M.I.G.R.O.S. Budget: Medical Image Generation, Retraining On a Strict Budget

Luca Drole Marina Crespo Aguirre Luca Anceschi

Abstract

In the context of medical imaging, generating synthetic data is a useful strategy to tackle data scarcity issues. To this end, fine-tuning Latent Diffusion Models to the medical domain represents a promising avenue, with Low-Rank Adaptation offering a compromise between computational cost and image quality. In this project, we extend the application of the recently proposed Self-Expanding LoRA (SeLoRA) algorithm on multiple Magnetic Resonance Imaging domains. To the best of our knowledge, we introduce the first data augmentation pipeline featuring a SeLoRA model. Our results suggest that SeLoRA can outperform more traditional image generation algorithms on relatively small training sets, generating images that hold diagnostic value for cancer imaging. Our code is available on github.

1. Introduction

Fully harnessing the capabilities of Deep Learning (DL) models is only possible when a sufficient number of training samples is used [2]. This is not always the case in the context of pathologic medical data due to privacy concerns or challenges in producing annotated data [20, 10]. Data augmentation has proven to be an effective way to address this issue [29]; consequently, the attention in the medical imaging field is turning to advanced techniques for generating synthetic images [9, 11]. In contrast to photorealistic images, where generative models may exhibit some degree of error, medical images need to be both anatomically and diagnostically accurate [7]. In the context of image generation, Latent Diffusion Models (LDMs) have gained popularity in recent years. Compared to more conventional generative approaches (GANs, VAEs), LDMs sample images from more complex and wider distributions, without the issues of training instability or mode collapse [14, 30]. Nonetheless, re-training an LDM to a particular data distribution is too computationally expensive [27], and in the context of medical data, often unpractical [18]. Additionally, the ability to perform image generation in the low-data regime is particularly meaningful in the case of rare diseases [21] or for institutions with limited resources. These considerations underscore the importance of establishing data-efficient and computationally feasible domain adaptation strategies for LDMs [15, 33, 24]. In this report, we seek to explore Low-Rank Adaptation as a strategy to efficiently fine-tune latent diffusion models in the context of medical imaging. We take particular interest in validating the recently proposed Self-Expanding Low-Rank Adaptation (SeLoRA) algorithm, presented by *Mao et al.* [22]. We seek to extend its domain of application to different Magnetic Resonance Imaging (MRI) data sources. While the SeLoRA preprint offers promising insights, it fails to evaluate the quality of the images on downstream tasks. We seek to comprehensively assess the quality of images generated from this procedure while providing a direct comparison with vanilla LoRA and GAN-based techniques and ensuring a degree of comparability with previous works exploring alternative approaches (*De Wilde et al.* [7]).

2. Models and Methods

2.1. Related Work

In the context of medical images, the adaptation of large diffusion models pre-trained on natural images, has proven to be effective for a wide range of domain-specific applications [6]. Textual Inversion (TI) [33] and Low-Rank Adaptation (LoRA) [15] are some of the strategies used to inject domain specificity to the models. *De Wilde et al.* [7] proposed to use TI on prostate MRI, assessing the quality of their images on a classification task. LoRA is based on optimizing small rank matrices injected into the frozen model layers to approximate the weights update [1, 15]. Higher ranks of the trainable matrices lead to improved result quality but come with increased computational cost. In this context, numerous studies have explored this trade-off, aiming to identify the optimal rank selection [31, 34, 8].

2.1.1. SeLoRA

In a recent development, Self-Expandable Low-Rank Adaptation [22] proposes a dynamic rank expansion method. More particularly, the output of the model when performing a forward pass in a SeLoRA layer can be formulated as:

$$f(x) = xW_0 + xAB + b_0,$$

Where $A \in \mathbb{R}^{d_{\text{in}} \times r}$ and $B \in \mathbb{R}^{r \times d_{\text{out}}}$ correspond to the low-rank trainable matrices, that get expanded $r \rightarrow r+1$ when the model improves its performance on noise prediction. This is assessed by the Fisher Information (FI) ratio between the FI score of consecutive time-step LoRA matrices:

$$\text{FI-Ratio} = \frac{\text{FI-Score}_r}{\text{FI-Score}_{r+1}}.$$

The FI score for a particular LoRA matrix is formulated as:

$$\text{FI-Score} = \sum_{i=1}^{d_{\text{in}}} \sum_{j=1}^{r} \Omega_{A_{i,j}} + \sum_{i=1}^{r} \sum_{j=1}^{d_{\text{out}}} \Omega_{B_{i,j}},$$

Where Ω_w represents the sensitivity of the model prediction to the parameters in the LoRA matrix, or:

$$\Omega_w = \frac{1}{|S|} \sum_{i=1}^{S} \left(\frac{\partial}{\partial w} \mathcal{L}(x_i; w) \right)^2$$

Only when the FI ratio is above a manually predefined threshold λ, should an update on the rank of that particular matrix be performed. Larger thresholds contribute to a decrease in model flexibility since the matrix expansion does not occur as frequently. On the contrary, a threshold of $\lambda = 1$ allows any type of improvement to the model.

2.2. Datasets

As we are interested in assessing the performance of different image generation strategies in the low-data regime, we select relatively small training sets from different MRI domains.

2.2.1. BraTS2021

We use the Brain tumour Segmentation (BraTS) 2021 Challenge dataset [19], which consists of a collection of 2000 brain tumour multi-parametric Magnetic Resonance Images (mpMRI) of size 512×512, coupled with ground truth tumour annotations. Out of all available modalities (Flair, T1, T2 and contrast T1), Flair was used for all the experiments, as the tumour volumes presented significantly better contrast. Experiments were done with 2D axial slices. For each volume, the mean slice was selected. The prompts were generated relying on the corresponding tumour masks. Negative prompts are of the kind *'Brain MRI Flair, healthy'*. For positive cases, the prompt was adjusted to the tumour size and position in the image (*e.g. 'Brain MRI Flair, with large tumour on the center left'*). The training set consists of a total of 160 slices randomly sampled from a balanced distribution of positive and negative cases. Similarly, the test set consists of an equally balanced distribution of 150 scans in total.

2.2.2. PI-CAI

The second dataset we worked with was made available as part of the PI-CAI (Prostate Imaging: Cancer AI) challenge [25]. It contains 1500 prostate MRI scans from 1476 patients. Only 1295 patients were used, as the experts' annotations are available for only 220 of 425 positive cases. T2-weighted ('t2') imaging modality was used, starting at an image size of 512×512. To align our methodology to some extent with the work by *De Wilde et al.* [7], we adopted a similar approach to process the dataset. This provides a basis for discussing their Textual Inversion fine-tuning approach in relation to our findings. To extract the positive cases, the available tumour segmentation masks [26] were used to select in each volume the axial slices featuring the maximum area of the lesion. For the negative cases, the median slice was used. Note that some cases were excluded due to their anatomy being out of the Field-of-View (FoV) or the images presenting different FoV (see Appendix C, Fig. 6), reducing the positive class to 134 cases. Prompts were generated relying on the tumour masks; contrary to BraTS prompts, only the particular tumour size was specified for this dataset (*e.g. 'Prostate MRI t2, with small tumour*'), as the tumour position is consistently aligned with the FoV centre. Similarly, negative prompts are of the kind '*Prostate MRI t2 healthy*'. The training set was created by randomly sampling 100 positive and 100 negative cases, while the testing set consisted of the remaining 34 positive cases and 34 negative ones sampled at random.

2.3. Fine-tuning

The experiments were conducted using the weights from 'runwayml/stable-diffusion v1-5' available on Hugging Face [23]. The model is built on three components: a Variational Autoencoder (VAE), a Denoising U-Net, and a text encoder; we followed the same approach used in *Mao et al.* [22] to introduce self-expandable LoRA matrixes into all linear layers of the denoising U-Net and Text Encoder. The VAE was not fine-tuned as previous work suggests that changing the weights in this component does not lead to significant performance increases [6]. Contrary to the original SeLoRA approach [22], we found that the loss on a validation set is not a reliable indicator of image quality. Therefore, we used no validation set. Due to the lack of metrics that comprehensively quantify the quality of medical image generation outputs [4], we conducted hyperparameter search by visual inspection of the generated images. After extensive experimentation, we selected a few significant model variations from the most promising training schemes for the two datasets. Training samples were resized to an image size of 224×224, and their intensity values were normalized. The initial rank of all LoRA matrices was set to 1 and Mean Squared Error (MSE) was used as the loss function during training. In both cases, Adam optimizer with a learning rate

of 10^{-4}, and $\beta_1 = 0.9, \beta_2 = 0.999$ was used. In the case of the BraTS2021 dataset, the models were trained for 100 epochs on batches of 3 images. We present experiments with $\lambda_1 = 1.0, \lambda_2 = 1.04$ and $\lambda_3 = 1.08$. The models fine-tuned on PI-CAI were trained for 150, 175 and 200 epochs with $\lambda = 1.0$. Furthermore, we set and froze the initial rank to 4 and 8 to assess the performance of vanilla Low-Rank Adaptation. Following the approach of *De Wilde et al.* [7], we took a pre-trained StyleGAN3 [1] [17, 16] model as our state-of-the-art baseline. This allowed us to compare the performance of a fine-tuned diffusion model with that of a fine-tuned generative model. As *De Wilde et al.*, we fine-tuned the model for the same number of steps as the LDMs. However, noticing how the visual quality of the outputs would increase with more iterations, we also present the results obtained with a total of 25000 iterations. Once the models were trained, the image generation for both prostate and brain images was performed with similar prompts, resulting in a balanced distribution of positive and negative samples of image size 224×224. A total of 1000 images were generated both for prostates and brains, with 500 positive cases and 500 negative cases. Notably, the synthetic brains obtained with forms of Low-Rank adaptation often presented artefacts in the background or shade distortions. For downstream classification, we applied a transformation to gray-scale and removed the background. This last step involved using the Canny edge detector [5] and some morphological operations on the obtained masks. We want to emphasize how these steps represent a straightforward post-processing operation requiring minimal tuning. Experiments were run with PyTorch, with the fine-tuning of the LDM networks taking around 3 hours on an NVIDIA T4 GPU, depending on the parameters.

2.4. Evaluation

Assessing the quality of generated images is not a trivial task, especially in the case of medical images [3]. First, we employed the Fréchet Inception Distance (FID) [13], which measures the similarity of a set of generated images to a set of real images. This is obtained by comparing the set-wise multivariate normal distributions estimated from Inception V3 features. Notably, the FID is computed using a model trained on natural images, therefore, its validity for medical image generation is still being discussed [28]. While in our proposal we mentioned computing the FID using a network pre-trained on a large number of medical images (MFID), recent results from *Woodland et al.* [32] seem to provide evidence against this practice. We computed the FID on images that had not undergone any post-processing. On a note, metrics such as the Peak Signal-to-Noise Ratio and the Structural Similarity Index Measure are not suited for our results, as the prompts are not detailed enough to establish a one-to-one relationship between synthetic and generated images. A proxy for expert visual assessment of the results is discussed in A.

[1] github.com/NVlabs/stylegan3

Table 1. FID scores for BraTS and PI-CAI synthetic data with simple LoRA, SeLoRA and StyleGAN3 models. r=4/8 indicates the fixed rank of the matrices.

PI-CAI	FID	BraTS	FID
LoRA (r=4)	299.61	LoRA (r=4)	195.7
LoRA (r=8)	258.83	LoRA (r=8)	211.98
SeLoRA 150 epochs	158.13	SeLoRA $\lambda = 1.0$	126.38
SeLoRA 175 epochs	171.67	SeLoRA $\lambda = 1.04$	125.29
SeLoRA 200 epochs	165.80	SeLoRA $\lambda = 1.08$	117.06
StyleGAN3 18k iter.	149.70	StyleGAN3 8k iter.	77.53
StyleGAN3 25k iter.	**140.30**	**StyleGAN3 25k iter.**	**65.71**

2.4.1. Classifier training

We evaluated the effectiveness of our methods by using synthetic data to perform data augmentation for a classification task. We did not consider the outputs from vanilla LoRA experiments as their quality was likely not sufficient to provide any benefit (see Appendix C, Fig. 5). We used a pre-trained ResNet-18 model [12] and trained it with the same set of images used for the Diffusion Model and StyleGAN3. Training was performed for 50 epochs with a batch size of 32 images and a learning rate of $1e-4$. We regularly employed a set of basic data augmentation techniques (horizontal flipping and Gaussian blurring). The network was then evaluated on the task of assessing whether an image contained a tumour or not. We computed the accuracy and the AUC score on the test sets using models trained with 5 different random seeds. This assessment provides a useful indication of the diagnostic value of our images.

3. Results

The FID scores are presented in Table 1. The results for samples generated with variations of StableDiffusion fine-tuned with vanilla LoRAs and SeLoRA are evaluated against the StyleGAN3 baseline. Similarly, the accuracy and AUC scores of the classifier trained with only real images, and its augmented version, are observed in Tables 2 and 3. Table 2 displays the results of the BraTS dataset augmentation with the StyleGAN3 and SeLoRA generation approaches. Equally, Table 3 shows the results from prostate data augmentation with StyleGAN3 and SeLoRA. Figure 1 displays positive and negative samples of the brains generated with the SeLoRA approach with the different expansion thresholds. Similar examples of prostates generated with SeLoRA for a different number of epochs are observed in Figure 2. Finally, Figure 3 in Appendix C shows multiple samples of brains and prostates generated with StyleGAN3.

Table 2. AUC and Accuracy scores for brain generation using StyleGAN3 and StableDiffusion fine-tuned with SeLoRA with different thresholds.

Augmentation Method	#Real	#Synthetic	AUC	Accuracy
No data aug.	160	0	0.93 ± 0.02	0.84 ± 0.02
SeLoRA $\lambda = 1.0$	160	1000	0.93 ± 0.02	$\mathbf{0.89 \pm 0.02}$
SeLoRA $\lambda = 1.04$	160	1000	$\mathbf{0.95 \pm 0.02}$	0.84 ± 0.03
SeLoRA $\lambda = 1.08$	160	1000	0.90 ± 0.02	0.82 ± 0.04
StyleGAN3 8k iter.	160	1000	0.89 ± 0.02	0.84 ± 0.05
StyleGAN3 25k iter.	160	1000	0.91 ± 0.04	0.84 ± 0.05

Table 3. AUC and Accuracy scores for prostate generation using StyleGAN3 and StableDiffusion fine-tuned with SeLoRA with different epochs.

Augmentation Method	#Real	#Synthetic	AUC	Accuracy
No data aug.	200	0	0.85 ± 0.02	0.72 ± 0.03
SeLoRA $epochs = 150$	200	1000	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.77 \pm 0.04}$
SeLoRA $epochs = 175$	200	1000	0.87 ± 0.03	0.73 ± 0.03
SeLoRA $epochs = 200$	200	1000	$\mathbf{0.87 \pm 0.02}$	0.75 ± 0.04
StyleGAN3 18k iter.	200	1000	0.81 ± 0.04	0.70 ± 0.05
StyleGAN3 25k iter.	200	1000	0.85 ± 0.03	0.72 ± 0.03

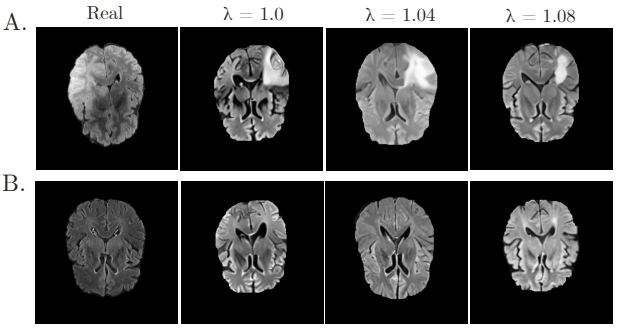


Figure 1. Examples of postprocessed positive (A) and negative (B) brain cases generated with Stable Diffusion fine-tuned with SeLoRA thresholds of $\lambda = 1.0$, $\lambda = 1.04$ and $\lambda = 1.08$.

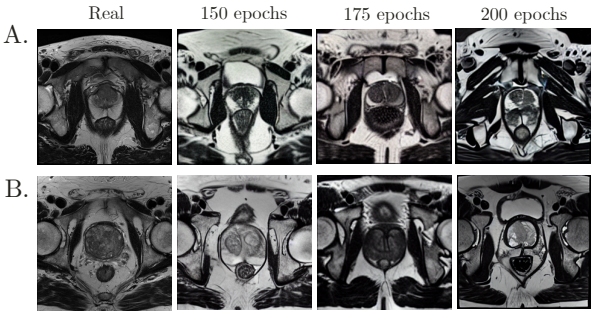


Figure 2. Examples of positive (A) and negative (B) prostate cases generated with Stable Diffusion fine-tuned for 150, 175 and 200 epochs SeLoRA.

4. Discussion

Our findings demonstrate that Stable Diffusion fine-tuned with SeLoRA successfully generates mostly realistic and diagnostically meaningful images for both prostate and brain MRI. Moreover, by leveraging text-based conditioning, we can control the size and position of the tumour within the generated image. These results are achieved at a relatively low computational cost. We successfully performed data augmentation even with relatively small training sets with no detailed textual descriptions available in the original datasets. The improved AUC and accuracy scores of the downstream classifier, as shown in Tables 2 and 3, further support our results. In contrast, the vanilla LoRA approach produced images that were far from being diagnostically meaningful or anatomically accurate, resulting in the least favourable FID metrics. While StyleGAN3 demonstrated better performance in terms of apparent visual quality, the AUC and accuracy scores reveal that augmenting the training dataset with images from StyleGAN3 either reduces or provides little improvement in the classifier performance. We observe that the StyleGAN images appear visually pleasing but present some artefacts at the lower scale (see Appendix C, Fig. 3) and often lack variability, hinting at a partial mode collapse during training and possibly justifying the worse classifier performance. Although no single optimal threshold was identified, our findings partially align with the claim by *Mao et al.* that smaller values of λ offer advantages in terms of image quality. Furthermore, during our experiments, we observed a degree of instability in the training performance, where even a small change in the parameters would produce rather different results. Moreover, reproducing the BraTS2021 images with a dark background presented some challenges, with some LDM configurations hallucinating background patterns (see Appendix, Figure 4) that were easily removed in post-processing. While our training and testing protocols differ slightly from those of *De Wilde et al.* [7] and it was not possible for us to recruit expert radiologists for visual assessment, our results suggest that SeLoRA outperforms their Textual Inversion approach in generating accurate Prostate MRI images from the same dataset. This improvement, however, comes at the cost of increased computational demands. While it was not possible within our timeline, it would be worth investigating in the future the application of the fine-tuned networks for inpainting real images and to provide a direct comparison with TI approaches.

5. Summary

In this work, we assessed the quality of synthetic MRI data produced with different fine-tuning strategies from small datasets lacking detailed radiology notes. Our findings suggest that SeLoRA is an effective method for fine-tuning diffusion models in our domain, yielding significantly more diagnostically accurate results than some alternative models.

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. arXiv preprint arXiv:2012.13255, 2020.

[2] Panagiotis Alimisis, Ioannis Mademlis, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Georgios Th Papadopoulos. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions. arXiv preprint arXiv:2407.04103, 2024.

[3] Ali Borji. Pros and cons of gan evaluation measures. Computer vision and image understanding, 179:41–65, 2019.

[4] Anna Breger, Ander Biguri, Malena Sabaté Landman, Ian Selby, Nicole Amberg, Elisabeth Brunner, Janek Gröhl, Sepideh Hatamikia, Clemens Karner, Lipeng Ning, Sören Dittmer, Michael Roberts, AIX-COVNET Collaboration, and Carola-Bibiane Schönlieb. A study of why we need to reassess full reference image quality assessment with medical images, 2024. URL https://arxiv.org/abs/2405.19097.

[5] John Canny. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, (6):679–698, 1986.

[6] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains, 2022. URL https://arxiv.org/abs/2210.04133.

[7] Bram De Wilde, Anindo Saha, Maarten de Rooij, Henkjan Huisman, and Geert Litjens. Medical diffusion on a budget: textual inversion for medical image generation. arXiv preprint arXiv:2303.13430, 2023.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems, 36, 2024.

[9] Laila El Jiani, Sanaa El Filali, et al. Overcome medical image data scarcity by data augmentation techniques: A review. In 2022 International Conference on Microelectronics (ICM), pages 21–24. IEEE, 2022.

[10] Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, and Kristin Mcleod. Generating synthetic labeled data from existing anatomical models: an example with echocardiography segmentation. IEEE Transactions on Medical Imaging, 40(10): 2783–2794, 2021.

[11] Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intelligence Review, 56(11):12561–12605, 2023.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

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

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In Proc. NeurIPS, 2021.

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems, 35:26565–26577, 2022.

[18] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis, 88:102846, 2023.

[19] Florian Kofler, Felix Meissen, Felix Steinbauer, Robert Graf, Eva Oswald, Ezequiel de da Rosa, Hongwei Bran Li, Ujjwal Baid, Florian Hoelzl, Oezguen Turgut, et al. The brain tumor segmentation (brats) challenge 2023: Local synthesis of healthy brain tissue via inpainting. arXiv preprint arXiv:2305.08992, 2023.

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. Medical image analysis, 42:60–88, 2017.

[21] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In

International conference on learning representations, 2020.

[22] Yuchen Mao, Hongwei Li, Wei Pang, Giorgos Papanastasiou, Guang Yang, and Chengjia Wang. Selora: Self-expanding low-rank adaptation of latent diffusion model for medical image synthesis. arXiv preprint arXiv:2408.07196, 2024.

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

[24] 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. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 22500–22510, 2023.

[25] Anindo Saha, Jasper J. Twilt, Joeran S. Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, and Henkjan Huisman. Artificial intelligence and radiologists at prostate cancer detection in mri: The pi-cai challenge. 2022. doi: 10.5281/zenodo.6522364.

[26] Anindo Saha, Joeran S Bosma, et al. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. The Lancet Oncology, 25(7):879–887, 2024. ISSN 1470-2045. doi: https://doi.org/10.1016/S1470-2045(24)00220-1.

[27] Wei Song, Wen Ma, Ming Zhang, Yanghao Zhang, and Xiaobing Zhao. Lightweight diffusion models: a survey. Artificial Intelligence Review, 57(6):161, 2024.

[28] Shenghuan Sun, Gregory M. Goldgof, Atul Butte, and Ahmed M. Alaa. Aligning synthetic medical images with clinical knowledge using human feedback, 2023. URL https://arxiv.org/abs/2306.12438.

[29] Beatriz Teixeira, Gonçalo Pinto, Vitor Filipe, and Ana Teixeira. Enhancing medical imaging through data augmentation: A review. In International Conference on Computational Science and Its Applications, pages 341–354. Springer, 2024.

[30] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In 2020 international joint conference on neural networks (ijcnn), pages 1–10. IEEE, 2020.

[31] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. arXiv preprint arXiv:2210.07558, 2022.

[32] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 87–97. Springer, 2024.

[33] Jianan Yang, Haobo Wang, Yanming Zhang, Ruixuan Xiao, Sai Wu, Gang Chen, and Junbo Zhao. Controllable textual inversion for personalized text-to-image generation. arXiv preprint arXiv:2304.05265, 2023.

[34] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512, 2023.

A. Visual evaluation

As comprehensively scoring the quality of synthetic images is hard, it is a common practice to perform blind comparisons with human experts. As we couldn't recruit a professional radiologist, we chose to provide a proxy score by submitting our images to a fifth-year medical student. While this does not provide a rigorous metric, we include this procedure as a blueprint on how to further assess the quality of similar results.

A.1. Procedure

For each dataset, the student was first shown a collection of 9 real healthy samples, and 9 samples that featured a tumour. Next, for each dataset, they were sequentially shown a set of $N = 10$ random pairs of synthetic healthy images, either generated by StyleGAN3 or by the SeLoRA model with the highest FID. The order of the images within the sample pair was also aleatoric. At each iteration, the student was asked to choose which image they believed to be more anatomically accurate. The same procedure was applied for $N = 10$ random pairs of synthetic images featuring a tumour.

A.2. Results

The medical student consistently preferred the StyleGAN3 brain images. In the case of prostate images, the StyleGAN3 images were preferred in 70% of cases. These results are consistent with the FID scores.

B. Hyperparamter tuning

As mentioned in Section 2.3, we conduct hyperparameter tuning by visually assessing the quality of the generated images. We attach here the working document used to keep track of our attempts. While it has not been curated to be used by an external audience, we deemed it potentially interesting to readers wanting to observe some samples from other trainings.

C. Image examples

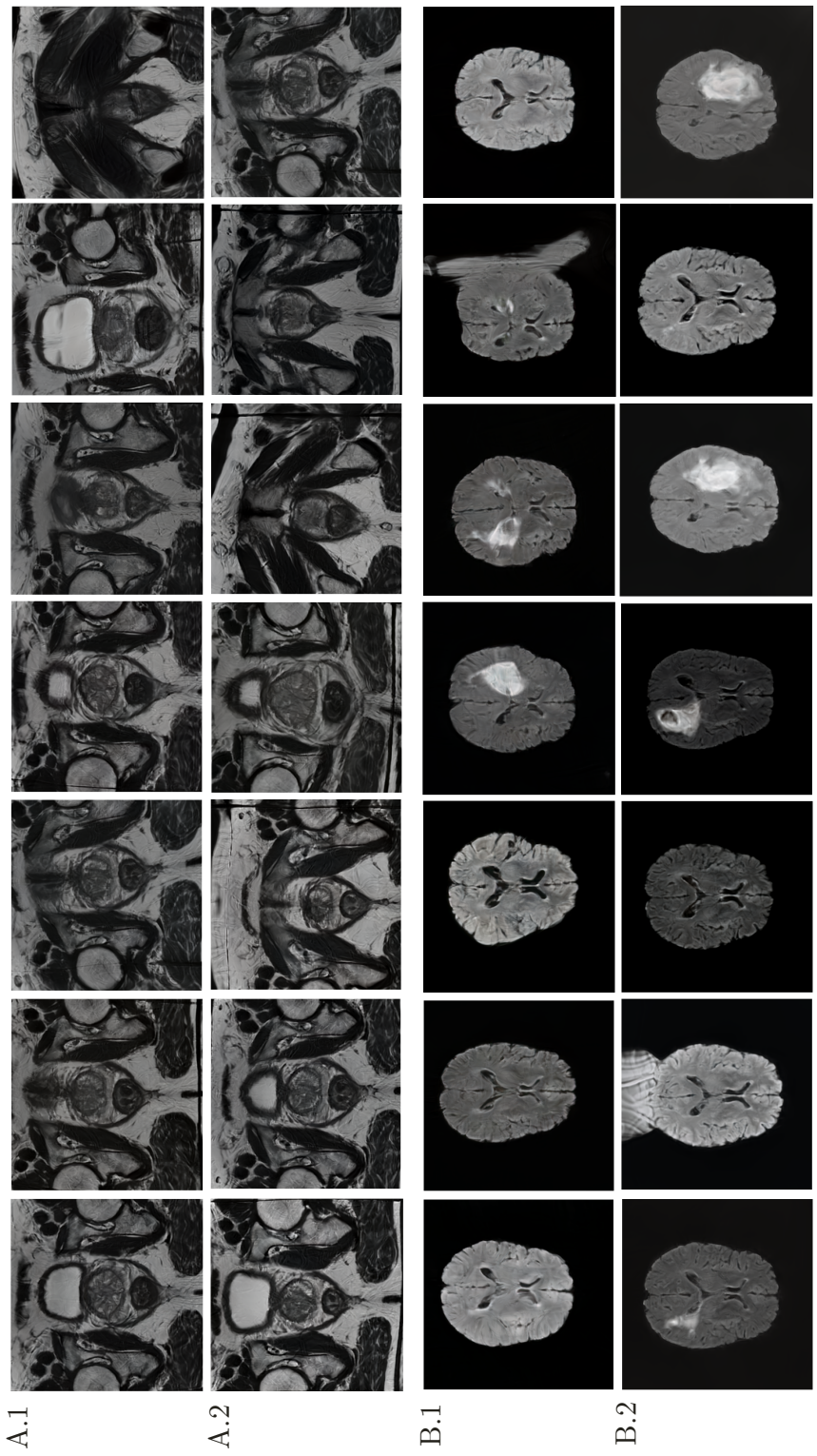


Figure 3. Prostates generated with StyleGAN3 trained for 18k (A.1) and 25k (A.2) iterations. Brains generated with StyleGAN3 trained for 8k (B.1) and 25k (B.2) iterations.

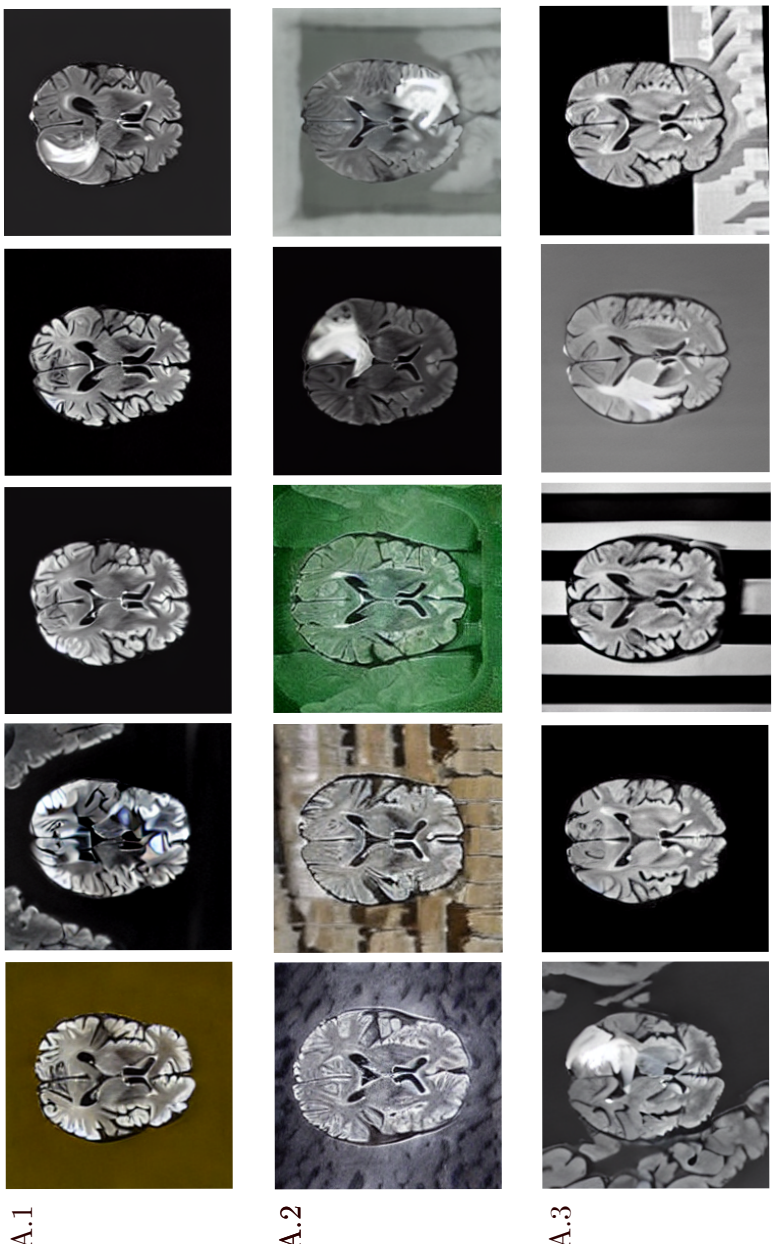


Figure 4. Examples of generated brains with expansion thresholds of $\lambda = 1.0$ (A.1), $\lambda = 1.04$ (A.2) and $\lambda = 1.08$ (A.3). We selected some examples featuring hallucinations in the background.

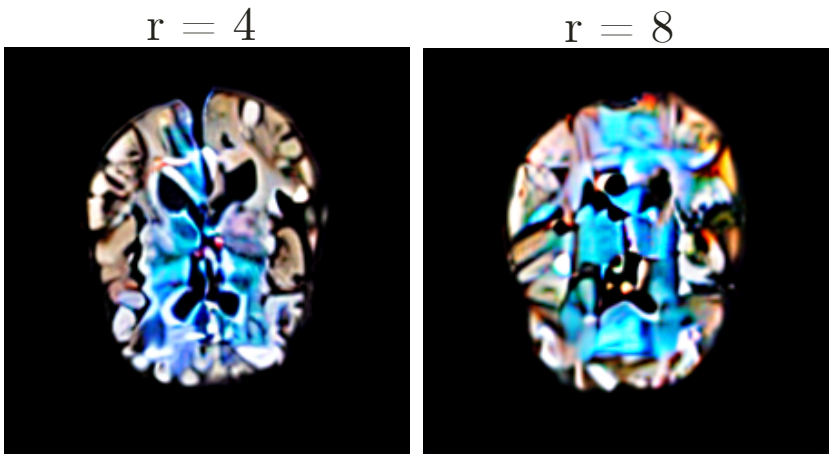
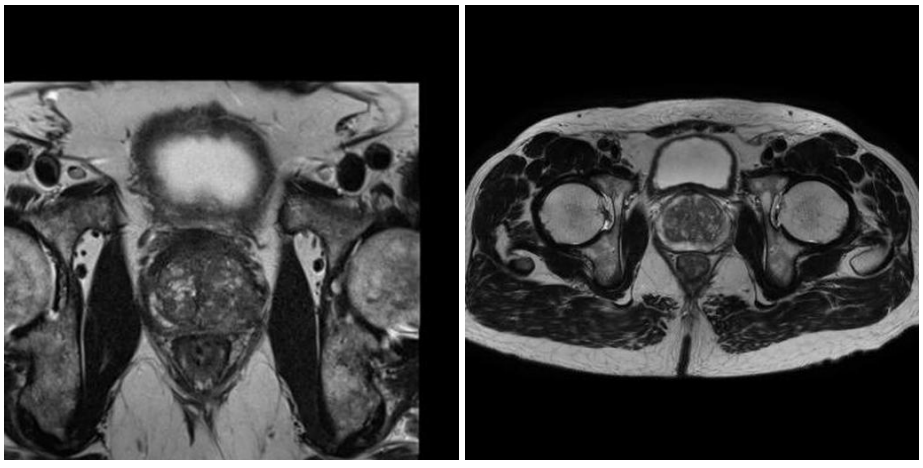


Figure 5. Examples of vanilla-LoRA outputs from experiments performed with ranks $r = 4$ and $r = 8$.



(a) Out of FoV example (b) Different FoV example

Figure 6. Examples of images discarded from the PI-CAI dataset