HistGen: Histopathology Report Generation via Local-Global Feature Encoding and Cross-modal Context Interaction

Zhengrui Guo[1], Jiabo Ma[1], Yingxue Xu[1], Yihui Wang[1], Liansheng Wang[3], and Hao Chen[1,2(✉)]

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
jhc@cse.ust.hk
[2]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology
[3]School of Information Science and Engineering, Xiamen University

Abstract. Histopathology serves as the gold standard in cancer diagnosis, with clinical reports being vital in interpreting and understanding this process, guiding cancer treatment and patient care. The automation of histopathology report generation with deep learning stands to significantly enhance clinical efficiency and lessen the labor-intensive, time-consuming burden on pathologists in report writing. In pursuit of this advancement, we introduce **HistGen**, a multiple instance learning-empowered framework for histopathology report generation together with the first benchmark dataset for evaluation. Inspired by diagnostic and report-writing workflows, HistGen features two delicately designed modules, aiming to boost report generation by aligning whole slide images (WSIs) and diagnostic reports at both local and global granularities. To achieve this, a local-global hierarchical encoder is developed for efficient visual feature aggregation from a region-to-slide perspective. Meanwhile, a cross-modal context module is proposed to explicitly facilitate alignment and interaction between distinct modalities, effectively bridging the gap between the extensive visual sequences of WSIs and corresponding highly summarized reports. Experimental results on WSI report generation show the proposed model outperforms state-of-the-art (SOTA) models by a large margin. Moreover, the results of fine-tuning our model on cancer subtyping and survival analysis tasks further demonstrate superior performance compared to SOTA methods, showcasing strong transfer learning capability. Dataset and code are available here.

Keywords: Histopathology Report Generation · Multiple Instance Learning · Cross-Modal Alignment

1 Introduction

Histopathology tissue analysis reveals critical details about tumor characteristics and is considered the gold standard in cancer diagnosis and prognosis [18]. In the

last decade, the increasing availability of digital WSIs has given rise to computational pathology (CPath), promoting diagnostics with computer-assisted tools. The recent success of deep learning has notably advanced this field, outperforming experienced pathologists on specific tasks [5,33,26]. However, pathologists are still required to compile findings and craft reports for each WSI, a task that is labor-intensive, time-consuming, and error-prone. Thus, the automated generation of pathology reports holds immense potential in facilitating the report-writing process and alleviating the heavy workloads for pathologists. Additionally, these automated reports can underscore abnormalities and offer diagnostic rationales, aiding pathologists in their analyses.

Contrasting with the field of radiology report generation [8,7,22], WSI report generation faces unique challenges. A primary obstacle is the lack of recognized benchmark datasets; existing datasets predominantly focus on patch-level images and captions [13,15,21], neglecting comprehensive, global descriptions of entire WSIs. Additionally, the gigapixel size of WSIs poses unique challenges for the visual encoding stage of report generation and hinders the direct adoption of existing report generation methods. Moreover, the contrast between enormous informative WSI patches of dense visual signals and succinctly summarized diagnostic reports of discrete textual tokens underscores the need for cross-modal interactions. Further, prevalent WSI analysis methods rely on multiple instance learning (MIL), necessitating pre-trained feature extractors, which we identify as a critical bottleneck, impeding optimal performance in WSI report generation.

Related Work:

Multiple instance learning in CPath. Given the gigapixel size of WSIs and GPU memory limitations [2], MIL framework is prevalently used for WSI analysis. This framework tiles a WSI into patches, uses a pre-trained feature extractor to convert them into feature vectors, and then aggregates these vectors into a WSI-level representation for prediction [5,12,16,26,19,33]. Several works [12,5] used max-pooling to locate the most suspicious instances for slide-level prediction. To engage information of all patches, a side network was employed in [16] to compute attention score of each patch and use it as the weight for aggregation. Li et al. [19] utilized cosine distance between instances as the weighting factor for aggregation. Shao et al. [26] further adopted transformer layers to explore the long-range instance correlations. However, recent studies in MIL mainly focus on WSI-level prediction tasks, with WSI report generation under-explored along this research direction, which is the focus of this work.

Image captioning and report generation. Image captioning [30,1,28,9] and radiology report generation [8,7] are the most relevant fields to this work. These tasks require models to generate descriptive captions or diagnostic reports for input natural or radiology images. Various studies [30,1] adopted the CNN-RNN framework for image encoding and caption decoding. Recent advent of Transformer [28] enabled capturing long-range dependencies, with further explorations proposed to bridge encoding and decoding stages [9]. This concept is similarly applied in radiology report generation [8,7]. However, the gigapixel size of WSIs requires immense computational capacity, making the direct application of these

methods to WSI data impractical. Pioneering efforts in WSI report generation have been made [25,14,6], yet they mainly involve simple combinations of existing MIL and captioning models (e.g., Chen et al. [6] combined TransMIL [26] with a transformer decoder), neglecting the connections between visual and textual modalities along with their distinct characteristics.

Contribution:

Building on the success of image captioning [30,1,9,31] and radiology report generation [8,7,22], and addressing unique challenges in WSI data, we present **HistGen**, a MIL-based histopathology report generation framework, designed to address WSI report generation from a cross-modal interaction perspective by aligning WSIs and diagnostic reports across local and global granularity. Our primary contributions are summarized as follows:

- We curate a benchmark WSI-report dataset of around 7,800 pairs by building a report collection and cleaning pipeline based on the TCGA platform [27].
- A local-global hierarchical visual encoder is proposed for effective encoding and aggregation of extensive WSI patch features in a region-to-slide manner.
- We develop a cross-modal context module to enable interactions between visual encoding and textual decoding, leveraging cross-modal information.
- Further, we pre-train a general-purpose MIL feature extractor on over 55,000 WSIs to boost MIL feature encoding.
- With extensive experimental validation, our framework outperforms SOTA methods on WSI report generation by a large margin. Moreover, experiment results on cancer subtyping and survival analysis tasks exhibit superior transfer learning capability of HistGen, with it achieving SOTA on both tasks.

2 Method

2.1 WSI-Report Dataset Curation

Differing from the well-established benchmarks in radiology report generation [17,10], there is a lack of accessible WSI-report datasets, hindering the advancement of this field. In light of this, our goal is to curate a publicly available WSI-report dataset to serve as an evaluation benchmark. Based on uncurated report data from the TCGA platform [27], we build a dataset curation pipeline detailed as follows. We begin by downloading diagnostic report PDFs from TCGA and extracting raw texts from them. Since raw texts are still noisy and cluttered with redundant information, we utilize GPT-4 for further cleaning and summarization. By matching the case IDs from the reports with corresponding WSIs, we form a WSI-report dataset with 7,753 pairs, encompassing various diseases from different primary sites (see **Supplementary Materials** for more details).

2.2 HistGen for Automated WSI Report Generation

Overview. Fig. 1 shows the overview of our HistGen framework. Fig. 1a denotes the local-global hierarchical visual encoder, which firstly extracts patch features

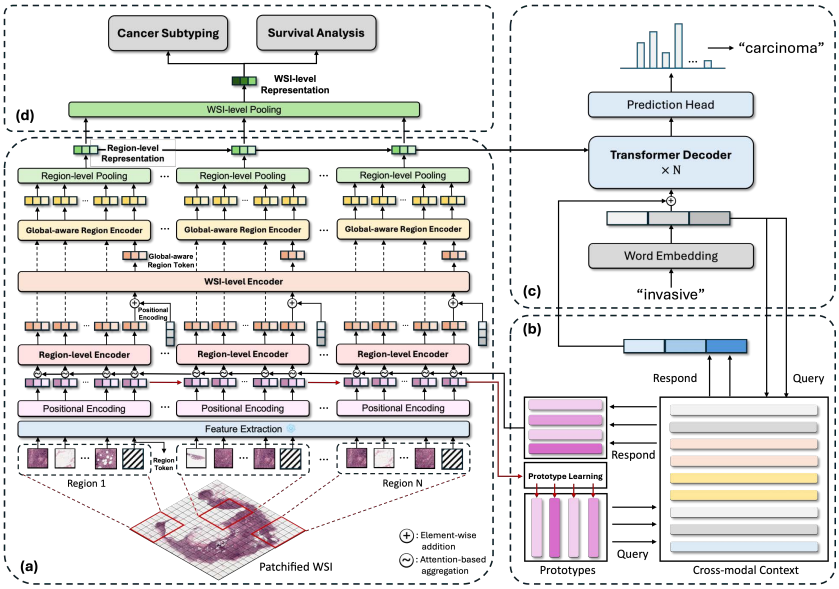

Fig. 1. Overview of the proposed HistGen framework: (a) local-global hierarchical encoder module, (b) cross-modal context module, (c) decoder module, (d) transfer learning strategy for cancer diagnosis and prognosis.

with our pre-trained backbone and then efficiently encodes them under a local-to-global manner. After that, the visual features are fed into the decoder module (shown in Fig. 1c) for report generation. Fig. 1b corresponds to the cross-modal context-aware learning module, designed to bridge different modalities and store the knowledge from previous iterations for the model to refer to. Further, Fig. 1d highlights our strategy for fine-tuning the model to WSI-level prediction tasks.

Pre-training feature extractor. To boost the MIL feature encoding process, we collect over 55,000 WSIs from public datasets (see **Supplementary Materials**) to develop a general-purpose feature extractor that offers informative and robust patch embeddings. After tissue segmentation and patch tiling, we leverage around 200 million patches to pre-train a ViT-L model $f(\cdot;\theta)$ with DINOv2 strategy [23], which is then used for feature extraction, as depicted in Fig. 1a.

Local-global hierarchical encoder (LGH). Efficiently leveraging the information and correlations among WSIs' extensive patch sequence is a vital problem. Given that pathology reports typically contain both region-level details and overall WSI diagnoses, we deduce that the visual encoder should process visual features from a local-to-global perspective to effectively capture this granularity. In this work, we propose the local-global hierarchical encoder (Fig. 1a) for region-aware representation learning. Specifically, we first segment the WSI into N regions $X_i = \{x_{i,1}, x_{i,2}, x_{i,S}\}, i = 1, \cdots, N$ with the region size S. Besides, we additionally add a region token $x_{i,S+1}$ at the end of each region sequence to learn a region-level representation. After that, we employ $f(\cdot;\theta)$ to obtain

patch embeddings and add positional encoding (PE) to each region sequence. Consequently, we design a region-level encoder $E_l(\cdot;\theta_l)$ to process X_i' to enable intra-region interactions:

$$X_i' = \{x_{i,1}', \cdots, x_{i,S+1}'\} = E_l\left(f(x_{i,1};\theta) + e_{i,1}, \cdots, f(x_{i,S+1};\theta) + e_{i,S+1};\theta_l\right).$$

Moreover, to capture inter-region information and long-range dependencies among patches from various regions, we introduce PE to each region-level representation token $x_{i,S+1}'$, followed by processing through a WSI-level encoder $E_g(\cdot;\theta_g)$ for global interactions. Subsequently, these global-aware region tokens are reorganized to infuse global information into original regions with $E_l(\cdot;\theta)$. Finally, we obtain region-level representations X_i'' by attentive pooling on each region:

$$\begin{aligned} X_i'' &= \text{Pooling}\left(\{x_{i,1}'', \cdots, x_{i,S+1}''\}\right) = \text{Pooling}\left(E_l(x_{i,1}', \cdots, \tilde{x}_{i,S+1}';\theta_l)\right) \\ &= \text{Pooling}\left(E_l(x_{i,1}', \cdots, x_{i,S}', E_g(x_{1,S+1}', x_{2,S+1}', \cdots, x_{N,S+1}';\theta_g)_i;\theta_l)\right), \end{aligned}$$

and then input region representations $\{X_1'', X_2'', \cdots, X_N''\}$ into the decoder (see Fig. 1c) for report generation. Moreover, to prove that our model learns diagnosis-related information during report generation, we further design a transfer learning mechanism (see Fig. 1d) by pooling the region-level representations X_i'' into a WSI-level representation X'' for downstream diagnosis and prognosis tasks.

Cross-modal context module (CMC). Given that report generation is an image-to-text task involving the transformation between distinct modalities, bridging the gap between pixel-based visual inputs and token-based textual outputs is crucial. Meanwhile, leveraging the model's accumulated knowledge and memory from previous iterations is expected to significantly enhance this task. The CMC module (Fig. 1b) is proposed in light of the above insights, denoted as $C = \{C_1, C_2, \cdots, C_m\}$, where C_i is the context vector stored in this module. For visual inputs $\{X_1', \cdots, X_N'\}$, we select key patches $\{p_1, \cdots, p_l\}$ via uniform sampling-based prototype learning, which aims to use representative patches as queries for the CMC module, addressing redundancy and computational complexity of the patch sequence. After querying, the responses generated by the CMC module are aggregated back into the original visual features, infusing them with cross-modal information. This step aims to help the model leverage learned knowledge from the training data that might not be captured by the input embeddings alone. Similar interactions are applied to textual embeddings $\{y_1, y_2, \cdots, y_t\}$ from the decoder, except for prototype learning. Additionally, the CMC module serves as an external memory, enabling the model to access and update iteratively, thereby progressively enhancing generation performance.

Loss function. Let $I = \{x_1, x_2, \cdots, x_n\}$ be a WSI with n denoting the patch number and $T = \{y_1, y_2, \cdots, y_t\}$ be the corresponding reports with t representing the number of tokens. The objective function is to maximize the conditional probability of the T given I: $\theta^* = \arg\max_\theta \sum_{i=1}^{t} \log P(y_i|y_1, y_2, \cdots, y_{i-1}, I;\theta)$, where $P(y_i|y_1, y_2, \cdots, y_{i-1}, I)$ is the probability of generating the next word y_i given the WSI I and the preceding words $y_1, y_2, \cdots, y_{i-1}$ in the report.

3 Experiments

3.1 Implementation Details

Datasets. In this study, we focus on WSI report generation as well as cancer subtyping and survival analysis (split set as 80%/10%/10% for train/val/test). For report generation, our model is evaluated on the WSI-report dataset (specified in Sec. 2.1), using NLG metrics including BLEU [24], METEOR [11], and ROUGE-L [20]. For cancer subtyping, since the TCGA reports may include phenotype-related information, we evaluate on three external classification datasets, including UBC-OCEAN[1], Camelyon [3,4], and TUPAC16 [29], using Monte Carlo cross-validated Accuracy and AUC. For survival analysis, we evaluate on six datasets including BRCA, STAD, KIRC, KIRP, LUAD, and COADREAD from TCGA platform, with Monte Carlo cross-validated concordance index (c-Index).
Model settings. For pre-trained DINOv2 ViT-L, the hidden dimension is 1,024 for feature extraction. For report generation model, we employ the LGH module with region size 96, and a 3-layer Transformer decoder with 8 attention heads, and 512 dimensions for hidden states. For the cross-modal context module, we set it with dimension 512×2048. Our model is trained under cross entropy loss with Adam optimizer, using learning rate 1×10^{-4} and weight decay by 0.8 per epoch. We adopt beam search with a beam size of 3 for inference.

3.2 WSI Report Generation Results

Pre-trained feature extractor. We compare our pre-trained DINOv2 ViT-L with ImageNet-pretrained ResNet50 and WSI-pretrained CTransPath [32]. Most MIL methods utilize ImageNet-pretrained ResNet50, and results in Tab. 1 show that this leads to relatively low generation performance. We then turn to CTransPath [32], which is pre-trained on 30,000 WSIs using contrastive learning strategy. However, Tab. 1 demonstrates only minor increases in the generation performance of baseline models. In contrast, as illustrated in Tab. 1, the baseline models show significant improvement by adopting our DINOv2 ViT-L, and several models such as R2GenCMN could already achieve plausible generation performance. This indicates that our DINOv2 ViT-L significantly outperforms both out-of-domain pretrained and domain-aligned pretrained methods, boosting subsequent tasks such as report generation.
WSI report generation results analysis. For SOTA report generation methods in Tab. 1, Show&Tell [30] and UpDownAttn [1] utilize CNN for image encoding and RNN for language decoding, manifesting lower performance compared to Transformer-based methods such as R2Gen [8]. M2Transformer [9] and R2GenCMN [7] further improve the performance by aligning encoding and decoding stages. However, visual token selection methods are needed to apply methods to WSI report generation due to the large sequence length, degrading report generation performance. Our proposed method addresses this problem and outperforms these methods by a large margin via efficiently exploring the

[1] `https://www.kaggle.com/competitions/UBC-OCEAN`

Table 1. Comparisons of the proposed **HistGen** with other SOTA models on WSI report generation w.r.t NLG metrics.

Feature Extractor	Methods	Metric					
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
ResNet50	**Show&Tell [30]**	0.249	0.099	0.047	0.025	0.086	0.165
	UpDownAttn [1]	0.250	0.115	0.065	0.043	0.096	0.180
	Transformer [28]	0.249	0.114	0.065	0.042	0.095	0.176
	M2Transformer [9]	0.250	0.115	0.065	0.042	0.095	0.180
	R2Gen [8]	0.240	0.105	0.058	0.036	0.089	0.177
	R2GenCMN [7]	0.225	0.095	0.047	0.022	0.094	0.151
CTransPath [32]	**Show&Tell [30]**	0.262	0.126	0.071	0.043	0.094	0.184
	UpDownAttn [1]	0.240	0.139	0.090	0.063	0.100	0.201
	Transformer [28]	0.271	0.165	0.112	0.082	0.113	0.227
	M2Transformer [9]	0.259	0.160	0.108	0.076	0.103	0.218
	R2Gen [8]	0.237	0.135	0.085	0.054	0.086	0.205
	R2GenCMN [7]	0.211	0.098	0.054	0.033	0.079	0.158
DINOv2 ViT-L (Ours)	**Show&Tell [30]**	0.189	0.094	0.056	0.039	0.070	0.165
	UpDownAttn [1]	0.320	0.206	0.147	0.112	0.131	0.271
	Transformer [28]	0.382	0.266	0.200	0.157	0.162	0.316
	M2Transformer [9]	0.321	0.213	0.152	0.112	0.131	0.266
	R2Gen [8]	0.274	0.166	0.107	0.071	0.102	0.234
	R2GenCMN [7]	0.381	0.264	0.199	0.157	0.164	0.313
	Ours	**0.413**	**0.297**	**0.229**	**0.184**	**0.182**	**0.344**

Table 2. Ablation study on the proposed LGH and CMC modules

Methods	Metric						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	AVG. Δ
Base	0.384	0.267	0.203	0.163	0.166	0.316	-
+ **CMC**	0.398	0.282	0.216	0.173	0.173	0.332	**5.18%**
+ **CMC** + **LGH**	**0.413**	**0.297**	**0.229**	**0.184**	**0.182**	**0.344**	**10.50%**

local and global information inside the patch sequence, interactively aligning the visual and textual context, and effectively utilizing the model's knowledge and memory from previous iterations. Tab. 1 confirms the superiority of our report generation model over SOTA models, with an average of around 3% improvement observed on all NLG metrics (**Qualitative Analysis** is shown in **Supplementary Materials**).

Ablation Studies. We conduct additional experiments to determine the effects of our proposed modules. Tab. 2 illustrates the performance of baseline models, in which *Base* represents a vanilla transformer with an additional pooling layer to accommodate the long sequence of WSI patches. Our findings reveal that progressive stacking of the proposed modules leads to incremental improvements in NLG metrics, indicating that both the LGH and CMC modules play essential roles in boosting WSI report generation.

Table 3. Transfer Learning Results on **Cancer Subtyping** (measured in Acc, AUC). $\mathbf{Mean_{Std}}$ and $\mathbf{Mean_{Std}}$ denote the best and second best performance, respectively.

Methods	UBC-OCEAN		CAMELYON		TUPAC16		Average	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
MaxMIL	$0.767_{\pm 0.005}$	$0.940_{\pm 0.001}$	$0.946_{\pm 0.001}$	$0.979_{\pm 0.000}$	$0.514_{\pm 0.002}$	$0.693_{\pm 0.001}$	$0.785_{\pm 0.002}$	$0.894_{\pm 0.001}$
MeanMIL	$0.771_{\pm 0.004}$	$0.947_{\pm 0.001}$	$0.811_{\pm 0.002}$	$0.898_{\pm 0.001}$	$0.571_{\pm 0.002}$	$0.724_{\pm 0.005}$	$0.738_{\pm 0.003}$	$0.866_{\pm 0.002}$
ABMIL [16]	$0.792_{\pm 0.004}$	$0.954_{\pm 0.000}$	$0.951_{\pm 0.000}$	$0.988_{\pm 0.000}$	$0.573_{\pm 0.002}$	$0.736_{\pm 0.002}$	$0.809_{\pm 0.002}$	$0.913_{\pm 0.001}$
TransMIL [26]	$0.782_{\pm 0.002}$	$0.929_{\pm 0.001}$	$0.944_{\pm 0.001}$	$0.976_{\pm 0.000}$	$0.541_{\pm 0.004}$	$0.689_{\pm 0.003}$	$0.795_{\pm 0.002}$	$0.889_{\pm 0.001}$
DS-MIL [19]	$0.785_{\pm 0.003}$	$0.944_{\pm 0.000}$	$0.937_{\pm 0.002}$	$0.979_{\pm 0.001}$	$0.559_{\pm 0.006}$	$0.719_{\pm 0.002}$	$0.797_{\pm 0.003}$	$0.902_{\pm 0.001}$
DTFD-MIL [33]	$0.785_{\pm 0.004}$	$0.955_{\pm 0.000}$	$0.943_{\pm 0.014}$	$0.978_{\pm 0.015}$	$0.555_{\pm 0.003}$	$0.739_{\pm 0.004}$	$0.799_{\pm 0.008}$	$0.910_{\pm 0.008}$
Ours	$0.808_{\pm 0.003}$	$0.946_{\pm 0.001}$	$0.961_{\pm 0.000}$	$0.996_{\pm 0.000}$	$0.604_{\pm 0.002}$	$0.738_{\pm 0.004}$	$0.827_{\pm 0.001}$	$0.915_{\pm 0.001}$

Table 4. Transfer Learning Results on **Survival Analysis** (measured with c-Index). $\mathbf{Mean_{Std}}$ and $\mathbf{Mean_{Std}}$ denote the best and second best performance, respectively.

Methods	BRCA	STAD	KIRC	KIRP	LUAD	COADREAD	Average
MaxMIL	$0.586_{\pm 0.010}$	$0.553_{\pm 0.006}$	$0.732_{\pm 0.002}$	$0.562_{\pm 0.022}$	$0.563_{\pm 0.004}$	$0.560_{\pm 0.012}$	$0.595_{\pm 0.009}$
MeanMIL	$0.608_{\pm 0.006}$	$0.569_{\pm 0.005}$	$0.710_{\pm 0.007}$	$0.609_{\pm 0.011}$	$0.584_{\pm 0.005}$	$0.589_{\pm 0.011}$	$0.612_{\pm 0.009}$
ABMIL [16]	$0.621_{\pm 0.005}$	$0.574_{\pm 0.008}$	$0.700_{\pm 0.009}$	$0.636_{\pm 0.017}$	$0.603_{\pm 0.005}$	$0.572_{\pm 0.006}$	$0.618_{\pm 0.008}$
TransMIL [26]	$0.615_{\pm 0.010}$	$0.582_{\pm 0.003}$	$0.672_{\pm 0.010}$	$0.532_{\pm 0.017}$	$0.584_{\pm 0.006}$	$0.525_{\pm 0.008}$	$0.593_{\pm 0.009}$
DS-MIL [19]	$0.624_{\pm 0.011}$	$0.600_{\pm 0.006}$	$0.672_{\pm 0.003}$	$0.641_{\pm 0.013}$	$0.591_{\pm 0.004}$	$0.561_{\pm 0.006}$	$0.593_{\pm 0.008}$
DTFD-MIL [33]	$0.592_{\pm 0.006}$	$0.601_{\pm 0.005}$	$0.702_{\pm 0.008}$	$0.572_{\pm 0.001}$	$0.575_{\pm 0.005}$	$0.561_{\pm 0.011}$	$0.601_{\pm 0.006}$
Ours	$0.630_{\pm 0.008}$	$0.607_{\pm 0.003}$	$0.698_{\pm 0.006}$	$0.648_{\pm 0.015}$	$0.623_{\pm 0.004}$	$0.568_{\pm 0.011}$	$0.628_{\pm 0.008}$

3.3 Transfer Learning for Cancer Diagnosis and Prognosis

WSI report generation could be considered as vision-language pre-training. To prove that our model learns diagnosis-related information during this task, we further fine-tune our model and evaluate on cancer subtyping and survival analysis with the strategy illustrated in Fig. 1d. Max-MIL, Mean-MIL, AB-MIL [16], TransMIL [26], DS-MIL [19], and DTFD-MIL [33] are included for comparison.

Cancer Subtyping. Tab. 3 summarizes the comparison of our model to SOTA MIL methods in cancer subtyping datasets, showing that our method outperforms all MIL approaches on all three classification tasks by a significant margin.
Survival Analysis. Results of six survival analysis tasks, shown in Tab. 4, reveal that our model outperforms others in most tasks and maintains the highest average score across all six datasets, manifesting superior transfer capabilities.

Tab. 3 and 4 highlight the effectiveness of our proposed model on WSI-level prediction tasks as well as the feasibility of transferring our model to these downstream tasks, confirming its ability to capture diagnostic information during WSI report generation.

4 Conclusion

In this work, we introduce **HistGen**, a MIL-empowered framework designed to enhance automated histopathology report generation, covering from model design to benchmark evaluation. To this end, we develop an advanced WSI report generation model by efficiently leveraging local-region and global-WSI information together with explicitly aligning the cross-modal encoding and decoding stages. To conduct experimental evaluation, we curate a benchmark WSI-report dataset, containing around 7,800 WSI-report pairs with high text quality. Extensive experiments on report generation, cancer subtyping, and survival analysis demonstrate the superiority of the proposed model as well as its strong capabilities of transfer learning. While this work lays the groundwork for WSI report generation, its scope is currently limited to histopathology. Future efforts will expand to include other fields such as radiology and ophthalmology, treating report generation across different fields as a unified problem within a single framework.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62202403), Hong Kong Innovation and Technology Fund (No. PRP/034/22FX), and Research Grants Council of the Hong Kong Special Administrative Region, China (No. R6003-22).

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: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 6077–6086 (2018)
2. Araujo, A., Norris, W., Sim, J.: Computing receptive fields of convolutional neural networks. Distill **4**(11), e21 (2019)
3. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Trans. Med. Imaging **38**(2), 550–560 (2018)
4. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
5. 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. Nat. Med. **25**(8), 1301–1309 (2019)
6. Chen, P., Li, H., Zhu, C., Zheng, S., Yang, L.: Mi-gen: Multiple instance generation of pathology reports for gigapixel whole-slide images. arXiv preprint arXiv:2311.16480 (2023)

7. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
8. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
9. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 10578–10587 (2020)
10. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
11. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation. pp. 85–91 (2011)
12. Feng, J., Zhou, Z.H.: Deep miml network. In: AAAI Conf. Artif. Intell. vol. 31 (2017)
13. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 16549–16559 (2021)
14. Guevara, B.C., Marini, N., Marchesin, S., Aswolinskiy, W., Schlimbach, R.J., Podareanu, D., Ciompi, F.: Caption generation from histopathology whole-slide images using pre-trained transformers. In: Medical Imaging with Deep Learning, short paper track (2023)
15. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nat. Med. **29**(9), 2307–2316 (2023)
16. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proc. Int. Conf. Mach. Learn. pp. 2127–2136. PMLR (2018)
17. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 1–8 (2019)
18. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., Srinivasan, B.: A generalized deep learning framework for whole-slide image segmentation and analysis. Scientific reports **11**(1), 11579 (2021)
19. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 14318–14328 (2021)
20. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
21. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 19764–19775 (2023)
22. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. 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)

25. Sengupta, S., Brown, D.E.: Automatic report generation for histopathology images using pre-trained vision transformers. arXiv preprint arXiv:2311.06176 (2023)
26. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Proc. Adv. Neural Inf. Process. Syst. **34**, 2136–2147 (2021)
27. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv Neural Inf Process Syst **30** (2017)
29. Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., et al.: Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. Med. Image Anal. **54**, 111–121 (2019)
30. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 3156–3164 (2015)
31. Wang, N., Xie, J., Wu, J., Jia, M., Li, L.: Controllable image captioning via prompting. In: AAAI Conf. Artif. Intell. vol. 37, pp. 2617–2625 (2023)
32. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. **81**, 102559 (2022)
33. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn. pp. 18802–18812 (2022)