Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-modal foundational models are trained on millions of pairs of natural images and texts, frequently obtained through web-crawling approaches. Although their performance is excellent, these models do not generalize well to other domains, such as medical imaging, especially when these domains do not resemble the centric-like images that can be found on the web. In this study, we assess the ability of the stable diffusion model to generate domain-specific images in the particular case of medical imaging. Based on quantitative and qualitative evaluations of the main components of the stable diffusion pipeline (the variational autoencoder, the U-Net and the text-encoder), we explore several approaches to fine-tune stable diffusion to generate radiological images, which accurately represent the clinical content of conditional text prompts. Our best-performing model improves upon the stable diffusion baseline and can be correctly conditioned to insert an abnormality on a synthetic radiology image.

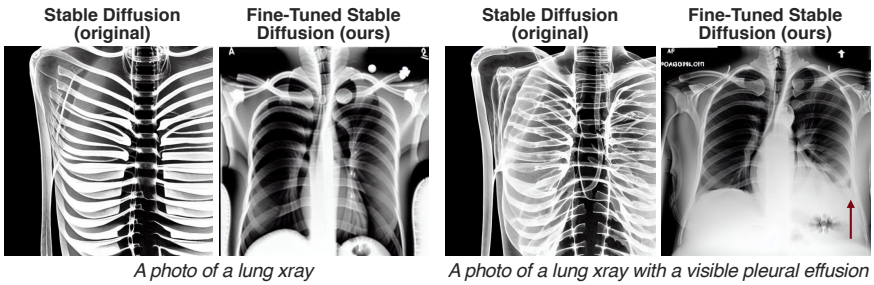

Figure 1: Generated images by both the original stable diffusion model and our fine-tuned model on radiology images. The prompts are designed to compare a standard radiology image with no particular findings, and the insertion of a frequent abnormality "pleural effusion" (red arrows).

1 Introduction

In recent months, latent diffusion models have gained immense popularity by enabling state-of-the-art image generation amenable to fine-grained control of the image generation process at inference time via conditioning of the denoising process (e.g., using text prompts) (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). Such models, termed foundation models (Bommasani, 2021), have been trained with large multi-modal curated datasets such as LAION-5B that consists of natural images and their captions (Schuhmann et al., 2022). The impressive generative capabilities of such models permits creation of high-fidelity synthetic datasets that may be used to augment traditional supervised machine learning pipelines in scenarios that lack training datasets.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

One particular area that such an advance would be beneficial in is the domain of medical imaging, where there is a paucity of high-quality labeled datasets. Annotating such medical imaging datasets typically requires trained medical experts who are capable in interpreting subtle, but semantically meaningful, image features. Despite the lack of large curated medical imaging datasets, one benefit that such medical imaging examinations have is that there is typically a text-based radiology report that describes pertinent findings from the imaging study. Leveraging the vision-language understanding capabilities of latent diffusion models could potentially provide an intuitive mechanism to create synthetic medical imaging data by prompting with relevant medical keywords or concepts of interest.

In this study, we explore the representational bounds of large vision-language foundation models and evaluate how to utilize pretrained foundational models to represent medical imaging studies and concepts, despite models never having been explicitly trained on these concepts. We utilize chest x-rays (CXRs) for this study as they are most common imaging modality globally. CXRs are fast to acquire, inexpensive, can provide important patient health insights, and can identify and monitor a variety of pathologies. We explore and quantify the representational capacity of the stable diffusion model (Rombach et al., 2022) to characterize the efficacy of both its language and vision encoders as applied to CXRs. We further explore different strategies for improving the representational capacity of non-domain-specific foundational models for representing medical concepts specific to CXRs. These experiments help provide novel decision making insights regarding whether such foundational models can accurately represent complex biomedical concepts for clinically-relevant downstream tasks, without explicit training on such concepts. In this study, we specifically show the following:

1. Training Stable Diffusion on LAION-2B learns a variational autoencoder (VAE) that can reconstruct CXR images out-of-the-box
2. A frozen CLIP text encoder can generate powerful medical embeddings with enough clinical context to allow accurate generated images, in conjunction with the methods below
3. Replacing the frozen CLIP encoder with a frozen in-domain text encoder with a projection head trained on LAION to map in-domain embeddings to CLIP embeddings, is not adequate to generate better images
4. Textual inversion can be used to learn complex medical concepts like pleural effusion in a few-shot manner
5. Fine-tuning the UNet component enables high-fidelity CXR image generation with the capability to insert custom pathologies (see examples in Figure 1).

We verify all our findings using using quantitative metrics of image quality as well as qualitative and domain-specific radiological interpretation from an expert thoracic radiologist.

2 Materials and Methods

2.1 Datasets

A large, publicly available chest x-ray (CXR) dataset (MIMIC-CXR, version 2.0.0) was used in this work, under institutional review board approval (Johnson et al., 2019). MIMIC-CXR contains a total of 377,110 images from studies performed at the Beth Israel Deaconess Medical Center in Boston, MA, USA, of which 700 frontal (i.e., anterior-posterior or posterior-anterior projection) radiographs were sampled randomly for this study. These images and their associated reports were used for experiments and study of the variational autoencoder and of text encoders.

In addition, we manually select 5 images with no findings, as well as 5 images that have visible pleural effusion, discarding any improperly cropped or colorized images (verified by a thoracic radiologist). Along a set of simple prompts generated synthetically, these form pairs of images and texts that are used for fine-tuning the stable diffusion model with various approaches. Finally, a sample of one million text prompts from the LAION-400M dataset (Schuhmann et al., 2021) is used for textual projection training and experiments.

2.2 Stable Diffusion

The stable diffusion model (depicted in Figure 2) is composed of a CLIP text encoder that parses text prompts to create a 768-dimensional latent representation (Radford et al., 2021a). This latent text

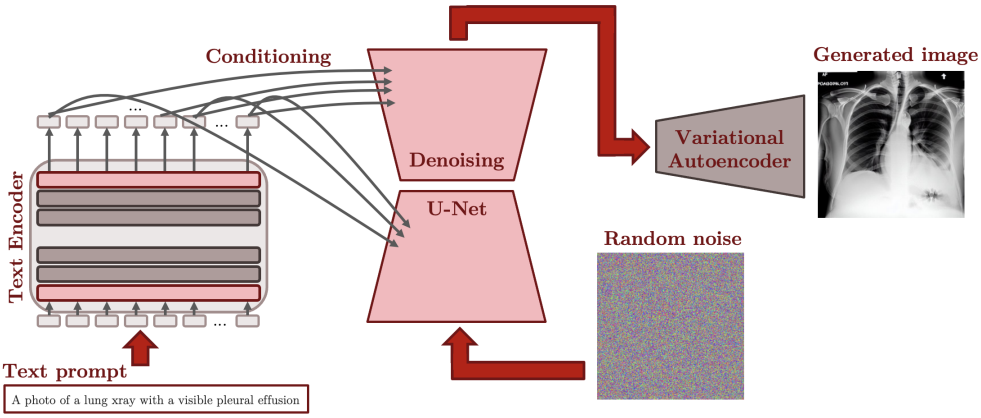

Figure 2: Stable diffusion architecture, run in the radiology setting to generate synthetic radiology images.

representation is used to condition a generative U-Net to generate images in the latent image space using random noise as an additional conditioning. Finally, the decoder component of a variational autoencoder is used to map this latent image projection to the output image space. While the original generative model has been trained with image and text captions arising from natural imaging domains, the extent of its capabilities for representing medical concepts and images remains unclear. To adapt the stable diffusion model for in-domain image generation, especially for radiology images and prompts, we can leverage each component and train it, or not, depending on its capabilities to represent in-domain data. More particularly, we can assess:

- Whether the variational autoencoder (VAE) alone is capable of reconstructing radiology images without losing general visual aspect as well as clinically important features.
- Whether the text encoder alone is capable of projecting clinical prompts to the text latent space while preserving clinically important features.

Section 2.3 presents the methods used to assess the reconstruction quality of the VAE, assessing whether it requires in-domain fine-tuning; Section 2.4 describes the experiments researching the quality of the CLIP text encoder and other in-domain text encoders; and Sections 2.5, 2.5, 2.7 present methods to fine-tune various components of the stable diffusion model for the radiology domain.

2.3 Variational Autoencoder

As latent diffusion model, stable diffusion translates image inputs into a latent space before performing the denoising process, using an encoder trained to remove perceptually negligible features ("perceptual compression")(Rombach et al., 2022). To analyze how well medical imaging information is preserved while passing through the VAE, CXR images sampled from MIMIC ("originals") were encoded to latent representations and reconstructed into images ("reconstructions").

Reconstruction quality was quantitatively assessed by calculating the root mean square error (RMSE), the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) for each image-reconstruction pair. Additionally, the Fréchet inception distance (FID, underlying model: Inception V3, 2048 features) was calculated on minibatches (batch size = 32) to compare the distribution of reconstructions to the distribution of original images(Szegedy et al., 2015; Heusel et al., 2017).

Qualitatively, the reconstruction quality compared to the original image input was assessed by a radiologist with 7 years of experience in reading CXR studies, using a scoring system ranging from 1 to 5 (5: Very good reconstruction with essentially non-inferior diagnostic quality to the original, 4: Good reconstruction with noticeable errors not negatively influencing diagnostic quality, 3: Moderate reconstruction errors with possible negative effects to diagnostic performance, 2: Severe reconstruction errors or errors of any level leading to hallucinated lesions, 1: Severe reconstruction errors yielding the image undiagnostic) on 100 randomly sampled original-reconstruction pairs.

The effect of the reconstruction process on classification performance was analyzed using a model pre-trained to detect 18 different pathologies commonly encountered in CXR (DenseNet-121, torchxrayvision library, version 0.0.37)(Cohen et al., 2022)(Cohen et al., 2020). Classification accuracy and F1 score were calculated for 12 of the labels included in both MIMIC-CXR and the pretrained model. For this step, uncertain findings (='-1') were considered positive findings, while missing values were treated as absence of the corresponding finding. Additionally, latent representations of original and reconstructed images were compared by calculating their pairwise cosine similarity.

2.4 Text Encoder

In the domain-specific setting of radiology reports and images, the goal is to be able to condition the generation of images on associated medical conditions, that can be represented through a text prompt or report. Therefore, the capability of the text encoder to correctly represent medical features in the latent space is critical for the rest of the stable diffusion process, in particular the U-Net operating in the latent space, to be able to generate images that are anatomically correct and representing the correct set of abnormalities.

A set of potential text encoders that could be interesting to accurately represent medical features was found through study of the previously published pre-trained language models in the field: PubMedBERT (Gu et al., 2022), ClinicalBERT (Huang et al., 2019), SapBERT-from-PubMedBERT-full text (Liu et al., 2021), RadBERT (huggingface.co/StanfordAIMI/RadBERT), CXR-BERT-general (Boecking et al., 2022), CXR-BERT-specialized (Boecking et al., 2022) and finally the Clip text encoder (Radford et al., 2021b).

As described in section 2.1, we can gather radiology report data from CXR, and the corresponding abnormality labels as output by the CheXpert model (Irvin et al., 2019). Then, for each particular $text\_encoder$ model and the corresponding $report\_list$ of elements $report$, one can run the $report$ through the model and get a representation $text\_encoder(report)$. Nevertheless, there exist several ways to extract embeddings from these text encoders, all based on a transformer architecture: extracting the last layer hidden state of the associated CLS token, "CLS hidden state"; extracting the last layer hidden states of each tokens and averaging these representations, "mean hidden states"; using the pooler output, "pooler output"; Using a model specific extraction method, if available, "model specific".

The combination of a $text\_encoder$ model and the associated extraction method $extraction$ gives a function $extraction \circ text\_encoder$ that takes an input report and outputs a document-level representation. This way, for a defined $text\_encoder$ model and $extraction$ method, one can obtain the document-level embeddings of radiology reports and then assess the quality of these embeddings and therefore the capability of a text-encoder to encode medical content.

For the evaluation, we first obtain the document-level embeddings on the impression section of each radiology report, obtained through regex parsing. This gives:

$$impression\_embeddings = extraction \circ text\_encoder(impression_sections)$$

For all the text-encoders that we study, the latent representations are of dimension 768. Therefore for 700 impression sections, $impression\_embeddings$ is a 700×768 matrix.

Then, we can compute the impression-impression similarities in the latent space

$$similarities = impression\_embeddings \times impression\_embeddings^T$$

We then compute a metric, that we denote the $CheXpert@k$ metric, that for each report i find the k most similar reports, and then measure the proportion of reports that share the same CheXpert label. If $chexpert\_labels$ is a list of the chexpert labels corresponding to the reports, we have:

$$CheXpert@k_i = \frac{sum(chexpert\_labels[argsort(similarities[i])[-k:]] == chexpert\_labels[i])}{k}$$

And then over all reports we get $CheXpert@k = \frac{\sum_{i=1}^{n} CheXpert@k_i}{n}$

Notice that in the implementation of this metric, a filter is added to $CheXpert@k_i$ so that among the k most similar reports, the report being compared to is not retrieved. In addition, the metric $CheXpert@k$ can be computed over each class instead: so for each abnormality class, we can average the $CheXpert@k_i$ scores, where the similarities are still computed over the reports of all classes. A macro-averaged score can then be retained for comparison purposes.

2.5 Textual Projection

Building up upon the stable diffusion work, we propose as a first method to generate domain-specific images to replace the CLIP text encoder, kept as frozen during the stable diffusion original training, with a domain-specific text encoder, typically pre-trained on biomedical or radiology data. The goal behind this architecture change is to hopefully rely on the better understanding the new text encoder has of radiology inputs and therefore provide better latent representations, that the U-Net will then be conditioned upon to generate synthetic images.

Nevertheless, simply replacing the CLIP text encoder with a new one should lead to catastrophic performance, given that latent spaces can be structured in a very different manner. There are no guarantee that any latent feature is redundant between the two text encoders. We therefore propose to train a projection capable of translating, in parts, the latent representations of one text encoder to the other. So that running radiology prompt through the in-domain text encoder, and then projecting these latent representations through this trained projection, should allow embeddings to be well-enough aligned for the U-Net conditioning to work, but still provide enhanced clinical representations through the in-domain text encoder added knowledge.

To train this projection, we use the LAION-400M dataset and define a $projection$ as a MLP model, taking a 768-dimensional input and mapping it to a 768-dimensional translated output. As a first approach, we take $projection = Linear \circ ReLU \circ LayerNorm \circ Linear$ and train it using MLE loss. At inference time, images can be generated by using the in-domain text encoder along the projection, and hopefully having enough clinical features passing through while keeping most of the CLIP latent space structure so that the U-Net conditioning allows for clinically correct generated images.

Notice that the prompts the model is trained on can have an impact on the performance, that we try to measure: we explore object-oriented prompts of the form "a photo of a ..." and style-oriented prompts of the form "a photo in the style of a ...", with lexical variants of these two base prompts.

2.6 Textual Embeddings Fine-tuning

Following the approach of Gal et al. (2022), the stable diffusion model can be further trained to generate better looking images for the radiology setting by focusing on the embeddings of the text encoder. In this case, during training, the variational autoencoder, the U-Net, as well as all the other layers of the text encoder are frozen. In addition, a new token gets introduced, that can either describe: patient-level features, such as gender, age and body weight; procedure-level features, such as body part and modality; abnormality-level features, such as "no findings" or "pleural effusion".

As an example, we could introduce the token $< lung - xray >$ that is supposed to describe both a body part, lungs, and a modality, X-ray. This learning approach, denoted *Textual Inversion*, zero out all the gradients associated with the embeddings of the already existing tokens, and in the end only learn the embedding of this newly introduced token.

Then, during training, input prompts with these new tokens are introduced, along associated radiology images. The rest is very similar to original training of the stable diffusion model, in that the model gets used to generate a synthetic image, and the noise at several timesteps in both the forward and backward process of the U-Net are passed through a MSE loss. Gradients are then used to only update the embeddings of the newly introduced tokens.

2.7 U-Net fine-tuning

Finally, in a similar approach to Ruiz et al. (2022), one can improve the baseline stable diffusion model to generate better domain-specific images by relying on a U-Net fine-tuning. Instead of switching text encoders and using a projection (see Section 2.5) or training the embeddings of new tokens (see Section 2.6), one could keep all components frozen and the original CLIP text encoder, to only further train the U-Net part. In this sense, the setting is very similar to the approach of Section 2.5, except that no new token gets added, and the freezing is over the set of parameters of the U-Net.

Then, the training is similar to the training of the original stable diffusion model, relying on MSE loss at several time steps of the denoising process to progressively converge to better generation of in-domain images.

3 Results

3.1 Training details

Experiments were conducted on several devices, depending on their compute hungriness. VAE and text encoder experiments were run in local, with M1 Pro and M1 Max GPUs. Textual projection relied on 3 NVIDIA Quadro P5000 GPUs, with a single run taking 3 hours for 10k training steps in the case of document-level training, and 8 hours for 10k training steps in the case of token-level training, when using only one of these GPUs. Textual embedding fine-tuning and U-net fine-tuning used a NVIDIA V100 GPU and took respectively 1 hour for 3k training steps and 15 minutes for 400 training steps.

To conduct our experiments and in particular access model weights, we relied heavily on the *Hugging Face* library (Wolf et al., 2019) and the recently released *diffusers* (von Platen et al., 2022). The stable diffusion weights we used come from the *CompVis/stable-diffusion-v1-4* repo. Weights of other in-domain text encoders are the ones associated with each corresponding publication.

3.2 Variational autoencoder

700 CXR images from MIMIC were encoded and decoded using the pretrained VAE from the Stable-Diffusion-v1.4 pipeline. Quantitative assessment showed a low reconstruction error (RMSE 41.0±8; median, 41.4; range, 20.4 - 76.3; PSNR, 33.6±1.8; median 33.3; range, 28.1 - 39.5) and a high structual similarity of original and reconstructed images (SSIM, 0.92±0.02; median, 0.93; range, 0.8 - 0.96). See Figure 3 for details. Image quality metrics did not depend on the class labels of the images (data not shown).

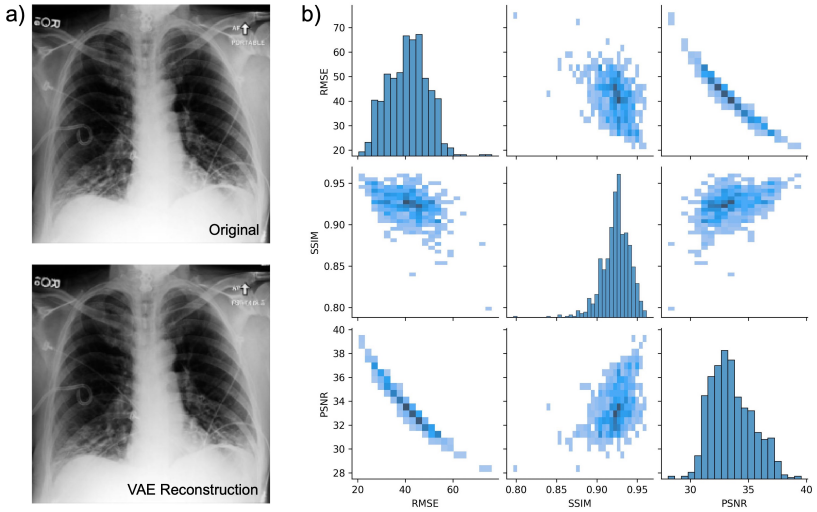

Figure 3: Image reconstruction analysis. a) Original (top) and reconstructed (bottom) image. The small burnt-in annotations in the top right corner get scrambled (seen in almost all samples), while the vast majority of other features (e.g., rib contours, devices) are well-preserved. b) Distribution of image quality metrics assessed for each image-reconstruction pair. RMSE: Root mean square error. SSIM: Structural similiarity index measure. PSNR: Peak signal-to-noise ratio.

Visual analysis yielded a generally good perceived reconstruction quality (Mean visual score 4.51±0.54; median score, 5; range, 3 - 5). No reconstruction resulted in a completely non-diagnostic image (score 1) or altered the diagnostic information in a potentially problematic way (score 2). Almost all burnt-in text annotations were scrambled beyond recognition, however, diagnostic features were well preserved in almost all cases. Most of the score deductions were for blurred device components, cerclages and wires that couldn't be traced reliably after reconstruction, or blurred rib contours.

Table 1: Classification results for original and reconstructed CXR images from the MIMIC-CXR dataset

Label	Prevalence	Acc. (%) original	Acc. (%) recon.	Acc. (%) %change	F1 (%) original	F1 (%) recon.	F1 (%) %change
Atelectasis	33.3	40.1	40.7	1.4	52.0	52.4	0.7
Cardiomegaly	34.0	45.9	47.4	3.4	54.4	54.8	0.7
Consolidation	13.1	22.9	23.9	4.4	25.0	25.2	1.0
Edema	20.7	37.1	42.7	15.0	39.6	41.3	4.4
En. Mediastinum	13.1	23.4	23.6	0.6	23.4	22.8	-2.7
Fracture	2.4	15.7	19.0	20.9	4.5	4.7	3.9
Lung Lesion	4.6	21.3	25.1	18.1	5.8	5.8	-1.0
Lung Opacity	32.4	39.4	39.6	0.4	50.0	49.6	-0.8
Pleural Effusion	38.1	51.4	52.7	2.5	60.6	61.3	1.1
Pleural Other	2.1	55.4	50.3	-9.3	5.5	4.9	-9.8
Pneumonia	14.6	30.0	33.9	12.9	26.9	25.4	-5.3
Pneumothorax	9.1	24.3	24.7	1.8	18.0	18.0	0.5

The reconstruction process negatively impacted the classification performance for the "Pleural Other" label (accuracy 50.3% for reconstructed vs. 55.4% for reconstructed and original images, respectively). Interestingly, most other labels were predicted with similar (Atelectases, Cardiomegaly, Enlarged Cardiomediastinum, Lung Opacity, Pleural Effusion, Pneumothorax) or higher accuracy (Edema, Fracture, Lung Lesion, Pneumonia) from the reconstructed images. See Table 1 for details. The latent embeddings generated by the pretrained DenseNet-121 pairs were highly similar for image-reconstruction pairs (mean cosine similarity, 0.99±0.01; median, 0.99; range, 0.94 - 1.00).

3.3 Text Encoder

Various text encoders and associated embedding methods are assessed on radiology reports in order to evaluate which method can retain optimum clinical knowledge in the latent representations.

Following the definitions in section 2.4 of the $text\_encoder$ models, the $extraction$ methods and the metric $CheXpert@k$, we compute in Table 2 for each model and each method the macro-average of the $CheXpert@k$ score aggregated per abnormality class, with $k = 10$. As seen in the table, the method "CLS hidden state" is in general the one that works best to maximize the quality of the document-level representations of the impression sections. In addition, the model CXR-BERT-specialized is the one that reaches highest performance, taking for each model the corresponding $extraction$ method that worked best.

Then, using the $extraction$ method that works best for each model, we can compute class-wise $CheXpert@k$ scores as well as the macro-averaged ones. These results are aggregated in Table 3. As a baseline, we use a bag-of-words approach that outputs a similarity measure between two reports using an intersection over union measure. This baseline does not create any embeddings, but provide a token-based similarity measure: we observe that the latent representations of the best models, on top of contracting the text space, better encode document-level content and result in higher scores.

We remark that PubMedBERT, ClinicalBERT and CXR-BERT-general are three models that perform significantly less well than the other models, and should therefore, if possible, not be preferred for tasks that involve radiology reports. On the contrary, the two best performing models are CXR-BERT-specialized and the CLIP text encoder. As CLIP text encoder was not specifically trained on radiology reports, this underlines the quality of the training and the associated model. Using CXR-BERT-specialized instead would only improve performance by +15%.

For these reasons, we explore the textual projection with CXR-BERT-specialized, but also assess CLIP performance to be high enough to not justify replacing the text encoder in the various textual inversion and U-Net fine-tuning experiments.

Table 2: Macro-average of $CheXpert@k$ scores computed per abnormality class, over the impression sections of a set of radiology reports. Models that are better are retaining medical features get a higher score.

Model	CLS hidden state	Mean hidden states	Pooler output	Model specific
PubMedBERT	**30.8**	23.6	20.6	None
ClinicalBERT	26.3	**35.1**	14.3	None
SapBERT	**49.1**	48.7	41	None
RadBERT	**54.2**	32.8	34.7	None
CXRBERTgeneral	**32.4**	25.4	31.6	None
CXRBERTspecialized	**61.1**	34.5	None	50.3
ClipTextEncoder	7.0	42.8	**52.9**	None

Table 3: For each text encoder and the associated best $extraction$ method as computed in Table 2, class-wise and macro-averaged $CheXpert@k$ scores are computed. Higher scores denote better capability at retaining important clinical features in the structure of the latent space.

Abnormality	Base	Pub.	Clin.	Sap.	Rad.	gen.	spe.	Clip.
Atelectasis	33.4	21.8	19.2	54.2	53	23.4	**64.2**	52.8
Cardiomegaly	21.6	20.8	10.2	51	53	21.6	**67.2**	47.6
Consolidation	36	13.4	35.8	**39.6**	38	35.4	27	38.4
Edema	62.8	54	62.6	64.6	67.2	47.4	**85.4**	72
Enlarged Cardiomediastinum	38	21.2	30.2	41.8	**44.8**	35.2	37.6	42.6
Fracture	49	36.2	35.6	73.2	72.6	50.8	**83.2**	74.2
Lung Lesion	30.2	24	21.2	32	37.8	24.8	**56.2**	33.8
Lung Opacity	20.4	16.2	20.6	20.4	**34.2**	20.4	23.2	25.6
No Finding	78.4	**82.2**	75.4	74.8	79.8	75.4	76.8	80.6
Pleural Effusion	46.4	25	39.4	42.6	65.8	24.2	**72.2**	68
Pleural Other	21.6	13.6	17.8	36	43.4	16.8	**54**	34.6
Pneumonia	53.8	33.8	40.4	42.6	44.4	24	45	**54**
Pneumothorax	56.4	39.6	60.6	65.2	73.6	28.6	**92.8**	72
Support Devices	32.6	29.2	23	49.6	50.8	25.8	**70.4**	44.8
Macro	41.5	30.8	35.1	49.1	54.2	32.4	**61.1**	52.9

3.4 Radiology Image Generation

Comparing the various methods introduced in Section 2, we use the Fréchet inception distance as introduced in Section 2.3 to measure the quality of the reconstructed images. The results are compiled in Table 4, along an empirical sample of images as produced by each method in Figure 4.

For the most simple prompt "A photo of a lung xray", progress is done only with the last method that consists in training the U-Net. In particular, no progress is observed with the token embedding training (also known as textual inversion). For more complex prompts such as "A photo of a lung xray with a visible pleural effusion", the stable diffusion baseline shows limitations, being outperformed by both textual inversion and U-Net fine-tuning.

The textual projection does not seem to converge well enough: samples from Figure 4 shows the generated images to be out-of-domain. Nevertheless, we estimate that a more complex architecture, instead of our simple 1-hidden-layer projection, could be worth exploring: if projection-based domain-adaptation turns out to produce interesting examples, this could open the door to very quick domain-adaptation for the large amount of pre-trained text encoders that are now available.

Out of all the methods, the U-Net fine-tuning seems by far the most promising: it gets the lowest FID-scores and obviously the most realistic outputs. Nevertheless, we notice that this underlines the limitations of our non-medical-based metric: samples clearly show that U-Net fine-tuning with prior leads the model to learn the difference between "no findings" and "pleural effusion", something a model trained without a prior can not do. As seen in Table 4, FID fails to capture this improvement. We assess that further progress in the domain-specific generation of images for radiology would

Table 4: Evaluation of the quality of generated images with different methods for adapting stable diffusion to the radiology domain. Scores represent the Fréchet inception distance (FID), and lower scores mean better generated images.

Training Strategy	A photo of a lung xray	A photo of a lung xray with a visible pleural effusion	A photo in the style of a lung xray
Original model			
Stable Diffusion	0.097	0.151	
Textual Projection			
CXR-BERT-specialized			
No Projection	0.124	0.144	
Document-level projection	0.266	0.104	
Token-level projection	0.201	0.257	
Token embedding training			
Object, radiology	0.108	0.058	0.092
Object, lung	0.135		0.135
Style, radiology	0.101	0.057	0.084
Style, lung	0.130		0.083
U-Net training			
Trained on no findings	0.057	0.043	
Trained on no findings and abnormality	**0.034**	**0.041**	
Trained on no findings and abnormality with prior	0.170	0.086	

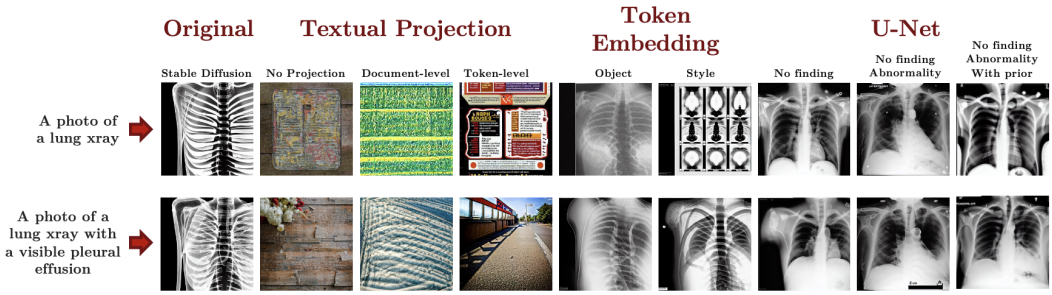

Figure 4: Images generated by various methods conditioned on radiology text prompts.

require the design/use of domain-specific metrics, that would be able to capture the ability of the model to correctly insert abnormalities that are coherent with the conditioning text prompt.

4 Conclusion

In this paper, we assessed the recently released stable diffusion model, including its variational autoencoder, the U-Net and the associated CLIP text encoder, and its capacity to produce clinically relevant images based on prompts that describe observable abnormalities. We conducted quantitative and qualitative evaluations, showing that: the variational autoencoder is powerful enough to reconstruct radiological images, including abnormalities and clinically relevant features; the CLIP text encoder accurately represents simple radiology-specific text prompts, outperforming 4 out of the 6 reviewed domain-specific text encoders. We explored textual projection, a domain-adaptation method that we designed, textual inversion and U-Net fine-tuning, and, with the latter, obtained a model capable of generating synthetic radiology images that are visually and quantitatively exceeding the baseline, and that can correctly represent abnormalities.

Building upon this work, we would like to further explore the potential of diffusion-based model to learn a wide-range of abnormalities, being able to combine them, as well as extending the research to other modalities and body parts. A limitation of our approach is that the employed metrics have limited capacity to assess the clinical correctness of the generated images. In addition, our fine-tuned stable diffusion model lacks diversity in the images it generates, probably due to the small range of samples they were trained on. Finally, the text prompts the models are conditioned on are synthetic and do not fully correspond to the wording used in the clinical setting, so that models capable of being conditioned on entire or partial radiology reports are an area of future research.

References

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing, 2022. URL https://arxiv.org/abs/2204.09817.

Rishi et al Bommasani. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction, 2020. URL https://arxiv.org/abs/2002.02497.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022. URL https://github.com/mlmed/torchxrayvision.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, jan 2022. doi: 10.1145/3458754. URL https://doi.org/10.1145%2F3458754.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL http://arxiv.org/abs/1706.08500.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019. URL https://arxiv.org/abs/1904.05342.

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.

Alistair E. W. 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), December 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online, June 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.naacl-main.334.

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. *CoRR*, abs/2103.00020, 2021a. URL https://arxiv.org/abs/2103.00020.

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, 2021b. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01042. URL https://doi.org/10.1109/cvpr52688.2022.01042.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022. URL https://arxiv.org/abs/2208.12242.

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, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. URL https://arxiv.org/abs/2111.02114.

Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, mehdi cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Mitchell Wortsman, Richard Vencu, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019. URL https://arxiv.org/abs/1910.03771.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Will be included in the non-double-blinded submission.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We included data collection, splits, compute details, number of training steps, code details when they could be explained at a high-level. We did not include hyperparameter details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] When relevant, especially for the autoencoder experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] We mentionned the reference, where the license can be found.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] No new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] No consent was needed.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Data from only already publicly available datasets was used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] Not applicable
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] Not applicable