HK-Fuse: Hilbert-Interleaved Kimi Delta Attention for Incomplete-Modality Brain Tumor Segmentation

Weizhi Zhang[1], Shumeng Li[1], Jian Zhang[1], Lei Qi[2], Ling Lin[3], Qian Yu[4], Yuqi Fang[5], and Yinghuan Shi[1⋆]

[1] State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China
zwz@smail.nju.edu.cn, syh@nju.edu.cn
[2] School of Computer Science and Engineering, Southeast University, Nanjing, China
[3] School of Network Security and Information Technology, YiLi Normal University, Yining, China
[4] School of Data and Computer Science, Shandong Women's University, Jinan, China
[5] School of Intelligence Science and Technology, Nanjing University, Suzhou, China

Abstract. Multimodal MRI is valuable for brain tumor segmentation because different modalities highlight different tumor subregions. In clinical practice, however, complete modalities are often unavailable due to shortened protocols or when contrast cannot be used. As a result, segmentation is often performed with incomplete-modality inputs, and the key challenge is to fuse available modalities. Existing approaches often introduce long-range dependency modeling for fusion, yet two challenges remain. First, 3D features are typically flattened into a 1D token sequences for global interaction, which could weaken spatial structure. Second, missing modalities are commonly replaced by zeros during fusion. Fusion then has fewer valid modality features to aggregate, which makes the fused features sparse in effective information. In this paper, we propose **HK-Fuse**, a Hilbert-interleaved Kimi fusion framework for incomplete-modality brain tumor segmentation. HK-Fuse organizes multimodal features with 3D Hilbert scanning to preserve spatial structure. It performs cross-modal fusion with Kimi Delta Attention, which supports a selective mechanism to reduce the impact of zero-filled inputs. We further introduce Key-Voxel Routing to apply cross-modal interaction selectively on informative key voxels. Extensive experiments on BraTS2023 validate the effectiveness of HK-Fuse across all 15 missing-modality scenarios. Code is available at https://github.com/weizhizhang7/HK-Fuse.

Keywords: Brain Tumor Segmentation · Incomplete Multimodal Segmentation · Kimi Delta Attention

⋆ Corresponding author: syh@nju.edu.cn

1 Introduction

In brain tumor segmentation, complementary contrasts from multi-sequence MRI (FLAIR, T1, T1ce, and T2) provide critical cues for accurately delineating the WT/TC/ET subregions. However, complete modalities are often unavailable in clinical practice due to shortened scanning protocols, motion artifacts, or contraindications to contrast agents. Evidence from multicenter glioma cohort studies and public data resources indicates that missing MRI sequences (i.e., missing modalities) are not uncommon in clinical data [14, 5]; for example, in a multimodal glioblastoma MRI resource released by The Cancer Imaging Archive (TCIA) [8], 54 out of 315 (about 17%) initially identified MRI scans were excluded due to missing MRI sequences. Missing modalities have been widely recognized as a key practical factor that limits the robustness of brain tumor segmentation models.

Current incomplete-modality segmentation approaches generally follow two paradigms: (1) explicit modality completion, which synthesizes missing modalities via *reconstruction-based* [17, 12, 25] or transfers knowledge via *distillation-based* [2, 11, 16, 18, 21, 23]; and (2) *latent-space feature fusion* [7, 13, 27, 6, 4, 10, 19], which learns robust shared representations directly from available inputs. Recently, sequence modeling methods based on Transformers [22] or state space models (SSMs) [9] have combined long-range dependency modeling with cross-modal feature fusion to mitigate the loss of complementary information caused by incomplete inputs, achieving competitive performance under this paradigm. Specifically, mmFormer [26] models long-range correlations with Transformers, while IM-Fuse [15] adopts SSMs for efficient long-sequence modeling.

Nevertheless, applying such sequence models to 3D incomplete multimodal data still exposes two challenges. First, a straightforward design flattens the 3D features of each modality and concatenates them into 1D token sequence. This sequence fusion does not reliably preserve local spatial context for 3D segmentation [3, 24]. Second, during modality fusion, zero tokens are typically used to replace missing modalities to maintain tensor-shape consistency. This practice reduces the number of effective modality features in the fused representation, leading to sparse feature information and degrading the segmentation of small-volume tumor subregions that rely more on strongly informative modalities [1].

These limitations motivate a cross-modal fusion method that maintains spatial structure in long-range interaction and suppresses the effect of zero-filled missing-modality tokens through a finer-grained selective mechanism.

To address these challenges, we propose HK-Fuse, a Hilbert-interleaved Kimi fusion framework for incomplete-modality brain tumor segmentation. Its core component is the Hilbert-Interleaved KDA Block (HK-Block). HK-Fuse serializes multimodal tokens using 3D Hilbert scanning and arranges them in a fixed-slot, modality-interleaved ordering, which better preserves 3D spatial locality and maintains cross-modal alignment during long-range interaction. On top of this ordering, we adopt Kimi Delta Attention (KDA) [20] as the core cross-modal fusion mechanism. We perform bi-directional scanning, allowing tokens to integrate context from both directions to capture long-range dependencies. Its selective

mechanism limits the impact of zero-filled missing-modality inputs, thereby alleviating effective-information sparsity and enabling stable fusion across diverse missing-modality settings. Moreover, we introduce a Key-Voxel Routing in the high-resolution skip connections which selects Top-K informative key voxels for sparse cross-modal interaction.

Contributions:

1. We propose the HK-Block, leveraging 3D Hilbert scanning and fixed-slot organization for spatially consistent long-range cross-modal alignment.
2. We employ bi-directional Kimi Delta Attention (Bi-KDA) for fusion, enabling context aggregation from both directions while suppressing zero-filled missing-modality effects.
3. We propose Key-Voxel Routing that selects Top-K informative key voxels on high-resolution skip features for sparse cross-modal interaction, supporting more reliable delineation under missing modalities.

Evaluations on the BraTS2023 dataset, across all 15 missing-modality combinations demonstrate that HK-Fuse achieves competitive segmentation performance, with average Dice scores of 90.4/85.9/77.3 for WT/TC/ET, respectively.

2 Method

2.1 Preliminary

Kimi Delta Attention [20] is a linear attention architecture that performs token mixing through an efficient state-based update rule inspired by the Delta Rule. A key property of KDA is its channel-wise decay together with a gated write, giving fine-grained control over how much each token retains past state and writes new information. This suits our fixed-slot setting: zero-filled slots carry no modality evidence and write little, while valid-slot context is retained or selectively decayed, yielding more stable fusion under incomplete inputs.

$$\mathbf{S}_t = \left(\mathbf{I} - \beta_t \, \boldsymbol{k}_t \boldsymbol{k}_t^\top\right) \mathrm{Diag}(\boldsymbol{\alpha}_t) \, \mathbf{S}_{t-1} + \beta_t \, \boldsymbol{k}_t \boldsymbol{v}_t^\top, \tag{1}$$

where, $\mathbf{S}_t$ denotes the memory state matrix at step t and $\mathbf{S}_{t-1}$ is the previous state; $\boldsymbol{k}_t$ and $\boldsymbol{v}_t$ are the key/value vectors of the current token, and $(\cdot)^\top$ indicates transpose so that $\boldsymbol{k}_t \boldsymbol{k}_t^\top$ and $\boldsymbol{k}_t \boldsymbol{v}_t^\top$ become outer-product terms; $\mathrm{Diag}(\boldsymbol{\alpha}_t)$ constructs a diagonal matrix from $\boldsymbol{\alpha}_t$ to decay $\mathbf{S}_{t-1}$ channel-wise (forgetting); β_t is a write gate that scales the rank-1 write-in term $\boldsymbol{k}_t \boldsymbol{v}_t^\top$ and modulates the suppression/correction along the current key direction via $\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top$, where $\mathbf{I}$ is the identity matrix.

2.2 Overview

As illustrated in Fig. 1, HK-Fuse is designed as a U-shaped hybrid architecture that follows established incomplete-modality fusion frameworks [26, 15]. We define the set of modality as $\mathcal{M} = \{\mathrm{FLAIR}, \mathrm{T1}, \mathrm{T1ce}, \mathrm{T2}\}$, with $M = |\mathcal{M}| = 4$.

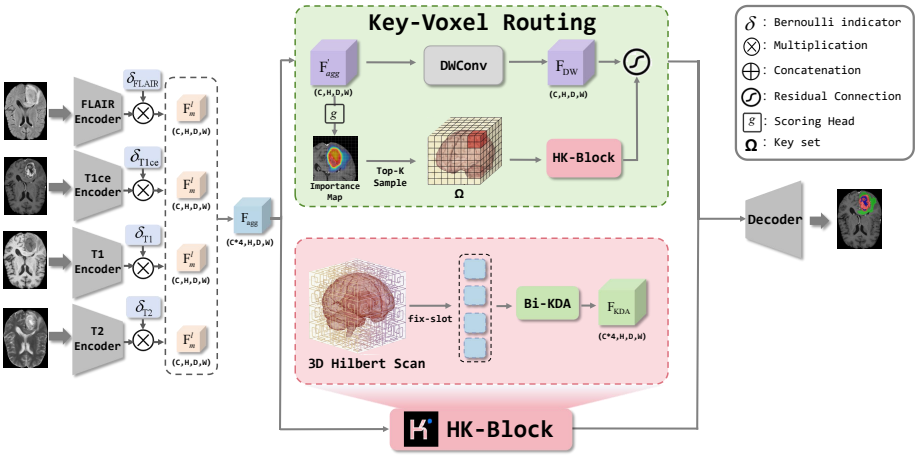

Fig. 1: Overview of the proposed HK-Fuse architecture.

We also define a fixed order of modality-slots $\pi = [m_1, m_2, m_3, m_4]$ (consistent with the order of the input channel) and allocate one slot per m_j for every voxel. Given the multimodal input set $X = \{\mathrm{X}_m \mid m \in \mathcal{M}\}$, each modality-specific volume is denoted as $\mathrm{X}_m \in \mathbb{R}^{H\times W\times D}$, where $H \times W \times D$ represents the spatial dimensions. To preserve modality-specific features, we employ M parallel non-shared convolutional encoders to extract hierarchical representations. Following mmFormer [26], we introduce a Bernoulli indicator $\delta_m \in \{0, 1\}$ during training to simulate missing modalities. Specifically, $\delta_m = 1$ indicates that modality m is present; if modality m is missing ($\delta_m = 0$), its input is zero-filled before fusion. Subsequently, cross-modal fusion is performed at the bottleneck via the HK-Block. The fused representation is mapped back to refine features for subsequent decoding. For skip connections, we introduce Key-Voxel Routing to conduct cross-modal fusion at selected key-voxel locations. Finally, a shared convolutional decoder progressively aggregates the bottleneck and multi-scale skip features to generate voxel-wise segmentation predictions.

2.3 Hilbert-interleaved KDA Block

Let $\mathrm{F}_m^l \in \mathbb{R}^{C_l\times H_l\times W_l\times D_l}$ denote the stage-l encoder output feature map for modality $m \in \mathcal{M}$, where $l \in \{1, 2, \ldots, 5\}$, C_l is the channel dimension, and (H_l, W_l, D_l) are the spatial dimensions; we apply the HK-Block only at the bottleneck stage ($l = 5$), while Key-Voxel Routing applies the same fusion operator at selected voxel locations on higher-resolution skip stages ($l \in \{1, 2, \ldots, 4\}$).

3D Hilbert-Interleaved Fixed-Slot. To better preserve 3D spatial locality during long-range interaction, we serialize bottleneck voxel features using a 3D Hilbert scan, denoted by a permutation index $\mathcal{I}_{\mathrm{Hilbert}}$. We then form a fixed-slot, modality-interleaved sequence by placing the M modality tokens of each voxel

consecutively according to the Hilbert order. Let $N = H_l W_l D_l$ be the number of voxels, and $Z_{\text{seq}} \in \mathbb{R}^{(NM)\times C_l}$ denote the Hilbert-ordered 1D sequence where each voxel yields M modality slots in fixed order; missing slots are zero-padded to preserve alignment. This voxel-wise interleaving keeps each voxel's M slots adjacent so co-located modalities interact directly.

Bi-directional Kimi Delta Attention. We apply bi-directional KDA on the flattened sequence Z_{seq} of length $T = N \times M$. We perform two scans with shared parameters: a forward scan ($\rightarrow$) that processes tokens from step $t = 1$ to T, and a backward scan ($\leftarrow$) that processes the reversed order. Let $\hat{\mathbf{o}}_t$ denote the fused output token at position t, obtained by combining the forward and backward readouts:

$$\hat{\mathbf{o}}_t = \phi\left(\mathbf{S}_t^{\rightarrow\top}\mathbf{q}_t^{\rightarrow} + \mathbf{S}_t^{\leftarrow\top}\mathbf{q}_t^{\leftarrow}\right), \tag{2}$$

where $\mathbf{S}_t^{\rightarrow}, \mathbf{S}_t^{\leftarrow} \in \mathbb{R}^{C_l\times C_l}$ are the memory states evolved from the input token at step t, and $\mathbf{q}_t^{\rightarrow}, \mathbf{q}_t^{\leftarrow}$ are the corresponding queries. Here $\phi(\cdot)$ denotes a linear projection. As in Eq. (1), the channel-wise selective mechanism controls how past states contribute at each step in the fixed-slot layout. Finally, the resulting sequence is mapped back to the 3D grid via the inverse mapping $\mathcal{I}_{\text{Hilbert}}^{-1}$.

2.4 Key-Voxel Routing

High-resolution skip features provide fine-grained anatomical and boundary cues. Because dense cross-modal fusion over all high-resolution voxels is costly and largely redundant, Key-Voxel Routing performs sparse interaction only at a fixed-budget set of the most informative key voxels.

To identify more informative regions, we form $\mathrm{F}_{\text{agg}} \in \mathbb{R}^{4C_l\times H_l\times W_l\times D_l}$ by stacking the modality features $\{\mathrm{F}_m^l\}_{m\in\mathcal{M}}$ along channels, and compute $\mathrm{F}'_{\text{agg}} \in \mathbb{R}^{C_l\times H_l\times W_l\times D_l}$ by mask-aware averaging over the available modalities. Next, a scoring head $g(\cdot)$ composed of a $1\times 1\times 1$ Conv is applied to F'_{agg} to predict a voxel-wise importance map $s = g(\mathrm{F}'_{\text{agg}}) \in \mathbb{R}^{H_l\times W_l\times D_l}$, from which we select the Top-K key interaction index set $\Omega = \text{Top-K}(s, K)$. Here, K is fixed to 200 to keep the number of key voxels at a comparable order of magnitude across scales and to prevent high-resolution levels from dominating the selection.

Meanwhile, another branch applies a depth-wise separable convolution to F'_{agg} to produce a local refinement output $\mathrm{F}_{\text{DW}} \in \mathbb{R}^{C_l\times H_l\times W_l\times D_l}$. For voxel locations in the set Ω, we organize the corresponding features into a fixed-slot format and feed them into the HK-Block for cross-modal fusion; the fused tokens are scattered back to the original grid, yielding a refined feature map $\mathrm{F}_{\text{KDA}} \in \mathbb{R}^{C_l\times H_l\times W_l\times D_l}$. Finally, we fuse the two branches in a residual manner:

$$\mathrm{F}_{\text{KVR}} = \mathrm{F}'_{\text{agg}} + \mathrm{F}_{\text{DW}} + \mathrm{F}_{\text{KDA}}. \tag{3}$$

We feed F_{KVR} into the subsequent decoder.

Table 1: Comparison of segmentation performance (Dice Score) under 15 different missing modality patterns. The available modalities are indicated by •, and missing ones by ∘. The best results are highlighted in **bold**.

Type	FLAIR	∘	∘	∘	•	∘	∘	•	∘	•	•	•	•	•	∘	•	Avg.
	T1	∘	∘	•	∘	∘	•	•	•	∘	∘	•	•	∘	•	•	
	T1ce	∘	•	∘	∘	•	•	∘	∘	∘	•	•	∘	•	•	•	
	T2	•	∘	∘	∘	•	∘	∘	•	•	∘	∘	•	•	•	•	
WT	U-HVED	80.1	63.5	71.2	82.8	85.1	79.0	86.8	83.9	87.9	86.8	88.4	89.3	90.0	86.8	90.8	83.5
	Rob.Seg	85.3	76.1	74.7	88.7	89.3	79.7	90.6	87.3	90.8	91.2	91.6	91.4	92.0	88.7	92.2	87.3
	mmFormer	88.5	83.7	82.8	91.4	90.1	85.5	92.2	89.8	91.3	92.7	92.8	92.5	93.0	90.5	93.0	90.0
	SFusion	87.0	77.6	78.5	89.1	88.7	81.3	90.8	88.2	91.3	91.2	91.7	91.6	92.1	88.8	92.2	88.0
	M3AE	88.5	82.5	81.7	91.5	89.7	83.6	91.9	89.2	92.2	92.5	92.6	92.1	93.0	90.0	92.9	89.6
	M3FeCon	88.5	81.2	81.2	87.7	89.5	83.9	89.4	89.8	92.1	90.0	90.5	92.7	92.6	90.7	93.1	88.8
	IM-Fuse	88.7	83.7	83.0	**91.8**	90.2	85.5	92.4	**90.1**	**92.6**	92.8	92.8	**93.0**	**93.1**	90.5	**93.3**	90.2
	HK-Fuse	**89.0**	**84.4**	**83.1**	91.7	**90.6**	**86.1**	**92.5**	**90.1**	**92.6**	**93.0**	**93.1**	92.9	**93.1**	**90.7**	93.2	**90.4**
TC	U-HVED	61.7	81.9	47.1	58.9	84.8	84.9	68.0	60.2	66.5	81.9	84.4	68.9	84.4	85.3	86.0	73.7
	Rob.Seg	69.8	85.9	66.9	70.2	88.4	87.7	76.3	73.9	76.0	89.6	90.3	78.1	90.0	89.6	90.5	81.5
	mmFormer	74.5	89.2	73.6	78.3	90.6	90.1	80.6	77.5	80.0	90.7	90.9	80.8	91.0	90.8	91.0	84.7
	SFusion	74.3	86.7	70.4	74.0	89.4	88.4	77.4	76.5	77.8	88.6	89.1	78.5	89.3	89.5	89.5	82.6
	M3AE	**77.9**	89.9	75.9	76.8	90.8	90.5	79.9	78.9	79.2	90.9	91.2	79.9	**91.5**	91.0	91.5	85.1
	M3FeCon	75.9	90.1	74.1	72.3	90.7	90.6	77.1	79.5	79.1	90.9	91.2	80.9	91.2	91.4	91.1	84.4
	IM-Fuse	76.5	**90.5**	75.4	**78.8**	**91.2**	**91.2**	80.9	79.1	79.9	**91.4**	**91.6**	81.4	91.3	**91.5**	**91.5**	85.5
	HK-Fuse	77.8	90.4	**77.2**	78.7	91.1	**91.2**	**81.2**	**80.7**	**81.4**	90.9	91.4	**82.4**	91.1	91.3	91.3	**85.9**
ET	U-HVED	40.1	76.6	21.5	39.9	78.2	73.3	43.9	36.1	47.5	76.5	78.5	47.6	78.2	79.2	79.3	59.8
	Rob.Seg	50.5	80.2	44.8	50.3	82.4	81.7	55.4	53.9	57.4	82.7	83.7	59.0	82.8	82.9	83.6	68.8
	mmFormer	58.3	84.1	54.7	58.8	84.7	84.7	62.5	62.5	63.9	84.9	84.8	66.0	84.2	85.9	84.7	73.6
	SFusion	54.3	82.2	48.9	52.2	83.8	83.6	57.4	57.1	59.0	83.9	83.8	60.1	83.9	84.0	84.0	70.6
	M3AE	58.8	82.5	56.0	56.7	85.8	84.9	60.7	60.6	61.2	85.6	85.9	62.3	85.7	85.1	85.6	73.2
	M3FeCon	56.6	82.8	53.8	53.2	83.9	84.3	58.0	61.2	60.4	83.5	84.0	62.4	84.0	84.4	84.2	71.8
	IM-Fuse	59.6	83.5	56.3	59.5	84.9	84.5	63.9	63.6	64.3	85.0	85.1	67.0	85.2	85.6	85.8	74.3
	HK-Fuse	**62.0**	**87.2**	**59.6**	**62.4**	**88.2**	**88.6**	**66.7**	**65.3**	**67.2**	**88.9**	**88.9**	**68.8**	**88.9**	**88.8**	**88.8**	**77.3**

2.5 Decoder and Loss Function

We adopt a U-Net–style convolutional decoder to progressively upsample the fused representation and merge it with the corresponding skip connections, yielding the final segmentation logits. To encourage each modality-specific encoder to produce informative high-level features on its own, we follow the mmFormer training paradigm and introduce a shared-weight auxiliary decoder. This shared decoder performs tumor segmentation independently on the output of each modality encoder, while reusing the same decoder architecture as the main convolutional decoder to keep the supervision consistent. Moreover, we generate additional predictions from intermediate decoder stages by interpolating their feature maps to the full resolution and applying supervision on these outputs.

The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{m \in \mathcal{M}} \mathcal{L}_m^{\text{encoder}} + \sum_{i=1}^{L-1} \mathcal{L}_i^{\text{decoder}} + \mathcal{L}^{\text{output}}, \tag{4}$$

where $\mathcal{M}$ denotes the modality set and L is the number of decoder stages.

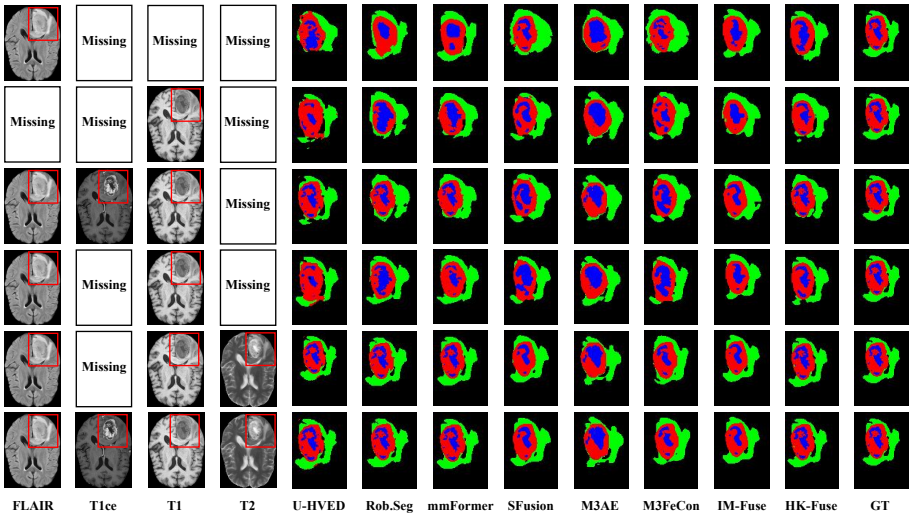

Fig. 2: Visual comparison of different segmentation methods on BraTS2023.

3 Experiments

Datasets and Implementation Details. We evaluate HK-Fuse on the BraTS2023 dataset, utilizing four MRI modalities (FLAIR, T1ce, T1, T2) to segment three tumor subregions (WT, TC, ET). The framework is implemented in PyTorch 2.5 and trained on NVIDIA A100 GPUs (80GB). Optimization is performed using RAdam and a cosine annealing scheduler with an initial learning rate of 3×10^{-4} for 1000 epochs. During training, the modality-availability mask $\delta \in \{0, 1\}^4$ is uniformly sampled from the 15 missing-modality combinations for each.

Comparison with the State-of-the-Art. We compare HK-Fuse with representative methods for incomplete-modality brain tumor segmentation, including U-HVED [7], Rob.Seg [4], mmFormer [26], SFusion [13], M3AE [12], M3FeCon [25], and IM-Fuse [15]. For fair evaluation, we follow the BraTS2023 benchmark protocol provided by [15]. Table 1 reports the quantitative results under this setting. HK-Fuse achieves the best average performance across all 15 missing-modality combinations and attains the top score in most scenarios. Over repeated training runs, HK-Fuse yields average WT/TC/ET Dice of 90.42 ± 0.07 / 85.89 ± 0.31 / 77.34 ± 0.18, indicating stable performance. Notably, on the most challenging enhancing tumor region, our method outperforms the runner-up by 3.0pp, indicating improved robustness when discriminative cues are incomplete. Furthermore, under extreme single-modality settings (FLAIR-only, T1ce-only, T1-only, and T2-only), HK-Fuse achieves the best ET performance in all four cases and remains consistently competitive on the tumor core (TC), demonstrating its ability to capture features critical for segmenting small and complex tumor structures. A per-combination analysis shows that HK-Fuse improves ET over the strongest

baseline (IM-Fuse) in all 15 settings by +1.7–4.1 pp, with the largest gains when T1ce—the dominant ET cue—is available (e.g., T1+T1ce, FLAIR+T1ce). When T1ce is missing, HK-Fuse still gains +2.5 pp on ET on average, and TC improves by +1.0–1.8 pp in five such settings, so the benefit is not limited to ET.

Qualitative visualizations (Fig. 2), generated following [15] benchmark implementation, corroborate these findings: as modalities are progressively removed, competing methods exhibit boundary leakage or false-positive spread, whereas HK-Fuse produces more spatially consistent TC/ET regions with fewer leakage artifacts and large false-positive blobs. Fig. 3 shows that HK-Fuse attains higher segmentation accuracy while maintaining a comparable parameter budget. On a single A100 (batch size 4), HK-Fuse trains in 112 h vs. 120 h for IM-Fuse, with the same 1.5 s/case inference, so the accuracy gains incur no extra runtime cost.

Ablation Study. We conduct a stepwise ablation study to evaluate the contributions of the HK-Block and KVR, with results summarized in Table 2. As components are progressively integrated, the overall performance consistently improves: adding the HK-Block to the baseline increases the average Dice by 1.4pp. Furthermore, adding KVR on top of the HK-Block (i.e., the full HK-Fuse) improves the ET Dice by 4.7pp compared with the HK-Block-only variant. These results are consistent with the role of KVR: by routing cross-modal interaction to informative voxels on high-resolution skips, it improves ET performance under missing modalities and reduces leakage artifacts in qualitative comparisons. For the routing budget, the Avg. Dice for $K = 100/200/300/400$ is 84.16/84.34/83.63/84.32; we adopt $K = 200$ as it achieves the best result, with no consistent gain from larger budgets.

Table 2: Ablation study of HK-Fuse on BraTS2023. (a) represents using only baseline, (b) adds HK-Block, and (c) adds KVR. (A) represents HK-Block and (B) represents KVR.

Method	A	B	Dice Score (%)			
			WT	TC	ET	Avg.
(a)			86.7	81.4	71.2	79.8
(b)	✓		87.9	83.0	72.6	81.2
(c)		✓	88.8	83.6	73.6	82.0
HK-Fuse	✓	✓	90.4	85.9	77.3	84.3

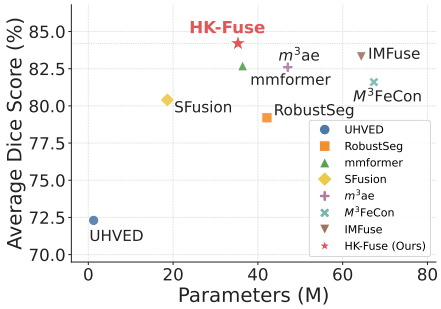

Fig. 3: Efficiency comparison of state-of-the-art methods on BraTS2023. Avg. Dice (%) vs parameters (M).

4 Conclusion

In this work, we propose a Hilbert-interleaved Kimi fusion framework HK-Fuse, for incomplete-modality brain tumor segmentation. The method targets two practical issues in long-sequence fusion: maintaining 3D spatial structure during token interaction and reducing effective-information sparsity caused by zero-filled missing-modality inputs. HK-Fuse serializes multimodal tokens with 3D Hilbert scanning using a fixed-slot, modality-interleaved organization, fuses them with bi-directional Kimi Delta Attention, and introduces Key-Voxel Routing on high-resolution skip connections to restrict interaction to Top-K informative voxels. Experiments on BraTS2023 across all 15 missing-modality combinations show that HK-Fuse achieves competitive segmentation performance under the same evaluation protocol.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Nos. 62506162, 62506005, 62192783), the Jiangsu Science and Technology Project (No. BK20251241), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), and the "111 Center" (No. B26023).

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

References

1. Abidin, Z.U., Naqvi, R.A., Haider, A., Kim, H.S., Jeong, D., Lee, S.W.: Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey. Frontiers in Bioengineering and Biotechnology **12**, 1392807 (2024)
2. Azad, R., Khosravi, N., Merhof, D.: Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. In: International conference on medical imaging with deep learning. pp. 48–62. PMLR (2022)
3. Baljer, L., Briski, U., Leech, R., Bourke, N.J., Donald, K.A., Bradford, L.E., Williams, S.R., Parkar, S., Kaleem, S., Osmani, S., et al.: Gambas: Generalised-hilbert mamba for super-resolution of paediatric ultra-low-field mri (2025)
4. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International conference on medical image computing and computer-assisted intervention. pp. 447–456. Springer (2019)
5. Chrysochoou, D., Gandhi, D.B., Adib, S., Familiar, A.M., Vunnava, B., Varshochi, S., Khalili, N., Ware, J.B., Tu, W., et al.: Ai-powered segmentation and prognosis with missing mri in pediatric brain tumors. npj Precision Oncology (2026)
6. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021)
7. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–82. Springer (2019)

8. Gagnon, L., Gupta, D., Nguyen, U., Correia de Verdier, M., Saluja, R., Mastorakos, G., White, N., Goodwill, V., McDonald, C.R., Beaumont, T., et al.: The university of california san diego post-treatment glioblastoma (ucsd-ptgbm) annotated multimodal mri dataset. Scientific Data (2026)
9. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
10. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: Hetero-modal image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (2016), https://arxiv.org/abs/1607.05194, arXiv:1607.05194
11. Karimijafarbigloo, S., Azad, R., Kazerouni, A., Ebadollahi, S., Merhof, D.: Mmcformer: Missing modality compensation transformer for brain tumor segmentation. In: Medical imaging with deep learning. pp. 1144–1162. PMLR (2024)
12. Liu, H., Wei, D., Lu, D., Sun, J.: M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 1657–1665 (2023)
13. Liu, Z., Wei, J., Li, R., Zhou, J.: Sfusion: Self-attention based n-to-one multimodal fusion block. In: International conference on medical image computing and computer-assisted intervention. pp. 159–169. Springer (2023)
14. Pemberton, H.G., Wu, J., Kommers, I., Müller, D.M., Hu, Y., Goodkin, O., Vos, S.B., Bisdas, S., Robe, P.A., Ardon, H., et al.: Multi-class glioma segmentation on real-world data with missing mri sequences: comparison of three deep learning algorithms. Scientific reports **13**(1), 18911 (2023)
15. Pipoli, V., Saporita, A., Marchesini, K., Grana, C., Ficarra, E., Bolelli, F.: Im-fuse: A mamba-based fusion block for brain tumor segmentation with incomplete modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 225–235. Springer (2025)
16. Qiu, Y., Chen, D., Yao, H., Xu, Y., Wang, Z.: Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21317–21326 (2023)
17. Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P.L., et al.: Multi-domain image completion for random missing input data. IEEE transactions on medical imaging **40**(4), 1113–1122 (2020)
18. Shi, J., Shang, C., Sun, Z., Yu, L., Yang, X., Yan, Z.: Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 456–465 (2024)
19. Shi, J., Yu, L., Cheng, Q., Yang, X., Cheng, K.T., Yan, Z.: Mftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. IEEE journal of biomedical and health informatics **28**(1), 379–390 (2023)
20. Team, K., Zhang, Y., Lin, Z., Yao, X., Hu, J., Meng, F., Liu, C., Men, X., Yang, S., Li, Z., et al.: Kimi linear: An expressive, efficient attention architecture. arXiv preprint arXiv:2510.26692 (2025)
21. Vadacchino, S., Mehta, R., Sepahvand, N.M., Nichyporuk, B., Clark, J.J., Arbel, T.: Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In: Medical imaging with deep learning. pp. 787–801. PMLR (2021)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

23. Wang, S., Yan, Z., Zhang, D., Wei, H., Li, Z., Li, R.: Prototype knowledge distillation for medical segmentation with missing modality. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
24. Wu, H., Li, H., Su, Y.: Bridging the perception-cognition gap: Re-engineering sam2 with hilbert-mamba for robust vlm-based medical diagnosis. arXiv preprint arXiv:2512.24013 (2025)
25. Zeng, Z., Peng, Z., Yang, X., Shen, W.: Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–433. Springer (2024)
26. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 107–117. Springer (2022)
27. Zhang, Z., Lu, Y., Ma, F., Zhang, Y., Yue, H., Sun, X.: Incomplete multi-modal brain tumor segmentation via learnable sorting state space model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25982–25992 (2025)