FOURIER PRINCIPAL PROJECTED TRANSFORMER FOR EFFICIENT AND ACCURATE MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate medical image segmentation requires reliable modeling of global anatomical structure in high-resolution medical images. However, existing Transformer-based methods lack explicit global structural representation and suffer from quadratic attention complexity, limiting their effectiveness and efficiency in dense prediction settings. To address these limitations, we propose the Fourier Principal Projected Transformer (FPP-Former). Specifically, FPP-Former leverages the intrinsic global correlations of spatial tokens to derive principal component representations that characterize the major variations in global anatomical appearance, enabling explicit modeling of global anatomical organization. For efficient principal component modeling, we reformulate this process in the Fourier domain, where a low-dimensional principal subspace is constructed from dominant frequency components, avoiding explicit eigen-decomposition on high-dimensional feature representations. The resulting Fourier principal components are then fed into the attention model to guide key–value interactions, where keys and values are projected into a shared principal space for structure-guided attention computation and value aggregation, thereby focusing attention on spatially continuous and structurally coherent anatomical regions. This formulation results in linear-complexity attention computation, improving efficiency for high-resolution segmentation. Extensive experiments on BraTS 2017, Synapse, and ACDC show that FPP-Former achieves average Dice scores of 83.2%, 84.84%, and 91.94%, with relative improvements of 2.34%, 0.11%, and 0.12%, respectively, while consistently improving training and inference efficiency.

1 INTRODUCTION

Accurate medical image segmentation is essential for a wide range of clinical applications, including computer-aided diagnosis, radiotherapy planning, and surgical navigation. Compared with natural images, volumetric medical images exhibit complex anatomical structures, strong spatial continuity, and substantial inter-slice variation, requiring segmentation models to capture both long-range dependencies and fine-grained structural details. Transformer-based architectures have therefore become increasingly prominent in medical image segmentation due to their ability to model long-range contextual dependencies Vaswani et al. (2017); Dosovitskiy et al. (2021); Chen et al. (2024). Recent studies have further improved volumetric representation learning through hierarchical attention, local–global interaction, and efficient token modeling Hatamizadeh et al. (2022); Tang et al. (2022); Zhou et al. (2023); Shaker et al. (2024). Nevertheless, most of these approaches still rely on spatial-domain self-attention as the core mechanism for global interaction, where contextual dependencies are established through pairwise correlations among spatial tokens.

Despite substantial progress, this attention paradigm remains limited in both structural representation and computational efficiency for high-resolution segmentation. Standard self-attention constructs dense pairwise token correlations to capture global context, but such correlations contain considerable redundancy and do not explicitly distinguish dominant anatomical structure from noisy or less informative interactions. This formulation leads to two major limitations. *(i) Structural redundancy.* Pairwise feature similarity does not explicitly distinguish dominant anatomical organization from redundant or noisy token interactions. This issue is particularly pronounced in volumetric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

medical images, where anatomically coherent regions exhibit strong spatial continuity, while weak boundaries and adjacent structures can introduce ambiguous correlations. Although recent structure-aware Transformers incorporate explicit structural priors Chen et al. (2022); Kim et al. (2024), they typically require additional graph construction or relation modeling, increasing the computational burden. *(ii) Computational complexity.* Dense self-attention requires interactions among all spatial tokens, leading to quadratic complexity with respect to the token number. This cost becomes prohibitive for high-resolution volumes. A key observation is that volumetric representations are highly redundant, with most of their information concentrated in a small number of dominant components. This suggests that global anatomical structures can be effectively represented within a compact principal subspace while suppressing redundant variations. PCA provides a natural formulation to capture dominant variations and construct a compact representation by preserving the most informative directions while discarding redundant dimensions Jolliffe (2002); Wold et al. (1987); Abdi & Williams (2010). However, directly constructing such a principal subspace for dense volumetric representations requires expensive covariance computation and eigendecomposition. Therefore, an efficient alternative is needed to approximate the principal subspace without explicitly decomposing large volumetric correlation matrices.

Motivated by this observation, we propose **FPP**-Former, a **F**ourier **P**rincipal **P**rojected transformer for efficient and accurate medical image segmentation. FPP-Former reformulates dense token interaction as attention over a compact principal structural subspace. Specifically, keys and values are projected from the original token space onto a small set of dominant structural components, allowing each query to interact with representative global modes instead of all spatial tokens. This principal-projected attention suppresses redundant correlations while preserving globally coherent anatomical information.

Direct construction of the principal subspace, however, would still require costly spatial-domain decomposition. To avoid this overhead, we further exploit the spectral regularity of ordered volumetric features and construct an efficient approximation of the principal structural subspace in the Fourier domain. Dominant Fourier modes are selected according to their spectral energy and used as a compact basis for global interaction, replacing explicit eigendecomposition with efficient frequency-domain basis construction. This compact spectral representation reduces the interaction complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, providing near-linear computational scaling for a compact principal subspace while preserving global structural information.

Our contributions are summarized as follows:

- We introduce a **principal structural subspace** for medical image segmentation and propose principal-projected attention that captures globally coherent anatomical organization while suppressing redundant token interactions.
- We develop a **Fourier-based principal subspace approximation** that constructs dominant structural components without explicit spatial-domain eigendecomposition. Combined with principal projection, it reduces dense attention from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, enabling efficient global modeling of high-resolution volumes.

2 Related Work

Transformer-based Medical Image Segmentation. Transformers have been widely adopted in medical image segmentation to capture long-range dependencies beyond local convolutional receptive fields. UNETR Hatamizadeh et al. (2022) formulates volumetric encoding as sequence modeling, while Swin UNETR Tang et al. (2022) and nnFormer Zhou et al. (2023) introduce hierarchical and multi-scale attention for improved contextual modeling. Subsequent studies increasingly focus on computational efficiency. UNETR++ Shaker et al. (2024) employs efficient spatial–channel interaction, whereas SegFormer3D Perera et al. (2024) and Slim UNETR Pang et al. (2024) reduce volumetric modeling cost through lightweight hierarchical designs. More recent methods, including nnWNet Zhou et al. (2025) and SlimFormer-3D Hong et al. (2025), further improve efficiency by restructuring local–global interaction or suppressing redundant computation. These approaches improve global modeling and network efficiency, but generally do not explicitly construct a data-dependent principal subspace to guide attention compression.

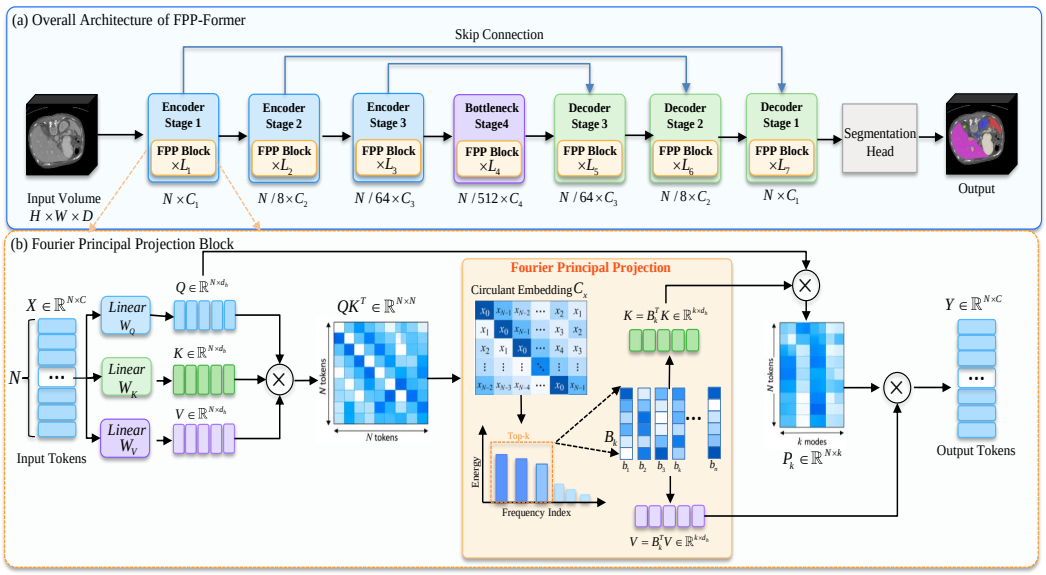


Figure 1: Overview of FPP-Former. (a) Overall architecture of FPP-Former. (b) Fourier Principal Projection Block.

Efficient Attention and Spectral Compression. Standard self-attention incurs quadratic complexity with respect to the number of tokens Vaswani et al. (2017). Efficient alternatives reduce this cost through different forms of interaction compression. Linformer Wang et al. (2020) projects keys and values into a reduced sequence dimension, Performer Choromanski et al. (2021) approximates softmax attention with random features, and Nyströmformer Xiong et al. (2021) reconstructs attention from landmark tokens. Recent methods further improve the information capacity of efficient attention, including PolaFormer Meng et al. (2025), RALA Fan et al. (2025), and Contract-and-Broadcast Self-Attention Wen et al. (2025). Agent Attention Han et al. (2024) similarly introduces compact agent tokens to mediate global token interaction.

Low-rank and spectral representations provide another route to compact global modeling. PCA and SVD concentrate correlated representations into dominant components Jolliffe & Cadima (2016), while Primal-Attention Chen et al. (2023) derives efficient attention from low-rank kernel decomposition. SpectFormer Patro et al. (2025) combines spectral token mixing with attention, and Frequency-Dynamic Attention Modulation Chen et al. (2025) explicitly models frequency-dependent attention responses. Related spectral analyses further investigate the conditioning and singular-value structure of attention operators Saratchandran & Lucey (2025). For medical image segmentation, WaveFormer Al Hasan et al. (2025) employs wavelet decomposition to construct compact multi-scale representations. Different from these formulations, FPP-Former uses a data-dependent principal subspace to guide key–value compression and exploits Fourier diagonalization to efficiently construct the projection basis.

3 Method

3.1 Overview

FPP-Former efficiently models global structural dependencies in medical images through compact principal-space representations. It adopts a hierarchical encoder–decoder architecture with skip connections and integrates Fourier Principal Projection Blocks at multiple resolution stages. As shown in Fig. 1, the proposed block projects keys and values onto selected principal modes to support efficient global interaction.

3.2 PRINCIPAL-SUBSPACE GUIDED ATTENTION COMPRESSION

Given ordered volumetric token features $X \in \mathbb{R}^{N\times d_h}$, standard self-attention produces $Q, K, V \in \mathbb{R}^{N\times d_h}$ and computes

$$A = \frac{QK^\top}{\sqrt{d_h}} \in \mathbb{R}^{N\times N}, Y = \mathrm{softmax}(A)V. \tag{1}$$

Standard self-attention explicitly evaluates interactions between all token pairs, resulting in an $N \times N$ interaction matrix. However, volumetric representations exhibit strong correlations across spatial tokens, while each pairwise score is generated from query and key embeddings of only d_h dimensions. This suggests that the full set of N^2 pairwise responses does not necessarily span N independent interaction directions, and that the dominant interactions may be represented in a substantially smaller subspace. We formalize this observation for the pre-softmax interaction matrix below.

Theorem 1 (Low-rank structure of dense attention scores). Let $A = QK^\top/\sqrt{d_h}$, where $Q, K \in \mathbb{R}^{N\times d_h}$. Then

$$\mathrm{rank}(A) \le d_h. \tag{2}$$

Furthermore, if the singular-value energy of A is concentrated in its leading k components, A admits an accurate rank-k approximation.

Proof. Since scalar multiplication does not affect matrix rank,

$$\mathrm{rank}(A) = \mathrm{rank}(QK^\top) \le \min\left(\mathrm{rank}(Q), \mathrm{rank}(K)\right) \le d_h. \tag{3}$$

Let $A = U\Sigma W^\top$, and let A_k denote its optimal rank-k approximation. According to the Eckart–Young–Mirsky theorem Eckart & Young (1936); Mirsky (1960),

$$\|A - A_k\|_F^2 = \sum_{i>k} \sigma_i^2. \tag{4}$$

Therefore, when the spectral energy is concentrated in the leading singular components, the dense interaction matrix can be accurately represented by a compact low-dimensional subspace. □

The complete proof of Theorem 1 is provided in Appendix A.1. Theorem 1 demonstrates that dense attention interactions admit a compact low-dimensional representation. Therefore, the key problem becomes identifying a projection subspace that preserves the dominant variations while suppressing redundant token dimensions.

Given a token representation $X \in \mathbb{R}^{N\times d_h}$, we seek an orthogonal basis $B \in \mathbb{R}^{N\times k}$ that maximizes the retained feature energy:

$$\underset{B^\top B = I_k}{\arg\max} \|B^\top X\|_F^2. \tag{5}$$

To characterize this objective, we define the token-domain second-order matrix as

$$C_X = \frac{1}{d_h} XX^\top = \frac{1}{d_h} \sum_{c=1}^{d_h} x^{(c)} x^{(c)\top}, \tag{6}$$

where $x^{(c)} \in \mathbb{R}^N$ denotes the ordered token signal of the c-th feature channel. Since the constant scaling factor does not affect the optimizer, Eq. (5) is equivalently expressed as

$$\underset{B^\top B = I_k}{\arg\max} \mathrm{Tr}\left(B^\top C_X B\right). \tag{7}$$

We denote the resulting principal basis by B_k. the optimization in Eq. 5 follows the principal-subspace criterion of PCA. Accordingly, $B_k^\star$ is given by the eigenvectors associated with the k largest eigenvalues of C_X. The resulting principal basis provides a data-dependent subspace for guiding attention compression. The principal subspace is incorporated into the attention computation through a shared token-space projection. Using the data-dependent basis B_k, the key and value representations are mapped from the original N-token space to k dominant principal modes:

$$\widetilde{K} = B_k^\top K \in \mathbb{R}^{k\times d_h}, \widetilde{V} = B_k^\top V \in \mathbb{R}^{k\times d_h}. \tag{8}$$

The attention operation is then performed in the resulting principal space:

$$Y_{\mathrm{PCA}} = \mathrm{softmax}\left(\frac{Q\widetilde{K}^{\top}}{\sqrt{d_h}}\right)\widetilde{V}. \tag{9}$$

This projection reduces the interaction space from N original tokens to k principal modes, each encoding a dominant variation pattern of the current volumetric representation. Consequently, global interactions are concentrated on structure-bearing directions while redundant token-space components are suppressed. The principal subspace therefore provides a unified mechanism for structural guidance and attention compression.

With B_k available, the query–key interaction map is reduced from $N \times N$ to $N \times k$. The corresponding computational complexity decreases from

$$\mathcal{O}(N^2 d_h)\text{to}\mathcal{O}(Nkd_h), k \ll N, \tag{10}$$

while the memory complexity of the interaction map is reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$.

The remaining computational bottleneck is the estimation of the principal basis. Explicit construction of $C_X \in \mathbb{R}^{N\times N}$ and extraction of its dominant eigenvectors become costly as the number of volumetric tokens increases. We therefore reformulate principal-subspace estimation in a structured circulant domain, enabling efficient recovery of the dominant modes through Fourier analysis.

3.3 Fast Fourier Principal Projection in the Circulant Domain

The principal-subspace formulation above relies on extracting the dominant eigenvectors of the token-domain second-order matrix, which becomes computationally expensive for dense volumetric representations. To make principal-subspace estimation tractable, we construct a structured circulant representation whose eigensystem can be efficiently obtained in the Fourier domain.

Circulant-Domain Construction. For the c-th feature channel, let

$$x^{(c)} = \left[x_0^{(c)}, x_1^{(c)}, \ldots, x_{N-1}^{(c)}\right]^{\top} \in \mathbb{R}^N \tag{11}$$

denote the ordered token signal, and let $S \in \mathbb{R}^{N\times N}$ denote the cyclic-shift operator. We introduce cyclic symmetrization to impose a structured eigensystem on the token-domain second-order representation:

$$\overline{C}_X = \frac{1}{N}\sum_{s=0}^{N-1} S^s C_X (S^s)^{\top}. \tag{12}$$

This operation maps the second-order statistics to a cyclic-shift-invariant representation, yielding a circulant structure while retaining the shift-averaged correlations of the original features.

Following the standard construction of circulant matrices Gray (2006), the corresponding embedding for each feature channel is defined as

$$\mathcal{C}\left(x^{(c)}\right) = \begin{bmatrix} x_0^{(c)} & x_{N-1}^{(c)} & \cdots & x_1^{(c)} \\ x_1^{(c)} & x_0^{(c)} & \cdots & x_2^{(c)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1}^{(c)} & x_{N-2}^{(c)} & \cdots & x_0^{(c)} \end{bmatrix}. \tag{13}$$

Each column is a cyclic shift of the preceding one, making $\mathcal{C}(x^{(c)})$ circulant by construction. Using Eq. 6, the cyclically symmetrized representation can then be written as

$$\overline{C}_X = \frac{1}{Nd_h}\sum_{c=1}^{d_h} \mathcal{C}\left(x^{(c)}\right)\mathcal{C}\left(x^{(c)}\right)^H. \tag{14}$$

The equivalence is derived in Appendix A.2.

Fourier Principal Decomposition. Let $F \in \mathbb{C}^{N\times N}$ denote the unitary discrete Fourier transform (DFT) matrix,

$$F_{pq} = \frac{1}{\sqrt{N}}\exp\left(-\mathrm{j}\frac{2\pi pq}{N}\right). \tag{15}$$

Since circulant matrices are exactly diagonalized by the Fourier basis,

$$\mathcal{C}\left(x^{(c)}\right) = F^H \operatorname{diag}\left(\sqrt{N} F x^{(c)}\right) F. \tag{16}$$

Substituting Eq. 16 into Eq. 14 yields

$$\overline{C}_X = F^H \operatorname{diag}(e) F, \tag{17}$$

where

$$e_p = \frac{1}{d_h} \sum_{c=1}^{d_h} \left| (F x^{(c)})_p \right|^2, p = 0, \ldots, N-1, \tag{18}$$

denotes the energy associated with the p-th Fourier mode.

Eq. 17 provides a direct spectral characterization of the dominant eigenspace. Since $\overline{C}_X$ is Hermitian positive semidefinite and diagonalized by F, its eigenvectors correspond to Fourier modes and the associated eigenvalues are determined by e_p. The dominant modes are therefore selected according to

$$\Omega_k = \mathrm{TopK}(e, k). \tag{19}$$

As shown in Appendix A.3, Top-k selection recovers the dominant eigenspace. For real-valued volumetric features, conjugate frequency pairs are selected jointly and converted into an equivalent real orthonormal basis

$$B_k \in \mathbb{R}^{N \times k}, B_k^\top B_k = I_k. \tag{20}$$

The resulting B_k provides a Fourier-derived approximation to the dominant principal subspace and is subsequently used in the principal projection defined in Eq. 8.

Computational Complexity. The structured Fourier formulation avoids explicit construction and eigendecomposition of the dense token-domain second-order matrix. Under the formulation in Eq. 6, explicitly constructing $C_X \in \mathbb{R}^{N \times N}$ requires $\mathcal{O}(N^2 d_h)$ operations and $\mathcal{O}(N^2)$ memory, while full dense eigendecomposition incurs $\mathcal{O}(N^3)$ time complexity.

In contrast, the Fourier coefficients of all d_h channels are computed by FFT in

$$\mathcal{O}(d_h N \log N) \tag{21}$$

time, followed by $\mathcal{O}(d_h N)$ spectral-energy aggregation and top-k mode selection. Neither C_X nor $\overline{C}_X$ needs to be explicitly materialized.

Consequently, Fourier Principal Projection converts principal-subspace estimation from explicit dense matrix decomposition into a near-linear spectral procedure with respect to the token number, enabling efficient construction of the structure-aware principal basis used in Section 3.2.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We evaluate FPP-Former on three medical image segmentation benchmarks: BraTS2017, Synapse, and ACDC. BraTS2017 focuses on brain tumor segmentation with three standard evaluation regions, including enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Synapse contains eight abdominal organ classes, while ACDC targets three cardiac structures: the left ventricle (LV), right ventricle (RV), and myocardium (MYO). For preprocessing, BraTS2017 volumes are cropped to $128 \times 128 \times 128$, while Synapse and ACDC samples are cropped to 224×224.

Evaluation Metrics. We adopt two widely used metrics for quantitative evaluation, including the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95). A higher DSC indicates better volumetric overlap between the prediction and ground truth, whereas a lower HD95 reflects more accurate boundary localization.

Implementation Details. All experiments are implemented using PyTorch 1.13.1 with Python 3.9 and conducted on a server equipped with two NVIDIA GeForce RTX 3090 GPUs. We use the same experimental protocol across all datasets and competing methods whenever applicable to ensure fair comparisons.

Table 1: Quantitative comparison with representative CNN- and ViT-based methods. Higher values indicate better performance.

Type	Method	Average ↑	WT ↑	ET ↑	TC ↑
CNN	U-Net++ Zhou et al. (2020)	72.5	80.2	63.1	74.2
	3D U-Net Çiçek et al. (2016)	78.4	86.5	70.2	78.5
	SegNet Badrinarayanan et al. (2017)	76.0	85.7	70.6	71.6
	AResU-Net Zhang et al. (2020)	79.3	88.1	71.9	78.0
	BrainSeg-Net Rehman et al. (2021)	81.0	89.5	74.5	79.2
ViT	SETR MLA Zheng et al. (2021)	63.8	69.7	55.2	66.5
	TransUNet Chen et al. (2024)	64.4	70.6	54.2	68.4
	CoTr Xie et al. (2021)	68.3	74.6	55.7	74.8
	TransBTS Wang et al. (2021)	69.3	77.5	57.3	73.1
	UNETR Hatamizadeh et al. (2022)	71.1	78.9	58.5	76.1
	SegFormer3D Perera et al. (2024)	81.3	89.2	73.4	81.3
	FPP-Former	**83.2**	**89.7**	**75.3**	**84.6**

Table 2: Quantitative comparison of different segmentation methods. Dice scores (%) are reported for eight organs. Higher Dice scores and lower HD95 values indicate better performance.

Type	Methods	Average		Dice ↑							
		Dice ↑	HD95 ↓	Aorta	GB	KL	KR	Liver	PC	SP	SM
CNN	nnU-Net Isensee et al. (2021)	**86.99**	10.96	**93.01**	**71.77**	85.57	88.18	**97.23**	**83.01**	91.86	85.26
	Attention U-Net Oktay et al. (2018)	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
	MedNeXt Roy et al. (2023)	79.45	23.96	89.42	65.21	82.33	78.54	94.28	60.38	88.21	77.23
	U-Net++ Zhou et al. (2020)	78.13	25.65	89.27	62.35	83.00	78.98	94.53	56.70	85.99	74.20
ViT	TransUNet Chen et al. (2024)	78.02	30.69	87.23	63.13	81.87	77.02	94.08	58.04	87.06	75.75
	TransCASCADE Rahman & Marculescu (2023)	81.88	18.05	86.89	68.75	85.31	82.05	94.21	65.77	91.62	80.41
	Swin-Unet Cao et al. (2022)	77.55	27.42	81.66	65.85	82.42	79.32	93.63	53.71	88.14	75.69
	HiFormer Heidari et al. (2023)	80.39	21.55	85.47	66.53	83.28	89.61	94.29	56.58	90.66	76.60
	MISSFormer Huang et al. (2023)	82.56	15.80	89.46	67.59	84.21	79.16	95.47	69.28	90.95	84.39
	PVT-CASCADE Rahman & Marculescu (2023)	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
	Laplacian-Former Azad et al. (2023)	81.90	18.66	86.55	71.19	84.23	80.52	94.90	64.75	91.91	81.14
	MIST Rahman et al. (2024)	84.16	12.48	87.52	67.14	**90.93**	89.97	94.16	65.24	92.08	86.24
	RWKV-UNet Jiang et al. (2025)	81.48	19.60	91.43	56.67	86.40	84.03	96.65	64.63	91.62	80.43
	UNETR++ Shaker et al. (2024)	84.75	11.10	91.36	67.26	84.98	85.73	95.15	81.68	89.64	82.26
	FPP-Former	84.84	**10.87**	88.01	67.86	90.76	**90.61**	95.22	65.68	**93.33**	**87.23**

4.2 Comparison with Existing Methods

As shown in Table 1, FPP-Former achieves the highest average Dice score of 83.2% on BraTS2017. Compared with SegFormer3D, the strongest ViT baseline in terms of average performance, FPP-Former yields a relative improvement of 2.34%. It also improves ET and TC segmentation by 2.59% and 4.06%, respectively, while maintaining comparable performance on WT. These results indicate that FPP-Former provides more effective global structural modeling for challenging tumor regions. On the Synapse dataset, as shown in Table 2, FPP-Former achieves an average Dice score of 84.84%, slightly exceeding UNETR++ by 0.11% among the ViT-based methods. More importantly, it achieves the lowest HD95 of 10.87, corresponding to a 2.07% reduction compared with UNETR++. FPP-Former also obtains the best Dice scores on the right kidney, spleen, and stomach, improving over the strongest competing results by 0.71%, 1.36%, and 1.15%, respectively. This demonstrates that the proposed principal projection strategy is effective in preserving both global consistency and fine-grained anatomical boundaries. As reported in Table 3, FPP-Former achieves the best overall performance on ACDC with an average Dice score of 91.94%. Relative to nnFormer, FPP-Former improves the average Dice by 0.12%, while achieving relative gains of 0.27% and 0.55% on LV and RV segmentation, respectively. Although PVT-CASCADE attains a slightly higher MYO score, FPP-Former provides a more balanced performance across the three cardiac structures, leading to the highest overall accuracy among the compared methods. The quantitative improvements are further supported by qualitative comparisons in Fig. 2. FPP-Former consistently generates predictions with improved structural coherence and boundary alignment across BraTS 2017, Synapse, and ACDC, demonstrating the effectiveness of Fourier principal projection for structure-aware segmentation.

Table 3: Segmentation results on the ACDC dataset among different methods. Dice scores (%) are reported for the left ventricle (LV), right ventricle (RV), myocardium (MYO), and the average performance. Higher values indicate better performance.

Type	Methods	LV ↑	RV ↑	MYO ↑	Average ↑
CNN	U-Net++ Zhou et al. (2020)	94.82	89.58	89.52	91.31
	SegNet Badrinarayanan et al. (2017)	93.71	85.91	87.47	89.03
	MedNeXt Roy et al. (2023)	95.72	89.43	89.77	91.64
ViT	TransUNet Chen et al. (2024)	95.37	88.47	84.12	89.32
	MISSFormer Huang et al. (2023)	94.14	88.21	87.38	89.91
	Swin-Unet Cao et al. (2022)	95.29	88.46	85.61	89.78
	MedFormer Gao et al. (2022)	95.47	88.63	89.44	91.18
	nnFormer Zhou et al. (2023)	95.71	90.13	89.65	91.83
	PVT-CASCADE Rahman & Marculescu (2023)	95.51	88.85	**89.99**	91.45
	RWKV-UNet Jiang et al. (2025)	95.62	89.57	89.12	91.44
	FPP-Former	**95.97**	**90.63**	89.24	**91.94**

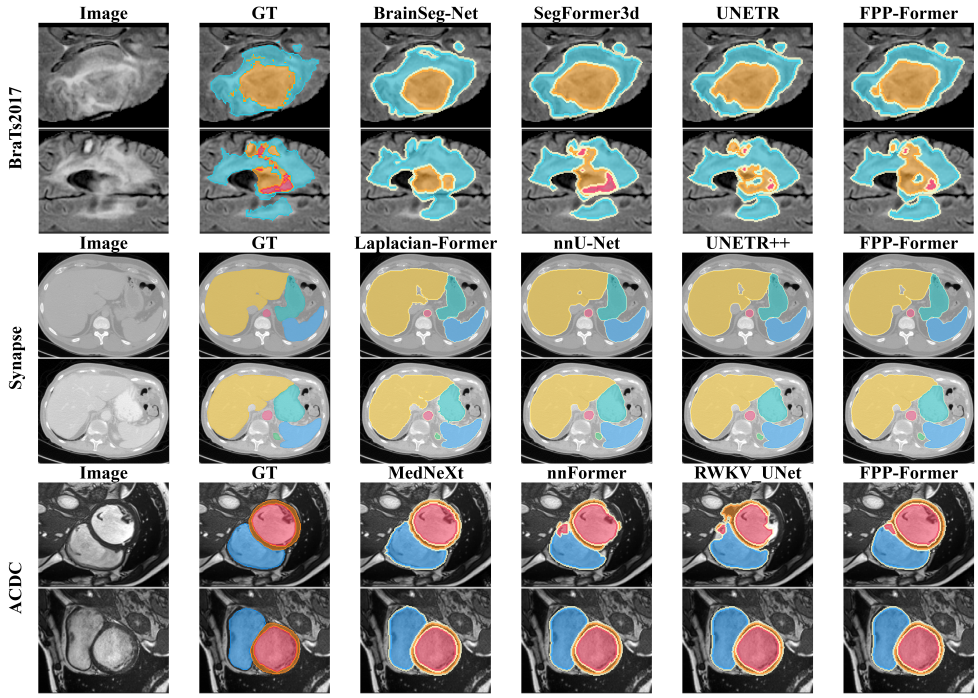


Figure 2: Visual comparison of segmentation results on BraTS 2017, Synapse, and ACDC datasets.

4.3 Ablation Study

We conduct ablation studies to evaluate the effectiveness of the proposed Fourier Principal Linear Attention (FPLA) and analyze the influence of the principal subspace dimension. We further compare FPLA with representative attention mechanisms in terms of computational complexity, memory consumption, training cost, and inference efficiency.

Effect of attention formulation. Table 4 compares different attention mechanisms under the same U-shaped Transformer architecture. FPLA achieves Dice scores of 83.20%, 84.84%, and 91.94% on BraTS2017, Synapse, and ACDC, respectively, outperforming other efficient attention variants on most datasets. These results demonstrate that Fourier principal projection effectively preserves global structural information while reducing redundant token interactions. Additional component-wise ablations and Fourier-energy analysis are provided in Appendix C.1 and Appendix C.4, respectively.

Sensitivity to the principal subspace dimension. We further evaluate the influence of the principal dimension k across all three datasets. Increasing k from $N/32$ to $N/8$ consistently improves segmentation performance by retaining richer structural variations. However, further increasing the

Table 4: Ablation study of different attention mechanisms and the principal subspace dimension. Higher Dice scores indicate better performance. Sensitivity to k is evaluated on all three datasets.

Attention	k/N	BraTS2017 Dice ↑	Synapse Dice ↑	ACDC Dice ↑
SA Vaswani et al. (2017)	–	**83.54**	84.17	91.52
LA Wang et al. (2020)	–	80.23	82.63	89.75
Agent Attention Han et al. (2024)	–	81.31	81.96	90.32
PolaFormer Meng et al. (2025)	–	82.83	84.06	91.68
FPLA	$\mathbf{1/8}$	83.20	**84.84**	**91.94**
FPLA	$1/32$	80.74	81.92	89.41
FPLA	$1/16$	82.06	83.37	90.76
FPLA	$1/4$	82.43	83.96	91.07

subspace dimension to $N/4$ degrades segmentation performance while introducing additional computation, suggesting that excessive components may reintroduce redundant variations. Therefore, $k/N = 1/8$ is selected as the default setting for balancing accuracy and efficiency. A detailed analysis of the computational and memory costs under different principal dimensions is provided in Appendix C.2.

Table 5: Computational efficiency comparison of different attention mechanisms.

Method	FLOPs (G)↓	Train Mem. (GB)↓	Infer Mem. (GB)↓	Train (ms/iter)↓	Infer (ms/image)↓	FPS↑
SA Vaswani et al. (2017)	105.93	25.35	10.14	510.59	200.83	4.98
LA Wang et al. (2020)	12.87	1.64	0.57	53.73	16.47	60.71
Agent Attention Han et al. (2024)	12.94	0.54	0.42	35.68	7.84	127.55
PolaFormer Meng et al. (2025)	13.87	0.53	0.30	**34.62**	8.23	121.51
FPLA	**12.81**	**0.49**	**0.22**	40.38	**7.68**	**130.21**

Computational efficiency comparison. Table 5 compares FPLA with representative attention mechanisms under an identical U-shaped Transformer architecture. Compared with SA, FPLA reduces FLOPs from 105.93G to 12.81G, corresponding to an 87.9% reduction. Training and inference memory are reduced from 25.35GB and 10.14GB to 0.49GB and 0.22GB, respectively. FPLA also decreases the inference latency from 200.83ms to 7.68ms and increases the throughput from 4.98 FPS to 130.21 FPS, demonstrating its substantially improved computational scalability for dense medical image representation.Additional module-level scaling comparisons across different feature resolutions are reported in Appendix C.3.

Compared with other efficient attention mechanisms, FPLA achieves the lowest computational cost and memory consumption, requiring 12.81G FLOPs, 0.49GB training memory, and 0.22GB inference memory. It also obtains the lowest inference latency of 7.68ms and the highest throughput of 130.21 FPS. Although its training latency of 40.38ms/iter is slightly higher than PolaFormer and Agent Attention, FPLA provides a favorable overall balance between computational efficiency, memory usage, and segmentation performance. These results indicate that Fourier principal projection effectively compresses redundant token interactions while maintaining efficient global structural modeling.

5 Conclusion

In this work, we propose FPP-Former, an efficient framework for medical image segmentation, which incorporates a novel Fourier-domain principal projection strategy for lightweight long-range dependency modeling. Motivated by the low-rank characteristics of volumetric feature interactions, we construct compact principal bases in the Fourier domain and project token representations onto a reduced subspace, thereby reducing the computational redundancy of conventional self-attention while preserving informative global representations. In addition, FPP-Former effectively integrates spectral principal components with spatial features to maintain fine-grained anatomical structures during segmentation. Extensive experiments on multiple medical image segmentation benchmarks demonstrate that FPP-Former achieves competitive segmentation performance with substantially reduced model complexity and computational cost, showing a favorable balance between segmentation accuracy and computational efficiency.

REFERENCES

Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

Md Mahfuz Al Hasan, Mahdi Zaman, Abdul Jawad, Alberto Santamaria-Pang, Ho Hin Lee, Ivan Tarapov, Kyle B. See, Md Shah Imran, Antika Roy, Yaser Pourmohammadi Fallah, Navid Asadizanjani, and Reza Forghani. WaveFormer: A 3d transformer with wavelet-driven feature representation for efficient medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15963, pp. 684–694. Springer Nature Switzerland, 2025.

Reza Azad, Amirhossein Kazerouni, Babak Azad, Ehsan Khodapanah Aghdam, Yury Velichko, Ulas Bagci, and Dorit Merhof. Laplacian-former: Overcoming the limitations of vision transformers in local texture detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14222, pp. 736–746. Springer, 2023. doi: 10.1007/978-3-031-43898-1_70.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, pp. 205–218. Springer, 2022. doi: 10.1007/978-3-031-25066-8_9.

Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformer for graph representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 3469–3489. PMLR, 2022.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

Linwei Chen, Lin Gu, and Ying Fu. Frequency-dynamic attention modulation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22620–22632, 2025.

Yingyi Chen, Qinghua Tao, Francesco Tonin, and Johan A. K. Suykens. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, volume 9901, pp. 424–432. Springer, 2016. doi: 10.1007/978-3-319-46723-8_49.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Qihang Fan, Huaibo Huang, and Ran He. Breaking the low-rank dilemma of linear attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25271–25280, 2025.

Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N. Metaxas. A data-scalable transformer for medical image segmentation: Architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022.

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006. doi: 10.1561/0100000006.

Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *European Conference on Computer Vision*, 2024.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6202–6212, 2023.

Yang Hong, Lei Zhang, Xujiong Ye, and Jianqing Mo. SlimFormer-3D: A layer-adaptive lightweight transformer for efficient 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15961, pp. 537–546. Springer Nature Switzerland, 2025.

Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. MISSFormer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5):1484–1494, 2023. doi: 10.1109/TMI.2022.3230943.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Juntao Jiang, Jiangning Zhang, Weixuan Liu, Muxuan Gao, Xiaobin Hu, Zhucun Xue, Yong Liu, and Shuicheng Yan. RWKV-UNet: Improving unet with long-range cooperation for effective medical image segmentation. *arXiv preprint arXiv:2501.08458*, 2025.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Manjin Kim, Paul Hongsuck Seo, Cordelia Schmid, and Minsu Cho. Learning correlation structures for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18941–18951, 2024.

Weikang Meng, Yadan Luo, Xin Li, Dongmei Jiang, and Zheng Zhang. PolaFormer: Polarity-aware linear attention for vision transformers. In *International Conference on Learning Representations*, 2025.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Yan Pang, Jiaming Liang, Teng Huang, Hao Chen, Yunhao Li, Dan Li, Lin Huang, and Qiong Wang. Slim UNETR: Scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources. *IEEE Transactions on Medical Imaging*, 43(3):994–1005, 2024.

Badri N. Patro, Vinay P. Namboodiri, and Vijay S. Agneeswaran. SpectFormer: Frequency and attention is what you need in a vision transformer. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 9525–9536, 2025.

Shehan Perera, Pouyan Navard, and Alper Yilmaz. SegFormer3D: An efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4981–4988, 2024.

Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6222–6231, 2023.

Md Motiur Rahman, Shiva Shokouhmand, Smriti Bhatt, and Miad Faezipour. MIST: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 404–413, 2024.

Mobeen Ur Rehman, SeungBin Cho, Jeehong Kim, and Kil To Chong. BrainSeg-Net: Brain tumor mr image segmentation via enhanced encoder–decoder network. *Diagnostics*, 11(2):169, 2021. doi: 10.3390/diagnostics11020169.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jäger, and Klaus H. Maier-Hein. MedNeXt: Transformer-driven scaling of convnets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023. doi: 10.1007/978-3-031-43901-8_39.

Hemanth Saratchandran and Simon Lucey. Spectral conditioning of attention improves transformer performance. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 109–119. Springer, 2021. doi: 10.1007/978-3-030-87193-2_11.

Qishuai Wen, Zhiyuan Huang, and Chun-Guang Li. Towards interpretable and efficient attention: Compressing all by contracting a few. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3):37–52, 1987.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 171–180. Springer, 2021. doi: 10.1007/978-3-030-87199-4_16.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14138–14148, 2021.

Jianxin Zhang, Xiaogang Lv, Hengbo Zhang, and Bin Liu. AResU-Net: Attention residual u-net for brain tumor segmentation. *Symmetry*, 12(5):721, 2020. doi: 10.3390/sym12050721.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890, 2021.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnWNet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20852–20862, 2025.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020. doi: 10.1109/TMI.2019.2959609.

AI USE STATEMENT

In this work, AI-assisted tools were used only for language refinement, manuscript formatting, and literature search support during manuscript preparation. AI tools were not used to generate the proposed methodology, experimental results, or scientific conclusions. All AI-assisted content was reviewed and verified by the authors.

A THEORETICAL DERIVATION OF FOURIER PRINCIPAL PROJECTION

This appendix provides the theoretical details underlying the proposed Fourier Principal Projection (FPP). We first prove the low-rank property of dense attention scores stated in Theorem 1 of Section 3.2. We then derive the circulant-domain principal-subspace formulation, its exact Fourier-domain eigendecomposition, and the resulting Fourier principal attention.

A.1 PROOF OF THEOREM 1: LOW-RANK STRUCTURE OF DENSE ATTENTION SCORES

Recall the pre-softmax attention score matrix

$$A = \frac{QK^\top}{\sqrt{d_h}}, \qquad Q, K \in \mathbb{R}^{N \times d_h}. \tag{22}$$

Since multiplication by the nonzero scalar $1/\sqrt{d_h}$ does not change matrix rank,

$$\begin{aligned} \operatorname{rank}(A) &= \operatorname{rank}\left(QK^\top\right) \\ &\le \min\left(\operatorname{rank}(Q), \operatorname{rank}(K)\right) \\ &\le d_h. \end{aligned} \tag{23}$$

Therefore,

$$\operatorname{rank}(A) \le d_h, \tag{24}$$

which proves the first statement of Theorem 1.

To characterize the approximation property, let the singular value decomposition of A be

$$A = U\Sigma W^\top, \tag{25}$$

where the singular values satisfy

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge 0. \tag{26}$$

Let

$$A_k = U_k \Sigma_k W_k^\top \tag{27}$$

denote the truncated rank-k approximation obtained from the leading k singular components. According to the Eckart–Young–Mirsky theorem Eckart & Young (1936); Mirsky (1960), A_k is an optimal rank-k approximation under the Frobenius norm and satisfies

$$\min_{\operatorname{rank}(\widehat{A}) \le k} \left\| A - \widehat{A} \right\|_F^2 = \|A - A_k\|_F^2 = \sum_{i>k} \sigma_i^2. \tag{28}$$

Hence, when the singular-value energy of A is concentrated in its leading k components,

$$\sum_{i>k} \sigma_i^2 \ll \sum_i \sigma_i^2, \tag{29}$$

the dense $N \times N$ interaction matrix can be accurately represented within a substantially lower-dimensional subspace. This completes the proof of Theorem 1. □

A.2 CIRCULANT-DOMAIN PRINCIPAL FORMULATION

We first formulate principal-subspace estimation in a structured circulant domain. The token-domain second-order statistics are cyclically symmetrized to obtain a shift-invariant representation whose dominant eigenspace defines the structured principal basis.

Given ordered token features

$$X \in \mathbb{R}^{N\times d_h}, \tag{30}$$

let

$$x^{(c)} = \left[x_0^{(c)}, x_1^{(c)}, \ldots, x_{N-1}^{(c)}\right]^\top \in \mathbb{R}^N \tag{31}$$

denote the ordered token signal of the c-th feature channel.

The token-domain second-order representation is

$$C_X = \frac{1}{d_h} X X^\top = \frac{1}{d_h} \sum_{c=1}^{d_h} x^{(c)} x^{(c)\top}. \tag{32}$$

For each feature channel, we construct the circulant embedding

$$\mathcal{C}(x^{(c)}) = \begin{bmatrix} x_0^{(c)} & x_{N-1}^{(c)} & \cdots & x_1^{(c)} \\ x_1^{(c)} & x_0^{(c)} & \cdots & x_2^{(c)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1}^{(c)} & x_{N-2}^{(c)} & \cdots & x_0^{(c)} \end{bmatrix} \in \mathbb{R}^{N\times N}, \tag{33}$$

whose entries satisfy

$$\left[\mathcal{C}(x^{(c)})\right]_{ij} = x^{(c)}_{(i-j) \bmod N}. \tag{34}$$

Thus, each column is a cyclic shift of the preceding one.

Let $S \in \mathbb{R}^{N\times N}$ denote the cyclic-shift operator. We define the cyclically symmetrized second-order representation as

$$\overline{C}_X = \frac{1}{N} \sum_{s=0}^{N-1} S^s C_X (S^s)^\top. \tag{35}$$

Using Eq. (32) and collecting all cyclic shifts of each channel signal gives the equivalent form

$$\overline{C}_X = \frac{1}{N d_h} \sum_{c=1}^{d_h} \mathcal{C}(x^{(c)}) \mathcal{C}(x^{(c)})^H. \tag{36}$$

Therefore, $\overline{C}_X$ corresponds to the cyclically symmetrized second-order statistics of the original token representation. Unlike the unrestricted matrix C_X, this representation possesses a structured circulant eigensystem.

We seek a k-dimensional orthonormal basis $B \in \mathbb{C}^{N\times k}$ that maximizes the retained energy of this structured representation:

$$B_k^\star \in \underset{B^H B = I_k}{\arg\max} \operatorname{Tr}\left(B^H \overline{C}_X B\right). \tag{37}$$

The corresponding Lagrangian is

$$\mathcal{L}(B, \Lambda) = \operatorname{Tr}\left(B^H \overline{C}_X B\right) - \operatorname{Tr}\left[\Lambda\left(B^H B - I_k\right)\right]. \tag{38}$$

Setting the derivative with respect to B^* to zero gives

$$\frac{\partial \mathcal{L}}{\partial B^*} = 0 \quad \Longrightarrow \quad \overline{C}_X B = B\Lambda. \tag{39}$$

Hence, the optimal basis is determined by the eigenvectors associated with the dominant eigenvalues of $\overline{C}_X$. The remaining problem is therefore to efficiently obtain the eigensystem of this structured second-order representation, which is derived in Appendix A.3.

A.3 Fourier-Domain Principal Decomposition

We next exploit the circulant structure of $\overline{C}_X$ to derive its eigensystem directly in the Fourier domain. This formulation converts dense eigendecomposition into spectral-energy estimation and enables efficient identification of the dominant principal modes.

Let $F \in \mathbb{C}^{N\times N}$ denote the unitary discrete Fourier transform (DFT) matrix,

$$F_{pq} = \frac{1}{\sqrt{N}} \exp\left(-\mathrm{j}\frac{2\pi pq}{N}\right), F^H F = FF^H = I_N, \tag{40}$$

for $p, q = 0, \ldots, N-1$.

A fundamental property of circulant matrices is that they are exactly diagonalized by the Fourier basis:

$$\mathcal{C}(x^{(c)}) = F^H \operatorname{diag}\left(\hat{x}^{(c)}\right) F, \hat{x}^{(c)} = \sqrt{N}\, F x^{(c)}. \tag{41}$$

Therefore,

$$\begin{aligned}\mathcal{C}(x^{(c)})\mathcal{C}(x^{(c)})^H &= F^H \operatorname{diag}\left(\hat{x}^{(c)}\right) FF^H \operatorname{diag}\left(\hat{x}^{(c)}\right)^H F \\ &= F^H \operatorname{diag}\left(|\hat{x}^{(c)}|^2\right) F.\end{aligned} \tag{42}$$

Substituting Eq. (**??**) into Eq. (36) yields

$$\begin{aligned}\overline{C}_X &= F^H \operatorname{diag}\left(\frac{1}{Nd_h}\sum_{c=1}^{d_h} |\hat{x}^{(c)}|^2\right) F \\ &= F^H \operatorname{diag}(\lambda) F.\end{aligned} \tag{43}$$

where

$$\lambda_p = \frac{1}{d_h}\sum_{c=1}^{d_h} \left|(Fx^{(c)})_p\right|^2, p = 0, \ldots, N-1. \tag{44}$$

Thus, the eigenvalues of $\overline{C}_X$ are exactly the spectral energies $\{e_p\}_{p=0}^{N-1}$, while the corresponding eigenvectors are Fourier modes.

Optimal principal-mode selection. Substituting Eq. (42) into Eq. (39) gives

$$F^H \operatorname{diag}(\lambda) FB = B\Lambda. \tag{45}$$

Left-multiplying by F yields

$$\operatorname{diag}(\lambda)(FB) = (FB)\Lambda. \tag{46}$$

Let

$$Z = FB. \tag{47}$$

Since F is unitary,

$$Z^H Z = B^H F^H FB = I_k. \tag{48}$$

Accordingly, the principal-subspace objective in Eq. (37) becomes

$$Z_k^\star \in \underset{Z^H Z = I_k}{\arg\max} \operatorname{Tr}\left(Z^H \operatorname{diag}(\lambda) Z\right). \tag{49}$$

Because $\operatorname{diag}(\lambda)$ is diagonal, the objective is maximized by selecting the coordinates associated with the k largest eigenvalues:

$$\Omega_k = \operatorname{TopK}(\lambda, k). \tag{50}$$

Let $E_{\Omega_k} \in \mathbb{R}^{N\times k}$ denote the corresponding column-selection matrix. One valid maximizer of Eq. (49) is

$$Z_k^\star = E_{\Omega_k}. \tag{51}$$

Transforming back to the token domain gives

$$B_k^\star = F^H E_{\Omega_k} = F^H[:, \Omega_k]. \tag{52}$$

Therefore,

$$\max_{B^H B = I_k} \operatorname{Tr}\left(B^H \overline{C}_X B\right) = \sum_{p \in \Omega_k} \lambda_p. \tag{53}$$

Hence, selecting the k largest eigenvalues, which correspond exactly to the dominant Fourier spectral energies, recovers a dominant k-dimensional eigenspace of the cyclically symmetrized representation $\overline{C}_X$. When eigenvalues are tied, the corresponding principal eigenspace is invariant to orthonormal rotations within the degenerate eigenspace.

Real-valued Fourier principal basis. For real-valued token features, the Fourier coefficients satisfy conjugate symmetry:

$$(Fx^{(c)})_{N-p} = (Fx^{(c)})_p^*. \tag{54}$$

Accordingly, non-self-conjugate Fourier modes are selected jointly with their conjugate counterparts. For a conjugate pair of Fourier vectors ϕ_p and $\phi_{N-p} = \phi_p^*$, an equivalent real orthonormal basis can be constructed as

$$b_p^{\cos} = \frac{\phi_p + \phi_{N-p}}{\sqrt{2}}, b_p^{\sin} = \frac{\phi_p - \phi_{N-p}}{\sqrt{2}\,\mathrm{j}}. \tag{55}$$

The DC component and, for even N, the Nyquist component are already real-valued and are retained directly. Collecting the selected real-valued modes produces

$$B_k \in \mathbb{R}^{N \times k}, B_k^\top B_k = I_k. \tag{56}$$

Therefore, Fourier Principal Projection provides an exact principal basis for the structured circulant representation $\overline{C}_X$, which serves as a computationally efficient, cyclically symmetrized approximation to the unconstrained token-domain second-order representation C_X.

Fourier principal attention and complexity. Using the resulting real-valued Fourier principal basis, the keys and values are projected into the selected principal subspace:

$$\widetilde{K} = B_k^\top K, \widetilde{V} = B_k^\top V, \tag{57}$$

where

$$\widetilde{K}, \widetilde{V} \in \mathbb{R}^{k \times d_h}. \tag{58}$$

Fourier Principal Linear Attention (FPLA) is then computed as

$$Y = \operatorname{softmax}\left(\frac{Q\widetilde{K}^\top}{\sqrt{d_h}}\right)\widetilde{V}. \tag{59}$$

The dense $N \times N$ interaction matrix is therefore replaced by an $N \times k$ interaction matrix.

Computing the Fourier spectra of all d_h feature channels using FFT requires

$$\mathcal{O}\left(d_h N \log N\right) \tag{60}$$

operations. Spectral-energy aggregation requires $\mathcal{O}(d_h N)$ operations, while the principal projection and compressed attention require

$$\mathcal{O}\left(N k d_h\right). \tag{61}$$

The overall complexity is therefore

$$\mathcal{O}\left(d_h N \log N + N k d_h\right), \tag{62}$$

with interaction-map memory complexity

$$\mathcal{O}(Nk). \tag{63}$$

For $k \ll N$, FPLA avoids the quadratic $\mathcal{O}(N^2 d_h)$ interaction of standard self-attention. Meanwhile, the selected Fourier modes constitute the exact dominant principal subspace of the structured representation $\overline{C}_X$, providing an efficient spectral approximation to principal-subspace attention in the original token domain.

Table 6: Dataset characteristics and implementation settings for BraTS 2017, Synapse, and ACDC.

Setting	BraTS 2017	Synapse	ACDC
Modality	MRI (FLAIR, T1w, T1Gd, T2w)	CT	Cine MRI
Number of cases	484	30	100
Training/validation/testing split	411/-/73	18/-/12	70/10/20
Segmentation targets	ET, TC, WT	8 abdominal organs	LV, RV, MYO
Number of classes	3 regions	8 organs	3 structures
Input size	4×128^3	224×224	224×224
Optimizer	AdamW	AdamW	AdamW
Initial learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	1×10^{-2}	1×10^{-4}	1×10^{-4}
Batch size	2	4	4
Training epochs	800	300	400
Loss function	Dice loss	$0.3\,CE + 0.7\,Dice$	$0.3\,CE + 0.7\,Dice$
Data augmentation	Crop, flip, rotation, noise	Rotation and flipping	Rotation and flipping

B Dataset details and training configurations

We provide additional details of the datasets and implementation settings used in our experiments. Table 6 summarizes the dataset characteristics and training configurations for BraTS 2017, Synapse, and ACDC. All settings are directly obtained from the corresponding experimental configurations.

BraTS 2017. BraTS 2017 provides multimodal brain MRI scans with four imaging sequences: T1, contrast-enhanced T1 (T1Gd), T2, and FLAIR. The labeled training set contains 411 subjects, including high-grade and low-grade glioma cases. Following the segmentation protocol used in our experiments, predictions are evaluated on three clinically relevant tumor regions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). The input volumes are cropped to $128 \times 128 \times 128$ for training and inference.

Synapse. The Synapse multi-organ dataset contains 30 abdominal CT volumes and a total of approximately 3,779 axial slices. Following the commonly adopted protocol, eight abdominal structures are evaluated: aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. The axial slices are processed at a spatial resolution of 224×224 in our experiments.

ACDC. The ACDC dataset contains cine-MR examinations from 100 patients with manual annotations provided for the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The annotations correspond to the cardiac structures used for quantitative evaluation in our experiments. The input slices are processed at 224×224 resolution.

C Additional Ablation and Efficiency Analysis

C.1 Additional Ablation of Fourier Principal Projection

To further investigate the contribution of different components in Fourier Principal Projection, we compare FPLA with two degraded variants: random Fourier mode selection and random subspace projection. As shown in Fig. 3, random subspace projection produces the largest performance degradation, indicating that arbitrary low-dimensional projections cannot effectively preserve structural information. Random Fourier selection improves over random subspace projection, demonstrating the benefit of the Fourier basis for modeling global token dependencies. By further selecting dominant modes according to spectral energy, FPLA achieves the best overall performance across the three datasets, verifying the effectiveness of energy-based Fourier principal selection.

C.2 Analysis of Different Principal Dimensions

The influence of the principal subspace dimension on computational efficiency is investigated in Table 7. As k increases, the principal projection space becomes larger, leading to higher computational and memory costs according to the $\mathcal{O}(Nkd_h)$ interaction term. When k/N increases from $1/32$ to

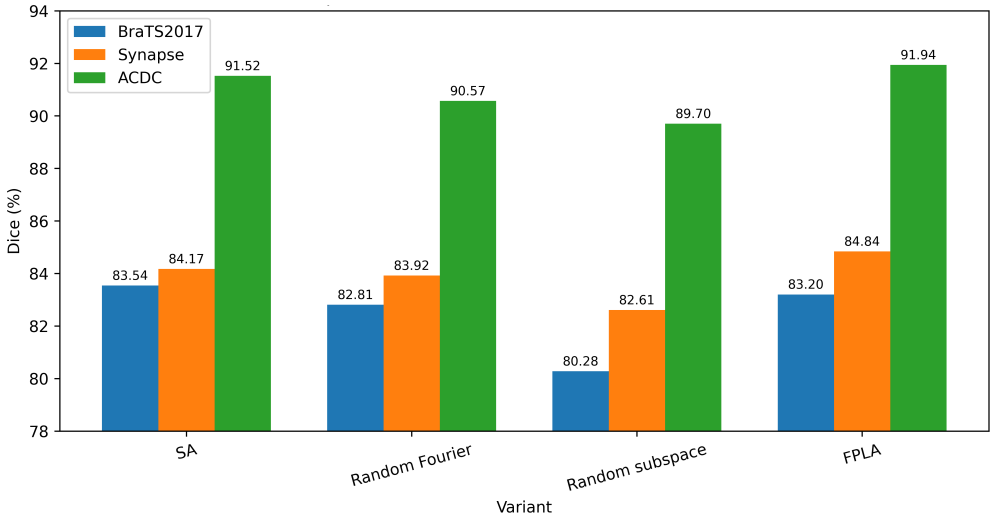


Figure 3: Visualization of component ablation results across BraTS2017, Synapse, and ACDC datasets. Compared with random Fourier mode selection and random subspace projection, the proposed FPLA consistently achieves superior segmentation performance, demonstrating the effectiveness of the Fourier principal basis and energy-based Top-k selection.

Table 7: Computational cost analysis with different principal subspace dimensions. Increasing k improves representation capacity but introduces additional computational overhead due to the enlarged principal projection space.

k/N	FLOPs (G)↓	Train Mem. (GB)↓	Infer Mem. (GB)↓	Infer Latency (ms)↓
$1/32$	6.73	0.31	0.14	5.21
$1/16$	8.65	0.38	0.17	6.37
$1/8$	12.81	0.49	0.22	7.68
$1/4$	16.49	0.73	0.38	13.93

$1/4$, the FLOPs rise from 6.73G to 16.49G, the training memory increases from 0.31GB to 0.73GB, and the inference latency increases from 5.21ms to 13.93ms. At the adopted setting of $k/N = 1/8$, FPLA requires 12.81G FLOPs, 0.49GB training memory, 0.22GB inference memory, and 7.68ms inference latency. This setting provides a favorable balance between principal-subspace capacity and computational efficiency.

C.3 Additional Efficiency Analysis

Table 8: Scaling comparison between standard self-attention (SA) and FPLA across different token resolutions.

Resolution	Method	FLOPs (G)↓	Train Mem. (GB)↓	Infer Mem. (GB)↓	Latency (ms)↓
112×112	SA	41.480	20.177	10.092	93.699
	FPLA	**0.625**	**0.058**	**0.048**	**1.214**
56×56	SA	5.495	1.286	0.653	5.077
	FPLA	**0.621**	**0.040**	**0.033**	**0.431**
28×28	SA	1.044	0.102	0.061	**0.401**
	FPLA	**0.619**	**0.028**	**0.026**	0.429
14×14	SA	**0.490**	0.030	**0.026**	**0.256**
	FPLA	0.618	**0.028**	0.026	0.416

As shown in Table 8, FPLA provides clear efficiency advantages at high-resolution stages. At 112×112, it reduces the FLOPs from 41.480G to 0.625G, the training memory from 20.177GB to 0.058GB, and the inference latency from 93.699ms to 1.214ms compared with SA. As the spatial resolution decreases, the gap becomes smaller because the dense attention matrix is already compact and the fixed Fourier projection overhead becomes more noticeable. Overall, these results show that FPLA is particularly suitable for high-resolution feature representations.

C.4 Analysis of Fourier Energy Distribution

To further understand why a compact Fourier principal subspace can preserve most of the structural information, we analyze the spectral-energy distribution of input from BraTS2017, Synapse, and ACDC. For each sample, Fourier coefficients are ranked according to their spectral energy, and the cumulative energy retained under different component ratios is measured. We further visualize representative reconstructions using only a few dominant Fourier modes.

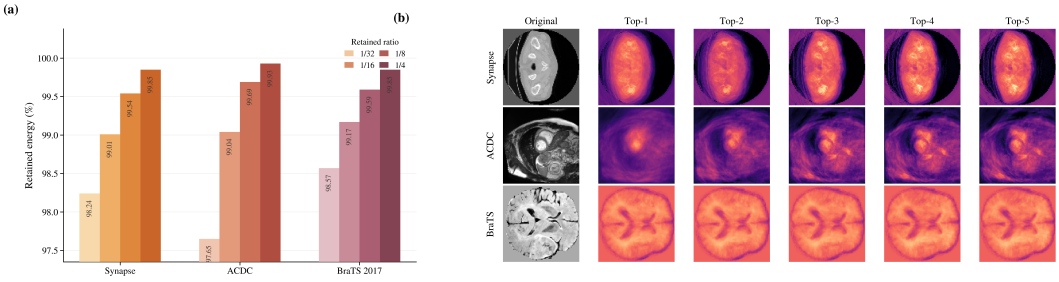


Figure 4: Analysis of Fourier energy concentration and reconstruction quality. (a) Cumulative spectral energy retained at different component ratios. (b) Reconstructions using the Top-1 to Top-5 dominant Fourier components. The results show that most spectral energy is concentrated in a compact subset of Fourier modes.

Table 9: Fourier energy concentration analysis from BraTS2017, Synapse, and ACDC. The left group reports the cumulative spectral energy retained under different component ratios, while the right group reports the minimum percentage of Fourier components required to preserve predefined energy levels.

Dataset	Retained Energy (%)				Required Components (%)		
	$1/32$	$1/16$	$1/8$	$1/4$	80% Energy	90% Energy	95% Energy
BraTS2017	98.57	99.17	99.59	99.85	0.02	0.09	0.36
Synapse	98.24	99.01	99.54	99.85	0.01	0.08	0.50
ACDC	97.65	99.04	99.69	99.93	0.16	0.55	1.41

Table 9 provides two complementary views of the Fourier energy concentration across the three datasets. When retaining only $1/32$ of the Fourier components, BraTS2017, Synapse, and ACDC already preserve 98.57%, 98.24%, and 97.65% of the total spectral energy, respectively. At the adopted ratio of $k/N = 1/8$, the retained energy further increases to 99.59%, 99.54%, and 99.69%. Conversely, only 0.36%, 0.50%, and 1.41% of the Fourier components are required to preserve 95% of the spectral energy for BraTS2017, Synapse, and ACDC, respectively. These complementary observations consistently demonstrate that the spectral energy is strongly concentrated in a compact subset of dominant Fourier components. Together with the reconstruction results in Fig. 4, this analysis supports the use of a compact Top-k Fourier representation in FPLA.

D Additional Qualitative Results

We provide additional qualitative comparisons on BraTS 2017, Synapse, and ACDC to further evaluate the segmentation behavior of FPP-Former across different anatomical structures and imaging modalities. Each row corresponds to a different representative case or slice. The additional results

complement the quantitative comparisons in the main paper and provide a more detailed visualization of boundary localization and structural consistency.

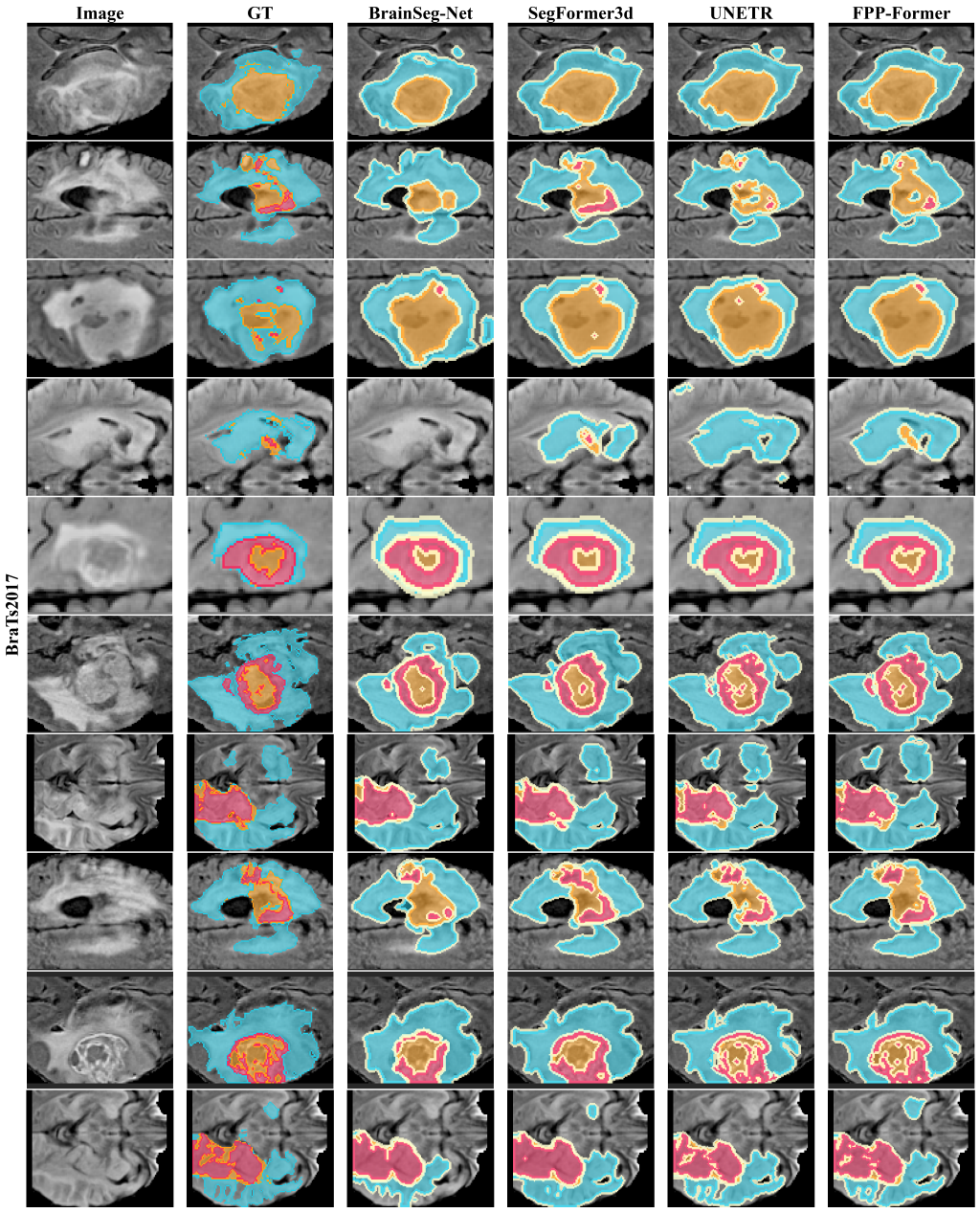


Figure 5: Additional qualitative comparison on BraTS 2017. From left to right: input image, ground truth (GT), BrainSeg-Net, SegFormer3D, UNETR, and FPP-Former. Different rows show representative tumor cases with varying morphology, location, and extent.

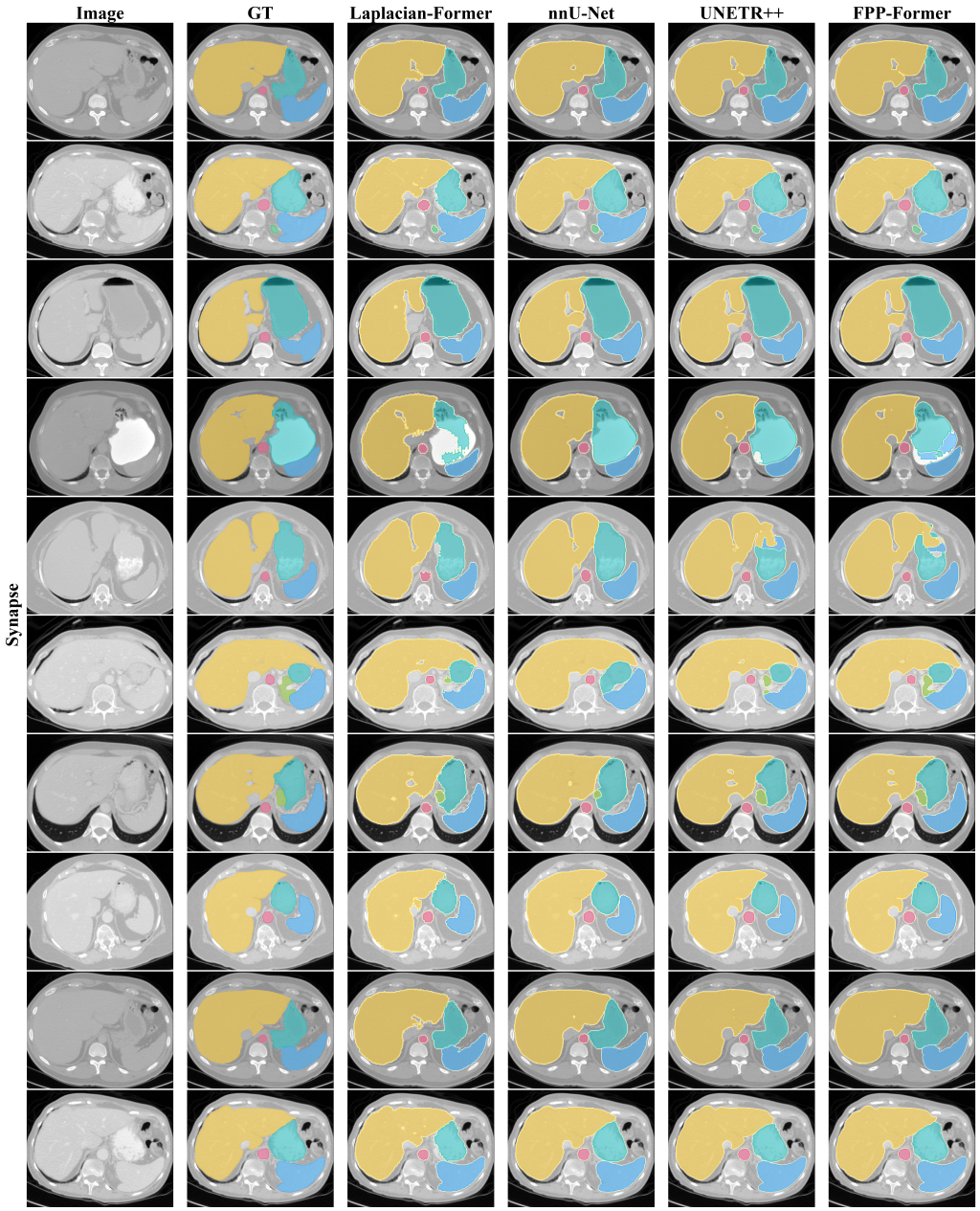


Figure 6: Additional qualitative comparison on the Synapse dataset. From left to right: input image, ground truth (GT), Laplacian-Former, nnU-Net, UNETR++, and FPP-Former. The examples cover different abdominal anatomies and illustrate the segmentation results for multiple organs with diverse shapes and spatial configurations.

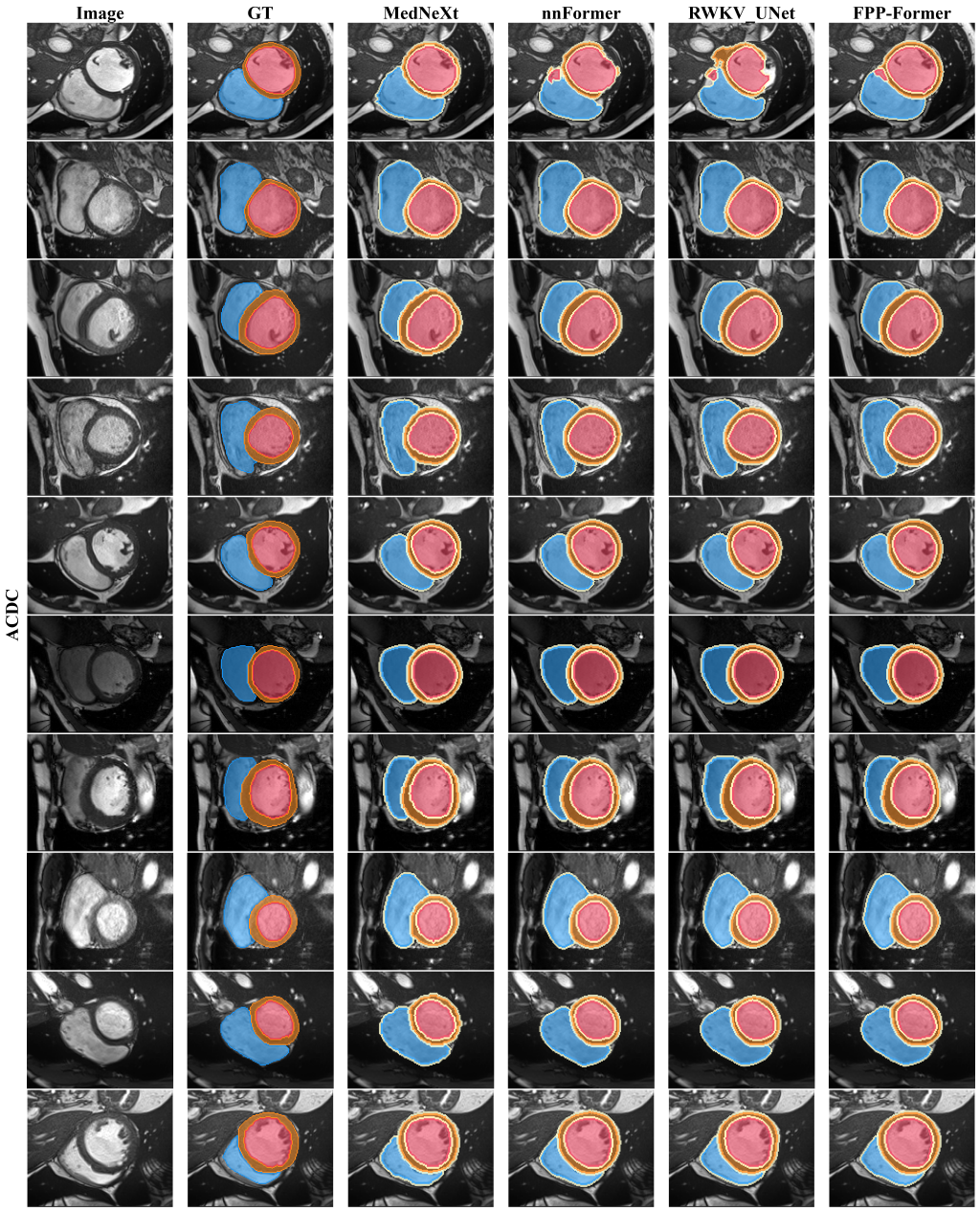


Figure 7: Additional qualitative comparison on the ACDC dataset. From left to right: input image, ground truth (GT), MedNeXt, nnFormer, RWKV-UNet, and FPP-Former. The selected cases exhibit substantial variations in cardiac morphology and demonstrate the segmentation of the left ventricle, right ventricle, and myocardium.