Prior-Anchored Debiasing for Long-Tailed Multi-Organ Pathology Report Generation

Feng Yang[1], Jie Liu[1], Yubo Pang[1], Peilin Chen[1], Xinheng Lyu[2], Shiqi Wang[1✉], Howard Leung[1✉], and Ping Chen[3]

[1] City University of Hong Kong, Hong Kong SAR
[2] University of Nottingham, United Kingdom
[3] University of Massachusetts Boston, United States
fenyang6-c@my.cityu.edu.hk, shiqwang@cityu.edu.hk, ping.chen@umb.edu

Abstract. Automated pathology report generation from Whole Slide Images (WSIs) has attracted increasing attention in digital pathology. However, existing methods are predominantly developed under single-organ settings, overlooking the multi-organ scenarios encountered in clinical practice, where organ types typically follow a long-tailed distribution. To address this gap, we identify two critical biases: (1) visual representation bias, where the encoder favors head-class patterns over tail-class discriminative features, and (2) textual decoding bias, where the decoder overfits to head-class narrative patterns, yielding diagnostically unreliable outputs for tail-class organs. To mitigate these two biases, we propose a novel **Pri**or-anchored multi-**Or**gan pathology report **Gen**eration framework (PriOrGen). Specifically, a Visual-Prototype Anchored Bottleneck module leverages the information bottleneck principle with learnable anchor representations to selectively retain diagnostically relevant visual information while filtering out head-biased redundancy. Secondly, a Meta-Report Anchored Bank module constructs an organ-specific meta-report anchored bank and retrieves organ-faithful textual priors to steer the decoder away from head-class narrative patterns. Extensive experiments on a multi-organ pathology dataset demonstrate that our method effectively mitigates long-tail biases and achieves superior report generation performance across both head and tail organ categories compared to state-of-the-art methods. Code is publicly available at https://github.com/yangfeng0216/PriOrGen.

Keywords: Whole Slide Image · Pathology Report Generation · Long-Tail Learning.

1 Introduction

In the rapidly evolving field of digital pathology, the automated analysis of Whole Slide Images (WSIs) has become a pivotal area of research [8,15,3]. Among its various applications, pathology report generation aims to directly translate gigapixel-level visual patterns into structured, diagnostic-aware natural language. However, existing pathology report generation methods [4,9,21,26]

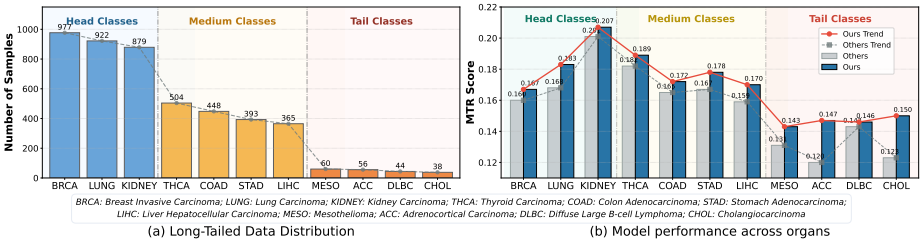

Fig. 1. (a) Long-tailed data distribution. The sample count per organ type exhibits a long-tailed distribution grouped into head, medium, and tail classes. **(b) Model performance across organs**. We compare the MTR score between our method and others.

have been predominantly designed and validated under *single organ* settings, where models are trained and evaluated on WSIs from the specific organ.

In real-world clinical scenarios, WSIs are derived from *multiple organs* rather than a *single organ*. To address this gap, this work investigates pathology report generation in a multi-organ setting, where organ types follow a long-tailed distribution (Fig. 1(a)). Recent efforts have tackled long-tailed distributions in medical image classification [11,18,2,17], where the output space is a fixed set of categorical labels. However, the impact of long-tailed distributions on pathology report generation remains largely unexplored. Unlike classification, report generation requires producing free-form texts that faithfully capture organ-specific characteristics, which renders the direct adoption of existing long-tail classification strategies insufficient.

Specifically, this introduces two biases into the pathology report generation pipeline. First, **visual representation bias**: when trained on long-tail data, the visual encoder tends to learn feature representations that are predominantly aligned with the histomorphological patterns of head-class organs [13,14]. Consequently, the discriminative visual cues of tail-class organs, which often exhibit distinct yet underrepresented morphologies, are inadequately captured, leading to degraded visual grounding for downstream report generation. Second, **textual decoding bias**: the language decoder, exposed to an overwhelming volume of head-class reports during training, tends to overfit to their dominant narrative patterns, including recurring sentence templates, high-frequency diagnostic phrases, and organ-specific terminologies [7,16]. As a result, when generating reports for tail-class organs, the decoder is prone to producing linguistically fluent but diagnostically erroneous outputs that implicitly borrow the stylistic and semantic patterns from head classes, undermining clinical reliability.

To address the above challenges, our method introduces two key components: (1) We propose a **Visual-Prototype Anchored Bottleneck** module that leverages the information bottleneck principle with learnable anchor representations to selectively filter out redundant or head-biased visual information, retaining diagnostically relevant patches from each WSI for more balanced and

clinically grounded visual representations. (2) We propose a **Meta-Report Anchored Bank** module that constructs an organ-specific meta-report anchored bank by distilling representative report templates from the training reports of each organ type, and retrieves the most relevant meta-reports to provide the decoder with organ-faithful textual priors, steering the generation away from head-class narrative patterns. The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, we propose the first framework that explicitly addresses the long-tail distribution bias in multi-organ pathology report generation, which manifests in two critical aspects: visual representation bias and textual decoding bias.
- We propose a dual prior mechanism consisting of a Visual-Prototype Anchored Bottleneck (VPAB) and a Meta-Report Anchored Bank (MRAB), which jointly mitigate long-tail bias from two perspectives.
- Extensive experiments on the constructed ML-Path dataset demonstrate that our method achieves significant improvements over existing baselines, particularly on tail-class organs.

2 Method

2.1 Overview

As illustrated in Fig. 2, our framework consists of three key components. Given patch features extracted from a WSI, the Visual-Prototype Anchored Bottleneck (VPAB, Sec. 2.3) performs anchor-guided compression to selectively retain diagnostically relevant patches while suppressing head-biased redundancy. The Meta-Report Anchored Bank (MRAB, Sec. 2.4) retrieves organ-faithful textual priors from a pre-constructed meta-report anchored bank to mitigate textual decoding bias. Finally, a cross-modal decoder (Sec. 2.5) integrates the compressed visual representations and retrieved textual priors to generate diagnostic reports.

2.2 Problem Formulation

Following standard practice in computational pathology [4,26], each WSI $\mathcal{W}$ is partitioned into N non-overlapping patches at 10× magnification. Each patch is encoded by a pre-trained pathology foundation model, yielding patch features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{N \times d}$. Each WSI is associated with an organ label $o \in \mathcal{O} = \{o_1, \ldots, o_C\}$, where C is the total number of organ categories and the training set follows a long-tail distribution across organs. The goal is to generate an L-length diagnostically accurate report $\mathbf{Y} = \{\mathbf{y}_l\}_{l=1}^{L}$.

2.3 Visual-Prototype Anchored Bottleneck

A WSI typically contains thousands of patches, many of which are redundant or diagnostically irrelevant, particularly diluting the discriminative information

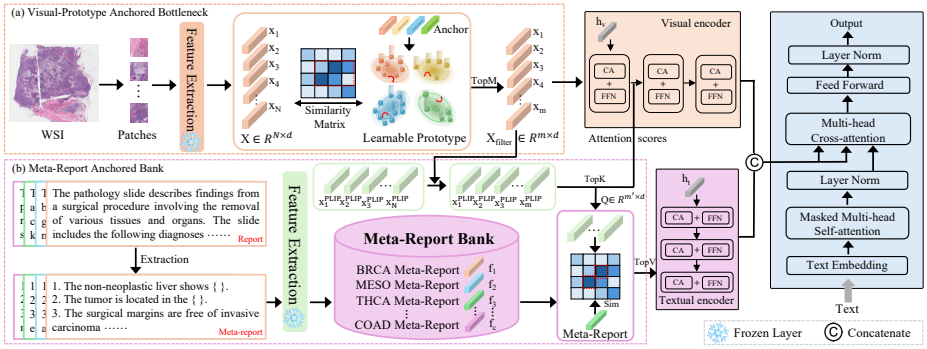

Fig. 2. Overview of our proposed model PriOrGen. CA and FFN refer to Cross-Attention and Feed-Forward Network. (a) Visual-Prototype Anchored Bottleneck module. (b) Meta-Report Anchored Bank module.

of tail-class organs. Inspired by the information bottleneck (IB) principle [20], we propose a Visual-Prototype Anchored Bottleneck (VPAB) that compresses redundant patch features while preserving diagnostically critical information, where prototypes clustered from labeled multi-organ WSI data serve as semantic anchors to guide the compression toward discriminative tissue patterns.

I. Learnable Diagnostic Prototypes: Directly applying the Variational Information Bottleneck [1] to WSIs is intractable, as each WSI constitutes a bag of thousands of patch instances and modeling the full posterior $p(z|\mathbf{x})$ over such high-dimensional bags is computationally prohibitive [27]. We introduce a set of K learnable diagnostic prototypes $\mathbf{P} = \{\mathcal{N}(\hat{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ to approximate the bag-level distribution and convert the intractable IB optimization into a tractable prototype-guided selection process. The prototype means $\{\boldsymbol{\mu}_k\}$ and covariances $\{\boldsymbol{\Sigma}_k\}$ are both learnable parameters. To provide semantically meaningful supervision, we extract region-of-interest patch-level features across diverse organs. Then, k-means clustering is performed over the resulting feature pool to obtain K centroids $\{\mathbf{c}_k\}_{k=1}^{K}$, which capture representative tissue patterns in the histopathological feature space.

The *compression* term $I(Z, X)$ in the IB framework is realized by regularizing each prototype toward its domain-informed prior $\mathcal{N}(\mathbf{c}_k, \mathbf{I})$ via KL divergence:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{k=1}^{K} \mathrm{KL}\big[\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,\|\, \mathcal{N}(\mathbf{c}_k, \mathbf{I})\big]. \tag{1}$$

The *predictive* term $I(Z, Y)$ is fulfilled by aligning prototype samples with clinical report semantics. We sample $\hat{\mathbf{z}}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ via Monte Carlo sampling and project it through an MLP layer to match the dimension of the text embedding space, then maximize its similarity with the corresponding organ report

embedding $\mathbf{e}_Y$ from report ground truth using a pre-trained text encoder :

$$\mathcal{L}_{\text{report}} = -\frac{1}{K}\sum_{k=1}^{K}\text{sim}(\hat{\mathbf{z}}_k, \mathbf{e}_Y), \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity.

The overall Visual-Prototype Anchored Bottleneck objective combines the two terms as: $\mathcal{L}_{\text{VPAB}} = \mathcal{L}_{\text{report}} + \beta\,\mathcal{L}_{\text{KL}}$, where β is a trade-off hyperparameter controlling the compression–prediction trade-off. Minimizing $\mathcal{L}_{\text{VPAB}}$ encourages the prototypes to retain maximal diagnostic information from clinical reports while compressing task-irrelevant variations toward priors.

II. Prototype-Guided Patch Selection: Given patch features from a WSI $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{N\times d}$, we leverage the learned prototypes to identify diagnostically relevant patches. Specifically, we sample $\hat{\mathbf{z}}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and compute the relevance score of each patch by max-pooling its cosine similarity over all prototypes:

$$r_i = \max_{k\in\{1,\ldots,K\}} \text{sim}(\mathbf{x}_i, \hat{\mathbf{z}}_k). \tag{3}$$

We rank all patches by their relevance scores $\{r_i\}_{i=1}^{N}$ and retain the top-m ($m = \rho \cdot N$, $\rho \in (0,1)$) to form the filtered representation $\mathbf{X}_{\text{filter}} \in \mathbb{R}^{m\times d}$. This yields a compact, diagnostically informative bag representation where head-biased redundant patches are suppressed.

2.4 Meta-Report Anchored Bank

Pathology reports within the same organ type naturally share recurrent linguistic patterns and standardized terminologies. We exploit this by constructing an organ-specific meta-report anchored bank and retrieving organ-faithful textual priors to steer the decoder away from head-class narrative patterns.

I. Meta-Report Anchored Bank Construction: To provide organ-faithful textual priors, we construct an organ-specific meta-report anchored bank from the training split of the dataset. For each organ type o, we first prompt a powerful LLM (Gemini-3 Pro) to extract M_o candidate meta-report templates from the training reports. To ensure representativeness, we compute the similarity between each candidate and the original reports from both biomedical-semantic (BioSimCSE) [12] and general-semantic (BGE) [24] perspectives, and fuse the two scores into a unified ranking. Only the top-M_o candidates with the highest fused scores are retained as meta-report templates. We further manually verify the selected templates for each organ to ensure clinical accuracy. The construction procedure and the meta report are provided in our code to ensure reproducibility. The resulting knowledge bank is denoted as $\mathbf{B} = \{\mathbf{b}_o\}_{o=1}^{C}$, where $\mathbf{b}_o = \{t_1^o, \ldots, t_{M_o}^o\}$. Each template is encoded by the PLIP [10] text encoder, yielding the organ-specific feature bank $\mathbf{F} = \{\mathbf{f}_o\}_{o=1}^{C}$ with $\mathbf{f}_o \in \mathbb{R}^{M_o\times d}$.

II. Knowledge Retrieval: To ensure cross-modal compatibility with the PLIP-encoded meta-reports, we re-extract patch features using the PLIP image encoder and apply the VPAB filtering indices to obtain $\mathbf{X}_{\text{filter}}^{\text{PLIP}} \in \mathbb{R}^{m\times d}$. Then we

leverage the cross-attention scores from the first decoder layer to further rank selected patches by diagnostic informativeness, retaining the top-ρ' proportion ($m' = \lfloor \rho' \cdot m \rfloor$) as the query set $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{m'} \in \mathbb{R}^{m' \times d}$. The cosine similarity between each query $\mathbf{q}_i$ and the meta-report features $\mathbf{f}_o$ is computed, and max-pooled across queries to score each template. The top-v templates are selected as the retrieved meta-reports $\mathbf{F}_{\text{filter}} \in \mathbb{R}^{v \times d}$.

2.5 Report Generation

Given the filtered visual features $\mathbf{X}_{\text{filter}}$ from VPAB and the retrieved meta-report features $\mathbf{F}_{\text{filter}}$ from MRAB, we adopt the bi-modal encoder [26] to further compress both modalities. Specifically, a learnable visual token $\mathbf{h}_v$ and a learnable textual token $\mathbf{h}_t$ iteratively attend to $\mathbf{X}_{\text{filter}}$ and $\mathbf{F}_{\text{filter}}$ respectively via cross-attention layers, yielding a compact joint representation $\mathbf{H} = [\mathbf{h}_v; \mathbf{h}_t] \in \mathbb{R}^{2 \times d}$. The decoder is trained with the negative log-likelihood loss $\mathcal{L}_{\text{NLL}} = -\sum_{l=1}^{L} \log P(\mathbf{y}_l \mid \{\mathbf{y}_i\}_{i<l}, \mathbf{H})$, where $\{\mathbf{y}_i\}_{i<l}$ denotes the previously generated tokens. Combined with the VPAB loss defined in Sec. 2.3, the overall training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \alpha \mathcal{L}_{\text{VPAB}}$, where α is a balancing hyperparameter.

3 Experiments and Results

3.1 Implementation Details

I. Datasets: We construct a Multi-organ Long-tailed Pathology report generation dataset (ML-Path) by selecting 11 cancer types from The Cancer Genome Atlas (TCGA) based on sample frequency. The resulting dataset comprises 4,686 WSI-report pairs, naturally exhibiting a long-tailed distribution. Specifically, we categorize the cancer types into three groups: Head types including BRCA, LUNG, and KIDNEY with 879–977 samples each, Medium types including THCA, COAD, STAD, and LIHC with 365–504 samples each, and Tail types including MESO, ACC, DLBC, and CHOL with 38–60 samples each. The corresponding ground-truth reports are sourced from Chen et al. [4]. We randomly split the dataset into training, validation, and testing sets, and ensure no patient-level overlap across splits and long-tail property in all sets. The split can be found in our code.

II. Model Settings: Each WSI is partitioned into non-overlapping 256×256 tissue patches at 10× magnification using CLAM [15], and patch-level features are extracted by UNI [5]. The number of prototypes is set to $K = 4$, and the filtering ratios ρ and ρ' are set to 0.4. The number of retrieved meta-report templates v is set to 2. The loss balancing coefficients α and β are both set to 0.1. We use the Adam optimizer with a learning rate of 1e-4, weight decay of 5e-5, and beam search with a beam size of 3. All experiments are conducted on a single NVIDIA L40-48GB GPU.

Table 1. Results on long-tailed dataset across 11 organ types. B-M: BLEU-Mean (average of BLEU-1 to BLEU-4); MTR: METEOR; R-L: ROUGE-L. **Bold**: best; Underline: second best. In the Mean column, shaded cells denote significant improvements over the baseline (paired t-test, n=11 organs): $p < 0.001$, $p < 0.01$, $p < 0.05$; no shading: $p \geq 0.05$.

Method	BRCA			LUNG			KIDNEY			THCA			COAD			STAD		
	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L
CNN-RNN[23]	0.190	0.146	0.238	0.208	0.161	0.255	0.252	0.189	0.289	0.243	0.175	0.265	0.216	0.171	0.260	0.207	0.153	0.243
att-LSTM[25]	0.176	0.137	0.253	0.225	0.153	0.288	0.286	0.195	0.341	0.211	0.162	0.296	0.202	0.145	0.288	0.208	0.139	0.269
Transformer[22]	0.180	0.137	0.261	0.252	0.168	0.298	0.254	0.174	0.326	0.235	0.163	0.295	0.249	0.167	0.309	0.221	0.147	0.284
Wsicaption[4]	0.207	0.150	0.280	0.239	0.163	0.295	0.293	0.191	0.336	0.239	0.168	0.304	0.229	0.155	0.305	0.227	0.146	0.281
HistoCap[19]	0.187	0.144	0.276	0.242	0.168	**0.302**	0.278	0.190	**0.342**	0.232	0.163	0.304	0.218	0.153	0.297	0.241	0.155	**0.301**
R2Gen[7]	0.224	0.157	0.279	0.237	0.164	0.296	0.233	0.204	0.322	0.244	0.165	0.290	0.242	0.166	0.306	0.218	0.173	0.286
R2GenCMN[6]	0.198	0.148	0.279	0.232	0.167	0.299	0.279	**0.218**	0.338	0.245	0.172	0.306	0.254	0.164	0.304	0.255	0.173	0.296
BiGen[26]	0.240	0.160	0.287	0.256	0.168	0.296	0.279	0.201	0.319	0.279	0.182	0.311	0.251	0.165	0.293	0.277	0.167	0.300
Ours	**0.250**	**0.167**	**0.291**	**0.269**	**0.183**	**0.302**	**0.324**	0.207	**0.342**	**0.286**	**0.189**	**0.312**	**0.259**	**0.172**	**0.311**	**0.279**	**0.178**	0.290

Method	LIHC			MESO			ACC			CHOL			DLBC			Mean		
	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L
CNN-RNN[23]	0.206	0.151	0.245	0.156	0.122	0.210	0.176	0.139	0.245	0.168	0.139	0.216	0.152	0.092	0.226	0.215	0.164	0.250
att-LSTM[25]	0.229	0.158	0.279	0.163	0.103	0.241	0.177	0.109	0.237	0.156	0.120	0.242	0.160	0.084	0.214	0.221	0.156	0.287
Transformer[22]	**0.247**	0.163	0.293	0.183	0.117	0.247	0.188	0.118	0.259	0.176	0.136	0.247	0.161	0.113	0.242	0.230	0.157	0.293
Wsicaption[4]	0.225	0.152	0.291	0.185	0.116	0.258	0.189	0.118	0.264	0.184	0.130	0.261	0.168	0.131	0.241	0.239	0.162	0.298
HistoCap[19]	0.221	0.150	0.287	0.200	0.117	0.266	0.217	0.129	0.282	0.147	0.120	0.239	0.203	0.125	0.282	0.231	0.161	0.301
R2Gen[7]	0.204	0.153	**0.294**	0.150	0.114	0.232	**0.222**	0.146	**0.290**	0.161	0.126	0.231	0.200	0.143	0.246	0.247	0.170	0.295
R2GenCMN[6]	0.228	0.160	0.282	0.153	0.119	0.242	0.197	0.136	0.267	0.171	0.139	0.243	0.145	0.126	0.261	0.259	0.172	0.300
BiGen[26]	0.226	0.159	0.279	0.175	0.131	0.251	0.202	0.120	0.266	0.156	0.123	0.239	**0.230**	0.143	**0.287**	0.268	0.172	0.297
Ours	**0.247**	**0.170**	**0.294**	**0.203**	**0.143**	**0.273**	0.220	**0.147**	0.268	**0.213**	**0.150**	**0.272**	0.200	**0.146**	0.273	**0.273**	**0.181**	**0.305**

3.2 Multi-Organ Report Generation Results

For all comparative methods, we use exactly the same data split to ensure fairness of the experiments. We compare our method with eight existing approaches in Table 1, including two LSTM-based methods: CNN-RNN [23] and att-LSTM [25], and six Transformer-based methods: Transformer [22], R2Gen [7] and R2GenCMN [6] (both originally designed for radiology report generation), Wsicaption [4] and HistoCap [19] (designed for pathology report generation), and BiGen [26] which also serves as a naive retrieval baseline. Following [4,26], three widely used Natural Language Processing metrics are adopted: BLEU-Mean (the average of BLEU-1 to BLEU-4), METEOR, and ROUGE-L.

As shown in Table 1, our method achieves the best overall Mean scores across all three metrics. Existing methods such as BiGen [26] perform competitively on head-class organs but degrade notably on tail classes due to the lack of explicit long-tail handling. In contrast, our method yields substantial improvements on tail classes: on CHOL, we achieve 0.213 in B-M, surpassing Wsicaption [4] (0.184) by +15.8%; on MESO, we obtain 0.203 in B-M and 0.143 in MTR, outperforming BiGen [26] by +16.0% and +9.2%, respectively. Importantly, these tail-class gains do not compromise head-class performance, confirming that our framework effectively mitigates long-tail bias without sacrificing overall generation quality.

Discussion. We attribute the improvements to the complementary design of VPAB and MRAB. VPAB compensates for the scarce visual representations of tail-class organs by providing organ-specific prototypes, while MRAB prevents the decoder from defaulting to head-class expressions by injecting organ-faithful

Table 2. Ablation study of different components.

Components			Overall			Head			Medium			Tail		
BASE	MRAB	VPAB	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L	B-M	MTR	R-L
✓	✗	✗	0.259	0.170	0.290	0.266	0.176	0.297	0.254	0.165	0.285	0.202	0.121	0.240
✓	✓	✗	0.271	0.178	0.301	0.275	0.183	0.305	0.266	0.175	0.299	0.212	0.127	0.250
✓	✗	✓	0.268	0.174	0.295	0.276	0.181	0.298	0.261	0.171	0.296	0.211	0.132	0.245
✓	✓	✓	**0.273**	**0.181**	**0.305**	**0.278**	**0.186**	**0.308**	**0.270**	**0.178**	**0.302**	**0.220**	**0.141**	**0.264**

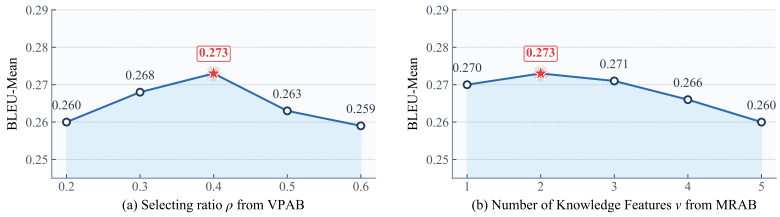

Fig. 3. Experimental results of BLEU-Mean with different parameters.

textual priors. The two modules jointly address the long-tail bias, explaining the consistent tail-class gains without head-class degradation.

3.3 Ablation Study

To validate the contribution of each component, we conduct ablation experiments by integrating MRAB and VPAB into the baseline model. As shown in Table 2, the baseline alone reveals a significant Head–Tail performance gap (B-M: 0.266 vs. 0.202), confirming the severity of long-tail bias. Adding MRAB alone improves overall B-M from 0.259 to 0.271, with the Tail group gaining +4.9% in B-M and +4.2% in R-L, validating that organ-faithful textual priors effectively alleviate decoding bias. Incorporating VPAB alone yields an overall B-M of 0.268, with Tail MTR improving by +9.1%, demonstrating that prototype-guided visual compression suppresses head-biased redundancy. When both modules are combined, the full model achieves the best results across all groups. The Tail group benefits most significantly (B-M: +8.9%, MTR: +16.5%, R-L: +10.0% over the baseline), while Head and Medium groups also exhibit consistent gains.

Fig. 3 illustrates the sensitivity of two key hyperparameters: the selecting ratio ρ in VPAB and the number of retrieved meta-reports v in MRAB. Both exhibit a rise-then-fall pattern, peaking at $\rho = 0.4$ and $v = 2$, respectively. A small ρ discards useful visual information while a large value retains redundancy. Similarly, excessive retrieved meta-reports introduce noise and conflicting textual priors. Additional sensitivity analysis on the number of VPAB prototypes further shows that $K = 4$ yields the best performance. We thus adopt $\rho = 0.4$, $v = 2$, and $K = 4$ as the default settings throughout all experiments.

4 Conclusion

In this paper, we present PriOrGen, a dual prior-anchored framework that addresses the long-tail distribution challenge in multi-organ pathology report generation. The proposed VPAB and MRAB modules jointly mitigate visual representation bias and textual decoding bias from complementary perspectives. Experiments on 11 organ types demonstrate consistent improvements, particularly on tail-class organs. This work establishes the first systematic study of long-tail bias in multi-organ pathology report generation and offers a general dual-prior paradigm that jointly mitigates distributional imbalance from both visual and textual perspectives, with potential extensions to other long-tailed medical report generation scenarios.

Acknowledgements. This work was supported by the Institute of Digital Medicine, City University of Hong Kong, and the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CityU11208324].

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

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016)
2. Cai, Z., Wei, T., Lin, L., Chen, H., Tang, X.: Bpaco: Balanced parametric contrastive learning for long-tailed medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 383–393. Springer (2024)
3. 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)
4. 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)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
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. El Nahhas, O.S., van Treeck, M., Wölflein, G., Unger, M., Ligero, M., Lenz, T., Wagner, S.J., Hewitt, K.J., Khader, F., Foersch, S., et al.: From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nature protocols **20**(1), 293–316 (2025)

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. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
11. Ju, L., Yan, S., Zhou, Y., Nan, Y., Xing, X., Duan, P., Ge, Z.: Monica: Benchmarking on long-tailed medical image classification. arXiv preprint arXiv:2410.02010 (2024)
12. Kanakarajan, K.R., Kundumani, B., Abraham, A., Sankarasubbu, M.: Biosimcse: Biomedical sentence embeddings using contrastive learning. In: Proceedings of the 13th international workshop on health text mining and information analysis (LOUHI). pp. 81–86 (2022)
13. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)
14. Li, M., Cheung, Y.m., Lu, Y., Hu, Z., Lan, W., Huang, H.: Adjusting logit in gaussian form for long-tailed visual recognition. IEEE Transactions on Artificial Intelligence **5**(10), 5026–5039 (2024)
15. 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)
16. 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)
17. Pan, L., Zhang, Y., Yang, Q., Li, T., Chen, Z.: Combat long-tails in medical classification with relation-aware consistency and virtual features compensation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–23. Springer (2023)
18. Pan, L., Zhang, Y., Yang, Q., Li, T., Chen, Z.: Long-tailed medical diagnosis with relation-aware representation learning and iterative classifier calibration. Computers in Biology and Medicine **188**, 109772 (2025)
19. Sengupta, S., Brown, D.E.: Automatic report generation for histopathology images using pre-trained vision transformers and bert. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
20. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
21. Tran, M., Schmidle, P., Wagner, S.J., Koch, V., Lupperger, V., Feuchtinger, A., Böhner, A., Kaczmarczyk, R., Biedermann, T., Eyerich, K., et al.: Generating highly accurate pathology reports from gigapixel whole slide images with histogpt. Medrxiv pp. 2024–03 (2024)
22. 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)
23. 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)

24. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.Y.: C-pack: Packed resources for general chinese embeddings. In: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval. pp. 641–649 (2024)
25. 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)
26. 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)
27. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. arXiv preprint arXiv:2401.01646 (2024)