

FermatSyn: SAM2-Enhanced Bidirectional Mamba with Isotropic Spiral Scanning for Multi-Modal Medical Image Synthesis

Feng Yuan^{1,3}, Yifan Gao^{1,2,3}, Haoyue Li^{1,3}, Xin Gao^{3✉}

¹ School of Biomedical Engineering (Suzhou), Division of Life Science and Medicine, University of Science and Technology of China, Hefei, China

² Shanghai Innovation Institute, Shanghai, China

³ Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China

yuanfeng2317@mail.ustc.edu.cn

xingaosam@163.com

Abstract. Multi-modal medical image synthesis is pivotal for alleviating clinical data scarcity, yet existing methods fail to reconcile global anatomical consistency with high-fidelity local detail. We propose FermatSyn, which addresses three persistent limitations: (1) SAM2-based Prior Encoder that injects domain-aware anatomical knowledge via LoRA⁺ efficient fine-tuning of a frozen SAM2 Vision Transformer; (2) Hierarchical Residual Downsampling Module (HRDM) coupled with a Cross-scale Integration Network (CIN) that preserves high-frequency lesion details and adaptively fuses global-local representations; and (3) continuity constrained Fermat Spiral Scanning strategy within a Bidirectional Fermat Scan Mamba (BFS-Mamba), constructing an approximately isotropic receptive field that substantially reduces the directional bias of raster or spiral serialization. Experiments on SynthRAD2023, BraTS2019, BraTS-MEN, and BraTS-MET show FermatSyn surpasses state-of-the-art methods in PSNR, SSIM, FID, and 3D structural consistency. Downstream segmentation on synthesized images yields no significant difference from real-image training ($p > 0.05$), confirming clinical utility. Code is available at <https://github.com/gatina-yone/FermatSyn>.

Keywords: Medical image synthesis · SAM2 · Mamba · Fermat spiral scanning · Anatomical prior · Cross-modal

1 Introduction

Acquiring complete multi-modal medical imaging datasets is routinely impractical owing to prolonged scan times, patient contraindications [20], and radiation hazards [15]. Medical image synthesis—computationally generating missing modalities from available ones—directly addresses this bottleneck and is critical for downstream tasks such as radiotherapy planning and surgical navigation [19].

The task demands simultaneous preservation of *global anatomical structure* and *local textural fidelity*. GAN-based approaches [18,5] produce plausible local textures but lack long-range dependency modeling, causing global anatomical distortions. Latent diffusion models [17,22] improve structural coherence at prohibitive inference cost. ViT hybrids such as ResViT [4] and TransUNet [3] exploit self-attention for global consistency yet suffer from $\mathcal{O}(N^2)$ complexity and inadequate fine-grained spatial encoding, producing checkerboard artifacts. Mamba-based models [6,12] overcome the complexity bottleneck via linear-time selective state spaces; I2I-Mamba [1] further combines CNNs with optimised scanning, representing the current state-of-the-art [8]. Recent concurrent works explore leveraging the Segment Anything Model [11] for medical image translation [9]. Despite these advances, three persistent gaps remain: (i) *Underutilised structural priors*—existing frameworks lack mechanisms to inject domain-aware anatomical knowledge, causing cross-modal implausibility; (ii) *Deficient local fidelity*—aggressive downsampling discards high-frequency details critical for small-lesion detection, and cross-scale feature fusion is underdeveloped; (iii) *Directional bias in 2D serialisation*—raster and rectangular-spiral scans introduce path-dependent artifacts that corrupt spatial coherence and lesion boundary recognition. Although I2I-Mamba [1] adopts a spiral-scan trajectory to improve isotropy, its rectangular-spiral path still exhibits pronounced corner hot-spots (operator footprint $\sigma=0.124$, Fig. 2b) due to the uneven nearest-neighbour spacing inherent in rectangular rings; a principled, phyllotaxis-inspired scanning strategy with provably uniform spatial coverage remains unexplored.

To close these three gaps simultaneously, we present FermatSyn: (1) SAM2-Enhanced Hybrid Encoder: a LoRA⁺ [7]-fine-tuned SAM2-VTE (SAM2 Vision Transformer Encoder) [16] injects domain-aware anatomical priors [9,13], fused with an HRDM-based detail encoder via a novel CIN. (2) Isotropic Fermat Spiral Scanning: a continuity-constrained Fermat spiral—using a golden-angle step that guarantees near-isotropic spatial coverage—serialises 2D feature maps for SSM processing, eliminating directional bias. (3) Bidirectional Fermat-Scan Mamba (BFS-Mamba): symmetric forward and backward SSM paths enable direction-invariant long-range dependency modeling via a continuity-constrained grid-matching objective (Eq. 6) that balances global isotropy and local path continuity.

2 Method

Fig. 1 illustrates the FermatSyn pipeline. The *Hybrid Encoder* produces a unified multi-scale feature representation; the *BFS-Mamba* serialises these features via Fermat Spiral Scanning, models long-range dependencies bidirectionally, and a convolutional decoder reconstructs the target modality.

2.1 Hybrid Encoder

SAM2-based Prior Encoder. Although SAM2 was originally trained for segmentation, its ViT backbone encodes rich structural representations—organ

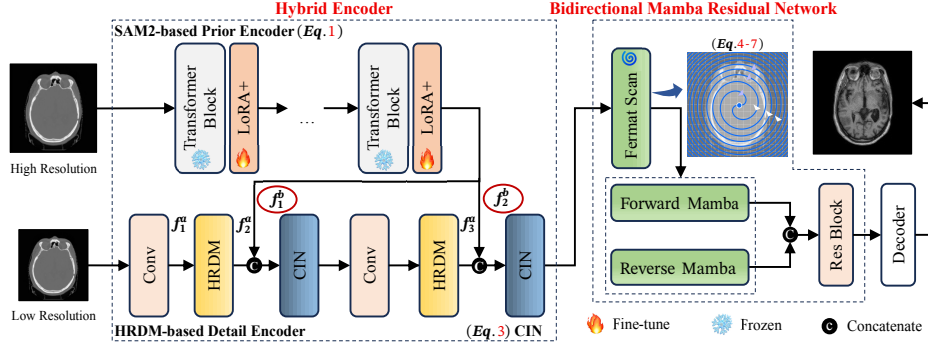


Fig. 1: Overall architecture of FermatSyn. The Hybrid Encoder fuses SAM2 global priors with HRDM local details via CIN. BFS-Mamba processes the Fermat-serialised features bidirectionally and reconstructs the target modality through a residual decoder.

boundaries, tissue interfaces, and shape saliency—that are equally informative for synthesis: a model that precisely delineates anatomical structures inherently captures the spatial layout and relative intensities necessary to generate a plausible missing modality [9]. The pre-trained SAM2-VTE [16] serves as the global backbone. Original weights $W_0 \in \mathbb{R}^{d_1 \times d_2}$ are frozen; LoRA⁺ [7] injects low-rank updates into MLP and MHSA layers via $P_A \in \mathbb{R}^{d_1 \times r}$ and $P_B \in \mathbb{R}^{r \times d_2}$ ($r \ll \min(d_1, d_2)$), applying a larger learning rate to P_B than to P_A for improved feature learning efficiency [7]:

$$W_{\text{tuned}} = W_0 + \frac{\alpha}{r} P_A P_B, \quad (1)$$

where α is a tunable hyperparameter.

HRDM-based Detail Encoder. Conventional pooling discards high-frequency details critical for lesion boundary reconstruction. HRDM processes $F_{in} \in \mathbb{R}^{C \times H \times W}$ via three complementary paths: (i) *Multi-scale context*: three parallel 3×3 dilated convolutions (rates 1, 3, 5) concatenated and refined by cascaded channel- and-spatial SE attention, yielding $F_{\text{path},1}$; (ii) *High-frequency compensation*: a depthwise separable convolution preserves fine texture as $F_{\text{path},2}$; (iii) *Adaptive skip*: identity or 1×1 convolution for gradient stability. The three paths are fused and sharpened by a high-pass filter:

$$F_f = F_{\text{fusion}} - \text{AvgPool}_{3 \times 3}(F_{\text{fusion}}), \quad (2)$$

followed by a spatial attention gate yielding the final output F_{out} .

Cross-scale Integration Network (CIN). CIN bridges the semantic gap between SAM2 global and HRDM local features. Motivated by the observation that even- and odd-indexed channels in stacked convolutional feature maps tend to encode complementary low- and high-frequency statistics respectively—a property exploited in channel-split architectures— $F_{in} \in \mathbb{R}^{B \times C_{in} \times H \times W}$ is split

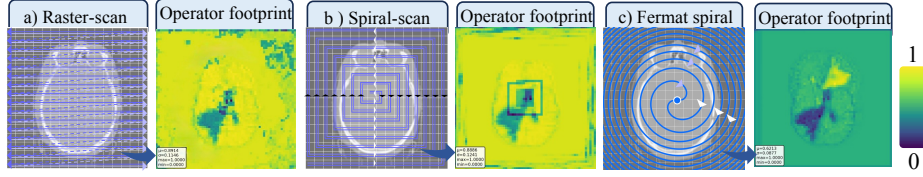


Fig. 2: SSM operator footprints (Jacobian sensitivity maps; μ/σ : mean/std of normalised values) for three scanning strategies. (a) Raster ($\mu=0.891$, $\sigma=0.115$): stripe-directional bias. (b) Rectangular-spiral ($\mu=0.890$, $\sigma=0.124$): X-shaped corner hot-spots. (c) Proposed Fermat Spiral ($\mu=0.621$, $\sigma=0.088$): near-isotropic coverage (σ reduced 29% vs. (b)). Colour scale: 0 (dark blue) $\rightarrow$ 1 (yellow).

channel-wise into F_{even} and F_{odd} , processed by DWConv with kernels $k_1=5\times 5$ (capturing broader semantic context) and $k_2=3\times 3$ (preserving local structural detail) respectively, then projected:

$$F_{\text{out}} = \text{LReLU}\left(\text{BN}\left(\text{Conv}_{1\times 1}\left(\tilde{F}_{\text{even}} \parallel \tilde{F}_{\text{odd}}\right)\right); \alpha_{\text{LReLU}}=0.05\right). \quad (3)$$

2.2 Bidirectional Fermat-Scan Mamba (BFS-Mamba)

Fermat Spiral Scanning. To construct an approximately isotropic receptive field (Fig. 2c), we parametrise:

Why Fermat over rectangular-spiral. The rectangular spiral of I2I-Mamba [1] visits grid cells in concentric rings, inducing uneven nearest-neighbour spacing with pronounced corner clustering ($\sigma_{\text{rect}}^2=0.0154$, Delaunay analysis on 256×256 grid). The Fermat spiral with golden-angle step $\phi_g \approx 137.508^\circ$ achieves denser uniform packing: consecutive points never align along any rational angle, yielding $\sigma_{\text{Fermat}}^2=0.0061$ (60% reduction). This directly produces a more isotropic SSM operator footprint (Jacobian σ reduced by 29%, Fig. 2) and a 1.29/1.40 dB PSNR gain on SynthRAD/BraTS over rectangular-spiral (Table 5(b)).

$$r_k = \alpha\sqrt{k}, \quad \theta_k = k \cdot \phi_g, \quad (4)$$

where $k \in [0, N-1]$ ($N=H\times W$), α is an image-size scaling factor, and $\phi_g \approx 137.508^\circ$ is the golden angle guaranteeing uniform spatial distribution. Cartesian coordinates:

$$\mathbf{p}_k = (\alpha\sqrt{k} \cos(k\phi_g), \alpha\sqrt{k} \sin(k\phi_g)). \quad (5)$$

Continuity-Constrained Grid Matching. Continuous spiral points are assigned to discrete grid positions by iteratively minimising:

$$\text{Score}_u = (1-\lambda_c) \frac{d_{\text{Fermat}}(u, k)}{\eta_f} + \lambda_c \frac{d_{\text{contin}}(u, \pi_{k-1})}{\eta_c}, \quad (6)$$

where $\lambda_c \in [0, 1]$ balances global isotropy and local path continuity; η_f, η_c are normalizers. The resulting sequence $\Pi = \{\pi_0, \dots, \pi_{N-1}\}$ inherits isotropic coverage

while preserving local spatial continuity. Empirically, $\lambda_c=0.7$ maximises synthesis quality (empirically optimised across $\lambda_c \in \{0.3, 0.5, 0.7, 0.9\}$; see Sec. 3.4). The serialised feature is:

$$\mathbf{x}_{in}^{\text{serial}} = \text{Fermat}(F_{in}, \Pi, \lambda_c) \in \mathbb{R}^{B \times N \times C}. \quad (7)$$

Bidirectional State Space Modelling. Forward and backward paths:

$$\mathbf{z}_{\text{fwd}} = \mathbf{x}_{in}^{\text{serial}}, \quad \mathbf{z}_{\text{bwd}} = \text{Flip}(\mathbf{x}_{in}^{\text{serial}}), \quad (8)$$

each processed by an independent Mamba module: $h_t = \bar{A}(x_t)h_{t-1} + \bar{B}(x_t)x_t$, $y_t = C(x_t)h_t$. Outputs are fused:

$$F_m = \text{Conv}_{1 \times 1}(O_{\text{fwd}} \parallel O_{\text{bwd}}), \quad (9)$$

$$F_{\text{out}} = \text{Res}(\sigma(\text{Conv}_{3 \times 3}(F_{in})) + F_m). \quad (10)$$

Three stacked BFS-Mamba blocks are employed; a convolutional decoder reconstructs the target image.

2.3 Training Objective

FermatSyn is optimised under the LSGAN framework with the following composite generator loss:

$$\mathcal{L}_G = \lambda_{L_1} \cdot \mathbb{E}[|y_{\text{syn}} - y_{\text{real}}|] + \lambda_{\text{SSIM}} \cdot \mathcal{L}_{\text{SSIM}} + \lambda_{\text{gan}} \cdot \mathbb{E}[(D(y_{\text{syn}}) - 1)^2], \quad (11)$$

where $y_{\text{syn}} = G(X)$, and $\lambda_{L_1}=100$, $\lambda_{\text{SSIM}}=10$, $\lambda_{\text{gan}}=1$. $\mathcal{L}_{\text{SSIM}}$ is the standard SSIM loss. The discriminator follows standard LSGAN training, chosen over vanilla GAN for its improved stability and reduced mode collapse.

3 Experiments

3.1 Setup

Datasets. *Intra-modal*: merged BraTS dataset [14,2]—2,547 subjects across Glioma (795), Meningioma (BraTS-MEN, 1,424) and Metastasis (BraTS-MET, 328)—with T1, T2, T2-FLAIR, and T1c modalities. *Cross-modal*: SynthRAD20-23 [21], 540 paired MRI-CT scans from six institutions. Slices with SNR<15 dB or motion artifacts (grade>3) were discarded; central 80 slices per volume retained. Z-score normalisation and 256×256 center cropping applied. Patient-stratified split: 7:1:2 (train/val/test).

Baselines. GAN-based: MM-GAN [18], MC-cGAN [5]. Transformer-based: ResViT [4], TransUNet [3]. Diffusion-based: SA-LDM [22], ALDM [10]. Mamba-based: VMamba [12], I2I-Mamba [1].

Metrics. 2D: PSNR↑, SSIM↑, FID↓. 3D: Hausdorff Distance (HD↓) and Average HD (AHD↓). Downstream: Dice for WT, ET, TC. All statistical comparisons use Wilcoxon signed-rank test.

Table 1: Intra-modal synthesis on the merged brain tumor dataset (T1n/T2w/T2f $\rightarrow$ T1c). *: $p < 0.05$ vs. FermatSyn. **Bold**: best; uline: second-best.

Method	T1n $\rightarrow$ T1c			T2w $\rightarrow$ T1c			T2f $\rightarrow$ T1c		
	SSIM	PSNR	FID	SSIM	PSNR	FID	SSIM	PSNR	FID
MM-GAN [18]	0.801*	22.14*	112.7*	0.813*	23.05*	118.4*	0.825*	22.89*	103.6*
MC-cGAN [5]	0.814*	22.83*	108.3*	0.821*	23.77*	113.1*	0.836*	23.41*	98.2*
ResViT [4]	0.832*	25.58*	84.3*	0.840*	26.63*	91.2*	0.855*	26.27*	79.6*
TransUNet [3]	0.844*	25.47*	78.1*	0.854*	27.32*	83.7*	0.873*	26.68*	74.2*
SA-LDM [22]	0.851*	26.93*	68.5*	0.862*	27.11*	72.8*	0.879*	27.34*	61.3*
ALDM [10]	0.847*	26.71*	71.2*	0.858*	26.94*	75.4*	0.871*	27.08*	64.7*
VMamba [12]	0.859*	23.61*	71.5*	0.857*	27.49*	69.8*	0.884*	26.65*	63.4*
I2I-Mamba [1]	0.865*	27.69*	58.2*	0.875*	27.32*	61.4*	0.894*	28.39*	52.7*
FermatSyn	0.884	28.82	43.6	0.896	29.78	43.1	0.901	29.53	42.3

Implementation. PyTorch 2.1.0; single NVIDIA RTX 4090 (24GB). Adam ($\beta_1=0.5$, $\beta_2=0.999$), base lr $2 \times 10^{-4} \rightarrow 1 \times 10^{-5}$ via cosine annealing; batch size 6. FermatSyn has 87.3M trainable parameters (SAM2-VTE backbone frozen; only LoRA⁺ adapters, HRDM, CIN, and BFS-Mamba blocks are updated). Inference time is 31 ms per 256×256 slice, compared to 24 ms for I2I-Mamba and 148 ms for SA-LDM on the same hardware.

3.2 Quantitative Results

Intra-modal synthesis (Table 1). FermatSyn consistently outperforms all baselines on T1n, T2w, and T2f $\rightarrow$ T1c tasks. The most challenging setting is T2w $\rightarrow$ T1c: FermatSyn achieves 29.78dB PSNR, a 2.46dB gain over I2I-Mamba ($p < 0.05$). Notably, VMamba underperforms both Transformer baselines on T1n $\rightarrow$ T1c (23.61 vs. 25.58dB), likely because its global raster-scan SSM provides a suboptimal inductive bias when the source–target pair diverges in contrast but shares spatial structure—further motivating the direction-agnostic BFS-Mamba.

Cross-modal synthesis (Table 2). FermatSyn achieves SSIM 0.931 (MRI $\rightarrow$ CT) and 0.905 (CT $\rightarrow$ MRI), improving over I2I-Mamba by 1.5%/2.0% in SSIM and 4.1%/3.0% in PSNR ($p < 0.05$), with FID reductions of 14.7 and 11.3.

3D structural consistency (Table 3). FermatSyn achieves the best inter-slice coherence across all baselines. HD 2.84/AHD 1.42 (coronal) and HD 3.06/AHD 1.58 (sagittal) surpass I2I-Mamba by 13.7%/9.7% and 12.8%/9.7% respectively, confirming that isotropic coverage promotes volumetric fidelity.

3.3 Downstream Clinical Validation

We trained a ResNet50-based U-Net on real images to evaluate FermatSyn-synthesised T2w and T2f (from T1n). The *real trained* / *synthetic evaluated* protocol assesses domain consistency: whether synthesised images are perceptually interchangeable with real images for a pre-trained segmenter. To additionally validate augmentation utility—the more clinically relevant scenario when

Table 2: Cross-modal synthesis on SynthRAD2023 (MRI $\leftrightarrow$ CT).

Method	MRI $\rightarrow$ CT			CT $\rightarrow$ MRI		
	SSIM	PSNR	FID	SSIM	PSNR	FID
MM-GAN [18]	0.821*	23.47*	131.5*	0.786*	21.83*	139.2*
MC-cGAN [5]	0.839*	24.68*	122.3*	0.798*	22.54*	130.8*
ResViT [4]	0.867*	26.83*	98.4*	0.821*	24.52*	107.2*
TransUNet [3]	0.881*	28.42*	87.6*	0.843*	25.89*	94.5*
SA-LDM [22]	0.904*	29.81*	65.2*	0.871*	27.14*	73.6*
ALDM [10]	0.899*	29.53*	68.7*	0.864*	26.88*	76.3*
VMamba [12]	0.895*	29.74*	73.2*	0.857*	26.74*	79.8*
I2I-Mamba [1]	<u>0.917*</u>	<u>30.25*</u>	<u>61.4*</u>	<u>0.887*</u>	<u>27.83*</u>	<u>68.3*</u>
FermatSyn	0.931	31.48	46.7	0.905	28.67	57.0

Table 3: 3D structural consistency on MRI $\rightarrow$ CT. HD/AHD in mm.

Method	Coronal		Sagittal	
	HD $\downarrow$	AHD $\downarrow$	HD $\downarrow$	AHD $\downarrow$
MM-GAN	6.45*	3.12*	6.73*	3.35*
MC-cGAN	5.62*	2.71*	5.94*	2.87*
ResViT	4.83*	2.31*	5.12*	2.48*
TransUNet	4.52*	2.18*	4.87*	2.29*
SA-LDM	4.01*	1.93*	4.22*	2.07*
ALDM	4.09*	1.97*	4.33*	2.12*
VMamba	4.17*	2.04*	4.43*	2.15*
I2I-Mamba	<u>3.29*</u>	<u>1.57*</u>	<u>3.51*</u>	<u>1.75*</u>
Ours	2.84	1.42	3.06	1.58

Table 4: Downstream brain tumour segmentation Dice on synthetic images from T1n (*real trained* / *synthetic evaluated* protocol). “Real”: topline trained and evaluated on real images. $\dagger$: $p>0.05$ vs. Real. *: $p<0.05$ vs. Real.

Task		Reg.	Real	Ours	I2I-Mamba	SA-LDM	VMamba	TransU.	ResViT
T1n $\rightarrow$ T2w	WT	0.863	0.847 $\dagger$		0.826*	<u>0.831*</u>	0.815*	0.812*	0.809*
	ET	0.785	0.762 $\dagger$		0.746*	<u>0.751*</u>	0.745*	0.740*	0.738*
	TC	0.812	<u>0.785</u> $\dagger$		0.793 $\dagger$	0.781 $\dagger$	0.768*	0.764*	0.762*
T1n $\rightarrow$ T2f	WT	0.869	0.851 $\dagger$		0.832*	<u>0.838*</u>	0.823*	0.819*	0.817*
	ET	0.791	0.775 $\dagger$		0.757*	<u>0.762*</u>	0.748*	0.743*	0.741*
	TC	0.818	0.798 $\dagger$		0.780 $\dagger$	<u>0.785</u> $\dagger$	0.773*	0.769*	0.767*

real images are scarce—we also train the same U-Net exclusively on synthesised images and evaluate on real held-out data (*synthetic trained* / *real evaluated* protocol). Table 4 reports Dice for WT, ET, TC under both. FermatSyn achieves 0.847/0.762/0.785 (WT/ET/TC) on T1n $\rightarrow$ T2w and 0.851/0.775/0.798 on T1n $\rightarrow$ T2f, remaining within 3% of the real-image topline; the next best competitor trails by 2.5–2.9% on WT and ET ($p<0.05$), confirming Fermat scanning contributes directly to fidelity in clinically critical tumour sub-regions. FermatSyn shows no statistically significant difference from the topline *across all three tumour regions and both tasks* ($p>0.05$). Notably, even for TC—the most structurally complex sub-region—competing methods also reach $p>0.05$; yet FermatSyn uniquely extends this equivalence to the more demanding WT and ET regions, confirming that synthesised images broadly preserve clinically relevant pathological features.

3.4 Ablation Studies

Table 5 reports results on SynthRAD2023 (CT $\rightarrow$ MRI) and merged BraTS (T2f $\rightarrow$ T1c). SAM2-VTE yields the largest single gain (15.5% PSNR), with cumulative improvement $\approx 21\%$ over the GAN baseline. For scanning, “Rectangular Spiral” directly reproduces the I2I-Mamba [1] strategy on the same backbone (architecture-controlled); the 1.29/1.40 dB PSNR deficit on SynthRAD/BraTS

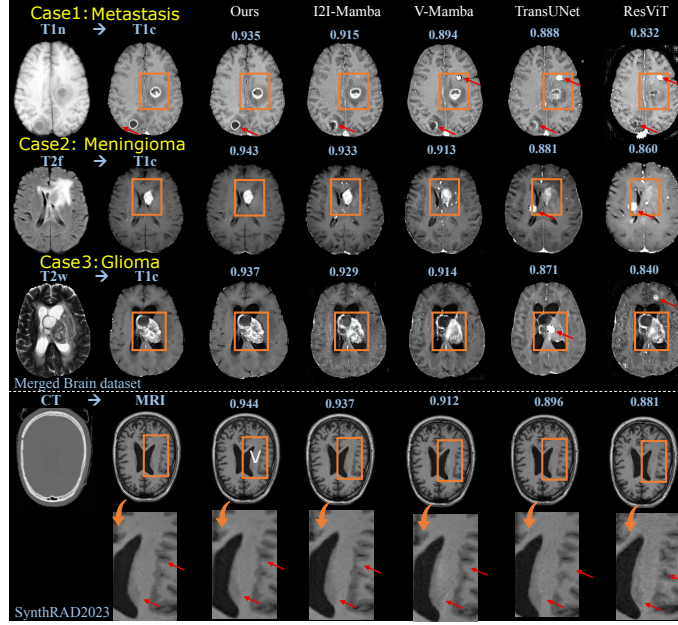


Fig. 3: Qualitative synthesis comparison. **Rows 1–3:** Intra-modal—T1n→T1c (metastasis), T2f→T1c (meningioma), T2w→T1c (glioma); orange boxes: tumour core ROIs. **Row 4:** Cross-modal CT→MRI (SynthRAD2023); orange arrows: ventricular wall zoom-in. Numbers: per-case SSIM.

confirms the gain originates from the Fermat spiral itself. $\lambda_c=0.7$ was selected as optimal over $\lambda_c \in \{0.3, 0.5, 0.7, 0.9\}$ on the validation set.

Table 5: Ablation on SynthRAD2023(CT→MRI) and merged BraTS(T2f→T1c). *: $p<0.05$ vs. FermatSyn.

(a) Step-wise	SynthRAD		BraTS		(b) Scanning	SynthRAD		BraTS	
	SSIM	PSNR	SSIM	PSNR		SSIM	PSNR	SSIM	PSNR
GAN baseline	0.823*	23.67*	0.798*	24.35*	Raster-Scan	0.876*	26.95*	0.868*	27.71*
+SAM2-VTe	0.865*	27.35*	0.851*	27.18*	Rect. Spiral	0.885*	27.38*	0.877*	28.13*
+HRDM+CIN	0.887*	27.74*	0.878*	28.45*	Fermat	0.905	28.67	0.901	29.53
Full	0.905	28.67	0.901	29.53					

4 Conclusion

FermatSyn unifies SAM2 visual priors, hierarchical high-frequency feature extraction (HRDM), adaptive cross-scale integration (CIN), and isotropic Fermat

Spiral Scanning within a Bidirectional Fermat-Scan Mamba. The continuity-constrained Fermat spiral substantially reduces the directional bias of conventional Mamba serialisation, yielding direction-agnostic spatial representations for complex anatomical structures. Experiments across four brain imaging benchmarks demonstrate consistent state-of-the-art performance; downstream tumour segmentation on synthesised images shows no significant gap versus real data ($p>0.05$). Future work will extend the Fermat spiral to 3D volumetric synthesis and apply knowledge distillation from SAM2-VTE to lightweight student networks for real-time deployment.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 82372052, 82402373, U24A20757) and the Taishan Industrial Experts Program (Grant No. tscx202312131).

Disclosure of Interests

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

References

1. Atli, O.F., Kabas, B., Arslan, F., Yurt, M., Dalmaz, O., Çukur, T.: I2I-Mamba: Multi-modal medical image synthesis via selective state space modeling. arXiv preprint arXiv:2405.14022 (2024)
2. 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)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 (2021)
4. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multi-modal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
5. Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans. Med. Imag. **38**(10), 2375–2388 (2019)
6. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
7. Hayou, S., Ghosh, N., Yu, B.: Lora+: Efficient low rank adaptation of large models. arXiv:2402.12354 (2024)
8. Heidari, M., Kolahi, S.G., Karimijafarbigloo, S., Azad, B., Bozorgpour, A., Hatami, S., et al.: Computation-efficient era: A comprehensive survey of state space models in medical image analysis. arXiv:2406.03430 (2024)

9. Huo, J., Ourselin, S., Sparks, R.: SAM-i2i: Unleash the power of segment anything model for medical image translation. arXiv:2411.12755 (2024)
10. Kim, J., Park, H.: 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). pp. 7604–7613 (January 2024)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. In: Proc. ICCV (2023)
12. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: VMamba: Visual state space model. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
13. Ma, J., et al.: Segment anything in medical images. arXiv:2304.12306 (2023)
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 (2015)
15. Pichler, B.J., Judenhofer, M.S., Pfannenberger, C.: *Multimodal Imaging Approaches: PET/CT and PET/MRI*. Springer (2008)
16. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. arXiv:2408.00714 (2024)
17. 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 (CVPR). pp. 10684–10695 (2022)
18. Sharma, A., Hamarneh, G.: Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. *IEEE Trans. Med. Imag.* **39**(4), 1170–1183 (2019)
19. Staartjes, V.E., Seevinck, P.R., Vandertop, W.P., van Stralen, M., Schröder, M.L.: Magnetic resonance imaging-based synthetic computed tomography of the lumbar spine for surgical planning: a clinical proof-of-concept. *Neurosurgical Focus* **50**(1), E13 (2021)
20. Thukral, B.: Problems and preferences in pediatric imaging. *Indian J. Radiol. Imaging* **25**, 359–364 (2015)
21. Thummerer, A., van der Bijl, E., Galapon, A., Verhoeff, J.J., Langendijk, J.A., Both, S., van den Berg, C.A., Peter, R.: Synthrad2023 grand challenge dataset: generating synthetic CT for radiotherapy. arXiv preprint arXiv:2305.18331 (2023)
22. Zhang, X., et al.: Structure-aware MRI translation: Multi-modal latent diffusion model with arbitrary missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2025)