The More, The Better? A Critical Study of Multimodal Context in Radiology Report Summarization

Mong Yuan Sim[1,2] **Wei Emma Zhang**[1] **Xiang Dai**[2] **Biaoyan Fang**[3]
Sarbin Ranjitkar[1] **Arjun Burlakoti**[4] **Jamie Taylor**[1] **Haojie Zhuang**[5]
[1]The University of Adelaide [2]CSIRO Data61 [3]Oracle
[4]University of South Australia [5]The Australian National University
{mongyuan.sim;wei.e.zhang}@adelaide.edu.au

Abstract

The IMPRESSION section of a radiology report summarizes critical findings of a radiology report and thus plays a crucial role in communication between radiologists and physicians. Research on radiology report summarization mostly focuses on generating the IMPRESSION section by summarizing information from the FINDINGS section, which typically details the radiologist's observations in the radiology images. Recent work start to explore how to incorporate radiology images as input to multimodal summarization models, with the assumption that it can improve generated summary quality, as it contains richer information. However, the real effectiveness of radiology images remains inconclusive. To address this, we conduct a thorough analysis to understand whether existing multimodal models can effectively utilize radiology images in generating a summary of the FINDINGS section. Our analysis reveals that multimodal models might not actually make use of radiology images. For example, masking the image inputs leads to minimal or no performance drop compared to the original images when they are used as input to a trained multimodal summarizer. An expert annotation study on two widely used datasets also shows that radiology images are often unnecessary for writing the IMPRESSION section if FINDINGS section is provided [1].

1 Introduction

Radiology report acts as a bridge of communication between radiologists and physicians (Kahn et al., 2009; Gershanik et al., 2011). Having reliable radiology reports could improve communication, enhance patient care, and facilitate research and data analysis in the field of radiology (Pesapane et al., 2023). A standard radiology report usually contains REASON FOR THE EXAM, COMPARISON (with any available previous exams), FINDINGS, and IMPRESSION sections (Naik et al., 2001; Wallis and McCoubrie, 2011). Specifically, FINDINGS section, describes what the radiologist observes in the image(s), and IMPRESSION section, summarizes important findings and possible causes (differential diagnosis) (Wallis and McCoubrie, 2011) are the two most critical sections in the radiology report analysis.

Earlier work on radiology report summarization aims to generate IMPRESSION given the FINDINGS section as input (Zhang et al., 2018; Miura et al., 2021; Ben Abacha et al., 2021). More recently, researchers (Kim et al., 2023; Wang et al., 2023; Nicolson et al., 2023) explored the usage of radiology images as additional context, aiming to improve the generated IMPRESSION, under the assumption that radiology images could provide richer information to generate a more accurate IMPRESSION section. However, the effectiveness of incorporating radiology images in radiology report summarization remains inconclusive.

In this paper, we critically examine the effectiveness of using radiology images in radiology report summarization. Our key research question is: *To what extent does the radiology image contribute to generating more accurate and informative* IMPRESSION *section?* To this end, we systematically perform controlled experiments with several multimodal summarization models. We first observe that text-only models often outperform multimodal models but masking image inputs lead to minimal or no performance drop, raising concerns about whether visual information is genuinely utilized by multimodal summarization models. We then explore two-stage fine-tuning strategies aimed at encouraging image utilization and introduce an exclusive set, by removing sentences describing critical diagnosis from the text input, enforcing the reliance of radiology images. Finally, we conduct an expert annotation study to assess the actual necessity of ra-

[1] Data pre-processing code, note IDs, and data split are available at https://github.com/raymondsim/rrs-analysis.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19116–19131
November 4-9, 2025 ©2025 Association for Computational Linguistics

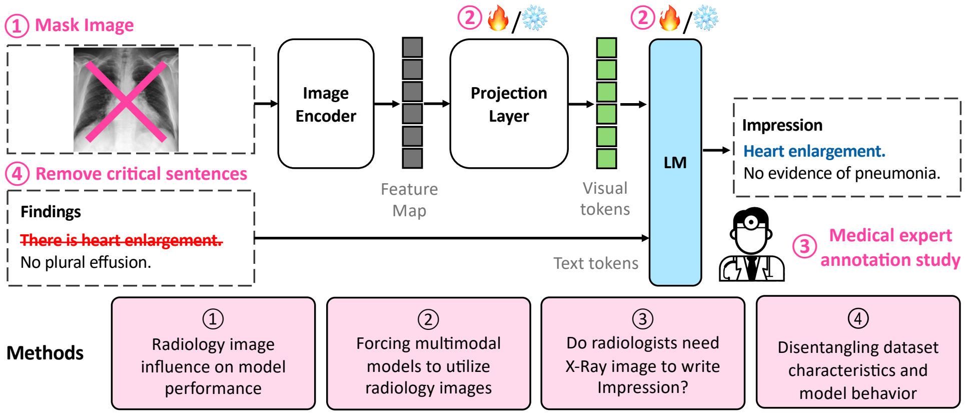
① Mask Image
②🔥/❄️
②🔥/❄️
Image Encoder
Feature Map
Projection Layer
Visual tokens
LM
Impression
Heart enlargement.
No evidence of pneumonia.
④ Remove critical sentences
Findings
~~There is heart enlargement.~~
No plural effusion.
Text tokens
③ Medical expert annotation study
Methods
① Radiology image influence on model performance
② Forcing multimodal models to utilize radiology images
③ Do radiologists need X-Ray image to write Impression?
④ Disentangling dataset characteristics and model behavior

Figure 1: An overview of our study. In this study, we (1) train summarization models under multimodal setting, mask image during inference time, and observe performance changes and how model attends to visual tokens, (2) to track the bottleneck, we train projection layer and language model separately, (3) we conducted an annotation study, to understand if radiologists need radiology image to write the IMPRESSION section, (4) to disentangle model behavior and dataset characteristics, we constructed an exclusive set, where text input has missing critical information, in order to assess a model's ability in utilizing visual input.

diology images for radiology report summarization task, using two widely used datasets, MIMIC-HIST-AUG and CheXpert.

Based on our analysis, we argue that assessing the utility of different input modalities is crucial when building a multimodal model (Sim et al., 2025). This is particularly important for domain-specific tasks like radiology report summarization, where the accuracy of generated content can directly influence clinical decision-making and patient outcomes.

To review, our contributions in this paper are three-fold:

- A thorough analysis of model behavior and dataset characteristics to investigate the (in)effectiveness of visual information in radiology report summarization.
- An expert annotation study that aligns with our analysis findings and provides suggestions on the additional context that is needed to generate a more accurate IMPRESSION.
- We explore a two-stage fine-tuning strategy and introduce an **exclusive set**, designed to disentangle dataset characteristics from model behavior, providing insights into the utility of image inputs in multimodal models.

2 Analysis Setup

2.1 Problem Formulation

The task of multimodal radiology report summarization is defined as: given a text X_T and a radiology image X_V, we aim to generate an IMPRESSION section of a radiology report that summarizes critical findings. In our study, we use the BACKGROUND section (if available) concatenated with the FINDINGS section as input text, X_T.

2.2 Multimodal Model Architecture

Most multimodal summarization models, including recent Large Vision Language Models (LVLMs) such as LLaVA, DeepSeek-VL, and Qwen-VL, typically follow a three-stage architecture that integrates visual and textual inputs. The model comprises: (1) a frozen vision encoder f_{vision}, (2) a trainable projection layer f_{proj}, and (3) a pretrained language model f_{LM}, as illustrated in Figure 1.

Given an input image X_V and a sequence of text tokens X_T, the image is first encoded into visual features:

$$V = f_{\text{vision}}(X_V)$$

These features are then mapped into the language embedding space using a projection layer:

$$V' = f_{\text{proj}}(V)$$

Finally, the language model conditions on both the text X_T and the projected image features V' to generate the output:

$$\hat{Y} = f_{\text{LM}}(X_T, V')$$

This architecture enables the model to perform multimodal processing by attending to both modalities. In our experiments, we leverage this modular structure to explore whether and how the model utilizes image features by applying different fine-tuning strategies and controlled experiments.

2.3 Models

Text-only Summarization Models We selected four text-only summarization models as baseline methods. We use (1) **PG** (See et al., 2017), pointer-generator model introduces copy and coverage mechanism to the standard seq2seq framework, (2) **BART** (Lewis et al., 2020), a pretrained encoder-decoder summarization model whose pre-training objective is reconstructing corrupted documents, (3) **GSum** (Trienes et al., 2023) which extends pretrained BART with a guidance encoder, (4) **WGSum** (Hu et al., 2021) is a domain-specific method for radiology report summarization that uses a graph-guided decoder to attend a graph of clinical entities. These models were selected as they are either radiology domain-specific or widely used for radiology report summarization tasks.

Multimodal Summarization Models We selected six radiology domain-specific multimodal models. (1) **Vilmedic** (Delbrouck et al., 2021) uses an RNN encoder-decoder model and fuses visual information extracted from pre-trained Densenet101 (Huang et al., 2017) using dot-product attention, (2) **CvT-BERT** (Nicolson et al., 2023) uses pretrained Convolutional vision Transformer (CvT) as image encoder to extract image features, then uses BERT as a decoder to generate IMPRESSION section conditioned on FINDINGS, (3) **VG-BART** (Yu et al., 2021) uses dot-product attention and multi-head attention methods to incorporate visual information into BART. In addition, we also adapt some representative Large Vision Language Models (LVLMs), including (4) **LLaVA-1.5** (Liu et al., 2023), **LLaVA-Med** (Li et al., 2023), (5) **Qwen-VL** (Bai et al., 2023), and (6) **DeepSeek-VL** (Lu et al., 2024). We use 7b version for all LVLMs.

2.4 Datasets

MIMIC-CXR (Johnson et al., 2019) is a large-scale radiology report summarization dataset, consisting of 377,110 chest X-Rays and 227,827 associated radiology reports, collected from Beth Israel Deaconess Medical Center between 2011 - 2016. We follow the official train/validation/test split.

OpenI (Demner-Fushman et al., 2015) is a small-scale radiology report collected by Indiana University with 3,346 reports after preprocessing. We follow the official train/validation/test split.

	MIMIC-CXR	OpenI	CheXpert
# exp	122k / 963 / 1.5k	2.2k / 386 / 653	42k / 1k / 2k
EXCLUSIVE	22,224 / - / 767	159 / - / 70	14,906 / - / 1,309
# sent$_x$	5.48	5.83	5.81
# word$_x$	56.35	52.38	65.85
# sent$_y$	1.61	1.42	4.47
# word$_y$	15.89	9.78	36.93

Table 1: Dataset statistics including number of samples for train/val/test split, number of sentences and tokens in FINDINGS and IMPRESSION.

CheXpert (Irvin et al., 2019) is a large-scale chest radiograph dataset collected from Stanford Hospital, for studies conducted between October 2002 and July 2017. [2]

2.5 Evaluation Metrics

We assess the quality of generated summaries from two dimensions: i) similarity to reference summaries and ii) correctness using domain-specific faithfulness metrics. For similarity, we use ROUGE score (Lin, 2004), which computes the overlapping n-grams, word pairs, and word sequences between generated summaries and ground truth. We report F_1 score for ROUGE-1, ROUGE-2 and ROUGE-L.

For factual correctness, we use pretrained ChexBert (Smit et al., 2020a) to obtain 14-class labels for generated summaries and ground truth, whose results will then be used to compute F_1 score. We also use RadGraph [3] score (Jain et al., 2021) to measure factual correctness and completeness.

3 Text-Only Summarization vs. Multimodal Summarization

To understand whether incorporating radiology images in summarization models can improve summary quality, we compare the performance of representative text-only models and multimodal models, including Large Vision Language Models (LVLMs). In addition, we design an ablation experiment on multimodal models trained on paired FINDINGS and radiology images to generate the IMPRESSION section. At inference time, we mask

[2] As the official dataset does not include text reports, we use the shared task version of the dataset from RRG24 shared task (Xu et al., 2024), which provides train and test set. We randomly sampled 2000 instances from the training set and use them as the validation set.

[3] RadGraph-XL is used and partial reward are used.

	MIMIC-CXR				OpenI				CheXpert			
	R-1	R-L	RadG	CBert	R-1	R-L	RadG	CBert	R-1	R-L	RadG	CBert
Text-Only Models												
PG	40.14	36.84	26.72	54.44	57.80	57.27	48.76	84.53	42.70	37.49	14.39	42.00
BART	**50.86**	**47.00**	40.90	65.39	**68.51**	**67.73**	**63.35**	87.29	**57.46**	**51.94**	19.84	56.85
GSum	45.32	42.98	35.18	59.60	61.76	60.10	52.32	85.28	51.92	47.09	15.22	49.01
WGSum	43.92	41.55	31.17	58.92	60.52	58.67	50.36	84.73	44.64	41.02	14.89	45.00
Multimodal Models												
VG-BART Dot	50.69	46.68	40.92	**65.59**	51.38	51.21	42.24	84.84	55.07	49.84	18.63	56.45
VG-BART MHA	50.64	46.89	**41.10**	65.33	56.30	56.14	47.08	84.23	54.63	49.02	18.70	55.60
CvT-BERT	43.85	44.32	37.76	58.52	60.12	57.91	50.16	84.10	49.01	44.56	18.90	51.00
Vilmedic	35.58	34.13	28.13	54.13	63.95	63.88	60.25	84.53	50.94	46.15	18.63	52.70
LLaVA-1.5	47.32	44.32	38.81	62.14	50.78	50.13	43.96	87.44	51.22	45.26	37.63	53.90
Qwen-VL	47.62	44.30	38.83	63.08	49.67	49.52	43.41	86.62	51.54	45.45	**38.76**	53.81
DeepSeek-VL	46.70	43.12	37.79	62.95	49.51	49.08	42.50	86.83	49.19	42.86	35.93	54.90
LLaVA-Med	47.58	44.53	38.81	62.98	50.91	50.93	44.18	**88.41**	51.64	46.01	38.61	**57.27**

Table 2: The performance of all text-only and multimodal baseline methods on the test sets of MIMIC-CXR, OpenI, and CheXpert datasets. F1 score is reported for all evaluation metrics. The best score and upper bound for all metrics are 100. RadG refers to RadGraph, CBert refers to ChexBert.

the image input and provide the FINDINGS section only. In other words, we report results of models trained on multimodal inputs, but with the image masked during inference time. We conjecture that any performance drop from this masking operation indicates the extent to which the model relies on image input.

Results From Table 2, we found that incorporating visual information (i.e., multimodal summarization models) performs worse than text-only models on most evaluation metrics, including ROUGE scores and radiology factuality measures. This finding could partially be explained by how a radiology report is constructed. Radiologists first write the FINDINGS section based on the radiology image, and then write the IMPRESSION section based on the FINDINGS section. If a radiologist writes the FINDINGS section with sufficient details, this section should include all key information from radiology images. Another possible reason is that multimodal models failed to learn meaningful cross-modal interactions between text and visual representations.

In addition, we observe that LVLMs like LLaVA, Qwen-VL, and DeepSeek-VL perform slightly worse than smaller models like BART and VG-BART. We hypothesize that large models require more data during fine-tuning. We leave this for future work, as the aim of this study is to understand the utilization of radiology images in generating IMPRESSION section.

We assess the contribution of visual information in multimodal radiology report summarization by masking the image input at inference time. Specifically, we replace the X-Ray image with a black image (all-zero pixel values) and measure the performance drop compared to using the original images. From Figure 2, we observe that masking images on multimodal models has minimal to no negative effect on overall model performance. This indicates that multimodal models do not genuinely utilize image input when generating IMPRESSION section, they heavily rely on the text input (i.e., FINDINGS section).

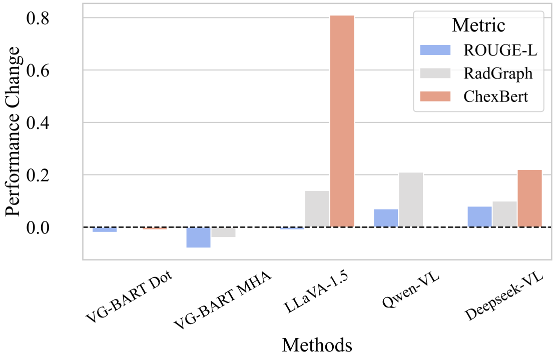

Figure 2: Performance change (Δ score) when the image is masked, evaluated on MIMIC-CXR. Upper bounds for all metrics are 100. Masking images only causes minimal to no negative effect on the model's overall performance. Complete results are reported in Appendix E.

4 Two-Stage Fine-Tuning: Encourage Image Utilization

Most multimodal models consist of three main components: an image encoder, a language model, and a projection layer that projects visual embeddings

		MIMIC-CXR			OpenI			CheXpert		
Model	Setting	R-L	RadG	CBert	R-L	RadG	CBert	R-L	RadG	CBert
LLaVA-1.5	Proj	36.09	30.62	55.44	6.72	6.79	78.10	35.35	24.18	48.10
	LM	**44.32**	**38.80**	**62.14**	49.98	43.96	87.44	45.26	37.63	53.90
	Proj + LM	44.27	38.38	61.76	**59.13**	**53.53**	86.68	**45.46**	38.10	**54.10**
	Proj → LM	44.06	38.34	61.76	50.33	44.05	87.90	45.42	37.79	53.20
	Proj → Proj + LM	43.83	37.98	61.76	50.75	44.29	**88.36**	45.35	**38.45**	53.45
Qwen-VL	Proj	36.76	31.44	56.45	6.69	8.33	85.45	40.25	29.13	50.50
	LM	44.30	38.83	63.08	52.66	45.78	87.14	45.45	38.41	54.25
	Proj + LM	**44.41**	**39.26**	63.52	**62.90**	**58.24**	87.60	45.43	38.50	53.95
	Proj → LM	44.10	38.76	**63.58**	52.33	45.75	87.44	45.31	**38.57**	54.65
	Proj → Proj + LM	44.10	38.62	63.14	52.11	45.61	**87.90**	**45.59**	38.30	**54.85**
DeepSeek-VL	Proj	35.47	30.59	56.20	13.41	8.06	84.53	31.20	18.30	48.20
	LM	**43.12**	37.79	62.95	49.01	42.50	**86.83**	42.86	34.11	50.25
	Proj + LM	42.80	37.39	62.39	29.80	23.95	52.57	**43.03**	34.56	50.20
	Proj → LM	42.72	37.80	63.45	48.81	42.54	86.06	42.67	34.15	**52.40**
	Proj → Proj + LM	42.80	**37.87**	**63.70**	**49.47**	**42.68**	**86.83**	42.94	**35.10**	52.15

Table 3: The performance of LVLMs on the test sets of MIMIC-CXR, OpenI and CheXpert datasets, fine-tuned with different strategies. F1 score is reported for all evaluation metrics. The best score and upper bound for all metrics are 100. RadG refers to RadGraph, CBert refers to ChexBert.

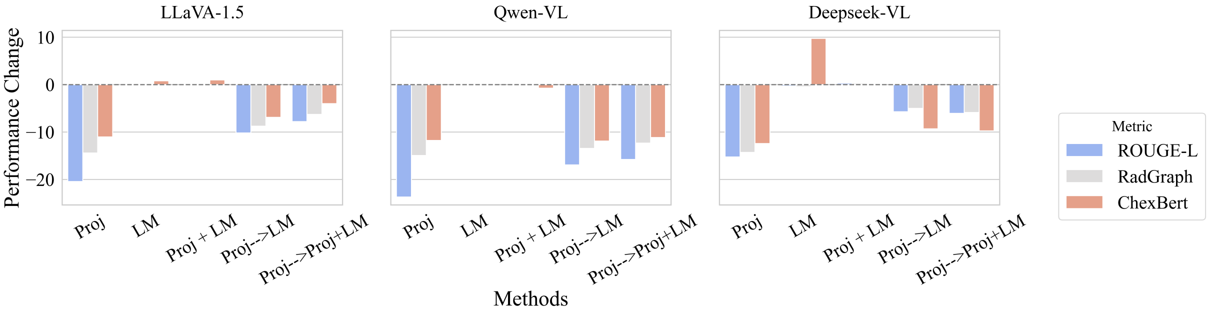

Figure 3: Performance changes on LLaVA-1.5, Qwen-VL, and DeepSeek-VL trained on MIMIC-CXR with different training strategies. Results showed that training the projection layer only, or two-stage fine-tuning, encourages models to utilize image inputs, as masking the image causes a significant drop.

to the language model representation space. We suspect that the FINDINGS section in these datasets might contain sufficient information to generate the IMPRESSION section. Hence, multimodal summarization models might learn to rely on text input while ignoring image input, as the latter is harder to interpret (Verma et al., 2024). To track the bottleneck where image input is ignored, we conducted a controlled experiment by training the projection layer and the language model separately, together, and sequentially. Specifically, there are five settings as shown in Table 3, which are: i) projection layer only, ii) language model only, iii) projection layer and language model, iv) projection layer first, then language model, v) projection layer first, then projection layer and language model.

We hypothesize that training the projection layer forces the model to map visual features to corresponding medical concepts in vocabulary space. Training the projection layer and the language model together will cause the text input to dominate, and the model will take shortcuts and heavily rely on text input only, ignoring image input.

Results Our results are presented in Table 3. We can see that training the projection layer only forces the model to align image representation to the language model's textual space, and masking the image input leads to a significant performance drop (Figure 3). Training the language model along with the projection layer enables the model to utilize text input only; masking image input does not affect the overall performance. Two-stage fine-tuning, where training the projection layer first, and then the language model and/or the projection layer, preserves the model's ability in projecting image input to language model textual space, and the language model learns to utilize this image input; masking image input leads to a performance drop.

We also visualized the attention weights on visual tokens at inference as shown in Figure 4. Specifically, we averaged the attention weights on visual tokens at each generated token. From the visualization plot, we observe that LLaVA-1.5 trained on the projection layer only relies heavily on visual features. Training the language model only or the language model with the projection layer on LLaVA-1.5 causes the model to ignore visual features, as it has extremely low attention weights on visual tokens. Models trained with two-stage fine-tuning (i.e., fine-tune the projection layer first, then the language model) attend to visual tokens similar to training the projection layer only.

However, it is worth noting that the overall performance of such a two-stage fine-tuning strategy is very close to training the language model and/or projection layer together. We suspect this is because text input (i.e., FINDINGS section) contains rich information that is sufficient to generate IMPRESSION section.

5 Do Medical Experts Need Image To Write IMPRESSION section?

Dataset	Raw Agreement	Avg. "Yes" %
MIMIC-CXR	96%	94%
CheXpert	86%	88%

Table 4: Inter-annotator agreement and results on the annotation question: "Can you write the IMPRESSION section using only the FINDINGS section, without seeing the radiology image?"

To understand the utilization of text and vision modalities by medical practitioners, we conducted an annotation study with the help of two medical experts with over 20 years of experience. We provided 100 samples from MIMIC-CXR and CheXpert (50 randomly selected samples from each dataset) to two annotators. Specifically, we provide X-Ray images, FINDINGS, and IMPRESSION section, and asked whether it is possible to write a complete IMPRESSION section from the FINDINGS section only, without using the corresponding radiology image. We perform a medical expert annotation study on MIMIC-CXR and CheXpert datasets only, as OpenI dataset is relatively smaller compared to MIMIC-CXR and CheXpert. It might not reflect the real characteristics of radiology report datasets in a real clinical setting. The annotation interface and guidelines can be found in Appendix B.

Results Our annotation study result is presented in Table 4. We find medical experts agree that radiology images are unnecessary in writing the IMPRESSION section most of the time (94% for MIMIC-CXR and 88% for CheXpert). Out of the annotated samples, samples that require information from radiology images are those FINDINGS section that do not contain sufficient information of the current study. That is, some diagnosis are mentioned in IMPRESSION section but not in FINDINGS section. This annotation study results align with our model performance from the previous section, forcing multimodal summarization models to use radiology images using a two-stage fine-tuning strategy leads to similar or slightly worse performance compared to training the language model only or the language model and projection layer together. Our annotation achieves 96% and 86% agreement for MIMIC-CXR and CheXpert datasets, respectively.

Our annotators pointed out that one critical information missing from MIMIC-CXR and CheXpert datasets is imaging request, which typically indicates specific clinical questions or reasons for the exam. This context is essential in a real-world clinical setting as it guides radiologists to focus on specific concerns when interpreting radiology images or writing reports. Future work on constructing a radiology report summarization dataset could include this information to guide the summarization model in generating a more accurate IMPRESSION section.

6 Isolating Dataset Characteristics And Model Behavior With The Exclusive Set

As IMPRESSION section is written based on the FINDINGS section, we hypothesize that FINDINGS section contains enough information to generate the IMPRESSION section. To test the effectiveness of radiology images, we construct a controlled dataset called the **exclusive set**, where some critical sentences are removed from the FINDINGS section, and can only be obtained from the image input. We hypothesize that two-stage fine-tuning strategy will outperform other methods on the exclusive set, as it enables the model to leverage visual information as discussed in the previous section.

To construct the exclusive set, we first obtain three sets of diagnosis (binary) labels from FINDINGS, IMPRESSION (using ChexBert (Smit et al., 2020b)) and radiology images (using TorchXRayVision (Cohen et al., 2022)) respectively. For FINDINGS and IMPRESSION, each diagnosis label is associated with a target sentence, which is used to justify whether a diagnosis is positive or negative. If a diagnosis is consistent across

		MIMIC-CXR			OpenI			CheXpert		
Model	Setting	R-L	RadG	CBert	R-L	RadG	CBert	R-L	RadG	CBert
LLaVA-1.5	Proj	18.52	14.16	17.34	8.32	**7.98**	27.14	17.50	11.52	19.25
	LM	27.27	22.32	28.03	8.31	7.22	27.14	36.10	23.67	20.47
	Proj + LM	27.06	22.05	28.03	8.36	7.18	25.71	36.21	23.36	20.02
	Proj → LM	**30.19**	**26.12**	**38.20**	**8.40**	7.06	**31.43**	**38.71**	**29.72**	**24.68**
	Proj → Proj + LM	**30.19**	25.86	36.63	8.33	6.70	30.00	30.65	29.49	24.37
Qwen-VL	Proj	13.46	15.37	16.20	7.20	8.52	27.14	28.77	13.90	19.33
	LM	23.77	25.05	22.54	**11.66**	**9.49**	**31.43**	37.72	23.98	21.70
	Proj + LM	**23.89**	24.95	22.69	11.43	9.32	28.57	37.81	24.02	22.23
	Proj → LM	22.73	25.06	25.59	10.85	5.62	30.00	39.79	29.39	25.67
	Proj → Proj + LM	23.03	**25.53**	**26.59**	10.09	5.39	30.00	**40.25**	**30.03**	**26.81**
DeepSeek-VL	Proj	18.76	14.27	17.60	12.85	**9.01**	27.14	9.33	4.64	15.05
	LM	25.67	20.70	22.95	12.84	8.72	25.63	31.59	18.87	17.49
	Proj + LM	26.06	21.27	24.51	12.83	8.72	25.63	31.21	18.52	17.27
	Proj → LM	30.58	**26.47**	38.10	13.11	8.61	**28.97**	37.67	**28.97**	**21.16**
	Proj → Proj + LM	**30.63**	26.32	**38.20**	**13.25**	8.69	**28.97**	**37.99**	28.81	19.94

Table 5: The performance of LVLMs on the test sets of exclusive sets for MIMIC-CXR, OpenI, and CheXpert, fine-tuned with different strategies. F1 score is reported for all evaluation metrics. The best score and upper bound for all metrics are 100. RadG refers to RadGraph, CBert refers to ChexBert.

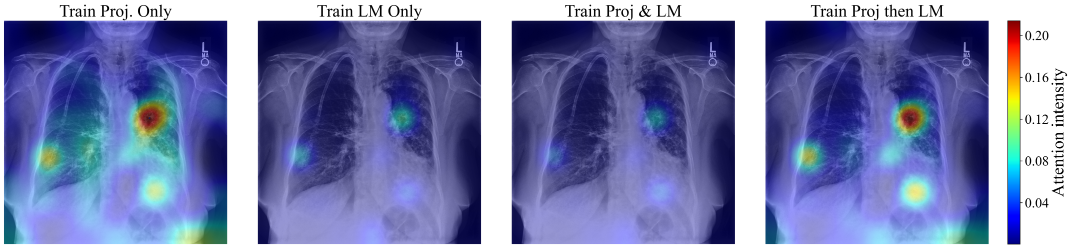

Figure 4: LLaVA-1.5 averaged attention weights on the visual token at generation.

all three label sets (i.e., all positive or all negative), we remove the target sentence from FINDINGS, such that the only source of information for this diagnosis is the radiology image.

For instance, if the label, "Heart Enlargement" is positive for FINDINGS, IMPRESSION, and the corresponding radiology image, we create the exclusive set by removing the target sentence from the original FINDINGS. With this, the augmented FINDINGS and the radiology images contain exclusive information, where FINDINGS do not mention about heart enlargement, while the radiology image indicates heart enlargement.

The previous section shows that encouraging LVLMs to utilize radiology images does not lead to a performance gain. Two-stage fine-tuning strategy achieves similar performance to training the language model only or training the language model and the projection layer together, though masking image input leads to a significant performance drop. Our medical expert annotation study also shows that radiology images are not needed to write the IMPRESSION section. To disentangle model performance and dataset characteristics, we further train and test models on the exclusive set. This exclusive set enables us to test whether a multimodal model is utilizing image input or not, as the radiology image is the only source of information for the target diagnosis.

Results From Table 5, we observe that models trained with two-stage fine-tuning strategies (i.e., projection layer first, then the language model only or with the projection layer) can achieve better overall performance. This suggests that such a training strategy encourages the model to utilize image input and integrate information from both modalities more effectively. The first stage of training fine-tunes the projection layer only with the language model frozen, forcing the model to map visual features to the language model representation space. The second stage training trains the language model and optionally with the projection layer, enables the model to integrate multimodal information and produce a more complete IMPRESSION section.

7 Discussion

Multimodal Radiology Summarization Is Ill-Defined In a clinical setting, radiologists first write the FINDINGS section based on a radiology image and write the IMPRESSION section based on the FINDINGS section. Due to the nature of how the IMPRESSION section is written, radiology image is often unnecessary when writing the IMPRESSION section. This is confirmed by our medical annotators (Section 5). Therefore, we argue that a multimodal radiology report summarization task is ill-defined, as in most examples, FINDINGS section (text input) have provided enough information to generate an accurate IMPRESSION section. As we have shown that X-Ray images are not required to write the IMPRESSION section in general, future work on multimodal radiology report summarization should focus on identifying cases that require additional input, other than the FINDINGS section. We believe this insight may extend to other multimodal tasks as well, particularly those where multimodal inputs are added to tasks that are originally designed as single modality task, resulting in limited contribution from the additional modality. In such a case, the role of each modality should be carefully evaluated before assuming its necessity and effectiveness in model design, to avoid modality collapse (Sim et al., 2025).

Multimodal Models Learn Shortcuts and Completely Ignore One Modality From the medical expert annotation study, we confirm that multimodal inputs (FINDINGS section and radiology image) in radiology report summarization dataset are highly redundant, where the images are not needed to generate the IMPRESSION section. These training data causes the model to heavily rely on text input, and ignore image input completely. This is shown by our ablation experiments, where masking the image leads to minimal or no performance changes, indicating that the model relies on text cues in generating the IMPRESSION section.

Two-Stage Fine-tuning Forces Model to Align Medical Concepts Our experimental results showed that training the projection layer and the language model sequentially encourages the model to utilize both text and image modalities, mitigating the problem of shortcut learning for multimodal tasks. Specifically, the first training stage trains the projection layer to project visual features to the language model representation space, aligning medical concepts to the vocabulary space. The second training stage trains the language model, and optionally the projection layer, to better integrate multimodal inputs in the language model space. This could be a solution for multimodal tasks that require multimodal input integration, but are often dominated by text modality during model training.

8 Related Work

8.1 Radiology Report Summarization

Early work on radiology report summarization focus on utilizing background information (Zhang et al., 2018), attending medical entities from FINDINGS section (Sotudeh Gharebagh et al., 2020), using Graph Neural Networks (Hu et al., 2021, 2022) and RadGraph score (Xie et al., 2023) to guide radiology report summarization.

Some efforts have been made to incorporate visual information from radiology images into the summarization model (Delbrouck et al., 2021; Wang et al., 2023; Nicolson et al., 2023), using methods like medical vision-language model pretraining (Kim et al., 2023), and retrieval-based strategies (Wang et al., 2023). While these methods yield marginal gains (Delbrouck et al., 2021; Kim et al., 2023), it remains unclear whether models truly attend to and utilize visual input.

8.2 Modality Contribution

Recent studies have investigated the contribution of an input modality in multimodal models (Goyal et al., 2017). Parcalabescu and Frank (2023, 2025) use Shapley-values to randomly mask image patches and text tokens to measure the output changes which reflects modality importance. Liang et al. (2023a) proposed an information theory-based framework to estimate a dataset's multimodal interaction and conducted an annotation study on general domain datasets (Liang et al., 2023b). To the best of our knowledge, our work is the first to investigate the role of radiology image in radiology report summarization, from model behavior to medical expert annotation study.

8.3 Multimodal Summarization

Multimodal summarization attracts the interest of the research community as it can utilize multiple modalities and generate an informative summary (Li et al., 2020; Im et al., 2021; Delbrouck et al., 2021; Li et al., 2018; Atri et al., 2021). Some works focus on pretraining vision-language models (VLMs) (Dosovitskiy et al., 2020; Van Veen et al., 2023; Radford et al., 2021). Specifically, Yu et al. (2021) proposed the first general framework to fuse visual information into pretrained language models (PLMs) such as BART and T5, and this framework has been used as the baseline in multimodal summarization tasks.

9 Conclusion & Future Work

In this work, we critically examined the effectiveness of radiology images in multimodal radiology report summarization. Through extensive model analysis, input ablations, and medical expert annotations, our results highlight the need to reassess how multimodal models utilize different modalities, and how multimodal tasks are formulated, along with the necessity of multimodal inputs. Future work should focus on reassessing multimodal task definition, the modality imbalanced problem in datasets, and involving expert annotation to better reflect and align with real-world clinical decision-making.

Limitations

Our study is limited to chest X-ray images within the radiology domain. Hence, findings presented in this paper may not generalize to other imaging modalities such as CT or MRI, which may exhibit different characteristics and cause different model behavior. Additionally, the datasets used in this study are collected from institutions in the United States, where reporting styles and clinical workflows may differ from those in other regions. These factors may influence the behavior of multimodal models across diverse settings. Nonetheless, we believe our empirical analysis provides insights that are broadly informative and could guide future research on multimodal learning in other domain-specific tasks and datasets.

References

Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*, 227:107152.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain. In *BioNLP@NAACL*.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. 2022. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*.

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.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

E. F. Gershanik, R. Lacson, and R. Khorasani. 2011. Critical finding capture in the impression section of radiology reports. *AMIA Annu Symp Proc*, 2011:465–469.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4980–4990, Online. Association for Computational Linguistics.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688, Dublin, Ireland. Association for Computational Linguistics.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. 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)*, pages 388–403, Online. Association for Computational Linguistics.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Preprint*, arXiv:1901.07031.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, D. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curt P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *ArXiv*, abs/2106.14463.

A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Y. Deng, R. G. Mark, and S. Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(1):317.

C. E. Kahn, C. P. Langlotz, E. S. Burnside, J. A. Carrino, D. S. Channin, D. M. Hovsepian, and D. L. Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Gangwoo Kim, Hajung Kim, Lei Ji, Seongsu Bae, Chanhwi Kim, Mujeen Sung, Hyunjae Kim, Kun Yan, Eric Chang, and Jaewoo Kang. 2023. KU-DMIS-MSRA at RadSum23: Pre-trained vision-language model for radiology report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 567–573, Toronto, Canada. Association for Computational Linguistics.

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.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llavamed: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023a. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*.

Paul Pu Liang, Yun Cheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023b. Multimodal fusion interactions: A study of human and automatic quantification. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 425–435, New York, NY, USA. Association for Computing Machinery.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. In *NAACL*.

Sandeep S. Naik, Anthony Hanbidge, and Stephanie R. Wilson. 2001. Radiology reports. *American Journal of Roentgenology*, 176(3):591–598. PMID: 11222186.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. e-health CSIRO at RadSum23: Adapting a chest X-ray report generator to multimodal radiology report summarisation. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 545–549, Toronto, Canada. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2025. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *The Thirteenth International Conference on Learning Representations*.

Filippo Pesapane, Priyan Tantrige, Paolo De Marco, Serena Carriero, Fabio Zugni, Luca Nicosia, Anna Carla Bozzini, Anna Rotili, Antuono Latronico, Francesca Abbate, and 1 others. 2023. Advancements in Standardizing Radiological Reports: A Comprehensive Review. *Medicina*, 59(9):1679.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can VLMs actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020a. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020b. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *Preprint*, arXiv:2004.09167.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Jan Trienes, Paul Youssef, Jörg Schlötterer, and Christin Seifert. 2023. Guidance in radiology report summarization: An empirical evaluation and error analysis. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 176–195, Prague, Czechia. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, Toronto, Canada. Association for Computational Linguistics.

Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. 2024. Cross-modal projection in multimodal LLMs doesn‘t really project visual attributes to textual space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 657–664, Bangkok, Thailand. Association for Computational Linguistics.

A. Wallis and P. McCoubrie. 2011. The radiology report–are we getting the message across? *Clin Radiol*, 66(11):1015–1022.

Tongnian Wang, Xingmeng Zhao, and Anthony Rios. 2023. UTSA-NLP at RadSum23: Multi-modal retrieval-based chest X-ray report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 557–566, Toronto, Canada. Association for Computational Linguistics.

Qianqian Xie, Jiayu Zhou, Yifan Peng, and Fei Wang. 2023. Factreranker: Fact-guided reranker for faithful radiology report summarization. *Preprint*, arXiv:2303.08335.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and “discharge me!”. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng

Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

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.

A The Exclusive Set Construction

The exclusive set is constructed by obtaining three sets of diagnosis labels from FINDINGS, IMPRESSION and the radiology images using the pretrained ChexBert (Smit et al., 2020b) and TorchXRayVision (Cohen et al., 2022). If a diagnosis label agrees in three sets of labels (i.e., all positive or all negative), we remove the sentence that is used to justify the presence of absence of a diagnosis from the FINDINGS section. We visualize the overview of the construction process in Figure 5.

B Medical Expert Annotation Study

B.1 Annotation Guideline

> Question: Can you write the IMPRESSION section with FINDINGS section only, without using radiology image? (Y / N)
>
> - For this question, you need to read FINDINGS (column B) and Impression (column C), and decide if the IMPRESSION section can be written based on the provided FINDINGS section only, without any other information like radiology images and historical data.
> - Answer Y (Yes) if the FINDINGS section contains enough information to write the provided IMPRESSION section.
> - Answer N (No) if other information is required to write the provided IMPRESSION section, and is missing from the FINDINGS section.

B.2 Annotation Interface

We use Microsoft Excel for the annotation study. We include a screenshot of the annotation interface of our task in Figure 6

C Prompt Used For Summarization

For training and inference, we use the following prompt for LLaVA-1.5, Qwen-VL, DeepSeek-VL, and LLaVA-Med:

> You are an expert multimodal medical language model specializing in radiology. Your task is to analyze both the Findings section of a radiology report and the corresponding radiograph images to generate a concise, clear, and accurate Impression section. The Impression should: 1. Integrate information from the written report and visual findings in the radiographs. 2. Highlight the most critical findings while ensuring consistency between the text and image analysis. 3. Be written in a formal tone suitable for medical professionals. 4. Avoid unnecessary repetition or extraneous details.
>
> Here is the radiology image: <image>
>
> Here is the Findings section: <Findings>

D Experimental Details

All experiments were run on a single A100 40GB GPU. We use LVLMs weights from Huggingface, including LLaVA-1.5 (`llava-1.5-7b-hf`), LLaVA-Med (`llava-med-v1.5-mistral-7b`), Qwen-VL (`Qwen2-VL-7B`) and DeepSeek-VL (`deepseek-vl-7b-chat`). All LVLMs are fine-tuned with LoRA framework with 1 epoch and batch size 4 on all three datasets due to computational constraint. For all other models, we train 25 epochs for MIMIC-CXR, 20 epochs for CheXpert, and 10 epochs for OpenI, with batch size 4 and early stopping based on ROUGE-2 score. Unless otherwise specified, we followed the default hyperparameters and settings reported in the respective original papers.

E Baseline Model Results

In Table 6, we report ROUGE-1, ROUGE-L, Radgraph, and ChexBert for all baseline models, with and without image input (black image).

F Using Image Only As Input

We report results inference using image only as input in Table 7.

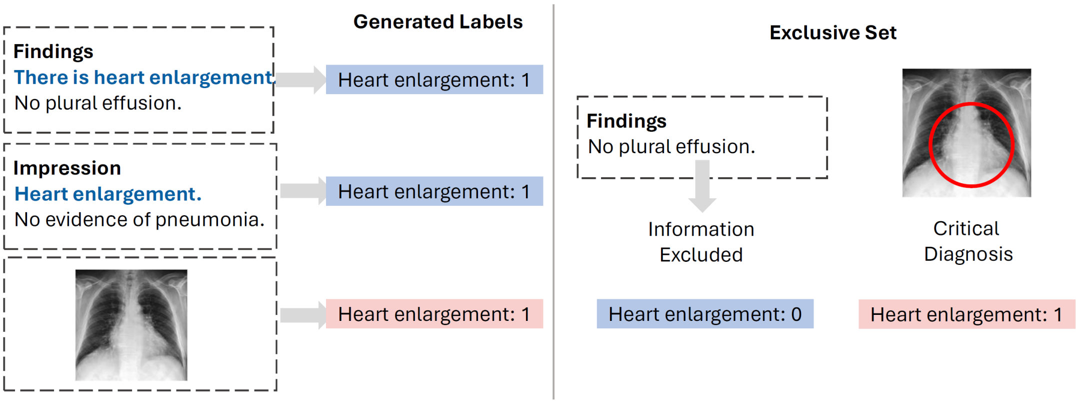

Figure 5: Overview of the dataset preprocessing and augmentation pipeline. Labels for FINDINGS, IMPRESSION, and radiology images are obtained from ChexBert (Smit et al., 2020b) and TorchXRayVision (Cohen et al., 2022). For diagnosis that aligns across all three label sources (e.g., "Heart Enlargement" is positive in three sets of labels in this example), the target sentence is augmented (negation for conflicting set, removal for exclusive set).

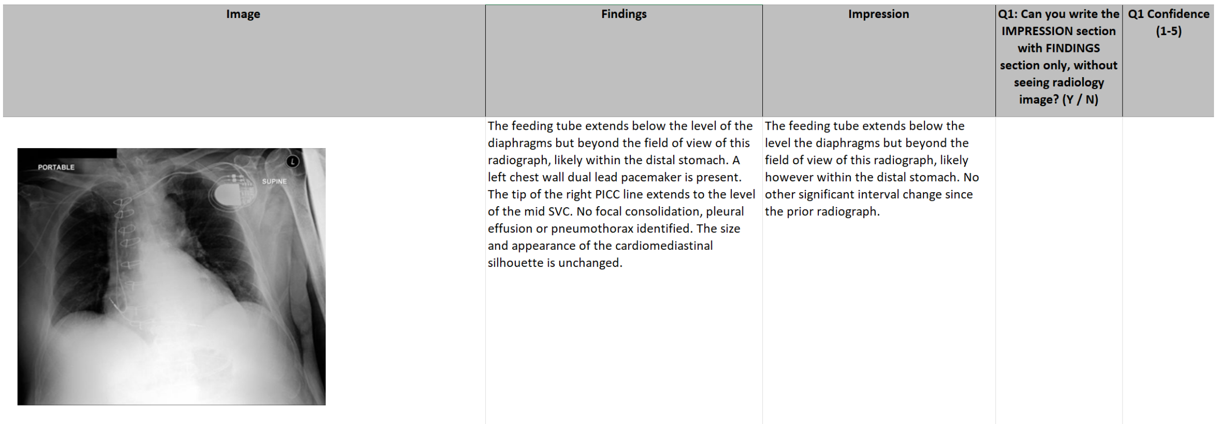

Image	Findings	Impression	Q1: Can you write the IMPRESSION section with FINDINGS section only, without seeing radiology image? (Y / N)	Q1 Confidence (1-5)
	The feeding tube extends below the level of the diaphragms but beyond the field of view of this radiograph, likely within the distal stomach. A left chest wall dual lead pacemaker is present. The tip of the right PICC line extends to the level of the mid SVC. No focal consolidation, pleural effusion or pneumothorax identified. The size and appearance of the cardiomediastinal silhouette is unchanged.	The feeding tube extends below the level the diaphragms but beyond the field of view of this radiograph, likely however within the distal stomach. No other significant interval change since the prior radiograph.		

Figure 6: An example of the annotation interface.

G Medoid Chest X-Ray As Input

Other than using a black image as input to test the utilization of the image in multimodal summarization models, we also report results using medoid chest x-ray image as input in Table 8.

H Two-Stage Fine-tuning Results

In Table 9, we report ROUGE-L, Radgraph and ChexBert score for LVLMs trained with different training strategies, inference with and without image (black image).

I LLM as a judge

In Table 10, we report evaluation results on MIMIC-CXR and OpenI using LLM-as-a-judge. Specifically, we use the ChexPrompt framework (Zambrano Chaves et al., 2025), which uses GPT-4 as a backbone model to automatically evaluate generated reports and categorize six types of errors, separately for clinically significant and insignificant cases.

We find that the text-only model BART generates the fewest errors in most categories, consistent with n-gram overlapping (ROUGE) and radiology factuality metrics (F1-RadGraph and F1-ChexBert). For instance, on MIMIC-CXR, BART has the fewest false positive findings and omission of findings compared to other LVLMs, but it has more insignificant errors compared to other LVLMs.

J Case Studies

Based on the expert annotation study, we categorize studies into three categories in Table 11: i) average case, the FINDINGS section contains sufficient information to construct the IMPRESSION section; ii) FINDINGS section shorter than IMPRESSION section, and iii) FINDINGS section refers to previous studies.

	MIMIC-CXR				OpenI				CheXpert			
	R-1	R-L	RadG	CBert	R-1	R-L	RadG	CBert	R-1	R-L	RadG	CBert
Text-Only Models												
PG	40.14	36.84	26.72	54.44	57.80	57.27	48.76	84.53	42.70	37.49	14.39	42.00
BART	**50.86**	**47.00**	40.90	65.39	**68.51**	**67.73**	**63.35**	87.29	**57.46**	**51.94**	19.84	56.85
GSum	45.32	42.98	35.18	59.60	61.76	60.10	52.32	85.28	51.92	47.09	15.22	49.01
WGSum	43.92	41.55	31.17	58.92	60.52	58.67	50.36	84.73	44.64	41.02	14.89	45.00
Multimodal Models												
VG-BART Dot	50.69	46.68	40.92	**65.59**	51.38	51.21	42.24	84.84	55.07	49.84	18.63	56.45
VG-BART MHA	50.64	46.89	**41.10**	65.33	56.30	56.14	47.08	84.23	54.63	49.02	18.70	55.60
CvT-BERT	43.85	44.32	37.76	58.52	60.12	57.91	50.16	84.10	49.01	44.56	18.90	51.00
Vilmedic	35.58	34.13	28.13	54.13	63.95	63.88	60.25	84.53	50.94	46.15	18.63	52.70
LLaVA-1.5	47.32	44.32	38.81	62.14	50.78	50.13	43.96	87.44	51.22	45.26	37.63	53.90
Qwen-VL	47.62	44.30	38.83	63.08	49.67	49.52	43.41	86.62	51.54	45.45	**38.76**	53.81
DeepSeek-VL	46.70	43.12	37.79	62.95	49.51	49.08	42.50	86.83	49.19	42.86	35.93	54.90
LLaVA-Med	47.58	44.53	38.81	62.98	50.91	50.93	44.18	88.41	51.64	46.01	38.61	**57.27**
Multimodal Models (Mask Image)												
VG-BART Dot	50.67	46.59	40.92	65.58	51.38	51.21	42.25	84.84	55.15	49.85	18.65	56.50
VG-BART MHA	50.56	46.82	41.06	65.33	56.22	56.13	47.17	84.38	54.63	49.03	18.71	55.60
CvT-BERT	43.84	44.35	37.41	59.01	59.80	56.97	50.23	83.98	49.04	44.68	18.95	51.03
Vilmedic	35.66	34.18	27.63	54.08	63.87	63.72	60.33	84.15	50.94	46.14	18.63	52.70
LLaVA-1.5	47.31	44.08	38.95	62.95	50.78	50.13	43.96	87.44	50.91	44.93	36.18	54.50
Qwen-VL	47.69	44.38	39.04	63.08	46.68	49.55	43.41	86.62	50.97	44.71	35.93	54.90
DeepSeek-VL	46.78	43.13	37.89	63.17	48.95	48.50	42.10	86.68	47.14	40.94	26.99	51.90
LLaVA-Med	47.59	44.52	38.89	62.91	50.91	50.94	44.20	**88.45**	51.60	46.20	38.57	57.00

Table 6: The performance of all text-only and multimodal baseline methods on the test sets of MIMIC-CXR, OpenI, and CheXpert datasets. F1 score is reported for all evaluation metrics. The best score and upper bound for all metrics are 100. RadG refers to RadGraph, CBert refers to ChexBert.

K Dataset License

MIMIC-CXR is under the PhysioNet Credentialed Health Data License 1.5.0. OpenI is publicly available and no license terms are stated. CheXpert is also publicly available under the Stanford University Dataset Research Use Agreement.

		MIMIC-CXR			OpenI			CheXpert		
Model	Setting	R-L	RadG	CBert	R-L	RadG	CBert	R-L	RadG	CBert
LLaVA	Proj	13.43	7.67	44.62	4.27	2.38	25.50	2.71	1.47	57.58
	LM	17.29	14.22	44.74	16.11	5.97	17.00	47.78	41.88	84.53
	Proj + LM	17.30	14.22	44.74	14.90	5.77	14.75	47.25	39.29	84.53
	Proj → LM	17.24	14.10	44.74	14.07	4.68	11.80	47.83	41.97	84.53
Deepseek-VL	Proj	17.44	14.27	44.62	8.72	8.22	20.75	3.37	8.60	32.00
	LM	17.99	14.76	44.93	25.51	9.92	18.35	47.90	42.10	84.53
	Proj + LM	18.06	14.82	44.62	25.08	9.72	18.40	47.91	42.10	84.53
	Proj → LM	18.05	14.92	44.91	17.82	6.62	22.40	47.91	42.10	84.53
Qwen-VL	Proj	17.32	14.00	44.68	17.41	6.38	21.25	4.48	1.46	71.82
	LM	16.89	13.25	44.81	17.67	5.50	14.80	47.78	42.00	84.53
	Proj + LM	16.16	12.34	44.74	17.93	6.69	17.80	43.44	31.79	84.38
	Proj → LM	15.00	9.83	44.74	15.90	5.67	11.35	47.92	42.13	84.53

Table 7: The performance of LVLMs on test sets of MIMIC-CXR, OpenI, and CheXpert datasets, using image only as input. LVLMs are tested on FINDINGS and X-Ray image pairs. F1 score is reported for all evaluation metrics. The best score and upper bound for all metrics are 100. RadG refers to RadGraph, CBert refers to ChexBert.

	MIMIC-CXR				OpenI				CheXpert			
	R-1	R-L	RadG	CBert	R-1	R-L	RadG	CBert	R-1	R-L	RadG	CBert
VG-BART Dot	50.69	46.68	40.92	65.59	51.38	51.21	42.24	84.84	55.07	49.84	18.63	56.45
VG-BART MHA	50.63	46.85	41.11	65.33	56.31	56.15	47.10	84.30	54.59	49.11	18.70	55.60
CvT-BERT	43.85	44.32	37.76	58.52	60.08	57.87	50.12	84.04	49.05	44.61	19.01	51.03
Vilmedic	35.56	34.11	28.13	54.13	63.87	63.81	60.23	84.49	50.98	46.17	18.70	52.81
LLaVA-1.5	47.30	44.29	38.81	62.14	50.78	50.13	43.9	87.44	51.23	45.28	37.65	53.89
Qwen-VL	47.61	44.32	38.81	63.00	49.66	49.50	43.45	86.58	51.55	45.48	38.80	53.83
DeepSeek-VL	46.71	43.13	37.79	62.95	49.51	49.08	42.50	86.83	49.19	42.86	35.93	54.90
LLaVA-Med	47.60	44.55	38.85	63.01	50.93	50.94	44.21	88.45	51.62	46.00	38.60	57.28

Table 8: The performance of LVLMs using medoid image and FINDINGS section as input on test sets of MIMIC-CXR, OpenI, and CheXpert datasets.

		MIMIC-CXR			OpenI			CheXpert		
Model	Setting	R-L	RadG	CBert	R-L	RadG	CBert	R-L	RadG	CBert
LLaVA-1.5	Proj	36.09	30.62	55.44	6.72	6.79	78.10	35.35	24.18	48.10
	LM	**44.32**	**38.80**	**62.14**	49.98	43.96	87.44	45.26	37.63	53.90
	Proj + LM	44.27	38.38	61.76	**59.13**	**53.53**	86.68	**45.46**	38.10	**54.10**
	Proj → LM	44.06	38.34	61.76	50.33	44.05	87.90	45.42	37.79	53.20
	Proj → Proj + LM	43.83	37.98	61.76	50.75	44.29	**88.36**	45.35	**38.45**	53.45
Qwen-VL	Proj	36.76	31.44	56.45	6.69	8.33	85.45	40.25	29.13	50.50
	LM	44.30	38.83	63.08	52.66	45.78	87.14	45.45	38.41	54.25
	Proj + LM	**44.41**	**39.26**	63.52	**62.90**	**58.24**	87.60	45.43	38.50	53.95
	Proj → LM	44.10	38.76	**63.58**	52.33	45.75	87.44	45.31	**38.57**	54.65
	Proj → Proj + LM	44.10	38.62	63.14	52.11	45.61	**87.90**	**45.59**	38.30	**54.85**
DeepSeek-VL	Proj	35.47	30.59	56.20	13.41	8.06	84.53	31.20	18.30	48.20
	LM	**43.12**	37.79	62.95	49.01	42.50	**86.83**	42.86	34.11	50.25
	Proj + LM	42.80	37.39	62.39	29.80	23.95	52.57	**43.03**	34.56	50.20
	Proj → LM	42.72	37.80	63.45	48.81	42.54	86.06	42.67	34.15	**52.40**
	Proj → Proj + LM	42.80	**37.87**	**63.70**	**49.47**	**42.68**	**86.83**	42.94	**35.10**	52.15
					Mask Image					
LLaVA-1.5	Proj	15.67	16.23	44.43	6.72	6.79	78.10	19.55	12.64	48.35
	LM	44.12	38.95	62.95	49.98	43.96	87.44	44.88	36.18	54.50
	Proj + LM	44.17	38.37	62.75	59.10	53.80	86.98	45.21	38.40	54.10
	Proj → LM	33.87	29.58	54.88	50.30	43.92	89.13	33.12	15.98	50.60
	Proj → Proj + LM	36.05	31.70	57.76	50.41	44.21	88.97	32.39	15.73	50.20
Qwen-VL	Proj	13.07	16.52	44.68	5.99	8.00	82.54	17.29	13.69	48.15
	LM	44.42	39.04	63.08	51.56	45.04	87.29	44.70	35.93	44.69
	Proj + LM	44.54	39.29	62.77	62.41	57.84	86.98	44.82	37.95	44.72
	Proj → LM	27.18	25.33	51.69	51.87	45.09	88.06	22.77	15.53	45.85
	Proj → Proj + LM	28.34	26.31	52.00	52.07	45.11	87.75	23.23	15.36	46.80
DeepSeek-VL	Proj	20.23	16.30	43.80	12.68	7.39	85.30	24.99	12.19	47.85
	LM	42.85	37.36	62.70	48.46	42.10	86.68	40.93	26.99	51.90
	Proj + LM	43.11	37.66	62.58	30.14	24.13	52.32	40.79	27.24	51.05
	Proj → LM	37.01	32.80	54.13	47.86	41.44	86.06	32.97	15.34	50.55
	Proj → Proj + LM	36.74	31.99	53.94	48.63	41.69	86.37	32.71	15.31	51.20

Table 9: The performance of LVLMs on the test sets of MIMIC-CXR, OpenI and CheXpert datasets, fine-tuned with different strategies. F1 score is reported for all evaluation metrics. The best score and upper bound for all metrics are 100. RadG refers to RadGraph, CBert refers to ChexBert.

		Number of Clinically Significant Error					
MIMIC-CXR		FP. Findings	Omission Findings	Inc. Location	Inc. Severity	FP Comparison	Omission Comparison
	BART	414	637	16	157	2	244
	LLaVA	443	755	11	118	1	253
	QWen	449	702	13	134	1	266
	Deepseek	481	663	17	132	2	256
		Number of Clinically Insignificant Error					
		FP. Findings	Omission Findings	Inc. Location	Inc. Severity	FP. Comparison	Omission Comparison
	BART	41	1	3	1	10	0
	LLaVA	37	1	3	2	6	0
	QWen	27	0	2	4	2	0
	Deepseek	36	1	1	2	6	0
		Number of Clinically Significant Error					
OpenI		FP. Findings	Omission Findings	Inc. Location	Inc. Severity	FP. Comparison	Omission Comparison
	BART	138	149	6	15	0	15
	LLaVA	133	196	2	7	0	14
	QWen	139	207	5	16	0	15
	Deepseek	128	206	0	8	0	12
		Number of Clinically Insignificant Error					
		FP. Findings	Omission Findings	Inc. Location	Inc. Severity	FP. Comparison	Omission Comparison
	BART	6	0	1	0	1	0
	LLaVA	15	0	0	1	0	0
	QWen	16	0	2	1	2	0
	Deepseek	15	0	0	1	3	0

Table 10: Evaluation results on MIMIC-CXR and OpenI using ChexPrompt. We report 6 types of clinically significant and insignificant errors as designed in ChexPrompt. We observe that the text-only model, BART, generates the fewest errors in most categories, consistent with n-gram overlapping (ROUGE) and radiology factuality metrics (F1-RadGraph and F1-ChexBert). * FP. refers to False Positive, Inc. refers to Incorrect.

a. Average case: FINDINGS *section itself is sufficient to construct the* IMPRESSION *section*
Findings: Single portable chest radiograph dated 12-30-2011 at 0303 demonstrates midline appearance of the trachea. The lung volumes are decreased, and there is mild widening of the superior mediastinum likely secondary to position and technique. The lungs are otherwise clear. No focal pleural or bony abnormalities are identified. Exam is limited by overlying trauma backboard.
Impression: 1. no radiographic evidence of acute cardiopulmonary disease. 2. mild widening of the superior mediastinum is likely secondary to technique and projection. 3. no acute fracture or pneumothorax identified.
b. Short FINDINGS *section*
Findings: There is a persistent right-sided pleural effusion
Impression: Right sided pleural effusion persistent cardiomegaly and interstitial edema
c. Report referring to previous studies
Findings: A right subclavian venous catheter is seen with tip in the mid superior vena cava. Low lung volumes are appreciated. Vasculature and cardiomediastinal silhouette is within normal limits.
Impression: 1. low lung volumes. cardiomediastinal silhouette and pulmonary vasculature are within normal limits. 57793358474 single portable view of the chest: 1-19-2007 findings: interval increase in pulmonary edema is noted. impression: 1. interval increase in pulmonary edema.

Table 11: Based on the medical expert annotation, we split radiology reports into three categories: i) Average Case, where FINDINGS section itself contains sufficient information to construct the IMPRESSION section; ii) Short FINDINGS section, that cannot infer the IMPRESSION; and iii) Report referring to previous studies, require access to previous studies to construct IMPRESSION section.