UniXGen: A Unified Vision-Language Model for Multi-View Chest X-ray Generation and Report Generation

Hyungyung Lee[1], **Da Young Lee**[2*], **Wonjae Kim**[3], **Jin-Hwa Kim**[3,4]
Tackeun Kim[5], **Jihang Kim**[5], **Leonard Sunwoo**[5], **Edward Choi**[1]
[1] KAIST, [2] Deep-in-Sight Co., [3] Naver AI Lab, [4] SNU AIIS,
[5] Seoul National University Bundang Hospital
ttumyche@kaist.ac.kr, dyan.lee717@gmail.com, wonjae.kim@navercorp.com,
j1nhwa.kim@navercorp.com, tackeun.kim@snu.ac.kr, radio622@gmail.com,
leonard.sunwoo@gmail.com, edwardchoi@kaist.ac.kr

Abstract

Generated synthetic data in medical research can substitute privacy and security-sensitive data with a large-scale curated dataset, reducing data collection and annotation costs. As part of this effort, we propose UniXGen, a unified chest X-ray and report generation model, with the following contributions. First, we design a unified model for bidirectional chest X-ray and report generation by adopting a vector quantization method to discretize chest X-rays into discrete visual tokens and formulating both tasks as sequence generation tasks. Second, we introduce several special tokens to generate chest X-rays with specific views that can be useful when the desired views are unavailable. Furthermore, UniXGen can flexibly take various inputs from single to multiple views to take advantage of the additional findings available in other X-ray views. We adopt an efficient transformer for computational and memory efficiency to handle the long-range input sequence of multi-view chest X-rays with high resolution and long paragraph reports. In extensive experiments, we show that our unified model has a synergistic effect on both generation tasks, as opposed to training only the task-specific models. We also find that view-specific special tokens can distinguish between different views and properly generate specific views even if they do not exist in the dataset, and utilizing multi-view chest X-rays can faithfully capture the abnormal findings in the additional X-rays. The source code is publicly available at: https://github.com/ttumyche/UniXGen.

1 Introduction

Patient privacy, imbalanced class distribution, the need for trained clinicians, and the lack of large publicly available datasets are chronic problems in medical research. In an effort to alleviate these problems, research on synthetic data generation has been actively explored. Among them, the combination of chest radiographs and radiology reports has made significant progress, as it provides a comprehensive examination of the patient's condition that helps diagnose and detect various diseases. In response, we propose a unified chest X-ray and radiology report generation model with multiple views, namely UniXGen, with the following improvements.

First, we design a unified model for bidirectional chest X-ray and report generation. Although these tasks are bidirectional, existing works propose task-specific architectures and develop them separately. We simplify the design effort by utilizing VQ-GAN [7] as an image tokenizer. This approach allows

*Work done at KAIST

Preprint.

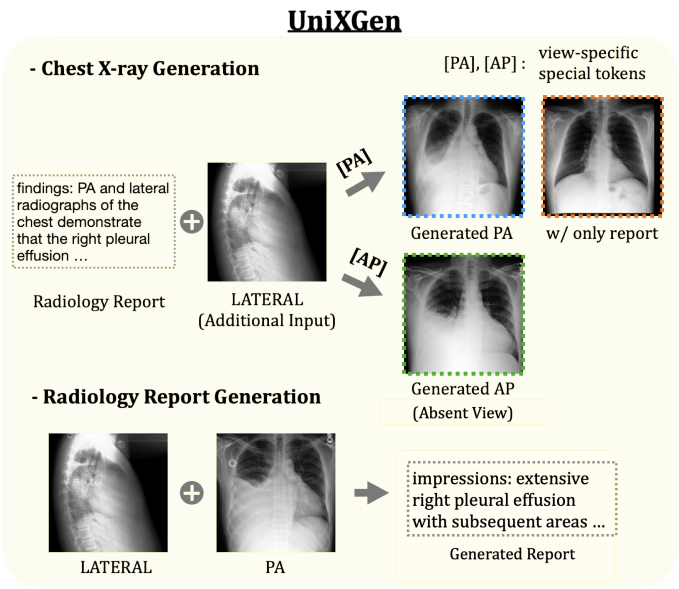

Figure 1: Overview of UniXGen. We propose a unified chest X-ray and report generation model. UniXGen can generate view-specific X-rays of the user's choice. We also boost the generation ability with multi-view chest X-ray input.

us to convert an image into a sequence of discrete tokens; thus both X-ray generation and report generation can be performed as sequence generation tasks with a unified model.

Next, UniXGen can generate chest X-rays with specific views of the user's choice, such as posterior-anterior (PA), anterior-posterior (AP), and lateral views. This ability can be useful when the desired view is unavailable. In addition, it allows for generating various views, which can provide additional findings that cannot be seen in a single view and also help radiologists make more accurate diagnoses. Despite these advantages, no prior efforts have been made to control the view position in chest X-ray generation. To achieve this, we introduce a set of special tokens according to the different view positions.

Lastly, we utilize multi-view chest X-rays and study their effects on the generation quality. X-rays from multiple views can provide more valuable information during generation, compared to a single view input. There are; however, no previous chest X-ray generation studies considering the relations between them. As a result, they do not take advantage of the additional findings available in other X-ray views of the same study. Meanwhile, in report generation, [40, 39] only use studies with a single frontal and a single lateral view. Therefore, their architectural design is not flexible to various input formats. In this work, we design a model that can flexibly take arbitrary input from single to multi-view images and empirically show that using multi-view images can faithfully capture abnormal findings in both chest X-ray and report generation.

Specifically, we adopt Performer [4], an efficient transformer-based architecture. Although the vanilla Transformer [34] is a task-agnostic design that can perform both image and text generation tasks, the computational cost increases quadratically by the input sequence length. We adopt the Performer for computational and memory efficiency as we utilize long paragraph reports and high resolution multi-view chest X-rays, which result in long-range sequences.

We evaluate our model on MIMIC-CXR [17]. The experimental results show that UniXGen achieves better performance on both standard metrics such as FID [13] and BLEU [25] and clinical efficacy metrics such as 14-diagnosis Classification and CheXpert scores with CheXpert labeler [16] over several baselines. Furthermore, human evaluation shows that UniXGen 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) We first propose a unified model for chest X-ray and radiology report generation in the medical domain.
2) UniXGen can generate view-specific X-rays with a set of special tokens according to the different view positions, even without the desired views.
3) We utilize multi-view chest X-rays and study their effects on chest X-ray and radiology report generation. Also, we design a flexible model that can take both single and multi-views as input.

2 Related Work

Chest X-ray Generation With the growing demand to access high quality medical data and the success of generative models such as VAE [18], GANs [8], and diffusion models [10], chest X-ray generation has gained a lot of attention. [22, 42] use unconditional GANs to augment a dataset for abnormality classification. [2, 24] adopt a latent diffusion model [31] 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, [1] has 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.

Radiology Report Generation Automatic report generation can relieve the heavy burden on radiologists. Most existing works adopt the CNN-RNN encoder-decoder framework based on standard image captioning approaches. Specifically, [37] integrates multi-level attention. [40, 21] adopt the hierarchical architecture, and the latter introduces reinforcement learning. [39] incorporates a hierarchical retrieval-based method. However, these task-specific architectures make it difficult to scale and generalize to other tasks. Also, they design a model for a specific input format. [37, 21] and [40, 39] only support single-view and two multi-view inputs, respectively. Recently, [23] proposed a Transformer-based architecture for various tasks, including report generation. They learn joint representations of chest X-rays and reports for report generation but do not extend their work to chest X-ray generation. In addition, they only use a subset of studies with a single frontal view.

Bidirectional Image-Text Generation Traditionally, image-to-text and text-to-image generation have been developed separately with task-specific architectures in the general domain. However, with the success of the Transformer-based model in both tasks [43, 28], several works [15, 41, 14] have argued the bidirectional nature of these tasks and propose a unified model for the bidirectional image-text generation. They discretize images into discrete tokens to formulate both tasks as sequence generation tasks. [15] uses K-means clustering and [41, 14] adopt vector quantization methods [33, 7].

Efficient Transformer Transformer [34] 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 [4], 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.

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 [33], VQ-VAE-2 [30] improved VQ-VAE by introducing a multi-scale hierarchical encoder. [7] introduced VQ-GAN with a discriminator and a perceptual loss for high-resolution images. [28] proposed DALL-E to better optimize VQ-VAE with Gumbel-softmax relaxation. Recently, diffusion models have achieved promising performance in generating high-quality samples in continuous domains (*e.g.*, image and audio) [27, 32]. However, diffusion models for text generation are still underdeveloped and have not achieved significant success compared to image generation. In addition, the models are not flexible to take arbitrary input from single to multiple images. Thus, we adopt the transformer-based auto-regressive model with VQ-GAN to generate both image and text with a unified model.

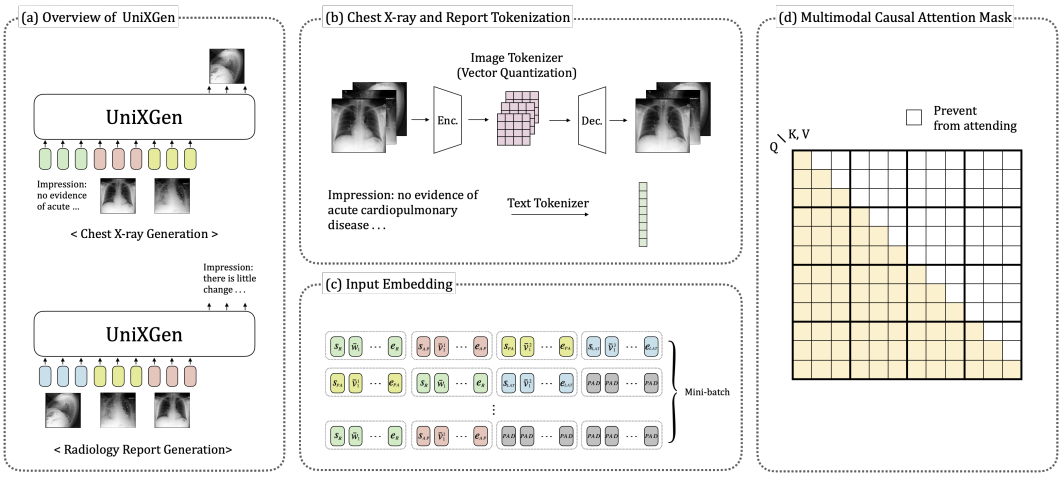

Figure 2: Overview of UniXGen architecture. (a) UniXGen is a unified model that can generate both reports and view-specific X-rays. (b) Images are tokenized via VQ-GAN, and reports are tokenized via a text tokenizer. (c) A minibatch consists of input sequences consisting of AP/PA/Lateral X-rays and a report in random order. (d) We use a causal attention mask to simultaneously handle multi-view X-rays and a report.

3 Method

Figure 2 shows the overall depiction of UniXGen. Notably, 1) We design a unified model for bidirectional chest X-ray and report generation. 2) To generate chest X-rays with specific views, we incorporate various special tokens according to the view type. 3) UniXGen takes a series of chest X-rays of different views and a report from the same study as input for better generation quality.

3.1 Input Embedding

Image Tokenization We adopt VQ-GAN [7] as an image tokenizer. This model consists of an encoder, a decoder, and a fixed-size learnable codebook. Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, the encoder embeds the input image into a continuous feature space $\mathbf{z} \in \mathbb{R}^{h \times w \times d_z}$, which is then quantized into a sequence of discrete tokens $\{v_1, \ldots, v_{h \times w}\}$ by finding the closest code embedding in the codebook via the nearest neighbor search. The decoder then maps the discrete codes back to the original input. This method allows the model to learn a compact and discrete representation of the images.

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 [11, 19]:

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

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

Radiology Report Embedding We first split a report into word tokens with a byte-level BPE tokenizer [36] 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:

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

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

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

3.2 Bidirectional Generative Model

We design a unified model for chest X-ray and report generation by formulating both tasks as sequence generation tasks. Incorporating the Transformer architecture [34], 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 Figure 2 (d). This attention mechanism differs from the sequence-to-sequence attention mask [6] 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.

Furthermore, to handle long-range sequences with limited resources, we adopt Performer [4] as an efficient attention mechanism. They propose 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, they utilize a prefix-sum mechanism to avoid storing an explicit lower-triangular regular attention matrix. Please refer to Appendix B for a visual explanation.

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. UniXGen 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:

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

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.). For radiology report generation, the model input is $[\mathbf{v}^1; ...; \mathbf{v}^k]$, assuming there are k views of X-rays available for this study.

4 Experiments

4.1 Dataset

MIMIC-CXR [17] 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]. Table 1 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).

Table 1: Composition of chest X-ray views in each study. S w/1, S w/2, and S w/2 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

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

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

4.2 Evaluation Metrics

We evaluate the generated chest X-rays and reports in various aspects, from sample quality to clinical efficacy. FID and BLEU are the standard evaluation metrics in generative models, but they are not appropriate to capture complex medical concepts. Therefore, we use additional metrics, including 14-diagnosis classification for the chest X-rays and CheXpert scores for the reports. We also perform human evaluations.

4.2.1 Statistical Evaluation

FID [9] 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 [5].

BLEU [25] We report BLEU-4 between the original and the generated reports.

4.2.2 Clinical Efficacy Evaluation

14-diagnosis Classification We train DenseNet-121 with positive labels extracted from the Findings and Impression sections using CheXpert labeler [16]. The model then predicts the classes of the generated chest X-rays. We report micro-averaged F1 and AUROC.

CheXpert Scores We extracted diagnosis labels from the original and generated reports with the CheXpert labeler. We then compare these labels and measure micro-averaged Precision, Recall, and F1.

4.2.3 Human Evaluation

Using 100 triples of an original chest X-ray generated chest X-rays 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

Experiments are designed to investigate the followings:

The Advantage of the Unified Model
We evaluate the advantage of a bidirectional unified model compared to separate models for chest X-ray and radiology report generation. There are four variants: 1) Single$_{AP}$, 2) Single$_{PA}$, 3) Single$_{LAT.}$, and 4) Single$_{report}$. Each model is only trained to maximize the log-likelihood of the tokens corresponding to its specific modality.

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. Similarly, for radiology report generation, we first use only one of the X-rays as input and then both X-rays.

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. Furthermore, the correctness of the view position is indirectly evaluated with the FID score as well. The generated chest X-rays are divided based on their original view position, and we calculate the FID scores to measure the similarity to the original view position.

Comparison with Fine-tuned Stable Diffusion
We compare UniXGen with a fine-tuned Stable Diffusion for chest X-ray generation as proposed in [1]. While various chest X-ray generation models have been proposed, only [1] utilize radiology reports as an input condition. In addition, Stable Diffusion has shown great performance in image generation.

4.4 Implementation Details

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.

Text tokenizer We train a byte-level BPE tokenizer [36] with a minimum frequency of 2 on reports converted to lowercase. We then obtain 14,526 unique tokens, including three special tokens $[SOS]$, $[EOS]$, $[TXT\_PAD]$.

UniXGen We set the length of word tokens n=256 and visual tokens m=1,026, including special tokens. In this work, UniXGen 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 [4]. 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. 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 [12] with p=0.9 and temperature=0.7.

Finetuned Stable Diffusion Following [1], we replace the CLIP text encoder with SapBERT [20] 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. Please refer to Appendix C for more details.

5 Results and Discussion

5.1 The Advantage of the Unified Model

We first study the advantage of training a bidirectional unified model for both generation tasks.

For chest X-ray generation, we compare our model with Single$_{AP}$, Single$_{PA}$, and Single$_{LAT.}$. Table 2 shows the results. Although UniXGen achieves slightly lower performance than single models (except for Single$_{PA}$) in terms of the statistical metric (FID), it outperforms all single models in the clinical efficacy metric (14-diagnosis Classification). This result suggests that combining report generation as a target can learn the abnormal findings from the report to generate chest X-rays that correctly capture the abnormalities described in the report.

For radiology report generation, we compare our model with Single$_{report}$. As shown in Table 3, we can observe that UniXGen significantly outperforms Single$_{report}$ across all metrics. This indicates that combining chest X-ray generation as a target can effectively capture local regions that encourage the model to generate more precise reports containing abnormal findings. We can conclude that bidirectional training has a synergistic effect on both generation tasks, as opposed to training only the single models.

Table 2: Comparison of UniXGen 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 (↓)				14-diagnosis Classification							
					F1				micro AUROC			
	ALL	AP	PA	LAT.	ALL	AP	PA	LAT.	ALL	AP	PA	LAT.
Original Image	0.361	0.776	0.874	0.487	0.510	0.540	0.492	0.460	0.810	0.808	0.812	0.793
Single$_{AP}$	-	**15.458**	-	-	-	0.382	-	-	-	0.689	-	-
Single$_{PA}$	-	-	7.189	-	-	-	0.365	-	-	-	0.697	-
Single$_{LAT.}$	-	-	-	**7.991**	-	-	-	0.326	-	-	-	0.667
UniXGen	10.436	17.331	**5.935**	9.328	0.416	**0.465**	**0.368**	**0.353**	0.728	**0.755**	**0.704**	**0.695**

Table 3: Comparison of UniXGen with a single model to evaluate the impact of the unified model in radiology report generation.

Models	BLEU-4	CheXpert Labeler		
		Precision	Recall	F1
Single$_{report}$	0.038	0.364	0.343	0.353
UniXGen	**0.050**	**0.431**	**0.410**	**0.420**

5.2 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. Tables 4 and 5 show FID and 14-diagnosis classification results, respectively. In the ALL view of the S w/2 group (1 of 2 vs 2 of 2), we can observe that 2 of 2 achieves better performance than 1 of 2 in both statistical (FID 16.416 vs 8.757) and clinical efficacy (F1 0.309 vs 0.400, AUROC 0.664 vs 0.713) metrics. Also, 2 of 2 is better than 1 of 2 in the individual views (AP, PA, and LATERAL). Similarly, in the ALL view of S w/3 group (1 of 3 vs 2 of 3 vs 3 of 3), 3 of 3 outperforms 2 of 3 and 1 of 3 across all metrics: FID (19.252 vs 11.450 vs 11.136), F1 (0.313 vs 0.362 vs 0.376), and AUROC (0.668 vs 0.694 vs 0.699). In the individual views, 2 of 3 or 3 of 3 usually perform better than 1 of 3 (except for the F1 score in the AP view). In addition, 3 of 3 performs slightly lower than 2 of 3 in some metrics, which we believe that this is because the studies with three chest X-rays account for only a small percentage of the entire train dataset (8.5%). Therefore, there is less opportunity to learn the 3 of 3 input format during training. We can conclude that utilizing multiple X-ray views as input helps the model generate more accurate chest X-rays that can capture the abnormal findings in the report and other chest X-rays.

Table 6 shows the radiology report generation result. We also increase the input chest X-rays to generate the target report. In the S w/2 group, the 1 of 2 and the 2 of 2 show the same statistical score (BLEU-4 0.056), but the 2 of 2 shows a better clinical efficacy score (F1 0.415 vs 0.422). Also, in the S w/3 group, 3 of 3 outperforms 1 of 3 and 2 of 3 across all metrics. These results also show that using multi-view chest X-rays encourages the model to generate more precise reports.

5.3 The Ability to Generate Specific Views

View Position column in Table 8 indicates that UniXGen can generate the view-specific X-rays comparable to the original images. In addition, FID scores in Tables 2 and 4 imply that the generated chest X-rays are similar to their original view positions. These results demonstrate the usefulness of view-specific special tokens that effectively capture the specific properties of each view.

Table 4: Evaluations of generated chest X-rays using FID 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.

Group	Input	(src. $\rightarrow$ tar.)	FID (↓)			
			ALL	AP	PA	LAT.
S w/1	1 of 1	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	25.441	25.742	76.753	37.380
S w/2	1 of 2	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	16.416	24.907	16.779	20.185
	2 of 2	$(\mathbf{w}, \mathbf{v}^2 \rightarrow \mathbf{v}^1)$	**8.757**	**20.851**	**7.547**	**8.469**
		$(\mathbf{w}, \mathbf{v}^1 \rightarrow \mathbf{v}^2)$	6.583	9.312	5.838	9.511
S w/3	1 of 3	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	19.252	33.584	23.753	22.786
	2 of 3	$(\mathbf{w}, \mathbf{v}^2 \rightarrow \mathbf{v}^1)$	11.450	**19.781**	**11.853**	15.116
		$(\mathbf{w}, \mathbf{v}^1 \rightarrow \mathbf{v}^2)$	23.302	33.22	59.485	14.941
	3 of 3	$(\mathbf{w}, \mathbf{v}^2, \mathbf{v}^3 \rightarrow \mathbf{v}^1)$	**11.136**	19.865	12.002	**14.587**
		$(\mathbf{w}, \mathbf{v}^1, \mathbf{v}^3 \rightarrow \mathbf{v}^2)$	10.26	26.545	8.793	13.845
		$(\mathbf{w}, \mathbf{v}^1, \mathbf{v}^2 \rightarrow \mathbf{v}^3)$	8.83	19.973	9.507	11.444

Table 5: Evaluations of generated chest X-rays using 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.

Group	Input	(src. $\rightarrow$ tar.)	F1				micro AUROC			
			ALL	AP	PA	LAT.	ALL	AP	PA	LAT.
S w/1	1 of 1	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	0.457	0.460	0.615	0.255	0.747	0.751	0.728	0.563
S w/2	1 of 2	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	0.309	0.447	0.280	0.227	0.664	0.741	0.642	0.634
	2 of 2	$(\mathbf{w}, \mathbf{v}^2 \rightarrow \mathbf{v}^1)$	**0.400**	**0.485**	**0.362**	**0.370**	**0.713**	**0.753**	**0.692**	**0.702**
		$(\mathbf{w}, \mathbf{v}^1 \rightarrow \mathbf{v}^2)$	0.394	0.487	0.382	0.334	0.726	0.781	0.715	0.696
S w/3	1 of 3	$(\mathbf{w} \rightarrow \mathbf{v}^1)$	0.313	**0.396**	0.333	0.245	0.668	0.714	0.667	0.643
	2 of 3	$(\mathbf{w}, \mathbf{v}^2 \rightarrow \mathbf{v}^1)$	0.362	0.371	**0.376**	0.350	0.694	0.691	0.716	**0.679**
		$(\mathbf{w}, \mathbf{v}^1 \rightarrow \mathbf{v}^2)$	0.377	0.464	0.363	0.358	0.703	0.73	0.715	0.69
	3 of 3	$(\mathbf{w}, \mathbf{v}^2, \mathbf{v}^3 \rightarrow \mathbf{v}^1)$	**0.376**	0.394	0.373	**0.369**	**0.699**	**0.722**	**0.716**	0.673
		$(\mathbf{w}, \mathbf{v}^1, \mathbf{v}^3 \rightarrow \mathbf{v}^2)$	0.373	0.44	0.349	0.365	0.72	0.796	0.707	0.706
		$(\mathbf{w}, \mathbf{v}^1, \mathbf{v}^2 \rightarrow \mathbf{v}^3)$	0.381	0.505	0.333	0.353	0.705	0.749	0.688	0.704

Table 6: Evaluations of generated reports using BLEU and CheXpert scores to quantify the effect of using multi-view chest X-rays on radiology report generation. src. is short for source, and tar. for target.

Group	Input	(src. $\rightarrow$ tar.)	BLEU-4	CheXpert Labeler		
				Precision	Recall	F1
S w/1	1 of 1	$(\mathbf{v}^1 \rightarrow \mathbf{w})$	0.042	0.419	0.405	0.412
S w/2	1 of 2	$(\mathbf{v}^1 \rightarrow \mathbf{w})$	0.056	0.431	0.400	0.415
	2 of 2	$(\mathbf{v}^1, \mathbf{v}^2 \rightarrow \mathbf{w})$	**0.056**	**0.434**	**0.410**	**0.422**
S w/3	1 of 3	$(\mathbf{v}^1 \rightarrow \mathbf{w})$	0.054	0.461	0.413	0.435
	2 of 3	$(\mathbf{v}^1, \mathbf{v}2 \rightarrow \mathbf{w})$	0.060	0.447	0.424	0.435
	3 of 3	$(\mathbf{v}^1, \mathbf{v}2, \mathbf{v}^3 \rightarrow \mathbf{w})$	**0.063**	**0.471**	**0.431**	**0.450**

5.4 Comparison with Stable Diffusion

Table 7 shows the chest X-ray generation performances of UniXGen and the fine-tuned Stable Diffusion. For a fair comparison, our model generates chest X-rays using only radiology reports as input, without adding any additional chest X-rays. As seen from the table, UniXGen outperforms

the fine-tuned Stable Diffusion across all metrics. In addition, our model proves again that using additional chest X-rays can effectively generate more realistic and accurate chest X-rays as shown in Table 2 (FID 18.994 vs 10.436, F1 0.396 vs 0.416, AUROC 0.711 vs 0.728).

Table 7: Comparison of UniXGen and the fine-tuned Stable Diffusion for chest X-ray generation.

Models	FID (↓)	14-diagnosis Classification	
		F1	micro AUROC
Stable Diffusion	78.857	0.29	0.589
UniXGen	**18.994**	**0.396**	**0.711**

5.5 Human Evaluation

Table 8 confirms that UniXGen can generate realistic chest X-rays close to the original, and the view-specific special tokens can capture refined features of each view, encouraging the model to generate view-specific X-rays. In addition, our model scores higher than the baseline for both realism and alignment.

Table 8: Human evaluation for chest X-ray generation.

Models	Realism	Alignment	View Position (%)		
			AP	PA	LATERAL
Original Image	4.177	3.977	84.3	58.3	100
UniXGen	4.193	3.583	75.5	66.7	100
Stable Diffusion	2.09	1.827	-	-	-

5.6 Qualitative Examples

Figure 3 shows that UniXGen can generate realistic chest X-ray images even when conditioned only on the report, showing a small consolidation in the lingula as described by the report. When given an additional view, UniXGen generates an image that is more similar to the original image, showing its ability to take advantage of both input modalities. Figure 4, on the other hand, shows an example where UniXGen cannot correctly capture all the facts in the report when the image is generated conditioned on the report only. Although the report says "large right pleural effusion", the generated image depicts a rather small pleural effusion. When given an additional view, however, UniXGen 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 enables UniXGen to generate the desired views, even when they do not exist in reality.

Figures 5, and 6 show that using multiple X-rays encourages the model to generate more precise reports. Please refer to Appendix A for more examples. All examples are confirmed by the clinicians.

6 Limitation & Conclusion

In this paper, we propose for the first time a unified view-specific X-ray and report generation model. Our approach has several limitations, each providing opportunities for future work. First, due to the nature of the real-world patient dataset, the report contains many comparison phrases with previous studies (e.g. unchanged, increase, and compared to previous radiographs), so the model simply memorizes these patterns without understanding their meaning and generates historical phrases. In the future, we plan to use CXR-PRO [29], a refined dataset that removes historical phrases, to generate clinically accurate reports. Second, the human evaluation confirms that our model generates chest X-rays that are as realistic as the original images, but lack precise alignment with the report. We plan to improve the alignment between each sentence in the report and the image regions. 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. This will also be improved in the future. We hope that our approach suggests a new direction for medical synthetic data generation and can be useful in the medical domain.

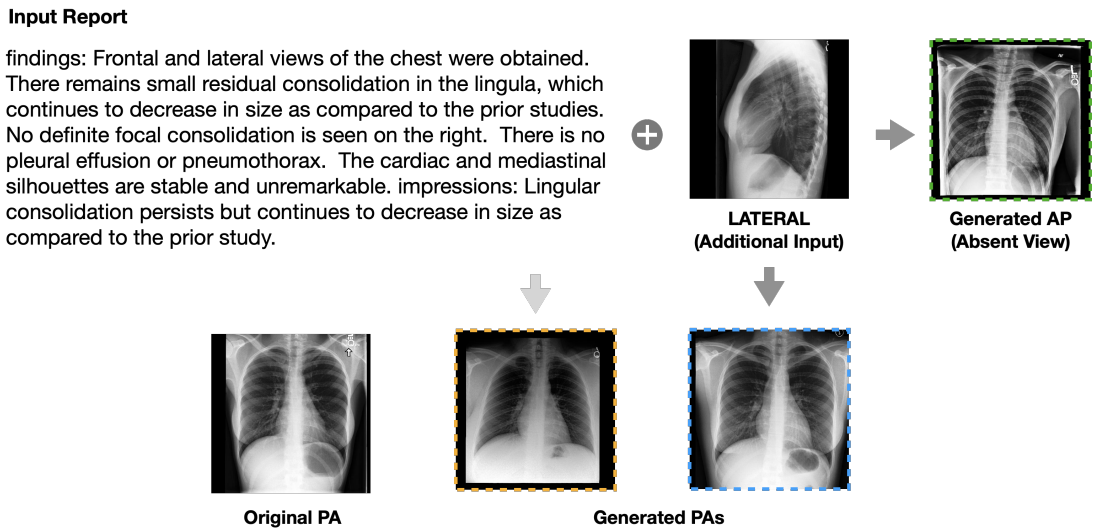

Figure 3: Generated chest X-rays of UniXGen. Based only on the report, the generated PA in the orange dashed box draws a rather small portion of the consolidation in the lingula, as is written in the report. Based on an additional lateral view, the generated PA in the blue dashed box draws a consolidation that is of more similar size as that of the original PA.

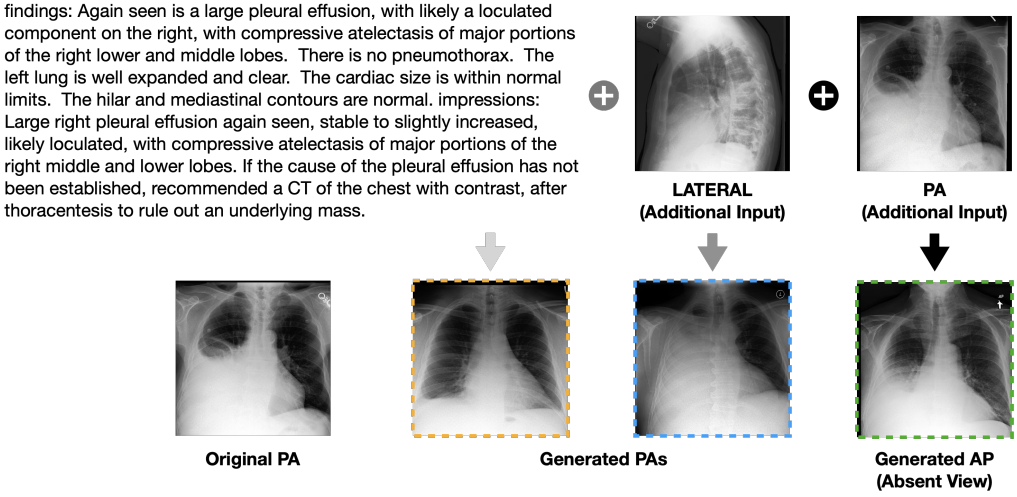

Figure 4: Generated chest X-rays of UniXGen. The generated PA conditioned only on the report (orange dashed box) draws relatively small-sized pleural effusion while the report says “large right pleural effusion”. However, by adding an additional lateral view (blue dashed box), UniXGen can properly generate the PA view with large pleural effusion.

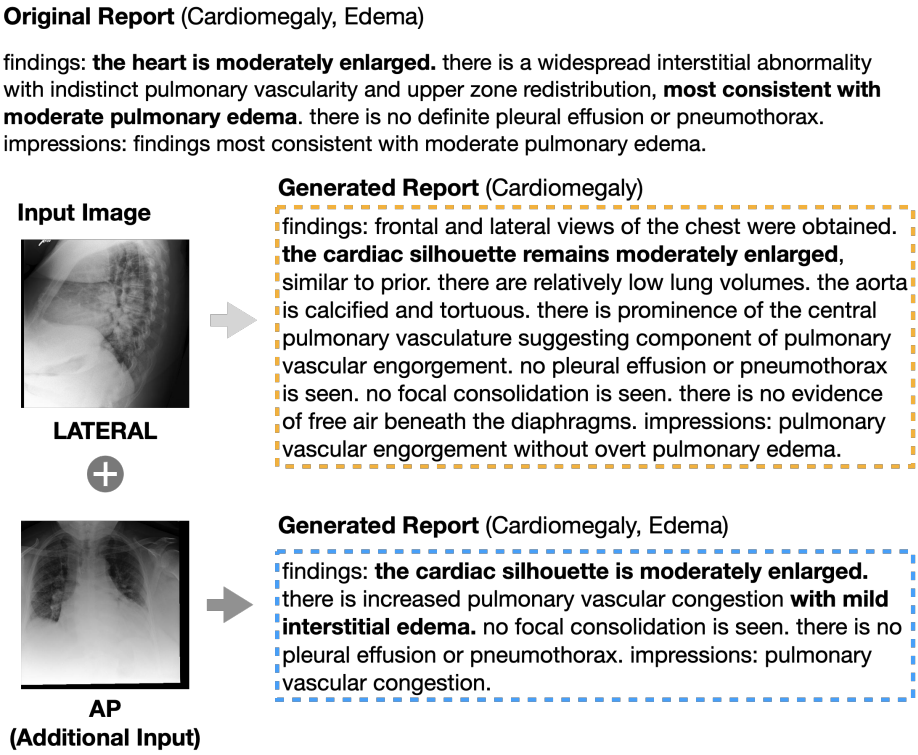

Figure 5: Generated radiology reports of UniXGen.

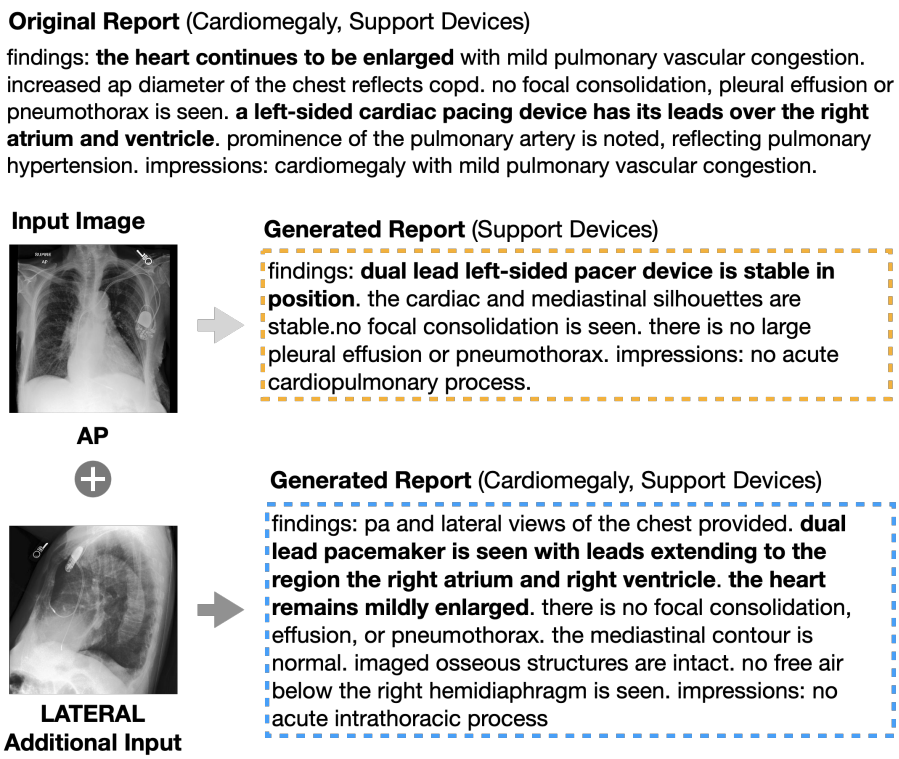

Figure 6: Generated radiology reports of UniXGen.

References

[1] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[2] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

[3] P. Chambon, T. S. Cook, and C. P. Langlotz. Improved fine-tuning of in-domain transformer model for inferring covid-19 presence in multi-institutional radiology reports. *Journal of Digital Imaging*, pages 1–14, 2022.

[4] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[5] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. 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] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] P. Esser, R. Rombach, and B. 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.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[10] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[11] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans. Axial attention in multidimensional transformers, 2019.

[12] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[14] L. Huang, G. Niu, J. Liu, X. Xiao, and H. Wu. Du-vlg: Unifying vision-and-language generation via dual sequence-to-sequence pre-training. *arXiv preprint arXiv:2203.09052*, 2022.

[15] Y. Huang, H. Xue, B. Liu, and Y. Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1138–1147, 2021.

[16] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. 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.

[17] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[20] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*, 2020.

[21] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[22] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical imaging 2018: Image processing*, volume 10574, pages 415–420. SPIE, 2018.

[23] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. 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.

[24] K. Packhäuser, L. Folle, F. Thamm, and A. Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. *arXiv preprint arXiv:2211.01323*, 2022.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. 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.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[28] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[29] V. Ramesh, N. A. Chi, and P. Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022.

[30] A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[32] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[33] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf. Diffusers: State-of-the-art diffusion models, 2022.

[36] C. Wang, K. Cho, and J. Gu. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160, 2020.

[37] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[38] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[39] X. Yang, M. Ye, Q. You, and F. Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. *arXiv preprint arXiv:2106.06471*, 2021.

[40] J. Yuan, H. Liao, R. Luo, and J. Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer, 2019.

[41] H. Zhang, W. Yin, Y. Fang, L. Li, B. Duan, Z. Wu, Y. Sun, H. Tian, H. Wu, and H. Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *arXiv preprint arXiv:2112.15283*, 2021.

[42] T. Zhang, H. Fu, Y. Zhao, J. Cheng, M. Guo, Z. Gu, B. Yang, Y. Xiao, S. Gao, and J. Liu. Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 777–785. Springer, 2019.

[43] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

A Generated Examples

Figures 7, and 8 show additional chest X-ray generation examples. Figures 9, and 10 show additional radiology report generation examples.

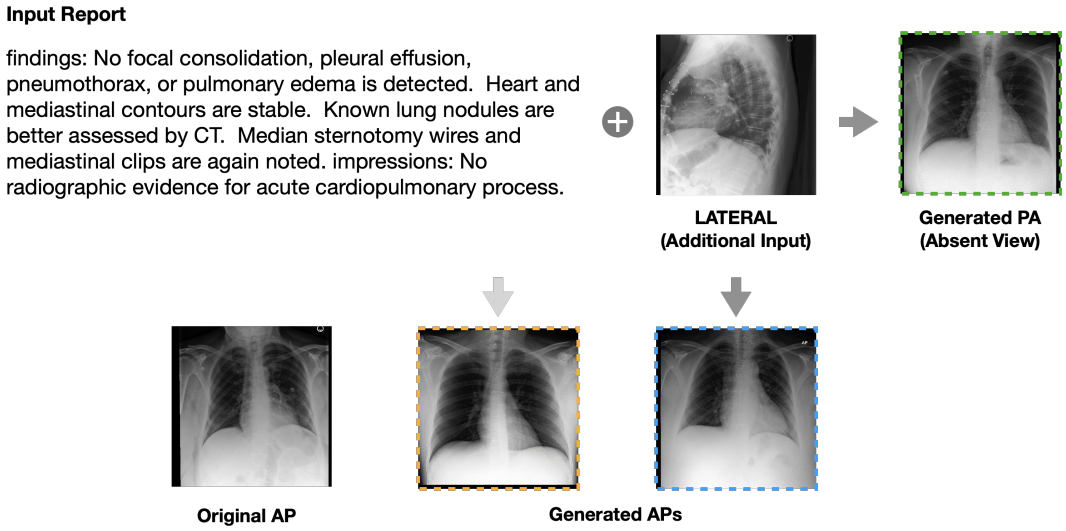

Figure 7: Generated chest X-rays of UniXGen. The AP view generated only from the report (orange dashed box) already reflects the facts described in the report quite well, capturing the wires and the clips. With an additional lateral view, the AP view in the blue dashed box shows an increased similarity with the original PA in terms of overall structure.

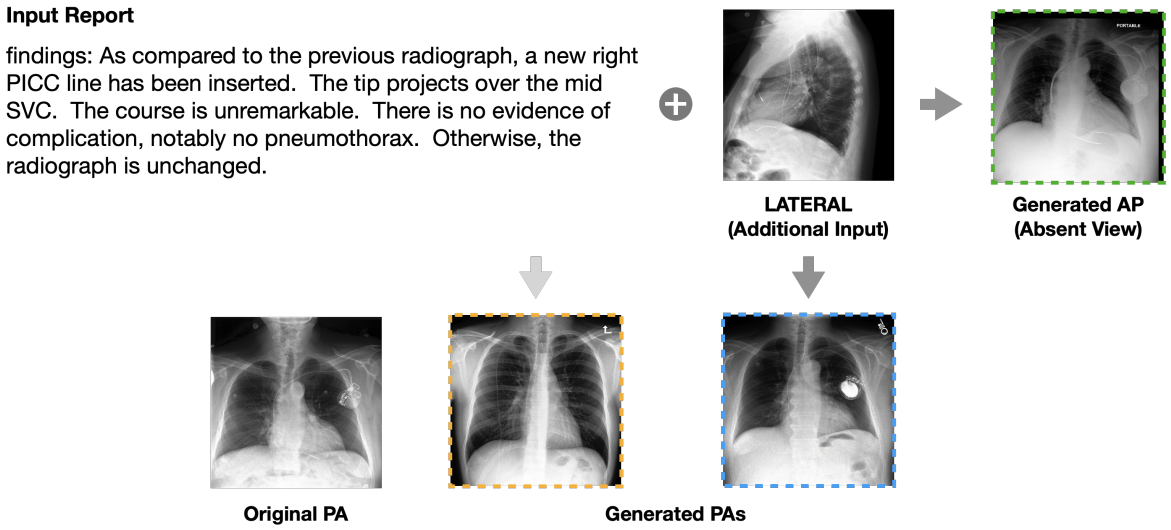

Figure 8: Generated chest X-rays of UniXGen. Although the PA view generated based only on the report (orange dashed box) captures most of the facts in the report, it is missing the implanted device. Given an additional lateral view, however, UniXGen can generate a PA view (blue dashed box) that correctly draws the implanted device.

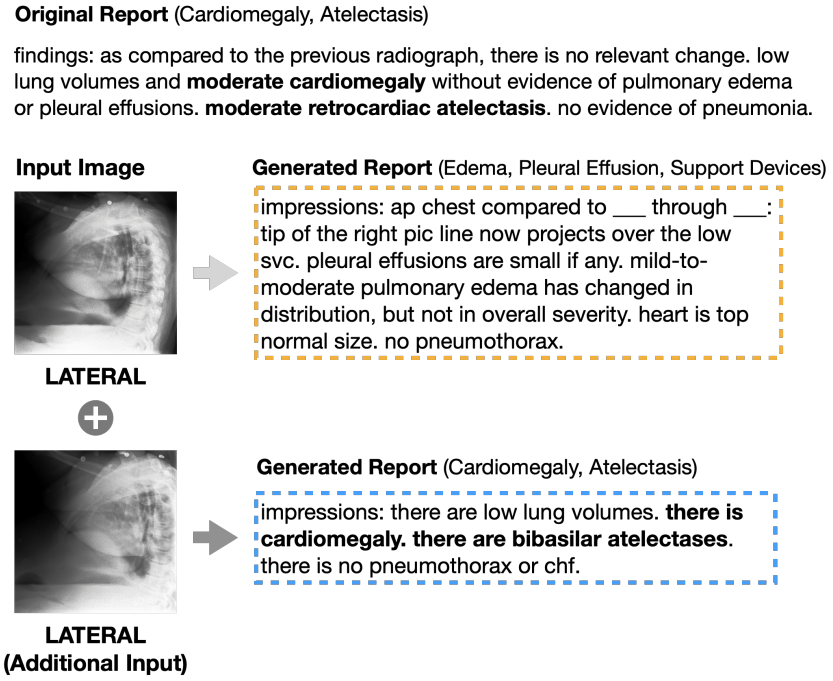

Figure 9: Generated radiology reports of UniXGen

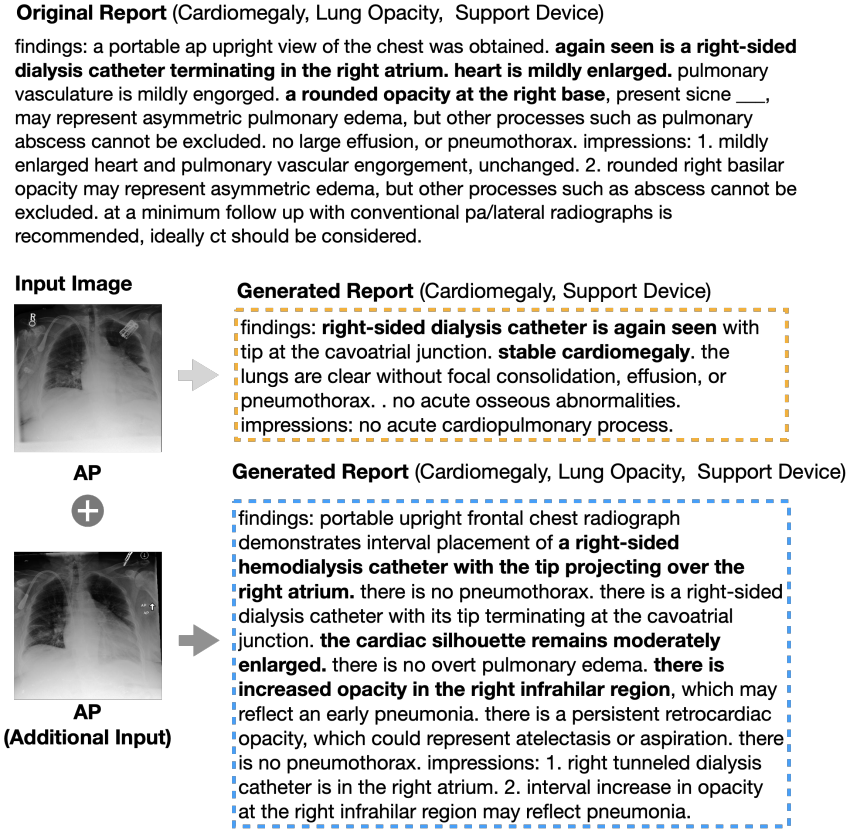

Figure 10: Generated radiology reports of UniXGen

B Causal Attention Mechanism

Figure 11 shows a visual explanation of the causal attention mechanism.

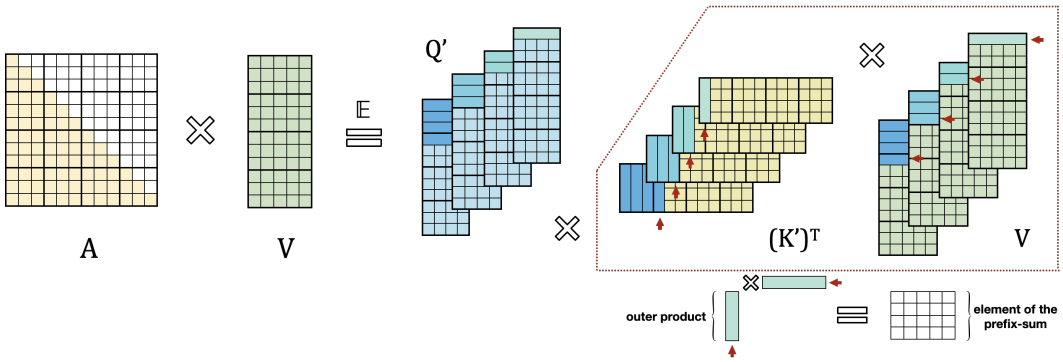

Figure 11: Visualization of multimodal causal attention mask and its prefix-sum algorithm. Illustrations inspired by [4]. This algorithm builds the prefix-sum of the outer-products of random feature maps for keys $\mathbf{K}'$ and value vectors $\mathbf{V}$ on the fly and left-multiplies it with the query random feature vector $\mathbf{Q}'$ to obtain the new row in the resulting matrix $\mathbf{AV}$.

C Details of the Fine-tuned Stable Diffusion

Following [1], we fine-tune Stable Diffusion [31] as a baseline model for chest X-ray generation. Since the implementation code has not been released, we re-implement it with the descriptions provided in the original paper.

The code is built on the diffusers library [35]. We load the stable diffusion pipeline from the Hugging Face repository [38] (CompVis/stable-diffusion-v1-4).

To evaluate reproducibility, we use only PA views, limit the number of "No finding" reports, and select only the Impression section. We train three different models: (a) train both U-Net and CLIP [26], (b) train U-Net from scratch and freeze RadBERT [3], (c) train U-Net from scratch and freeze SapBERT [20]. The models are trained for 1k steps with batch size of 256, learning rate of 5e-5. The FID scores for each model are 23.534, 68.162, and 76.39, respectively. We measure the FID score with chest X-ray pretrained DenseNet-121 [5] and P19 testset. Figure 12 shows our re-implementation results.

For a fair comparison with our model, we fine-tune Stable Diffusion with multiple views and both Findings and Impression sections. We replace the CLIP text encoder with SapBERT to handle both Findings and Impression sections (the CLIP tokenizer is only limit to 77 tokens) and keep frozen the text encoder and VAE and only train U-Net from scratch. The model is trained as the same hyper-parameters with our model. We evaluate the generated chest X-rays every 1k steps and select the best performance model. We try to make this model learn specific view information by prompting the view position along with the report, but this model fails to learn the view position. Thus, we train this model without the specific view information. Figure 13 shows the generated chest X-rays.

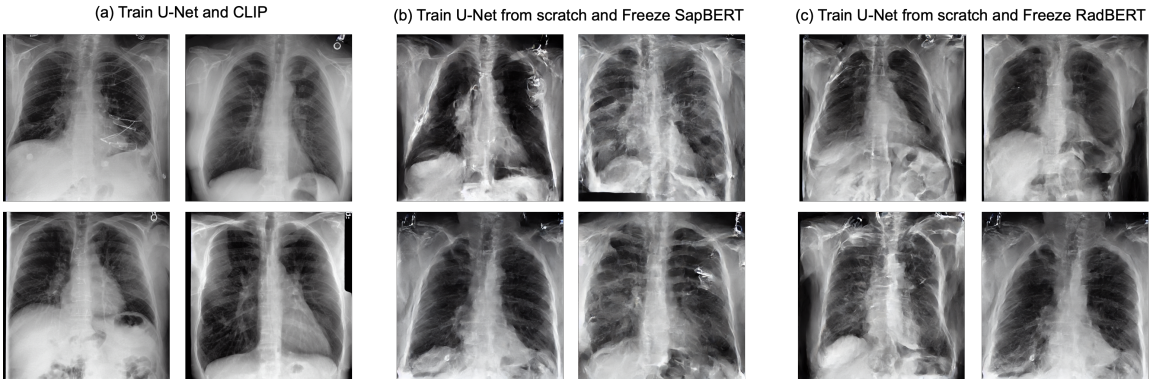

Figure 12: Re-implementation results of the fine-tuned Stable Diffusion as proposed in [1].

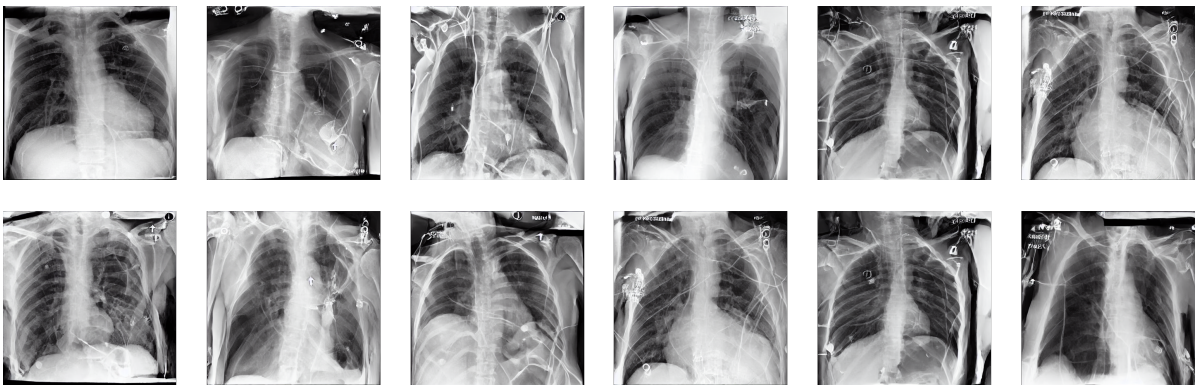

Figure 13: The generated chest X-rays of the fine-tuned Stable Diffusion with our dataset.