LaGuadia: Language-Guided Adaptive Distillation from Pathology Foundation Models

Gangsu Kim and Won-Ki Jeong†

Department of Computer Science and Engineering, College of Informatics, Korea University, Republic of Korea
{gangsu1813, wkjeong}@korea.ac.kr

Abstract. Pathology Foundation Models (PFMs) offer powerful Whole Slide Image (WSI) representations but suffer from massive computational costs. While Knowledge Distillation (KD) can create efficient student models, existing multi-teacher methods often use suboptimal uniform weighting that ignores tissue heterogeneity. We propose LaGuadia (Language-Guided Adaptive DistillAtion), a framework that develops a compact pathology image encoder by dynamically integrating expertise from multiple PFMs under clinical linguistic guidance. Our approach utilizes a multi-stage pipeline: first, extracting visually observable clinical keywords from pathology reports; second, aligning visual features with these keywords via a Vision-Language meta-teacher (MedSigLIP) to provide dense semantic guidance; and finally, performing adaptive KD where teacher contributions are weighted based on their semantic alignment with the clinical narrative. Experiments on WSI captioning, visual question answering, and slide-level classification tasks demonstrate that an 87M parameter LaGuadia student model matches or exceeds foundation-scale models such as GigaPath and UNI, achieving strong factual consistency and robust generalization. These results highlight clinical language as an effective semantic anchor for building efficient and reliable digital pathology systems. Code is available at https://github.com/hvcl/LaGuadia.

Keywords: Whole Slide Image · Knowledge Distillation · Clinical Language Knowledge

1 Introduction

Computational pathology has rapidly advanced with the emergence of Pathology Foundation Models (PFMs), which leverage hundreds of millions of patches extracted from large-scale Whole Slide Image (WSI) datasets. Early studies primarily focused on vision-only PFMs that learn fine-grained morphological patterns in a self-supervised manner [4, 6, 26, 29]. More recently, Vision-Language PFMs have further enhanced pathology representation learning by integrating visual information with clinical reports, either through joint embedding spaces [12] or

† Corresponding author: wkjeong@korea.ac.kr

fused architectures [25]. This integration enables richer diagnostic context beyond pure morphology, substantially improving representational expressiveness. However, the massive parameter scale of such PFMs incurs high computational costs, limiting their practicality in real-world clinical settings. Thus, developing lightweight yet expressive student models is crucial for improving the accessibility of digital pathology.

Knowledge distillation (KD) has emerged as an effective strategy for transferring the representational capacity of large PFMs to compact student networks. In pathology, KD methods can be broadly categorized into single-teacher [5, 27, 24] and multi-teacher approaches [14, 23, 7, 21]. While single-teacher distillation is constrained by the bias of a fixed expert, multi-teacher KD seeks to leverage complementary strengths across models. However, existing multi-teacher methods typically rely on uniform or coarse confidence-based aggregation, failing to reflect the strong spatial and semantic heterogeneity of pathology images. As a result, the optimal teacher may vary substantially across tissue regions.

Despite the need for adaptive teacher selection, assessing teacher expertise solely from visual features remains ambiguous due to the lack of a reliable ground truth, as visually similar tissue regions may correspond to distinct diagnostic interpretations depending on clinical context. Since pathological diagnosis is ultimately summarized in linguistic reports, we argue that such reports provide a more objective semantic reference for identifying clinically relevant knowledge. Although Rasa [24] leverages report-derived linguistic information during distillation, it remains confined to a single-teacher model, limiting the diversity of transferable expertise. Based on this insight, we propose a computationally efficient pathology image encoder trained via language-guided multi-teacher knowledge distillation, in which teacher contributions are dynamically weighted based on the semantic alignment between visual embeddings and pathological text. This context-aware expert selection enables more robust and generalizable representations across diverse pathology tasks.

In this work, we make several key contributions to the field of computational pathology. First, we introduce LaGuadia, a novel KD framework that adaptively integrates the specialized expertise of multiple PFMs by aligning visual features with clinical keywords in a shared semantic space, selecting teachers per patch via clinical language rather than the fixed or vision-only weighting of prior multi-teacher distillation, thereby resolving which teacher to trust for each tissue region under semantic heterogeneity. Second, we demonstrate that a compact student model with only 87M parameters, fine-tuned with Low Rank Adaptation (LoRA) [8], can achieve superior performance compared to PFM teachers such as GigaPath [26] (1.13B) and UNI [4] (303M) in complex generative tasks, including WSI captioning and Vision Question Answering (VQA). Finally, we provide extensive empirical evidence that our language-guided approach significantly enhances factual consistency in generated pathological narratives while maintaining robust generalizability across diverse slide-level diagnostic tasks, such as survival analysis and cancer subtyping.

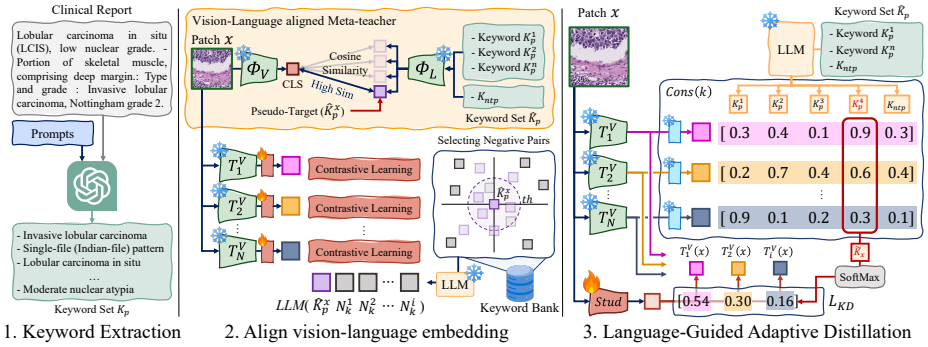

Fig. 1. Overview of the proposed framework comprises three stages.

2 Method

Figure 1 illustrates the overall framework of LaGuadia. The pipeline extracts clinically observable keywords, aligns visual embeddings through a vision–language meta-teacher, and adaptively distills teacher knowledge based on semantic agreement with the clinical narrative. The details of each stage are presented below.

2.1 Stage 1: Keyword Extraction

The first stage extracts clinical keywords from unstructured pathology reports to guide the image encoder's representation learning. To avoid non-visual noise—such as IHC or genetic alterations that are not visually observable in H&E-stained images—we prioritize keywords with high visual-semantic correlation. We establish three pathological constraints: **(i)** observability on H&E-stained WSIs, **(ii)** identifiability at 20x magnification, and **(iii)** minimal redundancy to maximize discriminative power in the semantic space. We employ GPT-5-mini [20] as a controlled keyword extractor using pathology-specific prompts designed to favor visually observable concepts. Subsequently, the keywords extracted across all cohorts are aggregated to construct a Keyword Bank, which serves as a global semantic reference for negative samples during the contrastive learning process.

2.2 Stage 2: Align Vision-Language embeddings via Meta-Teacher

While existing vision-only PFMs exhibit exceptional capabilities in representing morphological features, they lack a mechanism to directly link these features with clinical language. To address this limitation, we introduce a shallow projection layer integrated with the frozen visual backbones to perform an alignment process that incorporates linguistic knowledge into the visual representations.

Patch-level Pseudo-labeling via Pathology Expert Meta-Teacher. To

incorporate visual features with clinical semantic information in an environment where patch-level labels are unavailable, we perform weakly supervised learning via a Meta-Teacher Φ. Specifically, we employ MedSigLIP [17] as our Meta-Teacher, leveraging its capabilities as a pathology-specific Vision-Language Model (VLM) to provide dense semantic guidance for the image patches.

A WSI contains diverse background regions that do not directly correspond to diagnostic keywords, in addition to the actual tumor areas. Consequently, it is unrealistic to assume that every patch within a WSI reflects the diagnostic characteristics described in the pathology report. To account for these domain-specific characteristics, we extend the patient-specific keyword set K_p extracted in Section 2.1 by appending a No Tumor Present keyword k_{ntp}, representing non-diagnostic regions. This results in an expanded candidate set $\hat{K}_p = K_p \cup \{k_{ntp}\}$.

Given an input image patch x, the pseudo-label for the patch is defined as the keyword within the expanded set $\hat{K}_p$ of patient p that has the highest cosine similarity in the VLM embedding space:

$$\hat{K}_p^* = \arg\max_{k \in \hat{K}_p} \cos(\Phi_V(x), \Phi_L(k)) \tag{1}$$

Through this process, the model enhances semantic discriminability by aligning both tumor-specific morphologies and normal tissue patterns with their corresponding linguistic concepts.

Similarity-based Contrastive Learning Strategy. To strengthen the alignment between the selected $\hat{k}_p^*$ and the visual embeddings, we perform contrastive learning using a SigLIP [28] loss. To maximize the discriminability of the model, we dynamically reference N negative samples from the Keyword Bank established in Section 2.1.

Rather than employing simple random sampling for the negative set, we propose an average similarity-based thresholding strategy to enable the model to effectively learn pathological differences. Specifically, we compute an average similarity *threshold* by measuring the cosine similarity between the selected pseudo-label embedding and the patient-specific keyword embeddings. Keywords from the Keyword Bank whose embeddings fall below this threshold are selected as negative samples.

$$threshold(\hat{K}_p^*, \hat{K}_p) = avg_{k \in \hat{K}_p} \left(\cos(\Phi_L(\hat{K}_p^*), \Phi_L(k)) \right) \tag{2}$$

This strategy encourages the model to consistently repel concepts that are clearly distinct from the current patient's pathological distribution, thereby reinforcing the separation within the representation space. Furthermore, this serves as a critical foundation for enhancing the stability and accuracy of the language-guided weighting to be performed in Section 2.3

2.3 Stage 3: Language-Guided Adaptive Knowledge Distillation

In the final stage, we perform adaptive knowledge distillation to develop a unified student model by integrating the specialized expertise of multiple teachers.

The core of this process lies in determining adaptive weights that effectively incorporate the disparate representational strengths of each foundation model based on their clinical relevance.

Consensus-based Semantic Target Selection. To identify the core clinical concepts consistently highlighted by the ensemble of experts, we perform a soft voting process that aggregates similarity scores across the patient-specific keyword set $\hat{K}_p$. For each keyword $k \in \hat{K}_p$, which encompasses both diagnostic terms from the pathology report and the No Tumor Present concept, we calculate the consensus score $Cons(k)$ as follows:

$$Cons(k) = \sum_{i=1}^{n} \cos(Proj_i(T_i^V(x)), \Phi_L(k)) \tag{3}$$

The keyword with the highest soft voting score is selected as the Pseudo-Target Keyword ($\tilde{K}_x$) for the given patch. This process prevents knowledge conflicts between disparate models and establishes the most plausible clinical concept—as agreed upon by the ensemble of experts—as the anchor for knowledge transfer.

Semantic Similarity Based Weighting. To quantify each teacher's expertise, we compute the final distillation weights $\omega_i(x)$ by applying a Softmax function over the alignment scores, which represent the semantic proximity between the i-th teacher's visual representation and the target $\tilde{K}_x$:

$$\omega_i(x) = \frac{\exp(\cos(Proj_i(T_i^V(x)), \Phi_L(\tilde{K}_x))/\tau)}{\sum_{j=1}^{n} \exp(\cos(Proj_j(T_j^V(x)), \Phi_L(\tilde{K}_x))/\tau)} \tag{4}$$

This mechanism ensures that the teacher model most closely aligned with the clinical narrative receives a higher relative weight, allowing the student to adaptively distill knowledge from the most contextually relevant expert.

Adaptive Knowledge Distillation. Finally, the student model distills the class token representations from each teacher model based on the computed weights ω_i. The distillation loss $\mathcal{L}_{KD}$ is designed to maximize the cosine similarity, encouraging the student model to integrally learn the morphological, contextual, and clinical expertise of the three expert teachers under the consistent guidance of clinical language.

$$L_{KD} = \sum_{i=1}^{n} \omega_i(x) \cdot (1 - \cos(MLP_i(student(x)), T_i^V(x))) \tag{5}$$

3 Experiments and Results

3.1 Experimental Details

Datasets and Evaluation Protocol. For a large-scale analysis, we extracted approximately 28 million patches from 1,910 WSIs and 1,801 corresponding

pathology reports across three The Cancer Genome Atlas (TCGA) cohorts: BRCA, STAD, and THCA. To evaluate the generalizability of our learned representations, we conduct experiments across three diverse downstream tasks. For WSI Captioning, we employ the PathText dataset [2], and for VQA, we utilize WSI-VQA [3] for BRCA and WSI-Bench [10] for STAD and THCA. Lastly, for the MIL task, we perform subtyping, survival, molecular, and progression analysis using the same TCGA cohorts. All experiments follow a slide-level 5-fold cross-validation protocol, and we report the mean performance across all folds. Notably, LaGuadia is self-supervised, using no downstream labels, and its Keyword Bank is built only from the train/val splits, preventing test-split leakage.

Evaluation Metrics. We assess generated pathological narratives using METEOR (M) [1], BLEU-4 (B-4) [16], and ROUGE-L (R-L) [11]. Furthermore, we adopt $Fact_{ent}$ [15] to assess the factual consistency and clinical reliability of the generated text, ensuring its alignment with ground-truth medical knowledge. For the MIL tasks, we utilize the C-Index for survival analysis and the weighted F-1 score for other classification tasks, such as cancer subtyping. In all tables, **bold** and underlined values indicate the **best** and second-best results, respectively.

Implementation Details. For preprocessing, CLAM [13] extracted valid tissue from WSIs, segmented into 224×224 patches at $20\times$ magnification. Our teacher ensemble comprises frozen backbones of GigaPath [26], UNI [4], and Virchow2 [29]. The student, a ViT-B (86M) initialized with DINOv3 [19], was optimized using LoRA [8] on a single NVIDIA RTX A6000 (48GB). Stage 2 was trained for 10 epochs with a batch size of 4096 and 1 positive sample with 15 negative samples using AdamW with $lr = 1e-3$. Stage 3 was optimized for 150k iterations with a batch size of 256. We applied AdamW with $lr = 2e-4$ for LoRA and $lr = 1e-3$ for the distillation MLP layer. For comparison, we benchmark our method against PFMs including GigaPath (1.13B), Virchow2 (631M), MedSigLIP (429M) [17], UNI (303M), and MUSK (303M) [25], as well as distillation baselines such as the single-teacher framework H0-mini (86M) [5] and the multi-teacher framework GPFM (303M) [14].

3.2 Results

WSI Captioning. To assess the linguistic expressiveness of our visual representations, we replace the encoder in the HistoGPT [22] with various encoders. Table 1 compares the generated captions against ground-truth on the PathText dataset. Despite its compact 87M parameters, our encoder achieves an overall score of 0.2461, comparable to far larger models like UNI (303M) and GigaPath (1.13B). Specifically, our model excels in the STAD and THCA cohorts. In STAD, it achieves the highest scores across all metrics, including METEOR (0.2896) and BLEU-4 (0.1031). For the THCA cohort, it records the top BLEU-4 score (0.1015), indicating that our features capture precise morphological details necessary for accurate description. These results show that LaGuadia matches billion-parameter models while remaining highly parameter-efficient.

WSI VQA. The visual encoder's capacity for complex clinical reasoning is evaluated on the WSI-VQA and WSI-Bench datasets by leveraging the WSI-VQA

Table 1. Performance of WSI captioning on PathText benchmarks.

Model	BRCA ($n = 974$)			STAD ($n = 316$)			THCA ($n = 325$)			OVR
	M	B-4	R-L	M	B-4	R-L	M	B-4	R-L	
GigaPath	**0.2886**	0.0867	**0.3575**	0.2883	0.0973	0.3508	0.2755	0.1008	0.3615	0.2452
MUSK	0.2481	0.0636	0.3264	0.2620	0.0882	0.3272	0.2783	0.0985	0.3629	0.2284
Virchow2	0.2801	0.0864	0.3572	0.2857	0.1004	0.3497	0.2745	0.0983	0.3654	0.2442
MedSigLIP	0.2703	0.0776	0.3417	0.2874	0.0986	0.3466	0.2765	0.1001	0.3602	0.2399
UNI	0.2822	0.0856	0.3565	0.2887	0.0986	0.3519	0.2818	0.0996	0.3675	0.2458
GPFM	0.2794	0.0855	0.3562	0.2891	0.0999	0.3502	**0.2826**	0.1003	**0.3679**	0.2457
H0-mini	0.2833	**0.0872**	0.3563	0.2850	0.0975	0.3498	0.2734	0.0981	0.3612	0.2435
Ours	0.2815	0.0852	0.3520	**0.2896**	**0.1031**	**0.3524**	0.2818	**0.1015**	0.3676	**0.2461**

[3]. This task requires the model to not only identify localized features but also to understand the spatial and relational context of the tissue. As shown in Table 2, our encoder achieves the highest overall (OVR) score of 0.4184, albeit by a narrow margin over the second-best model (0.4182). This result demonstrates that the features extracted by our encoder possess superior reasoning capabilities across diverse pathological contexts. Notably, while showing a slight gap in linguistic fluency, our encoder yields significantly superior $Fact_{ent}$, reaching the highest scores in BRCA (0.9268) and STAD (0.4737). This divergence indicates that our language-guided adaptive distillation effectively prioritizes the alignment of visual features with precise pathological semantics, enabling the generation of reliable, fact-grounded responses.

Table 2. Performance of WSI Vision Question Answering on WSI-VQA benchmarks.

Model	BRCA ($n = 8,599$)			STAD ($n = 6,496$)			THCA ($n = 6,427$)			OVR
	M	R-L	$Fact_{ent}$	M	R-L	$Fact_{ent}$	M	R-L	$Fact_{ent}$	
GigaPath	0.2390	0.4709	0.9166	0.1714	0.3908	0.4700	0.1831	0.4067	0.4968	0.4161
MUSK	0.2290	0.4608	0.9007	0.1754	**0.3970**	0.4693	0.1827	**0.4244**	**0.5067**	0.4162
Virchow2	0.2445	**0.4903**	0.9086	0.1702	0.3846	0.4652	0.1812	0.4165	0.5027	0.4182
MedSigLIP	0.2208	0.4298	0.8913	**0.1767**	0.3968	0.4713	0.1822	0.4113	0.5004	0.4090
UNI	0.2343	0.4626	0.9178	0.1745	0.3912	0.4691	0.1782	0.4148	0.4926	0.4150
GPFM	**0.2448**	0.4770	0.9127	0.1747	0.3808	0.4710	0.1846	0.4191	0.4942	0.4176
H0-mini	0.2358	0.4624	0.8890	0.1667	0.3822	0.4546	0.1809	0.4121	0.4942	0.4087
Ours	0.2373	0.4716	**0.9268**	0.1700	0.3848	**0.4737**	**0.1855**	0.4123	0.5036	**0.4184**

MIL Classification. To evaluate representations in classification tasks, we adopt the ABMIL [9] architecture within the MIL-LAB [18]. Patch-level embeddings are aggregated to perform patient-level classification across multiple downstream tasks. Table 3 summarizes the performance of our encoder on various MIL classification benchmarks. Despite being explicitly optimized for alignment with consensus-derived clinical keywords, our model exhibits strong generalization across conventional vision-based tasks, without any loss in performance. Notably, it achieves the highest overall (OVR) score of 0.6538, outperforming all compared pathology foundation models. In particular, our approach attains superior results on diagnostically demanding tasks, including LAUREN classifi-

cation and N-stage prediction in STAD, as well as survival and progression in THCA. These results indicate that anchoring representations to clinically relevant semantics not only preserves but can even enhance the discriminative power of visual features in vision-centric downstream tasks.

Table 3. Performance of MIL Classification on various benchmarks.

Model	TCGA-BRCA		TCGA-STAD			TCGA-THCA		OVR
	Survival	Subtype	Survival	LAUREN	N-Stage	Survival	Progression	
GigaPath	0.5467	0.8909	0.5119	0.7277	<u>0.2920</u>	0.5162	0.8187	0.6149
MUSK	0.5625	0.7585	0.4939	0.6382	0.1894	0.4974	0.8243	0.5663
Virchow2	<u>0.5754</u>	<u>0.8966</u>	0.5122	0.7839	0.2379	0.5040	0.8216	<u>0.6201</u>
MedSigLIP	0.5671	**0.8974**	0.5188	0.7107	0.2783	0.4307	0.8194	0.5974
UNI	0.5536	0.7978	0.5320	0.8049	0.2469	0.5227	0.8261	0.6165
GPFM	0.5184	0.8936	<u>0.5332</u>	0.7599	0.2632	<u>0.5269</u>	0.8347	0.6181
H0-mini	0.5613	0.8890	0.4883	<u>0.8181</u>	0.2603	0.4841	<u>0.8350</u>	0.6199
Ours	**0.5801**	0.8930	**0.5436**	**0.8303**	**0.3007**	**0.5843**	**0.8446**	**0.6538**

<u>Ablation Studies.</u> We conducted ablation studies on the THCA cohort to compare the baseline, uniform Knowledge Distillation (**KD**), and our Language-Guided (**LaGu**) mechanism. While naive KD marginally enhances surface-level linguistic metrics by mimicking descriptive patterns, it inadvertently leads to performance degradation in critical clinical benchmarks, such as $Fact_{ent}$ (0.5017 to 0.4988) and survival analysis (0.5425 to 0.5180), compared to the baseline. These results suggest that uniform teacher aggregation may not effectively differentiate between clinically relevant and less informative teacher signals, potentially leading to suboptimal knowledge transfer. As shown in Table 4, incorporating LaGu into the distillation process consistently outperforms naive KD, enhancing METEOR (0.2818) while significantly boosting $Fact_{ent}$ (0.5036) and survival analysis (0.5843). These results demonstrate that LaGu effectively balances linguistic fluency with diagnostic accuracy, grounding visual representations in clinical semantics.

Table 4. Ablation study on the effect of Language-Guided Adaptive Distillation.

Method		WSI Captioning			WSI VQA			Classification	
KD	LaGu	M	B-4	R-L	M	R-L	$Fact_{ent}$	Survival	Progression
✗	✗	0.2699	0.0943	0.3560	0.1834	0.4187	0.5017	0.5425	0.8122
✓	✗	0.2739	0.0983	0.3610	**0.1856**	**0.4190**	0.4988	0.5180	0.8105
✓	✓	**0.2818**	**0.1015**	**0.3676**	0.1855	0.4123	**0.5036**	**0.5843**	**0.8446**

<u>Limitations/Future Work.</u> We adopt class token-based distillation for simplicity and training stability. Extending the framework to patch-level supervision remains a promising direction for capturing finer morphological details. Finally, distilling per-fold rather than once across all folds could further reduce cross-fold leakage, which we leave as future work.

4 Conclusion

In summary, we introduced LaGuadia, a framework that adaptively distills expertise from multiple PFMs via vision-language alignment. By matching or exceeding foundation-scale models (e.g., GigaPath, UNI) in WSI captioning and VQA with only 87M parameters, we demonstrate that clinical linguistic guidance is a powerful surrogate for massive scale. Furthermore, the high factual consistency of our model highlights its potential as a reliable tool for digital pathology in resource-constrained settings.

Acknowledgments. This work was supported in part by the National Research Foundation of Korea under Grant RS-2024-00349697 and Grant RS-2021-NR060143; in part by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2026-RS-2020-II201819; in part by the Technology Development Program funded by the Ministry of SMEs and Startups (MSS) under Grant RS-2024-00437796; in part by the Commercialization Promotion Agency for R&D Outcomes(COMPA) funded by the Ministry of Science and ICT(MSIT) under Grant RS-2026-25539491; and in part by Korea University Grant.

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

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909/
2. 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)
3. 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 (2024)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
5. Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., et al.: Distilling foundation models for robust and efficient models in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 162–172. Springer (2025)
6. 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)

7. Grashei, C., Brechenmacher, C., Umer, R.M., Liu, J., Marr, C., Szczurek, E., Schüffler, P.J.: Pathryoshka: Compressing pathology foundation models via multi-teacher knowledge distillation with nested embeddings. arXiv preprint arXiv:2511.23204 (2025)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
10. Liang, Y., Lyu, X., Chen, W., Ding, M., Zhang, J., He, X., Wu, S., Xing, X., Yang, S., Wang, X., et al.: Wsi-llava: A multimodal large language model for whole slide image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22718–22727 (2025)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
12. 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)
13. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
14. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., Yan, F., Cai, Y., Zhu, Z., Jin, C., et al.: A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nature Biomedical Engineering pp. 1–20 (2025)
15. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304 (2021)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
18. Shao, D., Chen, R.J., Song, A.H., Runevic, J., Lu, M.Y., Ding, T., , Mahmood, F.: Do multiple instance learning models transfer? In: International conference on machine learning (2025)
19. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
20. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
21. Tan, J.W., Kim, G., Jeong, W.K.: Pathme: Hierarchical multi-expert knowledge distillation for whole slide image analysis. IEEE Access (2025)
22. Tran, M., Schmidle, P., Guo, R.R., Wagner, S.J., Koch, V., Lupperger, V., Novotny, B., Murphree, D.H., Hardway, H.D., D'Amato, M., et al.: Generating dermatopathology reports from gigapixel whole slide images with histogpt. Nature communications **16**(1), 4886 (2025)
23. Wang, Q., Zhang, Y., Lu, J., Li, C., Zhang, Y.: Semi-supervised lung adenocarcinoma histopathology image classification based on multi-teacher knowledge distillation. Physics in Medicine & Biology **69**(18), 185012 (2024)

24. Wang, Z., Liu, H., Wang, Z., Li, D., Cen, M., Magnier, B., Liang, L., Wang, L.: Enhancing wsi-based survival analysis with report-auxiliary self-distillation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–207. Springer (2025)
25. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al.: A vision–language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)
26. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
27. Yu, M., Zhong, Z., Zhou, X., Wang, Y., Liang, T., Chen, J., Huang, H., Zhou, J., Zhao, D., Lei, B., et al.: Human visual attention-inspired knowledge distillation underlying interpretable computational pathology. Scientific Reports **15**(1), 42124 (2025)
28. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
29. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)