Think Global, Look Focal: Dual-Stream Retrieval-Augmented Pathology Report Generation for Whole Slide Images

Liang Xiong[1⋆], Yi Zhang[1⋆], HaoJie Song[2], QianYao Peng[1], and ShaoGuo Cui[1(✉)]

[1] School of Computer and Information Science, Chongqing Normal University, Chongqing, China
2024210516130@stu.cqnu.edu.cn, {zhangyii, csg}@cqnu.edu.cn
[2] School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, Guangdong, China

Abstract. Automatically generating pathology reports from gigapixel Whole Slide Images (WSIs) is essential for reducing pathologist workload and improving clinical efficiency. However, the gigapixel resolution and semantic complexity of WSIs expose two critical limitations in existing methods: (1) Visual dilution: Predominant approaches treat WSIs as single-granularity patch sequences, causing fine-grained focal details to be overwhelmed by background noise; (2) Retrieval noise: Existing methods rely on raw corpora with non-diagnostic redundancy, and their fusion mechanisms fail to suppress interference from low-similarity evidence, leading to factual hallucinations. To address these issues, we propose DR-Gen, a Dual-Stream Retrieval-Augmented Pathology Report Generation framework. Inspired by pathologists' workflow, DR-Gen decouples global tissue context from focal cellular details via two visual streams: a global stream for slide-level architecture and a focal stream for diagnostic regions. A Visual Priming Adapter leverages these decoupled features as visual anchors, grounding generation in slide morphology. To safeguard medical validity, we construct a CAP-protocol-guided diagnostic knowledge bank and propose an Adaptive Confidence Gating module that dynamically reweights evidence by retrieval confidence. Experiments on the PathText (BRCA) benchmark show that DR-Gen consistently outperforms existing methods in both report quality and subtype classification, achieving a relative improvement of 10.4% in BLEU-4. Ablation studies further validate each proposed component. Code is publicly available at https://github.com/jasonlightx/DR-Gen.

Keywords: Whole Slide Images · Pathology Report Generation · Image Captioning.

⋆ Equal contribution.

1 Introduction

Histopathological assessment using Whole Slide Images (WSIs) remains the gold standard for tumor diagnosis and staging, yet manual interpretation of these gigapixel-scale images is both labor-intensive and highly subjective [3,4,21]. Automated pathology report generation has thus emerged as a significant research focus. Technically, this field traces its evolution from natural Image Captioning, progressing from early CNN-RNN encoder-decoder architectures [1,13,24,26] to modern Transformer-centric sequence generation models [8, 10, 11, 23, 27]. This paradigm has achieved notable success in Radiology Report Generation (RRG), yielding landmark works like R2Gen [7] that enhance vision-language alignment via memory networks. Recently, to combat factual hallucinations and domain knowledge deficiency, incorporating retrieval-augmented and knowledge-driven mechanisms has become a prevailing trend in RRG [15, 17, 25, 28].

However, transferring these radiological methods to computational pathology encounters a fundamental paradigm mismatch: radiological images are typically single-frame, semantically dense, and low-resolution, whereas pathological WSIs are gigapixel-scale inputs characterized by extreme semantic sparsity—diagnostically valuable morphological cues (e.g., mitotic figures) often occupy less than 0.1% of the tissue area [12], deeply embedded within complex stromal and background structures. Existing WSI report generation methods such as HistGen [9] and Wsicaption [5] treat the entire slide as a single patch sequence. This indiscriminate all-in-one strategy feeds tens of thousands of visual tokens into the encoder, entangling global tissue architecture with focal lesions and severely diluting critical micro-morphological details with background noise. Furthermore, existing methods typically rely on randomly initialized queries. Lacking explicit semantic directionality, these queries struggle to localize sparse diagnostic regions, thereby exacerbating visual dilution and early-layer attention diffusion. To mitigate data scarcity and improve factual accuracy, recent works like BiGen [29] introduce Retrieval-Augmented Generation (RAG) [14]. However, existing RAG frameworks suffer from an inherent noise propagation chain: they rely on unstructured raw corpora laden with non-diagnostic redundancies, and employ quality-unaware fusion strategies. Consequently, low-confidence retrieval noise is indiscriminately absorbed into the generative context, inducing factual hallucinations inconsistent with the actual slide morphology.

To overcome these bottlenecks, we propose DR-Gen (Dual-stream Retrieval-augmented report Generation). By decoupling visual perception, our framework emulates the "think global, look focal" diagnostic paradigm of pathologists—who first evaluate the overall tissue architecture, subsequently pinpoint focal lesions, and ultimately synthesize these findings with established medical knowledge to formulate the diagnostic report. Our core contributions are as follows:

- We propose a global-focal dual-stream architecture with a Visual Priming Adapter (VPA). By injecting slide-specific morphological anchors into initial queries, VPA replaces random initialization with semantic directionality, suppressing early-layer attention diffusion and hallucinations.

– We design a CAP-protocol-guided knowledge bank and an Adaptive Confidence Gating (ACG) module. ACG calibrates retrieved evidence via semantic confidence scores, curbing non-diagnostic redundancy and retrieval noise.
– Extensive experiments on the PathText (BRCA) benchmark demonstrate state-of-the-art performance in both natural language generation metrics (BLEU-4: **0.138**, ROUGE-L: **0.298**) and downstream clinical subtype prediction (F1: **0.926**).

2 Method

2.1 Dual-Stream Visual Decomposition

I. Global Stream. Given patch-level features $X = \{x_i\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$, a two-layer perceptron f_{attn} computes a diagnostic relevance score s_i for each patch. After Softmax normalization, the attention weights α_i produce the global feature by weighted aggregation: $z_g = \sum_{i=1}^{N} \alpha_i x_i \in \mathbb{R}^d$.
II. Focal Stream. Using the same weights α_i, we retain the top-$K = \lfloor \kappa \cdot N \rfloor$ scoring patches (κ is a retention ratio) to form the local focal feature set: $z_f = \{x_i \mid i \in I_{\text{top}}\} \in \mathbb{R}^{K\times d}$.
III. Hybrid Visual Context. We concatenate both stream outputs along the sequence dimension to form a unified key-value sequence $\mathbf{H}_{\text{vis}} \in \mathbb{R}^{(1+K)\times d}$, serving as the shared multi-granularity feature source for downstream queries.

2.2 Visual Priming Adapter

To prevent the diffuse attention and factual hallucinations caused by randomly initialized queries, we propose the Visual Priming Adapter (VPA). VPA replaces random initialization with slide-specific visual anchors to provide semantic directionality for visual queries. We extract two anchors from the dual-stream decomposition: the global anchor $a_g = z_g \in \mathbb{R}^{1\times d}$, taken directly from the global stream, and the focal anchor $a_f = \frac{1}{K}\sum_{i=1}^{K} z_f^{(i)} \in \mathbb{R}^{1\times d}$, computed as the mean of all focal patches. They are concatenated into the anchor matrix $\mathbf{A} = [a_g; a_f] \in \mathbb{R}^{2\times d}$. VPA performs additive fusion between a nonlinear projection of the anchors and a learnable bias, allowing the model to interpolate between data-driven visual priors and task-specific semantic adjustments. The anchor matrix is projected through a two-layer MLP with intermediate layer normalization and GELU activation, added element-wise with a learnable bias $\mathbf{E}_{\text{bias}} = [e_g; e_f] \in \mathbb{R}^{2\times d}$ (where e_g and e_f are the learnable bias terms for the global and focal streams, respectively), and layer-normalized to yield the initial visual queries:

$$\mathbf{Q}_v^{(0)} = \left[\mathbf{Q}_g^{(0)}; \mathbf{Q}_f^{(0)}\right] = \mathrm{LN}\big(\mathrm{MLP}(\mathbf{A}) + \mathbf{E}_{\text{bias}}\big) \in \mathbb{R}^{2\times d} \tag{1}$$

where $\mathrm{MLP}(\cdot) = W_2\,\mathrm{GELU}(\mathrm{LN}(W_1(\cdot)))$ and $W_1, W_2 \in \mathbb{R}^{d\times d}$ are projection weights. This initialization biases the global query toward architecture-level regions and the focal query toward cytology-level patches, while the learnable bias retains flexibility for task-specific adaptation.

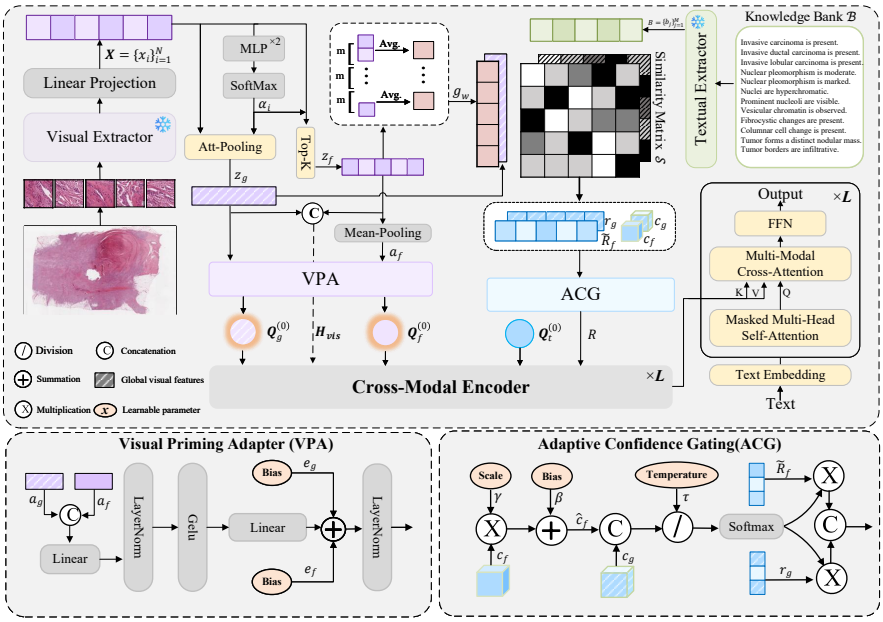

Fig. 1. The architecture of DR-Gen. Visual features are decoupled into global (z_g) and focal (z_f) streams. Visually, they serve as VPA anchors to initialize visual queries ($\mathbf{Q}_g^{(0)}, \mathbf{Q}_f^{(0)}$). Textually, they retrieve diagnostic priors calibrated by ACG, yielding purified evidence ($\mathbf{R}$). Finally, all components enter the Cross-Modal Encoder, where the visual and textual queries respectively cross-attend with the hybrid visual context ($\mathbf{H}_{vis}$) and $\mathbf{R}$ to drive report generation.

2.3 Confidence-Gated Knowledge Retrieval

I. CAP-Guided Atomic Knowledge Bank. We construct a diagnostic memory bank $\mathcal{B} = \{b_j\}_{j=1}^{M} \in \mathbb{R}^{M\times d}$, where each entry b_j is the encoded representation of an atomic diagnostic statement decomposed from source reports following the College of American Pathologists (CAP) protocol.

II. Dual-Stream Retrieval. Corresponding to the dual-stream visual encoding, retrieval operates independently at two granularities. For the local focal features $z_f \in \mathbb{R}^{K\times d}$, considering that spatially adjacent patches in a WSI typically belong to the same tissue type, we uniformly partition them into $W = \lceil K/m \rceil$ tissue regions (m is the region size) and compute the mean of each group to obtain region representations $g_w \in \mathbb{R}^d$. We then compute the cosine similarity between each region feature and all entries in the knowledge bank, select the top-v most similar entries, and average them to obtain the retrieval result $\bar{r}_w$ for that region. These W region-level results are passed through a linear projection layer to form a sequence of local retrieval features $\tilde{\mathbf{R}}_f \in \mathbb{R}^{W\times d}$. Notably, we take the maximum of all region-level mean similarities as the confidence score c_f rep-

resenting the entire focal stream, ensuring that high-confidence focal diagnostic cues dominate the weighting of the focal stream.

For the global feature $z_g \in \mathbb{R}^d$, we directly retrieve the top-n most similar entries from the knowledge bank and average them to obtain the global retrieval feature $r_g \in \mathbb{R}^{1\times d}$ and its confidence c_g.

III. Adaptive Confidence Gating. The two streams operate at different retrieval granularities, introducing systematic bias in raw confidences—the focal stream yields lower cosine similarities due to finer-grained queries. We apply learnable affine calibration to the focal confidence: $\hat{c}_f = \gamma \cdot c_f + \beta$, where γ, β are learnable scalars, keeping global confidence unchanged ($\hat{c}_g = c_g$). The calibrated scores become fusion weights via temperature-scaled Softmax:

$$[\omega_f,\ \omega_g] = \mathrm{Softmax}\left(\frac{[\hat{c}_f,\ \hat{c}_g]}{\tau}\right) \tag{2}$$

where τ is a learnable temperature parameter controlling weight allocation sharpness: a smaller τ favors the higher-confidence stream, while a larger τ encourages balanced fusion. The gated retrieval evidence is concatenated as:

$$\mathbf{R} = [\omega_f \cdot \tilde{\mathbf{R}}_f;\ \omega_g \cdot \mathbf{r}_g] \in \mathbb{R}^{(W+1)\times d} \tag{3}$$

2.4 Cross-Modal Encoding

To avoid the prohibitive $\mathcal{O}(N^2)$ complexity of applying global self-attention directly over tens of thousands of WSI patches, based on [29], we design a Cross-Modal Encoder driven by learnable queries. These query vectors act as information bottlenecks to efficiently extract and fuse visual and textual features.

I. Learnable Visual Query. The visual queries $\mathbf{Q}_v \in \mathbb{R}^{2\times d}$ extract morphological features from the slide. Starting from the VPA-initialized $\mathbf{Q}_v^{(0)}$, they are iteratively updated at each encoder layer via cross-attention with $\mathbf{H}_{\text{vis}}$:

$$\mathbf{Q}_v^{(l)} = \mathrm{FFN}\Big(\mathrm{SelfAttn}\big(\mathrm{FFN}(\mathrm{CrossAttn}(\mathbf{Q}_v^{(l-1)},\ \mathbf{H}_{\text{vis}},\ \mathbf{H}_{\text{vis}}))\big)\Big) \tag{4}$$

where $l = 1, \ldots, L$ is the layer index. Cross-attention allows each query to selectively attend to the most relevant regions of $\mathbf{H}_{\text{vis}}$ given its current state, while self-attention enables information exchange between the global and focal queries. After L layers, $\mathbf{Q}_v^{(L)}$ encodes multi-granularity visual information.

II. Learnable Textual Query. A learnable textual query $\mathbf{Q}_t \in \mathbb{R}^{1\times d}$ is designated to incorporate external diagnostic knowledge. Following the same structure as the visual query refinement, the gated retrieval evidence $\mathbf{R}$ from ACG is injected into $\mathbf{Q}_t$ via cross-attention exclusively at the second encoder layer, with $\mathbf{R}$ serving as Key/Value in place of $\mathbf{H}_{\text{vis}}$. In subsequent layers, $\mathbf{Q}_t$ receives no further retrieval input and is refined only through self-attention for internal consolidation, preventing cascading noise amplification. Finally, $\mathbf{Q}_v^{(L)}$ and $\mathbf{Q}_t^{(L)}$ are concatenated into a unified key-value representation and fed into the decoder to drive autoregressive pathology report generation.

2.5 Objective Function

The ultimate goal of DR-Gen is to generate a pathology report $Y = \{y_n\}_{n=1}^{N}$ of length N. The entire framework is trained end-to-end by minimizing the negative log-likelihood (NLL) loss. The objective function is formulated as follows:

$$\mathcal{L} = -\sum_{n=1}^{N} \log P\big(y_n \mid \{y_i\}_{i<n}, \mathbf{H}_{\text{vis}}, \mathbf{R}\big) \tag{5}$$

where $\{y_i\}_{i<n}$ denotes the previously generated tokens, $\mathbf{H}_{\text{vis}}$ represents the hybrid visual context derived from the dual-stream decomposition, and $\mathbf{R}$ indicates the confidence-gated retrieval evidence.

3 Experiments

Dataset and Evaluation Metrics. We evaluate DR-Gen on the PathText (BRCA) benchmark to validate its effectiveness in complex pathological scenarios. Constructed by Wsicaption [5], PathText (BRCA) is the largest and most representative subset of the PathText dataset. It contains 1,041 cleaned and aligned WSI-report pairs. We strictly follow its official data split: 845 pairs for training, 98 for validation, and 98 for testing. For evaluation, we employ standard NLP metrics: BLEU [22], METEOR [2], and ROUGE [16] to assess generation quality, and Fact_{ent} [20] to measure completeness and consistency of biological entities in the generated reports. Additionally, Precision, Recall, and F1 score are used as classification metrics to evaluate carcinoma subtyping prediction performance.

Model Setting. Following [5], WSIs are segmented by CLAM [19] into non-overlapping 256×256 patches at $10\times$ magnification. Both visual and text features are extracted by CONCH [18], pretrained on 1.17M histopathology image-text pairs to align pathology images and medical text in a shared semantic space. DR-Gen adopts an encoder and a decoder with 3 layers each, featuring 4 attention heads and an embedding dimension of 512. The focal stream filtering ratio is $\kappa = 0.6$; retrieval parameters are $m = 80$, $v = 5$, and $n = 5$. The model is optimized with Adam (lr 1e-4, weight decay 5e-5) and decoded via beam search (beam size 3). All experiments are run on a single NVIDIA RTX 5000 Ada (32GB) GPU.

4 Results and Analysis

Performance Comparison. We compare DR-Gen with seven image captioning methods in Table 1. Foundational architectures, such as CNN-RNN [24] and Vanilla Transformer [23], along with radiology-specific R2Gen [7] and R2Gen-CMN [6], exhibit suboptimal performance. This stems from their inability to handle the extreme semantic sparsity and visual dilution inherent in gigapixel WSIs. While pathology-tailored methods, including HistGen [9], MI-Gen [5], and

Table 1. Performance comparison of DR-Gen with existing methods on the PathText (BRCA) test set. Bold: best score. Underline: second best.

Model	NLG metrics							Classification		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	Fact_{ent}	Precision	Recall	F1
CNN-RNN [24]	0.341	0.211	0.122	0.069	0.135	0.246	0.397	0.623	0.612	0.617
Vanilla Transformer [23]	0.390	0.231	0.135	0.093	0.142	0.254	0.453	0.648	0.665	0.657
R2Gen [7]	0.378	0.243	0.160	0.076	0.163	0.251	0.421	0.671	0.692	0.681
R2GenCMN [6]	0.390	0.229	0.128	0.074	0.154	0.248	0.470	0.835	0.873	0.854
HistGen [9]	0.407	0.250	0.149	0.090	0.146	0.264	0.466	0.821	0.855	0.838
MI-Gen [5]	0.408	0.262	0.171	0.117	0.164	<u>0.287</u>	0.470	0.865	0.805	0.834
BiGen [29]	<u>0.437</u>	<u>0.282</u>	<u>0.186</u>	<u>0.125</u>	<u>0.172</u>	0.282	<u>0.494</u>	<u>0.897</u>	<u>0.884</u>	<u>0.891</u>
DR-Gen (Ours)	**0.441**	**0.291**	**0.198**	**0.138**	**0.173**	**0.298**	**0.519**	**0.958**	**0.896**	**0.926**

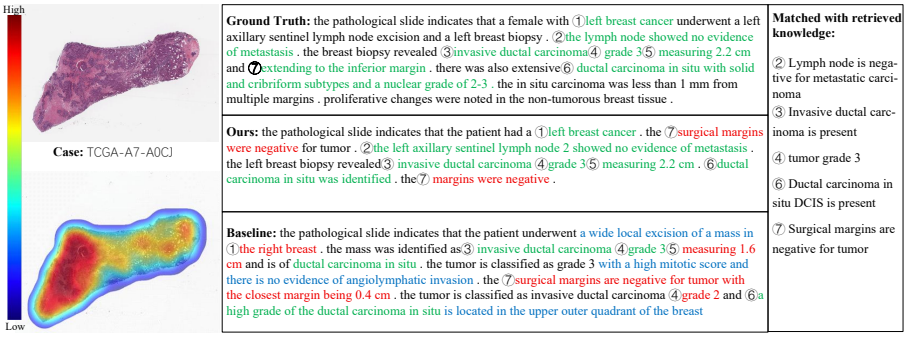

Fig. 2. Visualization: high and low attention weights are indicated by red and blue regions, respectively. Report: correct, incorrect, and hallucinated descriptions are highlighted in green, red, and blue, respectively.

the retrieval-augmented BiGen [29], show improvements, they still struggle with background redundancy and non-diagnostic retrieval noise.

In contrast, DR-Gen effectively mitigates these challenges by integrating multi-granularity visual features with high-quality medical priors, strictly anchoring generation in reliable facts. Consequently, our model elevates BLEU-4 to 0.138 (+10.4% versus BiGen [29]). Significant gains in Fact_{ent} and classification metrics demonstrate that DR-Gen not only ensures linguistic fluency but also substantially improves factual correctness and clinical diagnostic reliability.

Qualitative Analysis. To qualitatively assess our proposed method, we compare the generated reports with the ground truth and Baseline outputs in Fig. 2. Our method accurately captures core pathological entities while providing precise, reliable descriptions of tumor grade and exact size. In contrast, the Baseline fails to maintain clinical coherence, generating severe hallucinations. These results illustrate our model's superior performance; it effectively integrates retrieved diagnostic knowledge to offer comprehensive, accurate pathological descriptions, thereby improving overall diagnostic reliability.

Ablation Study. Table 2 presents the ablation study results for DR-Gen. Compared to the baseline model Vanilla Transformer [23], the progressive integration of the Dual-Stream (DS) visual decomposition and the VPA module yields steady

Table 2. Ablation study on key components. Focal and Global denote the two streams of the Dual-Stream (DS) visual decomposition. RAG indicates retrieval-augmented generation via direct text concatenation.

DS		VPA	RAG	ACG	Metric						
Focal	Global				BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	Fact_{ent}
					0.390	0.231	0.135	0.093	0.142	0.254	0.453
✓					0.401	0.265	0.163	0.103	0.164	0.271	0.471
	✓				0.396	0.251	0.162	0.108	0.159	0.262	0.462
✓	✓				0.409	0.267	0.176	0.113	0.160	0.275	0.478
✓	✓	✓			0.422	0.274	0.183	0.127	0.169	0.282	0.491
✓	✓	✓	✓		0.418	0.271	0.179	0.123	0.170	0.276	0.482
✓	✓	✓	✓	✓	**0.441**	**0.291**	**0.198**	**0.138**	**0.173**	**0.298**	**0.519**

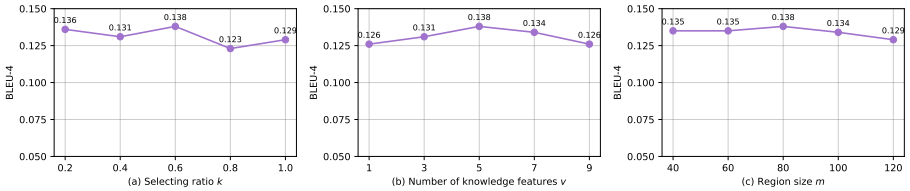

Fig. 3. Impact of varying (a) the selection ratio κ, (b) the number of knowledge features v, and (c) the region size m on model performance.

relative improvements of 21.5% and 12.4% in BLEU-4, respectively, validating the necessity of capturing multi-scale contexts. Notably, the direct introduction of naive retrieval-augmented generation (RAG) degrades performance, as non-diagnostic textual noise severely disrupts the decoding process. By replacing naive concatenation with our ACG, the model effectively filters out irrelevant priors. Ultimately, by deeply aligning visual features with purified textual priors, the model achieves optimal performance.

Fig. 3 shows three hyper-parameter sensitivities, with default values $\kappa = 0.6$, $v = 5$ and $m = 80$. As shown in Fig. 3(a), performance remains stable across $[0.2, 0.6]$ but degrades notably when $\kappa \geq 0.8$: retaining too many patches causes the focal stream to degenerate into an indiscriminate all-in-one representation, defeating the purpose of dual-stream decomposition. As shown in Fig. 3(b), performance peaks at $v = 5$ and deteriorates at both extremes, reflecting a fundamental trade-off between knowledge richness and retrieval noise. As shown in Fig. 3(c), performance stays robust within $m \in [40, 80]$ and declines gradually beyond this range, suggesting that moderately fine region granularity strikes the best balance between preserving local diagnostic cues and computational efficiency.

5 Conclusion

In this paper, we propose a dual-stream retrieval-augmented pathology report generation framework to tackle the critical challenges of visual dilution and retrieval noise in gigapixel WSIs. Specifically, on the visual end, our method extracts multi-granularity features and provides clear semantic directionality; on the textual end, it filters out non-diagnostic retrieval noise to inject reliable medical priors. Extensive experiments on the PathText (BRCA) benchmark demonstrate that DR-Gen achieves state-of-the-art performance in both report quality and clinical subtype classification. These results validate the effectiveness of integrating multi-granularity visual perception and retrieval noise filtering for generating accurate and reliable pathology reports.

Acknowledgments. This work was supported by the Key Project of Chongqing Natural Science Foundation Innovation Development Joint Fund (Grant No. CSTB2025NSCQ-LZX0073), the Chongqing Normal University Foundation (Grant No. 24XLB006), the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202400539), and the National Natural Science Foundation of Chongqing (General Program) (Grant No. CSTB2025NSCQ-GPX0253).

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

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. 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)
3. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A.: Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature reviews Clinical oncology **16**(11), 703–715 (2019)
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
5. 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)
6. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). pp. 5904–5914 (2021)

7. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 1439–1449 (2020)
8. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
9. Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H.: Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–199. Springer (2024)
10. He, S., Liao, W., Tavakoli, H.R., Yang, M., Rosenhahn, B., Pugeault, N.: Image captioning through image transformer. In: Proceedings of the Asian conference on computer vision (2020)
11. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. Advances in neural information processing systems **32** (2019)
12. Jahanifar, M., Shephard, A., Zamanitajeddin, N., Graham, S., Raza, S.E.A., Minhas, F., Rajpoot, N.: Mitosis detection, fast and slow: robust and efficient detection of mitotic figures. Medical Image Analysis **94**, 103132 (2024)
13. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
15. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3334–3343 (2023)
16. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
17. Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 269–280 (2021)
18. 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)
19. 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)
20. 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)
21. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The lancet oncology **20**(5), e253–e261 (2019)
22. 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)

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
25. Wang, F., Du, S., Yu, L.: Hergen: Elevating radiology report generation with longitudinal data. In: European Conference on Computer Vision. pp. 183–200. Springer (2024)
26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
27. Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. IEEE transactions on circuits and systems for video technology **30**(12), 4467–4480 (2019)
28. Yun, H., Maeng, J., Kang, E., Suk, H.I.: Diff-rrg: Longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 152–161. Springer (2025)
29. Zhang, L., Yun, B., Li, Q., Wang, Y.: Historical report guided bi-modal concurrent learning for pathology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 343–352. Springer (2025)