UTSA-NLP at RadSum23: Multi-modal Retrieval-Based Chest X-Ray Report Summarization

Tongnian Wang*, **Xingmeng Zhao***, and **Anthony Rios**
Department of Information Systems and Cyber Security
The University of Texas at San Antonio
{Tongnian.Wang, Xingmeng.Zhao, Anthony.Rios}@utsa.edu

Abstract

Radiology report summarization aims to automatically provide concise summaries of radiology findings, reducing time and errors in manual summaries. However, current methods solely summarize the text, which overlooks critical details in the images. Unfortunately, directly using the images in a multimodal model is difficult. Multimodal models are susceptible to overfitting due to their increased capacity, and modalities tend to overfit and generalize at different rates. Thus, we propose a novel retrieval-based approach that uses image similarities to generate additional text features. We further employ few-shot with chain-of-thought and ensemble techniques to boost performance. Overall, our method achieves state-of-the-art performance in the F1RadGraph score, which measures the factual correctness of summaries. We rank second place in both MIMIC-CXR and MIMIC-III hidden tests among 11 teams, with F1RadGraph scores of 42.86 and 36.31.

1 Introduction

Radiology reports are written documents that record and interpret the results of radiological exams. A radiology report usually consists of a *findings* section, which contains multiple sentences presenting detailed observations and discoveries made by the radiologist regarding the images acquired from the examination, and an *impression* section that summarizes the findings with major observations, conclusions or diagnosis (Kahn Jr et al., 2009). However, manually summarizing radiology findings into impressions is a time-consuming and error-prone process (Xue et al., 2018; Gershanik et al., 2011). Hence, developing tools to augment (i.e., help) radiologist is important. The resulting systems have the potential to significantly improve the efficiency of clinical communication and accelerate the radiology workflow (Delbrouck et al., 2021).

Current radiology report summarization methods have generally focused on summarizing radiology reports only from the report findings. Zhang et al. (2018) developed a model for automatically summarizing radiology findings into natural language impression statements. More recently, some studies used pre-trained language models such as BART (Lewis et al., 2020) or PEGASUS (Zhang et al., 2020) for radiology report summarization (Zhu et al., 2021; Xu et al., 2021). Dai et al. (2021) effectively used a domain adaptation module, an ensemble module, and text normalization heuristics to achieve better summarization results. However, the initial radiology findings may be inaccurate and overlook certain image information. Therefore, it is important to explore the use of multimodal frameworks, i.e., methods that incorporate the image, for radiology report summarization.

Multimodal training faces significant constraints, including uncertainty regarding the necessity of visual-linguistic reasoning and comprehension in completing tasks involving images and text (Delbrouck et al., 2022b). Training multimodal models has been recognized as a challenging task because it requires the model to comprehensively understand the content of images and bridge the gap between the image and its description (Delbrouck et al., 2022b). Multimodal models are also susceptible to overfitting due to their increased capacity, and modalities tend to overfit and generalize at different rates (Wang et al., 2020).

To overcome the limitations of traditional multimodal training and radiology report summarization, we employ a multimodal encoder where the final model uses the text alone using similarity matching. Overall, this leads to a more accurate and comprehensive representation of findings than using the images directly. We use the joint information to fine-tune pre-trained T5-based models and leverage few-shot with chain-of-thought and model ensembles to further bolster our model's performance.

*These authors contributed equally to this work.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 557–566

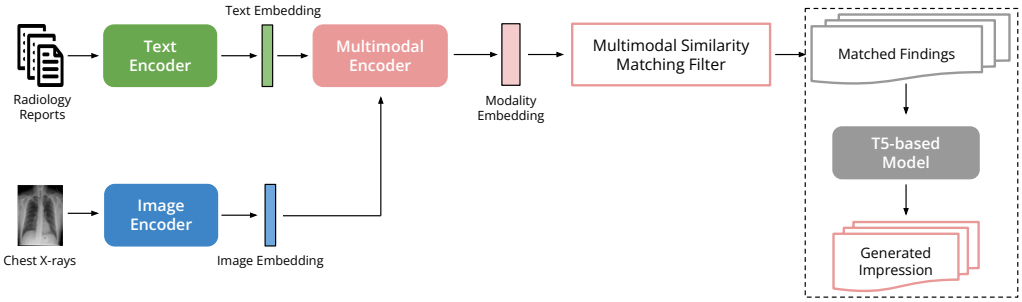

Figure 1: An overview of our proposed multimodal radiology report summarization approach.

2 Method

The core of our approach involves two stages shown in Figure 1. In the first stage, we retrieve the most similar radiology report based on chest X-ray images and findings from a medical corpus using a pre-trained multi-modal encoder. We then prepend the top k most similar report's findings and impressions to the input example's findings. In the second stage, a text-only model generates the impression based on the new inputs. This approach is inspired by previous work Endo et al. (2021); Jeong et al. (2023) that treated medical report generation as an image-text retrieval task and utilized a large retrieval corpus to provide sufficient coverage of potential diagnoses of an input chest X-ray image (Yu et al., 2022). By incorporating information from similar radiology reports, our approach aims to generate a more precise impression.

2.1 Stage 1: Multimodal Retrieval

Formally, given an input instance x_i consisting of a text input w and image m, we retrieve the most similar examples $\{x_1, \ldots, x_{\mathcal{N}(x_i)}\}$, where $\mathcal{N}(x_i)$ are the k most similar examples to x_i. Our approach uses an image-text model to obtain multimodal representations instead of two separate unimodal encoders representing images and texts for similarity search. Specifically, our multimodal image-text retrieval model consists of two encoders shown in Figure 1: unimodal image and text encoders and a multimodal encoder to obtain modality embeddings. First, the image is passed through the image encoder to generate image embeddings. We use the pre-trained Vision Transformer (ViT) model (Dosovitskiy et al., 2020) to encode the images. Since some findings contain multiple images, we average all image embeddings corresponding to the same findings. Next, we modify the pre-trained Transformer encoder-decoder T5 model (Raffel et al., 2020) to handle multimodal input. Specifically, we pass the findings as input to the T5 encoder and initialize the hidden state of the averaged embeddings produced by the image encoder. Finally, the final EOS token from the T5 encoder is used as the multimodal embedding.

Unfortunately, this model cannot be used as-is with the initial pre-trained models. Instead, we train this model where the T5 encoder outputs are passed to the T5 decoder to generate the impressions. After training the joint model, we remove the decoder and use the embeddings for Stage 2 of our pipeline.

2.2 Stage 2: Unimodal Model for TEXT

Next, using the Stage 1 model, we employ similarity matching to obtain a more accurate description of the given images and findings. This involves comparing each joint finding-image embedding in embeddings for each example in the training corpus using cosine similarity. We return the top k neighbors based on the cosine similarity. In our experiments, we set k to 1 to avoid long sequences that increases the training time. Given the most similar example $x_{\mathcal{N}(x_i)}$ to our input instance x_i, we concatenate $x_{\mathcal{N}(x_i)}$'s findings $w_{\mathcal{N}(x_i),f}$ and impression $w_{\mathcal{N}(x_i),i}$ with the x_i's findings $w_{x_i,f}$ into a single sequence defined as $[w_{\mathcal{N}(x_i),f},\ SEP,\ w_{\mathcal{N}(x_i),i},\ SEP,\ w_{x_i,f}]$. In our final ensemble, we use both Clinical-T5 (Lehman and Johnson, 2023) and Flan-T5 (Chung et al., 2022) encoder-decoder models.

2.3 Model Ensemble

We employ an ensemble approach to improve the robustness of our summarization process by combining diverse predictions from multiple models. Candidate predictions for each example (i.e., generated impressions) $p_i, \ldots p_N$ are generated using various pre-trained T5-based models and beam search sizes, resulting in diverse outputs. We com-

pute the similarity score, $sim(p_i, p_j)$, between each pair of predictions using the F1RadGraph score. These scores are then aggregated to determine the overall similarity of each prediction p_i against all others as $s_i = \frac{1}{N-1} \sum_{k \neq i} sim(p_i, p_k)$. We also calculate the similarity between each prediction p_i and its corresponding example's findings, represented as $sim(p_i, x_f)$, to ensure semantic proximity to the original findings. We calculate the final score for each prediction using a weighted sum between the two similarities defined as $f(l_i) = [\frac{1}{N-1} \sum_{k \neq i} sim(p_i, p_k)] + \lambda * sim(p_i, x_f)$, where λ is the weight of the prediction-finding similarity. We select the predicted impression with the highest overall similarity score as the final summary.

2.4 Back-translation

Back-translation is used for data augmentation on MIMIC-III. We then train two models for the model ensemble: one using the original data and the other using augmented data. To increase the size and diversity of the training data, we employ back-translation (Sennrich et al., 2016). This involves translating the English findings into Spanish and then translating the Spanish sentence back into English using a machine translation model. The resulting rephrased findings are then added to the training data for augmentation with the original impressions. We use MarianNMT (Junczys-Dowmunt et al., 2018) as our translation model.

2.5 Few-Shot with Chain-of-Thought

Large language models (LLMs) perform well in many natural language processing tasks, but their effectiveness in specialized fields like radiology is limited due to complex terminology and language. Fine-tuned models face generalization and transfer issues when applied to out-of-domain data (Wu et al., 2023). We apply few-shot learning using Chain of Thought (CoT) as a domain adaptation strategy for specialized radiology summarization tasks inspired by Wei et al. (2022). This balances using unified LLMs and fine-tuned models to generate tailored predictions for the target domain. We use Stanford Alpaca (Taori et al., 2023) with few-shot prompting and manual reasoning demonstrations, fine-tuned from Meta's LLaMA 7B model (Touvron et al., 2023). The demonstrations consist of a task instruction with a role, findings, and reasoning chain with a rationale and impression separated by a special token [SEP]. The role is a sentence that assigns a persona to a language model within the prompt, providing it with a specific context to generate text. The rationale comprises a series of intermediate reasoning steps. These manually designed demonstrations form a paradigm known as Manual-CoT (Wei et al., 2022). We provide a figure depicting this process in Figure 2 in the Appendix.

Source	Train	Validation	Test	Hidden
MIMIC-III	59320	7413	6526	6531
MIMIC-CXR	125417	991	1624	1000

Table 1: Statistics of the radiology report summarization task datasets.

3 Experiments and Results

In this section, we describe the data, evaluation metrics, and report results.

3.1 Data Description

We conduct our experiments on the 2023 BioNLP shared task (Delbrouck et al., 2023) version of the MIMIC-III (Johnson et al., 2016) and MIMIC-CXR (Johnson et al., 2019) Radiology Report Summarization datasets. The statistics and sources of the data splits are summarized in Table 1.

3.2 Implementation

As the MIMIC-CXR dataset contains more data than the MIMIC-III dataset, we have chosen to use a pre-trained MIMIC-CXR model for fine-tuning on the MIMIC-III dataset, as this approach has been shown to enhance overall accuracy.

3.2.1 MIMIC-CXR

For MIMIC-CXR, we fine-tune the multimodal retrieval model using the clinical T5 base model and fine-tune the language model using various pre-trained T5 model variants, including Clinical-T5 base and large models (Lu et al., 2022), Flan-t5-xl (Chung et al., 2022), and two Flan-t5 model variants (flan-t5-3b-summarizer and flan-t5-small-finetuned-open-summarize_from_feedback). Additionally, we explore the few-shot learning model with CoT and finally ensemble seven models that generate a total of nine predictions. Specifically, we obtain two predictions from the clinical T5 base models, three predictions from the clinical T5 large models, three predictions from the Flan-t5 variant models, and one prediction from the few-shot model with CoT. To obtain different predictions from the clinical T5 base and large models, we use

	Hidden Test				Open Test			
Team	**BLEU4**	**ROUGE-L**	**Bscore**	**F1Rad**	**BLEU4**	**ROUGE-L**	**Bscore**	**F1Rad**
shs-te-dti-mai	**18.36**	**35.32**	**57.26**	**36.94**	**17.33**	33.93	55.49	34.93
aimi	16.61	33.43	55.54	35.12	1.25	24.45	45.54	21.24
sinai	17.38	32.32	55.04	33.96	17.12	31.62	54.33	32.65
knowlab	13.23	32.02	55.64	33.39	13.86	32.22	54.91	32.49
nav-nlp	15.13	32.39	55.34	33.37	15.31	32.33	54.49	32.68
elirf	18.06	30.19	53.94	32.58	17.41	29.57	52.24	31.40
utsa-nlp	16.05	34.41	57.08	36.31	15.99	**34.07**	**56.30**	**35.25**

Table 2: Official results on MIMIC-III open test and hidden test set.

	Hidden Test					Open Test				
Team	**BLEU4**	**ROUGE-L**	**Bscore**	**F1CheX**	**F1Rad**	**BLEU4**	**ROUGE-L**	**Bscore**	**F1CheX**	**F1Rad**
dmis-msra	**18.62**	34.57	**55.90**	**72.36**	**43.20**	25.58	47.75	**64.80**	76.29	50.96
knowlab	14.41	33.63	54.72	67.20	39.98	22.97	46.15	63.43	75.14	48.04
shs-te-dti-mai	14.59	32.43	53.99	68.99	38.40	25.32	47.48	63.61	74.34	49.00
aimi	5.15	31.84	47.83	64.18	32.05	—	—	—	—	—
iuteam1	1.99	26.08	46.75	40.28	27.35	10.10	40.44	56.44	58.01	39.48
e-health csiro	4.12	21.58	43.86	53.46	23.86	17.97	44.14	61.47	71.67	44.95
nlpaueb	5.03	19.87	41.84	50.69	23.26	11.69	36.80	55.50	59.53	36.92
utsa-nlp	16.33	**34.97**	55.54	69.41	42.86	**25.87**	**47.86**	64.74	**77.93**	**51.84**

Table 3: Official results on MIMIC-CXR open test and hidden test sets.

beam search decoding and choose the top two or three predictions to be used for ensembling.

3.2.2 MIMIC-III

The MIMIC-III impression prediction ensemble consists of six models: 1) fine-turned on retrieval-based MIMIC-CXR clinical T5 model with original and back-translated data, 2) fine-tuned on MIMIC-CXR text-only model with original data, 3-5) fine-tuned on retrieval-based MIMIC-CXR clinical T5 model with original data using varying epochs, and 6) the best of models 3-5 fine-tuned with back-translated data. This approach enhances prediction accuracy by leveraging multiple models.

3.3 Evaluation Metrics

Performance on the MIMIC-CXR datasets is evaluated using the following metrics: BLEU4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), Bscore (Zhang* et al., 2020), F1CheXbert (Delbrouck et al., 2022b), and F1RadGraph (Delbrouck et al., 2022a). Performance on the MIMIC-III dataset is evaluated using the following metrics: BLEU4, ROUGE-L, Bscore, and F1RadGraph. We primarily try to balance BLEU4 and ROUGE-L to achieve a higher F1RadGraph score.

4 Results

Table 2 shows the official results of participating systems on the MIMIC-III test splits, where systems are ranked by F1RadGraph score. Overall, our system achieved the second-highest results for all metrics on the Hidden Test set and the highest results on the Open Test set, with the exception of the BLEU4 score.

Table 3 present the results on the MIMIC-CXR tests split ranked by the F1RadGraph score. For this dataset, we had the highest result for the ROUGE-L score on the hidden test set, and second place for all other metrics on the hidden test set. For the Open Test set, we had the highest scores for all metrics except for the BertScore (BScore) metric, in which we placed second. Please refer to Appendix A.2 for additional ablation details.

5 Error Analysis

Based on our error analysis, we observe that our model is less accurate at generating negative impressions for radiology reports. Positive impressions typically indicate the presence of certain symptoms, whereas negative impressions suggest the absence of symptoms with limited contextual information. However, negative impressions share the same meaning across different anatomies, despite varying input forms. Possibly due to the

negative impression's simplistic format, our model struggles to capture anatomy and observation semantics meaning, leading to a higher likelihood of predicting "no acute cardiopulmonary process."

Negative Impression Example: "findings: No focal consolidation is seen. There is no pleural effusion or pneumothorax. Cardiac and mediastinal silhouettes are stable with possible slight decrease in right paratracheal prominence."

For example, the ground truth states "no radiographic findings to suggest pneumonia" while our model generates "no acute cardiopulmonary process" as its prediction. The F1RadGraph score confirms that there is no similarity between the two. This problem affects both general and clinical domain models, indicating their difficulty in accurately recognizing "pneumonia".

Negative Impression Example: "findings: An endotracheal tube, ... There is no change in left lower lobe opacity. There is no large pleural effusion, or pneumothorax. The cardiac silhouette remains moderately enlarged, mediastinal contours are notable for calcification of the aortic arch."

When comparing the clinical and general domain ensemble models, we notice that for the ground-truth of "mild residual retrocardiac opacification remains, pneumonia vs. atelectasis", the clinical model's prediction is "unchanged left lower lobe opacity, likely atelectasis, though pneumonia not excluded" (F1RadGraph is 30.76), while the general model predicts "no change in left lower lobe opacity" (F1RadGraph is 0). This shows the clinical model captures clinically specific semantics.

Short Impression Example: "findings: A Port-A-Cath terminates in the upper right atrium. The cardiac, mediastinal and hilar contours appear unchanged. Fine reticulation associated with pulmonary fibrosis appears very similar within each lung in extent and distribution with no significant superimposed change. The lung volumes are low. There is no pleural effusion or pneumothorax. Multiple compression deformities including lower thoracic vertebroplasties appear unchanged."

In this particular case, the ground truth indicates "no evidence of acute disease. Severe pulmonary fibrosis, not significantly changed." The clinical ensemble model predicts "stable appearance of the chest with no definite superimposed process", while the general ensemble model predicts "no evidence of acute disease". While the general ensemble model seems to perform better for negative impressions in this case, it's worth noting that the model's performance can vary depending on the context and specific examples. It is not always the case that the general model performs better than the clinical model for negative impressions.

Image-Referencing Example: **findings**: "PA and lateral views of the chest __ at 13:47 are submitted." **impression**: "overall cardiac and mediastinal contours ... lateral projection."

In the example above, the findings only mention a chest X-ray without useful text information, while the ground-truth impression is much more descriptive. Therefore, a text-only model would fail to generate an impression in this case. However, our retrieval model successfully handles this issue by extracting major information based on similar reports of the findings. The F1RadGraph score of 42.30 between our best prediction and the ground truth demonstrates the effectiveness of our model in capturing key information from images.

6 Conclusion

This paper presents our system at Task 1B: Radiology Report Summarization of the BioNLP 2023 shared task. We build our system on the basis of a multimodal encoder and pre-trained T5-based models. We employ similarity matching to deal with multimodal inputs and generate a more precise description of the given images and findings. Additionally, we use few-shot with chain-of-thought and ensemble techniques to further enhance the performance of our approach. As a result, our proposed approach achieved second place in both MIMIC-CXR and MIMIC-III hidden tests with F1RadGraph scores of 42.86 and 36.31, respectively. Furthermore, we ranked first in the open test with F1RadGraph scores of 51.84 on MIMIC-CXR and 35.25 on MIMIC-III, out of 11 teams.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. BDKG at MEDIQA 2021: System report for the radiology report summarization task. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 103–111, Online. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Cassie Zhang, and Daniel Rubin. 2021. QIAI at MEDIQA 2021: Multimodal radiology report summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 285–290, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR.

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 465. American Medical Informatics Association.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. 2023. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text (version 1.0.0). *PhysioNet*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443,

Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zihao Wu, Lu Zhang, Chao Cao, Xiaowei Yu, Haixing Dai, Chong Ma, Zhengliang Liu, Lin Zhao, Gang Li, Wei Liu, et al. 2023. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. *arXiv preprint arXiv:2304.09138*.

Liwen Xu, Yan Zhang, Lei Hong, Yi Cai, and Szui Sung. 2021. ChicHealth @ MEDIQA 2021: Exploring the limits of pre-trained seq2seq models for medical summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 263–267, Online. Association for Computational Linguistics.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, pages 2022–08.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Wei Zhu, Yilong He, Ling Chai, Yunxiao Fan, Yuan Ni, Guotong Xie, and Xiaoling Wang. 2021. paht_nlp @ MEDIQA 2021: Multi-grained query focused multi-answer summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 96–102, Online. Association for Computational Linguistics.

A Appendix

A.1 Implementation Detail

The multimodal encoder is fine-tuned on multimodal triplet datasets for a maximum of 10 epochs, while the language model is fine-tuned on the retrieved corpus for up to 20 epochs with a learning rate of 1e-4. We use the AdamW optimizer (Loshchilov and Hutter, 2017) with a mini-batch size of 16 and a sample dropout of 0.1 and projected the last hidden state of the image modality to the same dimension as the text modality. The input tokenizer has a maximum length of 512 for findings and 128 for impressions. For the language model, the output generation maximum length is set to 128 with a beam size of 6 and early stopping. All experiments are conducted on four Nvidia GeForce GTX 1080 Ti GPUs and one Nvidia A6000.

A.2 Ablation Analysis

Effect of Ensemble. We examine the impact of model ensemble (as shown in Table 4) by categorizing pre-trained models into three groups: pre-trained on MIMIC-III/IV, pre-trained on general domain, and few-shot with chain of thought prompt.

We can see in Table 7 in the Appendix, the few-shot with CoT model has a competitive F1RadGraph score for hidden tests (34.20 F1RadGraph score for few-shot with CoT vs 42.86 for our best model), but lower performance for open tests, likely due to prompt learning sensitivity case by case. Due to time constraints during the competition, we could not test all combined models. As a result, we added CoT model to our ensemble model, which decreased the F1RadGraph score by 0.39. Therefore, in our next ensemble ablation tests, we will remove few-shot learning. The F1RadGraph score for our clinical T5 ensemble model is 49.70, while the score for the general domain ensemble model is slightly lower at 49.19. However, these two ensemble models exhibit different prediction behavior, as we will discuss further in the error analysis seen in Appendix.

Effect of Retrieval Model on MIMIC-CXR. We evaluate the performance of our best clinical T5 base retrieval model by comparing it with text-only, image-only, and multimodal models, as shown in Table 5. We can see that our retrieval model outperforms both the text-only model and the multimodal model that simply combines image and text embeddings to generate the impression. Notably, we find that the text-only model performs better than the multimodal model, indicating that directly concatenating image information in the T5 model may not enhance its summarization performance.

Model	BLEU	ROUGE-L	Bscore	F1CheXbert	F1RadGraph
Best Ensemble Model	25.87	47.86	64.74	77.93	51.84
— Chain-of-thought Few-shot	25.49	48.17	65.23	77.94	52.23
— General Domain	26.11	46.34	64.24	77.47	49.70
— Clinical Domain	23.16	46.72	63.80	74.28	49.19

Table 4: Ablation results for Ensemble Model.

Model	BLEU	ROUGE-L	Bscore	F1CheXbert	F1RadGraph
Best Clinical-T5 Base Model	22.34	47.01	64.60	75.97	49.49
— Retrieval (multi-model)	23.94	49.80	67.97	73.20	45.30
— Image (text-only)	22.06	47.20	64.83	76.33	48.76
— Text (image-only)	6.37	28.24	51.27	40.47	17.81

Table 5: Ablation results for Retrieval Model.

	Hidden Test				Open Test			
Model	**BLEU4**	**ROUGE-L**	**Bscore**	**F1RadGraph**	**BLEU4**	**ROUGE-L**	**Bscore**	**F1RadGraph**
Ensemble	16.05	34.41	57.08	36.31	15.99	34.07	56.30	35.25
fine-tuned cxr	14.69	33.51	56.16	34.18	14.61	33.25	55.22	33.21
fine-tuned cxr with back-translation	14.69	33.75	56.54	34.78	16.35	33.21	56.05	33.71

Table 6: Our results on MIMIC-III open test and hidden test sets.

	Hidden Test					Open Test				
Model	**BLEU4**	**ROUGE-L**	**Bscore**	**F1CheXbert**	**F1RadGraph**	**BLEU4**	**ROUGE-L**	**Bscore**	**F1CheXbert**	**F1RadGraph**
Ensemble	16.33	34.97	55.54	69.41	42.86	25.87	47.86	64.74	77.93	51.84
Few-shot with CoT	15.81	28.48	49.52	67.27	34.20	7.90	22.22	45.74	63.47	20.83
clinical t5 base	11.71	33.11	53.74	65.20	38.54	22.61	47.61	64.58	75.94	49.50
clinical t5 large	12.39	32.42	53.06	64.21	38.33	24.58	48.37	64.92	76.92	49.58

Table 7: Our results on MIMIC-CXR open test and hidden test sets.

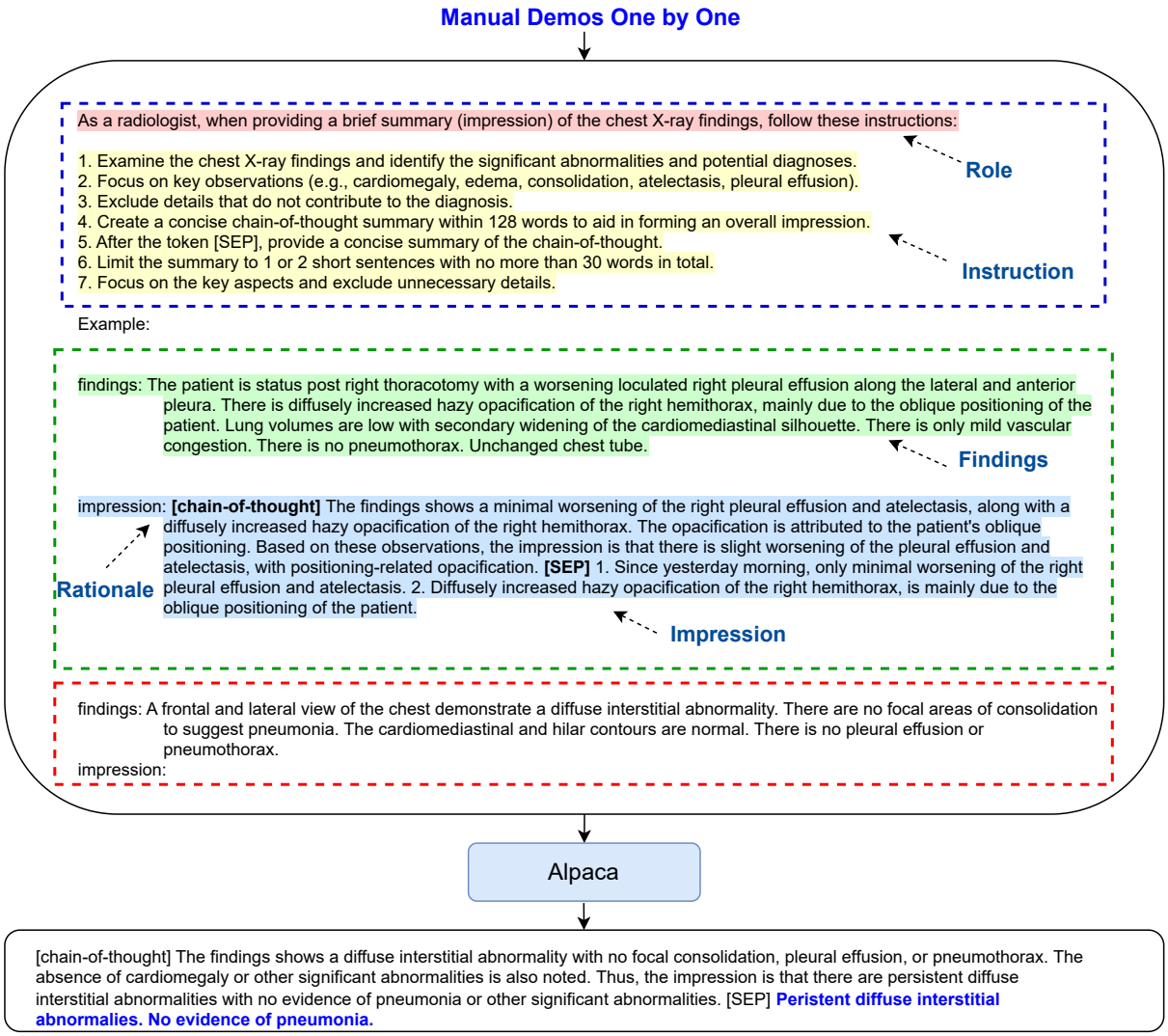

Figure 2: Few-Shot-CoT using manually demonstrations one by one with example inputs and outputs of an Alpaca LLM.