The Impact of Auxiliary Patient Data on Automated Chest X-Ray Report Generation and How to Incorporate It

Anonymous ACL submission

Abstract

This study investigates the integration of diverse patient data sources into multimodal language models for automated chest X-ray (CXR) report generation. Traditionally, CXR report generation relies solely on CXR images and limited radiology data, overlooking valuable information from patient health records, particularly from emergency departments. Utilising the MIMIC-CXR and MIMIC-IV-ED datasets, we incorporate detailed patient information such as aperiodic vital signs, medications, and clinical history to enhance diagnostic accuracy. We introduce a novel approach to transform these heterogeneous data sources into embeddings that prompt a multimodal language model, significantly enhancing the diagnostic accuracy of generated radiology reports. Our comprehensive evaluation demonstrates the benefits of using a broader set of patient data, underscoring the potential for enhanced diagnostic capabilities and better patient outcomes through the integration of multimodal data in CXR report generation.

1 Introduction

Chest X-ray (CXR) exams, which consist of multiple images captured during an imaging session, are essential for diagnosing and managing a wide range of conditions, playing a significant role in patient care. Radiologists interpret these exams and produce a written report with their findings. However, prompt reporting is hindered by a multitude of issues, including high patient volumes and limited availability of radiologists (Bailey et al., 2022).

Machine learning for automated CXR report generation is a promising solution that has garnered significant attention in the literature (Jones et al., 2021). By leveraging multimodal language models, exams can be rapidly interpreted and reported, potentially providing quick and reliable diagnostic insights crucial for decision-making, such as triaging patients. Models are often trained to generate

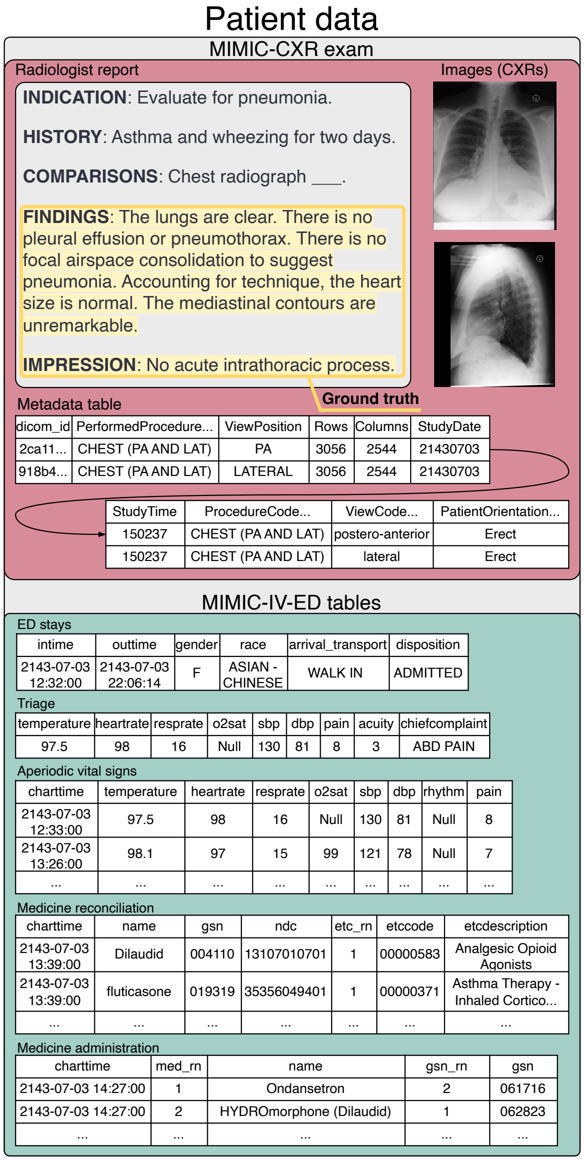


Figure 1: The patient data from MIMIC-IV-ED associated with a CXR exam from MIMIC-CXR. This includes the exam's images, the corresponding radiology report, and the associated image metadata. The findings and impression sections of the radiology report form the ground truth for CXR report generation. ED-specific data, such as medicine reconciliation and aperiodic vital signs, is also available for the patient.

the *findings* and *impression* sections of a radiology report (Figure 1), where the former details the interpretation of a patient's exam and the latter summarises the most important findings. Potential benefits include enhanced radiologist effectiveness, a reduced workload, alleviation of the burden of report writing, and improved patient outcomes (Shen, 2021; Irmici et al., 2023).

Early methods for CXR report generation produced a separate report for each image within an exam (Wang et al., 2018). Later methods improved on this by considering all images of an exam to generate a single report (Miura et al., 2021; Nicolson et al., 2024a), and incorporating prior exams for a patient (Wu et al., 2022; Nicolson et al., 2024a). Additionally, including the reason for conducting the exam (the *indication* section in Figure 1) offered a further improvement (Nguyen et al., 2023). This indicates that CXR report generation could benefit from the inclusion of a more comprehensive set of patient data.

Modern patient record systems are another rich source of patient data, containing detailed information that may be valuable for CXR report generation. However, (1) the utility of this data has not been empirically investigated, and (2) it is unclear how to harmonise this heterogeneous data into a unified multimodal language model. This paper aims to address these two points. To achieve this, we combine CXR exams from MIMIC-CXR (Johnson et al., 2019) with emergency department (ED) patient records from MIMIC-IV-ED (Johnson et al., 2023). This means that for a single exam, a wide variety of multimodal data is available, as shown in Figure 1. From MIMIC-CXR, we utilise the images, their metadata, and several sections of the radiology report. Notably, incorporating the comparison or history section is a novel approach in the literature. From MIMIC-IV-ED, we investigate triage information, aperiodic vital signs, medications, and other data to provide a wider clinical context.

We explore combining these sources of patient data as patient embeddings to prompt a multimodal language model. We demonstrate that complementary information from different data sources can improve the diagnostic accuracy of CXR report generation. To achieve this, we develop methods to transform tabular and aperiodic time series data into embeddings that can be used alongside token and image embeddings. We evaluate our model on MIMIC-CXR exams with accompanying patient data from MIMIC-IV-ED, using metrics shown to closely correlate with radiologists' assessments of reporting (Yu et al., 2023). The main contributions of this work are:

- An investigation into how patient data impacts CXR report generation, focusing on the effects of specific data sources, such as medications and vital signs.
- An empirical evaluation demonstrating that using multiple patient data sources — from a patient's CXR exams and their ED record — significantly improves diagnostic accuracy.
- Introducing methods to convert multimodal patient data into embeddings for a language model, including numerical, categorical, free text, temporal, and image data.
- A release of dataset splits based on MIMIC-CXR and MIMIC-IV-ED, linking patient exams with their associated ED records (available as a Hugging Face dataset). This, along with our code repository and Hugging Face checkpoint can be found at: `https://anonymous.4open.science/r/anon-D83E`, enabling others to experiment with new methods for multimodal patient data.

2 Background and Related Work

There is evidence to suggest that incorporating more patient data improves diagnostic accuracy in radiology reporting. Initial improvements came from using multiple images per exam, like EMNLI, which often includes complementary frontal and lateral views of the patient (Miura et al., 2021; Gaber et al., 2005). Methods such as CXRMate enhance diagnostic accuracy by incorporating a patient's prior exams to identify changes over time (Nicolson et al., 2024a; Wu et al., 2022; Kelly, 2012). Including the *indication* section of the radiology report to provide clinical context also provides an improvement (Nguyen et al., 2023). This trend indicates that providing more comprehensive patient data improves diagnostic accuracy, which we investigate in this work.

ED records contain a myriad of data, including vital signs such as respiratory rate, temperature, and blood pressure, which can aid in the identification of various diseases. A high respiratory rate and low blood oxygen saturation are indicative of conditions that compromise pulmonary function, such as pulmonary embolism. Similarly, an elevated body temperature is suggestive of an infectious process, such as pneumonia or tuberculosis. Incorporating such data into a CXR report generator could help corroborate subtle radiographic signs typical of these infections. Our findings demonstrate that patient data from the ED can indeed enhance CXR report generation.

Recent advancements in integrating multimodal patient data have enhanced diagnostic and predictive healthcare capabilities. A study showed that a Transformer encoder combining imaging and non-imaging data outperformed single-modality models, diagnosing up to 25 conditions with higher AUC scores (Khader et al., 2023b). Similarly, the MeTra architecture, which integrates CXRs and clinical parameters, demonstrated superior performance in predicting ICU patient survival compared to using either CXRs or clinical data alone (Khader et al., 2023a). ETHOS, using a zero-shot learning approach, outperformed single-modality models in predicting inpatient mortality, ICU length of stay, and readmission rates (Renc et al., 2024). These studies highlight the importance of multimodal data for improved healthcare analytics. Our work demonstrates that incorporating a comprehensive set of multimodal patient data enhances CXR report generation.

Recent advancements in multi-task learning have significantly improved biomedical models by leveraging shared knowledge. Med-PaLM M, a generalist biomedical model, excels in multiple tasks including classification, question answering, visual question answering (VQA), report summarisation, report generation, and genomic variant calling, using diverse input modalities like images, text, and genomics. It often outperforms specialised models, demonstrating superior performance and generalisation (Tu et al., 2024).

Similarly, MIMIC-CXR has been leveraged for multi-task learning with models like MedXChat, which integrates instruction-tuning and Stable Diffusion to perform CXR report generation, VQA, and report-to-CXR generation, outperforming other LLM multi-task learners (Yang et al., 2023). RaDialog, another LLM-based method, combines visual features and pathology findings to generate accurate radiology reports and support interactive tasks, significantly improving clinical efficacy. CXR-LLaVA, a multimodal LLM integrating a vision transformer with a language model, outperformed models like GPT-4 Vision and Gemini Pro Vision in CXR report generation (Lee et al., 2024).

Determining the state-of-the-art CXR report generation model can be challenging due to the unavailability of some models and the lack of comparison to recent methods. The 2024 Shared Task on Large-Scale Radiology Report Generation (RRG24) aimed to address this by benchmarking models on a common leaderboard. The winning model, CXRMate-RRG24 (Nicolson et al., 2024b), a derivative of CXRMate, emerged as a strong contender for state-of-the-art. In this work, we compare our model to established models (e.g., EMNLI) and recent benchmarks (e.g., CXRMate-RRG24, CXRMate, CXR-LLaVA, MedXChat, and RaDialog). We ensure a fair comparison by using available code or obtaining generated reports directly from the authors. Our findings indicate our model produces significantly better results than these models.

3 Dataset

We construct a dataset of 46 106 patients by linking individual patient information from two separate sources: (1) CXR exams from MIMIC-CXR and (2) emergency records from MIMIC-IV-ED. Thus we consider MIMIC-CXR exams that occurred during an ED stay from MIMIC-IV-ED. Both datasets are publicly available and originate from the Beth Israel Deaconess Medical Center in Boston, MA.

MIMIC-CXR was formed by first extracting patient identifiers for exams performed in the ED between 2011–2016, and then extracting all exams for this set of patients from all departments between 2011–2016. Each exam includes a semi-structured free-text radiology report (Figure 1) that describes the radiological findings of the images, written by a practising radiologist contemporaneously during routine clinical care. All images and reports were de-identified to protect privacy. Sections from the radiologist reports were extracted using a modification[1] of the official text extraction tool[2] in order to obtain the findings, impression, indication, history, and comparison sections.

MIMIC-IV-ED consists of de-identified data from ED stays between 2011–2019. The data was converted into a denormalised relational database with six primary tables: ED stays, diagnosis, medicine reconciliation, medicine administration, triage, and aperiodic vital signs. We do not consider the diagnosis table in this work, as it indicates the outcome of a patient's ED stay. The patients of

[1]https://anonymous.4open.science/r/anon-D83E

[2]https://github.com/MIT-LCP/mimic-cxr/tree/master/txt

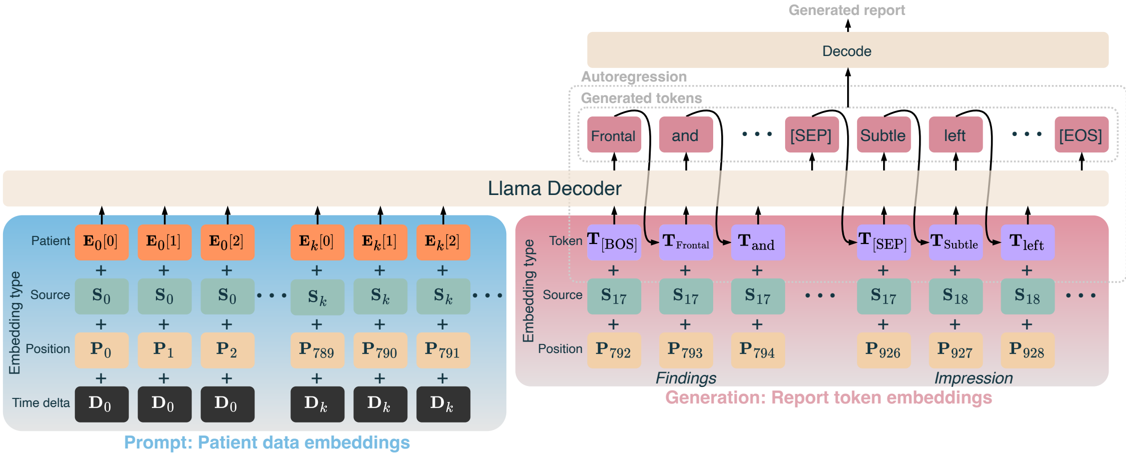


Figure 2: Multimodal language model for CXR report generation. The patient data embeddings prompt the decoder to generate the findings and impression sections of a radiology report.

MIMIC-CXR can be linked to MIMIC-IV-ED via an identifier, allowing an ED specific dataset to be formed.

Example tables for a patient's exam are shown in Figure 1. The dataset was formed by extracting a patient's exams whose times (formed by the 'StudyDate' and 'StudyTime' columns of the metadata table) occurred within the 'intime' and 'outtime' of one of their ED stays.[3] Exams with either a missing findings or impression section were not considered. Using the official splits of MIMIC-CXR, this gave a train/validation/test split of 45 527/343/236 patients, 76 398/556/958 exams, and 151 818/1 137/1 812 CXRs. Each of these exams had one ED stay and triage row; 53% had at least one medicine reconciliation row with up to 106 rows; 62% had at least one vital signs row with up to 69 rows; and 37% had at least one medication administration row with up to 52 rows. Exams had an indication section 66% of the time with a maximum of 75 words, a history section 34% of the time with a maximum of 74 words, and a comparison section 97% of the time with a maximum of 129 words. Only one exam had both an indication and a history section.

4 Methods

The patient data from MIMIC-CXR and MIMIC-IV-ED for an exam are transformed into embeddings, which are used to prompt a multimodal language model to generate the findings and impression sections of the radiology report, as illustrated in Figure 2. Additionally, 'Source' embeddings differentiate the source of the data (e.g., the 'chief complaint' column from the triage table, the indication section, etc.), and time delta embeddings represent the time difference between an event and the exam. Standard embeddings, such as position and token embeddings, are also included. The patient data embeddings originate from three main groups: the tables of MIMIC-IV-ED; the report, images, and metadata of the current exam from MIMIC-CXR; and the patient's prior exams (also originating from MIMIC-CXR). The prior exam and image embeddings are described in Section A and Subsection C.2, respectively.

4.1 Time, Position, & Source Embeddings

The ED information from MIMIC-IV-ED is typically recorded as discrete events, such as medications administered or vital signs measured, each with a specific timestamp. Events that occur closer to the time of the patient's exam are generally more relevant for diagnostic purposes. To capture this, a time delta is calculated by subtracting the time of an event from the time of the exam. The exam time originates from MIMIC-CXR's metadata table (Figure 3), whereas most of the MIMIC-IV-ED tables have event times for each row. As shown in Figure 3, the time delta is first converted to hours and then mapped using $1/\sqrt{\boldsymbol{\Delta}+1}$, assigning higher weights to events that occurred closer to the exam. The mapped time deltas are then passed through a feedforward neural network (FNN) defined as $f(\boldsymbol{\Delta W}_1)\boldsymbol{W}_2$, where $\boldsymbol{W}_1 \in \mathbb{R}^{1,2048}$, $\boldsymbol{W}_2 \in \mathbb{R}^{2048,H}$, $f(\cdot)$ is the sigmoid linear unit (SiLU) activation function (Hendrycks and Gimpel, 2016), and H is the hidden size of the decoder. This process generates the time delta embeddings, which are subsequently added to the embeddings of their respective sources. As shown in Figure 2, time delta embeddings are only applied

[3]Exam 59128861 was removed as it overlapped with two separate ED stays for the patient.

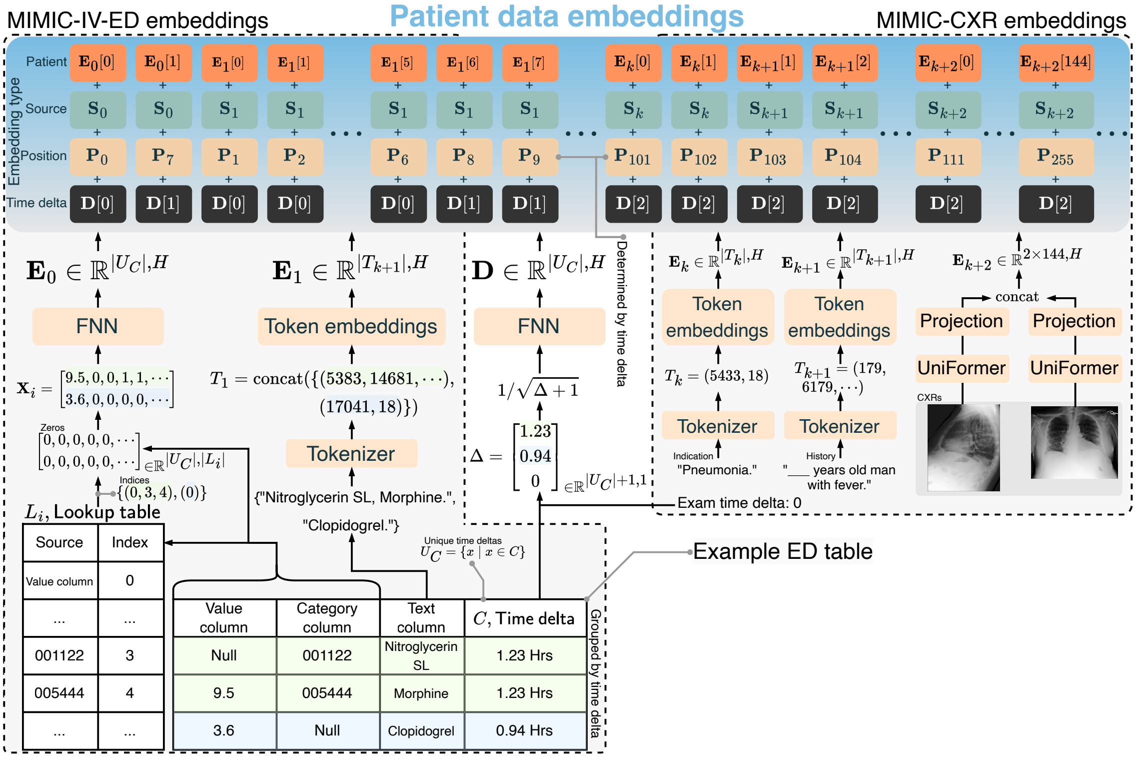


Figure 3: Proposed patient data embeddings from the multiple heterogeneous data types taken from MIMIC-IV-ED and MIMIC-CXR. The embeddings are formed from numerical, categorical, textual, temporal, and image data.

to the prompt. Patient data from the current exam, such as the images, have a time delta of zero, while data from prior exams have a positive time delta.

The position embeddings are ordered by the time delta (Figure 3). This is due to the rotary position embeddings of the decoder; tokens that are closer together are given more importance. Hence, the smaller the time delta, the closer the embedding's position is to the report token embeddings. Following Nicolson et al. (2024a), each unique patient data source is given its own source embedding. This includes the images, each report section, each table's text column and value-category columns (described in the next section), and prior images and report sections.

4.2 Tabular Data

An example table and its conversion to embeddings is shown in Figure 3. To convert an exam's tabular data to embeddings, columns were designated as value, category, text, or time columns. Value columns contained numeric data, while category columns contained categorical data. Datum from value and category columns were grouped by their time delta, with each group forming a feature vector. The feature vector initially consisted of zeros. Values and categories from the group were then used to set its values based on indices determined by a lookup table. For value columns, the lookup table determined the index where the numeric value was placed. For category columns, it determined which indices were activated (set to 1).

Next, the feature vector was passed through an FNN $f(\boldsymbol{X}_i\boldsymbol{W}_1)\boldsymbol{W}_2$ to form the embedding, where $\boldsymbol{X}_i \in \mathbb{R}^{|U_C|,|L_i|}$ are the grouped features, $\boldsymbol{W}_1 \in \mathbb{R}^{|L_i|,2048}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{2048,H}$, L_i is a lookup table, and i designates the table. Each table has a unique FNN and lookup table. Rows for a value column always had a unique time, preventing multiple values from the same column in a group. We investigated alternatives to form the value-category embeddings in Section 5. The described framework was found to be the most efficient. Columns with a high cardinality were set as text columns. Text embeddings were formed via the decoder's tokenizer and token embeddings. Text embeddings were given the time delta embedding from their respective row. The column designation for each table in Figure 1 is described in the Appendix B.

4.3 Report Section Embeddings

Here, we consider five sections of the radiology report: the findings, impression, indication, history, and comparison sections. The findings and impression sections serve as the ground truth to be generated. The remainder form part of the patient data embeddings. The indication section explains the reason for the exam, such as symptoms or suspected conditions. The history section provides relevant medical history, such as past conditions and treatments. The comparison section mentions any prior exams used to identify changes over time. These sections provide context that guides the interpretation of the exam, influencing the content of the findings and impression sections. The embeddings were formed via the decoder's tokenizer and token embeddings. Of these, the history and comparison sections have not been investigated for CXR report generation. The comparison section was used only when prior exams were considered.

4.4 Experiment Setup

Our multimodal language model, illustrated in Figure 2, is based on CXRMate-RRG24; it features a Llama decoder and the UniFormer as the image encoder. The training procedure for our model involved three stages: (1) initial training on the MIMIC-CXR training set using only images as input with Teacher Forcing (TF) (Williams and Zipser, 1989), (2) further training on the dataset described in Section 1 with the inputs detailed in Table 1, again using TF, and (3) reinforcement learning on the same dataset through self-critical sequence training (SCST) (Rennie et al., 2017) (only for Table 2). Our evaluation metrics included three that capture the semantics of radiology reporting — RadGraph-F1 (RG), CheXbert-F1 (CX), and CXR-BERT (CB) — as well as five natural language generation metrics: BERTScore-F1 (BS), CIDEr (C), METEOR (M), ROUGE-L (R-L), and BLEU-4 (B4). Comprehensive details on the model architecture, training procedure, significance testing, and comparison methods are provided in Appendix C.

5 Results & Discussion

The impact of different patient data sources on the performance of CXR report generation is summarised in Table 1. This analysis identifies which additional data sources enhance performance compared to using only images.

Significant improvements were observed by incorporating either the ED stays, triage, medicine reconciliation, or vital signs data from MIMIC-IV-ED dataset. Notably, the ED data markedly improved scores on the radiology report metrics (RG, CX, and CB). The medicine administration table did not significantly improve the scores overall, likely due to its infrequent occurrence in the exams (37%). (However, as shown in Table 4, it significantly improves performance when evaluated solely on exams that include a medicine administration table.) These findings demonstrate that ED patient data can enhance the diagnostic accuracy of CXR report generation.

Incorporating the indication or history section led to significant score improvements. This demonstrates the substantial influence these sections have on the findings and impression sections. Conversely, adding the metadata table did not result in significant score improvements, indicating it lacks valuable information for CXR report generation. While previous studies have established that the indication section boosts CXR report generation (Nguyen et al., 2023), our findings demonstrate that the history section is equally important.

When examining the impact of prior exams, we considered a maximum history size h of up to three, incorporating the findings and impression sections, and images from prior exams. Any history size significantly increases the scores compared to using solely the images, consistent with previous findings (Wu et al., 2022). However, performance gradually degrades as the history size increases, which contradicts earlier studies. Additionally, the comparison section appears to slightly degrade performance. We suspect this is due to the increasing number of inputs as h grows, combined with the limitations of our model architecture. $\overline{|\boldsymbol{\mathcal{E}}[:,0]|}$ in Table 1 is the average prompt length over the test set, where $\boldsymbol{\mathcal{E}} = [\mathbf{E}_0, \mathbf{E}_1, \cdots]$. It can be seen that $\overline{|\boldsymbol{\mathcal{E}}[:,0]|}$ increases substantially as h increases. Since we provide all inputs to the decoder's self-attention, a large input size may cause *attention dilution*. With more inputs, the attention weights must be distributed across a larger number of inputs, resulting in each input receiving a smaller share of the attention, making it harder for the model to focus on the most relevant inputs (Qin et al., 2022).

We then combined all the effective sources of patient data (those providing a significant improvement). This excluded 'medicine administration', 'metadata', and 'comparison'. The best performance was observed with no prior exams ($h = 0$), indicating that using any prior exams in combination with other sources is detrimental due to attention dilution. With $h = 0$, the combination of all effective sources outperformed each individual source. We then performed an ablation study

Table 1: Results of the various patient data sources on the test set described in Section 3. Results were calculated over ten training runs ($n = 9\,580$ exams; 958×10 runs). Underlined and Dashed underlined scores indicate a significant difference to the scores of 'Images' and 'Images + effective sources ($h = 0$)', respectively ($p < 0.05$). Evaluation is performed on both the **findings** and **impression** sections.

Patient data sources	RG	CX	CB	BS	C	M	R-L	B4	$\overline{\lvert\mathcal{E}[:,0]\rvert}$
Images only									
Images	26.00	29.24	58.87	24.10	12.24	14.35	24.34	6.33	272.4
Patient Emergency Department (ED) data (MIMIC-IV-ED)									
Images + ED stays	26.10	29.47	60.65	24.17	12.39	14.52	24.50	6.36	273.4
Images + triage	26.46	31.27	63.06	24.29	12.32	14.66	24.58	6.44	278.9
Images + vital signs	26.47	31.72	63.39	24.32	13.16	14.61	24.74	6.47	274.7
Images + medicine reconciliation	26.86	31.37	63.98	24.52	12.77	14.90	24.85	6.60	343.5
Images + medicine administration	26.15	29.47	59.21	24.25	12.30	14.44	24.47	6.38	273.0
Patient additional radiology data (MIMIC-CXR)									
Images + indication	26.94	32.13	65.43	24.74	14.16	15.19	25.16	7.02	279.5
Images + history	27.00	31.88	65.06	25.05	14.32	15.30	25.48	**7.33**	277.0
Images + metadata	26.34	29.63	59.55	24.37	12.40	14.55	24.50	6.43	273.4
Prior exams									
Images + $h = 1$	26.98	31.42	63.98	24.65	12.65	15.11	25.03	6.78	558.9
Images + $h = 1$ + comparison	26.76	31.55	64.20	24.42	13.36	15.03	24.82	6.74	563.4
Images + $h = 2$	26.67	30.48	61.27	24.53	13.60	14.94	24.85	6.72	810.6
Images + $h = 2$ + comparison	26.20	30.19	61.24	24.05	12.43	14.80	24.55	6.58	815.0
Images + $h = 3$	26.47	29.96	59.95	24.14	12.90	14.94	24.66	6.65	1037.1
Images + $h = 3$ + comparison	26.14	30.09	60.51	23.90	13.22	14.87	24.56	6.64	**1041.5**
All effective sources (no medicine administration, metadata, or comparison)									
Images + effective sources ($h = 0$)	27.11	32.23	64.80	25.07	14.48	15.15	25.40	7.07	365.0
Images + effective sources ($h = 1$)	26.78	31.83	63.85	24.75	14.10	15.15	25.25	7.01	651.7
Ablation from Images + effective sources ($h = 0$)									
- medicine reconciliation	26.78	**32.81**	**65.60**	24.84	14.44	15.21	25.33	7.19	293.9
- ED stays	26.94	31.56	64.87	25.02	14.08	15.14	25.37	7.09	364.0
- triage	27.15	32.45	65.18	25.15	**14.80**	15.27	**25.54**	7.25	358.5
- vital signs	**27.27**	31.78	65.44	25.14	14.07	**15.35**	25.49	7.22	362.6
- indication	26.89	31.25	64.65	24.99	13.87	15.07	25.39	7.00	357.9
- history	26.96	31.87	64.02	24.86	14.60	15.10	25.24	7.04	360.3
- time delta	27.17	32.11	65.10	**25.18**	14.64	15.24	**25.54**	7.16	365.0

using 'CXRs + effective sources ($h = 0$)'. Removing 'medicine reconciliation' significantly increased performance, specifically for CXR-BERT. This improvement was also likely due to attention dilution, as removing medicine reconciliation substantially decreased $\overline{|\mathcal{E}[:,0]|}$.

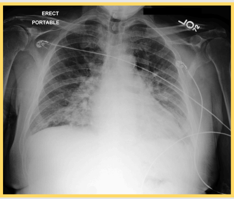
Next, we further trained 'Images + effective sources (h = 0) - medicine reconciliation' with reinforcement learning, as described in Subsection 4.4. This model, denoted as 'Ours' in Table 2, was compared to other benchmark CXR report generation models in the literature that included MIMIC-CXR in their training data. Despite having substantially fewer training samples than the other models, our model significantly outperformed them on CXR-BERT, BERTScore-F1, METEOR, ROUGE-L, and BLEU-4. This demonstrates the impact of incorporating a more comprehensive set of patient data on CXR report generation.

A case study is presented in Figure 4 demonstrating how a diverse set of patient data can impact report generation. Here, the first model is given the image only, and fails to identify key findings that the radiologist noted in their report. The second model is given the additional patient data available for this exam; the indication section and triage data. Hypoxia, as indicated by the low oxygen saturation ('o2sat'), along with the elevated respiratory rate ('resprate') and systolic blood pressure ('SBP'), are consistent with the physiological responses to pulmonary edema. Given this, the second model was able to identify the moderate pulmonary edema, echoing the radiologist's findings.

Table 3 compares different methods for converting value and category columns into embeddings. This evaluation includes images, the triage table, and the medicine reconciliation table, as these tables contain multiple value and category columns. The aforementioned method of producing embeddings by grouping data from value and category columns ('Grouped embeddings') is compared to two other methods. The first is separate embeddings for each datum, where each value column datum is separately transformed using the previously described FNN, while each category column datum is converted to an embedding using a learn-

Table 2: Benchmark models on the test set described in Section 3 ($n = 958$). Evaluation is on the **findings** section only. <u>Underlined</u> indicates statistical significance between the top two scores ($p < 0.05$). In the 'Train samples' column, 'images' means the model generates reports per image, while 'exams' means a report generated per exam.

Model	Train samples	RG	CX	CB	BS	C	M	R-L	B4
EMNLI (Miura et al., 2021)	152 173 exams	32.8	28.9	66.6	24.4	19.4	17.1	28.1	8.9
CMN (Chen et al., 2021)	270 790 images	25.3	24.3	49.4	19.7	16.9	15.1	26.4	7.6
TranSQ (Kong et al., 2022)	368 960 images	29.8	30.4	62.3	20.4	14.9	17.6	22.6	7.9
RGRG (Tanida et al., 2023)	166 512 images	23.2	22.8	37.9	23.4	7.6	12.4	21.1	5.4
CvT2DistilGPT2 (Nicolson et al., 2023)	270 790 images	25.8	29.3	59.8	24.8	20.9	16.0	27.3	8.8
RaDialog (Pellegrini et al., 2023)	276 778 images	26.8	**38.4**	60.7	26.2	14.6	14.7	25.4	6.9
MedXChat (Yang et al., 2023)	270 790 images	22.6	13.1	21.3	19.3	9.8	14.3	23.2	7.0
CXR-LLaVA-v2 (Lee et al., 2024)	193 513 images	20.7	20.7	44.1	23.6	5.2	11.3	19.9	2.7
CXRMate (Nicolson et al., 2024a)	125 395 exams	28.8	33.9	71.3	30.5	22.4	17.7	28.1	9.7
CXRMate-RRG24 (Nicolson et al., 2024b)	550 395 exams	30.4	31.2	58.2	31.0	20.6	16.7	27.5	9.1
Ours	76,398 exams	**33.7**	35.1	**<u>79.1</u>**	**<u>35.8</u>**	**24.1**	**<u>19.1</u>**	**<u>30.6</u>**	**<u>11.9</u>**



Case study

Indication: Hypoxia.
Radiologist findings: A portable frontal chest radiograph demonstrate an unchanged cardiomediastinal silhouette, which is top-normal in size. Bilateral opacities are consistent with moderate pulmonary edema. No definite focal consolidation or pneumothorax is identified. There are likely trace bilateral pleural effusions.
Radiologist impression: Moderate pulmonary edema.
Triage:

temperature	heartrate	resprate	o2sat	sbp	dbp	pain	acuity	chiefcomplaint
100.3	93	24	83	175	74	Null	1	ILI, Fever

Image *(Model: Images from Table 1)*
Generated findings: Cardiomediastinal silhouette is normal. There is no focal consolidation. There is no pneumothorax or pleural effusion. There is no significant pleural effusion.
Generated impression: No acute cardiopulmonary process.

Image + ***Indication*** + ***Triage*** *(Model: Images + effective sources (h=0) - medicine reconciliation from Table 1)*
Generated findings: There is moderate pulmonary edema. No definite focal consolidation is identified. There are probable small bilateral pleural effusions. The cardiac silhouette is mildly enlarged. There is no pneumothorax.
Generated impression: Moderate pulmonary edema and small bilateral pleural effusions.

Figure 4: Case study demonstrating how incorporating a diverse set of patient data can aid with report generation.

able weight matrix, akin to how token embeddings are produced ('Separate embeddings'). The second method modifies 'Separate embeddings' by instead converting the value column data to text and using the decoder's tokenizer and token embeddings ('Values-to-text, categories-to-tokens'). The results indicate that the grouped embeddings method generally works best and is useful for encoding heterogeneous patient data for multimodal models.

Table 3: Formatting strategies for the value-category columns. Four training runs were used ($n = 3\,832$; exams 958×4 runs). <u>Underlined</u> indicates a stat. sig. difference to 'Baseline' ($p < 0.05$).

Embeddings	CX	RG	CB	BS
Images				
Baseline	25.81	29.00	59.04	23.85
Images + triage + medicine reconciliation				
Grouped embeddings	**<u>26.72</u>**	**31.69**	**<u>64.01</u>**	24.38
Separate embeddings	25.32	25.28	<u>46.29</u>	23.51
Values-to-text, categories-to-embeddings	<u>26.46</u>	30.70	58.62	**<u>24.58</u>**

6 Conclusion

This paper demonstrates the value of incorporating diverse patient data into automated CXR report generation. By integrating patient data from the MIMIC-CXR and MIMIC-IV-ED datasets, we have shown significant improvements in the diagnostic accuracy of generated radiology reports. Our empirical evaluation uncovers new sources of patient information that enhance CXR report generation, including data from ED stays, triaging information, aperiodic vital signs, medications, and the history section of radiology reports. We present specific methods to convert multimodal patient data into embeddings for a language model, encompassing numerical, categorical, textual, temporal, and image data. We encourage further research and experimentation using our released dataset splits, code, and model checkpoints to explore innovative methods for multimodal patient data integration, with the ultimate goal of enhancing diagnostic accuracy and patient care.

7 Limitations

Despite the promising results demonstrated in this study, several limitations must be acknowledged. Firstly, the generalisability of our findings may be constrained by the datasets utilised, specifically MIMIC-CXR and MIMIC-IV-ED, which are derived from a single institution, the Beth Israel Deaconess Medical Center. This could introduce biases unique to the demographic and clinical practices of this institution, potentially limiting the applicability of our model to other healthcare settings with different patient populations or clinical workflows. Our reliance on these datasets is due to the fact that they are the only publicly available sources that link CXR exams with ED records.

Another limitation pertains to the completeness and quality of the patient data. Despite incorporating a wide range of data sources, the datasets still contain missing or incomplete information, which can affect model performance. For example, not all exams include a history section, and not all ED patient records have medicine administration details, leading to potential gaps in the data that the model can utilise. However, this reflects the nature of real patient records where issues of data quality and completeness are to be expected.

Our model's architecture, while effective, has certain limitations. It struggles with large input sizes, especially when incorporating multiple prior exams, likely due to attention dilution. Future work should explore advanced attention mechanisms or hierarchical models to better manage large input sequences.

The interpretability of the model also poses a challenge. While our model shows improved diagnostic accuracy, the decision-making process within the multimodal language model remains a black box. Developing methods to enhance the interpretability and explainability of the model's outputs would be beneficial, especially in clinical settings where understanding the rationale behind a diagnosis is critical.

Finally, while we provide a comprehensive set of metrics to evaluate our model's performance, these metrics focus primarily on the diagnostic accuracy and quality of the generated reports. Broader evaluations considering clinical outcomes, such as the impact on patient management or reduction in radiologist workload, would offer a more holistic view of the benefits and limitations of CXR report generation models in general. Conducting such assessments could help to better understand the practical implications of deploying these models in a clinical setting.

In summary, while our study provides valuable insights into the integration of multimodal patient data for CXR report generation, addressing these limitations will be crucial for further advancements and broader adoption of such models in clinical practice. Future research should explore alternative architectures and training strategies, find alternative datasets to evaluate generalisability, improve model interpretability, and comprehensively assess the practical impact on patient care and radiologist workflow.

8 Ethical Considerations

In this research, we used real-world patient data from the MIMIC-CXR and MIMIC-IV-ED datasets. Since these datasets are de-identified, we consider privacy leakage risks to be minimal. Our method employs a language model to generate medical reports from patient data. However, we acknowledge that language models can exhibit bias and produce hallucinations, which may result in incorrect content in the generated reports.

References

Christopher R. Bailey, Allison M. Bailey, Anna Sophia McKenney, and Clifford R. Weiss. 2022. Understanding and Appreciating Burnout in Radiologists. *RadioGraphics*, 42(5):E137–E139.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshops*, pages 65–72.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In *ECCV*, pages 1–21.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *IJCNLP*, pages 5904–5914.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In *EMNLP*, pages 4348–4360.

Mohamed Elgendi, Muhammad Umer Nasir, Qunfeng Tang, David Smith, John-Paul Grenier, Catherine

Batte, Bradley Spieler, William Donald Leslie, Carlo Menon, Richard Ribbon Fletcher, Newton Howard, Rabab Ward, William Parker, and Savvas Nicolaou. 2021. The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective. *Frontiers in Medicine*, 8.

Khalid A Gaber, Clive R McGavin, and Irving P Wells. 2005. Lateral Chest X-Ray for Physicians. *Journal of the Royal Society of Medicine*, 98(7):310–312.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415 [cs.LG]*.

Giovanni Irmici, Maurizio Cè, Elena Caloro, Natallia Khenkina, Gianmarco Della Pepa, Velio Ascenti, Carlo Martinenghi, Sergio Papa, Giancarlo Oliva, and Michaela Cellina. 2023. Chest X-ray in Emergency Radiology: What Artificial Intelligence Applications Are Available? *Diagnostics*, 13(2):216.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. 2023. MIMIC-IV-ED (version 2.2). PhysioNet.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. PhysioNet.

Catherine M Jones, Quinlan D Buchlak, Luke Oakden-Rayner, Michael Milne, Jarrel Seah, Nazanin Esmaili, and Ben Hachey. 2021. Chest radiographs and machine learning – Past, present and future. *Journal of Medical Imaging and Radiation Oncology*, 65(5):538–544.

Barry Kelly. 2012. The chest radiograph. *The Ulster Medical Journal*, 81(23620614):143–148.

Firas Khader, Jakob Nikolas Kather, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Karim Hamesch, Keno Bressem, Christoph Haarburger, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, and Daniel Truhn. 2023a. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports*, 13(1):10666.

Firas Khader, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Christoph Haarburger, Johannes Stegmaier, Keno Bressem, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. 2023b. Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers. *Radiology*, 309(1):e230806.

Ming Kong, Zhengxing Huang, Kun Kuang, Qiang Zhu, and Fei Wu. 2022. TranSQ: Transformer-Based Semantic Query for Medical Report Generation. In *MICCAI*, volume 13438, pages 610–620.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. 2024. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. ArXiv:2310.18341 [cs].

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2023. UniFormer: Unifying Convolution and Self-Attention for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *NAACL*, volume 1, pages 71–78.

Ilya Loshchilov and Frank Hutter. 2022. Decoupled Weight Decay Regularization. In *ICLR*.

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*, pages 5288–5304.

Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. 2023. Pragmatic Radiology Report Generation. In *Proceedings of the 3rd Machine Learning for Health Symposium*, pages 385–402. PMLR. ISSN: 2640-3498.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2024a. Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation. arXiv:2307.09758 [cs].

Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. 2024b. e-Health CSIRO at RRG24: Entropy-Augmented Self-Critical Sequence Training for Radiology Report G eneration. In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *ACL*, page 311.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance. ArXiv:2311.18681 [cs].

Zhen Qin, Xiaodong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes, and Yiran Zhong. 2022. The Devil in Linear Transformer. In *EMNLP*, pages 7025–7041.

Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. 2024. A Transformer-Based Model for Zero-Shot Health Trajectory Prediction. MedRxiv.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *CVPR*, pages 1179–1195.

Dinggang Shen. 2021. Grand Challenges in Radiology. *Frontiers in Radiology*, 1.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *EMNLP*, pages 1500–1519.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*, pages 7433–7442.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs].

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards Generalist Biomedical AI. *NEJM AI*, 1(3):AIoa2300138.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *CVPR*, pages 4566–4575.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *AAAI*, pages 9154–9160.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In *CVPR*.

Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280.

Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S. Kevin Zhou, and Li Xiao. 2022. DeltaNet: Conditional Medical Report Generation for COVID-19 Diagnosis. In *ICCL*, pages 2952–2961.

Ling Yang, Zhanyu Wang, and Luping Zhou. 2023. MedXChat: Bridging CXR Modalities with a Unified Multimodal Large Model. ArXiv:2312.02233 [cs].

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, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, page 100802.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

A Prior exam embeddings

The images, findings section, and impression section from previous exams were considered. For prior exams, the time delta was positive, calculated by subtracting the time of the prior exam from the current exam. The images, findings section, and impression section from prior exams were given distinct source embeddings, separate from the current exam, to enhance differentiation. The comparison section from the current exam was also investigated, anticipating that references to prior exams in this section would prompt the decoder to reflect this in the generated report. We explored prior exams with a history size h of up to three.

B Table column determination

The columns from the tables described in Figure 1 were given the following designations:

- For the ED stay table, the patients 'intime' was used as the event time. Gender (e.g., 'F'), race (e.g., 'HISPANIC OR LATINO'), and arrival transport (e.g., 'AMBULANCE') were designated as category columns. The disposition column was not considered.

- For the triage table, the 'intime' from the ED stay table was used. Temperature (e.g., '100.6'), heart rate (e.g., '93'), respiratory rate (e.g., '16'), O2 saturation (e.g., '94'), systolic blood pressure (SBP) (e.g., '110'), diastolic blood pressure (DBP) (e.g., '56'), and acuity (e.g., '2') were designated as value columns. Pain (e.g., '6-9' and 'yes.') and the chief complaint (e.g., 'BILATERAL FOOT PAIN') were designated as text columns.

- The column designations for the vital sign table were identical to the triage table, except for the rhythm column (e.g., 'Normal Sinus Rhythm'), which was treated as a category column. The vital signs table also had no chief complaint column and the 'charttime' column was used as the event time.

- For the medicine reconciliation table, the 'intime' from the ED stay table was used as the event time, as it pertains to the patient's medication history prior to the ED stay. The name column was designated as a text column, while the gsn, ndc, etc_rn, and etccode columns were designated as category columns. The etcdescription column was not considered, as it is a descriprion of the etccode column.

- For the medicine administration (`pyxis`) table, 'charttime' was used as the event time. The med_rn, name, gsn_rn, and gsn columns were all treated as category columns. The name column for the medicine reconciliation column did not have as high of a cardinality as the name column from the medicine reconciliation column, allowing it to be considered as a category column.

- For the metadata table, the 'PerformedProcedureStepDescription', 'ViewPosition', 'ProcedureCodeSequence_CodeMeaning', 'ViewCodeSequence_CodeMeaning', and 'PatientOrientationCodeSequence_CodeMeaning' columns were considered, and designated as category columns.

C Experiment setup

C.1 Metrics

CheXbert-F1 (Smit et al., 2020), RadGraph-F1 (Delbrouck et al., 2022), BLEU-4 (Papineni et al., 2001), and BERTScore-F1 (`roberta-large_L17_no-idf_rescaled`) (Zhang et al., 2020) have been found to correlate with radiologists' assessment of reporting (Yu et al., 2023) and were a part of our evaluation. Additionally, we include CXR-BERT (Boecking et al., 2022; Nicolson et al., 2024a), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin and Hovy, 2003) as part of our evaluation. CheXbert-F1, RadGraph-F1, and CXR-BERT were intended to capture the clinical semantic similarity between the generated and radiologist reports, while BERTscore-F1 was intended to capture general semantic similarity. Finally, CIDEr, METEOR ROUGE-L, and BLEU-4 were intended to capture the syntactic similarity between the generated and radiologist reports.

For the models in Table 2 that generate a report for each image in an exam, the average score was taken across all reports for an exam. Following this, the final average score was computed across all exams for both models that generate a report per image and those that generate a report per exam.

For CheXbert, the macro-averaged F1 was computed between the 14 CheXbert observations extracted from the generated and radiologist reports. "No mention", "negative", and "uncertain" were considered negative, while "positive" was considered positive. Here, the true positives, false positives, and false negatives were averaged over the reports of each exam for the models that generate a report per image.

We also perform statistical testing; first, a Levene's test was conducted to reveal if the variances across model scores was homogeneous or not. If the assumption of equal variances was upheld, a one-way ANOVA was conducted to determine if there was a significant difference between models. Finally, pairwise Tukey-HSD post-hoc tests were used for pairwise testing. If the assumption of equal variances was violated, a one-way Welch's ANOVA was conducted to determine if there was a significant difference between models. Finally, Games-Howell post hoc tests were used for pairwise testing. A p-value of 0.05 was used for all significance testing. Statistical testing was not performed for CheXbert, as it is a classification metric.

C.2 Model

Our model is illustrated in Figure 2; following (Nicolson et al., 2024b), we utilised UniFormer as the image encoder (in particular, the 384×384 base model warm started with its token labelling fine-tuned checkpoint) (Li et al., 2023). The image embeddings are formed by processing each image in the exam separately with the image encoder and then projecting its last hidden state to match the decoder's hidden size using a learnable weight matrix. Each image was resized using bicubic interpolation so that its smallest side had a length of 384 and its largest side maintained the aspect ratio. Next, the resized image was cropped to a size of $\mathbb{R}^{3 \times 384 \times 384}$. The crop location was random during training and centred during testing. Following (Elgendi et al., 2021), the image was rotated around its centre during training, where the angle of rotation was sampled from $\mathcal{U}[-5°, 5°]$. Finally, the image was standardised using the statistics provided with the UniFormer checkpoint. A maximum of five images per exam were used during training. If more were available, five were randomly sampled uniformly without replacement from the exam.

Again following (Nicolson et al., 2024b), we employed the Llama architecture for the decoder, which is notable for features such as its rotary positional encoding (RoPE), root mean square normalisation (RMSNorm), and SwiGLU activation function (Touvron et al., 2023). A byte-level byte pair encoding tokenizer (Wang et al., 2020) was trained with a vocabulary size of 30 000. It was trained on the findings, impression, indication, and history sections (not the comparison section) of the entire MIMIC-CXR training set, as well as the 'pain' and 'chiefcomplaint' columns from the triage table, the 'name' column of the medicine reconciliation table, and the 'pain' column from the vital signs table (from the entire MIMIC-IV-ED dataset). Newline, tab, repeated whitespaces, and leading and trailing whitespaces were removed from any text before tokenization.

The hyperparameters of the Llama decoder were six hidden layers, a hidden size of 768, 12 attention heads per layer, and an intermediate size of 3 072. The maximum number of position embeddings was set to 2048 to accommodate all the patient data embeddings and the report tokens. The maximum number of tokens that could be generated was set to 256, which was also the limit for the radiologist reports during training. During testing, a beam size of four was utilised. The Llama decoder allows a custom attention mask to be provided in current implementations.[4] This enabled non-causal masking to be utilised for the prompt and causal masking for the report token embeddings, as shown in Figure 5. This ensured that the self-attention heads were able to attend to all of the patient data embeddings at each position.

C.3 Training

Three stages of training were performed. Each stage used *AdamW* (Loshchilov and Hutter, 2022) for mini-batch gradient descent optimisation, where training and evaluation was performed on a 94GB NVIDIA H100 GPU. The three stages were

[4]https://huggingface.co/blog/poedator/4d-masks

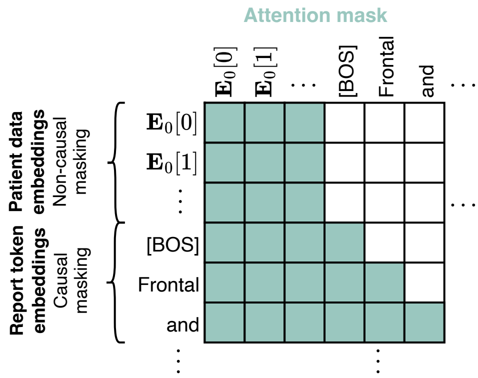


Figure 5: Attention mask for the decoder. Non-causal masking was used for the patient data embeddings and causal masking for the report token embeddings.

as follows:

1. Teacher forcing (TF) (Williams and Zipser, 1989) was performed on the MIMIC-CXR dataset with only the images for an exam as input, and exams that contained both a findings and impression section. This gave a training/validation split of 232 853/1 837 images, 125 416/991 exams, and 57 101/436 patients. Training was performed with an initial learning rate of 5e-5, a mini-batch size of 8, a maximum of 32 epochs, and with float16 automatic mixed precision. All model parameters were trainable during this stage. The validation macro-averaged CheXbert-F1 was the monitored metric for checkpoint selection. This stage was necessary, as the language model struggled to generate reports from multiple sources without prior learning.

2. TF on the dataset described in Section 3 with the inputs described in Table 1. The training strategy was identical to the previous stage, except that a maximum of 16 epochs was performed, and the image encoder's parameters were frozen (except for its projection). The models featured in Table 1 were trained using only the first two stages.

3. Reinforcement learning using self-critical sequence training (SCST) (Rennie et al., 2017) with CXR-BERT and BERTScore as the reward (each weighted with 0.5) was performed in the final stage of training. The sample report for SCST was generated with top-k sam-

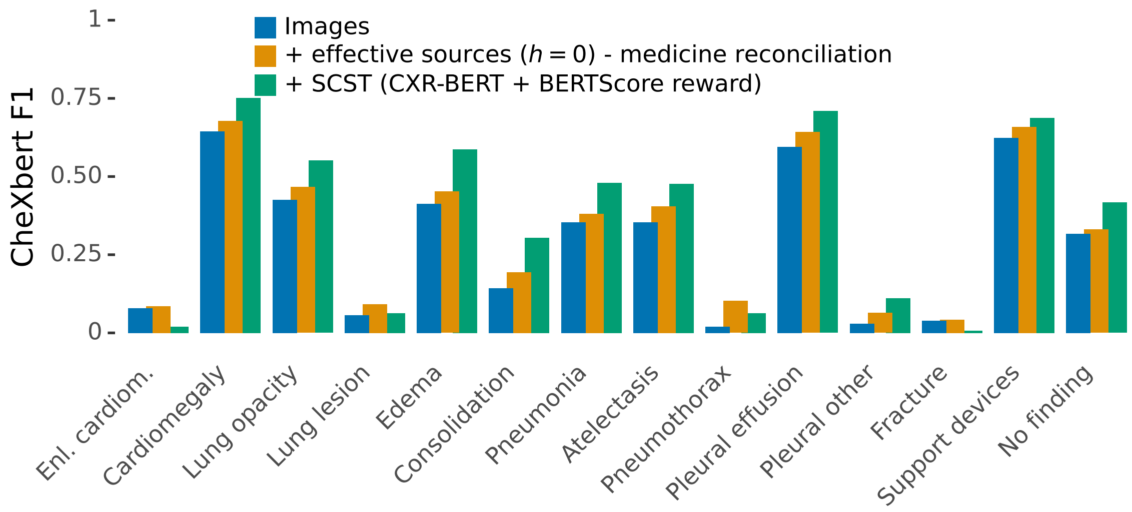


Figure 6: F1-score for each CheXbert label. ($n = 9\,580$ exams; 958×10 runs for 'Images' and 'Images + effective sources ($h = 0$) - medicine reconciliation' and $n = 2\,874$ exams; 958×3 runs for 'Images + effective sources ($h = 0$) - medicine reconciliation + SCST (CXR-BERT + BERTScore reward)'.)

pling ($k = 50$). Training was performed with an initial learning rate of 5e-6, a mini-batch size of 32, a maximum of 24 epochs, and with float32 precision. The image encoder's parameters were frozen during this stage (except for its projection). The validation BERTScore-F1 was the monitored metric for checkpoint selection, as it helped to select checkpoints less prone to repetitions. This stage of training was only applied to the best model from Table 1, 'Images + effective sources (h = 0) - medicine reconciliation', with the results presented in Table 2. This model had 161 185 728 parameters.

C.4 Comparison Models

The generated reports for the models in Table 2 were attained as follows:

- EMNLI reports were generated following `https://github.com/ysmiura/ifcc` (Miura et al., 2021).
- CMN reports were generated following `https://github.com/zhjohnchan/R2GenCMN` (Chen et al., 2021).
- TranSQ reports were kindly provided by the authors (Kong et al., 2022).
- RGRG reports were generated following `https://github.com/ttanida/rgrg` (Tanida et al., 2023).
- CvT2DistilGPT2 reports were generated following `https://github.com/aehrc/cvt2distilgpt2` (Nicolson et al., 2023).
- RaDialog reports were kindly provided by the authors (Pellegrini et al., 2023).
- MedXChat reports were kindly provided by the authors (Yang et al., 2023).
- CXR-LLaVA-v2 reports were generated following `https://huggingface.co/ECOFRI/CXR-LLAVA-v2` (Lee et al., 2024).
- CXRMate reports were generated following `https://huggingface.co/aehrc/cxrmate` (Nicolson et al., 2024a).
- CXRMate-RRG24 reports were generated following `https://huggingface.co/aehrc/cxrmate-rrg24` (Nicolson et al., 2024b).

CXRMate-RRG24 was trained on five datasets, including MIMIC-CXR. RGRG was trained on the ImaGenome dataset derived from MIMIC-CXR — which may have some overlap with our test set.

D Ancillary results

In Figure 6, the F1-scores for each CheXbert label are shown. The 'Images + effective sources ($h = 0$) - medicine reconciliation' model from Table 1 improves performance across all labels compared to the 'Images' model. This suggests that

Table 4: Results for exams that have a medicine administration table ($n = 3\,520;$ studies 352×10 runs). Underlined scores indicate a significant difference to the scores of 'Images' ($p < 0.05$).

Inputs	RG	CX	CB	BS
Images	26.24	28.36	57.17	24.33
Images + medicine administration	$\underline{\textbf{26.95}}$	**28.53**	$\underline{\textbf{58.94}}$	$\underline{\textbf{24.93}}$

incorporating ancillary data from MIMIC-IV-ED and MIMIC-CXR provides a general improvement, rather than benefiting any specific pathology.

Further improvements are seen when training the 'Images + effective sources ($h = 0$) - medicine reconciliation' model with SCST (i.e., our model from Table 2) for most pathologies. However, there are performance decreases for 'enlarged cardiomediastinum', 'lung lesion', 'pneumothorax', and 'fracture'. This might be due to these pathologies being underrepresented in the MIMIC-CXR dataset, leading the model to optimise for more common pathologies during SCST.

The results for exams that include a medicine administration table are show in Table 4. Adding the medicine administration table produced a significant improvement in the scores, indicating that it should be considered if available.