Vision-Language Generative Model for View-Specific Chest X-ray Generation

Hyungyung Lee TTUMYCHE@KAIST.AC.KR
KAIST, Republic of Korea

Da Young Lee* DYAN.LEE717@GMAIL.COM
Deep-in-Sight Co., Republic of Korea

Wonjae Kim WONJAE.KIM@NAVERCORP.COM
NAVER AI Lab, Republic of Korea

Jin-Hwa Kim J1NHWA.KIM@NAVERCORP.COM
NAVER AI Lab, Republic of Korea
AI Institute of Seoul National University, Republic of Korea

Tackeun Kim TACKEUN.KIM@SNU.AC.KR
Jihang Kim RADIO622@GMAIL.COM
Leonard Sunwoo LEONARD.SUNWOO@GMAIL.COM
Seoul National University Bundang Hospital, Republic of Korea

Edward Choi EDWARDCHOI@KAIST.AC.KR
KAIST, Republic of Korea

Abstract

Synthetic medical data generation has opened up new possibilities in the healthcare domain, offering a powerful tool for simulating clinical scenarios, enhancing diagnostic and treatment quality, gaining granular medical knowledge, and accelerating the development of unbiased algorithms. In this context, we present a novel approach called ViewXGen, designed to overcome the limitations of existing methods that rely on general domain pipelines using only radiology reports to generate frontal-view chest X-rays. Our approach takes into consideration the diverse view positions found in the dataset, enabling the generation of chest X-rays with specific views, which marks a significant advancement in the field. To achieve this, we introduce a set of specially designed tokens for each view position, tailoring the generation process to the user's preferences. Furthermore, we leverage multi-view chest X-rays as input, incorporating valuable information from different views within the same study. This integration rectifies potential errors and contributes to faithfully capturing abnormal findings in chest X-ray generation. To validate the effectiveness of our approach, we conducted statistical analyses, evaluating its performance in a clinical efficacy metric on the MIMIC-CXR dataset. Also, human evaluation demonstrates the remarkable capabilities of ViewXGen, particularly in producing realistic view-specific X-rays that closely resemble the original images.

Data and Code Availability We use the MIMIC-CXR dataset, which is available on the PhysioNet repository (Johnson et al., 2019). Our implementation code is available at this repository[1].

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Chest X-ray generation has become increasingly significant in the medical field, yet prior studies (Packhäuser et al., 2022; Chambon et al., 2022b,a) have notably missed two crucial aspects: First, there's a heavy reliance on radiology reports for generating chest X-rays, which disregards the rich information available in other X-ray views within the same study. Second, the importance of controlling view positions has been neglected, despite the fact that var-

* Work done at KAIST

1. https://github.com/ttumyche/UniXGen

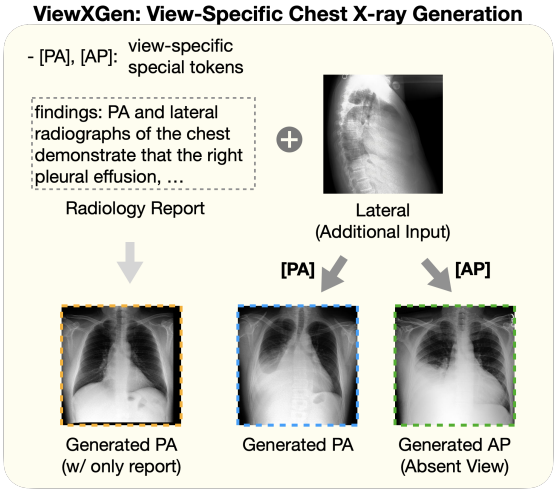

Figure 1: We introduce a view-specific chest X-ray generation model. ViewXGen leverages view-specific special tokens to empower its ability to capture unique features from different views. Additionally, the integration of multi-view chest X-rays as input enhances the overall generation quality.

ious views reveal diverse characteristics due to the angle of the X-ray beam (Puddy and Hill, 2007).

To address these, we introduce ViewXGen, a versatile generative model tailored for generating synthetic chest X-rays that are specific to view, symptom, and patient. Figs. 3, and 5 showcase detailed examples of the generated chest X-rays by our model, demonstrating its capabilities. This approach sets our work apart from earlier studies, showcasing a wide range of clinical applications: 1) Filling in Missing Data: Our model can address gaps by generating specific views that may have been mentioned in a report but are currently missing. Moreover, it enriches the generated images with patient information observed in other views, including gender, age, and obesity level. Upon investigating the presence of missing data in MIMIC-CXR (Johnson et al., 2019), it was discovered that among 27,859 studies where specific views were explicitly mentioned in the reports, 1,565 of these studies (5.62%) did not contain the mentioned views. 2) Reducing the Need for Additional Imaging: Our model provides a solution for scenarios where obtaining certain views is impractical due to patient conditions or limitations in medical equipment. By generating the necessary views, it conserves both time and resources, offering a way to acquire patient-specific images without additional imaging. 3) Enhancing Education and Training: The ability to create and analyze customized views and patient cases empowers medical students and professionals. This feature aids in deepening the understanding of how various conditions manifest across different X-ray views, thereby improving diagnostic capabilities and expanding anatomical knowledge. 4) Augmenting Data for Rare Conditions: Our model excels in generating images for a wide range of scenarios, including plausible yet rare conditions, enriching datasets with unique views that spotlight uncommon pathologies and aiding in the research and diagnosis of rare conditions.

To achieve these, we introduce a set of special tokens tailored to each view position, including posterior-anterior (PA), anterior-posterior (AP), and lateral views, and employ a simplified architectural design by combining VQ-GAN (Esser et al., 2021) and Performer (Choromanski et al., 2020), which is an efficient Transformer-based framework. Specifically, we utilize VQ-GAN as an image tokenizer, enabling the conversion of chest X-ray images into sequences of discrete tokens. The adoption of Performer enhances computational and memory efficiency, crucial for processing long paragraph reports and high-resolution multi-view chest X-rays that result in long-range sequences. By leveraging this approach, our model demonstrates the capability to handle diverse input formats, ranging from single to multi-view images.

We evaluate our model on MIMIC-CXR (Johnson et al., 2019). The experimental results show that ViewXGen achieves better performance on both standard metrics such as FID (Huang et al., 2017) and clinical efficacy metrics such as 14-diagnosis classification over several baselines. Furthermore, human evaluation shows that ViewXGen can generate realistic chest X-rays comparable to the original image, and the view-specific special tokens capture the refined features of each view, encouraging the model to generate appropriate view-specific X-rays.

Our contributions can be summarized as follows.

1) **Pioneering Approach**: Our work marks the first attempt to generate view-specific chest X-ray images with multimodal input in the medical domain. Additionally, we introduce special tokens that are simple yet effective for generating specific view positions. These tokens provide precise control over the view generation process, enabling our model to produce X-rays from various view positions.

2) **Novel Task**: We propose a novel task of generating chest X-rays with specific views, such as PA, AP, and Lateral views. This task addresses the limitations of previous approaches that primarily focused on generating frontal views and disregarded the multi-view nature of the dataset.
3) **Multi-View Integration**: By leveraging multi-view chest X-rays, our model demonstrates the potential to generate more accurate chest X-rays that capture abnormal findings and patient characteristics present in additional X-rays. This integration of multi-view information improves the fidelity and diagnostic quality of the generated chest X-rays.

2. Related Works

2.1. Chest X-ray Generation

With the growing demand to access high quality medical data and the success of generative models such as GANs (Goodfellow et al., 2020), and diffusion models (Ho et al., 2020), chest X-ray generation has gained a lot of attention. Chambon et al. (Chambon et al., 2022b) and Packhauser et al. (Packhäuser et al., 2022) adopt a latent diffusion model (Rombach et al., 2022) for class-conditional generation. However, these works only focus on specific diseases and do not utilize radiology reports that contain rich medical domain knowledge. Recently, Chambon et al. (Chambon et al., 2022a) have taken advantage of radiology reports for conditional generation, but they only use the impression section of the reports. Furthermore, they cannot generate view-specific chest X-rays or accept multiple views as input.

2.2. Image Tokenization

Many efforts have been made to convert images into discrete tokens like natural language, as this provides a compact and efficient representation compared to using raw pixels. Based on the success of VQ-VAE (Van Den Oord et al., 2017), Esser et al. (Esser et al., 2021) introduced VQ-GAN with a discriminator and a perceptual loss for high-resolution images. Recently, diffusion models have achieved promising performance in generating high-quality samples in continuous domains (*e.g.*, image (Ramesh et al., 2022a) and audio (Saharia et al., 2022)). However, the models are not flexible to take arbitrary input from single to multiple images.

2.3. Efficient Transformer

Transformer (Vaswani et al., 2017) has proven to be highly adaptable to both vision and language tasks with its task-agnostic design and generalization capabilities. However, the self-attention mechanism increases the computational and memory cost quadratically by the input sequence length. As we utilize long paragraph reports and high resolution multi-view chest X-rays, we adopt Performer (Choromanski et al., 2020), an efficient Transformer-based model to reduce the quadratic complexity to linear. They approximate the standard Transformer attention using positive orthogonal random features to kernelize the softmax operation.

3. Method

Fig. 2 shows the overall depiction of ViewXGen. Notably, 1) ViewXGen leverages a series of chest X-rays and a corresponding report from the same study as input, enhancing the quality of the generated chest X-rays. 2) To enable precise control over the generation of chest X-rays with specific views, we integrate special tokens tailored to each view type.

3.1. Input Embedding

3.1.1. Image Tokenization

We first train VQ-GAN (Esser et al., 2021) to encode chest X-rays into a discrete latent space, enabling us to represent each image as a sequence of discrete tokens. This model consists of an encoder E, a decoder G, and a fixed-size learnable codebook $C = \{e_m\}_{m=1}^{M}$ of size M, where $e_m \in \mathbb{R}^n$. Given an image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, the encoder encodes the input image into a continuous feature map $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^{h\times w\times n}$. Then, we obtain a quantized feature map $\hat{\mathbf{z}} \in \mathbb{R}^{h\times w\times n}$ and its sequence of visual tokens $\{v_1, \ldots, v_{h\times w}\}$, *a.k.a.*, discrete codes as follows:

$$\hat{\mathbf{z}}_{ij} = Q(\mathbf{z}_{ij}) = e_m, \quad m = \arg\min_k \|\mathbf{z}_{ij} - e_k\| = v_{ij}$$

where $Q(\cdot)$ denotes an element-wise quantization operation that performs the nearest neighbor search, $\mathbf{z}_{ij} \in \mathbb{R}^n$ is a feature vector at (i, j), and v_{ij} is its code. The decoder then maps the quantized feature map back to the original input $\hat{\mathbf{x}} = G(\hat{\mathbf{z}}) \in \mathbb{R}^{H\times W\times 3}$.

The encoder-decoder model and codebook are optimized using the following objectives:

$$L_{VQ}(E, G, C) = \|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 + \|sg[\mathbf{z}]-\hat{\mathbf{z}}\|_2^2 + \beta\|sg[\hat{\mathbf{z}}]-\mathbf{z}\|_2^2$$

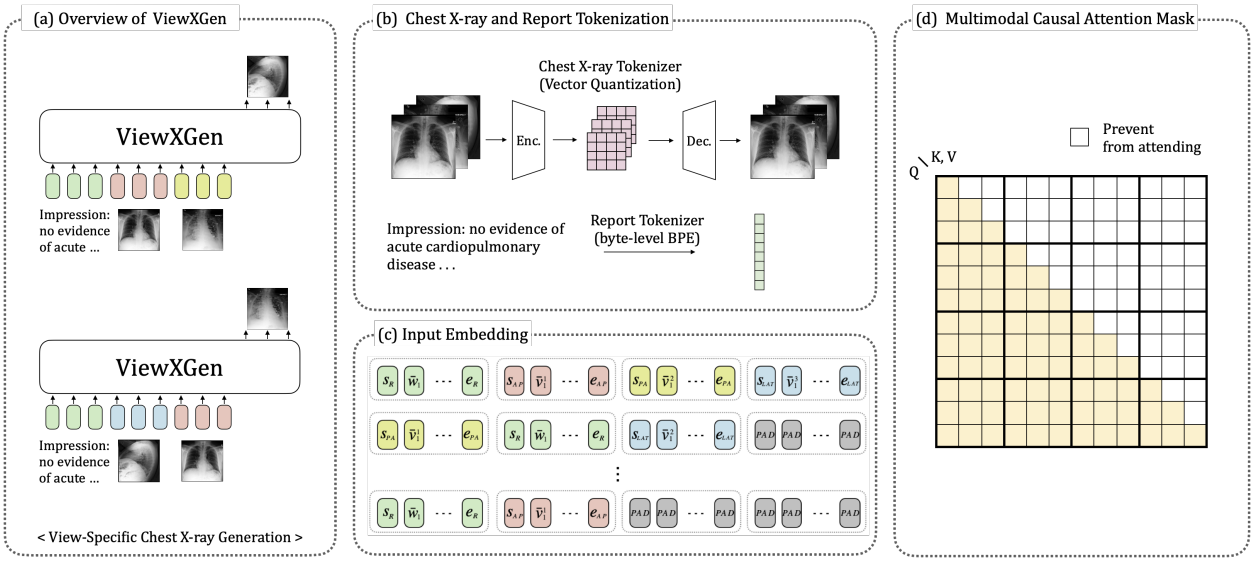

Figure 2: Overview of ViewXGen architecture. (a) ViewXGen is designed to generate chest X-rays with specific views, such as AP, PA, and Lateral views. (b) Images are tokenized via VQ-GAN, and reports are tokenized via a byte-level BPE tokenizer. (c) A minibatch consists of input sequences consisting of AP/PA/Lateral X-rays and a report in random order. (d) We use a causal attention mask to simultaneously handle multi-view X-rays and a report.

where the first term is a reconstruction loss, the second term optimizes the codebook embedding, the last term refers to a commitment loss with weighting factor β, and sg refers to a stop-gradient. To further enhance the reconstruction quality, VQ-GAN incorporates a discriminator D and perceptual loss as follows:

$$L_{GAN}(\{E, G, C\}, D) = [\log D(\mathbf{x}) + \log(1 - D(\hat{\mathbf{x}})]$$

Finally, the model is optimized as follows:

$$L_{VQGAN} = L_{VQ}(E, G, C) + \lambda L_{GAN}(\{E, G, C\}, D)$$

where λ is an adaptive weight. This method allows the model to learn a compact and discrete representation of the images.

3.1.2. Chest X-ray Embedding

Using the image tokenizer described above, chest X-rays of multiple views from the same study are individually tokenized into a sequence of discrete visual tokens, surrounded by special tokens to differentiate between different views, *e.g.* $\{[SOS_{PA}], v_1, \ldots, v_{h\times w}, [EOS_{PA}]\}$ for a PA-view X-ray. Additionally, if the study has fewer images than k[2], we add padding tokens to ensure that all input sequences have the same length. For example, the final embeddings of a PA-view X-ray is $\mathbf{v}_{PA} = \{\mathbf{s}_{PA}, \bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_{h\times w}, \mathbf{e}_{PA}\}$, where $\mathbf{s}_{PA}, \mathbf{e}_{PA} \in \mathbb{R}^d$ respectively denote the embeddings of the special tokens, $\bar{\mathbf{v}}_i \in \mathbb{R}^d$ is acquired by summing the visual embedding and axial positional embedding (Ho et al., 2019; Kitaev et al., 2020):

$$\bar{\mathbf{v}}_i = f_{VE}(v_i) + f_{VP}(i)$$

where $f_{VE}(\cdot)$ and $f_{VP}(\cdot)$ are the visual embedding and axial positional embedding functions, respectively.

3.1.3. Radiology Report Embedding

We first split a report into word tokens with a byte-level BPE tokenizer (Wang et al., 2020) and surround them with special tokens, *e.g.* $\{[SOS], w_1, \ldots, w_T, [EOS]\}$. The final embeddings for the report is $\mathbf{w} = \{\mathbf{s}_R, \bar{\mathbf{w}}_1, ..., \bar{\mathbf{w}}_T, \mathbf{e}_R\}$, where $\mathbf{s}_R, \mathbf{e}_R \in \mathbb{R}^d$ respectively denote the embeddings of the special tokens, $\bar{\mathbf{w}}_i \in \mathbb{R}^d$ is obtained by summing up the word embedding and sinusoid positional embedding:

$$\bar{\mathbf{w}}_i = f_{WE}(w_i) + f_{WP}(i)$$

where $f_{WE}(\cdot)$ and $f_{WP}(\cdot)$ are the word embedding and sinusoidal positional embedding functions, respectively.

2. In our work, we use $k = 3$ to include PA, AP, and Lateral view.

3.2. Multi-view Chest X-ray Generative Model

We design a model for multi-view chest X-ray generation by treating the task as a sequence generation task. Incorporating the Transformer architecture (Vaswani et al., 2017), our model is trained with a multimodal causal attention mask, which is designed to handle multimodal input while still maintaining the causal constraints of the standard causal mask as shown in Fig. 2 (d). The attention mask $M \in R^{S\times S}$ can be represented as follows:

$$M_{ij} = \begin{cases} 0, & \text{if i} \leq \text{j} \\ -\infty, & \text{otherwise} \end{cases} \qquad i, j = 1, ..., S.$$

where a value of 0 indicates allow to attend, while $-\infty$ prevents from attending, and $S = k \times (h \times w + 2) + T + 2$. This attention mechanism differs from the sequence-to-sequence attention mask (Dong et al., 2019) as it treats all modalities as targets for generation, allowing the model to simultaneously learn each modality conditioned on the preceding modalities along with the first modality which performs unconditional generation in each iteration.

The conventional self-attention mechanism is widely recognized for its expressive capabilities:

$$\begin{aligned} Attention(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \\ &= AV \end{aligned}$$

where Q, K, V, and d_k indicate queries, keys, values, and dimensions of queries and keys, respectively. However, when aiming for scalability and addressing long-range sequences, its computational demands can become a bottleneck.

To handle this with limited resources, we adopt the Performer (Choromanski et al., 2020) as an alternative that enhances computational efficiency. Following Performer, we utilize the FAVOR+ algorithm which uses positive orthogonal random features to approximate the softmax function with linear space and time complexity, allowing the model to compute the attention score more efficiently and reduce memory consumption. For causal attention, we also adopt a prefix-sum mechanism to avoid storing an explicit lower-triangular regular attention matrix. The mechanism of the FAVOR+ algorithm for unidirectional attention are delineated below:

- Outer Product Computation: For each key k_i and value v_i, compute the outer product using random features designated for keys:

$$\Phi_k(k_i)v_i^T$$

where Φ_k stands for the random features corresponding to key.

- Prefix-Sum Matrix Update: Iteratively accumulate the outer products to update the prefix-sum matrix:

$$P_i = P_{i-1} + \Phi_k(k_i)v_i^T$$

Notably, P_0 starts initialized to zero.

- Attention Matrix Row Generation: For every iteration, the most recent prefix-sum is multiplied with the random feature vector pertaining to a query. This yields a new row for the AV matrix:

$$AV_i = \Phi_q(q_i)P_i$$

where Φ_q stands for the random features corresponding to query.

To encapsulate the operation in matrix terminology:

$$AV_i = \Phi_q(q_i)\sum_{j=1}^{i}\Phi_k(k_j)v_j^T$$

with AV signifying the matrix generated by the attention mechanism.

During training, we concatenate a series of chest X-rays and report embeddings from the same study in random order to form a single input sequence as shown in Figure 2 (c), which is then fed into the model. ViewXGen is trained to minimize the negative log-likelihood of the next token given the previous tokens. Given $[\mathbf{w}; \mathbf{v}^1; ...; \mathbf{v}^k]$ as the input sequence, for example, the loss function is formulated as follows:

$$\begin{aligned} L = &\sum_{i=1}^{n} -logP(w_i|w_{0:i-1}) + \sum_{i=1}^{m} -logP(v_i^1|w, v_{0:i-1}^1) \\ &+ ... + \sum_{i=1}^{m} -logP(v_i^k|w, v^1, ..., v^{k-1}, v_{0:i-1}^k) \end{aligned}$$

where $n = T+2$ and $m = h\times w+2$, and $w_0, w_n, v_0^1, v_m^1, \ldots, v_0^k, v_m^k$ are special tokens.

At inference, for generating an X-ray of a specific view, the input to the model is $[\mathbf{w}; \mathbf{v}^1; ...; \mathbf{v}^{k-1}]$, meaning that the report embeddings are followed by X-ray embeddings of other views (if available for this study.).

4. Experiments

4.1. Dataset

MIMIC-CXR (Johnson et al., 2019) contains 377,110 chest X-rays from 227,835 radiology studies. Each study has one or multiple chest X-rays and a single report. We select a total of 208,534 studies that contain at most 3 chest X-rays composed of the most common views, namely PA, AP, and LATERAL[3]. Appendix A shows the statistics of chest X-ray view composition in each study. From the report, we use the two primary sections, namely Findings and Impression. We follow the official split of MIMIC-CXR (train 204,102, valid 1,659 test 2,773).

4.2. Evaluation Metrics

We evaluate the generated chest X-rays in various aspects, from sample quality to clinical efficacy. FID is the standard evaluation metrics in generative models, but it is not appropriate to capture complex medical concepts. Therefore, we use an additional metric, including 14-diagnosis classification We also perform human evaluation.

4.2.1. Statistical Evaluation

For FID (Heusel et al., 2017), we compute the distances of feature statistics between the original X-rays from the test set and the generated X-rays with the 1024-dimensional feature of the DenseNet-121 pretrained on chest X-ray datasets (Cohen et al., 2022).

4.2.2. Clinical Efficacy Evaluation

For 14-diagnosis classification, We train DenseNet-121 with positive labels extracted from the Findings and Impression sections using CheXpert labeler (Irvin et al., 2019). The model then predicts the classes of the generated chest X-rays. We report micro-averaged AUROC.

4.2.3. Human Evaluation

Using 100 triples of an original chest X-ray, a generated chest X-ray from our model, and a baseline, we ask three board-certified clinicians to evaluate each chest X-ray on three aspects: (1) realism, (2) alignment with the given report, and (3) the view position among PA, AP, and LATERAL views. Both (1) and (2) are rated on a scale from 1 (worst) to 5 (best). The triples consist of 33 triples from PA and AP and 34 triples from LATERAL. The clinicians consist of two radiologists and one neurosurgeon, and the X-rays are presented in random order for each triple.

4.3. Experiment Design

4.3.1. The Effect of Multi-view Chest X-rays

To evaluate the effect of using multi-view chest X-rays on the generation quality, we divide the test dataset into three groups based on the number of chest X-rays per study. These groups include studies with one X-ray (S w/1), two X-rays (S w/2), and three X-rays (S w/3). We evaluate our model by incrementally increasing the number of input chest X-rays within each group. For example, in the group of studies with two X-rays (S w/2), we first only use the report as the input condition for chest X-ray generation. Next, we use both the report and the remaining chest X-ray as the input condition. Then we compare the generated chest X-rays under these different conditions.

4.3.2. The Ability to Generate Specific Views

We evaluate the impact of the special tokens in generating specific views by asking the three clinicians to identify the view positions of the generated chest X-rays.

4.3.3. Comparison with Fine-tuned Stable Diffusion

We compare ViewXGen with a fine-tuned Stable Diffusion for chest X-ray generation as proposed in Chambon et al. (Chambon et al., 2022a). While various chest X-ray generation models have been proposed, only Chambon et al. (Chambon et al., 2022a) utilize radiology reports as an input condition. In addition, Stable Diffusion has shown great performance in image generation.

4.3.4. Comparison with a retrieval-based approach

Besides generating chest X-rays from reports and additional inputs, it is also possible to retrieve chest X-rays that closely match the contents of these reports. We qualitatively compare images X, generated by ViewXGen using reports R and additional inputs,

3. A study can have PA, PA, LAT or PA, LAT, or just AP.

with images X^* retrieved based on their pairing with the most similar reports R^* in the training set. This similarity is determined by the MedViLL approach (Moon et al., 2022), which identifies R^* as the most similar report sharing exactly matching 14 disease labels with R.

4.3.5. The Advantage of the Unified Model

We evaluate the advantage of a unified model compared to separate models for multi-view chest X-ray generation. There are three variants: 1) Single$_{AP}$, 2) Single$_{PA}$, 3) Single$_{LAT.}$, where each are trained to generate only the AP view, PA view, and the Lateral view images, respectively.

4.3.6. The Possibility for Radiology Report Generation

Due to our model's simple architectural design, it exhibits the capability to generate radiology reports in addition to chest X-rays. However, it is important to note that our primary focus and contributions lie in the generation of high-quality and view-specific chest X-rays. The generation of radiology reports serves as a proof of concept, demonstrating the versatility of our model. In Appendix C, we provide results showcasing the feasibility of generating radiology reports by swapping the order of text and image tokens in the input sequence.

5. Results and Discussion

The statistical significance is determined by calculating the confidence interval for the difference between the two group means. A 95% confidence interval (α = 0.05) is obtained by performing a non-parametric bootstrap. 1,000 bootstrap samples of the same size as the original test dataset are randomly taken from the dataset with replacement. In each table, numbers within parentheses indicate 95% CI. Diff.() indicates the confidence interval for the difference between the two means. Additionally, as the lower FID score indicates better performance, the negative mean FID difference reflects better performance.

5.1. The Effect of Multi-view Chest X-rays

We investigate the effect of inputting multi-view chest X-rays on the generation ability. As described in Section 4.3, we divide test dataset into three groups (S w/1, w/2, and w/3) and evaluate within each group.

For chest X-ray generation, we use the report as the input condition and also incrementally add the rest of the chest X-rays as input. Table 1 shows FID and 14-diagnosis classification results, respectively. In the ALL view of the S w/2 group, we can observe that *2 of 2* achieves significantly higher performance than *1 of 2* in both statistical (FID) and clinical efficacy (AUROC: *2of2* – *1of2* = 0.049, [95% CI 0.048, 0.049]) metrics. Also, *2 of 2* significantly outperforms *1 of 2* in the individual views (AP, PA and Lateral). In the ALL view of S w/3 group, using additional chest X-rays (*2 of 3* and *3 of 3*) shows significantly higher performance when compared to using only the report (*1 of 3*) across all metrics (AUROC: *3of3* – *1of3* = 0.032, [95% CI 0.030, 0.034] and *2of3* – *1of3* = 0.026, [95% CI 0.025, 0.028]). In the PA and Lateral views, both *2 of 3* and *3 of 3* significantly outperform *1 of 3* across all metrics. As for the AP view, on the other hand, although both *2 of 3* and *3 of 3* show significantly lower FID (the lower the better) than *1 of 3*, *2 of 3* does not show significantly superior 14-diagnosis classification performance than *1 of 3*. We believe this is partly due to the small number of AP views in the S w/3 group (refer to Table 5 more details), which also could be the cause for generally higher FID scores. Moreover, note that *3 of 3* does not always outperform *2 of 3* in some metrics. Specifically, the AP view and the PA view do not show statistically significant differences between *3 of 3* and *2 of 3* in terms FID. Also, *3 of 3* has significantly lower AUROC performance than *2 of 3* in the Lateral view (mean AUROC Lateral difference –0.005, [95% CI –0.007, -0.002]). We believe this is because the studies with three chest X-rays account for only a small percentage of the entire train dataset (8.5%, refer to Table 5 for more details.). Therefore, there is less opportunity for the model to learn the *3 of 3* input format during training. We can conclude that utilizing multiple X-ray views as input generally helps the model generate more accurate chest X-rays that can capture the abnormal findings in the report and other chest X-rays.

A key finding in this experiment is that considering the relations between the multi-view chest X-rays of the same study is important, as they provide valuable information. We observe that using multi-view chest X-rays can faithfully capture abnormal findings in chest X-ray generation, as *2 of 2* and *3 of 3* show statistically significant differences compared to *1 of 2* and *1 of 3* input formats. Although *2 of 3* sometimes demonstrates inferior performance than *1*

Table 1: Evaluations of generated chest X-rays using FID and 14-diagnosis classification to quantify the effect of using multi-view chest X-rays in chest X-ray generation. src., tar., and LAT. are short for source, target, and LATERAL, respectively. In each group, best values are emboldened and second-best underlined.

Group	Input	(src. → tar.)	FID (↓)				micro AUROC			
			ALL	AP	PA	LAT.	ALL	AP	PA	LAT.
S w/1	1 of 1	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	25.86 (25.727, 25.993)	25.986 (25.858, 26.113)	74.189 (73.425, 74.953)	41.322 (40.965, 41.679)	0.747 (0.747, 0.747)	0.751 (0.75, 0.751)	0.756 (0.751, 0.76)	0.565 (0.562, 0.569)
S w/2	1 of 2	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	16.965 (16.916, 17.013)	26.878 (26.699, 27.058)	17.778 (17.696, 17.859)	20.947 (20.872, 21.021)	0.664 (0.663, 0.664)	0.74 (0.739, 0.741)	0.642 (0.641, 0.643)	0.634 (0.633, 0.635)
	2 of 2	$(\mathbf{w}, \mathbf{v}^2 \rightarrow \mathbf{v}^1)$	**9.186** (9.133, 9.239)	**22.071** (21.827, 22.316)	**8.337** (8.301, 8.373)	**9.088** (9.054, 9.122)	**0.712** (0.712, 0.713)	**0.753** (0.752, 0.754)	**0.692** (0.691, 0.693)	**0.702** (0.701, 0.702)
	Diff. (2of2 - 1of2)	-	-7.779 (-7.851, -7.707)	-4.807 (-5.110, -4.504)	-9.441 (-9.530, -9.351)	-11.858 (-11.941, -11.776)	0.049 (0.048, 0.049)	0.013 (0.012, 0.014)	0.050 (0.049, 0.051)	0.067 (0.066, 0.068)
S w/3	1 of 3	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	21.148 (21.049, 21.246)	39.049 (38.714, 39.383)	27.051 (26.778, 27.325)	24.846 (24.699, 24.992)	0.668 (0.667, 0.669)	0.711 (0.709, 0.713)	0.666 (0.664, 0.668)	0.643 (0.642, 0.644)
	2 of 3	$(\mathbf{w}, \mathbf{v}^2 \rightarrow \mathbf{v}^1)$	12.792 (12.698, 12.887)	23.912 (23.524, 24.299)	14.606 (14.381, 14.83)	16.778 (16.677, 16.878)	0.694 (0.693, 0.695)	0.689 (0.687, 0.691)	**0.717** (0.716, 0.719)	**0.679** (0.678, 0.681)
	3 of 3	$(\mathbf{w}, \mathbf{v}^2, \mathbf{v}^3 \rightarrow \mathbf{v}^1)$	**12.684** (12.588, 12.781)	**23.695** (23.361, 24.03)	**14.517** (14.285, 14.75)	**16.499** (16.403, 16.595)	**0.699** (0.698, 0.7)	**0.72** (0.718, 0.722)	0.716 (0.714, 0.717)	0.675 (0.673, 0.676)
	Diff. (3of3 - 1of3)	-	-8.4631 (-8.6264, -8.2997)	-15.3531 (-15.9502, -14.756)	-12.5341 (-12.9476, -12.1205)	-8.3463 (-8.5437, -8.149)	0.0319 (0.0302, 0.0335)	0.0088 (0.0054, 0.0122)	0.0498 (0.0469, 0.0528)	0.0318 (0.0294, 0.0342)
	Diff. (3of3 - 2of3)	-	-0.1078 (-0.2712, 0.0555)	-0.216 (-0.8131, 0.381)	-0.0881 (-0.5017, 0.3254)	-0.2783 (-0.4756, -0.081)	0.0055 (0.0038, 0.0071)	0.0311 (0.0277, 0.0345)	-0.0016 (-0.0045, 0.0013)	-0.0046 (-0.007, -0.0022)
	Diff. (2of3 - 1of3)	-	-8.3553 (-8.5186, -8.1919)	-15.137 (-15.7341, -14.54)	-12.4459 (-12.8595, -12.0324)	-8.068 (-8.2654, -7.8707)	0.0264 (0.0248, 0.028)	-0.0223 (-0.0257, -0.0189)	0.0514 (0.0485, 0.0544)	0.0364 (0.034, 0.0388)

Table 2: Human evaluation Average means and standard deviations across three clinicians.

Models	Realism				Alignment				View Position		
	ALL	AP	PA	LATERAL	ALL	AP	PA	LATERAL	AP	PA	LATERAL
Original Image	4.177 ± 0.793	4.294 ± 0.703	4.281 ± 0.579	3.961 ± 0.912	3.977 ± 1.002	4.196 ± 0.855	4.156 ± 0.793	3.588 ± 1.123	0.843 ± 0.632	0.583 ± 0.487	1.0 ± 0.0
ViewXGen	4.193 ± 0.675	4.206 ± 0.659	4.188 ± 0.626	4.186 ± 0.674	3.583 ± 1.013	3.559 ± 1.028	3.719 ± 0.928	3.48 ± 1.043	0.755 ± 0.415	0.667 ± 0.461	1.0 ± 0.0
Stable Diffusion	2.09 ± 0.951	-	-	-	1.827 ± 0.812	-	-	-	-	-	-

of 3 on clinical efficacy metrics (AUROC of the AP view), the overall performance demonstrated by the ALL view suggests the effectiveness of utilizing more information rather than less information.

5.2. The Ability to Generate Specific Views

View Position column in Table 2 confirms that the view-specific special tokens can capture refined features of each view. Specifically, the lateral view result (Lateral: Original 1.0 vs ViewXGen 1.0) shows that the view-specific special tokens can properly capture the characteristics of the lateral view that are distinct from the frontal view. In addition, the 14-disease classification results in Table 1 support that our model does not simply generate the lateral appearance of the chest but generates the lateral chest X-rays that faithfully reflect the abnormal findings. The generated AP view images are certainly distinguishable from PA view images, but not as clearly as the original AP view images (AP: Original 0.843 vs ViewXGen 0.755), indicating that the AP view special tokens do not perfectly capture the characteristics of the AP view. On the other hand, given that the generated PA view images are more distinguishable than the original PA view images (PA: Original 0.583 vs ViewXGen 0.667), we can infer that the PA view special tokens are already capturing the characteristics of the PA view as best as possible. These results suggest that the view-specific special tokens are effective in generating chest X-rays in specific views, and that our model can even generate the desired views even if they do not exist in reality. The green dashed boxes in Fig. 3 show the generated chest X-rays that do not exist in the study. We can observe that the generated absent views have anatomical similarities to other existing views within the same study.

Table 3: Comparison of ViewXGen and the fine-tuned Stable Diffusion for chest X-ray generation.

Models	FID (↓)	micro AUROC
Stable Diffusion (S.D)	78.965 (78.883, 79.046)	0.589 (0.589, 0.589)
ViewXGen	**19.212** (19.157, 19.267)	**0.711** (0.711, 0.711)
Diff. (ViewXGen − S.D)	-59.753 (-59.852, -59.655)	0.122 (0.122, 0.122)

5.3. Comparison with Stable Diffusion

Table 3 shows the chest X-ray generation performances of ViewXGen and the fine-tuned Stable Diffusion. For a fair comparison, our model generates chest X-rays using only radiology reports as input, without inputting any additional chest X-rays (i.e. ViewXGen uses *1 of 1*, *1 of 2*, and *1 of 3*, respectively from S w/1, S w/2, and S w/3). We can observe that ViewXGen significantly outperforms the fine-tuned Stable Diffusion across all metrics (mean AUROC difference 0.122, [95% CI 0.122, 0.122]). We believe that these performance differences mainly arise from the

drastic difference in pixel distributions between the chest X-ray images and the general domain images used for originally training Stable Diffusion, and the difference in the length of input text (i.e. long radiology report VS short image captions). In addition, our model proves again that using additional chest X-rays can effectively generate more realistic and accurate chest X-rays when comparing Tables 4 and 3, with 14-diagnosis classification AUROC of 0.728 [95% CI 0.728, 0.729] VS 0.711 [95% CI 0.711, 0.711].

5.4. Human Evaluation

Table 2 confirms that ViewXGen can generate realistic chest X-rays comparable to the original. More specifically, the generated frontal X-rays score 0.091 points lower than the original image (Original 4.288 vs ViewXGen 4.197 on a 1-5 scale). One of the reasons of this difference can be the fact that the lines and tubes are sometimes generated in the wrong positions, and the details of the supporting device is insufficiently depicted. In terms of alignment, both the original and ViewXGen attain less than 4 points for the lateral view. This is because reports are usually written based on the frontal view, and since the lateral view plays an auxiliary role, much information cannot be found in the lateral view. Thus, focusing on the frontal view results, ViewXGen scores 0.538 points lower than the original image (Original 4.177 vs ViewXGen 3.629 on a 1-5 scale). This difference mainly arises because our model occasionally fails to fully reflect in the X-rays the abnormalities in the report. We can conclude that our model can generate chest X-rays similar to the original, but sometimes dose not faithfully reflect the contents in the report. Also, we can observe that the view-specific special tokens can capture refined features of each view, enabling the model to generate view-specific X-rays (AP: Original 0.843 vs ViewXGen 0.755, PA: Original 0.583 vs ViewXGen 0.667, Lateral: Original 1.0 vs ViewXGen 1.0). In addition, our model scores higher than the baseline for both realism (ViewXGen 4.193 vs Stable Diffusion 2.09 on a 1-5 scale) and alignment (ViewXGen 3.583 vs Stable Diffusion 1.827 on a 1-5 scale). Note that the baseline fails to learn view-specific information; thus, we do not evaluate its ability to generate images of specific views.

5.5. The Advantage of the Unified Model

We study the advantage of training a unified model for multi-view chest X-ray generation.

In Table 4, we compare our model with Single_{AP}, Single_{PA}, and $\text{Single}_{LAT.}$. In terms of the statistical metric (FID, the lower the better), ViewXGen outperforms the single models only in the PA case. In terms of the clinical efficacy metric (14-diagnosis classification), however, it shows significantly superior performance than all single models: ViewXGen – $\text{Single}_{AP} = 0.066$, [95% CI 0.065, 0.066], ViewXGen – $\text{Single}_{PA} = 0.007$, [95% CI 0.007, 0.008], and ViewXGen – $\text{Single}_{LAT.} = 0.027$, [95% CI 0.027, 0.028]. This suggests that training a model to generate multiple views helps the model to correctly capture the abnormalities described in the report.

From these results, we demonstrate that ViewXGen is comparable, if not superior, to the various single models tailored to generate only its specific modality. Specifically, only the mean FID difference of PA outperforms the single model in the statistical metric, but except for this, ViewXGen significantly outperforms the single models across all metrics. This suggests that our model can generate multi-view chest X-rays with clinically meaningful information. We can conclude that bidirectional training has a synergistic effect on generation tasks and also can save time and computational costs, as opposed to training multiple single models.

5.6. Comparison with a Retrieval-based Approach

Fig. 5 shows the results. The first sample shows that the image X, generated through our approach using the additional input view and the report R, accurately reflects the patient's gender information. In contrast, the retrieved image X^* paired with the report R^*, which is most similar to R, fails to incorporate this detail. Moreover, in identifying R*, even though disease label information was used, it did not capture the location of support devices, leading to the retrieved image X* inaccurately reflecting the precise position of support devices. This indicates the need for advanced techniques to consider all elements in retrieval effectively. The second sample demonstrates that X^* does not account for the patient's obesity level, as this information is absent in the report. Thus, it fails to reflect the patient's actual physical condition. In the third sample, the report R fails to mention a support device, yet the generated image X, enhanced by an additional lateral view, accurately includes this detail, in contrast to the retrieved image X^* which lacks it. Moreover, despite

Table 4: Comparison of ViewXGen with various single models to evaluate the impact of the unified model in chest X-ray generation. The FID scores for the original image are calculated with the same number of train set as the test set. Each AP, PA and LAT. column shows the performance measured by dividing the generated chest X-rays according to their original view position.

Models	FID (↓)				micro AUROC			
	ALL	AP	PA	LAT.	ALL	AP	PA	LAT.
Original Image	0.541 (0.531, 0.551)	1.15 (1.124, 1.177)	1.611 (1.59, 1.632)	1.082 (1.068, 1.096)	0.81 (0.809, 0.81)	0.808 (0.808, 0.808)	0.812 (0.812, 0.812)	0.793 (0.793, 0.794)
Single_{AP}	-	**16.172** (16.037, 16.307)	-	-	-	0.689 (0.689, 0.689)	-	-
Single_{PA}	-	-	7.579 (7.553, 7.605)	-	-	-	0.697 (0.696, 0.697)	-
$\text{Single}_{LAT.}$	-	-	-	**8.242** (8.222, 8.261)	-	-	-	0.667 (0.667, 0.668)
ViewXGen	10.582 (10.554, 10.609)	17.639 (17.572, 17.705)	**6.324** (6.302, 6.347)	9.553 (9.531, 9.575)	0.728 (0.728, 0.729)	**0.755** (0.755, 0.755)	**0.704** (0.704, 0.705)	**0.695** (0.694, 0.695)
Diff. (ViewXGen − $\text{Single}_{eachview}$)	-	1.467 (1.317, 1.618)	-1.255 (-1.290, -1.221)	1.311 (1.282, 1.341)	-	0.066 (0.065, 0.066)	0.007 (0.007, 0.008)	0.027 (0.027, 0.028)

the absence of gender information in the report, the generated image X correctly represents the patient's gender. These examples illustrate the advanced capabilities of our approach to generate images that accurately include details, even those not explicitly stated or omitted in the reports. In contrast, the retrieval-based approach often fails to capture details that are not explicitly mentioned. This comparison underscores the limitations of the retrieval method in handling complex clinical scenarios effectively.

5.7. Qualitative Examples

Fig. 3 (a) shows that ViewXGen can generate realistic chest X-ray images even when conditioned only on the report, describing a small consolidation in the lingula as described by the report. When given an additional view, ViewXGen generates an image that is more similar to the original image, showing its ability to take advantage of both input modalities. Fig. 3 (b), on the other hand, shows a scenario where the generated image, conditioned solely on the report, fails to accurately capture all the details described in the report. Although the report says "large right pleural effusion", the generated image depicts a rather small pleural effusion. When given an additional view, however, ViewXGen can draw pleural effusion that is of the similar size as that of the original image. Furthermore, both figures show that the view-specific special tokens enable ViewXGen to generate the desired views, even when they do not exist in reality. All figures are confirmed by the clinicians.

6. Limitation and Conclusion

Here, we propose for the first time a novel approach to generate chest X-rays with specific views, addressing the limitations of existing methods that primarily focus on generating frontal views. Our model introduces specialized tokens and leverages multi-view information to enable users to generate chest X-rays according to their desired views. Our approach has some limitations, each providing opportunities for future work. First, due to the nature of the real-world patient dataset, the report often contains references to previous studies (e.g. unchanged, increase, and compared to previous radiographs). These references have the potential to impact the quality of chest X-ray generation. In the future, we plan to use CXR-PRO(Ramesh et al., 2022b), a refined dataset that removes comparison phrases, to generate clinically accurate chest X-rays. Second, the human evaluation confirms that our model generates chest X-rays that sometimes fail to fully reflect the facts in the given report (Original 3.977, ViewXGen 3.583 on a 1-5 scale in Table. 2). In addition, the position and shape of the support device are slightly different from the original image, so we can infer that our model sometimes has difficulty capturing fine details. We defer addressing these challenges for the future.

Acknowledgments

This work was supported by Samsung Electronics (No.IO201211-08109-01), Institute of Information Communications Technology Planning Evaluation (IITP) grant (No.2019-0-00075), National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945), and the Korea Health Industry Development Institute (KHIDI) grant (No.HI21C1138) funded by the Korea government (MSIT, MOHW).

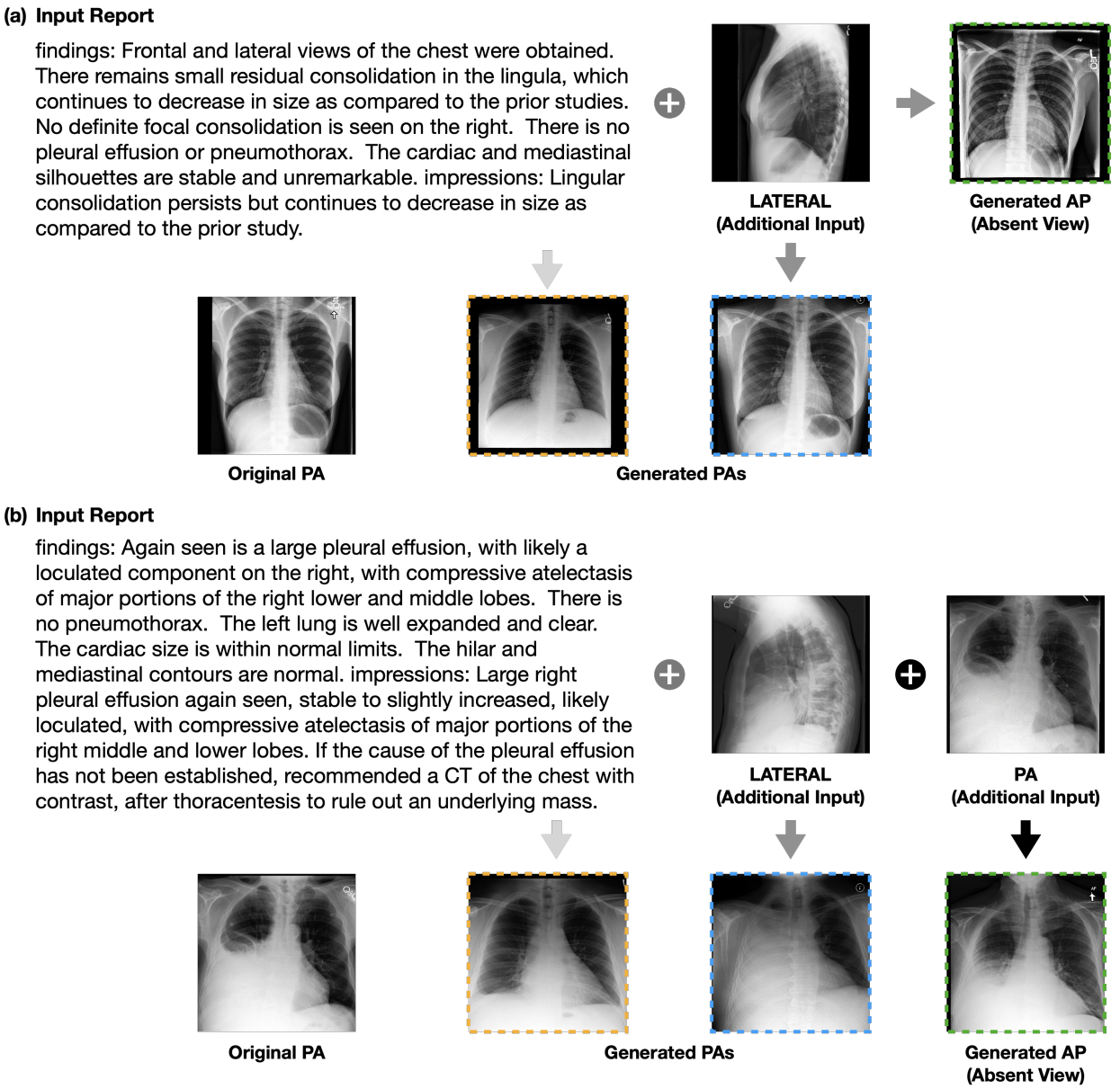

Figure 3: Generated chest X-rays of ViewXGen. (a) Based only on the report, the generated PA in the orange dashed box draws a rather small portion of the consolidation in the lingula, as is written in the report. Based on an additional lateral view, the generated PA in the blue dashed box draws a consolidation that is of more similar size as that of the original PA. (b) The generated PA conditioned only on the report (orange dashed box) draws relatively small-sized pleural effusion while the report says "large right pleural effusion". However, by adding an additional lateral view (blue dashed box), ViewXGen can properly generate the PA view with large pleural effusion.

References

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*, 2022a.

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*, 2022b.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32, 2019.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63 (11):139–144, 2020.

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.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*, 2020.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26 (12):6070–6080, 2022.

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.

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.

Elizabeth Puddy and Catherine Hill. Interpretation of the chest radiograph. *Continuing Education in Anaesthesia, Critical Care and Pain*, 7(3):71–75, 2007.

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

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022b.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160, 2020.

Appendix A. Dataset Statistic

Table 5 represents the compositions of chest X-ray views in each study.

Appendix B. Implementation Details

B.1. Image tokenizer

We adopt VQ-GAN with d_z=256 and a codebook size of 1024. The input image of size 512×512 is quantized into $32 \times 32 = 1024$ discrete visual tokens. The model is trained for 540k steps with a batch size of 8, a learning rate of 4.5e-6 with the Adam optimizer.

B.2. Text tokenizer

We train a byte-level BPE tokenizer (Wang et al., 2020) with a minimum frequency of 2 on reports converted to lowercase. We then obtain 14,526 unique tokens, including three special tokens $[SOS]$, $[EOS]$, $[TXT\_PAD]$.

B.3. ViewXGen

We set the length of word tokens n=256 and visual tokens m=1,026, including special tokens. In this work, ViewXGen takes up to three chest X-rays as input, as the majority of studies in the MIMIC-CXR dataset have three or fewer images. However, it is able to take more images if they are available. Our model is built on the Transformer architecture with generalized attention (Choromanski et al., 2020). The model has 12 layers, 12 heads, and 768 dimensions. We incorporate seven special tokens (in addition to three text special tokens), namely $[SOS_{AP}]$, $[EOS_{AP}]$, $[SOS_{PA}]$, $[EOS_{PA}]$, $[SOS_{LAT}]$, $[EOS_{LAT}]$, $[IMG\_PAD]$. Thus, the size of visual embedding function (*i.e.* lookup matrix) is $f_{VE}(\cdot) \in \mathbb{R}^{N \times d}$, where $N = 1024 + 7$, $d = 768$, and word embedding function is $f_{WE}(\cdot) \in \mathbb{R}^{M \times d}$, where M = 14,526, $d = 768$. We train the model for 337k steps with a batch size of 48 using four NVIDIA RTX A6000 GPUs. We use the AdamW optimizer with a learning rate of 1.7e-4, β_1=0.9, β_2=0.999, $e = 1e-8$, a weight decay of $1e-2$, and a cosine decay schedule. We generate all samples with Top-p sampling (Holtzman et al., 2019) with p=0.9 and temperature=0.7.

B.4. Finetuned Stable Diffusion

Following Chambon et al. (Chambon et al., 2022a), we replace the CLIP text encoder with SapBERT (Liu et al., 2020) to handle both Findings and Impression sections (the CLIP tokenizer is limited to 77 tokens) and keep frozen the text encoder and VAE and only train U-Net from scratch.

Appendix C. Radiology Report Generation

C.1. Evaluation Metrics

We evaluate the generated reports using metrics such as BLEU and CheXpert F1 score. For BLEU (Papineni et al., 2002), we report BLEU-4 between the original and the generated reports. For CheXpert F1 score, We extracted diagnosis labels from the original and generated reports with the CheXpert labeler. We then compare these labels and measure micro-averaged F1.

C.2. The Effect of Multi-view Chest X-rays

Table 6 shows the effect of using multi-view chest X-rays in the radiology report generation. We increase the input chest X-rays to generate the target report. In the S w/2 group, although *2 of 2* shows significantly lower performance than *1 of 2* in terms of the simple statistical metric (BLEU-4), *2 of 2* significantly outperforms *1 of 2* in the clinical efficacy metrics (mean CheXpert F1 difference, 0.007 [95% CI 0.006, 0.007]). In the S w/3 group, *3 of 3* performs significantly higher across all metrics. These results show that using multi-view chest X-rays encourages the model to generate more clinically precise reports. In particular, the use of multi-view chest X-rays in radiology report generation can be considered to follow the writing behavior of radiologists given that *2 of 2* and *3 of 3* show significantly superior performance than other input formats in clinical efficacy metric (CheXpert F1).

C.3. The Advantage of the Unified Model

As shown in Table 7, we compare our model with Single_{report}. We can observe that ViewXGen significantly outperforms Single_{report} in both statistical and clinical efficacy metrics (mean CheXpert F1 difference = 0.067, [95% CI 0.066, 0.067]). This indicates that combining chest X-ray image generation as a target can effectively capture local regions that encourage the model to generate more precise reports containing abnormal findings.

Table 5: Composition of chest X-ray views in each study. S w/1, S w/2, and S w/3 indicate the number of chest X-rays per study. LAT. is short for Lateral.

Group	Split	AP	PA	LAT.
S w/1	Train	91,736	85	1,596
	Valid	782	1	12
	Test	1,428	3	29

Group	Split	(PA, LAT.)	(AP, LAT.)	(AP, AP)	(LAT., LAT.)	(PA, PA)	(AP, PA)
S w/2	Train	68,600	13,971	9,853	471	315	105
	Valid	513	95	90	3	2	2
	Test	671	212	162	10	3	1

Group	Split	(PA, PA, LAT.)	(AP, LAT., LAT.)	(PA, LAT., LAT.)	(AP, AP, LAT.)	(AP, AP, AP)	Etc.
S w/3	Train	8,056	3,968	3,539	848	748	211
	Valid	66	36	36	9	7	5
	Test	82	89	52	11	14	6

Table 6: Evaluations of generated reports using BLEU and CheXpert F1 to quantify the effect of using multi-view chest X-rays on radiology report generation. src. is short for source, and tar. for target. Numbers within parentheses indicate 95% CI. Diff.() indicates the confidence interval for the difference between the two means.

Group	Input	(src. → tar.)	BLEU-4	CheXpert F1
S w/1	1 of 1	($\mathbf{v}^1 \rightarrow \mathbf{w}$)	0.042 (0.042,0.042)	0.412 (0.412, 0.412)
S w/2	1 of 2	($\mathbf{v}^1 \rightarrow \mathbf{w}$)	0.056 (0.056, 0.057)	0.415 (0.415, 0.415)
	2 of 2	($\mathbf{v}^1, \mathbf{v}^2 \rightarrow \mathbf{w}$)	**0.056** (0.056, 0.056)	**0.422** (0.421, 0.422)
	Diff. (2of2 − 1of2)	-	-0.001 (-0.001, -0.001)	0.007 (0.006, 0.007)
S w/3	1 of 3	($\mathbf{v}^1 \rightarrow \mathbf{w}$)	0.054 (0.054, 0.054)	0.435 (0.435, 0.436)
	2 of 3	($\mathbf{v}^1, \mathbf{v}^2 \rightarrow \mathbf{w}$)	0.060 (0.060, 0.061)	0.436 (0.435, 0.437)
	3 of 3	($\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3 \rightarrow \mathbf{w}$)	**0.063** (0.063, 0.063)	**0.451** (0.450, 0.452)
	Diff. (3of3 - 1of3)	-	0.009 (0.008, 0.009)	0.019 (0.014, 0.017)
	Diff. (3of3 - 2of3)	-	0.003 (0.002, 0.003)	0.016 (0.014, 0.017)
	Diff. (2of3 - 1of3)	-	0.006 (0.006, 0.007)	0.0003 (-0.001, 0.002)

Table 7: Comparison of ViewXGen with a single model to evaluate the impact of the unified model in radiology report generation. Numbers within parentheses indicate 95% CI. Diff.() indicates the confidence interval for the difference between the two means.

Models	BLEU-4	CheXpert F1
Single_{report}	0.038 (0.038 0.038)	0.353 (0.353, 0.353)
ViewXGen	**0.050** (0.050 0.050)	**0.420** (0.420, 0.420)
Diff. (ViewXGen − Single_{report})	0.012 (0.012, 0.012)	0.067 (0.066, 0.067)

C.4. Qualitative Examples

Fig. 4 (a) shows an example where ViewXGen generates accurate radiology reports when given one or two chest X-ray images. Fig. 4 (b) shows an example where the report generated based on only one view does not capture some findings, but additional input helps the model generate more precise reports. All examples are confirmed by the clinicians.

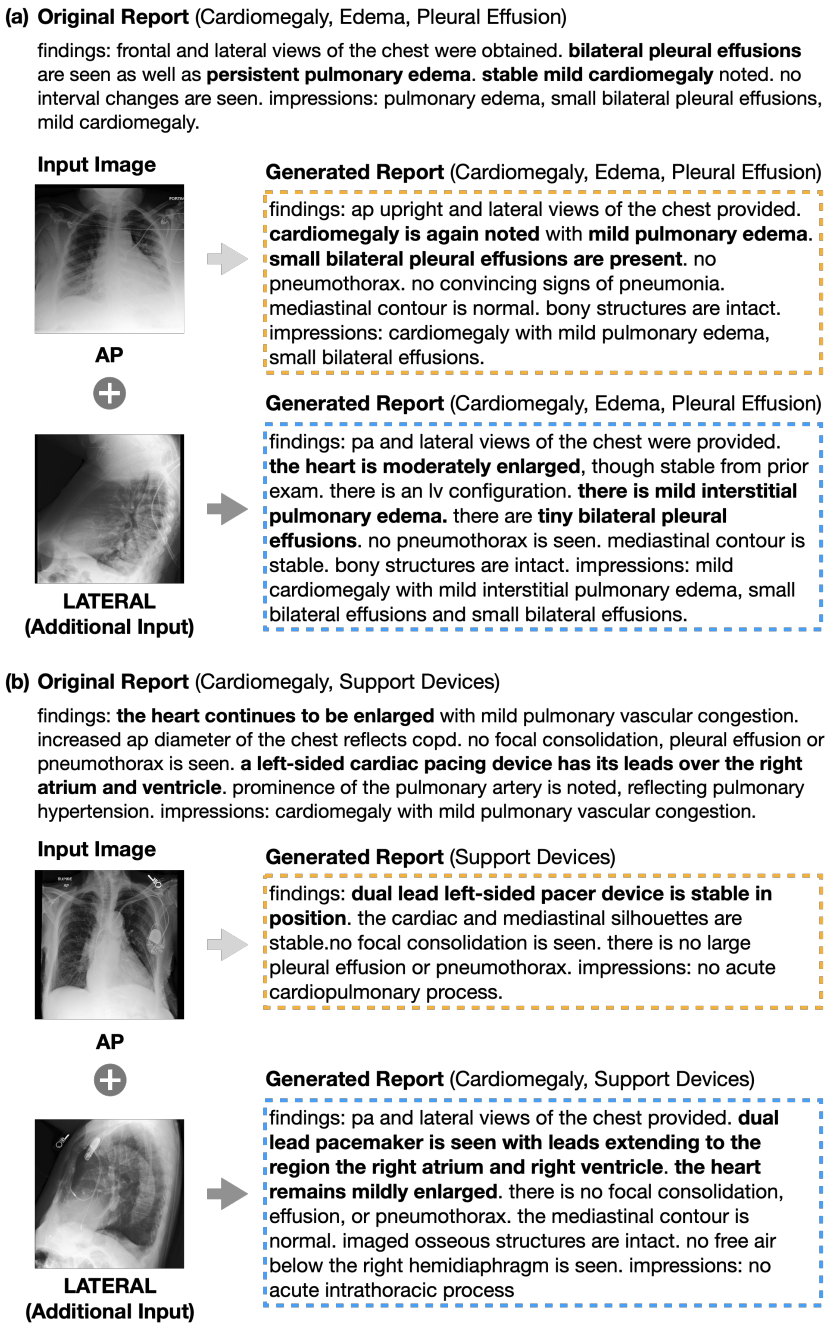

Figure 4: Generated radiology reports of ViewXGen. (a) Regardless of the number of chest X-rays input, ViewXGen can generate accurate radiology reports covering all diseases mentioned in the original report. (b) The generated report only from a single chest X-ray (orange dashed box) cannot fully capture the abnormalities in the given X-ray. With an additional chest X-ray, ViewXGen can generate a more precise report (blue dashed box) containing all diseases as described in the original report.

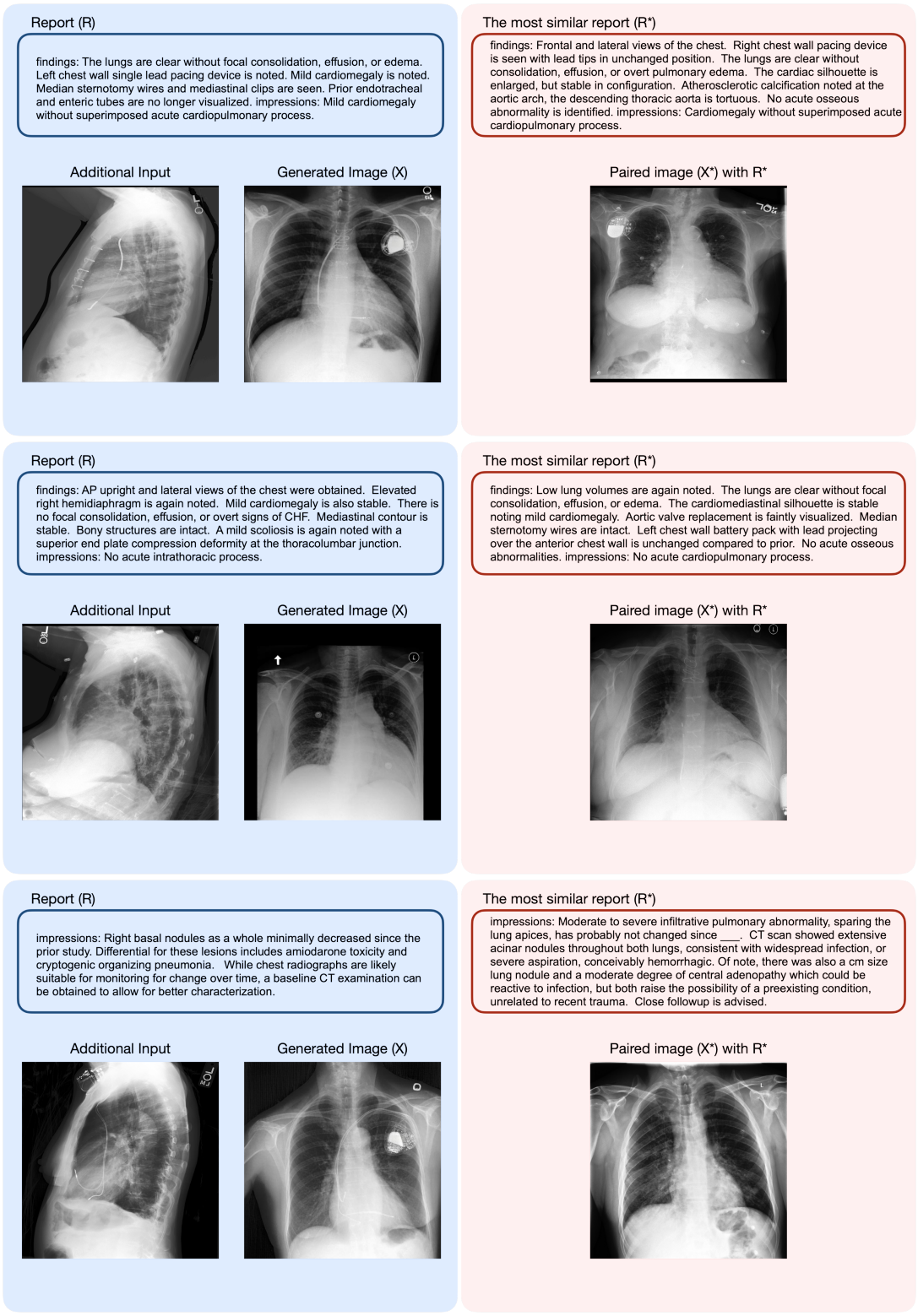

Figure 5: These examples highlight the advanced capabilities of our approach to generate images that accurately incorporate details, even those not explicitly stated or omitted in the reports. In contrast, they underline the limitations of a purely retrieval-based approach, which often fails to capture essential patient information such as gender or specific health conditions like obesity, especially when faced with incomplete or erroneous reports. This comparison demonstrates the inadequacy of the retrieval method in handling complex clinical scenarios.