CXR-IRGen: An Integrated Vision and Language Model for the Generation of Clinically Accurate Chest X-Ray Image-Report Pairs

Junjie Shentu, Noura Al Moubayed
Durham University
junjie.shentu, noura.al-moubayed@durham.ac.uk

Abstract

Chest X-Ray (CXR) images play a crucial role in clinical practice, providing vital support for diagnosis and treatment. Augmenting the CXR dataset with synthetically generated CXR images annotated with radiology reports can enhance the performance of deep learning models for various tasks. However, existing studies have primarily focused on generating unimodal data of either images or reports. In this study, we propose an integrated model, CXR-IRGen, designed specifically for generating CXR image-report pairs. Our model follows a modularized structure consisting of a vision module and a language module. Notably, we present a novel prompt design for the vision module by combining both text embedding and image embedding of a reference image. Additionally, we propose a new CXR report generation model as the language module, which effectively leverages a large language model and self-supervised learning strategy. Experimental results demonstrate that our new prompt is capable of improving the general quality (FID) and clinical efficacy (AUROC) of the generated images, with average improvements of 15.84% and 1.84%, respectively. Moreover, the proposed CXR report generation model outperforms baseline models in terms of clinical efficacy (F_1 score) and exhibits a high-level alignment of image and text, as the best F_1 score of our model is 6.93% higher than the state-of-the-art CXR report generation model. Our code is available at https://github.com/junjie-shentu/CXR-IRGen.

1. Introduction

Medical imaging plays a crucial role in medical practice by providing spatially resolved information about organs, tissues, and bones. The chest X-Ray (CXR) image is the most common medical image due to its cost-effectiveness and low radiation dose. Notably, on average, 238 CXR images are acquired per 1000 of the population annually in industrialized countries, with 129 million CXR images acquired in the United States in 2006 [4]. However, the large number of CXR images increases the workload and diagnosis time, posing a challenge for radiologists. Deep learning techniques provide huge support to this issue by demonstrating promising performance in AI-assisted medical applications, including segmentation and diagnosis [26, 38]. Nonetheless, the availability of high-quality medical data is still limited due to privacy protocols and imbalanced data distribution, which further constrains the deployment of deep learning models in the medical field [19, 27, 40].

For this purpose, deep generative models are utilized to augment the CXR image dataset. Previous studies have demonstrated the generation of CXR images using deep generative models, including generative adversarial networks (GANs) and diffusion models [2, 3, 5, 6, 19, 21, 27, 31, 43]. CXR images are typically annotated with radiology reports detailing clinical observations made by radiologists, as depicted in Fig. 1. However, the majority of previous studies have primarily focused on generating high-quality CXR images, overlooking the importance of paired radiology reports. To the best of our knowledge, no study has yet addressed the feasibility of generating paired CXR images and radiology reports in a unified workflow. The generated CXR image-report pairs can significantly extend the applications of the augmented dataset and provide substantial support for training deep learning models that handle data from various modalities.

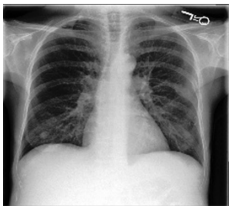
This work introduces *Chest X-Ray-Image Report Generation*(*CXR-IRGen*), an integrated model designed to generate CXR image-report pairs. In detail, *CXR-IRGen* is modularized and consists of a vision module and a language module (Fig. 2), providing high flexibility in generating multimodal CXR image-report pairs or unimodal images or reports. Furthermore, we evaluate the performance of *CXR-IRGen* on the test split of MIMIC-CXR dataset [18] and compare it with the baseline models concerning the general quality and clinical accuracy of the generated CXR image and report. Experimental results demonstrate that *CXR-IRGen* surpasses the baseline models in generating high-quality and clinically accurate CXR images and reports,

while ensuring clinical alignment of the generated image-report pairs. In summary, the contributions of our paper are as follows:

1. We propose *CXR-IRGen*, an integrated model that generates CXR image-report pairs based on a modularized structure comprising a vision module and a language module. The model supports multiple generative tasks, including the generation of unimodal images, reports, and multimodal image-report pairs.

2. We introduce a novel design of the prompt for the text-to-image diffusion model in the vision module by combining text embedding with image embedding of a reference image. The new prompt enhances the generation quality across different backbones of the diffusion model.

3. We propose a novel CXR report generation model as the language module, which utilizes a large language model and self-supervised learning strategy. The generated CXR reports exhibit promising performance in terms of both natural language metrics and clinical efficacy metrics.



FINDINGS:
There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.

IMPRESSION:
No acute cardiopulmonary process.

Figure 1. CXR image with radiology report

2. Related Work

2.1. Generative models for CXR image generation

In recent years, Generative Adversarial Networks (GANs) are frequently adopted for generating CXR images, and promising results were attained [2, 3, 19, 21, 27, 31, 39, 46]. Nonetheless, GANs exhibit problems including mode collapse and training instabilities, which increase training difficulties, and degrade generation quality. On the other hand, denoising diffusion models are proposed recently, which avoid these problems by adopting likelihood-based models and have been verified to outperform GANs in terms of the generation quality in general fields [10, 14, 32, 36]. In the medical domain, Chambon *et al*. [5, 6] sought the feasibility of adapting a pre-trained latent Diffusion Model [37](LDM) for generating CXR images, finding that fine-tuning the U-Net component of the LDM enables the domain adaption of a pre-trained LDM. They presented the *RoentGen* that can generate high-fidelity and diverse CXR images with radiology-specific text prompts. Packhäuser *et al*. [33] verified the performance of LDM in generating high-quality CXR images, and found that the images generated by LDM outperform those by PGGAN in an abnormality identification task. Weber *et al*. [43] proposed a cascaded LDM *Cheff* that can generate high-quality CXR images on a 1-megapixel scale. Based on the conclusions drawn by Chambon *et al*. [5,6], we adopt a pre-trained LDM as the backbone of the vision module, and attempt methods to further improve generation quality.

2.2. Generation of CXR reports

Many prior studies treat the generation of CXR reports as an image captioning task that generates natural language text conditioned on image input [25]. Image captioning models adopt an image encoder to extract information from the input image and a text decoder to synthesize corresponding text conditioned on the extracted vision information [41, 44]. Jing *et al*. [17] leveraged a CNN-RNN structure with a hierarchical LSTM [22] being the text decoder to generate corresponding descriptions and localize subregions. Xue *et al*. [45] used a stacked LSTM decoder in the CNN-RNN structure. Liu *et al*. [25] introduced a hierarchical generation strategy for CNN-RNN-RNN architecture, which enables the model to look at different parts of the image and enhance captioning accuracy. Ma *et al*. [29] introduced the contrastive attention mechanism that can better represent the visual features of abnormal regions. Chen *et al*. [7] proposed the memory-driven Transformer that uses transformers as backbones of the encoder and decoder. Based on Meshed-Memory Transformer ($\mathcal{M}^2$*Trans*) [9], Miura *et al*. [30] proposed two new rewards for capturing the factual completeness and report consistency, and optimized these rewards via reinforcement learning.

On the other hand, the presence of medically inconsistent and incoherent reports can still be frequently found in the reports generated by image captioning models [16]. Endo *et al*. [12] developed a retrieval-based CXR report generation method *CXR-RePaiR* that uses a Contrastive Language-Image Pre-training (CLIP [35]) model to retrieve the report with the highest similarity score. *CXR-RePaiR* gets a higher F_1 score than the baseline models, but much lower natural language metrics. Jeong *et al*. [16] also introduced a retrieval-based method *X-REM* that uses a novel image-text match score. Our work takes advantage of both the image captioning model and retrieval-based model, and applies a two-stage CXR report generation method in the language module, which further improves generation quality compared to the aforementioned models.

3. Method

The inference process of *CXR-IRGen* is depicted in Fig. 2. *CXR-IRGen* accomplishes a "label-to-image & report" task, taking the label from the MIMIC-CXR dataset as input, which alleviates the difficulties and complexities of input preparation. The input labels are subsequently converted into simple text to leverage the capabilities of the pre-trained CLIP text encoder. Simultaneously, a reference image with the same label is selected from the training set and encoded by a pre-trained CLIP image encoder. By combining the CLIP text and image embeddings, we obtain the conditional information for LDM sampling. The image embedding produced by the denoising backbone serves two purposes. First, it is decoded into the pixel space to create human-perceptible images. Additionally, it is projected into text embedding by a prior model for report generation. Consequently, we can obtain clinically accurate and aligned CXR image-report pairs by inputting simple labels.

3.1. Text-to-image generation and optimization with the diffusion model

The diffusion model is a probabilistic model specifically designed to describe the distribution of an observed sample $x_0 \sim q(x_0)$ by learning the reversal of a gradual and multi-step noising process, in which a Markov Chain of variables $x_1 \dots x_T$ is produced and expressed as [14, 28]:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{\alpha_t} x_{t-1}, \left(1-\alpha_t\right)\mathbf{I}\right) \quad (1)$$

where α is a noise schedule parameter. Furthermore, LDM applies the diffusion model in a latent space through the VAE (variational autoencoder) [20], which compresses the high-dimensional images into low-dimensional latent space. The denoising process is performed by a denoising backbone conditioned on the input information. The optimizing objective of LDM is given by:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, c\right)\|_2^2\right] \quad (2)$$

where x represents an input image $x \in \mathbb{R}^{H\times W\times 3}$ in pixel space, and c denotes the conditioning information. $\mathcal{E}$ is the VAE encoder, $t \in [1, T]$ is a timestep, and z_t is the image latent at timestep t of the Markov Chain. ε and ε_θ are standard Gaussian noise and predicted noise residue, respectively. In the vanilla LDM, the denoising backbone is a CNN-based U-Net consisting of down-sampling blocks and up-sampling blocks with skip connections between them. Besides, the feasibility of replacing the CNN layers with Vision Transformer (ViT) [11] was discussed, and a ViT-based backbone named U-ViT was proposed [1]. Following the conclusions drawn by Chambon *et al*. [5,6], we fine-tune the LDM on CXR images using a text-to-image approach to evaluate its domain-adapting performance. Both the U-Net and U-ViT backbones are involved and analyzed. To leverage the powerful capabilities of the pre-trained CLIP text encoder, we transform input labels into semi-structured text using the format of "*A chest X-Ray image with ..., without..., and unclear about ...*", where the three blanks are filled by pathology marked as 1.0, 0.0, and -1.0 in the label, respectively.

In text-to-image generation [37], the text prompts are projected into text embedding, and we additionally combine the CLIP reference image embedding of an image that shares the same label as the input label with the CLIP text embedding. We hypothesize that the inclusion of an additional reference image embedding is beneficial for generating high-quality CXR images, as the model can access more structural and semantic information from the input. Therefore, the optimization objective can be expressed as:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_t(y_t), \tau_i(y_i)\right)\|_2^2\right] \quad (3)$$

where y_t represents the input text, and y_i represents the reference image. τ_t and τ_i denote the CLIP text encoder and CLIP image encoder, respectively. Due to the difference in model architecture, the combination of the CLIP reference image embedding with the CLIP text embedding varies. For the U-Net backbone, we concatenate the image embedding and text embedding, while for the U-ViT backbone, we take the average value of them. Moreover, during the preparation of the reference image, we first search for a reference image with the same label in the training set. If none is found, then we search for an image with the same positive elements (marked as 1.0) but different negative elements (marked as -1.0) as the reference image.

The fine-tuning process follows the standard design of LDM fine-tuning and domain-adaption [5, 6] with the exception of the input design, as depicted in Fig. 3. We use a pre-trained Stable Diffusion model (checkpoint v1.4 [37]) with the U-Net backbone as the LDM, and a pre-trained U-ViT backbone [1].

3.2. CXR report generation with self-supervised learning

Image captioning models often exhibit inconsistency and incoherence between input images and generated reports, whereas retrieval-based models prioritize clinical accuracy, overlooking the consistency between retrieved and original reports [12]. We propose a two-stage CXR report generation method in the language module of *CXR-IRGen* that integrates the strengths of both models. In the first stage, we utilize a pre-trained large language model with an encoder-decoder architecture to process the CXR reports. Specifically, we encode the text into a sequence of text embedding and obtain the average value of all text embedding in the sequence as a representative text embedding. Subsequently,

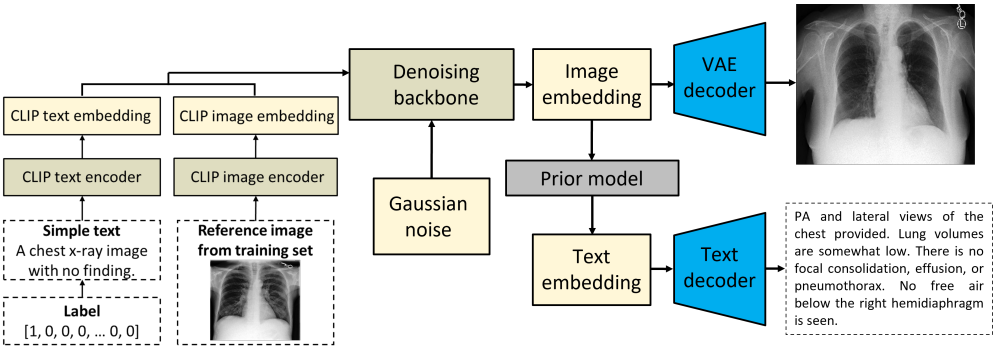


Figure 2. An overview of the inference process of *CXR-IRGen*

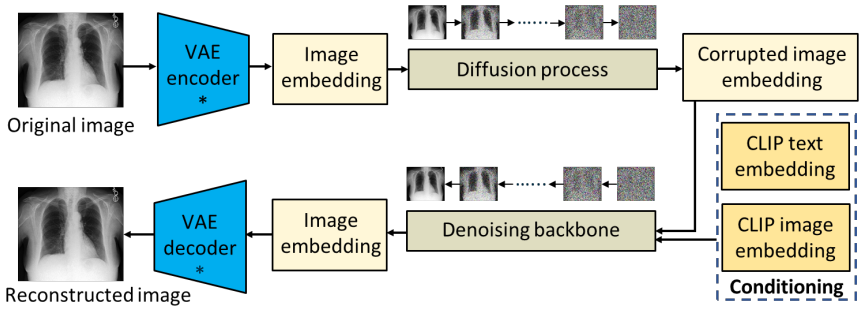


Figure 3. Illustration of the training process of the vision module (* denotes the frozen part)

we use this representative text embedding as the prompt for the decoder to reconstruct the input text. The loss function is calculated as the cross-entropy between the original and reconstructed text, expressed as:

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i) \tag{4}$$

where t_i and p_i are ith elements of the original and reconstructed text, respectively. n denotes the total sequence length.

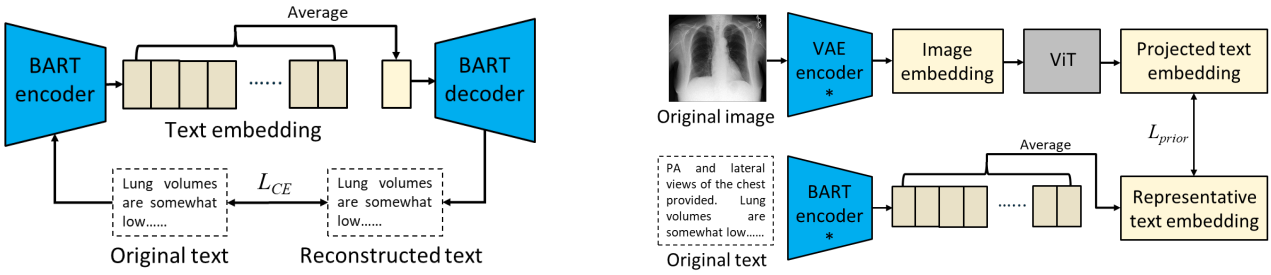
In the second stage, a prior model is employed to project the image embedding produced by the vision module into the corresponding text embedding. The training objective is to minimize the mean squared error and maximize the cosine similarity between the original and reconstructed text embeddings, which is given by:

$$L_{prior} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda\left[1 - \frac{\sum_{i=1}^{m} y_i \hat{y}_i}{\sum_{i=1}^{m}(y_i)^2 \sum_{i=1}^{m}(\hat{y}_i)^2}\right] \tag{5}$$

where y_i represents the text embedding projected by the prior model, and $\hat{y}_i$ represents the text embedding encoded from the input text. m is the dimension of text embedding, and λ is a scaling coefficient that aligns the magnitude of the cosine similarity with that of the mean squared error, set at 0.01 for this study. Other options for L_{prior} will be discussed in the ablation tests in Sec. 5.3.

Specifically, the first stage resembles the image captioning models that recurrently produce a sequence of text. However, in our approach, we utilize highly summarized text information from the representative text embedding as a prompt for the decoder, rather than using vision information extracted from images. This design enhances the consistency between the original and generated reports compared to retrieval-based models. Similar to contrastive learning, which is commonly used in retrieval-based models, the second stage operates on image and text embeddings. Both the image encoder and text encoder are pre-trained and frozen. Instead of comparing the image embedding and text embedding based on cosine similarity, we employ a prior model to directly project and match the image and text embedding pair using a novel loss function Eq. (5) under self-supervised learning, thereby strengthening their alignment. This approach ensures that the generated report exhibits high consistency with both the image and the original report.

For the large language model, we select Bidirectional and Auto-Regressive Transformers (BART [23]) as the backbone, and for the prior model, we utilize ViT as the backbone. The training process of both stages is depicted in Fig. 4.



(a) First stage

(b) Second stage

Figure 4. Illustration of the training process of the language module during (a) first stage and (b) second stage (* denotes the frozen part)

4. Experiments

4.1. Dataset

In this study, we use MIMIC-CXR [18] for training and evaluation. MIMIC-CXR is a publicly available large-scale dataset consisting of 377,110 images and 227,943 reports from 225,000 studies. Following Chambon *et al*. [5], we extract images in the "PA" (postero-anterior) view position from the training set to fine-tune the vision module. For training the language module, we extract the findings and impression sections separately from all reports in the first stage. In the second stage, we select images in the "PA" view position from each study that contains a report, and if "PA" is inapplicable, we consider images in the "AP" (antero-posterior) view position, as they are also taken from a frontal view and present the same content to those in the "PA" view position but in a mirrored position. All the extracted images are matched with the reports to form a dataset of image-report pairs.

For model testing, we utilize the official testing split of the MIMIC-CXR dataset. We randomly extracted 1000 images in the "PA" view position to evaluate the vision module. Subsequently, we select images in the "PA" or "AP" view position that are paired with reports and extract findings and impression sections, resulting in 2608 image-findings/impression pair samples and 1460 image-findings pair samples. The former is adopted to evaluate the clinical efficacy of generated reports, while the latter is employed to evaluate the natural language metrics.

4.2. Baselines and evaluation metrics

We conduct a comparative analysis between the vision module of *CXR-IRGen* and the vanilla Stable Diffusion model. Additionally, we compare the effects of different backbones fine-tuned with and without the CLIP reference image embedding. For the text module of *CXR-IRGen*, we employ three CXR report generation models that have been tested on MIMIC-CXR, including two image captioning models, namely, *R2Gen* [7] and $\mathcal{M}^2$*Trans* [30], as well as one retrieval-based model *CXR-RePaiR* [12]. Particularly, we re-implement *R2Gen* and $\mathcal{M}^2$*Trans* using publicly available code and checkpoints, and we cite the results of *CXR-RePaiR* from the original paper.

For the generated CXR images, the general quality is evaluated using image quality metrics, including Fréchet Inception Distance (FID) [13], Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) [42]. The clinical efficacy is assessed by the Area Under the Receiver Operating Characteristic (AUROC) value calculated in binary classification tasks on CXR images with and without specific pathologies. Moreover, the quality of the generated CXR reports is assessed using conventional natural language metrics, including BLUE [34] and ROUGE-L [24], alongside the clinical efficacy metric F_1 score. The F_1 score is calculated based on the labels generated by the CheXpert [15] labeler for both the original and generated reports.

4.3. Evaluation of CXR images

The image labels from the testing set are used as input to generate 1000 images for evaluation. The general image quality metrics are presented in Tab. 1. Compared to the vanilla LDM, all three metrics exhibit improvements after fine-tuning, confirming that fine-tuning on domain-specific data contributes to domain adaptation. When solely taking text embedding as input, the U-Net backbone variant fine-tuned for 5k steps outperforms the one fine-tuned for 10k steps. In contrast, for the U-ViT backbone, the variant fine-tuned for 5k steps demonstrates a better FID score but worse PSNR and SSIM scores compared to the variant fine-tuned for 10k steps. Furthermore, we investigate the effect of the reference image embedding, which shows an overall improvement in the generation quality of the U-Net backbone. As for the U-ViT backbone, the reference image embedding improves the FID score but slightly degrades the PSNR and SSIM scores. These different effects on general metrics could be attributed to the way we combine CLIP text embedding and reference image embedding, as taking the average value of the text embedding and image embedding may induce information loss.

We employ the U-Net backbone for investigating the

Table 1. General metrics of CXR images generated by different models (RIE: reference image embedding)

Model	FID↓	PSNR↑	SSIM↑
Baseline			
Vanilla LDM	303.4451	6.7723	0.9734
LDM with the U-Net backbone			
5k steps without RIE	54.0164	10.9598	0.9889
5k steps with RIE	**49.5479**	**11.2136**	**0.9897**
10k steps without RIE	59.8236	10.3455	0.9873
10k steps with RIE	53.1351	10.4316	0.9875
LDM with the U-ViT backbone			
5k steps without RIE	64.4917	11.1186	0.9896
5k steps with RIE	**43.4003**	10.4192	0.9876
10k steps without RIE	54.5434	**11.1798**	**0.9897**
10k steps with RIE	47.8233	10.5437	0.9878

clinical efficacy of the generated CXR images, considering its clear tendency and superior robustness in analyzing image general metrics, as elaborated in Tab. 1. To evaluate clinical efficacy, we select five pathologies, namely, *Atelectasis, Cardiomegaly, Lung opacity, Effusion*, and *Pneumonia* as positive labels, while *No finding* serves as the negative label. Each label is used to generate 500 CXR images, which are grouped together, resulting in five sub-testing sets, each containing 500 positive samples and 500 negative samples. Subsequently, a pre-trained classification model (DenseNet-121, XRV [8]) is applied to perform a binary classification task on each sub-testing set, and the AUROC value is calculated to assess the classification accuracy, with results presented in Tab. 2. It is observed that CXR images generated by vanilla LDM exhibit the worst performance, as all AUROC values are close to 0.5. Following fine-tuning, the AUROC values for all pathologies improve, and variants fine-tuned with reference image embedding achieve higher AUROC values than those without reference image embedding by an average value of 1.84%, indicating that the additional CLIP reference image embedding enhances clinical characteristics. Notably, the variant fine-tuned for 10k steps generates CXR images with higher AUROC scores than the original images extracted from the training set. This implies potential overfitting as the model might learn certain features highly discriminative to the XRV, therefore the training steps should be prudently designed, but the effect of the reference image embedding can still be reflected as the mean AUROC is improved by 1.89% for this variant.

4.4. Evaluation of CXR reports

We conduct a performance comparison of the language module of *CXR-IRen* with the baseline models. The evaluation results, presented in Tab. 3, are based on the original CXR images from the testing set. Unless specified otherwise, both the original and generated CXR reports refer to the findings section. We introduce two variants, namely *CXR-IRGen (F)* trained solely on the findings section, and *CXR-IRGen (F+I)* trained jointly on the findings and impression sections. The natural language metrics are evaluated using only the former, while both variants are used for assessing clinical efficacy. In comparison to the retrieval-based model *CXR-RePaiR*, *CXR-IRGen* demonstrates a dramatic improvement in BLUE-2 score. As the CXR reports in the dataset are highly diverse, the reports retrieved by *CXR-RePaiR* are clinically matched with images but usually different from the originals. On the other hand, *CXR-IRGen* learns the common textual description of the images in the same class and achieves good proficiency in generating reports consistent with the originals, resulting in the highest BLUE-1 score among all models, with the other four natural language metrics being on par with those of *R2Gen* but slightly below those of $\mathcal{M}^2$*Trans*. However, it should be noted that $\mathcal{M}^2$*Trans*'s image encoder is additionally trained on the CheXpert dataset [15], which may enhance CXR report generation quality and leads to unfair comparison. Furthermore, *CXR-IRGen* exhibits exceptional clinical accuracy, with the variant *CXR-IRGen (F+I)* achieving the highest F_1 score among all the models.

We also compare the clinical efficacy of all models on the CXR images generated by the vision module of *CXR-IRGen*. We utilize the 3000 generated CXR images introduced in Sec. 4.3 for report generation. The evaluation results are provided in Tab. 4. It is evident that *CXR-IRGen (F+I)* outperforms all other models in terms of clinical efficacy on the generated CXR images. While *CXR-IRGen (F)* demonstrates superior clinical efficacy to *R2Gen*, it falls short compared to $\mathcal{M}^2$*Trans*. This difference can be attributed to the fact that $\mathcal{M}^2$*Trans* employs an image encoder that is additionally trained on the CheXpert dataset [15], which aids in feature recognition and representation. The impact of reference image embedding on clinical efficacy is also reflected. For $\mathcal{M}^2$*Trans* and *CXR-IRGen*, the F_1 scores are higher on the CXR images generated by the vision module trained with reference image embedding by an average value of 6.58%.

5. Ablation

We conduct an analysis of various design choices in *CXR-IRGen* that might affect the generation quality, including (1) the strategy of extracting the representative text embedding; (2) utilizing the image or image embedding for report generation; and (3) different options for L_{prior}. Note that all the variants discussed in this section are trained using the findings section of the CXR report.

Table 2. AUROC values of the binary classification task on original CXR images and CXR images generated by different models (RIE: reference image embedding)

Source	Atelectasis	Cardiomegaly	Lung opacity	Effusion	Pneumonia	Mean
Baseline						
Original	0.7799	0.8197	0.8081	0.8921	0.7127	0.8025
Vanilla LDM	0.5504	0.5378	0.5876	0.5785	0.5458	0.5600
Proposed approach (U-Net backbone)						
5k steps without RIE	0.6303	0.7284	0.6397	0.8128	0.5769	0.6776
5k steps with RIE	0.6470	0.7326	0.6605	0.8126	0.5956	0.6897
10k steps without RIE	**0.8897**	0.9800	0.8938	0.9867	0.8267	0.9150
10k steps with RIE	0.8688	**0.9836**	**0.9537**	**0.9953**	**0.8602**	**0.9323**

Table 3. Comparison of *CXR-IRGen* and baselines models on original CXR images (Results with * are taken from the original paper [12])

Model	BLUE-1↑	BLUE-2↑	BLUE-3↑	BLUE-4↑	ROUGE-L↑	F_1 score↑
Baseline						
*CXR-RePaiR-2** [12]	-	0.0690	-	-	-	0.2560
*CXR-RePaiR-Select** [12]	-	0.0500	-	-	-	0.2740
R2Gen [7]	0.2870	0.1651	0.1072	0.0726	0.2093	0.1716
$\mathcal{M}^2$Trans [30]	0.3174	**0.1917**	**0.1195**	**0.0734**	**0.2252**	0.2665
Proposed approach						
CXR-IRGen (F)	**0.3200**	0.1760	0.1066	0.0669	0.2080	0.2695
CXR-IRGen (F+I)						**0.2930**

5.1. Extracting representative text embedding

During the first training stage of the language module in *CXR-IRGen*, we select a representative text embedding from a sequence of text embedding and use this representative embedding as input for the BART decoder. The goal is to ensure that the representative text embedding captures as much semantic information as possible. Several strategies for extracting the representative text embedding are considered, including using the text embedding of the `[BOS]` (beginning of sentence) token, the text embedding of the `[EOS]` (end of sentence) token, or the averaged text embedding of all tokens. The reconstruction quality is evaluated using different representative text embeddings, and the results are presented in Fig. 5. Notably, the averaged text embedding of all tokens outperforms the other strategies in terms of BLUE and ROUGE-L scores, displaying higher scores and a more consistent increase during the training process.

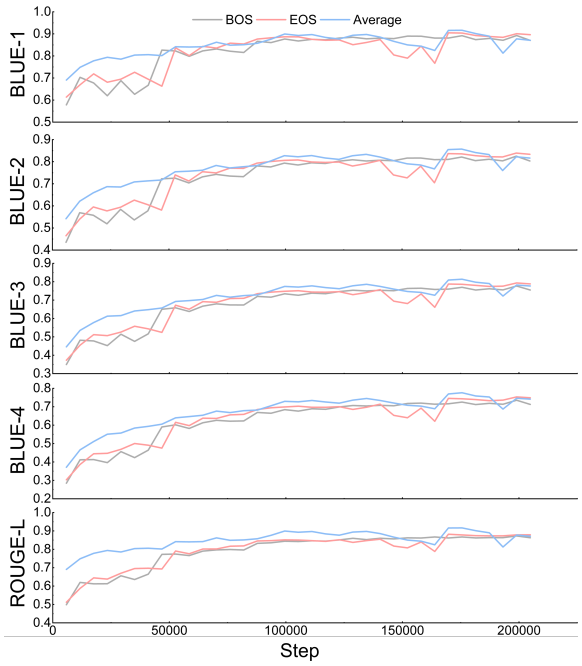


Figure 5. Comparison of different representative text embedding

5.2. Image vs. Image embedding

In the language module of *CXR-IRGen*, we utilize image embeddings as input for the prior model, whereas image captioning models typically take images directly as input. To compare the generation quality, we evaluate the performance using both image embeddings and images, and the

Table 4. F_1 scores of *CXR-IRGen* and baselines models on CXR images generated by the vision module *CXR-IRGen* (RIE: reference image embedding)

Model	5k steps without RIE	5k steps with RIE	10k steps without RIE	10k steps with RIE
Baseline				
R2Gen [7]	0.1095	0.1102	0.1895	0.1794
$\mathcal{M}^2$*Trans* [30]	0.2328	0.2347	0.3226	**0.3738**
Proposed approach				
CXR-IRGen (F)	0.2157	0.2280	0.3410	0.3627
CXR-IRGen (F+I)	**0.2390**	**0.2543**	**0.3603**	0.3719

Table 5. Comparison of different variants of *CXR-IRGen*

Model	BLUE-1↑	BLUE-2↑	BLUE-3↑	BLUE-4↑	ROUGE-L↑	F_1 score↑
Proposed approach						
CXR-IRGen (F)	**0.3200**	**0.1760**	**0.1066**	**0.0669**	**0.2080**	**0.2695**
Change input						
Input image	0.3096	0.1658	0.0975	0.0594	0.1996	0.2617
Change loss function						
Mean square error	0.3070	0.1616	0.0934	0.0558	0.1951	0.2433
Cosine similarity	0.0206	0.0055	0.0022	0.0008	0.0292	0.0696

results are detailed in Tab. 5. We observe that employing image embedding as input leads to higher scores in both natural language metrics and clinical efficacy metrics compared to using raw images, suggesting that the encoding process meaningfully compresses image information, emphasizing relevant details crucial for feature extraction and recognition by the prior model. Notably, this is consistent with the observation in image generation tasks reported by Weber *et al*. [43], who concluded that semantic features are more beneficial for a cascaded diffusion model in generating high-quality and high-resolution CXR images compared to low-resolution images.

5.3. Choice of L_{prior}

The prior model within the language module of *CXR-IRGen* is responsible for learning a projection from image embeddings to text embeddings. To achieve this, it is crucial to minimize the distance between the target embedding space of this projection and the pre-determined text embedding space. In the process of measuring this distance, several options are available, including the mean square error, the cosine similarity, or a combination of both, as shown in Eq. (5). The mean square error quantifies the Euclidean distance between two vectors, while the cosine similarity measures the angle between them. We trained the prior model with each of these metrics as loss functions, and the resulting model performances are presented in Tab. 5. The outcomes indicate that combining the mean square error and cosine similarity yields the best result. Particularly, solely using cosine similarity as the loss function severely limits model performance, but its inclusion alongside the mean square error with a coefficient λ that balances their values significantly improves performance.

6. Conclusion

In this study, we introduce an integrated model called *CXR-IRGen* designed for generating high-quality CXR image-report pairs. *CXR-IRGen* comprises a vision module for generating CXR images and a language module for generating corresponding reports. These modules can either be utilized together to produce CXR image-report pairs or independently to generate CXR images or reports separately. The vision module incorporates a novel prompt design for the text-to-image LDM by combining text embedding with a reference image embedding, which enhances the general quality and clinical efficacy of the generated CXR images. For the language module, we propose a new CXR report generation model that benefits from both image captioning and retrieval-based approaches, leveraging a large language model and self-supervised learning strategy. The proposed report generation model demonstrates the ability to produce coherent, consistent CXR reports, and it outperforms baseline models in terms of clinical efficacy. Furthermore, the CXR image-report pairs generated by *CXR-IRGen* exhibit a high level of clinical alignment.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22669–22679, 2023. 3

[2] Vedant Bhagat and Swapnil Bhaumik. Data augmentation using generative adversarial networks for pneumonia classification in chest xrays. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 574–579. IEEE, 2019. 1, 2

[3] Swathi Buragadda, Kodali Sandhya Rani, Sandhya Venu Vasantha, and M Kalyan Chakravarthi. Hcugan: Hybrid cyclic unet gan for generating augmented synthetic images of chest x-ray images for multi classification of lung diseases. *International Journal of Engineering Trends and Technology*, 70(2):229–238, 2022. 1, 2

[4] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021. 1

[5] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 1, 2, 3, 5

[6] Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022. 1, 2, 3

[7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 2, 5, 7, 8

[8] Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022. 6

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 2, 3, 5, 7

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 5, 6

[16] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023. 2

[17] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017. 2

[18] Alistair EW 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. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 1, 5

[19] Yash Karbhari, Arpan Basu, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. Generation of synthetic chest x-ray images and detection of covid-19: A deep learning based approach. *Diagnostics*, 11(5):895, 2021. 1, 2

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[21] Sagar Kora Venu and Sridhar Ravula. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet*, 13(1):8, 2020. 1, 2

[22] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017. 2

[23] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 4

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[25] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. 2

[26] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. *arXiv preprint arXiv:2301.00785*, 2023. 1

[27] Mohamed Loey, Florentin Smarandache, and Nour Eldeen M. Khalifa. Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning. *Symmetry*, 12(4):651, 2020. 1, 2

[28] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. 3

[29] Xuewei Ma, Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*, 2021. 2

[30] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020. 2, 5, 7, 8

[31] Saman Motamed, Patrik Rogalla, and Farzad Khalvati. Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images. *Informatics in Medicine Unlocked*, 27:100779, 2021. 1, 2

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[33] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. *arXiv preprint arXiv:2211.01323*, 2022. 2

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[39] Ilyas Sirazitdinov, Maksym Kholiavchenko, Ramil Kuleev, and Bulat Ibragimov. Data augmentation for chest pathologies classification. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 1216–1219. IEEE, 2019. 2

[40] Amirsina Torfi and Edward A Fox. Corgan: correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. *arXiv preprint arXiv:2001.09346*, 2020. 1

[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[43] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 180–191. Springer, 2023. 1, 2, 8

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[45] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer, 2018. 2

[46] Tianyang Zhang, Huazhu Fu, Yitian Zhao, Jun Cheng, Mengjie Guo, Zaiwang Gu, Bing Yang, Yuting Xiao, Shenghua Gao, and Jiang Liu. Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 777–785. Springer, 2019. 2