Pathologist Attention–Aligned Report Generation for Prostate Histopathology

Ruoyu Xue[1], Suryakant Singh[2†], Souradeep Chakraborty[1†], Pierre Marza[3,4], Oksana Yaskiv[5], Constantin Friedman[5], Natallia Sheuka[5], Paul Friedman[5], Bharat Ramlal[5], Beatrice Knudsen[6], Rajarsi Gupta[2], Joel Saltz[2], Prateek Prasanna[2], Gregory Zelinsky[7], and Dimitris Samaras[1]

[1] Department of Computer Science, Stony Brook University, USA
{ruoxue, souchakrabor, samaras}@cs.stonybrook.edu
[2] Department of Biomedical Informatics, Stony Brook University, USA
{suryakant.singh, Rajarsi.Gupta, Joel.Saltz, Prateek.Prasanna}@stonybrook.edu
[3] Université Paris-Saclay, CentraleSupélec, Gustave Roussy, INSERM, IHU PRISM, Cancer Data Science Unit, France
[4] Université Paris-Saclay, CentraleSupélec, MICS Laboratory, France
{pierre.marza}@centralesupelec.fr
[5] Department of Pathology and Laboratory Medicine, Northwell Health Laboratories
{OYaskiv, cfriedman7, nsheuka, pfriedman1, bramlal}@northwell.edu
[6] Department of Pathology, University of Utah School of Medicine
{beatrice.knudsen}@path.utah.edu
[7] Department of Psychology, Stony Brook University
{gregory.zelinsky}@stonybrook.edu
[†] Equal contribution.

Abstract. The allocation of visual attention by pathologists during cancer diagnosis is a highly selective process that critically shapes the information extracted from whole-slide images (WSIs). Human attention helps medical imaging tasks such as classification and segmentation, and becomes a strong semantic cue for identifying diagnostically informative regions for report generation. In this paper, we introduce human attention into the training of pathologist report generation models. To this end, we collected a multimodal human-attention dataset of 121 prostate WSIs annotated with pathologists' multi-scale viewport trajectories synchronized with the pathologists' verbal descriptions and cursor movements for five clinically relevant components. We then fine-tune two report generation models with an attention-alignment loss that regularizes the model attention over image patches to match the distribution of pathologist attention. We evaluate our approach on prostate cancer report generation and visual question answering using two models with different internal attention mechanisms (i.e., how image tokens are integrated into the language decoder). Experiments show average gains of 10.9% on NLP-based metrics and 19.3% in accuracy across five clinically relevant report components. Further, model attention maps extracted at inference time, with minimal computational overhead, align more closely with pathologist attention, providing stronger visual support for the

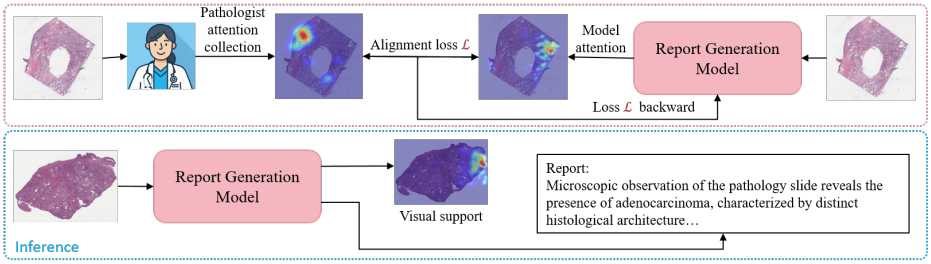


Fig. 1: Top-left: Collect pathologist attention during WSI examination. Top-right: Align model attention maps to human attention via a training-time loss. Bottom: At inference, the model generates reports with model-attention-based visual evidence.

generated reports by highlighting the regions that most influence the output. Dataset and Code are available at: https://github.com/cvlab-stonybrook/Pathologist-Attention-Aligned-Report-Generation.

Keywords: Report Generation · Human Attention · Explainability.

1 Introduction

Human visual attention is capacity-limited and shifts over time. During whole-slide image (WSI) examination, pathologists scan the slide sequentially, selectively focusing on task-relevant regions. These attention trajectories reflect the visual evidence a pathologist encodes during cancer diagnosis. Clinically, modeling pathologist attention can support trainee education by revealing how specialists examine slides and can also facilitate the assessment of pathologists' expertise levels [8,6,5,7], and understand decision-making [27]. Computationally, expert attention has increasingly been used as auxiliary supervision in medical imaging. Across pathology and radiology, attention signals have guided fully supervised and weakly supervised classification [21,4,2], segmentation [13], and text-to-image generation [3]. Beyond medical imaging, evidence from cognitive science and attention-aware image description [14,11,30] suggests that where people look shapes how they interpret and describe visual content. Together, these findings motivate leveraging pathologists' attention as an auxiliary cue for pathology report generation, which grounds generated text in plausible visual evidence, and improves both interpretability and report quality.

Despite rapid recent progress in pathology report generation and visual question answering [19,15,10,31,28], only a few works incorporate auxiliary supervision such as model-predicted semantic information, and none so far leverage human attention in pathology report generation. Kim *et al.* [18] introduces semantic cues via a multi-label classifier over key diagnostic elements, but the generated reports are limited to predefined categories and depend on predicted seman-

tics rather than ground-truth signals. Agent-based approaches such as CPathAgent [26] and PathChat+ [9] supervise the training by predicting the Region Of Interest(ROI)-based visual support with step-by-step reasoning; however, these ROIs are not derived from the attention of actual pathologists, and inducing such visual support typically requires large-scale instruction tuning.

In this paper, we align the visual attention of report generation models with pathologist attention. In vision–language transformers, model attention weights indicate how much each image patch contributes to token generation, providing a natural mechanism for grounding diagnostic statements in visual evidence. To supervise this alignment, we first collect a prostate cancer pathologist-attention dataset that captures multi-scale WSI examination behavior for five clinically critical components: Gleason patterns, perineural invasion, intraductal carcinoma, extraprostatic extension, and surgical margin. While the pathologists read the slides, we record viewport trajectories across magnification levels and synchronized, real-time verbal descriptions of observed findings, which provide soft labels for all the five components across the corresponding viewports.

Using this rich supervision, we introduce a plug-and-play module for report generation in the form of an auxiliary attention-alignment loss that regularizes model attention toward the corresponding pathologist attention distribution over image tokens (Figure 1). Pathologist attention is required only during training, with no additional inputs at inference, and the loss adds negligible overhead. We evaluate this design on both a report generation model and a WSI visual question answering model, covering two common vision–language fusion designs: cross-attention and self-attention. Experiments show 5.5%–12% average improvements on NLP-based metrics, and 8%–34% gains in accuracy on the five clinically related components across the two models. The model attention in the inference-time is aligned more closely with pathologist attention, providing stronger visual evidence for the generated reports. Our contributions are:

- To our knowledge, we are the first to incorporate pathologist-attention supervision into pathology report generation, showing consistent report-quality gains across architectures.
- We collect a multimodal pathologist-attention dataset of prostate WSIs with multi-scale viewport trajectories synchronized with spoken descriptions of observed findings.
- We improve interpretability by producing attention-based visual evidence that better matches how pathologists attend during cancer diagnosis.

As we show, fine-tuning a pathology report generation model with human attention improves performance with relatively small pathologist annotation effort ($\sim$ 7 hours in total for the reported experiments). This suggests that our plug-and-play module can easily become a part of model training for pathology report generation.

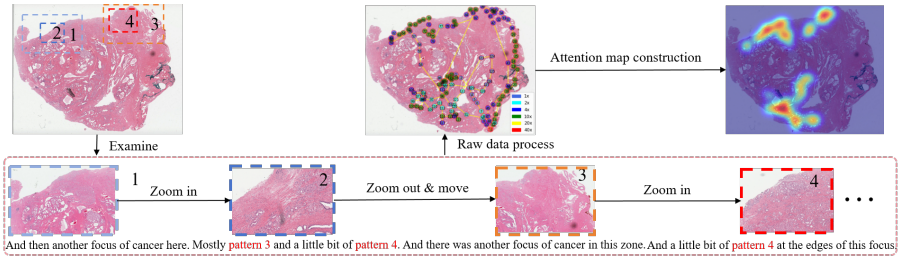


Fig. 2: **Data collection process**. Pathologists examine each WSI by iteratively zooming for detail and zooming out to navigate (bottom right). We record synchronized verbal descriptions, convert viewports center to scanpath (top center), filter points by target components, and generate a Gaussian-smoothed saliency map overlaid on the slide (top right).

2 Pathologist Attention Dataset

Dataset collection. We record pathologists' attention using QuIP caMicroscope [25]. Six pathologists (2 specialists, 2 generalists, 2 residents) reviewed 121 TCGA-PRAD prostate WSIs [33] and annotated seven elements: Gleason score, intraductal carcinoma (IDC), perineural invasion (PNI), extraprostatic extension (EPE), surgical margin, lymphovascular invasion, and seminal vesicle invasion. As shown in Figure 2, slides are viewed in a 1912×930 viewport with iterative zooming and panning, yielding multi-magnification viewports. We sample viewports at 20,Hz and convert each to its center point to form a magnification-aware scanpath; synchronized verbal descriptions provide soft component labels. We remove duplicates and keep only positively labeled points per component to reduce the influence of individual difference and noise. Due to sparsity, we exclude lymphovascular and seminal vesicle invasion and focus on five components. On average, scanpaths contain 59 points with 221,s per slide ($\sim$ 7 hours total).
Human attention map construction. We construct Gaussian-blurred saliency maps [17,6,8]. For each magnification (2×, 4×, 10×, 20×), we project scanpath points onto the corresponding grid, apply Gaussian smoothing, then resize and merge maps into a single multi-scale saliency map per slide.

3 Method

3.1 Preliminaries

Report generation models adopt transformers structure [29]. Given queries Q, keys K, and values V, scaled dot-product attention computes

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \qquad \text{Attn}(Q, K, V) = AV, \tag{1}$$

where A is the model attention weights, and $\mathrm{Attn}(Q, K, V)$ is the output feature of a transformer attention layer.
Multimodal report generation models typically use two image-text fusion designs. **Self-attention fusion** concatenates image and text tokens $Z = [\mathbf{X}; \mathbf{Y}_{<t}]$, where $\mathbf{X} \in \mathbb{R}^{I \times d}$ are image tokens and $\mathbf{Y}_{<t} \in \mathbb{R}^{t \times d}$ are previously generated text tokens, and projects Q, K, V from Z. **Cross-attention fusion** projects Q from the current text query, while K, V come only from $\mathbf{X}$. These fusion determine how slide features influence next-token prediction: WSI-LLaVA [19] uses self-attention fusion, whereas HistGen [15], Bi-Gen [31], and HistoGPT [28] adopt cross-attention.

3.2 Pathologist Attention Alignment

Pathologist attention encoding in token space. Pathology report generation models first encode each WSI into a sequence of image tokens $\mathbf{F} \in \mathbb{R}^{I \times d}$. Given a full-resolution human saliency map, we assign each token a saliency value by querying the map at the token's recorded patch coordinates, yielding $\mathbf{a}^{\text{human}} \in \mathbb{R}^{I}$. To encourage model attention consistent with pathologists' attention, we fine-tune WSI-LLaVA and HistGen with the standard autoregressive next-token objective augmented by an attention-alignment loss.
Human attention alignment in WSI-LLaVA [19]. We use the last-layer self-attention weights $\mathbf{A} \in \mathbb{R}^{(I+P+T)\times(I+P+T)}$, where I is the number of image tokens, P is the prompt length, and T is the number of answer tokens. We extract the answer-to-image attention block

$$\mathbf{A}_{\text{ans}\to\text{img}} = \mathbf{A}\big[\mathcal{T}_{\text{ans}}, \mathcal{I}\big] \in \mathbb{R}^{T \times I}, \tag{2}$$

where $\mathcal{I}$ indexes image-token positions and $\mathcal{T}_{\text{ans}}$ indexes answer-token query positions. We then aggregate token-wise attention into an image-level model attention vector by averaging over answer tokens:

$$\mathbf{a}^{\text{model}} = \frac{1}{T} \sum_{t \in \mathcal{T}_{\text{ans}}} \mathbf{A}_{\text{ans}\to\text{img}}[t, :] \in \mathbb{R}^{I}. \tag{3}$$

Given the corresponding human attention vector $\mathbf{a}^{\text{human}} \in \mathbb{R}^{I}$, we define the alignment loss as the KL divergence between the attention distributions following the design of GazeVLM [22],

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\Big(\mathrm{softmax}(\mathbf{a}^{\text{human}}) \;\Big\|\; \mathrm{softmax}(\mathbf{a}^{\text{model}})\Big). \tag{4}$$

Training objective. Given the ground-truth tokens $\{t_1, \ldots, t_T\}$, the base objective is the next-token negative log-likelihood $\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log p_\theta(t_t \mid t_{<t}, \mathbf{X})$. The overall loss is $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{KL}}$, where λ controls the strength of alignment, and is set to 1.0 in the experiment.
Human attention alignment in HistGen. Given the different attention mechanism of HistGen (cross-attention) from WSI-LLaVA (self-attention), we directly extract the model attention $\mathbf{A}^{\text{model}} \in \mathbb{R}^{T \times I}$ and then follow the same way to compute alignment loss.

3.3 Inference-Time Visual Support Generation

During inference, we generate reports and extract model attention to image tokens in each step, then average across steps to obtain a slide-level vector $\mathbf{a}^{\text{model}}$, then back-project it to a pixel-level heatmap by assigning each patch its token value, followed by normalization and Gaussian smoothing (Section 3.2).

4 Experiments

4.1 Setups

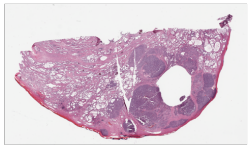
Baselines. We use HistGen [15] and WSI-LLaVA [19] as base models, which differ in multimodal fusion attention mechanisms and are designed for different tasks (report generation vs. visual question answering). **Zero-shot** denotes directly evaluating each pretrained model on its test set. **w/o alignment** denotes LoRA fine-tuning [16] using only the next-token prediction loss, without any pathologist-attention alignment.
Report Dataset. We use the official datasets and splits of HistGen and WSI-LLaVA. For HistGen, attention alignment is trained only on the 81 slides with our attention annotations, while evaluation follows the official val/test sets (86/76). Visual-support evaluation (requiring GT comparison) is performed on 20 test slides. For WSI-LLaVA, we use 107 training examples and 11 test examples for report-quality evaluation, and 15 test samples for visual-support evaluation.
Metrics. We evaluate both report quality and visual support. For report generation, we report standard NLP metrics: BLEU (B1–B4) [23], METEOR (M) [1], and ROUGE-L (R) [20]. Since our attention annotations target five pathological components, we use ChatGPT-5.2 to extract them and compute exact-match scores between generated and ground-truth reports: Gleason requires matching both primary and secondary patterns, while the remaining components are binary scored. We report ACC-G, ACC-PNI, ACC-EPE, and ACC-S for Gleason score, perineural invasion, extraprostatic extension, and surgical margin, respectively; we omit intraductal carcinoma because it does not appear in the ground-truth reports. For visual support, we measure attention heatmap similarity with Normalized Scanpath Saliency (NSS) and KL divergence (KL) [24], and quantify agreement with ground-truth Gleason regions using segmentation labels from [12], reporting ROC-AUC (the probability that a pixel inside the binary segmentation mask receives a higher heatmap score than a pixel outside).
Implementation details. We use LoRA finetuning for both HistGen and WSI-LLaVA with the same split and training setting as the original papers. The experiment runs on a single RTX 6000 GPU. We set $\lambda = 1.0$ for both models.

4.2 Results

Quantitative results. In Table 1, aligning pathologist attention achieves the best performance on both HistGen and WSI-LLaVA. In particular, our method

Backbone	Method	B1	B2	B3	B4	R	M	ACC-G	ACC-PNI	ACC-EDE	ACC-S
HistGen [15]	Zero-shot	0.329	0.162	0.087	0.048	0.172	0.150	25.0	25.7	30.0	25.7
	w/o Alignment	0.326	0.174	0.102	0.065	0.185	0.151	30.0	35.0	34.3	28.5
	Ours	**0.361**	**0.194**	**0.116**	**0.077**	**0.203**	**0.166**	**35.7**	**40.0**	**44.3**	**30.0**
WSI-LLaVA [19]	Zero-shot	0.469	0.283	0.187	0.129	0.328	0.211	45.5	81.8	45.5	-
	w/o Alignment	0.444	0.254	0.146	0.093	0.300	0.220	54.5	81.8	45.5	-
	Ours	**0.479**	**0.292**	**0.197**	**0.136**	**0.338**	**0.228**	**63.6**	**90.9**	**63.6**	-

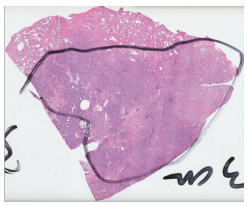
Table 1: Quantitative results on the test set of pathologist-aligned report generation performance on prostate cancer. ACC-S of WSI-LLaVA is not reported due to missing annotations in the ground-truth reports.



(Ours) biopsy showed moderately differentiated prostatic adenocarcinoma with a Gleason score of 9 in the left and right seminal vesicle and extraprostatic extension. Surgical margins were negative for tumor, …

(w/o alignment) gland with biopsy reveals a Gleason score of 3+3=6. Lymph nodes submitted for bladder cancer. while the bladder prostate gland weighs 5 g and shows multifocal prostatic hyperplasia, B) prostate prostate. The radical prostatectomy with the left seminal vesicles and vas deferens reveals intact but no carcinoma. Extraprostatic extension or prostate N findings noted. Tumor High-grade prostatic intraepithelial neoplasia is identified. Perineural invasion is not identified.

(GT) The pathology report describes an adenocarcinoma tumor with a Gleason score of 9 in the left posterior prostate, with extracapsular extension identified at the left posterior base. The seminal vesicles and bladder neck were not involved, and surgical margins were free of tumor, …



(Ours) The slide reveals a poorly differentiated adenocarcinoma with a Gleason score of 4+4, indicating a high-grade tumor with a predominant pattern of poorly formed glands and a secondary pattern of individual cells or sheets. There is extensive perineural invasion, suggesting aggressive behaviour, …

(w/o alignment) Microscopic observation of the pathology slide reveals adenocarcinoma characterized by a Gleason score of 4+3=7, indicating a primary pattern of poorly formed glands (score 4) and a secondary pattern of moderately differentiated glands (score 3). The tumor exhibits extensive perineural invasion, indicating that cancer cells are present in the spaces surrounding nerves, …

(GT) Microscopic observation of the pathology slide reveals poorly differentiated prostatic adenocarcinoma characterized by glandular architecture that significantly deviates from the normal prostate tissue structure. The Gleason score is recorded as 4+3=7, indicating a predominance of poorly formed glandular patterns with fused glands and cribriform structures. Noteworthy is the observation of perineural invasion, where cancer cells are seen encircling or tracking along nerves, …

Fig. 3: **Qualitative results of report generation.** Top: HistGen; Bottom: WSI-LLaVA. Components mentioned in the generated reports are highlighted in red, indicating more complete and higher-quality reports.

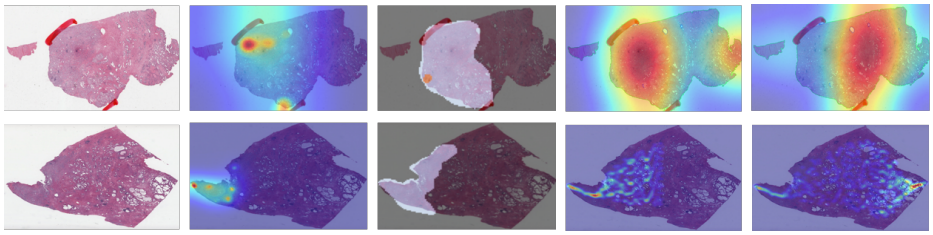


Fig. 4: **Qualitative results of visual support.** Visual support is extracted from the last decoder layer at inference (top: HistGen, blurred due to the small number of image tokens for better visualization; bottom: WSI-LLaVA). Compared to training without alignment (fifth column), our method (fourth column) suppresses non-Gleason regions and better matches ground-truth pathologist attention (second column) and Gleason segmentation [12] (third column).

Method	HistGen			WSI-LLaVA		
	NSS↑	KL↓	ROC-AUC ↑	NSS↑	KL↓	ROC-AUC ↑
w/o alignment	0.43	5.45	0.500	0.76	3.10	0.598
with alignment (Ours)	**0.59**	**4.76**	**0.537**	**0.87**	**2.98**	**0.615**

Table 2: **Quantitative results of visual support.** Our method produces model attention that aligns better with human and segmentation labels.

Method	HistGen			WSI-LLaVA		
	B4	R	ACC-G	B4	R	ACC-G
all layers	**0.079**	0.199	34.2	0.124	0.325	45.5
last layer (Ours)	0.077	**0.203**	**35.7**	**0.136**	**0.338**	**63.6**

Table 3: **Ablation on attention-extraction layer.** Applying the attention-alignment loss at every layer yields similar or worse performance than supervising only the last layer, supporting our choice of a lightweight last-layer design with minimal overhead.

yields higher accuracy on all five clinical components than the baselines, suggesting that human attention guides the model toward more informative regions. We also observe that fine-tuning WSI-LLaVA without attention alignment leads to a performance drop, likely due to overfitting in this low-data setting, as WSI-LLaVA is built on a large language model (Vicuna-7b-v1.5 [32]). By providing human attention as a visual constraint, this alignment discourages the model from memorizing report text. Despite the relatively low ACC-G of HistGen, our Gleason score classifier improves AUC from 0.71 to 0.77, suggesting stronger Gleason-related representation beyond free-text report evaluation.

Qualitative results of generated reports. In Figure 3, our model captures the clinically critical components targeted in our dataset collection more consistently than the baselines, suggesting it learns to leverage pathologist attention cues effectively.

Visual support. We report qualitative and quantitative results on visual support in Figure 4 and Table 2, evaluating (1) alignment with pathologist attention and (2) coverage of Gleason patterns. These results suggest our alignment loss encourages the model to focus on more informative, human-preferred regions, improving report quality. The resulting visual support also enhances interpretability by indicating which regions support the model's diagnostic statements. A user-study indicates that pathologists preferring it over baselines in 80% of cases for both overall quality and component-specific diagnosis.

Ablation. We ablate which transformer layer to use for extracting model attention in human-attention alignment. Results in Table 3 show that supervising only the last layer is sufficient to steer the model toward pathologist-attended regions, likely because last-layer attention most directly influences next-token prediction. In contrast, aligning multiple layers in WSI-LLaVA degrades performance below the zero-shot baseline, likely because it is much deeper than HistGen due to its large language model backbone.

5 Conclusion and Discussion

We propose a plug-and-play pathologist attention–alignment framework that uses multi-scale expert attention as auxiliary supervision for explainable pathology report generation. Across HistGen and WSI-LLaVA, it improves report quality and clinically relevant component accuracy while producing attention maps that better match pathologists' attention with minimal overhead. Our results highlight the value of auxiliary signals for report generation and suggest scalability to additional cancer types due to the modest data-collection burden. Future work will expand attention collection to more cancer types and develop models that generalize to new cancer types by leveraging shared examination behaviors of pathologists among different cancer types.

6 Acknowledgement

This research is supported by US National Science Foundation (NSF) grants IIS-2123920, IIS-2212046 and 2442053, National Institutes of Health (NIH), National Cancer Institute (NCI) grants 1R21CA25849301A1, 1R01CA297843-01, 3R21CA25849302S1, 1R03DE033489-01A1 and the *Health Data Hub* as part of the second edition of the *France-Québec* call for projects *Intelligence Artificielle en santé*. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
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: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Bhattacharya, M., Prasanna, P.: Gazediff: A radiologist visual attention guided diffusion model for zero-shot disease classification. In: Medical Imaging with Deep Learning (2024)
3. Bhattacharya, M., Singh, G., Jain, S., Prasanna, P.: Radgazegen: Radiomics and gaze-guided medical image generation using diffusion models. arXiv preprint arXiv:2410.00307 (2024)
4. Bhattacharya, M., Singh, G., Jain, S., Prasanna, P.: Gazelt: Visual attention-guided long-tailed disease classification in chest radiographs. arXiv preprint arXiv:2508.09478 (2025)
5. Brunye, T.T., Carney, P.A., Allison, K.H., Shapiro, L.G., Weaver, D.L., Elmore, J.G.: Eye movements as an index of pathologist visual expertise: a pilot study. PloS one **9**(8), e103447 (2014)
6. Chakraborty, S., Gupta, R., Ma, K., Govind, D., Sarder, P., Choi, W.T., Mahmud, W., Yee, E., Allard, F., Knudsen, B., et al.: Predicting the visual attention of pathologists evaluating whole slide images of cancer. In: International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis. pp. 11–21. Springer (2022)

7. Chakraborty, S., Gupta, R., Yaskiv, O., Friedman, C., Sheuka, N., Perez, D., Friedman, P., Zelinsky, G., Saltz, J., Samaras, D.: Decoding the visual attention of pathologists to reveal their level of expertise. In: MICCAI. pp. 120–130. Springer (2024)
8. Chakraborty, S., Xue, R., Gupta, R., Yaskiv, O., Friedman, C., Sheuka, N., Perez, D., Friedman, P., Choi, W.T., Mahmud, W., et al.: Measuring and predicting where and when pathologists focus their visual attention while grading whole slide images of cancer. Medical Image Analysis p. 103752 (2025)
9. Chen, C., Weishaupt, L.L., Williamson, D.F., Chen, R.J., Ding, T., Chen, B., Vaidya, A., Le, L.P., Jaume, G., Lu, M.Y., et al.: Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv preprint arXiv:2506.20964 (2025)
10. 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: MICCAI. pp. 546–556. Springer (2024)
11. Coco, M.I., Keller, F.: Scan patterns predict sentence production in the cross-modal processing of visual scenes. Cognitive science **36**(7), 1204–1223 (2012)
12. Codatta: Refined-tcga-prad prostate cancer pathology dataset. Hugging Face Datasets, `https://huggingface.co/datasets/Codatta/Refined-TCGA-PRAD-Prostate-Cancer-Pathology-Dataset`, accessed: 2026-02-24
13. Ge, R., Li, R., Wang, C., Liu, Y., Zhu, H., Coatrieux, J.L., Zhang, D., Lu, J., Chen, Y., Li, S., et al.: Adaptation follow human attention: Gaze-assisted medical segment anything model. IEEE Transactions on Circuits and Systems for Video Technology (2025)
14. Griffin, Z.M., Bock, K.: What the eyes say about speaking. Psychological science **11**(4), 274–279 (2000)
15. Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H.: Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In: MICCAI. pp. 189–199. Springer (2024)
16. 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)
17. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: CVPR. pp. 1072–1080 (2015)
18. Kim, K.A., Hong, S., Yoo, S., Kang, Y., Shim, H.S.: Enhancing structured pathology report generation with foundation model and modular design. IEEE Access (2025)
19. 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: CVPR. pp. 22718–22727 (2025)
20. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
21. Nan, T., Zheng, S., Qiao, S., Quan, H., Gao, X., Niu, J., Zheng, B., Guo, C., Zhang, Y., Wang, X., et al.: Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. Nature Communications **16**(1), 5493 (2025)
22. Pani, A., Yang, Y.: Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. arXiv preprint arXiv:2510.21356 (2025)
23. 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)

24. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision research **45**(18), 2397–2416 (2005)
25. Saltz, J., Sharma, A., Iyer, G., Bremer, E., Wang, F., Jasniewski, A., DiPrima, T., Almeida, J.S., Gao, Y., Zhao, T., et al.: A containerized software system for generation, management, and exploration of features from whole slide tissue images. Cancer research **77**(21), e79–e82 (2017)
26. Sun, Y., Si, Y., Zhu, C., Zhang, K., Shui, Z., Ding, B., Lin, T., Yang, L.: Cpathagent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. arXiv preprint arXiv:2505.20510 (2025)
27. Thai, V., Li, R., Ling, M., Jiang, S., Wolfe, J., Machiraju, R., Hu, Y., Li, Z., Parwani, A., Chen, J.: Eye-tracking, mouse tracking, stimulus tracking, and decision-making datasets in digital pathology. arXiv preprint arXiv:2510.24653 (2025)
28. 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)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, **L**., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
30. Xue, R., Le, H., Xu, J., Mondal, S., Leite, A., Zelinsky, G., Hoai, M., Samaras, D.: Personalized image descriptions from attention sequences. arXiv preprint arXiv:2512.06662 (2025)
31. Zhang, L., Yun, B., Li, Q., Wang, Y.: Historical report guided bi-modal concurrent learning for pathology report generation. In: MICCAI. pp. 343–352. Springer (2025)
32. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. NeurIPS **36**, 46595–46623 (2023)
33. Zuley, M.L., Jarosz, R., Drake, B.F., Rancilio, D., Klim, A., Rieger-Christ, K., Lemmerman, J.: Radiology data from the cancer genome atlas prostate adenocarcinoma [tcga-prad] collection. Cancer Imaging Arch **9**(10.7937), K9 (2016)