Multi-Modal Foundation Models for Computational Pathology: A Survey

Dong Li*
Baylor University
Waco, TX, USA
dong_li1@baylor.edu

Guihong Wan*
Harvard Medical School
Cambridge, MA, USA
guihong_wan@hsph.harvard.edu

Xintao Wu
University of Arkansas
Fayetteville, AR, USA
xintaowu@uark.edu

Xinyu Wu
Baylor University
Waco, TX, USA
xinyu_wu1@baylor.edu

Xiaohui Chen
Baylor University
Waco, TX, USA
xiaohui_chen1@baylor.edu

Yi He
William & Mary
Williamsburg, VA, USA
yihe@wm.edu

Zhong Chen
Southern Illinois University
Carbondale, IL, USA
zhong.chen@cs.siu.edu

Ninghui Hao
Harvard Medical School
Cambridge, MA, USA
nhao1@mgh.harvard.edu

Chen Zhao
Baylor University
Waco, TX, USA
chen_zhao@baylor.edu

Abstract

Foundation models have become a key paradigm in computational pathology (CPath), enabling scalable and generalizable analysis of histopathological images. Early work centered on uni-modal models trained solely on visual data, but recent advances highlight the potential of multi-modal approaches that integrate textual reports, structured knowledge, and molecular profiles. In this survey, we review 32 multi-modal foundation models built primarily on hematoxylin and eosin (H&E) whole-slide images (WSIs) and tile-level representations, categorizing them into vision–language, vision–knowledge graph, and vision–gene expression paradigms, with vision–language models further divided into non-LLM- and LLM-based variants. We also analyze 28 datasets, grouped into image–text pairs, instruction datasets, and image–other modality pairs, and summarize downstream tasks, training and evaluation strategies, and future challenges. This survey provides a comprehensive resource for advancing AI in pathology.

1 Introduction

The advent of foundation models has transformed computational pathology (CPath), enabling scalable analysis of histopathological images for improved diagnosis, prognosis, and biomarker discovery [53]. H&E-stained whole-slide images (WSIs) remain the most common modality, providing rich tissue morphology but requiring tiling into patches for computation [10, 29, 74, 22, 15]. Early uni-modal models [73, 14, 71] advanced classification, segmentation, and prediction by learning visual features, but their reliance on image-only data limited interpretability. Recent work therefore focuses on multi-modal models [48, 74, 49], which integrate pathology reports, knowledge graphs, and molecular profiles to capture richer context and deliver more clinically relevant insights. A detailed discussion of related background is provided in Appendix Section A.

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

Existing multi-modal foundation models for CPath (MMFM4CPath) fall into three paradigms: vision–language, vision–knowledge graph, and vision–gene expression. A roadmap of up-to-date MMFM4CPath is shown in Figure 1. Vision–language models [34, 37, 65, 62] use textual annotations such as WSI reports and captions to enrich visual features, supporting zero-shot learning and cross-modal integration; they are further divided into non-LLM- and LLM-based variants, with the latter leveraging LLMs for stronger language understanding and generation. Vision–knowledge graph models [90, 89] incorporate structured pathology ontologies to guide learning, while vision–gene expression models [78, 70] align image features with RNA and other omics data to capture genotype–phenotype associations.

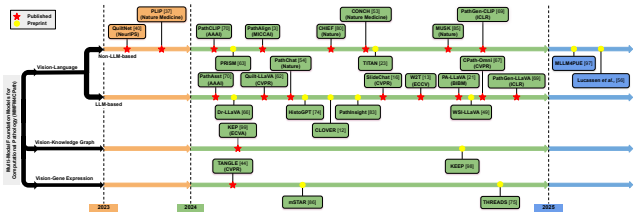

Figure 1: A roadmap of MMFM4CPath.

While existing surveys have explored FM4CPath [53, 10, 29, 8, 43], they often lack a comprehensive analysis tailored to multi-modal approaches. As shown in Table 1, our survey differentiates itself by systematically categorizing 32 of the most up-to-date MMFM4CPath and analyzing 28 available multi-modal datasets for pathology, with an emphasis on modalities beyond vision-language integration. Additionally, we provide an in-depth discussion on evaluation methodologies, training strategies, and emerging challenges in this field. The key contributions of this survey include:

Table 1: Comparison with related surveys.

Survey	# MMFM4CPath					# Datasets for MMFM4CPath				Tasks Taxo-nomy
	V-L		V-KG	V-GE	Total	Image-Text Pair	Instr-uction	I-OM	Total	
	Non-LLM	LLM								
Ochi *et, al.* [53]	4	✗	✗	1	5	4	✗	1	5	✓
Chanda *et, al.* [10]	7	4	1	✗	12	8	6	✗	14	✗
Guan *et, al.* [29]	3	11	✗	1	14	8	6	✗	14	✗
Bilal *et, al.* [8]	8	4	✗	2	14	✗	✗	✗	✗	✓
Li *et, al.* [43]	8	✗	2	✗	10	12	✗	2	14	✓
This Survey	13	14	2	3	32	12	12	4	28	✓

- **Comprehensive and Up-to-Date Survey.** We review 32 multi-modal foundation models in computational pathology across vision-language, vision-knowledge graph, and vision-gene expression paradigms, offering broader and more current coverage of architectures, pretraining strategies, and adaptations than prior surveys.
- **Analysis of Pathology-Specific Multi-Modal Datasets.** We curate 28 datasets and categorize them into three types: image-text pairs, multi-modal instructions, and image-other modality pairs, highlighting their roles in modality alignment and instruction tuning.
- **Overview of Multi-Modal Evaluation Tasks.** We provide a taxonomy of six evaluation categories—classification, retrieval, generation, segmentation, prediction, and VQA—and summarize how different MMFM4CPath are assessed under various settings.
- **Future Research Opportunities.** We outline three directions: integrating H&E with spatial omics, predicting MxIF markers from H&E for virtual staining, and establishing standardized benchmarks for fair comparison, aiming to improve clinical relevance and scalability.

2 Multi-Modal Foundation Models for CPath

2.1 Non-LLM-Based Vision-Language FM4CPath

Vision–language FM4CPath enhance pathology image understanding by aligning image–text pairs with SSL frameworks such as CLIP [54] and CoCa [81], supporting robust visual features, zero-shot learning, and cross-modal tasks. They typically pair a vision encoder with joint visual–language pretraining [91] and increasingly use LLMs or V-LLMs, either fine-tuned as text encoders or directly for generation and conversation.

CLIP-based Vision-Language FM4CPath. The success of CLIP on natural images has motivated its adaptation to CPath. PLIP [34], PathCLIP [65], and QuiltNet [37] fine-tune CLIP on tile–caption datasets, while CHIEF [74] combines a CPath-pretrained image encoder with CLIP's text encoder and a weakly supervised aggregator for WSI-level representations. Moving beyond off-the-shelf encoders, PathGen-CLIP [20] uses V-LLMs to generate high-quality captions and trains CLIP from scratch before fine-tuning on public datasets, whereas PathAlign-R [3] trains a CLIP framework from

scratch directly at the WSI-level. More recently, MLLM4PUE [88] employs V-LLMs as backbones to unify image and text into universal multi-modal embeddings for pathology.

CoCa-based Vision-Language FM4CPath. CoCa's multi-modal decoder bridges visual and linguistic features by transforming image embeddings into text-aware representations, thereby improving cross-modal integration for MMFM4CPath. Building on this, CONCH [48], PRISM [58], and Lucassen *et al.* [51] pre-train image encoders on pathology datasets before joint vision–language training, with PRISM and Lucassen extending to WSI-level via Perceiver[40] and clinical reports. MUSK [77] separately trains image and text encoders with BEiT-3 [72] and MIM [31], then aligns them under CoCa. TITAN [22] introduces a three-stage WSI framework: iBOT pre-training with positional encoding, CoCa-based refinement combining tile- and WSI-level features, and pathology-specialized V-LLM–generated captions and reports.

Other Vision-Language FM4CPath. Unlike previous methods that use CLIP or CoCa framework, PathAlign-G [3] first pre-trains a ViT-S using Masked Siamese Networks (MSN) [6], and then fine-tunes the model using the BLIP-2 framework. This enables PathAlign-G to utilize a shared pathology image-text embedding space, enhancing its cross-modal capabilities and making it more suitable for generative tasks.

2.2 LLM-Based Vision-Language FM4CPath

The fusion of vision and language allows MMFM4CPath to align pathology images with language signals, enabling LLMs to gain pathology knowledge and serve as generative assistants [49]. These models pair a pre-trained image encoder with an LLM via a lightweight alignment module and fine-tune the LLM using supervised or self-supervised objectives, which fall into instruction-based or non-instruction-based approaches.

Instruction-Based V-LLMs for CPath. Most visual LLMs for CPath are instruction tuned on curated datasets to adapt general LLMs and enhance cross-modal understanding. PathAsst [65] aligns a PathCLIP encoder with an LLM via QA instructions and light fine-tuning. Quilt-LLaVA [57] and PA-LLaVA [20] use public instruction sets, while PathChat [49] employs a broader corpus. SlideChat [15] and WSI-LLaVA [45] scale to gigapixel WSIs with WSI-level instructions. PathInsight [75] directly tunes existing VLLMs without separate encoders. Dr-LLaVA [61] combines instruction tuning with reinforcement learning for clinically valid multi-turn responses. CLOVER [11] improves efficiency with BLIP-2 and a lightweight Q-Former, freezing encoders and using GPT-3.5 plus template instructions. CPath-Omni [62] unifies tile- and WSI-level processing with four training stages over tile-caption, tile-instruction, and WSI-instruction datasets to enable generation and conversation.

Non-Instruction-Based V-LLMs for CPath. Non-instruction-based V-LLMs for CPath focus on generation tasks without explicit instruction tuning. PathGen-LLaVA [64] trains a CLIP model from scratch on tile–caption pairs, adds a fully connected layer to align image and text features, and uses supervised image captioning to generate pathology descriptions. W2T [12] combines four frozen visual and three text extractors, training with next-word prediction (NWP) on its WSI-VQA dataset for generative WSI question answering. HistoGPT [69] provides small, medium, and large variants: the smaller models use a Perceiver WSI encoder with MIL and NWP fine-tuning, while the large version applies a GNN to capture WSI-level positional information. HistoGPT supports multi-image report generation and incorporates prompts for expert knowledge guidance.

2.3 Enhancing FM4CPath with Other Modalities

Pathology-specific datasets are often small, noisy, and sourced from heterogeneous origins such as websites or videos [34, 37], leading to unstructured data that lacks domain knowledge. In contrast, large-scale multi-modal resources aligned with clinical practice, including gene expression profiles, remain underutilized for pretraining. To address this, recent studies explore incorporating modalities beyond vision and language to strengthen training signals.

Vision-Knowledge Graph FM4CPath. To integrate structured domain-specific knowledge, KEP [90] constructs a pathology knowledge graph and encodes it using a knowledge encoder, which then guides vision-language pretraining. They design a pathology knowledge encoding (PKE) method to align semantic groups in the latent space for training the knowledge encoder. Similarly, KEEP [89] builds a disease knowledge graph for encoding and employs knowledge-guided dataset structuring to

generate tile-caption pairs for pretraining within the CLIP framework, incorporating strategies such as positive mining, hardest negative sampling, and false negative elimination.

Vision-Gene Expression FM4CPath. Gene expression profiles provide WSI-level molecular insights that complement morphological features and capture biologically significant details. TANGLE [41] aligns WSIs with RNA sequences encoded by an MLP using contrastive loss, extending training to both human and rat tissues. THREADS [70] similarly leverages sequencing data but integrates both WSI–RNA and WSI–DNA pairs. mSTAR [78] incorporates WSIs, reports, and gene expression into an extended CLIP framework, applying inter-modality and inter-cancer contrastive learning and using self-distillation to transfer multi-modal knowledge to the patch extractor.

3 Multi-Modal Datasets for CPath

Larger, more diverse, and higher-quality datasets are key to the success of FM4CPath [71, 91], yet curating pathology-specific public datasets remains challenging. Numerous well-designed datasets have been introduced to address pathology tasks and advance CPath. We summarize existing multi-modal datasets, highlighting those of higher quality or proven utility, and categorize them into three groups in Appendix Table 3.

Image-Text Pair Datasets for CPath. Image–text datasets in CPath include tile–caption and WSI–report pairs, which support contrastive pretraining to enrich image embeddings and enable zero-shot and cross-modal tasks. Because expert annotations are costly and institutions prefer in-house data, many datasets are built from online sources, books, and educational resources, such as QUILT[37], OPENPATH[34], ARCH[24], and MI-ZERO[50]. Standardized pipelines filter non-pathology images, segment sub-figures, refine captions with LLMs, and align figures with text; for instance, QUILT also applies speech recognition to extract text from videos. Other datasets expand existing resources or rely on internal data to improve scale and diversity [65, 48, 64, 62, 20]. ARCH uniquely frames multiple-instance captioning by assigning one caption per image bag, while datasets such as PATHGEN[64], HISTGEN[30], and MASS-340K [22] generate WSI–report pairs using generative models or LLM-based processing.

Multi-Modal Instruction Datasets for CPath. Multi-modal instruction datasets train LLM-based vision–language FM4CPath as AI assistants in pathology. Since manual instruction design is costly, most rely on LLMs for scalable generation, with visual question answering (VQA) being the dominant form, covering both closed- and open-ended Q&A to build conversational ability. Different datasets adopt varied strategies: PATHINSTRUCT[65] enables LLMs to invoke other pathology models, Lu *et al.*[49] propose six instruction types for diverse conversations, and CLOVER INSTRUCTION[11] combines LLM-generated and template-matched QAs for efficiency. PATHMMU[63] leverages GPT-4V [1] to create professional multi-modal pathology Q&As with detailed explanations. Given the lack of large-scale datasets for WSI interpretation, recent efforts also generate WSI-level instructions, often by converting reports into VQAs with advanced LLM prompts.

Image-Other Modality Pair Dataset. Exploration beyond vision–language remains limited. Zhou *et al.*[89, 90] built disease knowledge graphs with hierarchical semantic groups, MBTG-47K[70] paired WSIs with DNA and RNA sequences, and Xu *et al.* [78] released a WSI–report–RNA dataset. These efforts represent early attempts to integrate pathology images with additional modalities.

4 Evaluation Tasks

Unlike uni-modal FM4CPath, multi-modal MMFM4CPath not only improve image understanding but also enable zero-shot and cross-modal tasks, and when combined with LLMs, they gain dialogue and generative capabilities. As shown in Figure 2, their evaluation tasks can be grouped into six types: classification, retrieval, generation, segmentation, prediction, and VQA, and further distinguished by input level (tile or WSI). Pre-training data and model design are closely linked to evaluation scope; for instance, CPath-Omni [62] covers the broadest range thanks to multi-scale training and diverse instructions.

We categorize the evaluation tasks of MMFM4CPath into six types: classification, retrieval, generation, segmentation, prediction, and VQA. Classification is the most common evaluation task for MMFM4CPath, as many pathology-related tasks, such as cancer subtyping and biomarker prediction,

are fundamentally classification problems. Most MMFM4CPath are evaluated on classification tasks across various settings. Models using tile-level inputs can perform WSI-level classification via multiple instance learning (MIL), treated as weakly supervised due to the lack of detailed region annotations. Multi-modal data enables zero-shot or few-shot classification with minimal reliance on costly annotations. Some methods also assess out-of-distribution (OOD) generalization to handle distribution shifts between training and test data (*e.g.*, data collected from different institutions). Additionally, CONCH [48] evaluates classification on rare diseases with imbalanced data.

In addition to basic image-to-image retrieval, non-LLM-based MMFM4CPath are widely used for cross-modal retrieval tasks, such as text-to-image and image-to-text retrieval. KEP [90] performs one-to-many disease retrieval, retrieving captions or tiles with the same disease label using disease names. MLLM4PUE [88] enables many-to-one composed retrieval by using pathology images and questions as queries. Moreover, due to its capability to understand gene expression data, THREADS [70] generates class prompts from gene expression profiles for WSI retrieval.

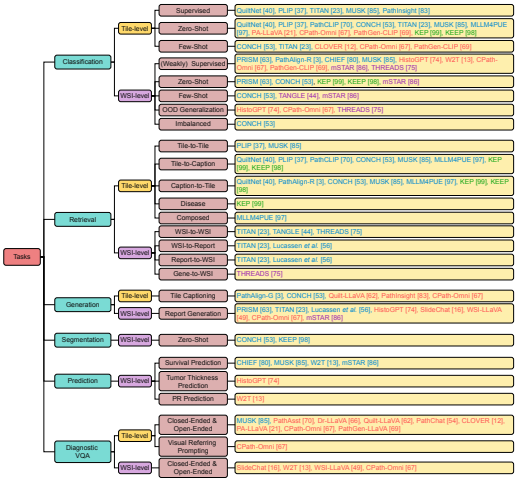

Figure 2: A taxonomy of MMFM4CPath by evaluation tasks, with non-LLM vision–language, LLM-based vision–language, vision–knowledge graph, and vision–gene expression models highlighted in different colors.

The integration of LLMs, whether by fine-tuning them as part of the model's architecture or by directly utilizing existing models, enables MMFM4CPath to generate captions/reports from tiles/WSIs. CONCH [48] and KEP [90] evaluate the segmentation capabilities of these models. Some MMFM4CPath have also been tested for prediction tasks, using WSIs to generate continuous value predictions.

LLM-based MMFM4CPath models focus on evaluating their diagnostic VQA ability. Compared to traditional QA tasks, VQA incorporates pathology images into its questions, challenging the image understanding capabilities of V-LLMs. Typically, VQA tasks involve answers from a fixed set, usually in the form of closed-ended questions, such as multiple-choice (single or multiple answers) or true/false questions, as well as open-ended questions with no predefined answer options. These tasks can also be divided into multi- and single-turn dialogues. The initial LLM-based MMFM4CPath only performed tile-level VQA tasks [65, 49], but recently, conversational abilities on WSI have gained increasing attention [15, 45]. Additionally, CPath-Omni [62] has been validated on the visual referring prompting task, where the regions of interest (ROIs) are highlighted, and both the question and answer are based on the these regions. It is worth noting that, due to its flexible format, the VQA task offers high adaptability: tasks like classification and generation can be transformed into VQA tasks via prompt engineering [75, 62]. Thus, LLM-based MMFM4CPath also encompass evaluation capabilities typical of non-LLM-based models. In addition to the quantitative analysis above, qualitative analysis is also frequently used to assess the performance of MMFM4CPath, especially their VQA and generation abilities. This is done by directly observing or through evaluation by professional pathologists to assess the quality of the generated text. Due to space constraints, the Future Directions are included in Appendix Section D.

5 Conclusion

In this survey, we have systematically reviewed the recent advances in multi-modal foundation models for computational pathology, focusing on three major paradigms: vision-language, vision-knowledge graph, and vision-gene expression models. We categorized 32 state-of-the-art models, analyzed 28 multi-modal datasets, and summarized key downstream tasks and evaluation strategies. Our comprehensive comparison highlights the growing impact and promise of integrating diverse data modalities in computational pathology. We hope this survey serves as a valuable reference for future research in developing generalizable, interpretable, and clinically useful multi-modal models.

References

[1] Gpt-4v(ision) system card. 2023.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] F. Ahmed, A. Sellergren, L. Yang, S. Xu, B. Babenko, A. Ward, N. Olson, A. Mohtashamian, Y. Matias, G. S. Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024.

[4] M. Alber, S. Tietz, J. Dippel, T. Milbich, T. Lesort, P. Korfiatis, M. Krügener, B. P. Cancer, N. Shah, A. Möllers, et al. A novel pathology foundation model by mayo clinic, charit\'e, and aignostics. *arXiv preprint arXiv:2501.05409*, 2025.

[5] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[6] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pages 456–473. Springer, 2022.

[7] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2, 2023.

[8] M. Bilal, M. Raza, Y. Altherwy, A. Alsuhaibani, A. Abduljabbar, F. Almarshad, P. Golding, N. Rajpoot, et al. Foundation models in computational pathology: A review of challenges, opportunities, and impact. *arXiv preprint arXiv:2502.08333*, 2025.

[9] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[10] D. Chanda, M. Aryal, N. Y. Soltani, and M. Ganji. A new era in computational pathology: A survey on foundation and vision-language models. *arXiv preprint arXiv:2408.14496*, 2024.

[11] K. Chen, M. Liu, F. Yan, L. Ma, X. Shi, L. Wang, X. Wang, L. Zhu, Z. Wang, M. Zhou, et al. Cost-effective instruction learning for pathology vision and language analysis. *arXiv preprint arXiv:2407.17734*, 2024.

[12] P. Chen, C. Zhu, S. Zheng, H. Li, and L. Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision*, pages 401–417. Springer, 2024.

[13] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022.

[14] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[15] Y. Chen, G. Wang, Y. Ji, Y. Li, J. Ye, T. Li, B. Zhang, N. Pei, R. Yu, Y. Qiao, et al. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. *arXiv preprint arXiv:2410.11761*, 2024.

[16] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[17] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[18] G. Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[19] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

[20] D. Dai, Y. Zhang, L. Xu, Q. Yang, X. Shen, S. Xia, and G. Wang. Pa-llava: A large language-vision assistant for human pathology image understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3138–3143. IEEE, 2024.

[21] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

[22] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv:2411.19666*, 2024.

[23] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.

[24] J. Gamper and N. Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16549–16559, 2021.

[25] R. H. Gindra, Y. Zheng, E. J. Green, M. E. Reid, S. A. Mazzilli, D. T. Merrick, E. J. Burks, V. B. Kolachalama, and J. E. Beane. Graph perceiver network for lung tumor and bronchial premalignant lesion stratification from histopathology. *The American Journal of Pathology*, 194(7):1285–1293, 2024.

[26] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[27] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[28] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[29] X. Guan, Z. Zhang, Y. Wang, and Y. Zhang. A systematic review on multimodal large language models (mllms) in computational pathology. *Authorea Preprints*, 2025.

[30] Z. Guo, J. Ma, Y. Xu, Y. Wang, L. Wang, and H. Chen. 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*, pages 189–199. Springer, 2024.

[31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[32] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[34] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 2023.

[35] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[36] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[37] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro. Quilt-1m: One million image-text pairs for histopathology. *NeurIPS*, 2024.

[38] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[39] A.-M. Istrate, D. Li, D. Taraborelli, M. Torkar, B. Veytsman, and I. Williams. A large dataset of software mentions in the biomedical literature. *arXiv preprint arXiv:2209.00693*, 2022.

[40] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *ICML*. PMLR, 2021.

[41] G. Jaume, L. Oldenburg, A. Vaidya, R. J. Chen, D. F. Williamson, T. Peeters, A. H. Song, and F. Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *CVPR*, 2024.

[42] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[43] D. Li, G. Wan, X. Wu, X. Wu, A. J. Nirmal, C. G. Lian, P. K. Sorger, Y. R. Semenov, and C. Zhao. A survey on computational pathology foundation models: Datasets, adaptation strategies, and evaluation tasks. *arXiv preprint arXiv:2501.15724*, 2025.

[44] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[45] Y. Liang, X. Lyu, M. Ding, W. Chen, J. Zhang, Y. Ren, X. He, S. Wu, S. Yang, X. Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. *arXiv preprint arXiv:2412.02141*, 2024.

[46] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[48] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 2024.

[49] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

[50] M. Y. Lu, B. Chen, A. Zhang, D. F. Williamson, R. J. Chen, T. Ding, L. P. Le, Y.-S. Chuang, and F. Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023.

[51] R. T. Lucassen, S. P. Moonemans, T. van de Luijtgaarden, G. E. Breimer, W. A. Blokx, and M. Veta. Pathology report generation and multimodal representation learning for cutaneous melanocytic lesions. *arXiv preprint arXiv:2502.19293*, 2025.

[52] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[53] M. Ochi, D. Komura, and S. Ishikawa. Pathology foundation models. *JMA journal*, 8(1):121–130, 2025.

[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[55] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[56] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[57] M. S. Seyfioglu, W. O. Ikezogwo, F. Ghezloo, R. Krishna, and L. Shapiro. 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*, pages 13183–13192, 2024.

[58] G. Shaikovski, A. Casson, K. Severson, E. Zimmermann, Y. K. Wang, J. D. Kunz, J. A. Retamero, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv:2405.10254*, 2024.

[59] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[60] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[61] S. Sun, A. Schubert, G. M. Goldgof, Z. Sun, T. Hartvigsen, A. J. Butte, and A. Alaa. Dr-llava: Visual instruction tuning with symbolic clinical grounding. *arXiv preprint arXiv:2405.19567*, 2024.

[62] Y. Sun, Y. Si, C. Zhu, X. Gong, K. Zhang, P. Chen, Y. Zhang, Z. Shui, T. Lin, and L. Yang. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. *arXiv preprint arXiv:2412.12077*, 2024.

[63] Y. Sun, H. Wu, C. Zhu, S. Zheng, Q. Chen, K. Zhang, Y. Zhang, D. Wan, X. Lan, M. Zheng, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer, 2024.

[64] Y. Sun, Y. Zhang, Y. Si, C. Zhu, Z. Shui, K. Zhang, J. Li, X. Lyu, T. Lin, and L. Yang. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*, 2024.

[65] Y. Sun, C. Zhu, S. Zheng, K. Zhang, L. Sun, Z. Shui, Y. Zhang, H. Li, and L. Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *AAAI*, 2024.

[66] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[67] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[68] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[69] M. Tran, P. Schmidle, S. J. Wagner, V. Koch, V. Lupperger, A. Feuchtinger, A. Böhner, R. Kaczmarczyk, T. Biedermann, K. Eyerich, et al. Generating highly accurate pathology reports from gigapixel whole slide images with histogpt. *medRxiv*, pages 2024–03, 2024.

[70] A. Vaidya, A. Zhang, G. Jaume, A. H. Song, T. Ding, S. J. Wagner, M. Y. Lu, P. Doucet, H. Robertson, C. Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

[71] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, et al. Virchow: A million-slide digital pathology foundation model. *arXiv:2309.07778*, 2023.

[72] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv:2208.10442*, 2022.

[73] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 2022.

[74] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 2024.

[75] X. Wu, R. Xu, P. Wei, W. Qin, P. Huang, Z. Li, and L. Luo. Pathinsight: Instruction tuning of multimodal datasets and models for intelligence assisted diagnosis in histopathology. *arXiv preprint arXiv:2408.07037*, 2024.

[76] P. Xia, K. Zhu, H. Li, T. Wang, W. Shi, S. Wang, L. Zhang, J. Zou, and H. Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.

[77] J. Xiang, X. Wang, X. Zhang, Y. Xi, F. Eweje, Y. Chen, Y. Li, C. Bergstrom, M. Gopaulchan, T. Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 2025.

[78] Y. Xu, Y. Wang, F. Zhou, J. Ma, S. Yang, H. Lin, X. Wang, J. Wang, L. Liang, A. Han, et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *arXiv preprint arXiv:2407.15362*, 2024.

[79] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[80] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[81] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[82] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[83] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.

[84] P. Zhang, X. Dong, B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, H. Duan, S. Zhang, S. Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[85] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[86] T. Zhao, Y. Gu, J. Yang, N. Usuyama, H. H. Lee, S. Kiblawi, T. Naumann, J. Gao, A. Crabtree, J. Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, pages 1–11, 2024.

[87] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[88] Q. Zhou, T. M. Dang, W. Zhong, Y. Guo, H. Ma, S. Na, and J. Huang. Mllm4pue: Toward universal embeddings in computational pathology through multimodal llms. *arXiv preprint arXiv:2502.07221*, 2025.

[89] X. Zhou, L. Sun, D. He, W. Guan, R. Wang, and L. o. Wang. A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. *arXiv:2412.13126*, 2024.

[90] X. Zhou, X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang. Knowledge-enhanced visual-language pretraining for computational pathology. In *ECCV*. Springer, 2024.

[91] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv:2408.00738*, 2024.

A Background

A.1 Computational Pathology

Computational Pathology (CPath) is an interdisciplinary field that applies computational techniques, including machine learning and computer vision, to analyze and interpret pathological data. By leveraging digital pathology, CPath enhances diagnostic accuracy, facilitates large-scale biomarker discovery, and supports personalized medicine. Among the various imaging modalities in pathology, Hematoxylin and Eosin (H&E) stained images serve as the most commonly used vehicle for studying CPath. These images capture essential morphological characteristics of tissues, making them fundamental for histopathological analysis. Within the realm of digital pathology, Whole Slide Images (WSIs) and tile images are two primary forms of data representation. WSIs, generated from high-resolution scanning of entire tissue slides, provide comprehensive visual information at gigapixel scale, allowing pathologists to examine cellular structures in detail. However, due to their enormous size and high computational demands, WSIs pose significant challenges in terms of storage, processing, and analysis. To mitigate these challenges, WSIs are often divided into smaller, more manageable tile images, which serve as the primary unit of analysis in many computational pathology studies.

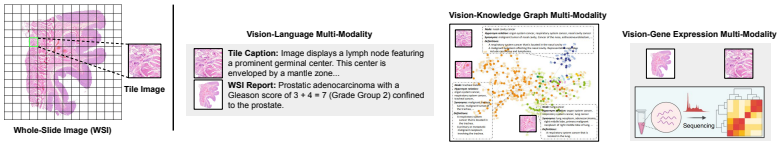

Figure 3: (Left) Illustration of whole-slide image and its corresponding tile images from H&E-stained tissue. (Right) The three primary types of multi-modal approaches in computational pathology.

While visual analysis remains central to CPath, researchers increasingly rely on multi-modal data to enhance interpretability and improve model performance. One major auxiliary modality is language, which includes both tile-level captions that describe specific regions of tissue and WSI-level pathology reports that provide global contextual information about a slide. Integrating text data with images enables vision-language models to learn richer feature representations and facilitate interpretability. Another important modality is structured domain knowledge, often represented in knowledge graphs, which encode relationships between diseases, biomarkers, and tissue structures, guiding AI models toward more biologically plausible interpretations. Additionally, molecular data, such as gene expression profiles, offer complementary insights by linking histopathological features to underlying genetic mechanisms. By aligning visual data with gene expression information, vision-gene expression models enable the discovery of novel genotype-phenotype associations. Figure 3 illustrates examples of WSIs and tile images alongside the three major multi-modal paradigms in CPath. The synergy of these multi-modal approaches, including vision-language, vision-knowledge graph, and vision-gene expression, has proven crucial in advancing the field of CPath, enabling more robust, generalizable, and interpretable AI-driven pathology models.

A.2 Pre-training Objective for Multi-Modal FMs

Unlike uni-modal models, which are primarily pre-trained through self-supervised contrastive learning (SSCL). Multi-modal FMs, due to their cross-modal nature, involve a more diverse set of self-supervised learning (SSL) objectives during their pre-training process. Furthermore, when fine-tuning LLMs to enable conversational abilities, supervised instruction tuning is usually required.

The primary pre-training objective for multi-modal FMs is SSCL. CLIP [54], as a pioneer in this field, ensures that the embeddings generated by the image encoder and text encoder are as similar as possible for paired image-text data by utilizing contrastive loss. CoCa [81] builds upon CLIP by adding a multi-modal encoder and an additional captioning loss to enable the mapping from the visual space to the language space. BLIP-2 [44] trains a lightweight Querying Transformer (Q-Former) using a two-stage strategy. In the first stage, a frozen image encoder bootstraps vision-language representation learning, while in the second stage, visual features are mapped to the language model input space, leveraging a frozen LLM for text generation. Additionally, next word prediction (NWP) is a text-specific SSL task commonly used for fine-tuning LLMs. It aims to predict the most likely

next token based on the given text sequence. Moreover, cross-modal alignment (CMA) multi-modal domain-specific task, which aims to build a unified semantic space where the embedding vectors from different modalities can reflect the same semantic content. In addition to contrastive learning, generative reconstruction and prediction are also commonly used SSL proxy tasks for CMA.

Instruction tuning (IT) is a method for fine-tuning LLMs to enable them to better understand and execute the instructions or task requirements provided by users. Unlike traditional pretraining objectives like NWP, the goal of IT is to enable the model to generate meaningful responses or actions based on specific instructions or questions. In Instruction Tuning, the model not only learns how to generate language but also learns how to adapt and generate different outputs according to various task requirements. This typically involves supervised training using a large number of instructions, ensuring that the model can understand the intent of the tasks and effectively perform them. Such tasks can include text generation, question answering, and conversation.

B Overview Table for Multi-Modal Foundation Models for CPath

We comprehensively summarize the network architectures and pre-training details across different stages for Multi-Modal Foundation Models for CPath (MMFM4CPath), as shown in Table 2

C Overview Table for Multi-Modal Datasets for CPath

We categorize the multi-modal datasets for CPath into three types based on data types and provide a summary of these datasets from multiple perspectives, as shown in Table 3.

D Future Directions

Developing MMFM4CPath Integrating H&E Images with Spatial Omics. The integration of H&E-stained histopathology images with spatial omics data, such as spatial transcriptomics and proteomics, represents a promising frontier in computational pathology. By coupling morphological context with spatially resolved molecular signatures, future multi-modal foundation models could enable precise cellular localization of gene and protein expression, bridging the gap between tissue architecture and molecular mechanisms. Developing such models would require addressing challenges like data sparsity, spatial resolution mismatch, and alignment between modalities, but could significantly enhance our understanding of disease heterogeneity and microenvironmental interactions.

Developing MMFM4CPath to Predict MxIF Markers from H&E Images. A compelling direction involves using H&E images to predict marker expressions captured by multiplexed immunofluorescence (MxIF), enabling cost-effective and scalable estimation of protein-level biomarkers. This line of research leverages the morphological cues from H&E to infer high-dimensional proteomic data, potentially reducing the need for expensive MxIF experiments. Multi-modal foundation models trained with paired H&E-MxIF data could facilitate virtual staining or marker imputation, supporting downstream tasks such as subtyping, immune landscape assessment, and therapy response prediction in a non-invasive manner.

Standardized Benchmarking for MMFM4CPath. As the field matures, there is a pressing need to establish standardized metrics and unified benchmarks for evaluating MMFM4CPath. Current evaluations are fragmented across tasks, modalities, and datasets, limiting comparability and reproducibility. Future work should focus on developing comprehensive evaluation protocols that span classification, retrieval, generation, and VQA across tile- and WSI-level inputs. Such efforts would guide model development, ensure fair comparisons, and accelerate the translation of multi-modal models into clinical practice.

E Scope and Exclusions

This survey focuses on multi-modal foundation models (FMs) developed specifically for computational pathology (CPath), with an emphasis on models built upon hematoxylin and eosin (H&E) stained whole-slide images (WSIs) and tile-level representations. We review 32 state-of-the-art

Table 2: Overview of architecture and pre-training details of MMFM4CPath (Due to space constraints, the references for the mentioned LLMs, V-LLMs, and off-the-shelf architectures are provided in the footnote of this table.)

	Model (Availability)	Year	Network Architecture†			Pre-training Details§					Input Image Type
			Vision (V)‡	Language (L) / Knowledge Graph (KG) / Gene Expression (GE)	Multi-Modal	Objective¶	Strategy* V	Strategy* O	Strategy* M	Data Short Description	
Vision-Language / Non-LLM-Based	QuiltNet (**author?**) [37] ✓	2023	T: ViT-B/32	**L**: Transformer Layers	-	SSL (CLIP)	D	D	-	438K Tiles and 802K Captions	Tiles
	PLIP [34] ✓	2023	T: ViT-B/32	**L**: Transformer Layers	-	SSL (CLIP)	D	D	-	208K Tile-Caption Pairs	Tiles
	PathCLIP [65] ✗	2024	T: ViT-B/32	**L**: Transformer Layers	-	SSL (CLIP)	D	D	-	207K Tile-Caption Pairs	Tiles
	PRISM [58] ✗	2024	T: ViT-H/14 W: Perceiver Net.	**L**: BioGPT (L1–12)	BioGPT (L13–24) with Cross-Attention Layers	SSL (CoCa)	F,S	F	F,S	587K WSIs with 195K Specimens	WSIs
	PathAlign-R [3] ✗	2024	T: ViT-S/16 W: Q-Former	**L**: Q-Former	-	SSL (MSN) SSL (CLIP)	S,N F,S	N S	- -	Tiles From 354,089 WSIs 434k WSI-Report Pairs	WSIs
	PathAlign-G [3] ✗	2024	T: ViT-S/16 W: Q-Former	**L**: Q-Former **L** (LLM): PaLM-2 S	MLP	SSL (MSN) SSL (BLIP-2) SSL (CMA)	S,N F,S F,D	N,N S,N N,F	N N S	Tiles From 354,089 WSIs 434k WSI-Report Pairs -	
	CHIEF [74] ✓	2024	T: Swin-T W: Aggregator Net.	**L**: Transformer Layers	MLP	WSL (CLIP)	D,S	D	S	60K WSIs with Labels	WSIs
	CONCH [48] ✓	2024	T: ViT-B/16	**L**: Transformer Layers	Transformer Layers	SSL (iBOT) SSL (NWP) SSL (CoCa)	S N D	N S D	N S D	16M Tiles Sampled From 21K WSIs >950K Pathology Text Entries 1.17M Tile–Caption Pairs	Tiles
	TITAN [22] ✓	2024	T: ViT-L W: ViT-S	**L**: Transformer Layers	Transformer Layers	SSL (iBOT) SSL (CoCa) SSL (CoCa)	F,S F,D F,D	N D D	N D D	336K WSIs 423K ROI-Caption Pairs 183K WSI-Report Pairs	WSIs
	MUSK [77] ✓	2025	T: V-FFN |⟵ Shared Attention Layers ⟶|	**L**: L-FFN	Cross-Attention Decoder	SSL (BEiT3) SSL (CoCa)	S D	S D	N S	1B Text Tokens and 50M Tiles 1.01M Tile–Caption Pairs	Tiles
	PathGen-CLIP [64] ✗	2025	T: ViT-B/32	**L**: Transformer Layers	-	SSL (CLIP) SSL (CLIP)	S D	S D	- -	1.6M High-Quality Tile-Caption Pairs 700K Tile-Caption Pairs	Tiles
	MLLM4PUE [88] ✗	2025	T: SigLIP	**L**: Qwen 1.5	MLP	SSL (CLIP)	D	D	D	594K Tile-Caption Pairs	Tiles
	Lucassen *et al.* [51] ✗	2025	T: ViT-L/14 W: Perceiver Net.	**L**: BioGPT (L1–12)	BioGPT (L13–24) with Cross-Attention Layers	SSL (CoCa)	F,S	F	F,S	42K WSIs and 19K Reports	WSIs
Vision-Language / LLM-Based	PathAsst [65] ✗	2024	T: ViT-B/32	**L** (LLM): Vicuna-13B	MLP	SSL (CMA) SL (IT)	F F	F I	S D	Description Part of PATHINSTRUCT 35K Samples From PATHINSTRUCT	Tiles
	Dr-LLaVA [61] ✓	2024	T: ViT-L/14	**L** (LLM): Vicuna-V1.5	MLP	SL (IT) & RL	D	I	D	Multi-turn Dialogues Based on 16K Tiles	Tiles
	Quilt-LLaVA [57] ✓	2024	T: ViT-B/32	**L** (LLM): GPT-4	MLP	SSL (CMA) SL (IT)	F F	F I	S D	723K Tile-Caption Pairs 107K Pathology-Specific Instructions	Tiles
	PathChat [49] ✓	2024	T: ViT-L/16	**L** (LLM): Llama 2-13B	MLP with Attention Pooling	SSL (CoCa) SSL (CMA) SL (IT)	D F F	N F I	S D D	1.18M Tile-Caption Pairs ~100K Tile-Caption Pairs 457K Instructions with 999K VQA Turns	Tiles
	HistoGPT-S/M [69] ✓	2024	T: Swin-T W: Perceiver Net.	**L** (LLM): BioGPT-B / BioGPT-L	-	WSL (MIL) SSL (NWP)	F,S F,F	F/F D/D	- -	15.1K WSIs with 6.7K Patient-Level Labels 15.1K WSI-Reports Pairs	Tiles
	HistoGPT-L [69] ✓	2024	T: ViT-L/16 W: GCN	**L** (LLM): BioGPT-L	-	SSL (NWP)	F,S	S	-	15.1K WSI-Reports Pairs	
	CLOVER [11] ✓	2024	T: EVA-ViT-G/14	**L**: Q-Former **L** (LLM): Vicuna 7B / FlanT5XL	Q-Former MLP	SSL (BILP-2) SL (IT)	F F	S,N/N N,I/I	S,N N,S	438K Tiles and 802K Captions 45K VQA Instructions	Tiles
	PathInsight [75] ✓	2024	|⟵ **V-LLM:** LLaVA / Qwen-VL-7B / InternLM ⟶|			SL (IT)		I / I / I		45K Instances Covering 6 Pathology Tasks	Tiles
	SlideChat [15] ✓	2024	T: ViT-L W: LongNet	**L** (LLM): Qwen2.5-7B	MLP	SSL (CMA) SL (IT)	F,S F,D	F I	S D	4.2K WSI-Report Pairs 176K Instruction-Following VQA Pairs	WSIs
	W2T [12] ✓	2024	T: ViT-S / Res-ResNet-50 / HIPT W: Transformer Layers	**L**: PubMedBERT / BioClinicalBERT / An Embedding Mapping	Transformer Layers	SSL (NWP)	T: F W: S	D D S	S	804 WSIs with 7.14K VQA Pairs	WSIs
	PA-LLaVA [20] ✓	2024	T: ViT-B/32	**L** (LLM): LLama3 with LoRA	Transformer Layers	SSL (CLIP) SSL (CMA) SL (IT)	D F F	F I I	F D D	827K Tile-Caption Pairs 518K Tile-Caption Pairs 35.5K VQA Pairs	Tiles
	WSI-LLaVA [45] ✗	2024	T: ViT-G/14 W: LongNet MLP	**L**: Bio_ClinicalBERT **L** (LLM): Vicuna-7b-v1.5	MLP	SSL (CLIP) SSL (CMA) SL (IT)	F,F,S F,F,F F,F,F	D,N N,F N,I	N S D	9.85K WSI-Report Pairs 9.85K WSI-Report Pairs 175K VQA Pairs	WSIs
	CPath-Omni [62] ✗	2024	T: ViT-H/14 ViT-L W: SlideParser	**L**: Qwen2.5-14B	MLP	SSL (CMA) SL (IT) SSL (CoCa) SL (IT)	F,F,F D,D,D F,F,D D,D,D	F I F I	S D F D	700K Tile-Caption Pairs 352K Tile Instructions 5.85K WSI-Report Pairs 53K Tile and 34K WSI Instructions	Tiles or WSIs
	PathGen-LLaVA [64] ✗	2025	T: ViT-B/32	**L**: Transformer Layers **L** (LLM): Vicuna	MLP	SSL (CLIP) SSL (CMA) SL (IC)	S F F	S,N N,F N,D	N S D	700K Tile-Caption Pairs 700K Tile-Caption Pairs 30K Detailed Tile Descriptions	Tiles
Vision-KG	KEP [90] ✓	2024	T: ViT-B/(16,32)	**L**: PubMedBERT ↪ **KG**: PubMedBERT	-	SSL (PKE) SSL (CLIP)	N D	N,S D,F	- -	A Pathology KG with 50.5K Attributes 715K Tile-Caption Pairs	Tiles
	KEEP [89] ✓	2024	T: ViT-L/16	**L**, **KG**: PubMedBERT	-	SSL (PKE) SSL (CLIP)	N D	S D	- -	A Pathology KG with 139K Attributes 143K Semantic Groups Through KG	Tiles
Vision-GE	TANGLE [41] ✓	2024	T: ViT-B (Rat) / Swin-T (Human) W: ABMIL	**GE**: A Three-Layer MLP	-	SSL (iBOT) SSL (CLIP)	S/N,N F/F,S	N S	- -	15M Rat Tiles From 47K WSIs 8.67K WSI- Gene Pairs	WSIs
	mSTAR [78] ✗	2024	T: ViT-L/16 W: Two-Layer TransMIL	**L**: BioBERT-Basev1.2 **GE**: scBERT	-	SSL (CLIP) SSL (SD)	F,S D,F	D,D N,N	- -	7.95K WSI-Report-Gene pairs 7.95K WSIs	WSIs
	THREADS [70] ✗	2025	T: ViT-L W: ABMIL	**GE**: scGPT (RNA), A Four-Layer MLP (DNA)	-	SSL (CLIP)	F,S	D,S	-	26.6K WSI-Gene (RNA) Pairs & 20.5K WSI-Gene (DNA) Pairs	WSIs

† Network architecture types: T: Tile Encoder, W: WSI Encoder, **L**: Text Encoder, **LLM**: Large Language Model, **V-LLM**: Multi-modal LLM, **KG**: Knowledge Graph Encoder, **GE**: Gene Expression Encoder.
§ Multi-modal foundation models are typically pretrained in multiple stages, with **each row in this column representing a distinct pretraining phase**.
¶ Training objectives are categorized into Supervised Learning (**SL**), Weakly Supervised Learning (**WSL**), Self-Supervised Learning (**SSL**), and Reinforcement Learning (**RL**). SL includes Image Captioning (**IC**) and Instruction Tuning (**IT**), WSL includes Multiple Instance Learning (**MIL**), and SSL encompasses Contrastive Learning (**CL**), Masked Siamese Networks (**MSN**), Next Word Prediction (**NWP**), Cross-Modal Alignment (**CMA**), Pathology Knowledge Encoding (PKE), and Self-Distillation (**SD**). CL is further divided based on its contrastive objectives into **CLIP**, **CoCa**, **BLIP-2**, **iBOT** and **BEIT3**.
* Pre-training strategies for different architectures (**V**: Vision, **O**: Other Modalities, **M**: Multi-Modal): **F**: Frozen, **S**: From Scratch, **D**: Domain-Specific Tuning, **I**: Instruction Tuning, **N**: Not Used, -: Not Existed.
• References of mentioned LLMs and V-LLMs in Table 2: BioGPT [52], PaLM-2 S [5], Qwen 1.5 [7], Vicuna-13B [16], Vicuna-V1.5 [67], GPT-4 [2], Llama 2-13B [68], Vicuna 7B [16], FlanT5XL [17], LLaVA [46], Qwen-VL-7B [7], InternLM [84], Qwen2.5-7B [79], LLama3 [26], LoRA [33], Vicuna-7b-v1.5 [87], Qwen2.5-14B [35], Vicuna [16]
• References of mentioned off-the-shelf architectures in Table 2: Perceiver Net. [40], GCN [25], EVA-ViT-G/14 [23], Q-Former [44], Swin-T [47], V-FFN [60], L-FFN [60], SigLIP [82], LongNet [21], PubMedBERT [27], BioClinicalBERT [28], HIPT [13], ABMIL [38], TransMIL [59], BioBERT-Basev1.2 [42], scGPT [19], scBERT [80]

models that integrate pathology images with auxiliary modalities such as textual reports, knowledge graphs, and molecular profiles, categorizing them into vision–language, vision–knowledge graph, and vision–gene expression paradigms. In addition, we analyze 28 pathology-specific multi-modal datasets, grouped into image–text pairs, instruction datasets, and image–other modality pairs, and summarize the evaluation tasks and strategies most relevant to CPath foundation models.

Several related directions are excluded from the scope of this survey. Specifically, methods that extend beyond pathology to broader biomedical imaging, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and X-ray [85, 83, 86, 76], are not covered in detail, as their primary goal is to build universal medical imaging models rather than enhance pathology image

Table 3: Multi-Modal Datasets for CPath.

	Dataset† (Availability)	Data Type	Description	Staining‡	Dataset Invariant	Data Source: Public	Data Source: Private	Method	LLM Assisted
Image-Text Pair	QUILT [37] ✓	Tile-Caption Pair	437,878 tiles paired with 802,144 captions extracted from 4,475 videos.	H, I, O	QUILT-1M: Combining QUILT with other pathology data sources to form 1M pairs.	YouTube	✗	QuiltNet [37]	✓
	PATHCAP [65] ✓	Tile-Caption Pair	207K pathology tile-caption pairs.	H, I, O	-	PubMed [27]	✗	PathCLIP [65]	✓
	OPENPATH [65] ✓	Tile-Caption Pair	208,404 tile-caption pairs.	H, I, O	PATHLAION: 32,041 additional tile–caption pairs scraped from the Internet and the LAION dataset [56]	WSI-Twitter, Replies, PATHLAION	✗	PLIP [34]	✗
	CONCH* [48] ✗	Tile-Caption Pair	1,170,647 tile–caption pairs.	H, I, O	-	PMC OA [39]	✓	CONCH [48]	✓
	HISTGEN [30] ✓	WSI-Report Pair	A WSI-report dataset with 7,753 pairs.	H	-	TCGA [66]	✗	-	✓
	MASS-340K [22] ✗	WSI	335,645 WSIs across 20 organs.	H, I	Synthetic captioning for 423,122 ROIs and curation of 182,862 WSI-report pairs.	GTEx [18]	✓	TITAN [22]	✓
	CPATH-PATCH CAPTION [62] ✗	Tile-Caption Pair	700,145 tile-caption pairs from diverse datasets.	H, I, O	-	PATHCAP, QUILT-1M, OPENPATH	✗	CPath-Omni [62]	✓
	PATHGEN [64] ✓	Tile-Caption Pair	1.6 million high-quality tile-caption pairs from 7,300 WSIs.	H	PATHGEN$_{instruct}$: 700K tile-caption pairs from PATHCAP, OPENPATH, and QUILT-1M	TCGA [66]	✗	PathGen-CLIP [64]	✓
	MUNICH [69] ✗	WSI-Report Pair	15,129 paired WSIs and pathology reports from 6,705 patients.	H	-	-	✓	HistoGPT [69]	✗
	PCAPTION-C [69] ✓	Tile-Caption Pair	1,409,058 tile-caption pairs.	H, I, O	PCAPTION-0.8M: removing non-human pathology data and PCAPTION-0.5M: further filter out pairs with <20 words.	PMC-OA [39], QUILT-1M	✗	PA-LLaVA [20]	✓
	ARCH [24] ✓	Bag-Caption Pair	11,816 bags and 15,164 images, with each bag containing multiple tiles.	H, I	-	PubMed [27], pathology textbooks	✗	-	✗
	MI-ZERO [50] ✓	Tile-Caption Pair	Diverse dataset of 33,480 tile-caption pairs.	H, I, O	-	educational resources, ARCH	✗	-	✗
Multi-Modal Instruction	PATHINSTRUCT [65] ✓	Tile-Level Instruction	180K pathology multi-modal instruction-following samples.	H, I, O	-	YouTube	✗	PathAsst [49]	✓
	CPATH-PATCH INSTRUCTION [62] ✗	Tile-Level Instruction	351,871 tile-level samples, including tile-caption pairs, VQA pairs, labeled images for classification, and visual referring prompting pairs.	H	CPATH-VQA: created by generating VQA pairs using GPT-4o [36], which combines classification labels with image data for datasets lacking captions.	CPATH-VQA, PATHGEN, CPATH-PATCHCAPTION, PATHINSTRUCT	✓	CPath-Omni [62]	✓
	CPATH-WSI INSTRUCTION [62] ✗	WSI-Level Instruction	7,312 WSI-level samples, including captioning, VQA, and classification.	H	Further generate a WSI VQA dataset by prompting GPT-4 [2].	HISTGEN	✗	CPath-Omni [62]	✓
	QULIT-INSTRUCT [57] ✓	VQA Pair	107,131 question/answer pairs.	H, I, O	QUILT-VQA: a Q&A dataset from Youtube videos, categorized into image-dependent and general-knowledge questions; QUILT-VQA-RED: QUILT-VQA with red circle marking the ROI in the pathology image.	YouTube	✗	Quilt-LLaVA [57]	✓
	PathChat* [49] ✗	Tile-Level Instruction	456,916 instructions with 999,202 question and answer turns.	H, I	PATHQABENCH: an expert-curated benchmark of 105 high-resolution pathology images, split into PATHQABENCH-PUBLIC and PATHQABENCH-PRIVATE subsets.	PMC-OA [39], TCGA [66]	✓	PathChat [49]	✓
	CLOVER INSTRUCTION [11] ✓	Tile-Level Instruction	45K question-and-answer instructions.	H	-	QUILT-VQA, PathVQA [32]	✓	CLOVER [11]	✓
	PATH-ENHANCEDS [75] ✓	Tile-Level Instruction	49K tile-level instructions, including captioning, VQA, classification and conversation.	H	-	OPENPATH, TCGA [66], PathVQA [32], *etc.*	✗	PathInsight [75]	✓
	SLIDE-INSTRUCTION [15] ✓	WSI-Level Instruction	44,181 WSI-caption pairs and 175,754 visual Q&A pairs.	H	SLIDEBENCH: 734 WSI captions along with a substantial number of closed-set VQA pairs to establish evaluation benchmark.	TCGA [66]	✗	SlideChat [15]	✓
	WSI-VQA [12] ✓	VQA Pair	977 WSIs and 8,672 Q&A pairs.	H	-	TCGA-BRCA [66]	✗	W2T [12]	✓
	PA-LLaVA* [20] ✓	VQA Pair	35,543 question-answer pairs.	H	-	PathVQA [32]	✗	PA-LLaVA [20]	✓
	WSI-BENCH [45] ✗	VQA Pair	179,569 WSI-level VQA pairs, which span across 3 pathological capabilities with 11 tasks.	H	SLIDEBENCH: 734 WSI captions along with a substantial number of closed-set VQA pairs to establish evaluation benchmark.	TCGA [66]	✗	WSI-LLaVA [45]	✓
	PATHMMU [63] ✓	VQA Pair	33,428 Q&As along with 24,067 pathology images.	H, I, O	SLIDEBENCH: 734 WSI captions along with a substantial number of closed-set VQA pairs to establish evaluation benchmark.	PubMed [27], QUILT-1M, Atlas [4], OPENPATH	✗	-	✓
Image-Other Modality	KEEP* [89] ✓	Pathology KG	KG contains 11,454 disease entities and 139,143 associated attributes.	-	-	DO [55], UMLS [9]	✗	KEEP [89]	✓
		Pathology Semantic Group	143K pathology semantic groups linked through the disease KG	H, I, O	-	QUILT-1M, OPENPATH	✗		
	PATHKT [90] ✓	Pathology KG	Pathology KG that consists of 50,470 informative attributes	-	-	OncoTree	✗	KEP [90]	✗
	mSTAR* [78] ✓	WSI-Report-RNA-Seq Pair	A dataset with 7,947 cases with image, text and RNA sequence modalities for pretraining.	H	-	TGCA [66]	✗	mSTAR [78]	✓
	MBTG-47K [70] ✗	WSI-RNA-Seq Pair WSI-DNA-Seq Pair	26,615 WSI-RNA pairs, and 20,556 WSI-DNA pairs.	H	-	TCGA [66], GTEx [18]	✓	THREADS [70]	✗

† Some methods introduced datasets without naming them, so we use the method name instead and marked with an asterisk (*).
‡ Staining type: **H**: H&E, **I**: IHC, **O**: Others.

representation. Similarly, we do not comprehensively review general-purpose multi-modal large language models (MLLMs) that incorporate pathology data only as a small subset of training, since their emphasis lies in broader generative AI capabilities rather than pathology-specific representation learning. By clearly defining these boundaries, we aim to provide a focused and coherent review of foundation models for computational pathology while acknowledging related but out-of-scope research directions.