# TEXT-TO-IMAGE RETRIEVAL AND GENERATION IN RADIOLOGY

**Jenny Chen, Kayla Hauessler, Rishika Randev**
ECE685D, Duke University

## ABSTRACT

Radiologists frequently need to reference past cases with similar findings when interpreting chest X-rays, but current retrieval systems rely on manual annotation and keyword-based search that fail to capture similarity between images and reports. This project develops a cross-modal retrieval system for chest X-rays and radiology reports by using a shared embedding space where semantically similar content, regardless of modality, is positioned nearby. Using the CheXpert dataset of 224,316 chest X-rays with associated clinical observations, vision and text encoders were fine-tuned through contrastive learning, maximizing similarity for matching image-text pairs while minimizing it for non-matching pairs. The system achieved recall@5 values of 64.5% for image-to-text and 68.2% for text-to-image retrieval, substantially outperforming the random baseline (2.13%) and demonstrating bidirectional consistency. Additionally, a conditional GAN was trained using the learned text embeddings to generate synthetic chest X-rays, achieving a Fréchet Inception Distance of 47.62 on training data and 97.33 on validation data, comparable to existing medical imaging generation literature. These results demonstrate that the approach successfully captures meaningful semantic alignment between radiographic images and clinical reports, enabling intuitive cross-modal search and synthetic image generation that could support diagnostic workflows, medical education, and clinical research.

## 1 INTRODUCTION

In the modern health care ecosystem, radiologists review and interpret countless numbers of chest X-rays. When they encounter unusual or atypical findings, it is often helpful to reference past cases with similar characteristics. Medical imaging databases contain very vast collections of chest X-rays paired with accompanying radiology reports. Currently, images and text reports are often stored and indexed separately, and retrieval systems often rely on manual annotation and keyword based searching to connect reports with images (Tagare et al., 1997). However, these methods fail to capture semantic similarity between scans and their text descriptions, possibly limiting searches to return all truly relevant cases.

Bridging this gap has the potential to significantly improve diagnostic efficiency and radiologist confidence. Additionally, it provides a valuable educational resource for medical students and residents learning to interpret radiological scans, and can facilitate clinical research by enabling large-scale retrieval of cases with specific characteristics.

Developing such systems introduces several challenges. Most image models are trained on natural image datasets, and may not directly transfer to the medical imaging domain with its vastly different visual features (Zhang et al., 2022). Additionally, our goal is to extend beyond a basic keyword search and learn a shared embedding space where semantically similar items are close together regardless of whether they originate as text or image.

Thus this paper focuses on developing a cross-modal retrieval system that retrieves relevant chest X-ray images given a text query, or relevant radiology text descriptions given an image query. The central challenge is aligning representations from both modalities so that semantically similar content is embedded nearby, enabling intuitive and clinically meaningful search across images and text.

This paper achieved this goal through fine-tuning of vision and text encoders to learn an aligned vision-language representation space. Utilizing contrastive loss, the encoders were trained to minimize distance for matching image-text pairs and maximize for non-matching pairs; recall@K was used to evaluate cross-modal retrieval capabilities based on the embeddings from the trained encoders. Finally, a generative model was trained to create realistic chest X-ray images based on input text describing specific lung pathologies and observations.

## 2 Previous Work

### 2.1 Chest X-Ray Classification

In 2019, researchers at Stanford University published the CheXpert dataset and paper, pivotal in the field of radiology image classification using deep learning (Irvin et al., 2019). This team designed CNN-based classifiers to automatically detect the presence of 14 different observations in chest X-rays with high accuracy, and the dataset is now widely used as a standard benchmark to evaluate the performance of chest radiograph interpretation models. Expanding on the use case of classification of X-ray conditions, we employed cross-modal embedding retrieval on text-image pairs from CheXpert.

### 2.2 Cross-Modal Retrieval in Medical Imaging

Cross-modal retrieval has garnered significant attention and exploration with the surge of data spanning different modalities across technologies. Various methods exist, from unsupervised and supervised real-value retrieval to hash-based retrievals; however, a notable problem in applying this technique to medical imaging has involved a shortage of annotated images (Wang et al., 2024). Recent work to address this includes applying unsupervised contrastive learning, such as through ConVIRT, which learns visual representations by exploiting naturally occurring paired descriptive text and image data (Zhang et al., 2022). ConVIRT maximizes the agreement between true image-text pairs by applying a bidirectional contrastive objective between the paired image and text data, performing better than baselines and requiring only 10% as much labeled data compared to previous methodologies. ConVIRT inspired subsequent frameworks like CLIP, which proposed an efficient method of learning from natural language supervision by learning a multi-modal embedding space; their approach was to jointly train an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings pairs while minimizing the cosine similarity of incorrect pairings (Radford et al., 2021). They then optimized a symmetric cross entropy loss over the similarity scores. In our project, we applied methodologies such as contrastive loss from ConVIRT and CLIP towards X-ray image and report text pairings to optimize cross-modal retrieval.

### 2.3 Generative Models

Generative Adversarial Networks (GANs) have been used in medical literature to augment and expand the training data for deep learning models. For example, Motamed et al. (2021) employed a GAN architecture to augment chest X-rays for semi-supervised detection of pneumonia and COVID-19 and found an improvement in classification accuracy of the diseases. Outside of the medical space, there has been use of GANs for cross-modal retrieval tasks, with Fréchet Inception Distance (FID) as a common evaluation metric (Bithel & Bedathur, 2023). Combining the two, we evaluated the performance of our multi-modal embedding retrieval of chest X-ray image-text pairs by utilizing GANs to generate images from text and comparing them to real images with FID.

## 3 Methods

### 3.1 Dataset

This project used the CheXpert dataset, which contains 224,316 chest X-rays from 65,240 patients who visited inpatient and outpatient centers at Stanford University Medical between 2002 and 2017. CheXpert is publicly accessible through the Stanford AIMI, and our work used a compressed version of it found on Kaggle. Each chest X-ray image in the dataset is accompanied by the patient's

age, sex, view of the X-ray (whether frontal / anteroposterior or lateral), and a set of 14 labels extracted from the actual radiology report associated with the X-ray. 12 of these labels corresponded to whether or not a specific pathology was found to be present by radiologist evaluation (enlarged cardiomegaly, cardiomegaly, lung lesion, lung opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, and pleural other). The last two labels corresponded to whether or not "no findings" or support devices were observed, respectively. Column values for these labels were either 1 (present), 0 (absent), or -1 (uncertain, indicating some degree of ambiguity in the report related to that observation). These labels were extracted using a rule-based labeler developed for the original CheXpert paper (Irvin et al, 2019). Our training dataset consisted of 223,415 samples, while the validation set was 235 samples.

### 3.2 Pre-Processing

For each sample, the 14 observation labels had to be converted into a concise string of text so that it could be fed into a text encoder to generate an embedding. This string of text was formatted as a small summary report, which contained the demographic information of the patient from whom the sample was collected (age and sex), the X-ray view, and then a comma-separated list of findings (including whether "no findings" was explicitly observed). For any samples that were missing all labels or had 0 values for all labels (including "no findings"), the list of findings simply consisted of "demonstrates no acute cardiopulmonary abnormality." Before adding any uncertain labels to the list of findings, the word "possible" was added prior to the label/observation name to clearly indicate a difference in the confidence around that observation. After this summary report was generated for each sample, it was tokenized using BioClinical BERT's associated tokenizer, with padding and a max length of 512 tokens, and the resulting token IDs and attention mask were used as input for the text encoder.

For chest X-ray images, pre-processing consisted of resizing the images to 256x256 so they matched the standard input dimensions for ResNet50, and then normalizing by RGB channel based on the means and standard deviations of the original ImageNet data; the resulting image tensors were used as input for the vision encoder.

### 3.3 Encoders

Our text encoder used a BioClinical BERT backbone, while our vision encoder used a ResNet50 backbone. BioClinical BERT is a specialized version of BERT that was fine-tuned on electronic health record data from the MMIMIC III database; while BERT-base was also experimented with, the clinical version led to better retrieval capabilities and thus it was used for our final text encoder. ResNet50 is trained on natural images and thus not specific to medical data (ImageNet). Both backbones were connected to their own projection heads which returned 512-dimensional text and vision embeddings, respectively. The 512-dimensional embedding space was selected to balance representational capacity with computational efficiency, following standard practices in vision-language models like CLIP.

For each training sample, the text labels and image were pre-processed appropriately, and then fed into their respective encoders, with the resulting embeddings used to calculate contrastive loss (described in the next section) and train the encoders. Training was conducted for 4 epochs using the whole training dataset, with a batch size of 32 and an Adam optimizer; differentiated learning rates of 3e-5, 5e-5, and 1e-3 were used for training the text encoder backbone, vision encoder backbone, and projection heads, respectively. Model checkpoints were saved every 1000 iterations, and training and validation set contrastive loss, as well as recall@1 and recall@5 metrics for both image-to-text and text-to-image retrieval, were tracked across training. The text and vision encoders from the checkpoint with the lowest validation loss, and highest recall (averaged across the two K values and two directions) were used to get our final recall@K values, as well as to train the conditional GAN.

### 3.4 Contrastive Loss

To learn aligned representations, contrastive loss is used following the CLIP framework (Radford et al., 2021). The objective maximizes similarity between matching image-text pairs while minimizing similarity for all non-matching pairs within each batch.

Given a batch of N paired images and texts, L2-normalized embeddings $\mathbf{I} \in \mathbb{R}^{N \times d}$ and $\mathbf{T} \in \mathbb{R}^{N \times d}$ are obtained from the vision and text encoders, respectively. The cosine similarity matrix is computed as

$$\mathbf{S} = \frac{\mathbf{I}\,\mathbf{T}^{\top}}{\tau}, \text{where } \tau = 0.2 \text{ is the temperature parameter that scales the logits.}$$

The temperature parameter $\tau = 0.2$ was selected to balance gradient stability and discriminative power during training at our batch size, as lower values (like 0.07 used in CLIP) led to unstable learning. Each row $\mathbf{S}_{i,:}$ represents the similarities between image i and all N texts, and the model is trained to assign the highest probability to the correct index i. This is formalized as cross entropy loss for image-to-text ($\mathcal{L}_{\mathrm{I}\to\mathrm{T}}$ ), and text-to-image($\mathcal{L}_{\mathrm{T}\to\mathrm{I}}$ ) matching. The final contrastive loss averages both directions: ($\mathcal{L} = \frac{\mathcal{L}_{\mathrm{I}\to\mathrm{T}} + \mathcal{L}_{\mathrm{T}\to\mathrm{I}}}{2}$).

### 3.5 RECALL@K

Cross-modal retrieval performance was evaluated using Recall@K, a standard metric that measures the proportion of queries for which the ground truth match appears in the top K retrieved results. Formally, Recall@K is the fraction of test queries for which at least one correct match is found in the top K predictions, defined as:

$$\mathrm{Recall@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[y_i \in \mathcal{T}_i^{(K)}\right], \text{where } y_i \text{ is the ground truth and } \mathcal{T}_i^{(K)} \text{ the top-}K \text{ predictions.}$$

During evaluation, all images and texts in the validation set were encoded using the trained encoders. For each query, the similarity score to all candidate items in the opposite modality were computed. All items of the opposite modality were then ranked by similarity to determine whether the ground truth match appeared within the top K positions. Recall was one of the metrics used to monitor performance during training. Once the text and vision encoders were fully trained, the final recall@K was computed for $K \in \{1, 5, 10\}$ for both image-to-text and text-to-image retrievals, and is reported in the Results section.

### 3.6 GAN AND FRÉCHET INCEPTION DISTANCE

Our conditional GAN used the text embeddings from the trained text encoder as input to the generator, which then outputted fake images. The discriminator took an image and the text embeddings and returned a logit. The generator architecture consisted of one fully connected layer and three deconvolution layers with batch normalization, ReLU activation, and then one final deconvolution with Tanh activation to upsample in order to output a 128x128 image. The discriminator consisted of four convolutional layers with batch normalization, Leaky ReLU, and then two linear layers to output its prediction. Both were trained using binary cross entropy loss over 5 epochs and an Adam optimizer, with a learning rate of 2e-4 for the generator, 1e-4 for the discriminator, and betas (0.5, 0.999) (See section 5.2 for additional detail on hyperparameter selection). Checkpoints were saved every 5000 iterations. Images were sized to 128x128 to reduce compute time with all other pre-processing the same as above.

Fréchet Inception Distance (FID) was used to evaluate generation quality. It is calculated by comparing the distribution of generated images to the distribution of real images. Given the real image feature distribution $\mathcal{N}(\mu_r, \Sigma_r)$ and the generated image feature distribution $\mathcal{N}(\mu_g, \Sigma_g)$, the Fréchet Inception Distance (FID) is defined as:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\Big).$$

Generated and real images were passed through a pre-trained Inception-V3 model to extract feature vectors. The feature vectors for real and fake images were treated as samples from a multivariate Gaussian, and the distributions' means and standard deviations were calculated for each. We then took 50,000 samples of real images from the validation set and 50,000 generated from our GAN and calculated the Fréchet Distance.

# 4 Experimental Results

## 4.1 Retrieval

Recall@K values from evaluation in both directions on the full validation set are reported below.

Table 1: Recalls for retrieval in both directions

| DIRECTION | RECALL@1 | RECALL@5 | RECALL@10 |
|---|---|---|---|
| Image-to-Text | 25.2% | 64.5% | 81.6% |
| Text-to-Image | 26.5% | 68.2% | 77.4% |

In general, recalls for retrieval using our multi-modal embedding system are similar across both directions. For reference, a random retrieval system would have recalls of 0.42%, 2.13%, and 4.26%, respectively, for K = 1, 5, and 10. By substantially outperforming a random baseline with only limited training, and by demonstrating consistent performance across both image-to-text and text-to-image retrieval, our system has likely captured some degree of real semantic alignment between the X-rays and their reports.

## 4.2 GAN and Fréchet Inception Distance

Real images from the validation set and images generated from the conditional GAN using the corresponding text embeddings are shown below.

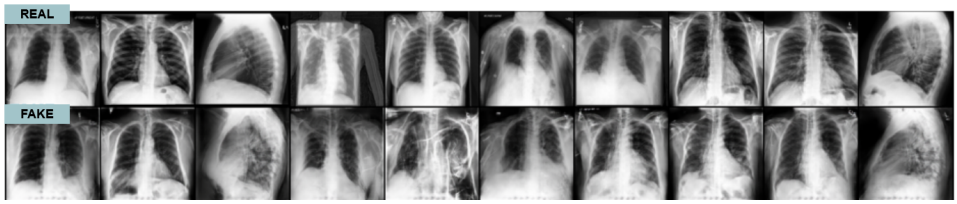


Figure 1: Real images and corresponding fake images from generator.

In general, images match similarly upon observation, with the generator's images aligning with the correct orientation and axis based on the text report. Using 50,000 samples from the training set, we achieved a FID of 47.62, which is comparable to results in literature (Motamed et al., 2021) and represents a strong performance for medical imaging generation. On the validation set, the FID had a higher value of 97.33, though this is still substantially lower than the initial baseline we achieved of 137 when trained on random text embeddings. Overall, the observed image generation and reasonable FID suggest that the conditional GAN performed well and the generator was able to produce realistic images based on the text provided.

# 5 Discussion

## 5.1 Encoders

The final recall values evaluated on the validation set show moderate success when it comes to retrieval in both directions, and there are several possible reasons for this. One major reason could be that we used ResNet, a CNN model pretrained on ImagNet, as the backbone for our vision encoder, so its ability to capture the nuances in medical images could be limited. Using a computer vision model already fine-tuned on medical imaging data could possibly lead to better quality image embeddings. It is also possible that the integration of triplet loss alongside contrastive loss could improve our encoders. This loss uses triplets consisting of an anchor (for example, an image embedding), a positive (correct text embedding), and a negative example (incorrect text embedding), so that the model can learn to distinguish between especially difficult or similar positive and negative examples (Schroff et al., 2015). Finally, due to time constraints we were only able to train our

encoders for 4 epochs; additional training time is likely to improve recalls further given that the loss and recalls had not fully plateaued by the end of our experiment.

### 5.2 GAN AND FRÉCHET INCEPTION DISTANCE

The GAN training process revealed an important limitation in using adversarial loss as a quality metric for GANs. Despite severe loss imbalance, with discriminator loss converging to 0.08 and generator loss exceeding 8.0, the model achieved competitive FID scores and produced anatomically plausible X-rays. This discrepancy suggests that the discriminator learned to exploit subtle artifacts imperceptible to humans, achieving high classification accuracy without preventing the generator from learning the overall data distribution (Brock et al., 2019). This finding aligns with recent work showing that adversarial losses can be misleading indicators of perceptual quality (Salimans et al., 2016), and highlights the importance of using distribution-based metrics like FID and human evaluation rather than loss values alone when assessing generation quality. However, even FID has limitations as an evaluation metric for medical imaging. While FID measures distributional similarity in feature space, it may not correlate with human perceptual quality or capture clinically relevant features such as fine-grained anatomical accuracy (Borji, 2018), (Bithel & Bedathur, 2023). This emphasizes the need for complementary evaluation approaches, such as retrieval-based metrics or radiologist assessment, in future work.

Additionally, due to our initial observations of severe loss imbalance (discriminator loss  0.08, generator loss  8.0), we attempted continued training beyond epoch 5 to see if extended optimization would improve generation quality. To address the dominance of the discriminator, we modified the training loop to update the discriminator only once every three iterations rather than every iteration, while maintaining the same learning rates. However, this adjustment led to rapid deterioration in image quality after epoch 5, despite the losses becoming more balanced (epoch 8:discriminator 0.45, generator  2.21). This outcome suggests that while the discriminator's dominance appeared problematic based on loss values, it was actually providing crucial gradients that guided the generator toward anatomically plausible outputs. By weakening the discriminator through less frequent training, we inadvertently removed this guidance, allowing the generator to exploit shallow patterns that fooled the weaker discriminator but failed to capture meaningful anatomical features. This reinforces our earlier observation that adversarial loss values are unreliable quality indicators in medical imaging GANs, and highlights the risk of over-correcting training dynamics based solely on loss imbalance.

## 6 CONCLUSION

This project focused on the projection of chest X-rays and their associated radiologist-annotated observations to a 512-dimensional embedding space, ultimately enabling cross-modal retrieval (relevant X-rays based on observations/pathologies, and vice versa). This provides a way for radiologists to find a more expansive set of relevant past cases and images based on observations they are searching for, capitalizing on semantic similarity as opposed to just simple keyword matching; finding cases that are semantically similar, even if not visually similar, can help provide a broad array of examples of what certain pathologies look like to radiologists in training. This approach also pre-trains representations for downstream tasks like X-ray classification. Our training of a conditional GAN using text embeddings to generate synthetic chest X-rays gave us an additional assessment of the quality of our shared embedding space; such a GAN can also serve as a potential on-demand source of realistic X-rays corresponding to different pathologies, which could be useful in an instructional setting. While our recall@K and FID values have room for improvement, they demonstrate a strong start towards the creation of a common embedding space between radiographic images and their corresponding reports, as well as an efficient way to retrieve and generate one of these modalities from the other.

### ACKNOWLEDGMENTS

We thank Dr. Vahid Tarokh, Mehak Arora, and the rest of the course staff for guidance throughout the project. We also thank the Data Science department at Duke University for supporting our work and acknowledge the use of Duke University's computational resources for model training. Finally, we thank the Stanford ML Group for providing the CheXpert dataset.

## REFERENCES


Shivangi Bithel and Srikanta Bedathur. Evaluating cross-modal generative models using retrieval task. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 1960–1965, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591979. URL https://doi.org/10.1145/3539618.3591979.

Ali Borji. Pros and cons of gan evaluation measures, 2018. URL https://arxiv.org/abs/1802.03446.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. URL https://arxiv.org/abs/1809.11096.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. URL https://arxiv.org/abs/1901.07031.

Saman Motamed, Patrik Rogalla, and Farzad Khalvati. Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images. *Informatics in Medicine Unlocked*, 27:100779, 2021. ISSN 2352-9148. doi: https://doi.org/10.1016/j.imu.2021.100779. URL https://www.sciencedirect.com/science/article/pii/S2352914821002501.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. URL https://arxiv.org/abs/1606.03498.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823. IEEE, June 2015. doi: 10.1109/cvpr.2015.7298682. URL http://dx.doi.org/10.1109/CVPR.2015.7298682.

Hemant D. Tagare, Christopher C. Jaffe, and James Duncan. Medical image databases: a content-based retrieval approach. *Journal of the American Medical Informatics Association*, 4(3):184–198, 1997. doi: 10.1136/jamia.1997.0040184.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions, 2024. URL https://arxiv.org/abs/2308.14263.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2022. URL https://arxiv.org/abs/2010.00747.


## A APPENDIX

### A.1 CODE REPOSITORY

The full implementation, including model training, experiments, and evaluation scripts, is available at: https://github.com/khaeuss808/ECE685D_FinalProject