HERGen: Elevating Radiology Report Generation with Longitudinal Data (Supplementary Material)

Fuying Wang[1], Shenghui Du[1], and Lequan Yu[1]

The University of Hong Kong, Hong Kong
{fuyingw@connect.,shenghui@connect.,lqyu@}hku.hk

The appendix is structured as follows. Section 1 offers implementation details of the HERGen architecture, along with comprehensive details on the implementation of our experiments. Additional experimental results are showcased in Section 2. Finally, Section 3 provides a qualitative analysis of our model, providing deeper insights into its performance.

1 More Technical Details

1.1 Modules of Our Framework

Image Encoder. Following the insights from CvT-212DistilGPT2 [8], we utilize the CvT architecture [15], pretrained on ImageNet-21K, as the foundation for our image encoder. This choice is supported by empirical evidence demonstrating its effectiveness in interpreting radiological images. Subsequently, an encoder projection layer E_{proj} is integrated to align the number and dimension of visual tokens with the text decoder's needs.

For each patient, indexed by i, our image encoder E processes its series of chest X-rays $\mathcal{I}_i = \{\mathbf{I}_1^{(i)}, \mathbf{I}_2^{(i)}, ..., \mathbf{I}_{N_i}^{(i)}\}$ into visual features $\mathcal{P}_i = \{\mathbf{P}_1^{(i)}, \mathbf{P}_2^{(i)}, ..., \mathbf{P}_{N_i}^{(i)}\}$. Each X-ray image $\mathbf{I}_j^{(i)} \in \mathbb{R}^{C\times W\times H}$ $(0 \leq j \leq N_i - 1)$, with C, H, and W representing the number of channels, height, and width respectively, is encoded into a feature representation $\mathbf{P}_j^{(i)}$ in $\mathbb{R}^{S\times F}$. Here, S and F denote the number of visual tokens and the feature dimension per token, respectively. To adjust the dimension of visual features and the number of visual tokens, we introduce an encoder projection layer E_{proj}, consisting of a 1×1 convolution layer followed by a linear projection layer. Specifically, E_{proj} transforms each $\mathbf{P}_j$ $(0 \leq j \leq N_i - 1)$ into a more compact visual representation $\mathbf{V}_j$, resulting in $\mathcal{V}_i = \{\mathbf{V}_1^{(i)}, \mathbf{V}_2^{(i)}, ..., \mathbf{V}_{N_i}^{(i)}\}$ in $\mathbb{R}^{S'\times F'}$, where S' and F' represent the adjusted number of visual tokens and their new dimensionality. In our experiments, we set $S' = 50$ and $F' = 768$.

Text Encoder. To accurately interpret the specialized language of radiology reports, we incorporates CXR-BERT [2], a domain-specific BERT-based model pretrained for biomedical literature and clinical domains, as the text encoder of our framework. This choice enables our model to produce detailed text embeddings that capture the clinical nuances and specific terminologies of radiology, laying a strong foundation for the subsequent stages of report generation.

Table 1: Detailed preprocessing steps of MIMIC-CXR in our study. The number of patients, images and reports after each step are shown in this table.

	Train			Val			Test		
	#. Patient	#. Image	#. Report	#. Patient	#. Image	#. Report	#. Patient	#. Image	#. Report
Original (same as [4,8])	59,799	270,790	152,173	459	2,130	1,196	289	3,858	2,347
+Remove lateral images	58,595	162,169	145,471	448	1,286	1,151	285	2,461	2,210
+Remove duplicated images	58,595	145,471	145,471	448	1,151	1,151	285	2,210	2,210

Formally, for each patient indexed by i, the text encoder transforms their series of radiology reports $\mathcal{R}_i = \{\mathbf{R}_1^{(i)}, \mathbf{R}_2^{(i)}, ..., \mathbf{R}_{N_i}^{(i)}\}$ into a corresponding series of global text embeddings $\{\mathbf{E}_1^{(i)}, \mathbf{E}_2^{(i)}, ..., \mathbf{E}_{N_i}^{(i)}\}$, where $\mathbf{E}_j^{(i)} \in \mathbb{R}^{F'}, 0 \leq j \leq N_i - 1$ with L being the number of text tokens in each embedding. Following [4], L is set to 128.

Text Decoder. Our model utilizes DistilGPT2 [11], a streamlined version of the GPT-2 architecture [10], pretrained on the diverse WebText corpus. This distilled version maintains the essential auto-regressive and self-attention features of GPT-2, crucial for nuanced language modeling and text generation. We have integrated the DistilGPT2 decoder with a multi-head cross-attention module, as [13], enabling it to effectively integrate visual context from projected image features with textual input, a key aspect in generating contextually rich radiology reports. Specifically, the implementation of this encoder-decoder framework leverages the HuggingFace Transformer library[1], ensuring robust and advanced architecture for our text generation tasks.

In our model, the text decoder generates a series of radiology reports $\mathcal{R}_i = \{\hat{\mathbf{R}}_1^{(i)}, \hat{\mathbf{R}}_2^{(i)}, ..., \hat{\mathbf{R}}_{N_i}^{(i)}\}$ from the visual embeddings $\mathcal{D}_i = \{\hat{\mathbf{D}}_1^{(i)}, \hat{\mathbf{D}}_2^{(i)}, ..., \hat{\mathbf{D}}_{N_i}^{(i)}\}$ for each patient i. This process is formally expressed as $\hat{\mathbf{R}}_j^{(i)} = \text{Decoder}(\hat{\mathbf{D}}_j^{(i)})$ for each image j $(0 \leq j \leq N_i - 1)$. Then, a cross entropy loss is optimized to classify each text token into ground truth token.

1.2 Implementation Details of Experiments

Dataset Preprocessing. We describe our preprocessing of the MIMIC-CXR dataset in Table 1. Starting from the same dataset in [4,8], we further process the dataset by removing lateral images and duplicates within the same study, while maintaining the original data split from prior research. The number of patients, images and reports after each step are shown in this table.

Implementation Details of Report Generation. Our model's optimization employs a three-stage learning rate strategy. In the first stage, following [8], we set the learning rate for the image encoder at $5e-4$ and for the encoder projection layer and text decoder at $5e-5$. In the second stage, we adopt the Cosine Annealing with Warmup scheduler for joint parameter optimization. Here, the learning rate oscillates between 0 and $5e-5$ over a cycle of 50 epochs. In the

[1] https://huggingface.co/blog/encoder-decoder

Table 2: Ablation results of hyperparameter λ on the Longitudinal MIMIC-CXR dataset. The best results are shown in **bold**.

λ	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
0.5	0.387	0.241	0.162	0.116	0.153	0.281
2	0.386	0.240	0.163	0.117	0.153	**0.283**
1	**0.389**	**0.242**	**0.163**	**0.117**	**0.155**	0.282

Table 3: Ablation results for Temporal Embedding (TE) on the Longitudinal MIMIC-CXR dataset. “Learnable" refers to the temporal embeddings that are learnable but do not consider absolute study dates. In contrast, “Ours" denotes the study time-aware temporal embeddings, as detailed in the main paper. The best-performing results are highlighted in **bold**.

TE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Learnable	0.389	0.240	0.161	0.116	0.153	0.280
Ours	**0.389**	**0.242**	**0.163**	**0.117**	**0.155**	**0.282**

third stage, we the same scheduler settings as the second stage. Interestingly, we found that only optimizing the temporal aggregation module achieves almost very close performance as optimizing all parameters. To improve efficiency, we only optimize the temporal aggregation module in this stage. Our model's architecture includes 2 group causal transformer blocks, each with 8 attention heads, and an embedding dimension of 768. For report generation, we utilize beam search with a beam size of 3.

Implementation Details of Temporal Image Classification. To assess our model's proficiency in discerning temporal progressions in longitudinal radiographs, we introduce a supervised temporal classification task. This task focuses on identifying disease progression using two consecutive radiographic scans from each patient. We utilize the MS-CXR-T dataset, comprising 1326 labeled observations for 5 common diseases, categorized into three classes: “worsening", “stable", “improving".

For this task, we employ HERGen, initially pretrained on the MIMIC-CXR dataset, freeze its visual encoder and temporal aggregation module, and finetune an additional linear layer to classify each disease's progression. To benchmark against other models, we perform similar finetuning on a randomly initialized ResNet, an ImageNet-pretrained ResNet, BioVil [2], and BioVil-T [1]. All models underwent the same experimental settings, with the dataset divided into training, validation, and testing sets at 70%, 10%, and 20%, respectively. We set the learning rates at 0.005 for ResNets and 0.00001 for BioVil models. The results, reported as averages over 5 different random seeds, are detailed in the main paper.

Table 4: Ablation results of joint training and curriculum learning strategies on the MIMIC-CXR dataset. Here, "CL" refers to the auxiliary contrastive learning objective, and "temporal" indicates the temporal aggregation module. The **best** results are highlighted in **bold**, while the second-best results are denoted with underline.

Strategy	CL	Temporal	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	AVG.Δ
Baseline			0.372	0.231	0.155	0.111	0.149	0.280	–
Jointly	✓		0.372	0.230	0.155	0.111	0.147	**0.295**	+0.92%
		✓	0.322	0.203	0.138	0.099	0.133	0.279	−9.55%
	✓	✓	0.363	0.223	0.149	0.106	0.144	0.277	−2.77%
Curriculum	✓		0.390	0.239	0.161	0.117	0.153	0.280	+3.23%
		✓	0.388	0.240	0.162	0.116	0.153	0.283	+3.39%
	✓	✓	**0.395**	**0.248**	**0.169**	**0.122**	**0.156**	0.285	+5.93%

2 More Experimental Results

Effect of Hyperparameters. In aligning with our discussion in the main paper, we employ a hyperparameter λ to balance between cross-entropy loss and contrastive loss. The influence of λ on report generation performance using the Longitudinal MIMIC-CXR dataset is detailed in Table 2. We examine our model's performance with varying λ values: $\{0.5, 1, 2\}$. The results indicate that a setting of $\lambda = 1$ yields the best outcomes across 5 of the 6 metrics evaluated. Notably, the performance variations across different λ settings are minimal, suggesting our model's robustness to hyperparameter λ.

Ablation on Temporal Embedding. Regarding temporal embeddings, Table 3 compares the efficacy of our proposed study time-aware embeddings against 'Learnable' embeddings, which are temporal but do not account for study dates. Our model demonstrates superior performance with our temporal embeddings, leading in 6 out of 6 metrics. Despite the close performance of both types of embeddings, our approach, which considers absolute study time, offers a more reasonable basis for radiology report generation.

More Analysis on Curriculum Learning Strategy. This section delves deeper into the curriculum learning strategy implemented for generating radiology reports from longitudinal data. Table 4 presents the results of our ablation study, which contrasts the baseline strategy (CvT-212DistilGPT2 model), our step-by-step curriculum strategy, and a jointly training strategy. The baseline strategy refers to the trained model after our stage 1. In contrast, our curriculum strategy, detailed in the main paper, involves a progressive training approach. The jointly training strategy, on the other hand, aims at simultaneous optimization from scratch with given objectives. Our findings demonstrate the superior performance of the curriculum strategy which integrates contrastive loss and a temporal aggregation module, excelling in 5 of 6 evaluated metrics. Interestingly, the joint training strategy performs worse (or comparable) than the baseline in all three settings, likely due to challenges in optimizing a large model for multiple complex tasks simultaneously. Furthermore, as discussed in the main paper,

Table 5: Micro-averaged metrics over 5 observations (denoted by mic-5) on MIMIC-CXR. The **Best** and 2nd best results are shown in bold and underline, respectively. † indicates the results are cited from the original papers. Note that they are not strictly comparable with us. For the baselines without †, their results are obtained by re-running the publicly released codebase on the same preprocessed dataset as we used.

Method	Year	$P_{\text{mic}-5}$	$R_{\text{mic}-5}$	$F_{\text{mic}-5}$
$\mathcal{M}^2$Transformer [5]	2019	0.443	0.275	0.309
R2Gen [4]	2020	0.295	<u>0.590</u>	0.393
R2GenCMN [3]	2021	<u>0.584</u>	0.408	0.443
$\mathcal{M}^2$TR.PROGRESSIVE [9]	2021	0.366	0.268	0.299
XProNet [14]	2022	**0.596**	0.353	0.444
CvT-212DistilGPT2 [8]	2022	0.525	0.567	<u>0.545</u>
DCL [6]	2023	0.541	0.412	0.468
HERGen(**Ours**)	-	0.519	**0.592**	**0.553**
Results below are not strictly comparable. For reference only.				
$\mathcal{M}^2$Trans w/ NLL† [7]	2021	0.489	0.411	0.447
$\mathcal{M}^2$Trans w/ NLL + BS + $\text{f}^{\dagger}_{C_E}$ [7]	2021	0.463	0.732	0.567
$\mathcal{M}^2$Trans w/ NLL + BS + $\text{f}^{\dagger}_{C_{\text{EN}}}$ [7]	2021	0.503	0.651	0.567
RGRG† [12]	2023	0.491	0.617	0.547

Table 6: Ablation study of different components with 95% confidence intervals.

CL	Temporal	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
		0.372 (0.367, 0.377)	0.231 (0.227, 0.235)	0.155 (0.151, 0.158)	0.111 (0.107, 0.114)	0.280 (0.276, 0.283)	0.149 (0.147, 0.151)
✓		0.388 (0.384, 0.393)	0.241 (0.237, 0.245)	0.162 (0.158, 0.166)	0.117 (0.113, 0.121)	0.281 (0.277, 0.284)	0.154 (0.152, 0.156)
	✓	0.389 (0.385, 0.394)	0.241 (0.237, 0.245)	0.162 (0.158, 0.166)	0.116 (0.112, 0.120)	0.281 (0.277, 0.284)	0.153 (0.152, 0.155)
✓	✓	**0.396** (0.392, 0.400)	**0.248** (0.244, 0.252)	**0.168** (0.164, 0.172)	**0.122** (0.118, 0.125)	**0.285** (0.281, 0.288)	**0.156** (0.154, 0.157)

the auxiliary contrastive alignment module and the temporal aggregation module are critical in enhancing the curriculum strategy, significantly outperforming the baseline. In summary, the comparative analysis of the jointly training and curriculum strategies underscores the effectiveness and necessity of our curriculum learning approach.

Additional Results of CE Metrics. We report micro-averaged precision, recall, and F1 score for 5 common observations, following the methodology in [7]. Our model demonstrates the best recall and F1 scores, along with compared precision that matches baseline models. These findings align with the comparisons detailed in the main paper, further showing the effectiveness of our HERGen in radiology report generation.

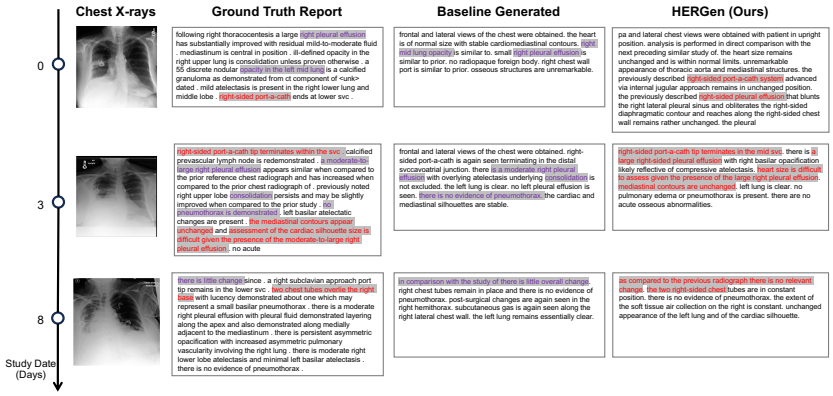

Fig. 1: This case study compares radiology report predictions for a patient by our model and CvT-212DistilGPT2. Text highlighted in gray indicates words or their synonyms found in both the predicted and ground truth reports. Purple highlights denote similar matches in the baseline-generated (CvT-212DistilGPT2) reports and ground truth, while red highlights show such matches in our model's reports and ground truth. From top to bottom, the chest X-rays are chronologically ordered.

Ablation Results with Confidence Intervals. We have computed the 95% confidence intervals using non-parametric bootstrap, resampling the test set 1000 times with replacement. Our metrics were calculated from these samples, deriving intervals from the 2.5 and 97.5 percentiles. The effectiveness of our components is further validated by our ablation study shown in Table 6.

3 More Visualization Results

Case Study. In this case study, we evaluate the performance of our proposed model, HERGen, against the CvT-212DistilGPT2 model by comparing the radiology reports they generate from longitudinal radiographs of a patient. The comparison is visually presented in Fig. 1. Our analysis reveals that HERGen outperforms CvT-212DistilGPT2 in producing radiology reports that are more consistent and precise in their clinical findings. Notably, the reports generated by HERGen align more closely with the ground truth, indicating superior prediction quality. This case study highlights the effectiveness of the contrastive alignment and temporal aggregation module incorporated in HERGen.

Visualization of Similarity of Visual Embeddings To intuitively understand the visual embedding distribution in our model, we analyzed the embedding similarity across different time steps, as depicted in Fig. 2. We utilized the MIMIC-CXR dataset, selecting 200 patients with a minimum of 5 longitudinal radiographs each. Then, we extract visual features via the image encoder and temporal aggregation module of the pretrained HERGen on MIMIC-CXR dataset. We grouped visual embeddings for each patient based on the time-step

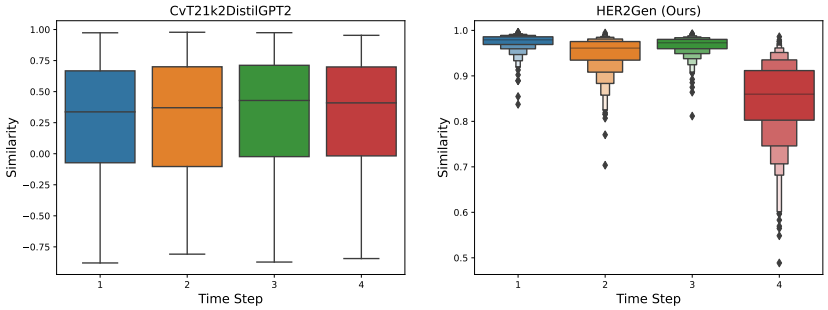

Fig. 2: Boxplot depicting the variation in embedding similarity across different time steps. The left subfigure visualizes results from the CvT-212DistilGPT2 model, while the right subfigure presents results from our model.

differences, resulting in four distinct image pair groups per patient. Subsequently, we calculated the similarity of embedding pairs within each group and visualized these findings. Our results indicate a significantly higher mean similarity in our model compared to the CvT2DistilGPT2 baseline, alongside a notably reduced variance. This suggests that our model achieves more consistent visual embeddings over various time steps compared to the baseline. The consistency underlines the effectiveness of our temporal aggregation module in integrating information from prior studies into current analyses, thereby enhancing the accuracy and consistency of generated radiology reports.

References

1. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023) 3
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022) 1, 3
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022) 5
4. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020) 2, 5
5. 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) 5
6. 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) 5

7. Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. arXiv preprint arXiv:2010.10042 (2020) 5
8. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023) 1, 2, 5
9. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777 (2021) 5
10. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 2
11. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019) 2
12. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023) 5
13. 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) 2
14. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision. pp. 563–579. Springer (2022) 5
15. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22–31 (2021) 1