+`[GRADE]`+`[RES(ER)]`+`[RES(PR)]`, but this structure reveals only which fields need to be filled, not whether the tissue is benign or malignant, nor whether any marker is positive or negative. Slot queries attend exclusively to visual features, so all diagnostic predictions are grounded in morphological evidence. In our experiments, templates are derived from the

ground-truth report structure; in deployment, they would be constructed from clinical requisition and accessioning metadata.

Training data. We collected 749 K pathology cases from three tertiary hospitals in China spanning 2011–2025. Hospitals A and C are general hospitals; Hospital B is a maternal-child hospital. The top five organs are uterus (9.3%), cervix (8.1%), stomach (6.1%), colon (5.0%), and breast (3.1%), covering 42 distinct organ sites. Each case pairs one or more pathologist-selected ROI images (median 1, P99=8) with a structured Chinese report. The vocabulary was curated to 304 standardised terms (184 DIAG + 93 GRADE + 27 RES) after manual synonym consolidation by two senior pathologists and frequency filtering ($\geq$10 occurrences); the top 10% of terms cover 90.8% of predictions. Cases were split 80/10/10 at the patient level independently per hospital, yielding approximately 600 K/75 K/75 K cases for training, validation, and testing. The 304-term vocabulary and safety boundaries were defined based on clinical standards prior to splitting. This study was conducted under IRB approval; all data were de-identified.

Feature extraction. Visual features: frozen UNI2 (ViT-Giant/14) tiles each image into 224×224 patches and encodes them to native 1536-d vectors, which we project to 1024-d. Text features: frozen Qwen3-30B-A3B [21] (MoE, 30B total/3B activated) encodes each report segment independently, yielding 2048-d embeddings $\{\mathbf{s}_l\}_{l=1}^{L}$ via mean pooling. Feature pre-extraction cost ∼120 GPU-h (H100).

2.1 Hierarchical Patch Aggregator (HiPA)

A pathology case often contains multiple ROI images capturing different tissue regions. Unlike whole-slide methods that apply self-attention within a single gigapixel image [2], HiPA uses cross-attention with learnable queries to aggregate across multiple independent images selected by pathologists. Patch embeddings are projected: $\mathbf{h}_i = \mathrm{LN}(\mathbf{W}_{\mathrm{proj}}\mathbf{v}_i) \in \mathbb{R}^d$ (d=512). Two cross-attention levels aggregate them:

Level 1 (Patch→Image). A learnable `[CLS]` token queries patches within image j:

$$\mathbf{r}_j = \mathrm{TransBlock}([\mathbf{q}_{\mathrm{patch}};\, \mathbf{h}_1, \ldots, \mathbf{h}_{N_j}])\big|_{\texttt{CLS}} \tag{1}$$

Level 2 (Image→Case). A second token aggregates per-image representations:

$$\mathbf{z}_{\mathrm{vis}} = \mathrm{TransBlock}([\mathbf{q}_{\mathrm{case}};\, \mathbf{r}_1, \ldots, \mathbf{r}_J])\big|_{\texttt{CLS}} \tag{2}$$

yielding case embedding $\mathbf{z}_{\mathrm{vis}} \in \mathbb{R}^d$ and per-image representations $\{\mathbf{r}_j\}$ for downstream modules.

2.2 Hierarchical Contrastive Learning (HiCL)

HiCL learns a shared vision-language embedding space via contrastive alignment. It serves as an auxiliary training signal that regularises HiPA's visual

representations and enables image-text retrieval. Qwen3 segment embeddings are projected via a two-layer MLP to case-level $\mathbf{z}_{\text{txt}}$ and segment-level $\{\mathbf{s}_l\}_{l=1}^{L}$.

Global loss. Symmetric InfoNCE on case-level pairs:

$$\mathcal{L}_{\text{global}} = \tfrac{1}{2}\big[\text{CE}(\mathbf{S} \cdot e^{\alpha}, \mathbf{y}) + \text{CE}(\mathbf{S}^{\top} \cdot e^{\alpha}, \mathbf{y})\big] \tag{3}$$

where $S_{ij} = \bar{\mathbf{z}}_{\text{vis},i}^{\top} \bar{\mathbf{z}}_{\text{txt},j}$ (L2-normalised), $\mathbf{y}$ is the identity target, and α is a learnable log-temperature clamped to $[1, 2.7]$.

Local loss. Per-image $\{\mathbf{r}_j\}_{j=1}^{J_c}$ and per-segment $\{\mathbf{s}_l\}_{l=1}^{L_c}$ embeddings form an unordered many-to-many correspondence with $J_c \neq L_c$ in general. We adopt optimal transport with entropic regularisation [5] rather than hard matching (*e.g.*, Hungarian algorithm, which requires $J_c = L_c$) or independent dot-product attention [15] (which treats each pair independently and does not enforce a global assignment). Sinkhorn OT produces a differentiable, dense transport plan that distributes probability mass across all image-segment pairs while respecting marginal constraints. The cost matrix $C_{jl} = 1 - \bar{\mathbf{r}}_j^{\top} \bar{\mathbf{s}}_l$ is solved with Sinkhorn-Knopp (ϵ=0.05, 3 iterations, uniform marginals). Let $\mathcal{B}_{\text{multi}}$ denote mini-batch cases with J_c>1 and L_c>1:

$$\mathcal{L}_{\text{local}} = \frac{1}{|\mathcal{B}_{\text{multi}}|} \sum_{c \in \mathcal{B}_{\text{multi}}} \frac{\langle \mathbf{T}_c^{*}, \mathbf{C}_c \rangle}{\min(J_c, L_c)} \tag{4}$$

where $\mathbf{T}_c^{*} \in \mathbb{R}^{J_c \times L_c}$ is the optimal transport plan. The combined loss is $\mathcal{L}_{\text{HiCL}} = \mathcal{L}_{\text{global}} + 0.5\,\mathcal{L}_{\text{local}}$.

2.3 Slot-based Masked Diagnosis Prediction (Slot-MDP)

Slot-MDP is the primary prediction module. Unlike masked language modelling over open vocabularies, it operates over a closed, type-specific vocabulary: each slot predicts from $\mathcal{V}_{\tau_k}$ via cosine matching against fixed Qwen3 embeddings $\mathbf{e}_w \in \mathbb{R}^{2048}$.

Each masked slot receives a learnable query $\mathbf{q}_k = \mathbf{E}_{\text{pos}}(k) + \mathbf{E}_{\text{type}}(\tau_k)$. Slots attend to per-image representations via cross-attention, then communicate via self-attention to encourage diagnostic consistency:

$$\hat{\mathbf{q}}_k = \text{CrossAttn}(\mathbf{q}_k, \{\mathbf{r}_j\}) + \mathbf{q}_k \tag{5}$$

$$\tilde{\mathbf{q}}_k = \text{SelfAttn}(\hat{\mathbf{q}}_k, \{\hat{\mathbf{q}}_{k'}\}) + \hat{\mathbf{q}}_k \tag{6}$$

Each slot is projected to $\mathbf{p}_k = \text{FFN}(\tilde{\mathbf{q}}_k)$ and scored:

$$p(w \mid k) = \frac{\exp(\beta \cdot \bar{\mathbf{p}}_k^{\top} \bar{\mathbf{e}}_w)}{\sum_{w' \in \mathcal{V}_{\tau_k}} \exp(\beta \cdot \bar{\mathbf{p}}_k^{\top} \bar{\mathbf{e}}_{w'})} \tag{7}$$

where β is learnable (initialised to 10). Training uses a focal loss [11] over the ground-truth term m_k of each slot:

$$\mathcal{L}_{\text{MDP}} = \frac{1}{K} \sum_{k=1}^{K} \big(1 - p(m_k \mid k)\big)^{\gamma}\, \text{CE}_{\epsilon}\big(p(\cdot \mid k),\, m_k\big), \tag{8}$$

Table 1. Main results (%). Top-1 = strict exact match. Accept. = clinically acceptable accuracy (Eq. 9). TP = trainable parameters. All UNI2-based methods share frozen features, data, and split. † Uses Qwen3 text during training only; all methods use only visual features and frozen vocabulary embeddings at inference. HiPath Cross = trained on Hospitals A+C, tested on Hospital B.

Method	Encoder	TP	Top-1	Top-5	Accept.	Safety
Text-only (no images)	Qwen3	0	21.52	54.27	—	—
Linear probe	PLIP	156 K	10.64	45.13	33.91	78.15
	CONCH	156 K	14.63	55.71	42.68	85.62
	UNI2	312 K	30.64	71.81	58.54	87.50
ABMIL [9]	UNI2	603 K	32.37	73.54	60.29	87.21
Deep MLP	UNI2	15.9 M	32.49	73.86	61.76	87.46
Flat CrossAttn†	UNI2	15.7 M	55.98	83.83	68.92	92.18
HiPath†	UNI2	15.0 M	**68.89**	**88.68**	**74.74**	**97.32**
HiPath Cross†	UNI2	15.0 M	65.47	80.99	68.52	97.06

with focusing parameter γ=2 and label-smoothed cross-entropy CE_ϵ (ϵ=0.1).

Training objective. The total loss is $\mathcal{L} = \mathcal{L}_{\mathrm{HiCL}} + 0.5\,\mathcal{L}_{\mathrm{MDP}} + 0.1\,\mathcal{L}_{\mathrm{aux}}$, where $\mathcal{L}_{\mathrm{aux}}$ is BCE for coarse category prediction (12 classes). Trainable parameters total 15.0 M (HiPA 5.2 M, HiCL 4.1 M, Slot-MDP 5.7 M). We trained for 50 epochs with AdamW (lr 10^{-4}, cosine schedule, batch 32) on 4×H100 (∼5.6 h). All experiments are repeated over three random seeds; we report mean values (std ≤0.3 pp for all metrics).

Evaluation protocol. We define four per-slot indicators in a hierarchical cascade: (1) strict: exact match; (2) semantic: match after synonym mapping; (3) coarse: same category (12 classes); (4) safe: no crossing of clinically dangerous boundaries (benign↔malignant, non-adjacent grade jump, IHC polarity reversal). These satisfy strict ⇒ semantic ⇒ coarse and strict ⇒ safe, but coarse and safe are independent. Clinically acceptable accuracy is defined as:

$$\mathrm{Acc}_{\mathrm{accept}} = \frac{1}{N}\left|\{k : \mathrm{strict}(k) \vee \mathrm{semantic}(k) \vee (\mathrm{coarse}(k) \wedge \mathrm{safe}(k))\}\right| \quad (9)$$

An exact or synonym match is always acceptable; a coarse match is acceptable only if safe. This reflects clinical practice [13,14]: identifying the broad diagnostic category suffices for the correct clinical pathway, whereas crossing the benign-malignant boundary causes direct harm. Safety boundaries were defined by consensus of two senior pathologists (>15 years each). All metrics are computed per slot (N=109,523 in the test set).

3 Results and Discussion

Baseline design. The baselines in Table 1 are designed to progressively isolate the factors contributing to HiPath's performance. (i) Linear probes (PLIP,

Table 2. Per-slot-type performance breakdown for HiPath (%). N = number of test slots of each type. Vocab = vocabulary size.

Slot type	Vocab	N	Strict	Semantic	Accept.	Safety
DIAG	184	95,372	69.68	74.25	76.14	96.93
GRADE	93	4,600	71.30	76.11	76.11	100.00
RES	27	9,551	59.88	59.98	59.98	100.00
All	304	109,523	68.89	73.08	74.74	97.32

Table 3. Ablation study with component configuration. ✓/× = present/removed. Flat CrossAttn from Table 1 included for reference. Removing HiPA also disables local OT (no per-image representations to align).

	HiPA		HiCL		Slot-MDP	Top-1	Accept.	i2t R@1	4+img Top-1
	L1	L2	Glob.	Local					
Full HiPath	✓	✓	✓	✓	✓	**68.9**	**74.7**	**21.2**	**55.1**
w/o HiCL	✓	✓	×	×	✓	67.6 (−1.3)	64.3 (−10.4)	0.0	53.2 (−1.9)
w/o HiPA	×	×	✓	×	✓	67.5 (−1.4)	63.3 (−11.4)	22.6	51.4 (−3.6)
w/o Slot-MDP	✓	✓	✓	✓	×	—	64.6 (−10.1)	21.1	—
Flat CrossAttn†	×	×	✓	×	✓	56.0	68.9	14.3	40.7

CONCH, UNI2) and ABMIL [9] are vision-only flat classifiers over the full 304-term vocabulary, establishing the ceiling achievable without text supervision or structured prediction. Deep MLP, the highest-capacity vision-only baseline, applies a 4-block 2048-d LN/GELU/Dropout MLP with a 304-way focal cross-entropy head to mean-pooled UNI2 features; the Text-only row instead classifies from frozen Qwen3 report embeddings without any image input, probing the language prior available in report text alone. English-pretrained encoders (PLIP, CONCH) reach ≤14.6% on Chinese data, reflecting the combined effect of domain shift and label-taxonomy mismatch. UNI2 classifiers plateau at 30–32% regardless of capacity (312 K → 15.9 M parameters yields +1.9 pp), indicating that frozen features alone are insufficient for structured prediction. (ii) Flat CrossAttn† adds text-supervised training and slot-based prediction (Slot-MDP + global contrastive loss) while replacing the two-level HiPA with a single `[CLS]` over all patches and removing local OT. It reaches 56.0%, isolating the combined contribution of text alignment and the slot formulation as ∼24 pp over the vision-only baselines. (iii) The w/o HiCL variant in Table 3 retains HiPA and Slot-MDP but removes all contrastive text alignment, serving as a purely vision-based structured predictor. It achieves 67.6%, confirming that the slot formulation with hierarchical aggregation is the primary driver of performance. Both HiPath and Flat CrossAttn use Qwen3 text only during training (marked †); at inference, only frozen vocabulary embeddings and visual features are used. The remaining 12.9 pp gap between Flat CrossAttn and the full model is decom-

posed by the ablation: hierarchical aggregation (HiPA: +1.4 pp overall, +3.6 pp on 4+ image cases) and contrastive alignment (HiCL: +1.3 pp strict, +10.4 pp acceptable).

Per-slot-type analysis. Table 2 reports performance broken down by slot type. DIAG slots, which draw from the largest vocabulary (184 terms) and require the most nuanced visual interpretation, show the lowest strict accuracy. GRADE and RES slots both maintain 100% safety: all grading errors are between adjacent grades, and non-strict RES errors are predominantly semi-quantitative confusions (*e.g.*, (1+) vs. (2+)) that do not alter clinical decisions. Among the 3.1% of DIAG slots classified as unsafe, the top confusions are: chronic inflammation↔dysplasia (847 slots), adenoma↔adenocarcinoma (312 slots), and leiomyoma↔leiomyosarcoma (201 slots). These are morphologically challenging distinctions with documented high inter-observer variability among expert pathologists [14].

Cross-hospital generalisation. HiPath Cross is trained on Hospitals A+C and tested on Hospital B, with no overlapping patients, staining protocols, or scanners. It achieves 65.5% strict accuracy and 97.1% safety (Table 1), only 3.4 pp below the in-distribution result. Hospital B is a maternal-child hospital with a narrower organ distribution (12 sites vs. 38–41 for A and C), so this experiment primarily tests generalisation to a different clinical population and staining protocol rather than to entirely unseen organ types.

Long-tail analysis. Head terms (top 10%, covering 90.8% of predictions) achieve 70.2% strict accuracy; mid-frequency terms (109 terms) reach 44.4%; tail terms (136 terms, 0.2% of predictions) drop to 2.6%. Three rebalancing strategies (square-root resampling, Focal loss adjustment, class-balanced loss [4]) all degraded head accuracy by 2–5 pp without meaningful tail improvement (+0.8 pp). Inspection of the top-10 tail-term confusions confirms that they are predominantly confused with morphologically similar head terms (*e.g.*, rare carcinoma subtypes misclassified as the dominant "adenocarcinoma" category), consistent with the hypothesis that tail errors reflect visual ambiguity rather than sampling imbalance.

Ablation. Table 3 confirms that each module targets a distinct capability. Removing HiCL collapses retrieval (R@1: 21.2 → 0) and reduces acceptable accuracy by 10.4 pp, while strict accuracy drops more modestly (−1.3 pp) as Slot-MDP compensates; HiCL's value is therefore primarily in cross-modal alignment and the auxiliary regularisation it provides to visual representations. Removing HiPA costs 3.6 pp on multi-image cases (4+ images) but only 0.8 pp on single-image cases, confirming that hierarchical aggregation specifically benefits multi-ROI interpretation. The gap between Flat CrossAttn (56.0%) and w/o HiPA (67.5%) shows that HiPA and local OT interact: without per-image representations, OT has nothing to operate on, creating a compounding deficit.

We note two limitations of the current experimental design. First, we do not compare against a generative baseline (*e.g.*, a visual-adapter-augmented Qwen3 with constrained decoding over the same vocabulary. Such a comparison would require fine-tuning the LLM backbone, which substantially exceeds the 15 M pa-

rameter budget of our framework; we consider this an important direction for future work. Second, while we motivate Sinkhorn OT by its ability to handle unequal-sized sets differentiably (§2.2), a systematic comparison against alternative soft-assignment strategies (*e.g.*, softmax attention) is not included.

4 Conclusion

We presented HiPath, a lightweight vision-language framework that treats structured pathology report prediction as its primary learning objective. With only 15 M trainable parameters on frozen backbones, HiPath achieves 68.9% strict accuracy, 74.7% clinically acceptable accuracy, and 97.3% safety on 749 K cases, with robust cross-hospital generalisation. Several limitations point to future work: the current evaluation covers three hospitals in one region (broader multi-centre validation is needed); the model operates on pathologist-selected ROIs rather than whole-slide images; and cases outside the 304-term vocabulary are assigned the nearest term rather than rejected. We release the evaluation protocol, including the vocabulary, synonym mappings, coarse taxonomy, and safety boundary definitions, to support reproducibility and further research.

Acknowledgments. R. Yuan was supported by the China Scholarship Council (CSC). G. Yang was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, the NIHR Imperial Biomedical Research Centre (RDA01), the Wellcome Leap Dynamic Resilience program (co-funded by Temasek Trust), UKRI guarantee funding for Horizon Europe MSCA Postdoctoral Fellowships (EP/Z002206/1), a UKRI MRC Research Grant, TFS Research Grants (MR/U506710/1), the Swiss National Science Foundation (Grant No. 220785), and the UKRI Future Leaders Fellowship (MR/V023799/1, UKRI2738). J. Luo was supported in part by the National Natural Science Foundation of China (No. 62372165).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **74**(3), 229–263 (2024). https://doi.org/10.3322/caac.21834
2. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16144–16155 (2022)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3

4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9268–9277 (2019)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 2292–2300 (2013)
6. Han, B., Zheng, R., Zeng, H., Wang, S., Sun, K., Chen, R., Li, L., Wei, W., He, J.: Cancer incidence and mortality in China, 2022. Journal of the National Cancer Center **4**(1), 47–53 (2024). https://doi.org/10.1016/j.jncc.2024.01.006
7. Huang, Z., et al.: A visual–language foundation model for pathology image analysis using medical Twitter. Nature Medicine **29**(9), 2307–2316 (2023). https://doi.org/10.1038/s41591-023-02504-3
8. Ikezogwo, W.O., et al.: Quilt-1m: One million image-text pairs for histopathology. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
9. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning (ICML). pp. 2127–2136 (2018)
10. Karasikov, M., van Doorn, J., Känzig, N., Cesur, M.E., Horlings, H.M., Berke, R., Tang, F., Otálora, S.: Training state-of-the-art pathology foundation models with orders of magnitude less data. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Lecture Notes in Computer Science, vol. 15967, pp. 573–583. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-04984-1_55
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (ICCV) (2017)
12. Lu, M.Y., Chen, B., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4
13. Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A., et al.: The 2019 WHO classification of tumours of the digestive system. Histopathology **76**(2), 182–188 (2019). https://doi.org/10.1111/his.13975
14. Raab, S.S., Grzybicki, D.M., Janosky, J.E., Zarbo, R.J., Meier, F.A., Jensen, C., Geyer, S.J.: Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. Cancer **104**(10), 2205–2213 (2005). https://doi.org/10.1002/cncr.21431
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
16. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 2136–2147 (2021)
17. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., Yang, L.: PathAsst: A generative foundation AI assistant towards artificial general intelligence of pathology. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5034–5042 (2024). https://doi.org/10.1609/aaai.v38i5.28308
18. Vorontsov, E., et al.: Virchow: A million-slide digital pathology foundation model (2023). https://doi.org/10.48550/arXiv.2309.07778

19. Wang, X., Zhao, J., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024). https://doi.org/10.1038/s41586-024-07894-z
20. Xu, C., Li, Y., Chen, P., Pan, M., Bu, X.: A survey on the attitudes of Chinese medical students towards current pathology education. BMC Medical Education **20**(1), 259 (2020). https://doi.org/10.1186/s12909-020-02167-5
21. Yang, A., et al.: Qwen3 technical report (2025). https://doi.org/10.48550/arXiv.2505.09388