

# Enhancing Image-to-Text Generation in Radiology Reports through Cross-modal Multi-Task Learning

Nurbanu Aksoy<sup>†</sup>, Serge Sharoff<sup>‡</sup>, Nishant Ravikumar<sup>†</sup>

School of Computing<sup>†</sup>, School of Languages, Cultures and Societies<sup>‡</sup>  
University of Leeds, United Kingdom

## Abstract

Image-to-text generation involves automatic generation of descriptive text from images, with this paper focusing on generation of reports from X-ray images. However, traditional approaches often exhibit a semantic gap between visual and textual information. In this paper, we propose a multi-task learning framework to leverage both visual and non-imaging data for generating radiology reports. Along with chest X-ray images, 10 additional features comprising numeric, binary, categorical, and text data were incorporated to create a unified representation. The model was trained to generate text, predict the degree of patient severity, and identify medical findings. Multi-task learning, especially with text generation prioritisation, improved performance over single-task baselines across language generation metrics. The framework also mitigated overfitting in auxiliary tasks compared to single-task models. Qualitative analysis shows more coherent narratives and more accurate identification of findings, though some repetition and disjointed phrasing remain. This study demonstrates the benefits of multi-modal, multi-task learning for image-to-text generation applications.

**Keywords:** image-to-text generation, cross-modal data fusion, multi-task learning

## 1. Introduction

With the development of deep learning techniques, the fields of computer vision and natural language processing have started to converge, as both images and texts can be represented via compatible embeddings. This convergence has led to success in the challenging cross-modal task of image-to-text generation, which involves automatically generating descriptive text from images, and it has numerous real-world applications, such as image captioning (Stefanini et al., 2022), medical report generation (Ramirez-Alonso et al., 2022), and assisting the visually impaired (SS et al., 2023).

Traditional approaches to image-to-text generation relied on independent models for image understanding and natural language generation, which often led to a semantic gap between the visual and textual information due to the lack of synergy between the two modalities. The use of multi-modal data has provided a way to improve the coherence and accuracy of generated text for image-to-text tasks. Additionally, multi-task learning, where a single model is trained to perform multiple related tasks simultaneously, has played an important role in harnessing the full potential of multi-modal data in image-to-text generation (Bayoukh et al., 2021).

In this work, we explore a novel approach for image-to-text generation using multi-modal data through multi-task learning. Specifically, we propose a framework that is trained on medical data to not only generate radiology reports describing the image content but also predict the patient's severity level (the criticality of a patient's medical condition) and indicate the presence or absence of specific findings from the image.

To fully leverage the multi-modal learning scenario, we incorporated a set of ten supplementary features along with the visual data. These additional features encompass a diverse range of data types, including numeric, binary, categorical, and textual information. These non-imaging data, both clinical and non-clinical in nature, were obtained from patient health records, thereby enabling the creation of a unified data representation. This unified representation was fed into cross-attention layers along with the visual features to generate attended visual features for each task-specific decoder. The proposed approach generates medical reports that are detailed, accurate, and useful for real-life applications. We evaluated the model on a large dataset of chest X-ray images and radiology reports, outperforming the single-task approach on quantitative natural language generation evaluation metrics.

This paper is structured as follows. Section 2 reviews previous related research and approaches in this field. Section 3 describes the data used in this study and outlines the methods proposed. Section 4 explains how the methods were implemented and the model architecture. Section 5 presents the results, both quantitative and qualitative and provides a discussion. Finally, Section 6 summarises the main conclusions drawn from this work.

## 2. Related Work

Multi-task learning (MTL) has become increasingly popular in natural language processing (NLP) because leveraging the commonalities and differences between related tasks improves overall per-

formance (Zhang and Yang, 2021). Recent work on MTL for NLP tasks has been categorised into two broad frameworks: joint training, where all tasks are trained concurrently, and multi-step training, where trains the tasks in a sequence of steps (Zhang et al., 2022).

This paper focuses on joint training for text generation, an area where MTL has been shown to be beneficial. By training text generation models on multiple objectives simultaneously, performance on downstream tasks like summarisation and translation can be improved (Zhang et al., 2019; Su et al., 2021).

In this context, Tang et al. (2017) studied the problem of joint question answering (QA) and question generation (QG) and proposed a training framework that trains the models of QA and QG simultaneously. They implemented a QG model based on sequence-to-sequence learning and a QA model based on recurrent neural networks. Guo et al. (2018) proposed a multi-task learning approach for abstractive summarisation, using auxiliary tasks of question generation and entailment generation. They introduced novel multi-task architectures with high-level layer-specific sharing across multiple encoder and decoder layers of the three tasks.

Sachan and Xing (2018) proposed a self-training method for jointly learning to ask and answer questions, leveraging unlabeled text along with labelled question-answer pairs for learning. They evaluated their approach on four benchmark datasets and showed significant improvements over established baselines. Zhang et al. (2019) proposed DIALOGPT, a large-scale generative pre-training approach for conversational response generation. They use a transformer-based architecture and pre-train the model on a large corpus of conversational data. They also propose a novel training objective that encourages the model to generate diverse and informative responses. Su et al. (2021) proposed a multi-task pre-training approach for plug-and-play task-oriented dialogue systems. They used a transformer-based architecture and pre-trained the model on multiple related tasks to improve the performance of each task.

While multi-task learning has demonstrated its potential to enhance text generation performance, effectively training a neural network to produce coherent narrative text from diverse multi-modal and cross-modal inputs remains a complex and challenging task. Cross-modal learning, which involves models understanding connections across modalities, presents inherent difficulties. Nonetheless, multi-task learning can mitigate some of these challenges in cross-modal text generation by enabling models to jointly learn representations across modalities while optimising multiple objectives.

More specifically, Li et al. (2020) introduced a new learning approach known as Oscar (Object-Semantics Aligned Pre-training) for Cross-modal tasks involving vision and language. This method capitalises on the insight that prominent objects in images can be accurately identified and are frequently referenced in associated text. Oscar employs object tags detected within images as reference points, facilitating the alignment learning process between images and text. More recently, Sharma et al. (2023) introduced a novel task called EXCLAIM, which generates explanations for visual semantic role labelling in memes. They also proposed a multi-modal, multi-task learning framework called LUMEN, which jointly learns to predict the correct semantic roles and generate suitable natural language explanations.

In the medical domain, image-to-narrative language generation is a more challenging task due to the need for more comprehensive paragraph annotations, the subtlety of distinctions in medical images, and the requirement for additional contextual information to analyse and interpret medical images, unlike the relatively straightforward nature of natural images. While multi-modal approaches have proven valuable in tackling some of these challenges in various vision language tasks such as visual question answering (Eslami et al., 2021; Wang et al., 2022a; Liu et al., 2023) or medical report generation (Yang et al., 2022; Wang et al., 2022b; Wu et al., 2023; Aksoy et al., 2023), there is a notable gap in the exploration of multi-task learning techniques for radiology report generation using diverse multi-modal data. Applying multi-task learning could potentially help models learn joint representations across text, images, and other clinical data to improve coherence and accuracy when generating narratives from multi-modal inputs in the medical domain.

### 3. Methodology

#### 3.1. Problem Formulation

The approach is designed to concurrently perform three tasks: Text Generation, Ordinal Classification, and Multi-Label Classification. These tasks involve processing various inputs and producing meaningful outputs while optimising for different loss functions. By integrating them within a single framework, the overarching objective is to leverage information and features from the other tasks, ultimately resulting in more precise and context-aware text generation capabilities.

Given an image  $I$ , unified additional features  $F$ , ground truth radiology report text sequences  $Y$ , ground truth ordinal acuity levels  $T$  (where  $T$  is an integer value between 1 to 5 indicating the sever-

ity of the medical finding), and ground truth labels for findings  $Z$  (where  $Z$  are multi-label categorical values with 5 possible labels indicating the presence or absence of specific medical findings), the objective is to learn an encoder-decoder model to minimise the loss for the three tasks:

**Image-to-text Generation (ITG)**, generates a radiology report  $\hat{Y}$  that maximises the probability of the ground truth text sequence  $Y$  given the image  $I$  and features  $F$ . The loss is defined as the cross-entropy loss between  $\hat{Y}$  and  $Y$ .

**Ordinal Classification (OC)** predicts ordinal acuity level  $\hat{T}$  given  $I$  and  $F$ . The loss is defined as the binary cross-entropy between  $\hat{T}$  and  $T$ .

**Multi-Label Classification (ML)** predicts multi-labels for findings  $\hat{Z}$  given  $I$  and  $F$ . The loss is the binary cross-entropy between  $\hat{Z}$  and  $Z$ .

**Overall Loss Function**, denoted as  $L$ , is composed of three task-specific loss components:

$$L = \alpha \cdot L_{\text{ITG}}(\hat{Y}, Y) + \beta \cdot L_{\text{OC}}(\hat{T}, T) + \gamma \cdot L_{\text{ML}}(\hat{Z}, Z)$$

Where:

$L_{\text{ITG}}(\hat{Y}, Y)$  is the loss function for ITG.

$L_{\text{OC}}(\hat{T}, T)$  is the loss function for OC.

$L_{\text{ML}}(\hat{Z}, Z)$  is the loss function for ML.

Here,  $\alpha$ ,  $\beta$ , and  $\gamma$  are hyperparameters that control the relative weighting of these losses. During the experiments to determine the optimal values for the loss hyperparameters, we explored various approaches. Prioritising  $\alpha$  in the context of the main task yielded better results. Therefore, with ITG having the highest weight, followed by OC, and then ML, we set  $\alpha > \beta > \gamma$ . This approach emphasises the prioritisation of the ITG task, followed by OC and ML. It leverages multi-task learning to enhance performance primarily in the main task of report generation.

During the training process, the model's encoder-decoder parameters  $\theta$  are optimised by minimising the overall loss  $L$  over the training dataset using the following objective:

$$\begin{aligned} \min_{\theta} L(\theta) = \min_{\theta} & \left( \alpha \cdot L_{\text{ITG}}(\hat{Y}, Y; \theta) \right. \\ & + \beta \cdot L_{\text{OC}}(\hat{T}, T; \theta) \\ & \left. + \gamma \cdot L_{\text{ML}}(\hat{Z}, Z; \theta) \right) \end{aligned}$$

### 3.2. Feature Extraction and Pre-processing

This section outlines the preparation and encoding of various data modalities employed in this study. Every image was resized and then normalised to ensure uniform intensity levels. Subsequently, a

pre-trained model extracted visual features from the images. The resulting visual feature vector was then input into a transformer-based encoder to gain a deeper understanding of the recognised elements in the image.

Numerical variables were initially preprocessed to remove potential outliers based on domain knowledge and clinical perspectives. For instance, data points with physiologically implausible values (e.g. a patient temperature of 67°C, which substantially exceeds normal human ranges) were identified as probable errors and excluded. The remaining data for each numerical feature was then standardised to a 0 to 1 range by rescaling based on the minimum and maximum observed values.

Binary variables were converted to numerical representations. Integer values were assigned to each group in categorical data, one-hot encoded, and reshaped into a 2D array for input to the encoder. Text-based variables were pre-processed by converting to lowercase, removing unnecessary punctuation using regular expressions, and condensing consecutive periods into single spaces. Double periods were replaced with single spaces to maintain consistent text formatting. Additional standardisation involved replacing shorthand phrases or abbreviations with their full-text equivalents, correcting errors, and addressing inconsistencies in pluralisation. This pre-processing pipeline standardised various data types for input into the model. Lastly, acuity levels were encoded using cumulative one-hot representation where ordinal levels are mapped to binary vectors. Each vector has a length equal to the number of ordinal levels minus one. The presence of a '1' in a specific position within the binary vector indicates the corresponding ordinal level. This encoding preserves the ordinal relationships between levels.

### 3.3. Cross-modal MTL Network

Our framework, as illustrated in Figure 1, comprises two primary blocks: shared layers that are common to all tasks and task-specific layers tailored to each individual task. Additionally, within this framework, there are 3 sub-blocks, namely the visual block, unified data block and cross-attention block.

The unified data block first encodes all numerical data into a single vector via the encoding process described in Section 3.2, then passes this through a dense layer to obtain a condensed representation of the integer outputs. Categorical data is one-hot encoded and passed through a separate dense layer to obtain categorical embeddings. Text-based data from each modality is passed through distinct embedding layers, with the resulting embeddings further processed by dense layers to obtain the final text data representations.

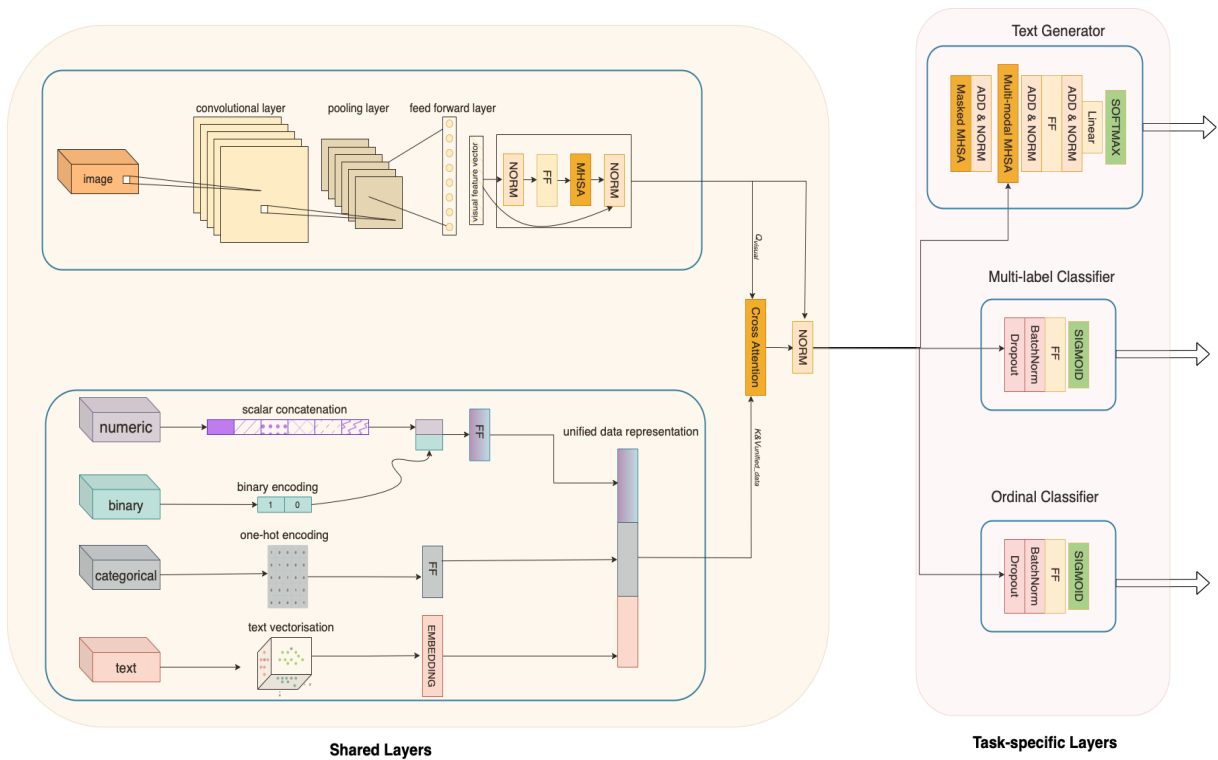


Figure 1: An overall framework of our proposed Cross-modal multi-task learning network

The visual block comprises two encoding stages. Initially, the pre-trained EfficientNet model, trained on ImageNet, functions as the Convolutional Neural Network (CNN) encoder for extracting the features from the image, denoted as 'x,' into L vectors. Each of these vectors is a D-dimensional representation corresponding to the features extracted from distinct spatial locations within the image. Subsequently, these visual vectors are utilised in two consecutive multi-head self-attention blocks to obtain a hybrid representation. The first self-attention block attends exclusively to the image features, while the cross-attention block focuses on the unified data representation, utilising the image features as queries. This allows the network to contextually focus on relevant aspects of the unified data conditioned on the image content.

The ordinal classification (OC) and the multi-label classification (MLC) decoders take the encoder outputs and normalise them. The normalised outputs are subsequently input to a dense layer with a sigmoid activation function to derive probabilities. These probabilities are then averaged across the sequence dimension, resulting in scalar predictions.

The text generator decoder employs causal masking, which is combined with padding masks. It comprises two consecutive multi-head self-attention layers. The first layer attends exclusively to the target sequence, while the second layer at-

tends to the encoder outputs using the decoder inputs as queries. The attended representations are processed through a two-layer positionwise feed-forward network. Finally, a linear layer produces predictions, which are used as inputs for the next time step.

## 4. Experimental Settings

### 4.1. Dataset

The dataset used in this research was created by combining three public databases - MIMIC-CXR, MIMIC-IV, and MIMIC-IV-ED. MIMIC-CXR contains over 377,000 chest X-ray images from multiple views and associated de-identified radiology reports for 63,473 patients. MIMIC-IV provides de-identified patient information like demographics for individuals admitted to Beth Israel Deaconess Medical Center (BIDMC). MIMIC-IV-ED contains detailed clinical data for emergency department visits at BIDMC from 2011-2019. Each database uses unique subject identifiers for patients. However, linking records across databases by patient ID was ineffective since patients may have multiple visits. Moreover, to generate accurate reports, non-imaging data must align time-wise with the chest x-ray. Therefore, we linked MIMIC-CXR and MIMIC-IV-ED records for patients in the ED during report generation. After data cleaning, the final dataset

had 65,813 entries with 10 features including oxygen saturation, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature, patient’s chief complaint, ICD (diagnosis) title, gender and ethnicity.

Challenges with this dataset include skewed distribution toward normal cases and duplicate reports for different patients. To address this, we selected a balanced subset of 10,000 samples - 7,000 for training, 2,000 for validation, and 1,000 for the test set. Since there are no comparable public datasets with similar non-imaging data, we used this dataset to develop and evaluate our approach.

## 4.2. Evaluation Metrics

### Natural Language Generation Metrics

To evaluate the quality of the generated reports, several automatic evaluation metrics were computed to compare the generated text to the reference reports. The first set of metrics used was the BLEU-1 to BLEU-4 scores, which assessed n-gram precision for unigrams up to 4 grams. These scores measured the local word-level similarity between the generated and reference texts, with higher scores indicating greater similarity. The second metric used was the ROUGE-L score, which measured the longest common subsequence and assessed the quality of the generated text in terms of recall and precision.

While the metrics above are traditionally used in radiology report generation, we have also evaluated semantic similarity using BERTScore and Bio-ClinicalBERT Score (Equation 1). These metrics used contextual embeddings from BERT and Bio-ClinicalBERT models to provide a more nuanced assessment of meaning compared to strict n-gram matching. The BERT-based metrics were able to capture whether the generated reports conveyed clinically coherent descriptions despite differing word usage compared to the reference. These automated evaluation metrics quantified linguistic similarity at word level, sentence level, and semantic meaning levels.

$$\text{BScore} = \frac{1}{N} \sum_{i=1}^N (F1(y_i, \hat{y}_i) + \text{Suff}(y_i, \hat{y}_i) + \text{Flu}(y_i, \hat{y}_i)) \quad (1)$$

Where:  $N$  is the number of sentence pairs in the evaluated dataset  $y_i$  is the  $i^{\text{th}}$  reference sentence  $\hat{y}_i$  is the  $i^{\text{th}}$  generated sentence  $F1(y_i, \hat{y}_i)$  is the F1 score between  $y_i$  and  $\hat{y}_i$  using both BERT and ClinicalBERT embeddings separately  $\text{Suff}(y_i, \hat{y}_i)$  is the sufficiency score between  $y_i$  and  $\hat{y}_i$   $\text{Flu}(y_i, \hat{y}_i)$  is the fluency score between  $y_i$  and  $\hat{y}_i$

### Classification Metrics

We utilised several metrics to evaluate model performance on multi-label and ordinal classification tasks. For multi-label classification, the metrics we employed included Precision, Recall, F1 Score, Hamming Loss, and Exact Match Ratio. Precision and Recall evaluated the accuracy of predicting positive labels and capturing all true positives respectively. F1 Score provided the balance between Precision and Recall. Hamming Loss quantified label prediction errors, and Exact Match Ratio measured how often the model correctly predicted all labels for a given instance.

For ordinal classification, our metrics consisted of Ordinal Classification Accuracy, Mean Absolute Error, Mean Squared Error, and the Accuracy-Correlation Hybrid Metric. Ordinal Classification Accuracy measured the accuracy by computing the total number of correct predictions divided by the total number of predictions. Mean Absolute Error and Mean Squared Error quantified the average magnitude of errors in predicted ordinal values. The Accuracy-Correlation Hybrid Metric combined aspects of accuracy and correlation to evaluate the preservation of the ordinal relationship.

## 4.3. Experimental Setup

The cross-modal multi-task learning model was implemented in TensorFlow 2.3.0 and Keras using a Transformer architecture. Transformer layers implemented with 3 attention heads, and 256 dimensional feedforward layers.

The model was trained on 7,000 data using an Adam optimiser with a learning rate warmup over 10% of steps up to  $3e-5$  and a batch size of 32. The validation, and test sets consist of 2,000, and 1,000 data, respectively. Training continued for 100 epochs with early stopping monitoring the validation loss with patience=10.

For parity in model optimisation, we maintained consistency in the choice of hyperparameters (e.g. learning rate, batch size, etc.) when training each of the assessed models. All models were trained on NVIDIA Tesla A100 GPUs with 40GB memory.

## 5. Results and Discussion

We evaluated the performance of our proposed multi-task learning (MTL) model on three tasks: text generation, ordinal classification, and multi-label classification. The MTL model was trained in two configurations - with equal weighting across tasks (MTL-EQ) and with task prioritisation for the text generation task (MTL-TP). We compared these MTL models to single-task learning baselines (STL).

Method	B_1	B_2	B_3	B_4	BS <sub>F1</sub>	Bio-CBS <sub>F1</sub>	R_L
STL	0.3326	0.2159	0.1488	0.0950	0.2056	0.7857	0.3096
MTL_EQ	0.3352	0.2229	0.1570	0.0983	0.1958	0.7883	0.3235
MTL_TP	<b>0.3424</b>	<b>0.2295</b>	<b>0.1616</b>	<b>0.1035</b>	<b>0.2065</b>	<b>0.7898</b>	<b>0.3366</b>

Table 1: Performance comparison of an image-to-text generator using different training approaches. B\_n for BLEU-n, R\_L for ROUGE-L, BS<sub>F1</sub> for BERT Score F1Score and CBS<sub>F1</sub> for Bio-ClinicalBERT Score F1Score. STL denotes Single Task Learning, MTL-TP represents Multi-Task Learning with Task Prioritisation for text generation, and MTL-EQ indicates Multi-Task Learning with equal task weights for each task

	B_1	B_2	B_3	B_4	BS <sub>F1</sub>	Bio-CBS <sub>F1</sub>	R_L
p-value	0.00117	0.00006	0.00035	0.02569	0.83824	0.00611	0.00000

Table 2: P-values from pairwise t-test between STL and MTL\_TP approaches (rounded to 5 decimal places)

Method	Precision	Recall	F1 Score	Hamming Loss	Exact Match Ratio
STL	0.7520	0.6552	0.7005	0.1466	0.8534
MTL_EQ	0.6502	0.6641	0.6562	0.1618	0.8382
MTL_TP	0.6603	0.6775	0.6694	0.1598	0.8402

Table 3: Comparing performance of the multi-label classifier in Single-Task, Task-Prioritised Multi-Task and Equal-Weight Multi-Task Learning

Method	Accuracy	MAE	ACC+Corr
STL	0.8790	0.1210	0.8197
MTL_EQ	0.8553	0.1447	0.7907
MTL_TP	0.8640	0.1360	0.8005

Table 4: Comparing performance of the ordinal classifier in Single-Task, Task-Prioritised Multi-Task and Equal-Weight Multi-Task Learning

Table 1 displays the results for text generation tasks, measured in terms of BLEU scores (B\_1 to B\_4), BERT Score F1Score (BS<sub>F1</sub>), Bio-ClinicalBERT Score F1Score (Bio-CBS<sub>F1</sub>), and ROUGE-L (R\_L). When employing Multi-Task Learning with Equal Task Weights (MTL-EQ), the model exhibits slightly improved performance across most metrics. Notably, the MTL-TP model achieved the best performance, outperforming STL and MTL-EQ on all metrics. This demonstrates the benefits of MTL with proper task weighting for improving text generation quality. It also validates the capability of MTL to leverage representations learned across related tasks.

To assess the statistical significance of the improvements achieved by the MTL-TP approach over the Single-Task Learning (STL) baseline, we conducted a pairwise t-test for each metric. Table 2 presents the p-values from these significance tests, rounded to 5 decimal places. The small p-values obtained for most metrics, particularly BLEU-

2, BLEU-3, and ROUGE-L, indicate that the improvements in text generation quality are statistically significant. This analysis focuses on the text generation task, as it was the primary task where substantial improvements were observed with the MTL-TP approach.

On ordinal classification, STL achieved the highest accuracy and lowest mean absolute error. However, the MTL-TP model was competitive, with only a 1.5% drop in accuracy compared to STL (Table 4). For multi-label classification, STL again performed the best in terms of precision, recall, F1, hamming loss, and exact match ratio. The MTL models achieved comparable but slightly lower performance (Table 3). However, it is important to note that STL suffered from overfitting after only 4 epochs on the multi-label classification task. Despite attempts to optimise hyperparameters, STL continued to overfit within a few epochs. In contrast, MTL helped prevent overfitting for 5-6 additional epochs compared to STL on this task. So even though computed multi-label classification metrics show better STL performance, this overfitting was inevitable with single-task training.

For a better comparison of the results, we have colour-coded the illustrated samples to match their respective ground truth labels, see Figure 2. Ambiguous or repeated expressions are denoted in italics, while incorrect predictions or expressions not present in the original report are underlined. If

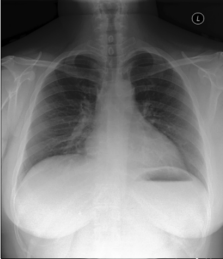

Input Image	Ground Truth	STL	MTL_EQ	MTL_TP
	the cardiac silhouette size is top normal mediastinal and hilar contours are normal lungs are clear and the pulmonary vascularity is normal no pleural effusion or pneumothorax is present no acute osseous abnormalities are detected	[UNK] for comparison there is no focal consolidation pleural effusion or pneumothorax the heart is normal in size the mediastinal and hilar contours are normal the pulmonary vascularity is normal there are no acute osseous abnormalities surgical clip is seen in	trace pleural effusion is seen the cardiac and mediastinal silhouettes are unremarkable there is no pneumothorax or pneumothorax the cardiac and hilar contours are normal there is no pneumothorax is no free air below the right hemidiaphragm is seen no focal	the lungs are clear the cardiac silhouette is normal in size the mediastinal and hilar contours are normal the pulmonary vasculature is normal no pleural effusion or pneumothorax is seen no acute osseous abnormalities identified there is seen the lungs are
	minimal basilar atelectasis is seen there is no focal consolidation no pleural effusion or pneumothorax is seen the cardiac and mediastinal silhouettes are unremarkable	change in the left lower lobe no focal consolidation pleural effusion or pneumothorax is seen the cardiac and mediastinal silhouettes are unremarkable <u>no displaced rib fracture is identified the visualized upper abdomen is unremarkable no displaced fracture is seen beneath the</u>	or size is normal the mediastinal and hilar contours are normal there is no pleural effusion or pneumothorax no free air below the right hemidiaphragm is seen no acute osseous abnormality is seen the right hemidiaphragm is seen no free air	aortic arch is again seen with known lower lobe atelectasis no focal consolidation pleural effusion or pneumothorax cardiac and mediastinal silhouettes are unremarkable pulmonary edema is detected no acute osseous structures are intact no acute osseous abnormalities identified no pulmonary edema

Figure 2: Illustrations of generated text from different approaches

the statement is for both, such as when the expression is both repeated and not in the original report, we used both italics and underlines. The results indicate promise in generating reports that capture many of the main findings mentioned in the ground truth. All approaches show a generally positive alignment in terms of grammar, however, some of the generated reports exhibit repeated words or phrases, which can affect the overall coherence.

Particularly, The STL approach struggles with unnatural wording like "[UNK] for comparison" and hallucinates findings that are not present in either the image or the ground truth. The MTL\_EQ output has disjointed phrasing and repetition indicating a lack of narrative coherence. In contrast, the MTL\_TP generates smooth, logical statements more similar to the ground truth, with some minor repetition. In the second example, the MTL\_TP text exhibits clearer structure, with sentences covering distinct findings. It includes details like "pulmonary edema", "aortic arch is again seen" and "no acute osseous abnormalities identified" not in the original text but present in the image.

Overall, results demonstrate good progress for the radiology report generation model, with accurate identification of key findings but also room for improvement. The STL model sometimes seems to include extraneous or inaccurate details where the equal-weighted MTL shows improvements in content quality over STL, but suffers from repetitiveness and disorganized narratives. The MTL\_TP generates the most coherent language with smoother transitions between ideas. The results demonstrate that task prioritisation during multi-task training can better capture logical relationships between medical imaging findings compared to single or equal-

weighted multi-task models. Further tuning to optimise content selection and narrative flow could help MTL models produce the text closer to ground truth quality.

## 6. Conclusion

In this paper, we proposed a novel framework for image-to-text generation in the medical domain. By leveraging multi-modal data and employing a multi-task learning (MTL) approach, our proposed model aimed to bridge the gap between image understanding and natural language generation, ultimately improving the quality and coherence of generated medical reports.

Our model was trained on chest X-ray images and radiology reports to generate text descriptions, predict patient severity levels, and identify medical findings. We incorporated 10 additional features along with visual data to create a unified representation for multi-modal learning.

The results demonstrate the benefits of multi-task learning, particularly with proper task weighting, for improving text generation quality. The multi-task model with text generation prioritisation (MTL-TP) outperformed single-task learning baselines across all language generation metrics. While single-task learning achieved better performance on the auxiliary tasks of ordinal and multi-label classification, it suffered from severe overfitting after only a few epochs. In contrast, multi-task learning helped prevent overfitting for these tasks and trained for additional epochs.

Qualitative analysis also showed that MTL-TP generated more coherent narratives that better captured logical relationships between medical find-

ings. The generated reports exhibited good identification of key findings from the images.

The ability to automatically produce accurate radiology reports from chest X-rays could greatly assist clinicians and reduce reporting bottlenecks. More broadly, our proposed approach and findings can inform future research on combining computer vision and natural language processing for medical applications.

Furthermore, our approach's success in the medical domain suggests the potential for its generalisation to other domains. The principles of the integrated learning approach, as demonstrated in our research, can be applied to a wide range of applications beyond medical image analysis. For example, in image captioning (Sirisha and Sai Chandana, 2022), our model's ability to understand the visual content and generate coherent textual descriptions could significantly enhance the accessibility and interpretability of images in various fields, including social media, e-commerce, and more. Moreover, in grounded story generation (Hong et al., 2023), our approach can be extended to create compelling and contextually relevant narratives based on visual cues, making it suitable for content generation in the entertainment and creative industries. The integrated learning approaches, which proved effective in our medical domain application, could play an important role in advancing these related domains, improving the quality and relevance of content generated from visual inputs.

Finally, the synergies between vision and language training could lead to more contextual, logical, and human-like computer-generated text. Our model, however, still has limitations in optimising content selection and flow that provide opportunities for improvement. However, the success demonstrated in this medical application underscores the potential of multi-modal, multi-task learning for a wide range of domains.

## 7. Ethics Statement

We declare that there are no specific ethical concerns or broader impact considerations associated with our work. The patient data used in this study is sourced from an open-source dataset for which we have obtained the necessary permissions and adhered to data usage guidelines.

## 8. Bibliographical References

- Nurbanu Aksoy, Nishant Ravikumar, and Alejandro F Frangi. 2023. Radiology report generation using transformers conditioned with non-imaging data. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, volume 12469, pages 146–154. SPIE.
- Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2021. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32.
- Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*.
- Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. *arXiv preprint arXiv:1805.11004*.
- Xudong Hong, Khushboo Mehra, Asad Sayeed, and Vera Demberg. 2023. [Visually grounded story generation challenge](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 17–22, Prague, Czechia. Association for Computational Linguistics.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.
- Gang Liu, Pengfei Li, Zixu Zhao, Jinlong He, Genrong He, and Shenjun Zhong. 2023. Cross-modal self-supervised vision language pre-training with multiple objectives for medical visual question answering.
- Graciela Ramirez-Alonso, Olanda Prieto-Ordaz, Roberto López-Santillan, and Manuel Montes-Y-Gómez. 2022. Medical report generation through radiology images: an overview. *IEEE Latin America Transactions*, 20(6):986–999.
- Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 629–640.
- Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role



- labelling in memes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9763–9771.
- Uddagiri Sirisha and Bolem Sai Chandana. 2022. Semantic interdisciplinary evaluation of image captioning models. *Cogent Engineering*, 9(1):2104333.
- Roshan Adhithya SS, M Priyadharshini, and Lekshmi Kalinathan. 2023. Image caption generation for blind users of social media websites.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.
- Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *arXiv preprint arXiv:2109.14739*.
- Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.
- Jianfeng Wang, Shuokang Huang, Huifang Du, Yu Qin, Haofen Wang, and Wenqiang Zhang. 2022a. Mhkd-mvqa: Multimodal hierarchical knowledge distillation for medical visual question answering. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 567–574. IEEE.
- Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. 2022b. An inclusive task-aware framework for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–577. Springer.
- Xing Wu, Jingwen Li, Jianjia Wang, and Quan Qian. 2023. Multimodal contrastive learning for radiology report generation. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):11185–11194.
- Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
- Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.
- Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. *arXiv preprint arXiv:2204.03508*.