Non-Adversarial Learning: Vector-Quantized Common Latent Space for Multi-Sequence MRI

Luyi Han[1,2][0000−0003−4046−2763], Tao Tan[3,2(✉)], Tianyu Zhang[1,2,4], Xin Wang[2,4], Yuan Gao[2,4], Chunyao Lu[1,2], Xinglong Liang[1,2], Haoran Dou[5], Yunzhi Huang[6], and Ritse Mann[1,2]

[1] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands
[2] Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[3] Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao Special Administrative Region of China
[4] GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, P. Debyelaan 25, 6202 AZ, Maastricht, The Netherlands
[5] Center for Computational Imaging and Simulation Technologies in Biomedicine within the School of Computing at the University of Leeds, LS2 9JT Leeds, UK
[6] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China
taotan@mpu.edu.mo

Abstract. Adversarial learning helps generative models translate MRI from source to target sequence when lacking paired samples. However, implementing MRI synthesis with adversarial learning in clinical settings is challenging due to training instability and mode collapse. To address this issue, we leverage intermediate sequences to estimate the common latent space among multi-sequence MRI, enabling the reconstruction of distinct sequences from the common latent space. We propose a generative model that compresses discrete representations of each sequence to estimate the Gaussian distribution of vector-quantized common (VQC) latent space between multiple sequences. Moreover, we improve the latent space consistency with contrastive learning and increase model stability by domain augmentation. Experiments using BraTS2021 dataset show that our non-adversarial model outperforms other GAN-based methods, and VQC latent space aids our model to achieve (1) anti-interference ability, which can eliminate the effects of noise, bias fields, and artifacts, and (2) solid semantic representation ability, with the potential of one-shot segmentation. Our code is publicly available [7].

Keywords: Latent Space · MRI synthesis · Multi-Sequence MRI.

1 Introduction

Multi-sequence magnetic resonance imaging (MRI) is a commonly used diagnostic tool that provides clinicians with a comprehensive view of tissue character-

[7] https://github.com/fiy2W/mri_seq2seq

istics [4, 14, 13]. However, some sequences may be unusable or absent in clinical practice for various reasons [3], leading to the need for rescanning or disrupting downstream processes. To avoid this, deep generative models can be used to synthesize these missing sequences, but require many paired training data to produce high-quality results. In cases lacking paired data among source and target sequence, most studies [16, 11, 6] rely on generative adversarial networks (GANs) [7] to minimize the distribution distance between the generated and the target sequence. However, it can also lead to training instability and mode collapse, harming the image quality and structure.

Using intermediate sequences in multi-sequence MRI can make unsupervised generation less challenging. For example, if we have paired T1-weighted (T1) and T2-weighted (T2) MRI for one population and paired T2 and fluid-attenuated inversion recovery (Flair) MRI for another, we can use T2 to establish the relationship between T1 and Flair without paired samples. Compared to single-task models [16, 6], dynamic models [11, 5, 8] controlled by a prompt branch can integrate multiple generation tasks to utilize intermediate sequences. Han *et al.* [8] use a shared encoder to extract structural features from images, which are then rendered to target images with the guidance of a one-hot code. Jiang *et al.* [11] disentangle images into structure and style features and reconstruct target images using target styles and source structures. These methods preserve the structure consistency but ignore the distribution differences between the latent spaces of distinct sequences, hindering the model from learning the mapping of the common latent space to the target sequence.

In this work, we construct a common latent space for multi-sequence MRI so that all sequences can be mapped from it. Specifically, we first utilize VQ-VAE [17] to compress images into a discrete latent space, then estimate the distribution of the vector-quantized common (VQC) latent space based on these representations. Finally, we leverage a dynamic model Seq2Seq [8] to generate arbitrary target sequences from the VQC latent space. The VQC latent space has three primary advantages: (1) achieving unsupervised synthesis without requiring adversarial learning; (2) preventing input interference, such as noise, artifacts, and field bias; and (3) having reliable semantic representation, which shows the potential of one-shot segmentation.

2 Methods

2.1 Preliminary

VQ-VAE Compared with the continuous latent space of VAE [12], the discrete latent space of VQ-VAE captures more structured features while ignoring some irrelevant details, *e.g.,* artifacts. Given an encoder $\mathbf{E}$ and a decoder $\mathbf{G}$, we can map image X into a continuous latent space $z_e = \mathbf{E}(X)$ with a latent dimension of D, while $\mathbf{G}$ can restore X from z_e. Then, using a codebook e_k with the embedding dimension of K to map z_e to the nearest vectors in the codebook.

$$z_q(z_e) = e_k, \quad k = \arg\min_i \|z_e - e_i\| \tag{1}$$

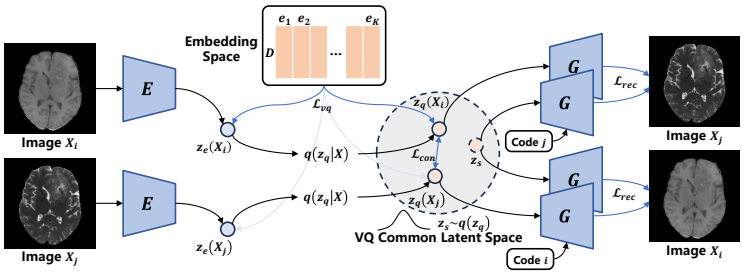

Fig. 1. Overview of the proposed VQ-Seq2Seq framework.

where the vector quantizing process is not differentiable, requiring an improved training loss,

$$\mathcal{L}_{\text{vqvae}} = \|X - \mathbf{G}(z_e + \text{sg}\left[z_q - z_e\right])\|_2^2 + \|\text{sg}\left[z_e\right] - z_q\|_2^2 + \beta \cdot \|\text{sg}\left[z_q\right] - z_e\|_2^2 \quad (2)$$

where $\text{sg}\left[\cdot\right]$ indicates a stop-gradient operation, and $\beta = 0.25$ ensures that z_e remains in proximity to z_q. To simplify the expression, we denote $z_e + \text{sg}\left[z_q - z_e\right]$ as z_q and merge the last two terms of Eq. 2 as $\mathcal{L}_{\text{vq}}$ in the following sections.

Dynamic Model Dynamic models [11, 5, 8] combine different generation tasks in a single model, which can utilize intermediate sequences. With a set of N sequences MRI $\mathcal{X} = \{X_i, f_i | i = 1, ..., N\}$, X_i is available if $f_i = 1$, otherwise $f_i = 0$ and the sequence is missing. The process of translating X_i to X_j is,

$$\hat{X}_{i \to j} = \mathbf{G}(\mathbf{E}(X_i), c_j) \quad (3)$$

where $\mathbf{G}$ refers to a dynamic decoder which input with structure feature $\mathbf{E}(X_i)$ and style feature c_j. In particular, c_j can be represented as a one-hot encoding for the target sequence [5, 8] or a style feature extracted from the target image [11]. In this work, we use Seq2Seq [8] as the baseline because the model is a simple autoencoder, which makes it easy to integrate the VQ module into the model.

2.2 VQ-Seq2Seq

Inheriting the advantages of discrete representations and dynamic models, we propose VQ-Seq2Seq to establish the VQC latent space for multi-sequence MRI. As shown in Fig. 1, continuous latent space z_e and corresponding discrete latent space z_q are extracted from the input images. By statistics on z_q, we can estimate a VQC latent space containing sampling points z_s that can reconstruct images of different sequences through the dynamic decoder $\mathbf{G}$.

Uncertainty Estimation It is challenging to strictly constrain multi-sequence MRI equal in latent space because one sequence involves specific information

that other sequences lack [16, 8]. To tolerate the sequence-specific attributes, we depict the probabilistic scope of z_q among different sequences by considering the uncertainty of the latent space. We propose a simple non-parametric method using the statistics of z_q for uncertainty estimation.

$$\begin{aligned} \mu_q(\mathcal{X}) &= \frac{1}{\sum_{i=1}^{N} f_i} \sum_{i}^{f_i \neq 0} z_q(X_i) \\ \sigma_q^2(\mathcal{X}) &= \frac{1}{\sum_{i=1}^{N} f_i - 1} \sum_{i}^{f_i \neq 0} (z_q(X_i) - \mu_q(\mathcal{X}))^2 \end{aligned} \tag{4}$$

VQC Latent Space After obtaining the uncertainty estimation, we can establish a Gaussian distribution for probabilistic statistics. To utilize randomness in further modeling the uncertainty, we use random sampling to draw the VQC latent space from the corresponding distribution randomly.

$$z_s = \mu_q(\mathcal{X}) + \epsilon \cdot \sigma_q^2(\mathcal{X}), \quad \epsilon \sim \mathcal{N}(0, 1) \tag{5}$$

Here, we use the re-parameterization trick to make the sampling operation differentiable, and ϵ follows the standard Gaussian distribution.

2.3 Loss Function

Pixel-Level Reconstruction We establish constraints between the generated image $\hat{X}$ and the target image X at the pixel, structural, and perceptual levels,

$$\mathcal{L}_{\text{rec}}(\hat{X}, X) = \lambda_1 \cdot \|\hat{X} - X\|_1 + \lambda_2 \cdot \mathcal{L}_{\text{ssim}}(\hat{X}, X) + \lambda_3 \cdot \mathcal{L}_{\text{per}}(\hat{X}, X) \tag{6}$$

where $\|\cdot\|_1$ refers to the L_1 loss, $\mathcal{L}_{\text{ssim}}$ indicates the SSIM loss [18], and $\mathcal{L}_{\text{per}}$ presents the perceptual loss [19] based on pre-trained VGG19. λ_1, λ_2, and λ_3 are weight terms and are experimentally set to be 10, 1, and 0.1.

Latent Space Consistency We ensure that z_q of sequences are close to narrowing the scope of VQC latent space. For two z_q (z_1 and z_2), we define a consistency loss composed with MSE and contrastive learning loss [15, 9].

$$\begin{aligned} \mathcal{L}_{\text{con}}(z_1, z_2) = &\|\text{sg}\left[z_1\right] - z_2\|_2^2 + \|\text{sg}\left[z_2\right] - z_1\|_2^2 \\ &- \sum_{p \in M} \log \frac{\exp\left(z_1^{(p)} \cdot z_2^{(p)} / \tau\right)}{\sum_{q \in M} \exp\left(z_1^{(p)} \cdot z_2^{(q)} / \tau\right)} \cdot \frac{\exp\left(z_2^{(p)} \cdot z_1^{(p)} / \tau\right)}{\sum_{q \in M} \exp\left(z_2^{(p)} \cdot z_1^{(q)} / \tau\right)} \end{aligned} \tag{7}$$

where p and q are features traversed from pixels in foreground of z_1 and z_2, $\tau = 0.07$ refers to the scalar temperature parameter.

Total Loss We formulate the total loss function using intermediate sequences without adversarial learning.

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\sum_{i}^{f_i\neq 0}\sum_{j}^{f_j\neq 0}\mathcal{L}_{\text{rec}}(\hat{X}_{i\to j}, X_j) + \sum_{i}^{f_i\neq 0}\mathcal{L}_{\text{rec}}(\hat{X}_{s\to i}, X_i) \\ &+\lambda_{\text{con}}\cdot\sum_{i}^{f_i\neq 0}\sum_{j}^{f_j\neq 0}\mathcal{L}_{\text{con}}(z_q(X_i), z_q(X_j)) + \lambda_{\text{vq}}\cdot\sum_{i}^{f_i\neq 0}\mathcal{L}_{\text{vq}}(z_e(X_i), z_q(X_i))\end{aligned} \tag{8}$$

where $\hat{X}_{i\to j} = \mathbf{G}(z_q(\mathbf{E}(X_i)), c_j)$ and $\hat{X}_{s\to j} = \mathbf{G}(z_s, c_j)$ are images generated from z_q and z_s, respectively. λ_{con} and λ_{vq} are both experimentally set to be 10.

2.4 Random Domain Augmentation

We use random domain augmentation for input images during training to further improve the stability of VQ-Seq2Seq and the anti-interference ability of VQC latent space. The domain augmentation process has three aspects: (1) simple intensity transformation $\mathcal{T}$ (*e.g.*, gamma transformation, random noise, and bias field); (2) cross-sequence translation with one-hot codes c_r; and (3) random domain translation with random target codes $c_r \sim \mathcal{U}(0, 1)$. The latter two augmentation methods allow us to generate an augmented image $X_{\text{aug}} = \mathbf{G}(z_q(X_i), c_r)$ from the input image X_i, and the first method makes $X_{\text{aug}} = \mathcal{T}(X_i)$. During training, we will randomly replace the input image X_i with one of X_{aug}.

3 Experiments

3.1 Experimental Settings

Dataset and Evaluation Metrics We utilize brain MRI images from the Brain Tumor Segmentation 2021 (BraTS2021) dataset [14, 2, 1], comprising 1,251 subjects with four aligned sequences: T1, T1Gd, T2, and Flair. From this dataset,

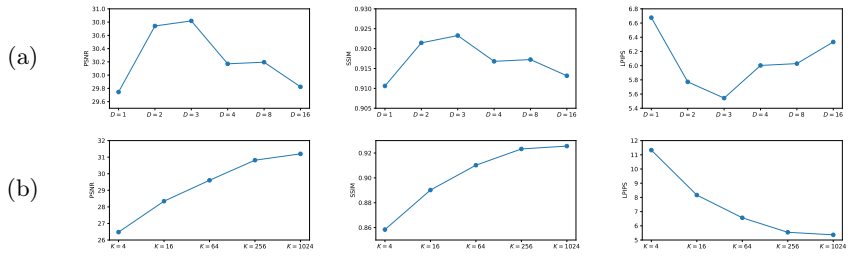

Fig. 2. Synthesis performance of VQ-Seq2Seq with different latent dimensions (D) and embedding dimensions (K). (a) Synthesis performance with different latent dimensions ($K = 256$); (b) Synthesis performance with different embedding dimensions ($D = 3$).

Table 1. The quantitative results of translating T1 to T1Gd, T2, and Flair with a single step or multiple steps. The best result is in bold, and the second best is underlined.

Step	Method	T1→T1Gd			T1→(T1Gd)→T2			T1→(T1Gd→T2)→Flair		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Single	MM-GAN [16]	26.5±2.0	0.871±0.044	11.4±4.2	22.2±1.0	0.779±0.028	20.5±3.4	21.7±1.2	0.791±0.039	17.8±3.1
	ResViT [6]	26.4±2.0	0.872±0.040	11.6±4.1	21.8±0.9	0.774±0.035	16.1±3.2	20.9±0.9	0.705±0.035	20.0±3.4
	Jiang *et al.* [11]	26.7±2.7	0.874±0.044	10.4±4.1	23.7±2.2	0.833±0.039	12.3±3.9	23.5±2.4	0.796±0.054	11.9±3.6
	Seq2Seq [8]	26.9±2.2	0.876±0.040	9.68±3.74	26.5±1.9	0.884±0.039	8.25±3.32	24.3±2.3	0.811±0.047	11.2±3.5
	+VQ	27.0±2.2	0.875±0.040	9.79±3.67	26.7±1.9	0.885±0.038	8.05±3.22	24.6±2.3	0.817±0.043	11.0±3.5
	+VQ+$\mathcal{L}_{con}$	27.1±2.1	0.876±0.039	**9.67±3.61**	26.8±1.9	0.886±0.038	7.85±3.30	24.7±2.2	0.823±0.040	10.8±3.4
	VQ-Seq2Seq	**27.1±2.1**	**0.876±0.039**	9.78±3.49	**27.1±1.9**	**0.890±0.036**	7.63±3.07	25.7±1.9	**0.847±0.033**	9.86±3.05
	w/o Aug	27.0±2.0	0.875±0.038	9.71±3.41	26.8±1.8	0.883±0.036	8.23±3.17	24.1±2.2	0.805±0.050	11.4±3.5
Multiple	MM-GAN [16]	-	-	-	25.4±1.7	0.866±0.037	10.9±3.4	24.9±1.3	0.826±0.032	14.2±3.4
	ResViT [6]	-	-	-	25.7±1.7	0.861±0.032	10.7±3.2	24.9±1.1	0.826±0.037	14.5±4.2
	Jiang *et al.* [11]	-	-	-	26.0±1.9	0.874±0.037	9.70±3.16	25.3±1.6	0.835±0.030	10.5±3.8
	Seq2Seq [8]	-	-	-	26.4±1.8	0.883±0.037	7.94±3.03	25.5±1.5	0.843±0.029	9.83±2.84
	+VQ	-	-	-	26.4±1.8	0.878±0.037	8.15±2.98	25.5±1.5	0.839±0.028	9.83±2.65
	+VQ+$\mathcal{L}_{con}$	-	-	-	26.6±1.8	0.881±0.036	7.74±2.89	25.7±1.6	0.843±0.028	9.63±2.67
	VQ-Seq2Seq	-	-	-	26.8±1.8	0.884±0.035	**7.63±2.76**	**25.9±1.5**	0.846±0.028	**9.47±2.56**
	w/o Aug	-	-	-	26.6±1.7	0.877±0.034	8.23±2.81	25.7±1.4	0.840±0.028	11.1±2.7

we allocated 830 subjects for training, 93 for validation, and 328 for testing. To simulate clinical settings with missing sequences, we divided the training set into three subsets, which contained paired sequences between (T1, T1Gd), (T1Gd, T2), and (T2, Flair), respectively. It can be simulated that there is no paired sample between T1 and Flair under this setting, but there are two partially paired intermediate sequences, T1Gd and T2. All images undergo intensity normalization to a range of [0, 1] and are subsequently centrally cropped to dimensions of 128×192×192. Synthesis performance is evaluated using metrics including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS).

Implementation Details We implemented the models using PyTorch and trained them on the NVIDIA GeForce RTX 3090 Ti GPU. The architecture of **E** and **G** is the same as Seq2Seq [8]. The proposed VQ-Seq2Seq is trained using the AdamW optimizer, with an initial learning rate of 10^{-4} and a batch size of 1 for 1,000,000 steps. All comparative experiments use domain augmentation, at least with simple intensity transformation $\mathcal{T}$, to ensure a fair comparison. $\mathcal{T}$ involves applying random gamma transformation with $\gamma \sim \mathcal{U}(0.95, 1.05)$, random Gaussian noise with $\sigma \sim \mathcal{U}(0, 0.1)$, and random bias field with scale of 0.2 and degree of intensity inhomogeneity $\alpha \sim \mathcal{U}(0, 2)$.

3.2 Experimental Results

Latent and Embedding Dimension Referring to Sec. 2.1, the latent dimension D represents the dimension of the compressed feature. The smaller D is, the greater the degree of compression. The embedding dimension K indicates the number of discrete vectors (clustering) in the codebook. The larger K is, the better a discrete vector fits the continuous features. We train VQ-Seq2Seq using the training sets with complete sequences to explore the optimal D and K before other experiments. As shown in Fig. 2, when $K = 256$, the proposed model

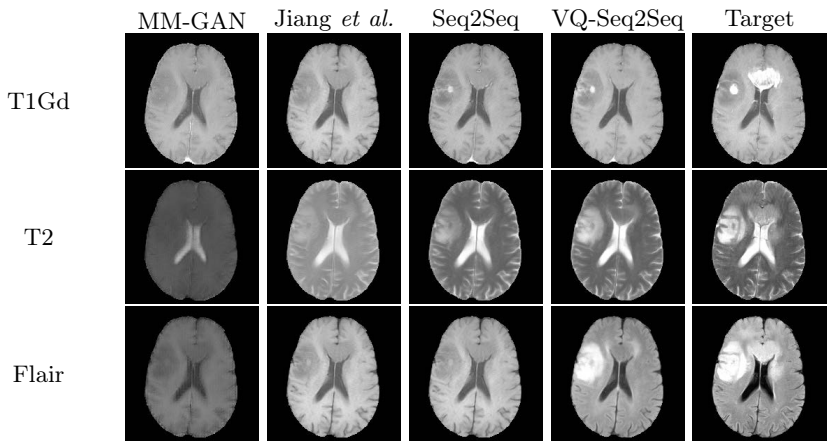

Fig. 3. Visualization of translating T1 to T1Gd, T2, and Flair with a single step.

Table 2. The quantitative results for comparisons of reconstructing images based on noise and bias field data. The best result is in bold, and the second best is underlined.

Method	Noise			Bias Field		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MM-GAN [16]	29.0±0.5	0.860±0.028	20.6±5.1	22.1±0.7	0.928±0.015	5.56±1.43
ResViT [6]	28.6±2.3	0.851±0.023	19.8±8.6	20.0±1.7	0.914±0.016	6.37±1.88
Jiang *et al.* [11]	30.1±1.8	<u>0.895±0.021</u>	10.3±3.1	21.2±0.8	0.924±0.019	5.98±1.18
Seq2Seq [8]	28.2±2.6	0.861±0.024	14.6±3.9	22.0±1.2	0.927±0.019	5.49±1.23
+VQ	29.0±2.9	0.877±0.027	10.7±3.1	22.4±1.2	0.928±0.018	5.14±1.31
+VQ+$\mathcal{L}_{\text{con}}$	<u>30.3±1.5</u>	0.891±0.017	<u>9.38±2.73</u>	22.6±1.4	**0.930±0.019**	**5.09±1.27**
VQ-Seq2Seq	**30.3±1.5**	**0.902±0.016**	**7.02±1.73**	**26.1±2.6**	<u>0.930±0.020</u>	<u>5.09±1.51</u>
w/o Aug	29.2±2.5	0.864±0.044	10.7±2.9	22.6±1.6	0.916±0.019	6.23±1.46

performs the best when $D = 3$. Additionally, when $D = 3$, the performance of the model continues to improve as K increases, but the rate of improvement slows down after $K > 256$. Thus, we set $D = 3$ and $K = 256$ in this work.

Latent Space Consistency To evaluate the effectiveness of the proposed VQC latent space for unsupervised cross-sequence generation, we compared VQ-Seq2Seq with other methods such as MM-GAN [16], ResViT [6], Jiang *et al.* [11], and Seq2Seq [8]. Additionally, we compared the three components of our method, which include VQ embedding, VQ with $\mathcal{L}_{\text{con}}$, and domain augmentation. There are two ways to implement a source→target generation: (1) generate the target directly from the source (single-step), and (2) first generate an intermediate sequence from the source and then generate the target (multi-step). Table 1 and Fig. 3 illustrate the synthesis performance of comparisons on translating T1→T1Gd, T1→T2, and T1→Flair. Note that, due to the settings of paired samples in the training set, the multi-step generation between T1 and T2 requires two steps, and between T1 and Flair requires three steps. As shown in Table 1, the comparison method achieves similar performance for the T1→T1Gd gener-

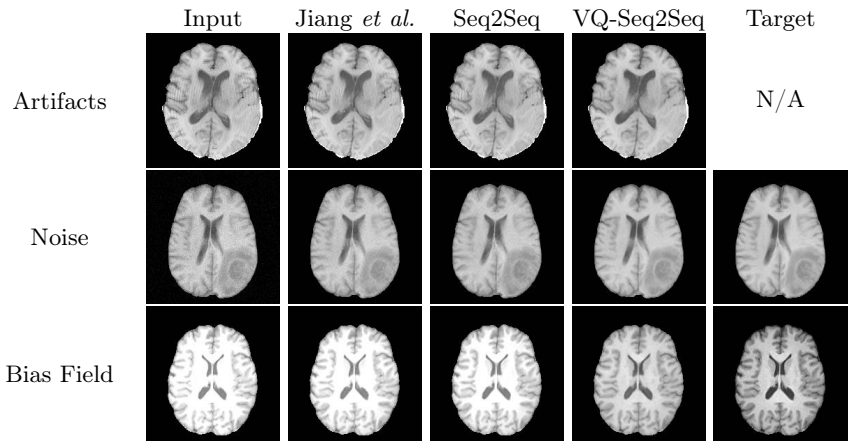

Fig. 4. Visualization of reconstruction from input images with artifacts, noise, and bias field. Artifacts exist in the original images, therefore, the target image is unavailable.

Table 3. The quantitative one-shot segmentation results for using latent space from comparisons. The best result is in bold. ET: enhanced tumor, TC: tumor core, WT: whole tumor.

Method	DSC↑			ASSD↓		
	ET	TC	WT	ET	TC	WT
nnU-Net [10]	0.481±0.298	0.457±0.308	0.463±0.231	12.7±11.9	13.7±11.2	14.2±7.1
Jiang *et al.* [11]	0.193±0.276	0.175±0.272	0.328±0.272	24.3±10.3	23.6±10.6	15.9±8.8
Seq2Seq [8]	0.276±0.234	0.274±0.224	0.386±0.206	19.3±9.4	19.4±8.4	15.6±5.4
VQ-Seq2Seq	**0.557±0.275**	**0.532±0.301**	**0.638±0.201**	**6.46±9.74**	**8.18±9.97**	**9.36±5.96**

ation task with paired samples. However, when it comes to unpaired T1→T2 and T1→Flair generation tasks, the performance of the comparison method decreases sharply when performing single-step generation compared to multi-step generation. In contrast, the proposed VQ-Seq2Seq shows only a minor performance penalty on T1→Flair task and improves on T1→T2 task. This shows that multi-step generation will lead to information loss and error accumulation, and our VQ-Seq2Seq can alleviate this problem through single-step generation.

Anti-Interference The proposed VQC latent space also has the anti-interference ability. We add fixed Gaussian noise and bias fields to the input image and reconstruct the input image using the comparisons. As shown in Table 2, the proposed method can effectively prevent the interference of noise and bias fields to reconstruct the original image. Fig. 4 shows the visualization results of the reconstruction, in which we found that the proposed model can also remove artifacts in images.

Compression and Representation The proposed VQC latent space showcases strong representation ability, indicating the potential of one-shot segmen-

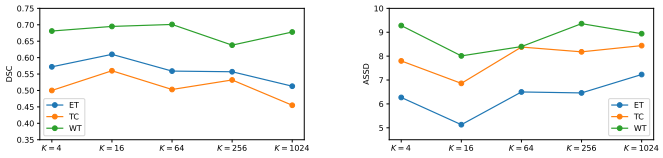

Fig. 5. One-shot segmentation performance of VQ-Seq2Seq with different embedding dimensions (K).

tation. To demonstrate this, we train the nnU-Net model based on the VQC latent space for brain tumor segmentation. For this purpose, we only use one subject containing all sequences for training. As shown in Table 3, the segmentation model trained based on the VQC latent space outperforms the model trained using only images. Furthermore, Fig. 5 shows that fewer VQ embedding dimensions $K = 16$ contribute towards the clustering of image semantics, which improves the segmentation performance.

4 Conclusion

In this work, we introduce a network for estimating the distribution of VQC latent space, which inherits the advantage of discrete representations and dynamic models. Experimental results based on BraTS2021 demonstrate that this latent space contributes to cross-sequence generation without adversarial learning and has substantial anti-interference and representation ability.

Acknowledgments. Luyi Han was funded by Chinese Scholarship Council (CSC) scholarship. This work is supported by Science and Technology Development Fund of Macao (0021/2022/AGJ).

Disclosure of Interests. The authors declare no competing interests.

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
3. Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A.: Multimodal mr synthesis via modality-invariant latent representation. IEEE transactions on medical imaging **37**(3), 803–814 (2017)

4. Chen, J.H., Su, M.Y., et al.: Clinical application of magnetic resonance imaging in management of breast cancer patients receiving neoadjuvant chemotherapy. BioMed research international **2013** (2013)
5. Chen, Z., Cai, L., Chen, C., Fu, X., Yang, X., Yuan, B., Lu, Q., Zhou, H.: Unsupervised image-to-image translation in multi-parametric mri of bladder cancer. Engineering Applications of Artificial Intelligence **124**, 106547 (2023)
6. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
8. Han, L., Tan, T., Zhang, T., Huang, Y., Wang, X., Gao, Y., Teuwen, J., Mann, R.: Synthesis-based imaging-differentiation representation learning for multi-sequence 3d/4d mri. Medical Image Analysis **92**, 103044 (2024)
9. Hu, X., Zhou, X., Huang, Q., Shi, Z., Sun, L., Li, Q.: Qs-attn: Query-selected attention for contrastive learning in i2i translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18291–18300 (2022)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Jiang, J., Veeraraghavan, H.: Unified cross-modality feature disentangler for unsupervised multi-domain mri abdomen organs segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 347–358. Springer (2020)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
13. Mann, R.M., Cho, N., Moy, L.: Breast mri: state of the art. Radiology **292**(3), 520–536 (2019)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
15. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
16. Sharma, A., Hamarneh, G.: Missing mri pulse sequence synthesis using multi-modal generative adversarial network. IEEE transactions on medical imaging **39**(4), 1170–1183 (2019)
17. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
19. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)