From Multi-Resolution Cells to Gigapixel Whole Slide Images Foundation Model for Computational Pathology

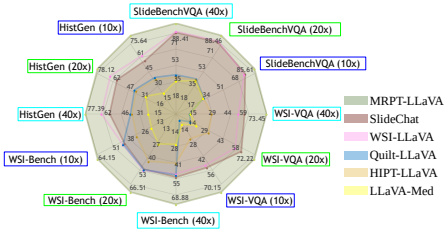
Basit Alawode[1], Moshira Ali Abdalla[1], Dwarikanath Mahapatra[1], Muhammad Muzammal Naseer[1,2], and Sajid Javed[1]

[1] Khalifa University of Science and Technology, UAE
[2] University of Western Australia, Australia
{basit.alawode,sajid.javed}@ku.ac.ae

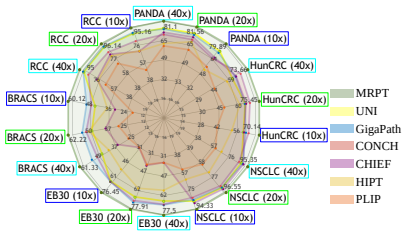
Abstract. Vision Transformers (ViTs) and their hierarchical variants have achieved strong performance in Computational Pathology (CPath). However, most are pre-trained on single-resolution Whole Slide Images (WSIs), limiting their generalization across arbitrary resolutions. Gigapixel WSIs inherently contain diagnostic patterns at multiple scales, including cellular morphologies, tissue architectures, and global context, mirroring how expert pathologists examine WSIs. We introduce Multi-Resolution Pyramid Transformer (MRPT), a model that hierarchically aggregates multi-resolution information from cellular to tissue and WSI levels. MRPT employs a biologically meaningful Consecutive Cross-Resolution Attention (CCRA) mechanism to capture scale-independent interactions and enforces multi-resolution semantic consistency by aligning embeddings across resolutions, yielding robust and generalizable WSI representations. Pre-trained in a multi-resolution self-supervised manner on 624M patches, 2.4M regions, and 36K WSIs, MRPT learns rich coarse-to-fine histopathology features. Extensive experiments on 34 diverse datasets show that MRPT surpasses recent foundation models and Multimodal Large Language Models (MLLMs) in cancer subtype classification, tissue phenotyping, and Visual Question Answering (VQA) for WSI understanding. Our code and models available on link.

1 Introduction

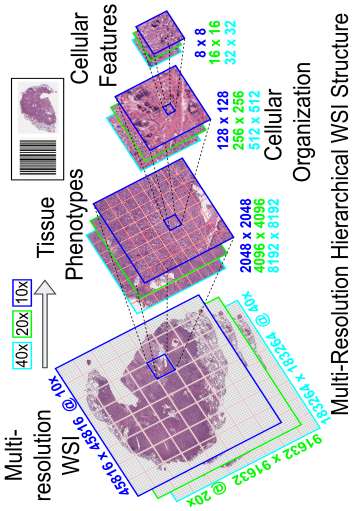
Diagnosing cancer and assessing patient outcomes from gigapixel Whole Slide Images (WSIs) remains the gold standard in clinical pathology [68, 71, 97]. The digitization of glass slides into WSIs has transformed modern pathology, enabling large-scale and systematic computational analysis of tissue morphology [27,31,56,72,80]. This digital transformation has given rise to the field of Computational Pathology (CPath), which leverages advances in machine learning and computer vision to automatically extract diagnostic and prognostic insights from gigapixel WSIs [3,43,63,81]. *A WSI inherently exhibits a hierarchical and multi-resolution structures (Fig. 1c) [1, 4, 23, 74].* The hierarchical structure reflects the nested organization of visual information: small regions comprising individual cellular entities (e.g., tumor cells, stroma, lymphocytes) form local clusters



(a) WSI VQA Task (Accuracy)



(b) WSI Classification Task (Balanced Accuracy)



(c) MRPT Hierarchical and Multi-resolution representation

Fig. 1: (a) and **(b)**: Our MRPT model remains consistent across resolutions and outperforms existing SOTA models by a significant margin. **(c)**: MRPT exploits hierarchical and multi-resolution representations to capture rich contextual information from WSIs.

that capture cell-cell interactions [23]. These clusters aggregate into larger tissue microenvironments, which collectively compose the full WSI, encapsulating the overall tissue heterogeneity [19, 67, 81]. Conversely, the multi-resolution structure refers to viewing the same spatial region at varying resolutions (e.g., 10×, 20×, or 40×), where each resolution maintains spatial alignment while offering distinct levels of visual detail [5, 38, 50, 95].

Expert pathologists routinely perform holistic analyses of WSIs across multiple resolutions, synthesizing evidence from different spatial scales and regions to generate comprehensive diagnostic reports [9, 69, 96]. For example, in colorectal cancer, low resolutions capture global tissue architecture such as mucosal orientation, intermediate resolutions reveal glandular organization and architectural distortion, and high resolution expose cellular features such as nuclear atypia, pleomorphism, and mitotic activity [39, 87]. *Therefore, multi-resolution analysis of WSIs is indispensable for accurate cancer diagnosis, as it integrates complementary information across scales (Fig. 1c) [34, 76].*

Recently, numerous CPath foundation models have been developed to help pathologists in diagnostic and prognostic decision-making [24,64,80]. Most SOTA models are patch-level at a single resolution such as CONCH [64], UNI [24], and QuiltNet [48]. MR-PLIP is the only multi-resolution patch-level model [4]; however, it does not fully exploit the hierarchical structure. In SOTA models, a WSI is divided into smaller patches during training [4, 48, 90]. *However, such patch-level representations often lack global spatial context of the entire WSI, limiting their ability to capture tissue-level organization and cross-regional dependencies [25, 29].* To overcome this limitation, several WSI-level models, such as

GigaPath [93], WSI-LLaVA [58], and SlideChat [25], have been introduced. These methods aggregate information across all patches to learn holistic WSI-level representations, thereby incorporating global contextual cues [29]. *While these models successfully encode global information, they often overlook the intrinsic WSI hierarchical organization, resulting in suboptimal WSI-level representations and reduced performance (Figs. 1 (a) & (b)) [29].* To address this, hierarchical models have been proposed, which exploit the nested structure of WSIs for improved feature learning [23, 36, 52]. For instance, the HIPT model captures the natural hierarchy within WSIs via a two-level Self-Supervised Learning (SSL) framework, producing enriched WSI representations [23]. While hierarchical modeling has improved WSI understanding, *the explicit integration of multi-resolution information within these hierarchies remains largely underexplored (see Table 1 in supplementary for more comparisons of SOTA models).*

To this end, *we propose a novel Multi-Resolution Pyramid Transformer (**MRPT**) that jointly models hierarchical and multi-resolution representations at the WSI level.* MRPT captures complementary diagnostic cues across multi-resolution cellular, tissue, and WSI scales, mirroring the diagnostic workflow of expert pathologists who continuously zoom in and out to assess both local details and global context. This design enables the model to preserve fine-grained morphological information while maintaining global spatial awareness, resulting in a more robust, generalizable, and clinically meaningful CPath model. *MRPT is hierarchically pre-trained on 624M patches, 2.4M regions, and 36K WSIs curated from TCGA [46] and CPTAC [32], leveraging multi-resolution SSL.* To maintain biological interpretability and semantic consistency, MRPT performs cross-attention only between consecutive resolutions, emulating the progressive zooming behavior of pathologists. At the cell-level pre-training stage, we introduce a *Consecutive Cross-Resolution Attention (CCRA)* mechanism that fuses information across resolutions using a novel ViT architecture designed to align token representations at multiple scales. The aggregated cell-level embeddings are used to represent corresponding patches, which are trained via multi-resolution SSL at the patch-level. Subsequently, the aggregated multi-resolution patch-level features are employed to represent higher-level regions. At the region-level, SSL is again applied to learn region-specific representations, which are ultimately aggregated into holistic, multi-resolution WSI-level embeddings. Finally, SSL is performed at the WSI-level to learn a unified, hierarchical representation encompassing all three scales. *Through this pipeline, MRPT effectively learns a multi-resolution, hierarchically structured representation.*

To further extend MRPT, we introduce a multimodal conversational model for WSI understanding, termed **MRPT-LLaVA**. This MLLM integrates MRPT representations with LLaVA [61] to facilitate cross-modal reasoning. Most existing MLLMs such as Quilt-LLaVA [77], operate only at the patch-level, thereby lacking global WSI-level context. Recent WSI-level models, including SlideChat [25] and WSI-LLaVA [58], incorporate WSI information but rely on a single resolution, limiting their ability to capture hierarchical and multi-resolution representations. *In contrast, MRPT-LLaVA leverages multi-resolution hierarchi-*

cal WSI-level representations to perform complex pathology VQA tasks. MRPT-LLaVA employs a three-stage training approach: *cross-modal alignment*, *feature space alignment*, and *visual-instruction tuning* on WSI-Bench dataset [58]. To evaluate the effectiveness of MRPT, we conduct extensive experiments on 34 publicly available datasets covering a range of CPath tasks, including classification, caption generation, and VQA. Our model achieves superior performance compared to the SOTA CPath foundation models. Our main contributions are:

1. We introduce a three-stage multi-resolution Self-Supervised Learning (SSL) paradigm that jointly models multi-resolution and hierarchical representations of WSIs, enabling comprehensive learning of both fine-grained cellular features and global tissue context.
2. We propose a novel multi-resolution cell-level ViT architecture based on the biologically inspored Consecutive Cross-Resolution Attention (CCRA) mechanism, which efficiently integrates cellular information across multi-resolution.
3. We develop MRPT-LLaVA, a CPath MLLM that leverages multi-resolution hierarchical WSI representations to perform complex VQA tasks.

2 Literature Review

1. Vision-only Foundation Models (VFMs): These models are pre-trained on large-scale histopathology datasets using SSL and serve as backbones for downstream tasks [42,44,53]. Most existing VFMs including UNI [24], REMEDIS [7], and CHIEF [91], operate at the patch-level [48, 64], and employ a DINO-based SSL framework using millions scale histology patches [17]. Many patch-level VFMs including Virchow [89] and GigaPath [93], aggregate patch-level embeddings to form holistic WSI representations [89, 91, 93]. *Patch-level learning inherently lacks the comprehensive global context embedded within the WSI.* Hierarchical models use SSL to combine local and global features [23,36,52]. A well-known example is HIPT [23], which is pre-trained at the cellular and tissue levels, and merges them into a WSI-level embedding [23]. *Nonetheless, existing hierarchical approaches neglect the inherent multi-resolution WSI information, which can lead to suboptimal representation and performance degradation.* Some patch-level VFMs, including RudolfV [30] and Virchow2 [89], utilize multi-resolution patches primarily to enhance dataset diversity [4, 53, 74]. However, these models do not explicitly exploit the cross-resolution dependencies within WSIs. *Our proposed MRPT introduces a WSI-level framework that hierarchically fuses information across multi-resolution, effectively combining the benefits of global context and multi-resolution information hierarchy.* **2. Vision-Language Models (VLMs):** CPath VLMs including PLIP [44], QuiltNet [48], and CONCH [64] etc., learn a shared embedding space between histology images and textual descriptions using contrastive learning [73], enabling cross-modal alignment, retrieval, and zero-shot classification [44, 51, 66]. Most existing VLMs, however, are pre-trained at a single resolution, failing to capture the coarse-to-fine contextual hierarchy in WSIs. MR-PLIP is the only multi-resolution patch-level

VLM [4]; however, it does not fully exploit the hierarchical structure. *Aggregating local features without respecting the hierarchical structure of WSIs inherently limits the global contextual understanding necessary for comprehensive WSI-level interpretation.* **3. MLLMs:** These models integrate histology images, textual information, and patient-level clinical data into a unified reasoning framework for complex VQA task [25, 77]. Most existing MLLMs operate at the *patch-level*, including PathChat [65], Quilt-LLaVA [77], and PathAsst [85] etc. WSI-level MLLMs, such as TITAN [29], SlideChat [25], and WSI-LLaVA [58], aggregate patch-level local features for WSI reasoning. While these MLLMs have significantly advanced complex WSI understanding, they still fail to leverage the intrinsic WSIs *hierarchical* and *multi-resolution* structures. *Our MRPT-LLaVA explicitly exploits these structures in a WSI.* See Table 1 in supplementary for more SOTA models comparisons.

3 Methodology

A schematic illustration of our proposed MRPT model is shown in Fig. 2. The multi-resolution representations are learned in a hierarchical manner, progressing from the cell-level to the patch-level, then to the region-level, and finally to the WSI-level. At each level, multi-resolution SSL is employed to learn rich, fine-grained, and coarse-grained representations. Our proposed MRPT-LLaVA is pre-trained in three different stages to address complex VQA tasks for WSI interpretation.

3.1 Proposed MRPT Model

Given an input multi-resolution WSI, $\mathbf{M} = \{\mathbf{X}_{\text{WSI}} \text{ at } 10\times,\ \mathbf{Y}_{\text{WSI}} \text{ at } 20\times,\ \mathbf{Z}_{\text{WSI}} \text{ at } 40\times\}$, we extract m non-overlapping regions each of size 4096×4096 at $20\times$, denoted as $\{\mathbf{Y}^i_{4096}\}_{i=1}^m$. For the corresponding $10\times$ and $40\times$ resolutions, we extract aligned regions of sizes $\{\mathbf{X}^i_{2048}\}_{i=1}^m$ and $\{\mathbf{Z}^i_{8192}\}_{i=1}^m$, respectively. These triplets of multi-resolution regions are represented as $\{\mathbf{R}_i\}_{i=1}^m = \{\mathbf{X}^i_{2048},\ \mathbf{Y}^i_{4096},\ \mathbf{Z}^i_{8192}\}_{i=1}^m$, where each $\mathbf{R}_i$ corresponds to the same anatomical region across resolutions, preserving spatial alignment and contextual integrity. Each region $\mathbf{R}_i$ is subdivided into smaller patches represented as $\{\mathbf{P}^j_i\}_{j=1}^{256} = \{\mathbf{X}^j_{128},\ \mathbf{Y}^j_{256},\ \mathbf{Z}^j_{512}\}_{j=1}^{256}$. Each $\mathbf{P}^j_i$ is further decomposed into multi-resolution cell-level patches denoted as $\{\mathbf{C}^j_{i,k}\}_{k=1}^{256} = \{\mathbf{X}^k_8,\ \mathbf{Y}^k_{16},\ \mathbf{Z}^k_{32}\}_{k=1}^{256}$. This co-registration allows the model to jointly capture global tissue context from low-resolution views and localized morphological information from high-resolution ones. Our objective is to pre-train a large-scale, multi-resolution, hierarchical WSI-level foundation model enabling unified representation learning across all resolutions.

3.2 MRPT Architecture

Learning a robust multi-resolution WSI-level representation presents two main challenges: (1) a multi-resolution WSI contains a significantly larger sequence length of tokens compared to a single resolution, and (2) aggregating and aligning information across multi-resolution is non-trivial.

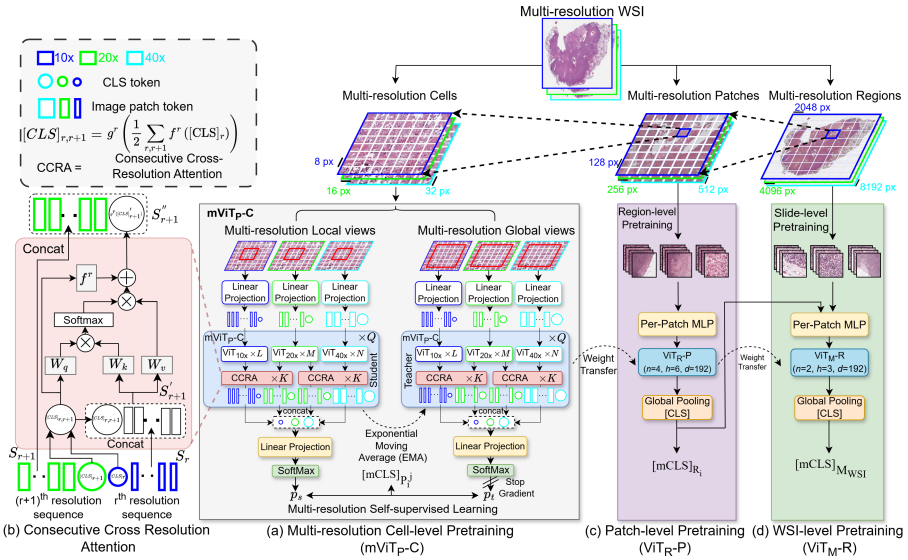


Fig. 2: An overview of proposed MRPT architecture, which comprises three hierarchical pre-training stages: (a)-(b) cell-level (mViT$_{\mathbf{P}}$-C), (c) patch-level (ViT$_{\mathbf{R}}$-P), and (d) region-level (ViT$_{\mathbf{M}}$-R). The cell-level ViT, mViT$_{\mathbf{P}}$-C, employs multi-resolution SSL paradigm. The teacher and student networks learn multi-resolution representations using the proposed Consecutive Cross-Resolution Attention (CCRA) mechanism that fuses features across consecutive resolutions (10×, 20×, and 40×). The learned multi-resolution [CLS] tokens are progressively aggregated from cell- to patch- to region-level through ViT$_{\mathbf{R}}$-P and ViT$_{\mathbf{M}}$-R to produce WSI-level embedding.

To address the first challenge, we adopt a hierarchical WSI structure inspired by HIPT [23]. We extend the ViT architecture for multi-resolution WSI representation learning by employing a nested aggregation of visual tokens that are recursively decomposed into smaller units. To address the second challenge, we propose a multi-resolution ViT that incorporates an efficient *Consecutive Cross-Resolution Attention (CCRA)* mechanism to effectively integrate information across consecutive resolutions. Our proposed MRPT follows a multi-resolution hierarchical structure ($\mathbf{M} \rightarrow \{\mathbf{R}_i\}_{i=1}^{m} \rightarrow \{\mathbf{P}_i^j\}_{j=1}^{256} \rightarrow \{\mathbf{C}_{i,k}^j\}_{k=1}^{256}$) (Fig. **??**). At the lowest hierarchical level, a ViT module aggregates cell-level features across consecutive resolutions using the CCRA mechanism. This design enables MRPT to capture hierarchical and multi-resolution structures of WSIs and to model clinically meaningful relationships across resolutions.

The pre-training of MRPT is performed in three distinct stages. In **Stage 1** (Fig. 2 (a)-(b)), we pre-train cell-level features $\{\mathbf{C}_{i,k}^j\}_{k=1}^{256}$ using multi-resolution SSL paradigm, where both teacher and student networks leverage a CCRA-enabled ViT architecture, referred to as mViT$_{\mathbf{P}}$-C. Here, C $\in \mathbf{C}_{i,k}^j$ represents multi-resolution cell-level tokens extracted from patches $\{\mathbf{P}_i^j\}_{j=1}^{256}$. In **Stage 2** (Fig. 2 (c)), the weights of mViT$_{\mathbf{P}}$-C are frozen and re-used as an embedding layer for patch-level pre-training using another ViT module, ViT$_{\mathbf{R}}$-P, where P $\in \mathbf{P}_i^j$ represents multi-resolution patch-level tokens. ViT$_{\mathbf{R}}$-P aggregates patch-

level embeddings to learn region-level representations $\mathbf{R}_i$. In **Stage 3** (Fig. 2 (d)), the weights of ViT$_{\mathbf{R}}$-P are frozen and re-used as a backbone for WSI-level pre-training using ViT$_{\mathbf{M}}$-R, where R $\in \{\mathbf{R}_i\}_{i=1}^{m}$ denotes multi-resolution region-level tokens. ViT$_{\mathbf{M}}$-R aggregates the region-level embeddings to produce the final WSI-level representation. The overall mathematical formulation of MRPT is defined as:

$$\begin{aligned} &\text{MRPT}(\mathbf{M}) = \text{ViT}_{\mathbf{M}}\text{-R}\left(\{[\text{mCLS}]_{\mathbf{R}_i}\}_{i=1}^{m}\right), \text{where} \\ &[\text{mCLS}]_{\mathbf{R}_i} = \text{ViT}_{\mathbf{R}}\text{-P}\left(\{[\text{mCLS}]_{\mathbf{P}_i^j}\}_{j=1}^{256}\right), \text{where} \\ &[\text{mCLS}]_{\mathbf{P}_i^j} = \text{mViT}_{\mathbf{P}}\text{-C}\left(\{\mathbf{C}_{i,k}^{j}\}_{k=1}^{256}\right), \end{aligned} \tag{1}$$

where $[\text{mCLS}]_{\mathbf{P}_i^j}$ denotes the multi-resolution [CLS] token computed over patch $\mathbf{P}_i^j$, and $[\text{mCLS}]_{\mathbf{R}_i}$ denotes the multi-resolution [CLS] token at the region level $\mathbf{R}_i$. Both mViT$_{\mathbf{P}}$-C and ViT$_{\mathbf{R}}$-P operate with a sequence length of 256 tokens, while ViT$_{\mathbf{M}}$-R processes approximately $m \approx 256$ region tokens per resolution. When small ViT backbones are employed at each stage, MRPT contains fewer than 12M parameters, making it computationally efficient and trainable on standard GPU workstations. Each pre-training stage of MRPT is detailed in the following sections.

Multi-Resolution Cell Aggregation (mViT$_{\mathbf{P}}$-C) The aggregation of multi-resolution cell-level tokens within $\mathbf{P}_i^j$ is performed using the proposed mViT$_{\mathbf{P}}$-C (Figs. 2(a)–(b)). This module employs the CCRA mechanism to integrate information across consecutive resolutions. To obtain the learnable embeddings $[\text{mCLS}]_{\mathbf{P}_i^j}$, the cell-level representations across all resolutions are aggregated along the token sequence. Multi-resolution SSL paradigm is employed in which local views are processed by the student network and global views are processed by the teacher. Both teacher and student learn to align the multi-resolution representation. Below, we discuss the general architecture of mViT$_{\mathbf{P}}$-C, which is the same for both student and teacher.

Given $\mathbf{P}_i^j$, mViT$_{\mathbf{P}}$-C decomposes it into r sequences, each of length 256, corresponding to non-overlapping $c_r \times c_r$ pixel tokens at different resolutions $r \in \{10\times, 20\times, 40\times\}$, where c_r denotes the cell-level patch size at r. For each resolution, a linear projection layer with positional embedding is applied to obtain a sequence embedding $\mathbf{S}_r$, which includes an appended $[\text{CLS}]_r$ token. A transformer ViT$_r$ is then pre-trained independently to learn resolution-specific embeddings. The number of encoder layers in each ViT$_r$ is adjusted according to the resolution to balance computation. Finally, the outputs of ViT$_r$ from two consecutive resolutions are fused via the proposed CCRA mechanism to yield the multi-resolution representation.

mViT$_{\mathbf{P}}$-C Architecture: mViT$_{\mathbf{P}}$-C consists of a stack of Q hierarchical fusion blocks. Each block contains three branches that process tokens from the three resolutions and fuse their outputs using the CCRA mechanism. As illustrated in Fig. 2 (a), the architecture is composed of Q multi-resolution transformer

encoders, where each encoder comprises: (1) a *10× branch* handling coarse-grained features (8×8 patch size) with L transformer layers, (2) a *20× branch* processing medium-grained features (16×16 patch size) with M layers, and (3) a *40× branch* focusing on fine-grained features (32×32 patch size) with N layers. The three branches are fused Q times through CCRA layers that enable the $[\text{CLS}]_r$ tokens of all branches to exchange contextual information across resolutions.

Consecutive Cross-Resolution Attention (CCRA): The proposed CCRA module (Fig. 2(b)) fuses information between two consecutive resolutions by averaging their $[\text{CLS}]_r$ tokens and appending them to the patch tokens of the finer sequence. This design allows efficient and interpretable cross-resolution information exchange. Specifically, the $[\text{CLS}]_r$ tokens from both sequences serve as information carriers between resolutions and are back-projected to their respective branches. Because $[\text{CLS}]_r$ tokens encode abstract-level semantics of their sequences, interacting with patch tokens of other resolutions enriches both coarse- and fine-grained contextual representations. The process continues across CCRA layers, where $[\text{CLS}]_r$ tokens iteratively refine their own sequence representations, thereby enhancing semantic alignment across resolutions. *This mechanism efficiently fuses coarse and fine features at the cell level while maintaining semantic consistency.*

While performing cross-attention between consecutive resolutions, we preserve the hierarchical diagnostic logic used by pathologists: 10× magnification provides a global overview of tissue architecture, 20× reveals local glandular and stromal organization, and 40× exposes nuclear and cellular morphology. *Pathologists examine WSIs sequentially across resolutions rather than skipping scales, as intermediate resolutions contextualize cellular detail within broader architecture.* Inspired by this workflow, MRPT restricts cross-attention to consecutive resolution pairs ($10\times \leftrightarrow 20\times$, $20\times \leftrightarrow 40\times$), ensuring semantic continuity and minimizing misalignment across distant representations. This design produces biologically interpretable and diagnostically meaningful embeddings (*see ablation*).

Class Token Fusion: The $[\text{CLS}]_r$ tokens represent global contextual summaries of their corresponding sequences and are used as final embeddings for downstream tasks [24, 93]. We first average the $[\text{CLS}]_r$ tokens of two consecutive resolutions (Fig. 2(d)) and append the averaged token to both sequences, enabling bidirectional feature exchange. Formally, the fused sequence $\mathbf{S}'_{r+1}$ is given by:

$$\mathbf{S}'_{r+1} = \left[g^r\left(\frac{1}{2}\sum_{r,r+1} f^r([\text{CLS}]_r)\right) \| \mathbf{S}_r \right], \tag{2}$$

where $f^r(\cdot)$ and $g^r(\cdot)$ denote projection and back-projection functions for dimensional alignment.

Multi-Resolution Fusion Using CCRA: We then perform CCRA between the averaged $[\text{CLS}]_{r,r+1}$ tokens and $\mathbf{S}'_{r+1}$. The query corresponds to $[\text{CLS}]_{r,r+1}$,

while keys and values correspond to $\mathbf{S}_{r+1}^{'}$, expressed as:

$$\mathbf{q} = [\text{CLS}]_{r,r+1}\mathbf{W}_q, \quad \mathbf{k} = \mathbf{S}_{r+1}^{'}\mathbf{W}_k, \quad \mathbf{v} = \mathbf{S}_{r+1}^{'}\mathbf{W}_v,$$
$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d}}\right), \quad \text{CCRA}(\mathbf{S}_{r+1}^{'}) = \mathbf{A}\mathbf{v}, \tag{3}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable projection matrices, and d is the embedding dimension. *Because the query consists of a single vector $[CLS]_{r,r+1}$, the computational complexity of CCRA is linear rather than quadratic, making it highly efficient.* We further employ multi-head CCRA (MHCCRA) for richer feature fusion. The updated sequence $\mathbf{S}_{r+1}^{''}$ for a given $\mathbf{S}_{r+1}$ is obtained with Layer Normalization (LN) and residual connections as:

$$[\text{CLS}]_{r+1}^{'} = f^r([\text{CLS}]_{r,r+1}) + \text{MHCCRA}\Big(\text{LN}(\mathbf{S}_{r+1}^{'})\Big),$$
$$\mathbf{S}_{r+1}^{''} = \Big[g^r\Big([\text{CLS}]_{r+1}^{'}\Big) \parallel \mathbf{S}_{r+1}\Big]. \tag{4}$$

Empirical results confirm that both class-token fusion and CCRA are critical for efficient and effective multi-resolution feature learning (*Table 5)*. For intermediate resolution (20×), two sequences are obtained: $\mathbf{S}_{10\times-20\times}^{''}$ and $\mathbf{S}_{20\times-40\times}^{''}$, whose corresponding [CLS] tokens ($[\text{CLS}]_{10\times-20\times}^{'}$, $[\text{CLS}]_{20\times-40\times}^{'}$) are concatenated and projected back to their original dimension to form $[\text{CLS}]_{20\times}^{'}$.

There are K stacked CCRA layers in total. At each layer, fine-grained features from 40× are propagated to 20×, then to 10×, and vice versa. The outputs of the final CCRA layer yield three sequences: $\mathbf{S}_{10\times}^{''}$, $\mathbf{S}_{20\times}^{''}$, and $\mathbf{S}_{40\times}^{''}$. The three [CLS] tokens—$[\text{CLS}]_{10\times}^{'}$, $[\text{CLS}]_{20\times}^{'}$, and $[\text{CLS}]_{40\times}^{'}$—are concatenated and passed through a projection layer to form multi-resolution patch representation $[\text{mCLS}]_{\mathbf{P}_i^j}$.

Multi-Resolution Patch Aggregation ($\text{ViT}_\mathbf{R}$-P) To obtain multi-resolution region-level embedding $[\text{mCLS}]_{\mathbf{R}_i}$, we aggregate the sequence $\{[\text{mCLS}]_{\mathbf{P}_i^j}\}_{j=1}^{256}$, derived via $\text{mViT}_\mathbf{P}$-C, by appending a [CLS] token and feeding it into $\text{ViT}_\mathbf{R}$-P architecture, to model broader tissue-level contextual relationships (Fig. 2 (c)).

Multi-Resolution Region Aggregation ($\text{ViT}_\mathbf{M}$-R) To obtain the multi-resolution WSI-level representation $[\text{mCLS}]_{\mathbf{M}_{\text{WSI}}}$, we aggregate sequence $\{[\text{mCLS}]_{\mathbf{R}_i}\}_{i=1}^{m}$, derived via $\text{ViT}_\mathbf{R}$-P, by appending a [CLS] token and feeding it into $\text{ViT}_\mathbf{M}$-R to model holistic representation (Fig. 2 (c)).

Hierarchical Pre-training of MRPT Model Stage 1: Multi-Resolution Patch-Level ($\mathbf{P}_i^j$) Pre-training. We pre-train $\text{mViT}_\mathbf{P}$-C using the multi-resolution SSL framework, wherein a student network learns to match the output distribution of a teacher network via cross-entropy loss: $H = -\sum_{i,j\in\{10\times,20\times,40\times\}} p_{t_i} \log p_{s_j}$, where p_{t_i} and p_{s_j} represent the outputs of teacher and student for resolutions i and j, respectively. For each multi-resolution patch $\mathbf{P}_i^j$, DINO generates a set of $L_r = 8$ local views (of sizes 48×48, 96×96, and 192×192 for 10×, 20×,

Table 1: Configuration of the mViT$_{\mathbf{P}}$-C architecture across Tiny (T), Small (S), and Base (B) variants. Each model processes three resolutions (10×, 20×, 40×) with corresponding embedding dimensions, number of attention heads (h_L, h_M, h_N), and encoder depths (L, M, N). For all variants, we fix CCRA layers to $K = 1$ and hierarchical fusion blocks to $Q = 4$.

Variants	Dimension			# heads	Encoders
	10×	20×	40×	h_L, h_M, h_N	L, M, N
mViT$_{\mathbf{P}}$-C-T	96	192	192	3,3,3	2,4, 4
mViT$_{\mathbf{P}}$-C-S	192	384	384	6,6,6	2,4,4
mViT$_{\mathbf{P}}$-C-B	384	768	768	12,12,12	2,4,4

and 40×, respectively) input to Ψ_s, and $G_r = 2$ global views (of sizes 112×112, 224×224, and 448×448) input to Ψ_t. This encourages local-to-global consistency across resolutions, minimizing:

$$\min_{\Psi_s} \sum_{r} \sum_{G_r} \sum_{L_r} H\big(p_{t_r}(G_r), p_{s_r}(L_r)\big). \tag{6}$$

Stage 2: Multi-Resolution Region-Level ($\mathbf{R}_i$) Pre-training. In this SSL stage, $[\text{mCLS}]_{\mathbf{P}_i^j}$ from mViT$_{\mathbf{P}}$-C serve as inputs to ViT$_{\mathbf{R}}$-P. The resulting $\{[\text{mCLS}]_{\mathbf{P}_i^j}\}_{j=1}^{256}$ are rearranged into a $16 \times 16 \times 384$ feature grid, on which local-global patch-level crops of sizes $[6 \times 6]$ and $[14 \times 14]$ are applied to form augmented views for contrastive pre-training.

Stage 3: Multi-Resolution WSI-Level Pre-training. In this stage, both mViT$_{\mathbf{P}}$-C and ViT$_{\mathbf{R}}$-P are frozen and $[\text{mCLS}]_{\mathbf{R}_i}$ from ViT$_{\mathbf{R}}$-P are input to ViT$_{\mathbf{M}}$-R. The resulting $\{[\text{mCLS}]_{\mathbf{R}_i}\}_{i=1}^{m}$ are rearranged into a $\sqrt{m} \times \sqrt{m} \times 192$ feature grid (with $m \approx 64$ for most WSIs). Local and global crops of sizes $[6 \times 6]$ and $[7 \times 7]$ are used in SSL. This stage enables the model to capture holistic WSI-level contextual dependencies while preserving fine-grained semantic consistency across resolutions. *We assess the contribution of each stage through ablation studies (see Table 2).*

3.3 MRPT-LLaVA Model Pre-Training

MRPT-LLaVA follows a three-stage training pipeline [21, 25, 58, 77]: *cross-modal alignment*, *feature space alignment*, and *instruction tuning*. In Stage I, 9,642 WSI–report pairs from WSI-Bench [58] are aligned via contrastive learning using the MRPT encoder and the Qwen2-1.5B text encoder [94]. In Stage II, the MRPT encoder and LLM are frozen while the projection layer is updated. In Stage III, we fine-tune the projection layer and LLM on all WSI-Bench tasks. For patch-level VQA, mViT$_{\mathbf{P}}$-C and ViT$_{\mathbf{R}}$-P backbones are trained using the aforementioned approach on QuiltNet-1M [48] and QuiltInstruct [77] datasets.

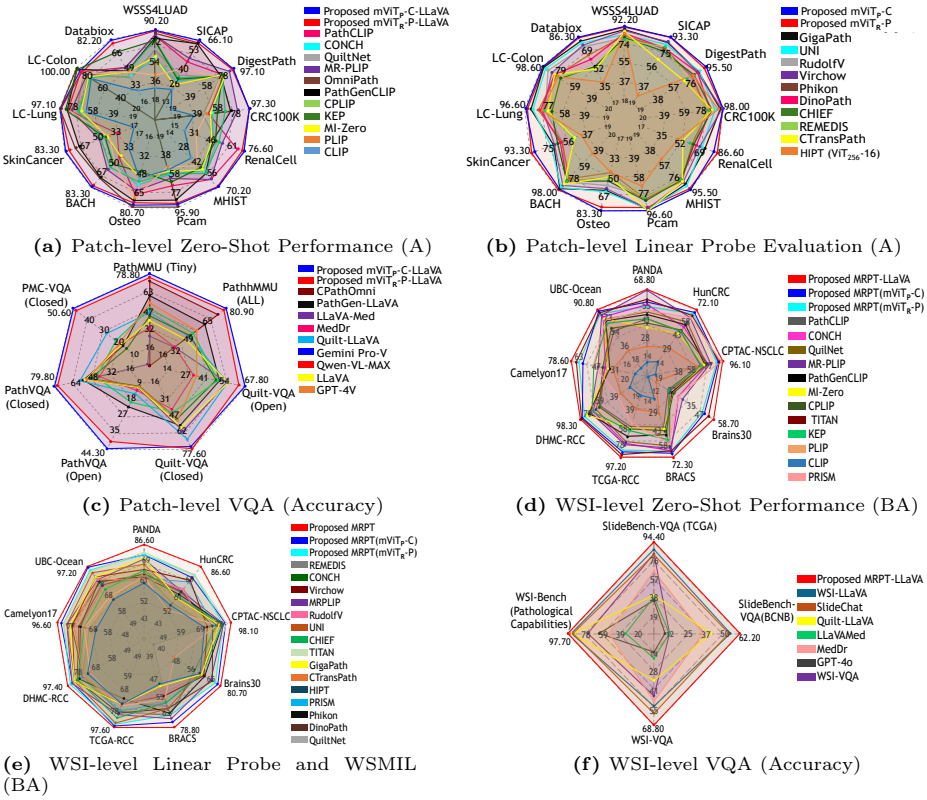
4 Experiments

Training Details: We pre-train **MRPT** on 36,000 WSIs in total, comprising 30,000 WSIs from diverse TCGA cancer types [46] and 6,000 WSIs from CPTAC [32]. **MRPT** is pre-trained on 624M patches using mViT$_{\mathbf{P}}$-C, 2.4M regions using ViT$_{\mathbf{R}}$-P, and 36K WSIs using ViT$_{\mathbf{M}}$-R. For mViT$_{\mathbf{P}}$-C, we employ three resolution-specific ViTs with $L = 2$, $M = 4$, and $N = 4$ encoder layers (see

Table 2: Performance of different hierarchical levels of the proposed MRPT model. Each variant combines the patch-level encoder mViT$_{\mathbf{P}}$-C with the region-level encoder ViT$_{\mathbf{R}}$-P and the WSI-level encoder ViT$_{\mathbf{M}}$-R. 'PF' denotes *Pre-trained and Frozen*, and 'LP' indicates *Linear Probe*. "Params" represent the number of learnable parameters during downstream fine-tuning task. Results are reported across three WSI benchmark datasets, comparing MRPT hierarchical variants (A–D) against the HIPT baseline [23].

Experiments	Models	Params	PANDA	BRAINS	UBC-OCEAN
A	HIPT (ViT$_{256}$-16$_{\text{PF}}$) [23]	0.505M	0.584	0.502	0.677
	mViT$_{\mathbf{P}}$-C-T$_{\text{PF}}$	0.505M	0.732	0.687	0.779
	mViT$_{\mathbf{P}}$-C-S$_{\text{PF}}$	0.505M	0.754	0.705	0.781
	mViT$_{\mathbf{P}}$-C-B$_{\text{PF}}$	0.505M	**0.773**	**0.723**	**0.811**
B	HIPT (ViT$_{256}$-16$_{\text{PF}}$, ViT$_{4096}$-256$_{\text{PF}}$) [23]	0.505M	0.623	0.567	0.723
	mViT$_{\mathbf{P}}$-C-T$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$	0.505M	0.773	0.704	0.855
	mViT$_{\mathbf{P}}$-C-S$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$	0.505M	0.768	0.712	0.842
	mViT$_{\mathbf{P}}$-C-B$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$	0.505M	**0.822**	**0.772**	**0.926**
C	HIPT (ViT$_{256}$-16$_{\text{PF}}$, ViT$_{4096}$-256$_{\text{PF}}$, ViT$_{\text{WSI}}$-4096) [23]	1.47M	0.576	0.514	0.653
	mViT$_{\mathbf{P}}$-C-T$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$, ViT$_{\mathbf{M}}$-R	1.47M	0.754	0.687	0.821
	mViT$_{\mathbf{P}}$-C-S$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$, ViT$_{\mathbf{M}}$-R	1.47M	0.767	0.706	0.827
	mViT$_{\mathbf{P}}$-C-B$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$, ViT$_{\mathbf{M}}$-R	1.47M	**0.807**	**0.740**	**0.892**
D	mViT$_{\mathbf{P}}$-C-T$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$, ViT$_{\mathbf{M}}$-R$_{\text{PF}}$ + LP	1.5K	0.786	0.728	0.854
	mViT$_{\mathbf{P}}$-C-S$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$, ViT$_{\mathbf{M}}$-R$_{\text{PF}}$ + LP	1.5K	0.803	0.720	0.862
	mViT$_{\mathbf{P}}$-C-B$_{\text{PF}}$, ViT$_{\mathbf{R}}$-P$_{\text{PF}}$, ViT$_{\mathbf{M}}$-R$_{\text{PF}}$ + LP	1.5K	**0.866**	**0.813**	**0.910**

Table 1), using an embedding dimension of 384 for ViT$_{10\times}$ and 768 for both ViT$_{20\times}$ and ViT$_{40\times}$. We use $K = 1$ CCRA layer and a total of $Q = 4$ blocks in mViT$_{\mathbf{P}}$-C. For ViT$_{\mathbf{R}}$-P, we set $n = 4$ encoders, $h = 6$ heads, and $d = 192$ embedding dimension; for ViT$_{\mathbf{M}}$-R, we use $n = 2$, $h = 3$, and $d = 192$, following HIPT [23]. We train mViT$_{\mathbf{P}}$-C for 250K iterations using AdamW [2] with batch size 256 and base learning rate 0.0007; the first 20 epochs are warm-up to the base rate, followed by cosine decay [62]. ViT$_{\mathbf{R}}$-P and ViT$_{\mathbf{M}}$-R are each trained for 200K iterations. In DINO [18], the teacher is updated via EMA of the student. For **MRPT-LLaVA**, stage 1 uses learning rate 0.001 and batch size 64; only a two-layer projection head is trained to align WSI-text features. We apply LoRA with rank 128 and $\alpha = 256$ for efficient tuning. All models are trained on four NVIDIA A100 GPUs. **Patch-level Tasks:** We evaluate patch-level classification and VQA across 17 datasets. Patch-level classification uses 13 independent datasets: WSSS4LUAD [37], SICAP [79], DigestPath [28], CRC100K [54], RenalCell [15], MHIST [92], PCam [88], Osteo [6], BACH [47], SkinCancer [55], LC-Lung [13], LC-Colon [13], and DataBiox [12], with both zero-shot and linear-probe settings. Zero-shot patch-level VQA is conducted on five datasets including QuiltVQA [77], PathVQA [41], PMC-VQA [98], and PathMMU [83]. **WSI-level Tasks:** We evaluate WSI-level classification, VQA, captioning, and report generation across 17 datasets. WSI-level classification uses 10 external datasets: PANDA [16], HunCRC [70], CPTAC-NSCLC [32], Brains30 [75], BRACS [14], TCGA-RCC [46], DHMC-RCC [100], Camelyon17 [10], and UBC-OCEAN [11], with zero-shot, WSMIL, and FSWL evaluations. WSI-level VQA and captioning are evaluated on SlideBench-Caption [26], SlideBench-VQA (TCGA) [26], SlideBench-VQA (BCNB) [26], WSI-VQA [22], and WSI-Bench [59]. WSI-level report generation uses HistGen [35] and WSI-Bench (Report) [59]. ***See suppl. mat. for more details, ablations, and results.***



(a) Patch-level Zero-Shot Performance (A)

(b) Patch-level Linear Probe Evaluation (A)

(c) Patch-level VQA (Accuracy)

(d) WSI-level Zero-Shot Performance (BA)

(e) WSI-level Linear Probe and WSMIL (BA)

(f) WSI-level VQA (Accuracy)

Fig. 3: Proposed mViT$_\mathbf{P}$-C-LLaVA, mViT$_\mathbf{R}$-P-LLaVA, mViT$_\mathbf{P}$-C, mViT$_\mathbf{R}$-P, MRPT, and MRPT-LLaVA outperform SOTA models.

4.1 Evaluation Setup

For classification task, we used Balanced Accuracy (BA), weighted F_1, and Accuracy (A) [24, 64]. For captioning and report generation: BLEU-2/4, ROUGE-L, and METEOR. For VQA: accuracy (closed-ended) and recall, where applicable. For classification, we compare against CLIP [73], PLIP [44], PathCLIP [85], HIPT [23], KEP [99], MI-Zero [66], UNI [24], CONCH [64], QuiltNet [48], CPLIP [51], MR-PLIP [4], OmniPath [82], CTransPath [90], REMEDIS [7], CHIEF [91], DinoPath [53], Phikon [33], PathGenCLIP [84], Virchow [89], GigaPath [93], TITAN [29], PRISM [78], and RudolfV [30]. Zero-shot prompts follow dataset-specific templates from CONCH [64] and QuiltNet [48]. For VQA, captioning, and report generation, we compare with general LMMs—GPT-4V [45], GPT-4o [45], LLaVA [60], Qwen-VL-Max [8], and Gemini Pro-V [86]—and CPath-specific MLLMs—Quilt-LLaVA [48], HistGen [35], MIGen [20], SlideChat [25], WSI-LLaVA [58], MedDr [40], LLaVA-Med [57], and PathGen-LLaVA [84]. We use official implementations, matched test splits, and consistent inference prompts [24].

Table 3: Performance comparison across different datasets with same training and testing splits.

Method	Camelyon16			Camelyon17			TCGA-NSCLC				TCGA-RCC				TCGA-BRCA			
	Train	Test	F_1	Train	Test	F_1	WSIs	Train	Test	F_1	WSIs	Train	Test	F_1	WSIs	Train	Test	F_1
Baseline HIPT	270	130	0.933	500	500	0.782	1041	80%	20%	0.851	1203	80%	20%	0.861	1048	80%	20%	0.866
ABMIL	270	130	0.886	500	500	0.766	1041	80%	20%	0.809	1203	80%	20%	0.831	1048	80%	20%	0.831
DSMIL	270	130	0.906	500	500	0.773	1041	80%	20%	0.867	1203	80%	20%	0.903	1048	80%	20%	0.871
TransMIL	270	130	0.924	500	500	0.796	1041	80%	20%	0.812	1203	80%	20%	0.911	1048	80%	20%	0.881
CLAM-MB	270	130	0.922	500	500	0.805	1041	80%	20%	0.844	1203	80%	20%	0.867	1048	80%	20%	0.873
Focus	270	130	0.902	500	500	0.802	1041	80%	20%	0.882	1203	80%	20%	0.882	1048	80%	20%	0.852
UNI	×	130	0.934	×	500	0.857	1041	×	20%	0.908	1203	×	20%	0.920	1048	×	20%	0.911
GigaPath	×	130	0.911	×	500	0.802	1041	×	20%	0.814	1203	×	20%	0.905	1048	×	20%	0.871
Vila-MIL	270	130	0.891	500	500	0.812	1041	80%	20%	0.852	1203	80%	20%	0.863	1048	80%	20%	0.876
SI-MIL	270	130	0.921	500	500	0.821	1041	80%	20%	0.820	1203	80%	20%	0.847	1048	80%	20%	0.837
MRPT (mViT$_{\mathbf{P}}$-C)	270	130	0.955	500	500	0.855	1041	80%	20%	0.908	1203	80%	20%	0.923	1048	80%	20%	**0.966**
MRPT (mViT$_{\mathbf{R}}$-P)	270	130	**0.971**	500	500	0.834	1041	80%	20%	0.918	1203	80%	20%	0.936	1048	80%	20%	0.951
MRPT	270	130	0.962	500	500	**0.883**	1041	80%	20%	**0.946**	1203	80%	20%	**0.956**	1048	80%	20%	0.964

Table 4: Component-wise ablation of MRPT model.

Variant	Δ Vs. Previous Row	Training Data	PANDA	BRAINS	UBC-OCEAN
(a) HIPT (20× only, 2-Stage Hier. SSL)	Baseline	36K WSIs	0.704	0.638	0.781
(b) + Multi-Resolution Inputs (Concat)	+ Multi-resolution data, MRPT, no CCRA	36K WSIs	0.762	0.701	0.834
(c) + CCRA (Consecutive Only)	+ CCRA Module	36K WSIs	0.823	0.776	0.882
(d) + ACT+[CLS] Concat	+ Token Fusion	36K WSIs	0.848	0.795	0.902
(e) + 3-Stage Hier. SSL (full)	+ Hierarchical SSL	36K WSIs	**0.866**	**0.813**	**0.910**

4.2 Ablation Studies

1. Hierarchical Variants of MRPT (Table 2). We analyze variants using mViT$_{\mathbf{P}}$-C alone and in combination with ViT$_{\mathbf{R}}$-P and ViT$_{\mathbf{M}}$-R. (A) Freeze mViT$_{\mathbf{P}}$-C and apply ABMIL [49] for WSI classification (no hierarchy). (B) Freeze mViT$_{\mathbf{P}}$-C and ViT$_{\mathbf{R}}$-P; use ABMIL for WSI classification. (C) Freeze mViT$_{\mathbf{P}}$-C and ViT$_{\mathbf{R}}$-P; fine-tune only ViT$_{\mathbf{M}}$-R. (D) Train a linear classifier on pre-extracted MRPT WSI-level features. ***Results:*** All MRPT variants outperform HIPT [23], validating multi-resolution hierarchical representations. mViT$_{\mathbf{P}}$-C-B yields the strongest results among tiny/small/backbone options; thus we report MRPT with mViT$_{\mathbf{P}}$-C-B thereafter. Experiment D averages 86.30%, exceeding B (84.0%) and C (81.30%), indicating richer WSI-level features from MRPT than from patch/region-only ViTs.

2. Multi-Resolution Cross-Attention (Table 5). We compare six fusion strategies: *[CLS] Concat* (concatenate all [CLS] tokens), *ACT* (average [CLS] tokens), *All-Token Attention* (self-attention over concatenated tokens across resolutions; strong but quadratic), *Pairwise Sum* (spatially aligned token-wise fusion with separate [CLS] fusion), *CCRA* (Consecutive Cross-Resolution Attention), and our combined *ACT+CCRA+[CLS] Concat*. *ACT+CCRA+[CLS] Concat* achieves the best accuracy, highlighting the complementary roles of averaging (Eq. (2)) and [CLS] concatenation within CCRA. **See more ablations in supplementary material.**

3. Disentangling Data Scale and MRPT Contributions (Table 3). To isolate data scale effects, we conduct a matched-data experiment shown in Table 3 where all WSI classification methods are trained and tested under identical splits. MRPT consistently outperforms baselines by a significant margin,

Table 5: Ablation of MRPT (Model D in Table 2) across cross-resolution token fusion methods.The combined ACT + CCRA + [CLS] Concat achieves the best accuracy.

Token Fusion Method	PANDA	BRAINS	UBC-OCEAN
[CLS] Concat	0.823	0.745	0.846
Average [CLS] Token (ACT)	0.848	0.789	0.882
All-Token Attention	0.850	0.795	0.902
Pairwise Sum	0.843	0.765	0.861
Consecutive Cross-Resolution Attention (CCRA)	0.862	0.810	0.907
ACT+CCRA+[CLS] Concat	**0.866**	**0.813**	**0.910**

Table 6: Report generation performance comparison across SOTA methods on the WSI-Bench and HistGen datasets.

Models	BLEU-2		BLEU-4		ROUGE-L		METEOR	
WSICaption	0.111	0.137	0.136	0.540	0.402	0.251	0.366	0.322
HistGen	0.307	0.297	0.208	0.184	0.448	0.344	0.416	0.182
GPT-4O	0.069	0.420	0.016	0.100	0.132	0.128	0.167	0.144
MIGen	0.306	0.466	0.209	0.234	0.446	0.322	0.407	0.412
Quilt-LLaVA	0.351	0.331	0.236	0.488	0.475	0.491	0.460	0.443
SlideChat	0.310	0.533	0.191	0.655	0.463	0.492	0.422	0.613
WSI-LLaVA	0.358	0.527	0.240	0.644	0.490	0.544	0.465	0.655
MRPT-LLaVA	**0.401**	**0.634**	**0.287**	**0.699**	**0.522**	**0.581**	**0.511**	**0.701**

showing gains stem from hierarchical multi-resolution SSL and cross-resolution alignment, not pre-training scale.

4. Component-wise Contribution Analysis (4). MRPT component-wise ablation is presented in Table 4, showing consistent gains from multi-resolution inputs, CCRA, token fusion, and 3-stage SSL, confirming improvements arise from structural design, not a coupled pipeline.

4.3 Main Results and Comparisons

1. Patch-level Zero-Shot Classification (Fig. 3a). We evaluate mViT$_{\mathbf{P}}$-C-LLaVA and mViT$_{\mathbf{R}}$-P-LLaVA against 11 SOTA VLMs on 13 datasets. *Across most datasets, our models achieve notably higher accuracy.* On average, mViT$_{\mathbf{P}}$-C-LLaVA attains 83.66% and mViT$_{\mathbf{R}}$-P-LLaVA 82.11%; the third-best, PathGen-CLIP, reaches 73.60%. *Thus, our patch-level model improves by ∼10.0% over the existing SOTA.* **2. Patch-level Linear Probe Evaluations (Fig. 3b).** With linear classifiers atop features from mViT$_{\mathbf{P}}$-C and mViT$_{\mathbf{R}}$-P, we compare against 10 SOTA VFMs. *Averages: 92.75% for mViT$_{\mathbf{P}}$-C and 92.22% for mViT$_{\mathbf{R}}$-P; the next best, GigaPath, achieves 86.12%.* This yields a 6.63% average gain for our patch encoder. **3. Patch-level Zero-shot VQA (Fig. 3c).** We compare our mViT$_{\mathbf{P}}$-C-LLaVA and mViT$_{\mathbf{R}}$-P-LLaVA with nine MLLMs. *On five closed-ended datasets: 73.12% (patch) and 71.52% (region) vs. 49.48% (PathGen-LLaVA) and 48.60% (Quilt-LLaVA). On two open-ended datasets: 56.05% (patch) and 51.95% (region) vs. 35.80% (PathGen-LLaVA).* Overall improvements average 23.64% (closed) and 20.25% (open), emphasizing our multi-resolution design. **4. Zero-Shot WSI Classification (Fig. 3d).** We compare MRPT-LLaVA with 12 CPath SOTA VLMs on 10 datasets. *Average balanced accuracy: 0.806*

for MRPT-LLaVA vs. 0.751 for the next-best TITAN, showcasing the benefits of multi-resolution WSI representations, whereas PRISM/TITAN rely on single-resolution pre-training. **5. WSI Classification Using WSMIL (Fig. 3e).** Following [24], we apply ABMIL [49] atop features from mViT$_{\mathbf{P}}$-C and mViT$_{\mathbf{R}}$-P. *Compared to 10 SOTA patch-level methods, our encoders achieve mean balanced accuracies of 0.863 and 0.851 across 10 datasets; MR-PLIP and UNI reach 0.812 and 0.798, respectively.* **6. WSI Linear Probe (Fig. 3e).** Using MRPT's WSI-level features with a linear classifier, we compare against PRISM, TITAN, GigaPath, and CHIEF. *MRPT averages 0.898 balanced accuracy vs. 0.845 for TITAN, underscoring the advantage of multi-resolution cues.* **7. WSI VQA (Fig. 3f).** Across four VQA datasets, we compare with seven CPath MLLMs. *MRPT-LLaVA averages 80.77% accuracy vs. 72.57% for WSI-LLaVA, indicating substantial gains from aligning MRPT with the LLM.* **8. WSI Report Generation (Table 6).** For report generation (WSI-Bench, HistGen), MRPT-LLaVA outperforms seven recent methods, including SlideChat and WSI-LLaVA (Tables 6), *evidencing strong generalization of multi-resolution features with LLMs.*

5 Conclusion

We introduced a multi-resolution hierarchical WSI foundation model (MRPT) that integrates information across resolutions to enhance diverse CPath tasks, including WSI classification, VQA, captioning, and report generation. At its core, MRPT employs a *Consecutive Cross-Resolution Attention (CCRA)* mechanism to fuse information across adjacent resolutions. A multi-resolution ViT captures cell-level features at varying granularities using CCRA, aggregates them at the patch level, and progressively encodes region and WSI-level representations. Each hierarchy is pre-trained using multi-resolution SSL. MRPT jointly models global tissue architecture and fine-grained cellular morphology, yielding substantial improvements over SOTA. Its biologically inspired, multi-resolution design reflects the diagnostic process of pathologists—integrating contextual and microscopic cues. Building on MRPT, we further proposed *MRPT-LLaVA*, a MLLM for patch and WSI-level reasoning. Extensive evaluations confirm that our models surpass single-resolution and existing MLLM baselines, achieving superior generalization. Future extensions incorporating multi-modal data, such as genomics and radiology, promise to advance holistic patient-level understanding.

6 Acknowledgment

This research was funded by Khalifa University of Science and Technology through the Faculty Start-Ups under Project ID: KU-INT-FSU-2005-8474000775.

References

1. Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., Van der Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., et al.: Computational pathology definitions, best practices, and recommendations for regulatory

guidance: a white paper from the digital pathology association. The Journal of pathology **249**(3), 286–294 (2019) 1
2. Adam, K.D.B.J., et al.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 **1412** (2014) 11
3. Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L.: A survey on graph-based deep learning for computational histopathology. Computerized Medical Imaging and Graphics **95**, 102027 (2022) 1
4. Albastaki, S., Sohail, A., Ganapathi, I.I., Alawode, B., Khan, A., Javed, S., Werghi, N., Bennamoun, M., Mahmood, A.: Multi-resolution pathology-language pre-training model with text-guided visual representation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25907–25919 (2025) 1, 2, 4, 5, 12
5. Ardon, O., Manzo, A., Spencer, J., Reuter, V.E., Hameed, M., Hanna, M.G.: Digital slide scanning at scale: Comparison of whole slide imaging devices in a clinical setting. Journal of Pathology Informatics p. 100446 (2025) 2
6. Arunachalam, H.B., Mishra, R., Daescu, O., Cederberg, K., Rakheja, D., Sengupta, A., Leonard, D., Hallac, R., Leavey, P.: Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. PloS one **14**(4), e0210706 (2019) 11
7. Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., MacWilliams, P., Mahdavi, S.S., Wulczyn, E., et al.: Robust and efficient medical imaging with self-supervision. arXiv preprint arXiv:2205.09723 (2022) 4, 12
8. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 12
9. Baidoshvili, A., Khacheishvili, M., van der Laak, J.A., van Diest, P.J.: A whole-slide imaging based workflow reduces the reading time of pathologists. Pathology International **73**(3), 127–134 (2023) 2
10. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018) 11
11. Bashashati, A., Farahani, H., Consortium, O., Karnezis, A., Akbari, A., Kim, S., Chow, A., Dane, S., Zhang, A., Asadi, M.: Ubc ovarian cancer subtype classification and outlier detection (ubc-ocean) (2023), https://kaggle.com/competitions/UBC-OCEAN 11
12. Bolhasani, H., Amjadi, E., Tabatabaeian, M., Jassbi, S.J.: A histopathological image dataset for grading breast invasive ductal carcinomas. Informatics in Medicine Unlocked **19**, 100341 (2020) 11
13. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142v1 (2019), https://arxiv.org/abs/1912.12142 11
14. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022) 11
15. Brummer, O., Pölönen, P., Mustjoki, S., Brück, O.: Integrative analysis of histological textures and lymphocyte infiltration in renal cell carcinoma using deep learning. bioRxiv pp. 2022–08 (2022) 11

16. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M.B., Evans, A.J., van der Kwast, T., Allan, R., Humphrey, P.A., Grönberg, H., Samaratunga, H., the PANDA challenge consortium: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature Medicine **28**, 154–163 (2022). https://doi.org/10.1038/s41591-021-01620-2, https://doi.org/10.1038/s41591-021-01620-2 11
17. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020) 4
18. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 11
19. Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F.: Fast and scalable search of whole-slide images via self-supervised deep learning. Nature Biomedical Engineering **6**(12), 1420–1434 (2022) 2
20. Chen, P., Li, H., Zhu, C., Zheng, S., Shui, Z., Yang, L.: Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 546–556. Springer (2024) 12
21. Chen, P., Zhu, C., Zheng, S., Li, H., Yang, L.: Wsi-vqa: Interpreting whole slide images by generative visual question answering. In: European Conference on Computer Vision. pp. 401–417. Springer (2025) 10
22. Chen, P., Zhu, C., Zheng, S., Li, H., Yang, L.: Wsi-vqa: Interpreting whole slide images by generative visual question answering. In: European Conference on Computer Vision (ECCV) 2024. pp. 401–417 (2025). https://doi.org/10.1007/978-3-031-72764-1_23, https://doi.org/10.1007/978-3-031-72764-1_23 11
23. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022) 1, 2, 3, 4, 6, 11, 12, 13
24. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024) 2, 4, 8, 12, 15
25. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Hu, M., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5134–5143 (2025) 2, 3, 5, 10, 12
26. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Zhang, B., Pei, N., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5134–5143 (June 2025), https://openaccess.thecvf.com/content/CVPR2025/html/Chen_SlideChat_A_Large_Vision-Language_Assistant_for_Whole-Slide_Pathology_Image_Understanding_CVPR_2025_paper.html 11
27. Cui, M., Zhang, D.Y.: Artificial intelligence and computational pathology. Laboratory Investigation **101**(4), 412–422 (2021) 1

28. Da, Q., Huang, X., Li, Z., Zuo, Y., Zhang, C., Liu, J., Chen, W., Li, J., Xu, D., Hu, Z., et al.: Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Medical Image Analysis **80**, 102485 (2022) 11
29. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666 (2024) 2, 3, 5, 12
30. Dippel, J., Feulner, B., Winterhoff, T., Milbich, T., Tietz, S., Schallenberg, S., Dernbach, G., Kunft, A., Heinke, S., Eich, M.L., et al.: Rudolfv: a foundation model by pathologists for pathologists. arXiv preprint arXiv:2401.04079 (2024) 4, 12
31. Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T., Kather, J.N.: Deep learning in cancer pathology: a new generation of clinical biomarkers. British journal of cancer **124**(4), 686–696 (2021) 1
32. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The cptac data portal: A resource for cancer proteomics research (2015). https://doi.org/10.1021/pr501254j, https://pubs.acs.org/doi/abs/10.1021/pr501254j 3, 10, 11
33. Filiot, A., Jacob, P., Mac Kain, A., Saillard, C.: Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 (2024) 12
34. Ghezloo, F., Wang, P.C., Kerr, K.F., Brunyé, T.T., Drew, T., Chang, O.H., Reisch, L.M., Shapiro, L.G., Elmore, J.G.: An analysis of pathologists' viewing processes as they diagnose whole slide digital images. Journal of Pathology Informatics **13**, 100104 (2022) 2
35. Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H.: Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–199. Springer (2024) 11, 12
36. Guo, Z., Zhao, W., Wang, S., Yu, L.: 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) 3, 4
37. Han, C., Pan, X., Yan, L., Lin, H., Li, B., Yao, S., Lv, S., Shi, Z., Mai, J., Lin, J., et al.: Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. arXiv preprint arXiv:2204.06455 (2022) 11
38. Hanna, M.G., Parwani, A., Sirintrapun, S.J.: Whole slide imaging: technology and applications. Advances in anatomic pathology **27**(4), 251–259 (2020) 2
39. Harada, S., Morlote, D.: Molecular pathology of colorectal cancer. Advances in anatomic pathology **27**(1), 20–26 (2020) 2
40. He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. CoRR (2024) 12
41. He, X., Zhang, Y., Mou, L., Xing, E.P., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020), https://arxiv.org/abs/2003.10286 11
42. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., et al.: Cellvit: Vision transformers for precise cell segmentation and classification. Medical Image Analysis **94**, 103143 (2024) 4

43. Hosseini, M.S., Bejnordi, B.E., Trinh, V.Q.H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., et al.: Computational pathology: a survey review and the way forward. Journal of Pathology Informatics p. 100357 (2024) 1
44. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023) 4, 12
45. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 12
46. Hutter, C., Zenklusen, J.C.: The cancer genome atlas: creating lasting value beyond its data. Cell **173**(2), 283–285 (2018) 3, 10, 11
47. ICIAR, B.: Grand challenge on breast cancer histology images. 2018 (2018) 11
48. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36** (2024) 2, 4, 10, 12
49. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018) 13, 15
50. Jain, E., Patel, A., Parwani, A.V., Shafi, S., Brar, Z., Sharma, S., Mohanty, S.K.: Whole slide imaging technology and its applications: Current and emerging perspectives. International journal of surgical pathology **32**(3), 433–448 (2024) 2
51. Javed, S., Mahmood, A., Ganapathi, I.I., Dharejo, F.A., Werghi, N., Bennamoun, M.: Cplip: Zero-shot learning for histopathology with comprehensive vision-language alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11450–11459 (2024) 4, 12
52. Jin, C., Luo, L., Lin, H., Hou, J., Chen, H.: Hmil: Hierarchical multi-instance learning for fine-grained whole slide image classification. IEEE Transactions on Medical Imaging (2024) 3, 4
53. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023) 4, 12
54. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue (2018). https://doi.org/10.5281/zenodo.1214456, https://doi.org/10.5281/zenodo.1214456 11
55. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Janssen, C., Meliss, R.R., Muley, T., Sack, U., Steinbuss, G., Kriegsmann, M.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. Frontiers in Oncology **12**, 1022967 (2022) 11
56. Van der Laak, J., Litjens, G., Ciompi, F.: Deep learning in histopathology: the path to the clinic. Nature medicine **27**(5), 775–784 (2021) 1
57. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023) 12
58. Liang, Y., Lyu, X., Chen, W., Ding, M., Zhang, J., He, X., Wu, S., Xing, X., Yang, S., Wang, X., Shen, L.: Wsi-llava: A multimodal large language model for whole slide image (2025), https://arxiv.org/abs/2412.02141 3, 4, 5, 10, 12

59. Liang, Y., Lyu, X., Ding, M., Chen, W., Zhang, J., Ren, Y., He, X., Wu, S., Yang, S., Wang, X., Xing, X., Shen, L.: Wsi-llava: A multimodal large language model for whole slide image. arXiv preprint arXiv:2412.02141 (2024), https://arxiv.org/abs/2412.02141 11
60. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024) 12
61. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 3
62. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 11
63. Louis, D.N., Feldman, M., Carter, A.B., Dighe, A.S., Pfeifer, J.D., Bry, L., Almeida, J.S., Saltz, J., Braun, J., Tomaszewski, J.E., et al.: Computational pathology: a path ahead. Archives of pathology & laboratory medicine **140**(1), 41–50 (2016) 1
64. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024) 2, 4, 12
65. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al.: A multimodal generative ai copilot for human pathology. Nature pp. 1–3 (2024) 5
66. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19764–19775 (2023) 4, 12
67. Lu, W., Graham, S., Bilal, M., Rajpoot, N., Minhas, F.: Capturing cellular topology in multi-gigapixel pathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 260–261 (2020) 2
68. McGuire, S.: World cancer report 2014. geneva, switzerland: World health organization, international agency for research on cancer, who press, 2015. Advances in nutrition **7**(2), 418 (2016) 1
69. Nan, T., Zheng, S., Qiao, S., Quan, H., Gao, X., Niu, J., Zheng, B., Guo, C., Zhang, Y., Wang, X., et al.: Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. Nature Communications **16**(1), 5493 (2025) 2
70. Pataki, B.Á., Olar, A., Ribli, D., Pesti, A., Kontsek, E., Gyöngyösi, B., Bilecz, Á., Kovács, T., Kovács, K.A., Kramer, Z., et al.: Huncrc: annotated pathological slides to enhance deep learning applications in colorectal cancer screening. Scientific Data **9**(1), 370 (2022) 11
71. Petersen, P.E.: Oral cancer prevention and control–the approach of the world health organization. Oral oncology **45**(4-5), 454–460 (2009) 1
72. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. Advances in Neural Information Processing Systems **36** (2024) 1
73. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4, 12

74. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., Ciompi, F.: Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Medical Image Analysis **68**, 101890 (2021). https://doi.org/10.1016/j.media.2020.101890, https://www.sciencedirect.com/science/article/pii/S1361841520302541 1, 4
75. Roetzer-Pejrimovsky, T., Moser, A.C., Atli, B., Vogel, C.C., Mercea, P.A., Prihoda, R., Gelpi, E., Haberler, C., Höftberger, R., Hainfellner, J.A., et al.: The digital brain tumour atlas, an open histopathology resource. Scientific Data **9**(1), 55 (2022) 11
76. Schüffler, P.J., Stamelos, E., Ahmed, I., Yarlagadda, D.V.K., Ardon, O., Hanna, M.G., Reuter, V.E., Klimstra, D.S., Hameed, M.: Efficient visualization of whole slide images in web-based viewers for digital pathology. Archives of pathology & laboratory medicine **146**(10), 1273–1280 (2022) 2
77. Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13183–13192 (2024) 3, 5, 10, 11
78. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al.: Prism: A multimodal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024) 12
79. Silva-Rodriguez, J., Colomer, A., Dolz, J., Naranjo, V.: Self-learning for weakly supervised gleason grading of local patterns. IEEE journal of biomedical and health informatics **25**(8), 3094–3104 (2021) 11
80. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023) 1, 2
81. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical image analysis **67**, 101813 (2021) 1, 2
82. Sun, Y., Si, Y., Zhu, C., Gong, X., Zhang, K., Chen, P., Zhang, Y., Shui, Z., Lin, T., Yang, L.: Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10360–10371 (2025) 12
83. Sun, Y., Wu, H., Zhu, C., Zheng, S., Chen, Q., Zhang, K., Yunlong, Z., Lan, X., Zheng, M., Li, J., Lyu, X., Lin, T., Yang, L.: Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology (2024), https://arxiv.org/abs/2401.16355 11
84. Sun, Y., Zhang, Y., Si, Y., Zhu, C., Shui, Z., Zhang, K., Li, J., Lyu, X., Lin, T., Yang, L.: Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. arXiv preprint arXiv:2407.00203 (2024) 12
85. 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) 5, 12
86. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 12
87. Tepus, M., Yau, T.O.: Non-invasive colorectal cancer screening: an overview. Gastrointestinal tumors **7**(3), 62–73 (2020) 2

88. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 210–218. Springer (2018) 11
89. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine pp. 1–12 (2024) 4, 12
90. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022) 2, 12
91. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature pp. 1–9 (2024) 4, 12
92. Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., et al.: A petri dish for histopathology image analysis. In: Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings. pp. 11–24. Springer (2021) 11
93. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024) 3, 4, 8, 12
94. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al., G.D.: Qwen2 technical report. Technical Report arXiv:2407.10671, CoRR, arXiv (2024), https://arxiv.org/abs/2407.10671 10
95. Zarella, M.D., Bowman, D., Aeffner, F., Farahani, N., Xthona, A., Absar, S.F., Parwani, A., Bui, M., Hartman, D.J.: A practical guide to whole slide imaging: a white paper from the digital pathology association. Archives of pathology & laboratory medicine **143**(2), 222–234 (2019) 2
96. Zarella, M.D., Rivera Alvarez, K.: High-throughput whole-slide scanning to enable large-scale data repository building. The Journal of Pathology **257**(4), 383–390 (2022) 2
97. Zech, D.F., Grond, S., Lynch, J., Hertel, D., Lehmann, K.A.: Validation of world health organization guidelines for cancer pain relief: a 10-year prospective study. Pain **63**(1), 65–76 (1995) 1
98. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023), https://arxiv.org/abs/2305.10415 11
99. Zhou, X., Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pretraining for computational pathology. In: European Conference on Computer Vision. pp. 345–362. Springer (2024) 12
100. Zhu, M., Ren, B., Richards, R., Suriawinata, M., Tomita, N., Hassanpour, S.: Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. Scientific reports **11**(1), 7080 (2021) 11