Efficient Many-to-Many MRI Modality Translation via a Latent-Conditioned Vector-Quantized Network*

Hassan Baker[1] and Austin J. Brockmeier[2]

Abstract— **Neuroimages are collected in different MRI modalities, such as T1 and T2, each of which may provide unique information, especially contrast-enhanced modalities. Collecting each takes time; additionally, in historical datasets missing modalities prevent holistic analysis. Translating between MRI modalities could play an important role in saving costs and leveraging existing data. In this paper, we propose a machine learning model that can translate a 3D neuroimage from one of many source modalities to one of many target modalities, which outperforms the current state of the art (SOTA), which rely on denoising diffusion models. Furthermore, our many-to-many translation model simplifies the majority of SOTA methods that use either one-to-one or one-to-many models, which require multiple models to handle different modalities. Our model consists of an attention U-Net with conditioning inputs based on the source and target modality labels and operates in a shared latent space based of a vector-quantized auto-encoder network (VQGAN). We find this to be sufficient to translate between MRI images and gives competitive results to SOTA. We evaluate our model on BRATs dataset which is composed of four different modalities.**

Clinical relevance—**Modality transfer can facilitate the generation of multiple MRI modalities, which has the potential to reduce cost and save time. The limitations of modality transfer can identify the unique information in MRI modalities.**

Keywords: neuroimaging, MRI, non-linear regression, image synthesis, machine learning.

I. Introduction

Capturing different MRI modalities is crucial to rightly diagnose potential diseases and abnormalities as each modality can highlight different tissues structure [1]. Nonetheless, there are several difficulties associated with obtaining all modalities such as time constraint, financial cost, and potential health risks [2], [3]. Thus, having modality translation models has the potential to positively impact neuroradiology across the world.

Brain MRI has several common modalities, T1-weighted imaging provides high-resolution anatomical detail, with cerebrospinal fluid (CSF) appearing dark and white matter brighter than gray matter. This sequence is particularly useful for assessing brain anatomy and detecting fat-containing lesions [4]. Injecting a contrast agent with T1-weighted imaging (T1-ce) enhances the visibility of vascular structures and areas with a disrupted blood-brain barrier, such as tumors or sites of inflammation [4]. T2-weighted imaging emphasizes fluid content, rendering CSF bright and aiding in the identification of edema, inflammation, and demyelination, making it instrumental in detecting lesions associated with conditions like multiple sclerosis and tumors [4]. Fluid-attenuated inversion recovery (FLAIR) sequences suppress the CSF signal, enhancing the visibility of periventricular and cortical lesions. This makes FLAIR particularly effective in identifying subtle abnormalities adjacent to CSF spaces, such as those seen in multiple sclerosis [4], [5].

While each these modalities are visually distinct, it is not clear the amount of unique information each provides considering that the images from some modalities can be transferred to other. Without a downstream task or expert evaluation it is not clear what is being lost in modality transfer. Nonetheless, it may be possible that information in each modality is shared, but subtly encoded, and each can be derived from the other. With this possibility we seek to advance MRI modality translation.

We propose a 3D GAN-based, many-to-many MRI modality translation model capable of generating any target modality from a single input modality at inference time. Our approach does not rely on multi-sequence inputs, making it more clinically flexible than existing multi-input methods. Our model is composed of two stages of training: first, we train a VQGAN [6], consisting of an encoder and decoder network, applicable to all modalities, then we adversarially train a U-Net in the latent space to facilitate the translation, outperforming more complicated diffusion models. As modality transfer in the 3D space is expensive, the encoder compresses the data to a lower dimensional space. VQGAN is well-known to focus on semantics, and incorporates both vector quantization in the latent space and adversarial training via a discriminator, which produces high fidelity images compared to auto-encoders.

II. Related Work

We categorize the current literature on MRI modality translation along several dimensions. Firstly, by data representation: 2D slice-based models versus 3D volume-based models. Secondly, by translation type: one modality to one modality (one-to-one), one modality to many modalities (one-to-many), and many modalities to many modalities (many-to-many) modality mappings. Thirdly, we consider

*This work was supported in part by the Delaware Community Foundation and the University of Delaware Graduate College

[1]H. Baker is affiliated with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, 19716 USA bakerh@udel.edu

[2]Austin J. Brockmeier is affiliated with the Department of Electrical and Computer Engineering, Department of Computer and Information Sciences, University of Delaware, Newark, DE, 19716 USA ajbrock@udel.edu

modality flexibility of how models utilize multiple modalities. In this last category, we distinguish between models that use multiple modalities as input, but require them at inference; approaches that operate on multiple modalities and can compensate for missing modalities at inference; and models that only use a single source input. Fourthly, we consider the usage of adversarial training versus dynamical generative model such as diffusion and conditional flow matching models.

A. 3D vs 2D Translation Models

3D models capture spatial features across volumetric context, whereas 2D models operate on individual slices and do not explicitly model inter-slice relationships. This effectively treats slices as independent, which can overlook important anatomical continuity across the volume. While 3D models are more computationally demanding due to the higher dimensionality, they are also more capable of leveraging spatial consistency and hence improve the performance, and can improve training and inference speed since only a single pass is required per 3D scan [7], [8]. For these reasons, we adopt a 3D model with a moderate parameter budget (under 60 million parameters) to balance modeling capacity and resource constraints.

B. Translation Type

Having a separate model for each pair of modalities (i.e., one-to-one) is a straightforward approach. Nonetheless, MRI modalities share information and having a carefully designed model can utilize shared features between each source-target translation task. Allowing the network to use shared features among modalities, may benefit the generalization performance and robustness against covariate shift [9], which is a common challenge in MRI [10]. In this work, we develop a many-to-many translation model to learn robust features. We encourage the model to learn a shared latent space.

C. Modality Flexibility

MRI modality translation models vary in how they utilize input and output modalities. Some models, like ours, is designed to take a single modality as input and then synthesize a single target modality as ouput. Other models operate on multiple modalities simultaneously—either as input, output, or both. Models trained with multiple input modalities often achieve superior synthesis performance by leveraging complementary information across sequences [11], [12], [13], [14], [15], [16], [17], [18]. For example, a model trained to synthesize FLAIR from both T1 and T2 requires access to both modalities at inference time [13]. However, this dependency can be a limitation in clinical practice, where not all sequences may be available. Recent approaches attempt to overcome this challenge by using multiple inputs and introducing mechanisms that compensate for missing modalities, such as modality dropout or feature-level fusion [14].

Despite the performance advantages of multi-input models, they come with trade-offs: (1) missing modalities at inference can degrade performance if not properly handled, (2) addressing this issue often requires more complex architectures, and (3) in the case of 3D training, using multiple modalities simultaneously imposes significant memory demands. For example, during training our model on a `NVIDIA A40` GPU with 48GB of VRAM, it is often not feasible to input more than one full-resolution MRI volume at once. This makes multi-input training impractical under typical hardware constraints.

Our model is trained across all source–target modality pairs using a unified framework, but performs single-source, single-target translation at inference. This setup enables us to learn shared representations across modalities during training while maintaining flexibility to operate under modality-limited conditions—accounting for worst-case clinical scenarios. Both strategies have their merits: fully multi-modal models offer high fidelity under ideal conditions, whereas our approach prioritizes robustness, generalizability, and real-world practicality.

D. Static versus Dynamical Training

Image synthesis models that rely solely on pixel-wise similarity losses—such as L1 or L2-norm of differences—often produce blurry and perceptually suboptimal outputs, due to the inability of these losses to capture high-frequency details and structural consistency [19]. In contrast, adversarial training has demonstrated superior performance in image-to-image translation tasks by encouraging photorealistic reconstructions [20], [21], [22], [11]. Nonetheless, adversarial training of static generating networks has a bad reputation of its instability [23]. Dynamical modeling techniques that are based on denoising probabilistic diffusion model (DDPM) [24], [25] and conditional flow matching (CFM) [26] have been deployed [27]. However, a generating network has faster inference time compared to dynamical models as it is a one shot mapping [28]. Based on current literature review for MRI modality transfer, we have seen that adversarially trained static models give superior performance to the dynamical models. We hypothesis that this is due to the paired data relationship that seems to be more suitable for static models as we demonstrate. Our model is adversarially trained and we show it gives superior performance compared to SOTA of diffusion models.

E. Prior MRI Modality Translation Methods

A summary of related work is shown in Table I. We begin with simpler methods and finish with dynamic models.

Yu et al. introduced a translation network that reconstructs a particular target modality given the source modality and is trained with a discriminator [29]. In order to preserve the edge information, the network also predict the target modality edges (obtained using Sobel filter) to preserve the structural information. Similarly, Dar et al. proposed using a conditional GAN with L1-loss to perform 2D translation model for one-to-one translation [30]. In a similar work, Xin et al. used a conditional GAN to train one-to-many translational models [31].

TABLE I

SUMMARY OF CAPABILITIES OF MRI MODALITY TRANSLATION METHODS. ✓ INDICATES CAPABILITY; ✗ INDICATES ABSENCE; N/A IS NOT APPLICABLE. ADVERSARIAL (ADV.) TRAINING AND FLEXIBILITY (FLEX.) ARE ABBREVIATED.

Ref.	Dim.	Input	Adv.	Dynamic	Flex.	Type
[29]	2D	Single	✓	✗	N/A	1–1
[30]	2D	Single	✓	✗	N/A	1–1
[31]	2D	Single	✓	✗	N/A	1–∞
[15]	2D	Multi.	✓	✗	✓	∞–1
[16]	2D	Multi.	✓	✗	✗	∞–1
[17]	2D	Multi.	✓	✗	✓	∞–1
[18]	2D	Multi.	✓	✗	✓	∞–1
[32]	2D	Multi.	✗	✗	✓	∞–1
[33]	2D	Multi.	✓	✗	✓	∞–∞
[12]	2D	Multi.	✓	✗	✓	∞–∞
[34]	2D	Multi.	✓	✗	✓	∞–∞
[35]	2D	Single.	✓	✗	N/A	∞–∞
[13]	2D	Single	✓	✗	N/A	∞–∞
[14]	2D	Multi.	✓	✗	✓	∞–∞
[36]	2D	Single	✗	✓	N/A	1–1
[27]	3D	Single	✓	✓	N/A	1–∞
Ours	3D	Single	✓	✗	N/A	∞–∞

Under the assumption of having multiple sequences available, Li et al., Zhou et al., Lee et al., and Yurt et al. proposed conditional GAN-based 2D many-to-one translation models that benefit from having multiple modalities providing complementary information to correctly predict the target modality [15], [16], [17], [18].

Chartsias et al. proposed a flexible many-to-many 2D translational model based on an autoencoder architecture [32], which can use multiple inputs but can handle missing ones using a modality-agnostic latent representation in the encoder space. Sharma et al. proposed a many-to-many 2D translation model using a conditional GAN to reconstruct the missing MRI modalities, given the available ones, from Gaussian noise inserted in the missing input channels [33].

Dalmaz et al. and Liu et al. (2023) treated the multiple modalities as a sequence and use transformer-based 2D conditional GAN model with multiple discriminators to generate the missing sequences [12], [34].

Liu et al. (2021) used many-to-many 2D translational model with cyclic loss and enforce a modality-agnostic features in the encoder space [35]. Following this work, Cho et al. (2024a) used a similar setup but used a transformer-based conditioning [13]. A subsequent work proposed a many-to-many 2D GAN-based model that can compensates for missing modalities during inference [14].

Several works advance dynamical models, including DDPM, for MRI modality translation. Jiang et al. proposed a one-to-one DDPM framework that needs less inference time and tested their method on T1 to T2 translation [36], achieving faster computation than standard DDPM inference and reported SSIM in the same range. Kim et al. proposed a three-stage training of a one-to-many translation model: first, a VQGAN is trained for all modalities; then a SPADE layer [37] that approximate the target latent embeddings given the source latent embeddings is optimized; finally, a latent diffusion model is used to map noise to the target embeddings conditioned on the target-like embeddings from the SPADE block [27]. This is the closest method to ours, as we also use a VQGAN, but instead we use a U-Net conditioned with the source-target labels to predict the target-like embeddings, achieving a many-to-many translation model.

III. METHODOLOGY

We present a 3D generative MRI modality translation framework that combines a vector-quantized generative adversarial network (VQGAN) with a conditional U-Net, trained on each ordered source-target modality pair in the VQGAN latent space. The model accepts a single 3D source input (along with two labels indicating the source and target modalities) and produces a single 3D output.

A. Stage 1: VQGAN Encoder and Decoder

Let $x \in \mathbb{R}^{I_C \times I_D \times I_H \times I_W}$ denote a volumetric MRI scan, where I_C is typically 1. The VQGAN consists of an encoder E, a decoder G, a learned codebook $Z = \{e_k\}_{k=1}^{K}$, and a discriminator D [6]. The encoder maps the input to a latent volume:

$$z_e = E(x) \in \mathbb{R}^{I'_D \times I'_H \times I'_W \times d_z} \tag{1}$$

The latent vector $z_e(i,j,k) \in \mathbb{R}^{d_z}$ at a given spatial location $i \in \{1,\dots,I'_D\}, j \in \{1,\dots,I'_H\}, k \in \{1,\dots,I'_W\}$ is broken into N_Q equal-sized chunks $z_e(i,j,k) = [z_e^1(i,j,k),\dots,z_e^{N_Q}(i,j,k)]$, and each chunk is quantized to the closest $d_z/N_Q = d_Q$-length vector in the codebook $\mathcal{C} = \{c_1,\dots,c_K\} \subset \mathbb{R}^{d_Q}$. Together, the quantization is

$$\begin{aligned} Q(z_e) = z_q &= [z_q(i,j,k)]_{i=1,j=1,k=1}^{I'_D,I'_H,I'_W}, \\ z_q(i,j,k) &= \underline{c}_* = \underset{\underline{c}\in\mathcal{C}\times\cdots\times\mathcal{C}}{\arg\min} \|z_e(i,j,k) - \underline{c}\|_2, \end{aligned} \tag{2}$$

where $\underline{c}_* = [c_*^1,\dots,c_*^{N_Q}]$ and

$$c_*^l = \underset{c\in\mathcal{C}}{\arg\min} \|z_e^l(i,j,k) - c\|_2, \tag{3}$$

as in product quantization [38]. Then, the decoder G reconstructs the image $\hat{x} = G(z_q) = G(Q(E(x)))$. During training, E, Q, G are optimized to minimize $\mathcal{L}_{\text{VQGAN}}$,

$$\mathcal{L}_{\text{VQGAN}} = \mathbb{E}_x[L_{\text{rec}}(x) + \beta L_{\text{commit}}(x) + \gamma L_{\text{adv}}^G(x)], \tag{4}$$

$$L_{\text{rec}}(x) = \|x - \hat{x}\|_1, \tag{5}$$

$$L_{\text{commit}}(x) = \|\text{sg}[z_e] - z_q\|_F^2 + \|z_e - \text{sg}[z_q]\|_F^2, \tag{6}$$

$$L_{\text{adv}}^G(x) = -D(\hat{x}), \tag{7}$$

where $\|\cdot\|_1, \|\cdot\|_F$ are L_1 and the Frobenius norm respectively, $\text{sg}[\cdot]$ is the stop gradient operator, and D is the discriminator jointly trained to minimize $\mathbb{E}_x[L_{\text{adv}}^D(x)]$, where

$$L_{\text{adv}}^D(x) = \max(0, 1 + D(\hat{x})) + \max(0, 1 - D(x)) \tag{8}$$

is the hinge loss. The different forms of $\mathcal{L}_{\text{adv}}^G = \mathbb{E}_{\hat{x}}[L_{\text{adv}}^G(x)]$ and $\mathcal{L}_{\text{adv}}^D = \mathbb{E}_x[L_{\text{adv}}^D(x)]$, means that it is not a minimax game. Importantly, $\mathcal{L}_{\text{adv}}^G$ does not saturate like the first term of $\mathcal{L}_{\text{adv}}^D$, a technique even used in the original GAN paper [39] and emphasized in later work [40]. The encoder E and decoder G follow the standard U-Net architecture but without

skip-connections [41], and the discriminator D follows the PatchGAN discriminator used in Pix2Pix [42].

B. Stage 2: Latent Space Translation

After training the VQGAN using the MRI scans from all M modalities, i.e., E, Q, G are shared, we train a 3D attention U-Net [43] in the latent space for cross-modality translation. For each unique pair of modalities $(s,t) \in \{1,\dots,M\} \times \{1,\dots,M\}$, the label pair is encoded as a learnable embedding vector $m_{s,t} \in \mathbb{R}^{d_m}$. The attention U-Net takes the source quantized latent $z_q^s = Q(x^s)$ and the label embedding to predict the target latent

$$\hat{z}_q^t = U(z_q^s, m_{s,t}), \tag{9}$$

with the resulting reconstruction $\hat{x}^{s\to t} = G(\hat{z}_q^t)$.

For training, paired images (x^s, x^t) from modalities (s,t) are mapped to quantized embeddings (z_q^s, z_q^t). During training, U and $\{m_{s,t}\}_{s,t}$ are updated to minimize the loss

$$\mathcal{L}_{\text{trans}} = \mathbb{E}_{(z_q^s,s,z_q^t,t)}[L_{\text{trans}}(z_q^s, s, z_q^t, t)], \tag{10}$$
$$L_{\text{trans}}(z_q^s, s, z_q^t, t) = \|z_q^t - \hat{z}_q^t\|_1 - \lambda \mathcal{D}_z(\hat{z}_q^t),$$

where $\mathcal{D}_z$ is a discriminator trained adversarially to minimize

$$\mathbb{E}_{\hat{z}_q^t, z_q^t}[\max\big(0,\, 1 + \mathcal{D}_z(\hat{z}_q^t)\big) + \max\big(0,\, 1 - \mathcal{D}_z(z_q^t)\big)]. \tag{11}$$

The discriminator follows the same PatchGAN architecture as D, but with the embedding dimensions.

C. Implementation Details

We train our models for 500 epochs for both the first stage (VQGAN) and second stage (translation) on a NVIDIA A40 GPU with a batch size of size 1 for the first stage and 3 for the second stage. We choose Adam as an optimizer for first and second stage with a learning rate of 2×10^{-6}. We set $\beta = 0.25$, $\gamma = \lambda = 0.8$, $K = 8192$ following the original VQGAN paper [6]. Architectural details are in Table II.

TABLE II
ARCHITECTURAL CONFIGURATIONS

VQGAN Encoder-Decoder (Stage 1)	Hierarchical encoder-decoder (U-Net style)
Base feature channels and multipliers	128 and [1, 1, 2] (3-level hierarchy)
Residual blocks per level	2
Latent embedding	$d_z = N_Q \times d_Q = 3 \times 64 = 192$
Codebook dimension and size	$K \times d_Q = 8192 \times 64$
PatchGAN Discriminator (Stage 1) or (Stage 2)	
Input channels	1 (Stage 1) or 192 (Stage 2)
Base conv. channels	32
Number of conv. layers	3
Latent-Space U-Net (Stage 2)	U-Net with attention (adapted from [43])
Base feature channels and multipliers	128 and [1, 1, 2] (3-level hierarchy)
Residual blocks per level	1
Conditioning	label embedding $d_m = 64$
Channels per self-attention head	32

IV. Evaluation and Results

We use the BraTs 2021 dataset which contains four modalities per subject: T1, T1 contrast-enhanced (T1ce), T2, and FLAIR, with 1251 subjects with tumors (ground truth labels are provided but are not used) for training and 219 subjects with tumors for testing [44]; however, the tumors on the test set are not labeled. Each MRI scan has a spatial resolution of 1 mm^3 per voxel and size of $240 \times 240 \times 155$.

We evaluate the translation performance using three standard image similarity metrics: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Normalized Mean Squared Error (NMSE) for all source-target modality pairs. We present both qualitative and quantitative evaluations of our method, comparing to the state-of-the-art in 3D translation [27] as a benchmark. Compared to the benchmark's one-to-many translation approach [27], our many-to-many architecture achieves competitive performance, as summarized in Table III. We highlight results that match or exceed the state-of-the-art [27] in bold. This suggests that for MRI modality translation a dynamic source-specific model may not be necessary to reach strong performance.

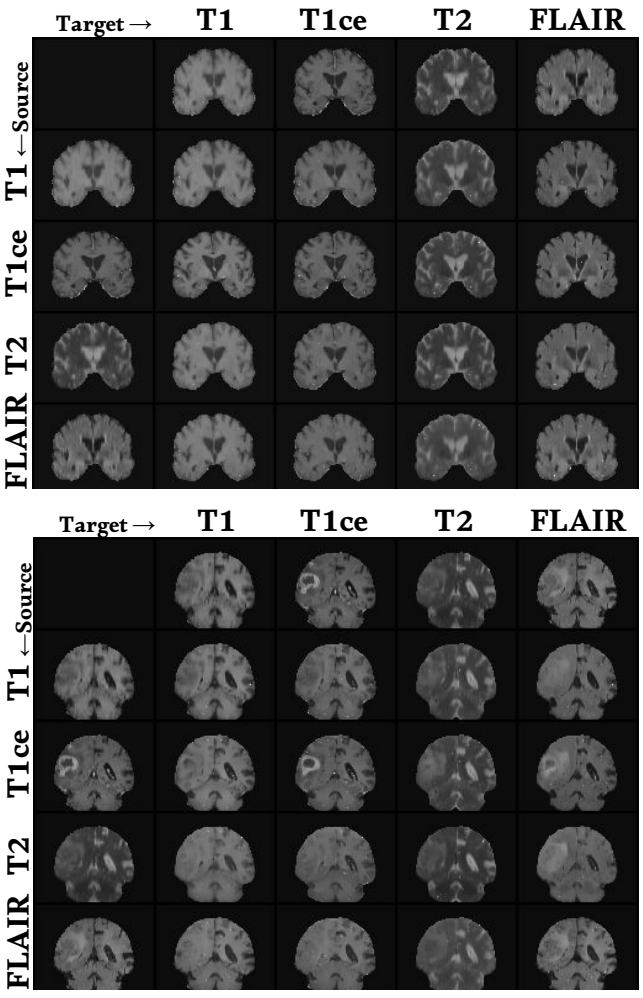

Fig. 1. Qualitative results on the 3D MRI translation benchmark. Each grid shows translations between source (rows) and target (columns) modalities. (Top) A typical case with no visible tumor. (Bottom) A case with a visible tumor region that is clearly visible in T1ce. Translations from T1ce to other modalities preserve the tumor to some extent. However, when translating *to* T1ce, none of the other source modalities can reconstruct the tumor, reflecting modality-specific limitations and the one-way nature of information flow.

For qualitative evaluation, Figure 1 shows two representative slices. The first case demonstrates a near-perfect

TABLE III

COMPARISON OF STATIC U-NET IMAGE TRANSLATION RESULTS AGAINST THE 3D BENCHMARK [27] IN PARENTHESES. MEANS AND STANDARD DEVIATIONS ACROSS THE 219 SUBJECTS ARE BOLDED WHEN OUR MODEL IS EQUAL TO OR BETTER THAN THE BENCHMARK.

Source \ Target	**T1**		**T1ce**		**T2**		**FLAIR**	
	Ours	[27]	Ours	[27]	Ours	[27]	Ours	[27]
T1 (PSNR↑)	**30.54 ± 2.35**	(29.00 ± 0.64)	**27.03 ± 2.61**	(26.12 ± 0.82)	24.17 ± 1.99	(25.82 ± 0.86)	24.68 ± 2.56	(24.84 ± 0.73)
T1 (NMSE↓)	**0.0010 ± 0.0005**	(0.0550 ± 0.0250)	**0.0024 ± 0.0018**	(0.0780 ± 0.0220)	**0.0044 ± 0.0030**	(0.1030 ± 0.0300)	**0.0041 ± 0.0029**	(0.1130 ± 0.0340)
T1 (SSIM↑)	**0.95 ± 0.01**	(0.94 ± 0.02)	**0.91 ± 0.03**	(0.91 ± 0.02)	**0.90 ± 0.04**	(0.90 ± 0.01)	**0.88 ± 0.04**	(0.86 ± 0.02)
T1ce (PSNR↑)	**27.11 ± 2.27**	(26.23 ± 0.79)	**31.63 ± 1.41**	(28.76 ± 0.89)	24.23 ± 2.04	(25.99 ± 0.86)	24.61 ± 2.48	(25.20 ± 0.81)
T1ce (NMSE↓)	**0.0022 ± 0.0011**	(0.0760 ± 0.0270)	**0.0007 ± 0.0002**	(0.0600 ± 0.0190)	**0.0044 ± 0.0036**	(0.0920 ± 0.0320)	**0.0041 ± 0.0026**	(0.0920 ± 0.0500)
T1ce (SSIM↑)	**0.92 ± 0.02**	(0.92 ± 0.03)	**0.94 ± 0.01**	(0.94 ± 0.01)	0.89 ± 0.04	(0.91 ± 0.01)	**0.88 ± 0.05**	(0.88 ± 0.04)
T2 (PSNR↑)	**26.38 ± 2.35**	(25.42 ± 0.85)	**26.84 ± 2.07**	(25.23 ± 1.15)	28.74 ± 1.71	(29.23 ± 0.72)	**25.26 ± 2.68**	(25.07 ± 1.09)
T2 (NMSE↓)	**0.0026 ± 0.0014**	(0.0850 ± 0.0260)	**0.0023 ± 0.0015**	(0.0870 ± 0.0340)	**0.0014 ± 0.0006**	(0.0480 ± 0.0180)	**0.0036 ± 0.0026**	(0.0980 ± 0.0210)
T2 (SSIM↑)	**0.91 ± 0.03**	(0.91 ± 0.02)	0.89 ± 0.03	(0.90 ± 0.03)	**0.94 ± 0.02**	(0.94 ± 0.01)	**0.87 ± 0.06**	(0.87 ± 0.02)
FLAIR (PSNR↑)	**26.40 ± 2.43**	(25.19 ± 0.76)	**26.73 ± 2.08**	(25.90 ± 1.04)	24.07 ± 1.98	(26.15 ± 0.64)	**30.19 ± 2.05**	(28.61 ± 0.77)
FLAIR (NMSE↓)	**0.0027 ± 0.0014**	(0.0900 ± 0.0280)	**0.0024 ± 0.0014**	(0.0940 ± 0.0250)	**0.0044 ± 0.0028**	(0.0860 ± 0.0280)	**0.0011 ± 0.0005**	(0.0580 ± 0.0250)
FLAIR (SSIM↑)	**0.91 ± 0.05**	(0.91 ± 0.05)	0.89 ± 0.05	(0.91 ± 0.03)	0.88 ± 0.06	(0.91 ± 0.01)	0.93 ± 0.02	(0.94 ± 0.03)

reconstruction of slices for all source-target pairs without an obvious tumor (the test set does not have tumors labeled, a tumor may be present). The second case has a tumor distinctly visible in the T1ce modality. When translating to FLAIR from T1 or T1ce, the tumor is relatively well-preserved across all source modalities. However, when translating to T1, only when the source is T1ce does the output preserve the tumor details present in T1, which aligns with the clinical relevance of T1ce in enhancing tumor contrast.

More notably, when T1ce is used as the *target*, none of the source modalities successfully reconstruct the tumor. We attribute this to an inherent information gap: since the contrast enhancement in T1ce is often due to injected contrast agents, this information is not present in other modalities. By the information processing inequality, one cannot recover modality-specific features that are not already encoded in the source. This limitation manifests as an 'inpainting' effect, where the tumor appears smoothed over or absent in the generated T1ce image.

V. DISCUSSION AND FUTURE WORK

While our many-to-many model is compact and scalable, it treats all source-target pairs equally during training. One direction for future research is to incorporate source-specific attention during the encoding to better model modality priors. Another future work is uncertainty estimation that may help identify cases where information is fundamentally missing from the source and translation is ill-posed.

Finally, the clinical relevance of modality transfer should be verified. It may be the case that modality transfer can help when certain modalities are unavailable. A first step, which avoids manual labeling, would be to see if translated images work for pretrained models used for tumor segmentation, or other neuroimaging tasks. Another step would be see if differences in modality transfer can be used to highlight meaningful differences between imaging modalities without simply highlighting spurious artifacts caused by the translation model.

VI. CONCLUSION

We presented a 3D generative model for MRI modality translation based on a vector-quantized GAN (VQGAN) backbone and a conditional U-Net operating in the latent space. Our architecture supports many-to-many modality translation by conditioning on discrete modality-pair embeddings. This design allows the model to flexibly translate between any pair of modalities within a unified framework, without requiring separate models for each direction.

Compared to the state-of-the-art in 3D MRI translation [27], we demonstrated that our model achieves competitive quantitative results compared to specialized one-to-many baselines, despite its broader scope. Qualitative results further highlight that the model can preserve fine anatomical and pathological structures when sufficient information is present in the source modality. Conversely, when modality-specific information is absent—such as the exogenous contrast that enhances tumor in T1ce—consistent with the information processing inequality and prior observations [17].

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 2108841. Effort at the University of Delaware was sponsored by the Department of the Navy, Office of Naval Research under ONR award number N00014-24-1-2259. This research was supported in part through the use of DARWIN computing system: DARWIN – A Resource for Computational and Data-intensive Research at the University of Delaware and in the Delaware Region, which is supported by NSF under Grant Number: 1919839, Rudolf Eigenmann, Benjamin E. Bagozzi, Arthi Jayaraman, William Totten, and Cathy H. Wu, University of Delaware, 2021, URL: https://udspace.udel.edu/handle/19716/29071.

REFERENCES

[1] P. P. Swoboda, A. Larghat, A. Zaman, T. A. Fairbairn, M. Motwani, J. P. Greenwood, and S. Plein, "Reproducibility of myocardial strain and left ventricular twist measured using complementary spatial modulation of magnetization," *J. Magn. Reson. Imaging*, vol. 39, no. 4, pp. 887–894, Apr. 2014.

[2] I. Ø. Brandsæter, E. R. Andersen, B. M. Hofmann, and E. Kjelle, "Drivers for low-value imaging: a qualitative study of stakeholders' perspectives in Norway," *BMC Health Serv. Res.*, vol. 23, no. 1, pp. 1–11, Dec. 2023.

[3] F. Reyes-Santias, C. García-García, B. Aibar-Guzmán, A. García-Campos, O. Cordova-Arevalo, M. Mendoza-Pintos, S. Cinza-Sanjurjo, M. Portela-Romero, P. Mazón-Ramos, and J. R. Gonzalez-Juanatey, "Cost Analysis of Magnetic Resonance Imaging and Computed Tomography in Cardiology: A Case Study of a University Hospital Complex in the Euro Region," *Healthcare*, vol. 11, no. 14, p. 2084, Jul. 2023.

[4] R. Bitar, G. Leung, R. Perng, S. Tadros, A. R. Moody, J. Sarrazin, C. McGregor, M. Christakis, S. Symons, A. Nelson, and T. P. Roberts, "MR Pulse Sequences: What Every Radiologist Wants to Know but Is Afraid to Ask1," *Radiographics*, Mar. 2006.

[5] P. Sati, I. C. George, C. D. Shea, M. I. Gaitán, and D. S. Reich, "FLAIR∗: A Combined MR Contrast Technique for Visualizing White Matter Lesions and Parenchymal Veins," *Radiology*, vol. 265, no. 3, p. 926, Dec. 2012.

[6] P. Esser, R. Rombach, and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," pp. 12 873–12 883, 2021, [Online; accessed 26. May 2025].

[7] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[8] A. Avesta, S. Hossain, M. Lin, M. Aboian, and S. Aneja, "Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation," *Bioengineering*, vol. 10, no. 2, p. 181, Feb. 2023.

[9] Y. Yan, H. Wang, Y. Huang, N. He, L. Zhu, and Y. Xu, "Cross-modal vertical federated learning for MRI reconstruction," *IEEE J. Biomed. Health Inf.*, vol. 28, no. 11, pp. 6384–6394, Jan. 2024.

[10] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk, "Domain generalization for medical image analysis: A review," *arXiv*, Oct. 2023.

[11] C. Li, K. Chen, and X. Huang, "Reg-gan: adaptive regularization of gans for unpaired image-to-image translation," *IEEE Transactions on Multimedia*, vol. 24, pp. 3806–3818, 2021.

[12] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *arXiv*, Jun. 2021.

[13] J. Cho, X. Liu, F. Xing, J. Ouyang, G. E. Fakhri, J. Park, and J. Woo, "Disentangled multimodal brain MR image translation via transformer-based modality infuser," in *Proceedings Volume 12926, Medical Imaging 2024: Image Processing*. SPIE, Apr. 2024, vol. 12926, pp. 602–607.

[14] J. Cho, and J. Park, "A Unified Framework for Synthesizing Multisequence Brain MRI via Hybrid Fusion," *arXiv*, Jun. 2024.

[15] H. Li, J. C. Paetzold, A. Sekuboyina, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze, "DiamondGAN: Unified Multi-modal Generative Adversarial Networks for MRI Sequences Synthesis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham, Switzerland: Springer, Oct. 2019, pp. 795–803.

[16] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Trans. Med. Imaging*, vol. 39, no. 9, pp. 2772–2781, Feb. 2020.

[17] D. Lee, W.-J. Moon, and J. C. Ye, "Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks," *Nat. Mach. Intell.*, vol. 2, pp. 34–42, Jan. 2020.

[18] M. Yurt, S. U. H. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, "mustGAN: multi-stream generative adversarial networks for MR image synthesis," *Med. Image Anal.*, vol. 70, p. 101944, May 2021.

[19] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *International Conference on Learning Representations (ICLR)*, 2016.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[21] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 319–345.

[22] Y. Chen, D. Nie, Y. Gao, and D. Shen, "Nice-gan: Noise injection and contrastive exploration for gan-based one-shot medical image synthesis," in *Medical Image Analysis*, vol. 75. Elsevier, 2022, p. 102293.

[23] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[25] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[26] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," *arXiv*, Oct. 2022.

[27] J. Kim and H. Park, "Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 7604–7613.

[28] C. Liu, M. Salzmann, T. Lin, R. Tomioka, and S. Süsstrunk, "On the loss landscape of adversarial training: Identifying challenges and how to overcome them," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 476–21 487, 2020.

[29] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-GANs: Edge-Aware Generative Adversarial Networks for Cross-Modality MR Image Synthesis," *IEEE Trans. Med. Imaging*, vol. 38, no. 7, pp. 1750–1762, Jan. 2019.

[30] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks," *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2375–2388, Feb. 2019.

[31] B. Xin, Y. Hu, Y. Zheng, and H. Liao, "Multi-Modality Generative Adversarial Networks with Tumor Consistency Loss for Brain MR Image Synthesis," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 03–07.

[32] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multi-modal MR Synthesis via Modality-Invariant Latent Representation," *IEEE Trans. Med. Imaging*, vol. 37, no. 3, pp. 803–814, Oct. 2017.

[33] A. Sharma and G. Hamarneh, "Missing MRI Pulse Sequence Synthesis Using Multi-Modal Generative Adversarial Network," *IEEE Trans. Med. Imaging*, vol. 39, no. 4, pp. 1170–1183, Oct. 2019.

[34] J. Liu, S. Pasumarthi, B. Duffy, E. Gong, K. Datta, and G. Zaharchuk, "One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation," *IEEE Trans. Med. Imaging*, vol. 42, no. 9, pp. 2577–2591, Mar. 2023.

[35] X. Liu, F. Xing, G. El Fakhri, and J. Woo, "A unified conditional disentanglement framework for multimodal brain mr image translation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 10–14.

[36] H. Jiang, M. Imran, T. Zhang, Y. Zhou, M. Liang, K. Gong, and W. Shao, "Fast-ddpm: Fast denoising diffusion probabilistic models for medical image-to-image generation," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[37] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[38] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[40] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium: Gans do not need to decrease a divergence at every step," in *International Conference on Learning Representations*, 2018.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[43] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, Dec. 2021.

[44] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," *arXiv*, Jul. 2021.