Supplementary Materials for
Adaptive Latent Diffusion Model for 3D Medical Image to Image Translation: Multi-modal Magnetic Resonance Imaging Study

A. Hyperparameter setting

In this section, we provide a detailed description of the model architecture and hyperparameters for the autoencoder used in image compression, as well as the structure of the diffusion model. The input dimension for all networks is 3D. For the autoencoder, we applied the network architecture of VQGAN [4], which is explained in Table A. The structure of the MS-SPADE block present in the bottleneck of the autoencoder is described in Table B. Additionally, we applied a UNet-based network architecture to the diffusion model used in previous studies [5, 8], which is explained in Table C.

Input Size	dim $\lvert \mathcal{Z} \rvert$	Channels	Embedding Size
$192 \times 192 \times 144$	8192	[256,512,512]	3
Batch Size	Epochs	Model Size	Param Size
1	500	749M	237M

Table A. Detailed Hyperparameters for latent diffusion model.

MS-SPADE Block							
Stream	Conv.	Act.	Norm.	Conv.	Act.	Norm.	Out ch.
In	C_7		IN		ReLU		128
ResBlock	C_3	ReLU	IN	C_3	ReLU	IN	[256,256]
SPADEBlock	C_3	ReLU	MS-SPADE	C_3	ReLU	MS-SPADE	[256,256,256,128]
Out	C_7						3

Table B. Detailed MS-SPADE Block. C_i is the convolution layer with $i \times i$ kernel. IN is the instance normalization layer, and MS-SPADE is the Multi switchable SPADE layer that is applied differently depending on the target modality. Out ch. represents the output channels, and both ResBlocks and SPADEBlocks are repeated 2 and 4 times, respectively

Stream	Condi	Batch Size	Model Size	Param Size
$48 \times 48 \times 36 \times 3$	[128,256,512]	1	722M	658M
Diffusion steps	Noise Scheculde	β_{start}	β_{end}	Epochs
1000	scaled-linear	0.0015	0.0195	800

Table C. Detailed hyperparameters for latent diffusion model.

B. Dataset

We trained our model on the BraTS 2021 training dataset, encompassing 1251 subjects and four MRI modalities (T1, T1ce, T2, FLAIR). Each MRI scan measures $240 \times 240 \times 155$ in dimensions, with a spatial resolution of $1 \times 1 \times 1mm^3$. To assess our model's image translation capabilities, we utilized the BraTS 2021 validation dataset, containing 219 subjects. Additionally, we tested our model using the IXI dataset, including T1, T2, and PD modalities. From the 574 subjects, 459 were allocated for training and 115 for testing. Each of these MRI scans measures $256 \times 150 \times 256$ in dimensions with a spatial resolution of $0.9375 \times 0.9375 \times 1.2mm^3$

C. Comparison Methods details

To validate the effectiveness of our model, we used commonly used methods in medical image-to-image translation as comparison models. For 2D methods, we employed Pix2Pix [6], CycleGAN [9], NICEGAN [2], RegGAN, [7] and ResViT [3]. For the 3D method, we employed the 3D versions of pix2pix and CycleGAN, as well as the Ea-GAN proposed as a 3D method, for comparison. We compared using the discriminator-induced Ea-GAN (dEa-GAN) model as presented in the reference [1]. 3D methods are not as commonly used and come with higher computational costs making it challenging to extend existing 2D models to 3D. 2D methods were executed with a batch size of 32 in the axial view. For the BraTS dataset, they operated on images sized 240×240, while for the IXI dataset, zero padding was added to process images at 256×160 dimensions. All 3D methods were conducted with a batch size of 1. On the BraTS dataset, images were cropped to $192 \times 192 \times 144$ after background removal. For the IXI dataset, images were cropped and padded to measure $256 \times 160 \times 224$.

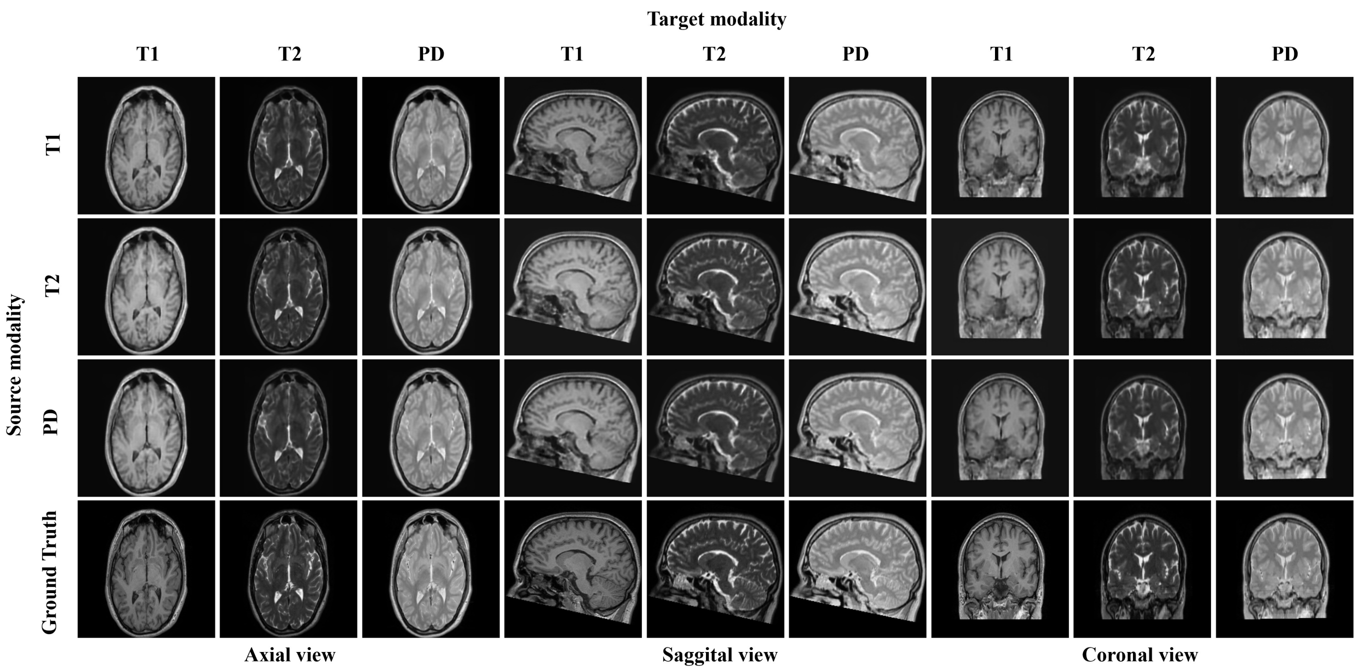

Figure A. The figures showcase the image translation results on the IXI dataset from each source modality to the corresponding target modality using our proposed model for all possible combinations.

Source \ Target	T1			T2			PD		
Metric	PSNR ↑	NMSE ↓	SSIM ↑	PSNR ↑	NMSE ↓	SSIM ↑	PSNR ↑	NMSE ↓	SSIM ↑
T1	*29.487 ±0.522*	*0.047 ±0.020*	*0.941 ±0.024*	27.265 ±0.629	0.071 ±0.022	0.921 ±0.015	27.729 ±0.685	0.072 ±0.025	0.922 ±0.018
T2	27.368 ±0.624	0.074 ±0.031	0.929 ±0.027	*29.259 ±0.582*	*0.045 ±0.017*	*0.937 ±0.015*	**27.913 ±0.659**	**0.067 ±0.023**	**0.927 ±0.019**
PD	**27.968 ±0.521**	**0.070 ±0.028**	**0.931 ±0.028**	**27.834 ±0.627**	**0.067 ±0.024**	**0.925 ±0.025**	*29.396 ±0.488*	*0.042 ±0.019*	*0.939 ±0.027*

Table D. The values present the quantitative evaluation of image translation results on the IXI dataset from source modalities to target modalities using our proposed model.

D. Additional Experimental Results

We also analyze which source modality is most effective in synthesizing the target modality within the IXI dataset. Figure A provides the qualitative evaluation results of this multi-modal translation, while Table D offers the quantitative assessment outcomes. From the qualitative evaluation, we observe that there are minimal differences between modalities, and most present a satisfactory translation performance. As for the quantitative evaluation, it is evident that PD is effective in image translation when generating T1, and similarly for T2. Conversely, T2 proves to be efficient when producing PD images.

References

[1] Yu Biting, et al. Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis. *IEEE transactions on medical imaging*, 38(7):1750–1762, 2019. 1

[2] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8168–8177, 2020. 1

[3] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022. 1

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial net-

works. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[7] Lingke Kong, Chenyu Lian, Detian Huang, zhenjiang li, Yanle Hu, and Qichao Zhou. Breaking the dilemma of medical image-to-image translation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1964–1978. Curran Associates, Inc., 2021. 1

[8] 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. 1

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1