MoCaf-Mamba: Modality Completion and Alignment in Feature Space for Missing-Modality Segmentation

Yongsong Zhou[1], Gui Wang[2], and Linlin Shen[3⋆]

[1] School of Computer Science and Software Engineering, Shenzhen University
[2] University of Nottingham, Nottingham, United Kingdom
[3] School of AI, Shenzhen University Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University
llshen@szu.edu.cn

Abstract. Multimodal medical image segmentation is often challenged by two factors: (i) missing modalities that disrupt cross-modal feature consistency, and (ii) residual inter-modality misalignment that degrades feature fusion. To this end, we propose MoCaf-Mamba (Modality Completion and Alignment in feature space), a Mamba-based framework that unifies feature-space modality completion and deformable feature alignment for missing-modality segmentation. Specifically, the **M**odality **C**ompletion in **F**eature **S**pace (**MCFS**) predicts missing-modality representations from available modalities and is trained with explicit reconstruction constraints to preserve segmentation-relevant semantics. The **D**eformable **F**eature **A**lignment (**DFA**) is integrated hierarchically into the encoder and predicts a bounded dense displacement field for feature-level warping, mitigating geometric discrepancies during fusion. Finally, the **S**pace **T**oken **M**ixer (**STM**) aggregates completed and aligned multi-scale features via a dual-branch design with a global consensus token, producing a robust representation for decoding under arbitrary missing patterns. Experiments on MM-WHS, BraTS2023, and Prostate158 show that MoCaf-Mamba consistently improves missing-modality segmentation and achieves state-of-the-art results across diverse missing settings. The code is available at: https://github.com/wekiqs/MoCaf-Mamba

Keywords: Missing-Modality Segmentation · Feature-Space Modality Completion · Deformable Feature Alignment · Multimodal Medical Image Analysis

1 Introduction

Multimodal medical imaging provides complementary evidence for accurate diagnosis and treatment planning [3]. For example, CT/MRI, multi-sequence MRI, and multiparametric MRI are widely used for heart, brain, and prostate segmentation, respectively [27, 2, 1]. In practice, however, one or more modalities are frequently unavailable due to protocol constraints, motion/artifact corruption, or

⋆ Corresponding author

cost, resulting in arbitrary missing-modality inputs at inference [10]. Designing segmentation models robust to such incomplete observations is crucial.

Existing missing-modality methods broadly fall into three categories: (i) *static filling* (e.g., mmFormer [24], IM-Fuse [16]), which is efficient but causes feature-statistics mismatch and loses modality-specific semantics; (ii) *learnable placeholders* (e.g., M3Ae [12]), which reduce discontinuities but are content-agnostic; and (iii) *pixel-level reconstruction* (GAN/diffusion/VAE) [22, 25], which is computationally heavy and may not align with segmentation-relevant objectives. Motivated by these limitations, we argue that effective missing-modality segmentation should *complete* missing information in feature space and *align* heterogeneous modalities before fusion. Feature-space completion targets task-relevant semantics without expensive synthesis, while alignment mitigates residual geometric discrepancies that accumulate across scales and blur boundaries.

Based on this principle, we propose **MoCaf-Mamba** (**Mo**dality **C**ompletion and **A**lignment in **f**eature space), a Mamba-based multi-scale framework unifying feature-space modality completion and deformable feature alignment. We design three modules: (1) **M**odality **C**ompletion in **F**eature **S**pace (**MCFS**) predicts missing-modality representations conditioned on available ones with reconstruction supervision. Its completion backbone uses bidirectional Mamba-style state-space blocks [8] to capture long-range spatial dependencies over tokenized feature sequences. (2) **D**eformable **F**eature **A**lignment (**DFA**) performs hierarchical feature-level alignment by predicting a bounded dense displacement field and warping features via differentiable sampling, reducing inter-modality discrepancies before fusion. (3) **S**pace **T**oken **M**ixer (**STM**) aggregates completed and aligned multi-scale features with a dual-branch design combining convolutional local cues and token-based global context via a consensus token, enabling robust decoding under arbitrary missing patterns.

Our contributions are three-fold: (1) We propose MoCaf-Mamba, a unified framework for feature-space completion and alignment in missing-modality segmentation. (2) We instantiate this with three modules: MCFS, DFA, and STM. (3) Experiments on BraTS2023, Prostate158, and MM-WHS show consistent improvements over recent baselines across diverse missing settings.

2 Related Work

Missing-Modality Learning. Existing methods mainly adopt robust fusion or reconstruction. Robust-fusion models such as mmFormer [24] and IM-Fuse [16] process available modalities with masking/zero-filling, improving tolerance to missing inputs but often without explicitly recovering modality-specific semantics from cross-modal correlations. Learnable placeholders (e.g., M3Ae [12]) reduce discontinuities yet remain content-agnostic. Reconstruction-based approaches synthesize missing modalities using GANs/diffusion/VAEs [5, 28, 14]; however, pixel-level objectives can be computationally heavy and may not align with segmentation-relevant representations. In contrast, we perform *feature-space* completion to recover missing-modality semantics efficiently.

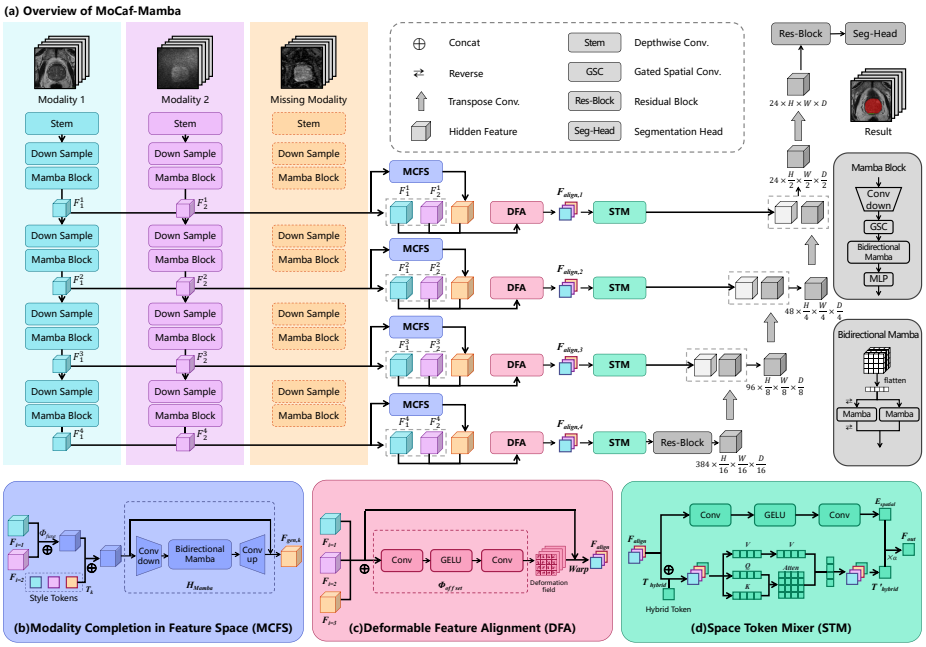

Fig. 1. Overview of MoCaf-Mamba. Given arbitrary missing-modality inputs (e.g., T2, DWI), MCFS completes missing latent features (e.g., ADC). DFA performs multi-scale deformable feature alignment to mitigate residual discrepancies, and STM conducts consensus-aware fusion for the UNETR-style decoder.

Multimodal Fusion and Alignment. While models such as HyPCA-Net [6] and Probabilistic Multimodal Learning [18] are true multimodal fusion architectures, they typically assume well-aligned inputs. In practice, motion or deformation causes residual misalignment, which global affine warping cannot fully address. We therefore introduce Deformable Feature Alignment (DFA) for hierarchical feature-level correction.

Visual State Space Models. Mamba-based models [8, 26] demonstrate strong long-range modeling with linear complexity. Recent segmentation variants include VM-UNet [17], U-Mamba [15], SegMamba [23], and small-lesion-oriented S^3-Mamba [20]; multimodal extensions like M-MambaS [21] also exist. However, existing multimodal methods such as DRIFA-Net [7] lack 3D missing-modality support. Our work fills this gap via feature-space completion and alignment for 3D volumetric data.

3 Method

Overview. We propose **MoCaf-Mamba**, a multi-scale framework for missing-modality segmentation that unifies *modality completion* and *feature-space alignment* under (i) incomplete multimodal inputs and (ii) residual inter-modality

discrepancies during fusion (Fig. 1). MoCaf-Mamba uses parallel bidirectional Mamba encoders [8, 26] for modality-specific feature extraction. Given inputs $\mathcal{X} = \{X_1, \ldots, X_K\}$ with an arbitrary missing subset, the pipeline follows completion–alignment–fusion: (1) **MCFS** performs *feature-space modality completion* with reconstruction supervision; (2) **DFA** performs multi-scale deformable *feature alignment* via a bounded displacement field; (3) **STM** conducts consensus-aware fusion of completed and aligned features for the UNETR-style decoder [9].

Modality Completion in Feature Space (MCFS). MCFS performs *feature-space modality completion* by predicting missing-modality features conditioned on the available modalities (Fig. 1). For a target modality k, we first fuse the available features into a compact shared context and then condition the completion backbone on a modality-specific Style Token $\mathcal{T}_k$ (one token per modality, shared across samples) to capture modality-dependent characteristics while preserving shared anatomical semantics. We treat each 3D feature map as a sequence of spatial tokens (flattened from the voxel grid), enabling efficient long-range dependency modeling in latent space.

Let $\mathcal{F}_{\text{in}} = \{F_i\}_{i=1}^{K}$ denote the encoder features, and let $m_i \in \{0, 1\}$ be an availability mask where $m_i = 1$ if modality i is present and 0 otherwise (unavailable slots zero-filled). The completed feature $F_{\text{gen},k}$ for a missing target modality k (aiming to reconstruct the ground-truth latent feature F_k) is obtained as

$$F_{\text{gen},k} = \mathcal{H}_{\text{Mamba}}\left(\Phi_{\text{fuse}}\left(\bigoplus_{i=1}^{K}(m_i \cdot F_i)\right) \oplus \mathcal{T}_k\right), \tag{1}$$

where $\oplus$ denotes channel-wise concatenation and Φ_{fuse} extracts a shared context from the fixed-dimensional ordered inputs. We implement Φ_{fuse} as a lightweight channel-mixing layer (e.g., 1×1×1 conv) to produce a compact context with the same spatial size as F_k, and employ $\mathcal{H}_{\text{Mamba}}$ as a lightweight U-shaped bidirectional state-space generator [8, 26] that preserves spatial resolution while modeling long-range dependencies for feature completion.

To supervise completion, we apply modality dropout during training and minimize the feature-level reconstruction loss

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}_{\text{drop}}|} \sum_{k \in \mathcal{M}_{\text{drop}}} \|F_{\text{gen},k} - F_k\|_2^2, \tag{2}$$

where $\mathcal{M}_{\text{drop}}$ denotes the set of dropped modalities. This objective encourages anatomically plausible, segmentation-relevant completions and stabilizes fusion under missing inputs. The training objective is $\mathcal{L} = \mathcal{L}_{seg} + \lambda_{rec}\mathcal{L}_{rec}$, where $\mathcal{L}_{seg}$ is the segmentation loss and λ_{rec} balances completion supervision.

Deformable Feature Alignment (DFA). After feature-space completion, DFA performs feature-level deformable alignment to mitigate residual inter-modality discrepancies. DFA give the concatenated features $\{F_j\}_{j=1}^{K}$, the offset predictor Φ_{offset} outputs modality-specific 3D displacement fields A_i for each modality. Each feature F_i is then warped by its own bounded field via STN [11]:

$$A_1, \ldots, A_K = \Phi_{\text{offset}}\left(\bigoplus_{j=1}^{K} F_j\right), \quad F_{\text{align},i} = \text{Warp}(F_i,\ \lambda \cdot \tanh(A_i)), \tag{3}$$

where Φ_{offset} is a shallow convolutional stack instantiated per scale, and Warp$(\cdot)$ denotes trilinear interpolation. DFA is applied hierarchically across multiple encoder resolutions because residual mismatch may occur at different semantic scales, requiring scale-specific correction. The tanh$(\cdot)$ nonlinearity bounds the offset magnitude to stabilize training and avoid overly large deformations.

Space Token Mixer (STM). STM performs consensus-aware fusion by combining convolutional local cues with voxel-wise cross-modal interactions. STM is applied on aligned multi-scale features (at selected resolutions). We explicitly define E_{spatial} as the output of the local 3D convolutional branch, which captures spatial context. To model cross-modal interactions efficiently, at each voxel v, we form K modality tokens $F^{(v)} = [f_1^{(v)}, \ldots, f_K^{(v)}]$ and prepend a single learnable Hybrid Token T_{hybrid} to build $Z^{(v)} = [T_{\text{hybrid}}; F^{(v)}]$. Crucially, self-attention [19] is applied *only* along this modality-token dimension (with a sequence length of $K+1$), rather than over the spatial voxels:

$$\text{Attn}(Z^{(v)}) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad Q, K, V = Z^{(v)}W_{q,k,v}. \tag{4}$$

We extract the updated token T'_{hybrid} and inject it back with a weighted residual:

$$F_{\text{out}} = E_{\text{spatial}} + \alpha \cdot T'_{\text{hybrid}}, \tag{5}$$

where α is learned end-to-end. Since attention is computed over only $K+1$ tokens (where K is small, e.g., ≤ 5), STM adds negligible overhead and scales linearly with spatial size.

4 Experiments

We evaluate MoCaf-Mamba on three multimodal benchmarks. *BraTS2023* includes 1,251 glioma cases with four MRI modalities (T1/T1ce/T2/FLAIR) and three tumor sub-regions (WT/TC/ET). *MM-WHS* contains 20 CT/MRI volumes for whole-heart segmentation. *Prostate158* has 139 cases with T2, DWI, and ADC maps, targeting Transition Zone (TZ) and Peripheral Zone (PZ) segmentation. We report Dice (%) averaged over all missing subsets per benchmark.

Implementation Details. We implement MoCaf-Mamba in PyTorch and train on NVIDIA V100 GPUs using RAdam [13] (lr=10^{-3}) with cosine annealing for 500 epochs (batch size=4). We use a 7:1:2 train/val/test split with standard augmentations (flip/rotation and intensity), simulating missing modalities via Bernoulli dropout ($p = 0.5$). For fair comparison, all baselines are trained from scratch under the same splits, augmentation, optimizer schedule, and missing-modality simulation protocol.

Results on BraTS2023.

Table 1 reports Dice across missing-modality subsets. MoCaf-Mamba achieves the best average performance on three targets (ET/TC/WT), with Avg Dice of 65.4%/83.6%/88.8%, and consistently outperforms mmFormer and IM-Fuse on

the subset-averaged metric. Moreover, MoCaf-Mamba remains competitive in severe missingness (single-modality inputs), indicating that feature-space completion in **MCFS** helps mitigate feature-statistics disruption under missing modalities.

Table 1. Quantitative results on the BraTS2023 dataset. The segmentation performance for Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT) is evaluated. ●/○: available/missing modality.

	FLAIR	●	○	○	○	●	●	●	○	○	○	●	●	●	○	●	
	T1	○	●	○	○	●	○	○	●	●	○	●	●	○	●	●	
	T1c	○	○	●	○	○	●	○	●	○	●	●	○	●	●	●	
	T2	○	○	○	●	○	○	●	○	●	●	○	●	●	●	●	**Avg.**
ET	M$_3$Ae [12]	47.0	47.6	73.1	51.3	50.8	73.7	54.7	73.7	53.0	73.4	73.9	55.1	73.8	73.6	73.8	63.2
	RobustSeg [4]	42.6	45.4	70.7	45.4	49.5	71.0	50.6	71.6	50.1	71.4	71.6	52.9	71.4	71.8	71.7	60.5
	SFusion [14]	**48.5**	49.2	65.4	49.8	**66.1**	58.4	**65.9**	59.1	**66.2**	57.5	66.5	58.8	58.2	66.9	67.1	60.2
	mmFormer [24]	48.2	51.0	72.0	52.0	53.6	72.8	54.4	72.6	54.7	73.0	73.2	56.0	73.3	73.3	73.3	63.6
	IM-Fuse [16]	47.1	46.4	57.8	44.6	58.3	49.8	58.1	49.0	57.8	47.6	58.1	57.8	49.9	57.9	57.9	53.2
	Ours	40.1	**52.8**	**76.2**	**52.1**	55.2	**76.7**	54.2	**75.9**	57.1	**77.3**	**75.9**	**59.2**	**76.6**	**76.2**	**76.3**	**65.4**
TC	M$_3$Ae [12]	**74.4**	71.8	**88.4**	73.5	75.9	**90.4**	75.7	88.9	74.9	90.1	**90.7**	76.2	**90.6**	90.5	91.0	82.9
	RobustSeg [4]	70.2	68.1	86.4	66.8	74.2	88.4	72.6	87.9	71.5	87.7	89.2	74.4	88.7	88.5	89.4	80.3
	SFusion [14]	73.5	**72.8**	86.5	**74.2**	**87.1**	76.9	**86.8**	77.4	**86.6**	77.8	87.5	**87.2**	78.5	87.6	87.8	81.9
	mmFormer [24]	73.4	72.1	88.3	73.8	76.4	89.6	76.0	**89.1**	75.8	89.3	90.2	77.2	89.6	89.7	90.2	82.7
	IM-Fuse [16]	72.2	71.3	84.3	72.7	84.4	74.5	84.4	74.6	84.2	74.7	84.4	84.3	75.4	84.3	84.3	79.3
	Ours	71.5	70.5	87.9	73.3	81.9	89.5	75.9	**89.1**	76.3	**90.2**	89.6	86.6	89.9	**90.7**	**91.2**	**83.6**
WT	M$_3$Ae [12]	89.6	78.9	77.9	86.5	90.4	90.5	90.8	79.6	87.6	87.7	90.5	91.1	91.2	88.0	91.1	87.4
	RobustSeg [4]	88.9	77.8	77.6	85.0	90.5	90.5	90.7	80.6	87.3	87.1	90.7	91.1	91.3	87.8	91.4	87.2
	SFusion [14]	85.2	79.5	79.1	86.8	85.5	85.2	82.4	87.1	86.9	87.5	85.8	87.4	87.6	87.2	87.9	85.4
	mmFormer [24]	**90.6**	**80.8**	79.8	**87.2**	**91.2**	**91.4**	**91.5**	82.8	88.5	88.2	91.5	**91.7**	**91.9**	88.7	91.9	88.5
	IM-Fuse [16]	83.4	78.7	78.4	85.2	84.1	84.1	80.7	85.9	85.7	85.8	84.3	86.1	86.1	85.9	86.2	84.0
	Ours	89.8	80.4	**80.1**	86.8	90.8	91.3	90.8	**87.3**	**89.5**	**88.7**	**91.8**	91.5	91.7	**89.6**	**92.1**	**88.8**

Results on Prostate158. When ADC is missing (T2+DWI subset in Table 2), MoCaf-Mamba maintains strong performance (Avg Dice: 81.9%) with only a limited drop from the full-modality setting (82.0%). Overall, MoCaf-Mamba achieves the best average Dice (75.8%), improving over RobustSeg (74.9%) and M$_3$Ae (75.2%), and is best (or tied-best) on both TZ and PZ.

Results on MM-WHS. Table 3 summarizes results on MM-WHS. MoCaf-Mamba achieves the best average Dice (78.7%), improving over RobustSeg by 2.7% under our protocol. It benefits from complementary CT/MRI cues when both modalities are available and remains competitive when one modality is missing, indicating stable fusion across input combinations.

Feature Completion and Efficiency. Figure 2 (left) shows MCFS yields high cosine similarity between zero-input and completed features in anatomically consistent regions, confirming effective cross-modal feature completion. The efficiency comparison (right) demonstrates that MoCaf-Mamba achieves favorable accuracy-efficiency trade-offs across all three datasets, with linear-scaling Mamba backbone maintaining competitive throughput despite increased parameters.

Table 2. Results on Prostate158 (Dice, %). We segment Transition Zone (TZ) and Peripheral Zone (PZ). ●/○: available/missing modality.

T2	●	○	○	●	●	○	●	●	○	○	●	●	○	●	
ADC	○	●	○	●	○	●	●	○	●	○	●	○	●	●	
DWI	○	○	●	○	●	●	●	○	○	●	○	●	●	●	
	TZ							**PZ**							**Avg.**
M_3Ae [12]	88.3	76.4	81.2	**88.4**	88.2	81.8	88.5	73.6	48.3	**57.8**	74.0	74.2	**57.1**	74.4	75.2
RobustSeg [4]	88.2	74.0	79.7	88.2	88.2	80.5	88.1	75.2	47.1	56.7	75.5	75.1	57.0	75.1	74.9
SFusion [14]	87.0	71.6	79.9	87.1	87.8	79.2	87.3	72.3	45.2	56.8	73.0	72.5	56.2	72.1	73.6
mmFormer [24]	87.6	73.4	80.5	87.3	87.2	80.6	86.7	72.2	47.8	56.2	72.4	72.4	56.5	72.4	73.8
IM-Fuse [16]	86.3	75.7	79.8	85.9	85.7	80.7	85.5	69.3	**50.0**	56.9	68.4	68.0	56.7	67.6	72.6
Ours	**88.5**	**77.2**	**82.3**	88.4	**88.6**	**82.8**	**88.5**	**75.2**	48.2	57.5	**75.5**	**75.2**	57.0	**75.5**	**75.8**

Table 3. Results on MM-WHS (Dice, %) for substructures: Left Ventricular Myocardium (LM), Left Atrium (LA), Left Ventricle (LV), Right Atrium (RA), Right Ventricle (RV), Ascending Aorta (AA), and Pulmonary Artery (PA). ●/○: available/missing CT/MRI.

CT	●	○	●	●	○	●	●	○	●	●	○	●
MRI	○	●	●	○	●	●	○	●	●	○	●	●
	LM			**LA**			**LV**			**RA**		
M_3Ae [12]	67.1	79.1	82.1	62.6	87.2	86.4	76.2	85.1	88.2	68.9	78.7	81.8
RobustSeg [4]	65.3	**81.7**	84.2	65.1	87.5	90.3	78.7	82.5	89.2	66.8	84.4	85.6
SFusion [14]	62.4	79.5	80.3	59.1	87.0	88.1	76.0	**87.7**	**90.0**	62.4	83.1	87.6
mmFormer [24]	64.8	78.0	82.0	**70.3**	87.8	89.7	78.8	77.8	86.1	70.1	82.7	83.1
IM-Fuse [16]	**68.5**	69.9	74.2	68.5	82.4	84.0	**85.9**	80.0	85.9	71.1	79.8	77.8
Ours	67.3	81.3	**85.7**	65.8	**90.1**	**90.6**	76.5	82.8	89.0	**78.2**	**87.5**	**88.5**
	RV			**AA**			**PA**			**Avg.**		
M_3Ae [12]	76.1	**83.6**	87.3	57.1	72.9	75.9	46.8	70.7	74.3	75.6		
RobustSeg [4]	78.1	75.1	85.0	49.0	70.6	78.0	49.4	73.2	75.3	76.0		
SFusion [14]	76.1	71.1	79.4	59.8	70.3	79.5	48.0	73.0	76.2	75.0		
mmFormer [24]	72.6	67.2	77.9	61.8	71.1	78.7	48.8	77.0	**78.1**	75.5		
IM-Fuse [16]	74.3	66.5	78.6	**65.6**	61.4	71.7	**51.4**	67.3	66.3	72.9		
Ours	**83.6**	70.8	**89.9**	65.1	**75.5**	**83.9**	50.9	**74.9**	75.0	**78.7**		

Ablation Studies. Table 4 reports ablations on Prostate158. Starting from the baseline (Avg Dice 74.1%), introducing **MCFS** consistently improves performance (Avg 74.7), indicating that feature-space modality completion alleviates the disruption caused by missing inputs. Adding **DFA** on top of MCFS further boosts accuracy to 75.2, with notable gains on PZ (65.0→65.8), suggesting enhanced cross-modality interaction and complementary cue aggregation. Replacing DFA with **STM** yields 75.3 Avg and the largest PZ improvement (65.0→66.1), highlighting the benefit of consensus-aware fusion for difficult structures under incomplete modalities. The **full model** achieves the best overall

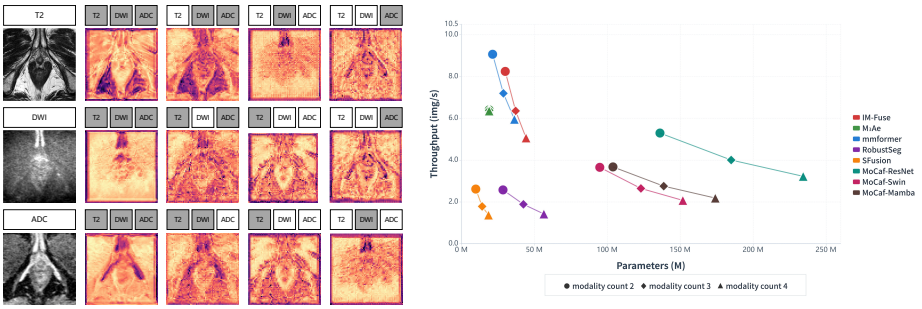

Fig. 2. Left: Cosine similarity between zero-input features and MCFS-completed features on Prostate158 under different missing-modality combinations. Gray/white boxes indicate available/missing modalities. **Right:** Throughput vs. parameter count across three datasets for 2 (●), 3 (◇), and 4 (▲) modality inputs.

Table 4. Ablation on Prostate158 (Dice, %). Top: Proposed modules. Bottom: Vision backbones.

Method	TZ	PZ	Avg.
Effectiveness of Proposed Modules			
Baseline	83.5	64.6	74.1
+ MCFS	84.3	65.0	74.7
+ MCFS+DFA	84.6	65.8	75.2
+ MCFS+STM	84.5	66.1	75.3
Full Model	**85.2**	**66.3**	**75.8**
Investigation of Vision Backbones			
ResNet3D	82.1	61.0	71.6
Swin3D	83.5	62.1	72.8
Mamba (Ours)	**85.2**	**66.3**	**75.80**

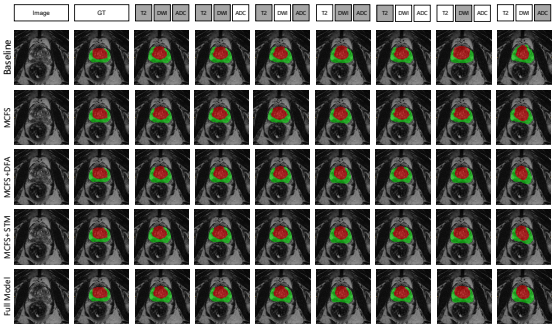

Fig. 3. Qualitative ablation on Prostate158 under different missing-modality settings. Gray/white boxes indicate available/missing modalities (T2/DWI/ADC). Red/green denote TZ/PZ, respectively.

performance (Avg 75.8), demonstrating that completion, interaction, and fusion modules provide complementary benefits. Furthermore, to justify our choice of vision backbone, we replace the Mamba encoders with a 3D CNN (ResNet3D) and a 3D Transformer (Swin3D) within the full framework. The Mamba backbone significantly outperforms both ResNet3D (Avg 71.6) and Swin3D (Avg 72.8), confirming its superior capability in modeling long-range dependencies for 3D missing-modality representation.

Qualitative Ablation Analysis. The qualitative results are consistent with the ablation trends. The baseline shows noticeable boundary leakage and fragmentation, especially on the thin PZ (green) under severe missingness. Adding MCFS yields more stable and GT-aligned contours across missing patterns, indicating that feature-space completion recovers segmentation-relevant cues from

available modalities. Further introducing DFA/STM refines fine boundaries and reduces local artifacts, improving cross-pattern consistency. Overall, the full model produces the cleanest and most accurate TZ/PZ delineations, particularly in single-modality cases.

5 Conclusion

We presented **MoCaf-Mamba**, a multi-scale framework for missing-modality segmentation that unifies *modality completion* and *feature-space alignment* under incomplete inputs and residual inter-modality discrepancies. Specifically, **MCFS** completes modalities in feature space, **DFA** provides bounded feature alignment, and **STM** ensures consensus-aware fusion for robust decoding across missing patterns. Experiments on BraTS2023, MM-WHS, and Prostate158 show that MoCaf-Mamba consistently improves performance over recent baselines, including extreme single-modality settings. These results indicate that feature-space completion and alignment provide an effective pipeline for practical scenarios with incomplete observations and subtle distortions. Future work will focus on improving DFA efficiency for real-time 3D deployment and strengthening feature completion under domain shifts.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62576216 and Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

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

References

1. Adams, L.C., Makowski, M.R., Engel, G., et al.: Prostate158 - An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. Computers in Biology and Medicine **148**, 105817 (2022)
2. Baid, U., et al. (eds.): Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation, vol. 14669. Springer (2024)
3. Basu, S., Singhal, S., Singh, D.: A systematic literature review on multimodal medical image fusion. Multimedia Tools and Applications **83**(6), 15845–15913 (2024)
4. Chen, C., Dou, Q., Jin, Y., et al.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019. vol. 11766, pp. 447–456. Springer (2019)
5. Dalmaz, O., Yurt, M., Cukur, T.: ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
6. Dhar, J., Pandey, M.K., Das Chakladar, D., et al.: HyPCA-Net: Advancing Multimodal Fusion in Medical Image Analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2026. pp. 1831–1840 (2026)
7. Dhar, J., Zaidi, N., Haghighat, M., et al.: Multimodal fusion learning with dual attention for medical imaging. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2025. pp. 4362–4371 (2025)
8. Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752 (2023)
9. Hatamizadeh, A., Tang, Y., Nath, V., et al.: UNETR: Transformers for 3D Medical Image Segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2022. pp. 1748–1758 (2022)
10. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: Hetero-Modal Image Segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2016. vol. 9901, pp. 469–477. Springer (2016)
11. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS) 2015. vol. 28 (2015)
12. Liu, H., Wei, D., Lu, D., et al.: M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 2023. vol. 37, pp. 1657–1665 (2023)
13. Liu, L., Jiang, H., He, P., et al.: On the variance of the adaptive learning rate and beyond. In: Proceedings of the International Conference on Learning Representations (ICLR) 2020 (2020)
14. Liu, Z., Wei, J., Li, R., Zhou, J.: SFusion: Self-attention Based N-to-One Multimodal Fusion Block. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023. vol. 14221, pp. 159–169. Springer (2023)
15. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv preprint arXiv:2401.04722 (2024)
16. Pipoli, V., Saporita, A., Marchesini, K., et al.: IM-Fuse: A Mamba-Based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2025. vol. 15967, pp. 225–235. Springer (2025)

17. Ruan, J., Li, J., Xiang, S.: VM-UNet: Vision Mamba UNet for Medical Image Segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2025)
18. Sastry, S., Choudhury, A., Madala, P., Su, G.M.: Probabilistic Multimodal Learning with Bayesian Disagreements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2026. pp. 7546–7555 (2026)
19. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS) 2017. vol. 30 (2017)
20. Wang, G., Li, Y., Chen, W., et al.: S^3-mamba: Small-size-sensitive mamba for lesion segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 2025. vol. 39, pp. 7655–7664 (2025)
21. Wang, G., Ren, J., Shen, L., Cheah, W.P., Qu, R.: M-MambaS: Multimodal Mamba for Small Lesion Segmentation. Pattern Recognition **180**, 113923 (2026)
22. Wang, H., Liu, Z., Sun, K., et al.: 3D MedDiffusion: A 3D Medical Latent Diffusion Model for Controllable and High-quality Medical Image Generation. IEEE Transactions on Medical Imaging **44**(12), 4960–4972 (2025)
23. Xing, Z., Ye, T., Yang, Y., et al.: SegMamba: Long-Range Sequential Modeling Mamba for 3D Medical Image Segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024. vol. 15008, pp. 578–588. Springer (2024)
24. Zhang, Y., He, N., Yang, J., et al.: mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022. vol. 13435, pp. 107–117. Springer (2022)
25. Zhao, C., Guo, P., Yang, D., et al.: MAISI-v2: Accelerated 3D High-Resolution Medical Image Synthesis with Rectified Flow and Region-specific Contrastive Loss. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 2026. vol. 40, pp. 13088–13098 (2026)
26. Zhu, L., Liao, B., Zhang, Q., et al.: Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In: Proceedings of the International Conference on Machine Learning (ICML) 2024. pp. 62429–62442. PMLR (2024)
27. Zhuang, X., Li, L., Payer, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. Medical Image Analysis **58**, 101537 (2019)
28. Özbey, M., Dalmaz, O., Dar, S.U.H., et al.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)