Recover Semantics First, Generate Better: Improved Latent Modeling for 3D MRI Reconstruction and Cross-Contrast Synthesis

Yonghao Chen[1,2*], Sicheng Yang[1*], Rui Tang[1], and Lei Zhu[1✉]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Xi'an Jiaotong University

Abstract. Multi-contrast magnetic resonance imaging (MRI) provides complementary information for clinical diagnosis. However, acquiring all MRI sequences is often time-consuming and costly. Recent generative models perform cross-contrast synthesis to address this issue by inferring absent contrasts from the available ones. Nevertheless, synthesizing 3D MRI presents significant challenges. Due to the massive volume sizes, operating directly in the pixel space is computationally prohibitive; therefore, a common approach is to first compress the 3D volumes into a latent space and subsequently train generative models in that space. We observe that existing compression architectures face several critical issues: they under-preserve long-range anatomical coherence, discard clinically meaningful semantics, and rely on optimization objectives that lead to over-smoothed reconstructions. Ultimately, these shortcomings compromise the performance of subsequent generative models. In this work, we propose a semantics-first latent modeling framework for 3D MRI reconstruction and cross-contrast synthesis. Specifically, we introduce a Latent Harmonization Encoder (LHE) to capture global anatomical dependencies, ensuring coherent volumetric representations. To mitigate semantic degradation during latent compression, we further design a Semantic Recovery Block (SRB) that injects high-level priors from a self-supervised semantic teacher, enhancing contrast-aware separability in the latent space. Additionally, we propose an Anatomy-aware Frequency Loss (AFL) to adaptively preserve diagnostically relevant high-frequency structures. Extensive experiments on two public multi-contrast MRI datasets demonstrate consistent improvements in reconstruction fidelity and cross-contrast synthesis quality. Our code is available at https://github.com/script-Yang/RSF.

Keywords: Latent representation learning · 3D MRI synthesis

1 Introduction

Multi-contrast MRI provides complementary information vital for diagnosis, but acquiring all sequences is resource-intensive, driving the need for cross-contrast

* Equal contribution.
✉ Lei Zhu (leizhu@hkust-gz.edu.cn) is the corresponding author.

synthesis [4,16,19]. While clinical applications increasingly demand **3D MRI synthesis** to preserve spatial continuity, massive volume sizes (e.g., $256 \times 256 \times 256$) make pixel-space generation computationally prohibitive [24,17]. Consequently, current methods adopt a hierarchical approach: compressing 3D volumes into a latent space via VAE/VQ models [18,27], followed by latent-space generation via GANs [9] or Diffusion [24]. Despite its efficiency, applying this pipeline to 3D MRI presents three primary hurdles.

(1) Long-range anatomical incoherence: Anatomical structures in 3D MRI exhibit strict spatial continuity across the volume. However, conventional latent models often fail to capture these long-range dependencies, leading to structural distortions and disjointed anatomy in the generated volumes [26,10]. **(2) Semantic entanglement across contrasts:** Different MRI protocols encode distinct contrast-specific semantics. Existing compressors often entangle these nuances, causing a semantic degradation that impairs downstream generative modeling [6]. **(3) Over-smoothing from pixel-wise objectives:** Standard reconstruction losses inherently prioritize low-frequency accuracy, yielding overly smooth representations that obscure clinically vital fine-grained details, such as subtle boundaries or small lesions [14,15].

To address these challenges, we propose a unified framework centered on the principle of *recovering semantics first to enable better generation*. Our framework integrates three cooperative components: (1) a **Latent Harmonization Encoder (LHE)** that facilitates global information interaction to ensure anatomical coherence; (2) a **Semantic Recovery Block (SRB)** that leverages high-level self-supervised priors to enhance contrast-specific discriminability; and (3) an **Anatomy-aware Frequency Loss (AFL)** that adaptively enforces high-frequency fidelity in clinically relevant regions. Extensive experiments on two public datasets demonstrate that our approach achieves state-of-the-art performance in both 3D MRI reconstruction and multi-contrast synthesis.

2 Methodology

2.1 Overall Framework

As shown in Figs. 1 and 3, our framework consists of three cooperative components: 1) a Latent Harmonization Encoder (LHE), which captures long-range anatomical dependencies; 2) a Semantic Recovery Block (SRB), which re-injects high-level semantic guidance into the latent representation; and 3) an Anatomy-aware Frequency Loss (AFL), which enforces high-frequency consistency in clinically important regions.

2.2 Latent Harmonization Encoder (LHE)

In 3D MRI synthesis, purely convolutional encoders often fail to capture the long-range spatial coherence of anatomical structures, leading to inconsistencies [28,10]. To address this, we introduce a context-aware harmonization stage

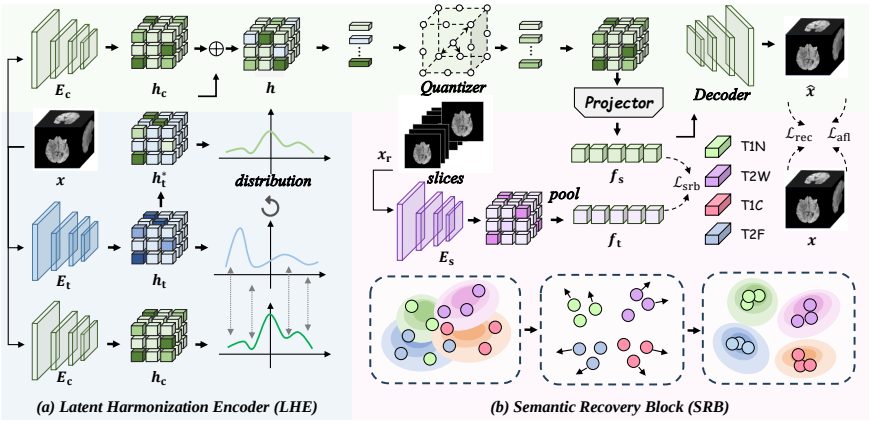

Fig. 1: Overview of our proposed semantics-first latent modeling framework. The model compresses 3D MRI volumes into a structured latent space using: (a) a Latent Harmonization Encoder (LHE), which captures long-range anatomical dependencies via heterogeneous encoders E_c and E_t; (b) a Semantic Recovery Block (SRB), which aligns latent features f_s with high-level priors f_t from a self-supervised teacher to restore lost semantic information and ensure contrast-aware separability; and (c) an Anatomy-aware Frequency Loss (AFL, detailed in Fig. 3) to mitigate the over-smoothing caused by reconstruction losses.

before latent discretization. Given an input volume x, we extract local convolutional features $h_c = E_c(x)$. To capture global dependencies, we employ a parallel, slice-wise Vision Transformer (ViT) [7] to extract context features $h_t = E_t(x)$.

To mitigate statistical mismatches between the two distinct pathways, we apply channel-wise feature alignment prior to fusion:

$$h_t^* = \frac{h_t - \mu_t}{\sigma_t + \epsilon} \cdot \sigma_c + \mu_c, \tag{1}$$

where μ and σ denote channel-wise mean and standard deviation. The aligned ViT features are then integrated via residual fusion: $h = h_c + h_t^*$.

Finally, the fused representation h is discretized via finite scalar quantization (FSQ) [22,1]. Quantizing these context-aware features ensures that the discrete latent codes retain long-range anatomical relationships, preventing the structural fragmentation often seen in standard compression.

2.3 Semantic Recovery Block (SRB)

Standard VAE/VQ-based compression prioritizes low-level fidelity over high-level anatomical semantics [27,29,31]. In multi-contrast MRI, this causes latent entanglement (Fig. 2(a)), impeding cross-contrast correspondence learning. To mitigate this, our Semantic Recovery Block (SRB) re-injects semantic priors into the latent space via a self-supervised teacher [5].

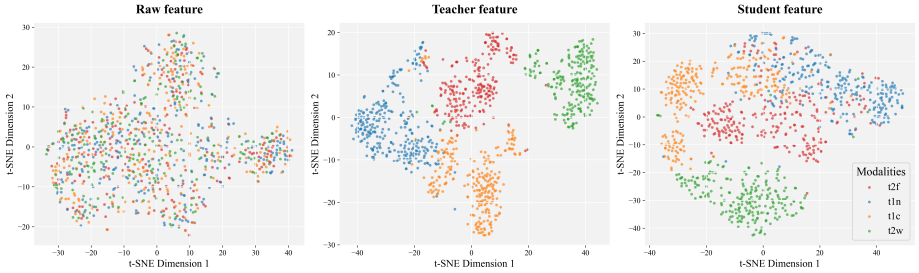

Fig. 2: T-SNE visualization of latent representations. (a) Raw FSQ latents exhibit severe entanglement, hindering cross-contrast learning. (b) DINO teacher features f_{t} show distinct contrast-wise clustering. (c) Our Semantic Recovery Block (SRB) aligns student embeddings f_{s} with the teacher, restoring semantic separability and enhancing the generative model's cross-contrast mapping.

As shown in Fig. 2(b), modern self-supervised models naturally extract contrast-specific features that exhibit clear clustering. Leveraging this, we feed a randomly sampled 2D slice x_r from the input volume x into a frozen DINO-pretrained ViT-B/16 [5]. The teacher semantic embedding f_t is obtained via global average pooling of normalized final-layer patch tokens:

$$f_{\mathrm{t}} = \mathrm{Pool}(\mathrm{E_s}(x_{\mathrm{r}})). \tag{2}$$

Simultaneously, the quantized latent z is mapped into the same semantic space through a learnable MLP projector $P(\cdot)$ to produce the student vector f_{s}:

$$f_{\mathrm{s}} = P(z). \tag{3}$$

Alignment is enforced via the semantic latent alignment loss:

$$\mathcal{L}_{\mathrm{srb}} = \|f_{\mathrm{s}} - f_{\mathrm{t}}\|_2^2 . \tag{4}$$

By optimizing $\mathcal{L}_{\mathrm{srb}}$, the latent embeddings achieve superior contrast-wise separability (Fig. 2(c)), facilitating the learning of more robust cross-contrast relations for downstream synthesis.

2.4 Anatomy-aware Frequency Loss (AFL)

Standard volumetric reconstruction objectives typically employ Mean Squared Error (MSE) loss, which prioritizes low-frequency components and often results in over-smoothed outputs [23,21]. To preserve high-frequency details essential for clinical diagnosis, we propose the Anatomy-aware Frequency Loss (AFL). As illustrated in Fig. 3, AFL adaptively enforces high-frequency consistency using a joint anatomy-semantic attention mechanism.

To emphasize diagnostically meaningful regions, we construct a joint attention weight $A_j = A_a \odot A_s$. This is derived from two parallel streams. First,

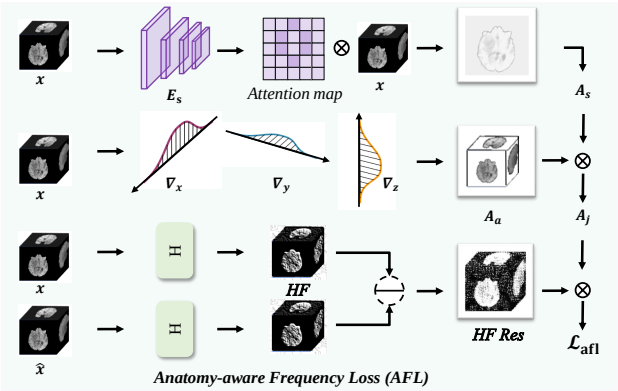

Fig. 3: Illustration of the proposed Anatomy-aware Frequency Loss (AFL) mechanism.

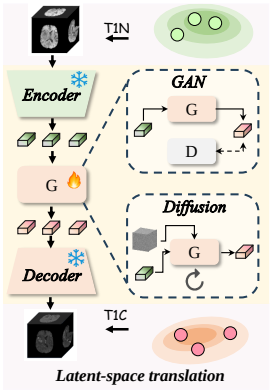

Fig. 4: Latent cross-contrast synthesis.

the anatomical attention map A_a captures morphological discontinuities using volumetric spatial gradients:

$$A_a = |\nabla_x x| + |\nabla_y x| + |\nabla_z x|, \tag{5}$$

where ∇_x, ∇_y, and ∇_z denote the first-order partial derivatives along the three spatial axes. Essentially, A_a acts as a local edge detector to explicitly highlight fine structural boundaries and subtle textures. Second, the semantic attention map A_s incorporates high-level structural relevance by aggregating and upsampling the final-layer self-attention responses from the pre-trained teacher network described in Section 2.3. This provides a global semantic filter, ensuring the high-frequency preservation prioritizes clinically significant tissues while suppressing irrelevant background artifacts.

Concurrently, we isolate the high-frequency residual $\mathcal{H}(x)$ of the input volume by subtracting its low-pass filtered representation:

$$\mathcal{H}(x) = x - \mathcal{S}_v(x), \tag{6}$$

where $\mathcal{S}_v(\cdot)$ denotes a local volumetric smoothing operator. Finally, the AFL objective computes the L_1 distance between the attention-weighted high-frequency components of the reconstruction $\hat{x}$ and the target x:

$$\mathcal{L}_{\text{afl}} = \left\| A_j \odot \mathcal{H}(\hat{x}) - A_j \odot \mathcal{H}(x) \right\|_1. \tag{7}$$

By restricting the high-frequency penalty specifically to anatomically and semantically significant regions, AFL effectively mitigates over-smoothing while maintaining robustness.

2.5 Training Objective

The final training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{srb}} + \mathcal{L}_{\text{afl}}. \tag{8}$$

Table 1: Quantitative comparison of reconstruction performance.

Method	BraTS Avg			IXI Avg		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
VQVAE [27]	27.43	0.8615	0.1342	28.61	0.8804	0.1128
VQGAN [8]	29.84	0.8953	0.0816	31.06	0.9127	0.0705
FSQ [1]	32.39	0.9244	**0.0557**	33.18	0.9305	0.0514
Ours	**32.80**	**0.9281**	0.0578	**33.65**	**0.9377**	**0.0450**

Table 2: Quantitative comparison on BraTS for two translation tasks.

Method	T2 → FLAIR			T1 → T1C		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
3D CycleGAN [33] (Pixel)	26.45	0.8751	0.1423	25.12	0.8224	0.1985
Latent CycleGAN [33] (Baseline)	26.61	0.8669	0.1342	26.93	0.8937	0.1272
Latent Diffusion [24] (Baseline)	25.17	0.7721	0.2276	25.44	0.7706	0.2478
Latent CycleGAN (Ours)	**26.65**	**0.8769**	**0.1229**	**30.41**	**0.9252**	**0.0851**
Latent Diffusion (Ours)	**27.95**	**0.7831**	**0.1623**	**26.55**	**0.7789**	**0.1874**

Here, $\mathcal{L}_{\text{rec}}$ is the baseline reconstruction loss consistent with 3D FSQ [22,1], while $\mathcal{L}_{\text{srb}}$ and $\mathcal{L}_{\text{afl}}$ are designed to enforce semantic consistency and maintain diagnostically relevant high-frequency structures, respectively.

3 Experiments

3.1 Datasets

We evaluate our framework on two public multi-contrast brain MRI datasets: **BraTS** [3,20], comprising T1-weighted (T1), post-contrast T1 (T1C), native T2-weighted (T2), and T2 fluid attenuated inversion recovery (FLAIR) sequences; and the **IXI** dataset [13], comprising healthy T1, T2, and PD sequences. For both datasets, all 3D volumes are resampled to a uniform spatial resolution. We use their respective official data splits for training, validation, and testing.

3.2 Implementation Details

Our framework is trained in two stages, both conducted on 8 RTX4090 GPUs. For latent space construction, we employ 3D FSQ [22,1] as the baseline, with E_{c} and the decoder adopting its structure. To evaluate the effectiveness of our proposed framework, we conduct comprehensive comparisons on two primary tasks: 3D MRI reconstruction and cross-contrast synthesis. We use PSNR, SSIM, and LPIPS to quantitatively assess both structural fidelity and perceptual quality.

Stage 1: 3D MRI Reconstruction. In the first stage, we optimize the model with the reconstruction objective in Sec. 2.5 to learn compact latent representations. We use the Adam optimizer with an initial learning rate of 1×10^{-4}

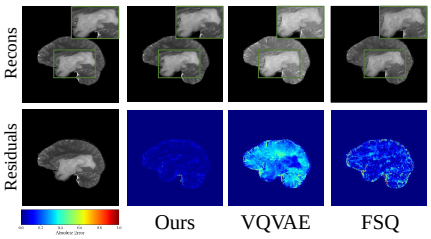

Fig. 5: Visual comparison of 3D MRI reconstruction performance.

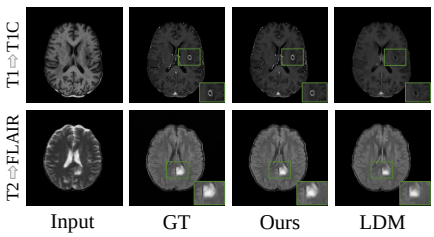

Fig. 6: Visual comparison of cross-contrast synthesis results.

and a total batch size of 2, and train the model for 500 epochs. After convergence, the modules responsible for compression and reconstruction (i.e., the encoder/decoder and quantization components) are frozen. To evaluate reconstruction fidelity, we compare our structurally enhanced latent space against representative volumetric compression methods, including VQ-VAE [27], VQ-GAN [8], and vanilla FSQ [1].

Stage 2: Cross-Contrast Synthesis in Latent Space. In the second stage, we perform latent-space translation for contrast-to-contrast synthesis (specifically, T1 → T1C and T2 → FLAIR), as illustrated in Fig. 4. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32, and train the translation model for 200 epochs. Additionally, we employ an L_1 loss as the translation objective. Importantly, our goal in this stage is not to propose a novel generative architecture, but rather to demonstrate that a well-structured, semantically separable latent space can significantly boost the performance of existing generative models. To this end, we select CycleGAN [33] and Latent Diffusion Models [24] as our GAN and diffusion-based comparison methods.

3.3 Reconstruction and Translation Results

Our method achieves state-of-the-art performance across both reconstruction and cross-contrast translation tasks. As summarized in Table 1, it improves PSNR by 0.41 dB over the strongest baseline on BraTS, alongside reaching 33.65 PSNR and the lowest LPIPS of 0.0450 on IXI. Furthermore, integrating this space into GAN and diffusion frameworks consistently outperforms existing baselines in Table 2; notably, it yields a PSNR increase of 3.48 dB for Latent CycleGAN (T1→T1C) and 2.78 dB for Latent Diffusion (T2→FLAIR), proving that these gains stem from the semantically aligned latent representation. Qualitative results in Fig. 5 and Fig. 6 further illustrate our model's ability to recover fine-grained details and maintain structural consistency across contrasts.

Table 3: Ablation study of the proposed modules.

LHE	SRB	AFL	PSNR ↑				PSNR ↑	SSIM ↑	LPIPS ↓
			T1	T2	FLAIR	T1C	Avg	Avg	Avg
×	×	×	31.60	30.38	32.88	34.69	32.39	0.9234	**0.0557**
✓	×	×	31.98	31.05	32.86	34.71	32.65	0.9256	0.0558
✓	✓	×	31.78	30.91	32.73	34.45	32.47	0.9253	0.0563
✓	✓	✓	**32.30**	**31.20**	**33.01**	**34.69**	**32.80**	**0.9281**	0.0578

Table 4: Ablation study on semantic teachers.

Encoder	PSNR ↑				PSNR ↑	SSIM ↑	LPIPS ↓
	T1	T2	FLAIR	T1C	Avg	Avg	Avg
ResNet-50	31.92	31.06	32.79	34.47	32.56	0.9268	0.0594
MAE (ViT-B)	31.75	31.00	32.76	34.33	32.46	0.9261	0.0595
DINO (ViT-B/16)	**32.30**	**31.20**	**33.01**	**34.69**	**32.80**	**0.9281**	**0.0578**

3.4 Ablation Studies

Module Analysis As shown in Table 3, our progressive ablation study on BraTS evaluates LHE, SRB, and AFL against a vanilla 3D FSQ baseline [1]. Adding LHE improves average PSNR (32.39 to 32.65) by capturing long-range anatomical dependencies. Introducing SRB slightly trades pixel-level reconstruction fidelity for improved semantic organization. Finally, integrating AFL achieves the peak performance of 32.80 dB PSNR and 0.9281 SSIM. Overall, these complementary modules combine to yield the strongest reconstruction results.

Effect of Semantic Teachers. We adopt a DINO-pretrained ViT-B/16 [5] as the semantic teacher due to its strong global representation capabilities [2,25,30]. To validate this design choice, we compare it against pre-trained ResNet-50 [12] and MAE (ViT-B) encoders [11] under identical settings on the BraTS dataset. As shown in Table 4, the DINO-based teacher achieves the best performance, obtaining the highest average PSNR (32.80) and SSIM (0.9281) alongside the lowest LPIPS (0.0578). These results confirm that DINO provides more robust and semantically aligned representations. In future work, we will explore whether VLMs/LLMs [32] can provide stronger semantic supervision.

4 Conclusion

In this paper, we propose a semantics-first latent modeling framework for 3D MRI reconstruction and cross-contrast synthesis. Our method improves the structural coherence and semantic organization of compressed latents before generative modeling. To this end, we introduce a Latent Harmonization Encoder (LHE) to capture long-range anatomical dependencies, a Semantic Recovery Block (SRB) to enhance contrast-specific separability, and an Anatomy-aware Frequency Loss (AFL) to preserve diagnostically important details. This design reduces structural inconsistency and semantic entanglement in conventional

VAE/VQ-based pipelines. Experiments on two public datasets demonstrate consistent improvements over volumetric compression and latent translation baselines, with ablations confirming the contribution of each component.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Project No.82572383) and the Guangdong Science and Technology Department (2024ZDZX2004).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., Peng, W.: Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. arXiv preprint arXiv:2312.02366 (2023)
3. 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)
4. Brown, R.W., Cheng, Y.C.N., Haacke, E.M., Thompson, M.R., Venkatesan, R.: Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons (2014)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
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. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
9. 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)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain mri segmentation combining label propagation and decision fusion. NeuroImage **33**(1), 115–126 (2006)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
15. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13919–13929 (2021)
16. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical image analysis **69**, 101950 (2021)
17. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., et al.: Denoising diffusion probabilistic models for 3d medical image generation. Scientific reports **13**(1), 7303 (2023)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
19. Kui, X., Fan, Z., Ji, Z., Li, Q., Liu, C., Si, W., Zou, B.: A comprehensive survey on magnetic resonance image reconstruction. Image and Vision Computing p. 105832 (2025)
20. LaBella, D., Baid, U., Khanna, O., McBurney-Lin, S., McLean, R., Nedelec, P., Rashid, A., Tahon, N.H., Altes, T., Bhalerao, R., et al.: Analysis of the brats 2023 intracranial meningioma segmentation challenge. arXiv preprint arXiv:2405.09787 (2024)
21. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
22. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple. arXiv preprint arXiv:2309.15505 (2023)
23. Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D.: Estimating ct image from mri data using 3d fully convolutional networks. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 170–178. Springer (2016)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
25. Song, X., Xu, X., Yan, P.: Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 608–617. Springer (2024)
26. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. Advances in neural information processing systems **33**, 18158–18172 (2020)
27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

28. Wang, Y., Li, Z., Mei, J., Wei, Z., Liu, L., Wang, C., Sang, S., Yuille, A.L., Xie, C., Zhou, Y.: Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 486–496. Springer (2023)
29. Yang, S., Hu, X., Wu, Q., Yang, D.: Vaevq: Enhancing discrete visual tokenization through variational modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11703–11711 (2026)
30. Yang, S., Wang, H., Xing, Z., Chen, S., Zhu, L.: Segdino: An efficient design for medical and natural image segmentation with dino-v3. arXiv preprint arXiv:2509.00833 (2025)
31. Yang, S., Xing, Z., Zhu, L.: Vq-seg: Vector-quantized token perturbation for semi-supervised medical image segmentation. arXiv preprint arXiv:2601.10124 (2026)
32. Yang, S., Zhou, H., Yang, Y., Wang, W., Chen, S., Yang, G., Fu, H., Zhu, L.: Lcm-net: Llm-driven cross-modality moe feature fusion network for cancer survival analysis. IEEE Transactions on Medical Imaging (2026)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)