M^{4}Fuse: Lightweight State-Space MoE with a Cross-Scale Gating Bridge for Brain Tumor Segmentation

Supplementary Material

1. Additional Results

1.1. Train-Valid-Eval

At a low input resolution (**IRS=1.04M**) and with the same number of experts, Table S1.1 presents an extended analysis of the metric results for each T-V-E mode of the **BraTS2021** dataset, building upon the findings in **Table 1**.

Table S1.1. Train-Valid-Eval Mode on the BraTS2021 Dataset using *Params (M)*, *Dice (%)* and *HD95 (mm)* Metrics: Results with Input Resolution (In Res = 64×128×128) and Input Resolution Sequential length (**IRS = 1.04M**).

Model	Params	Train			Valid			Eval			Train			Valid			Eval		
		WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
3D Unet	12.34	85.06	81.03	73.58	86.72	86.67	80.62	87.91	85.61	78.22	5.21	3.92	3.33	3.72	3.21	2.43	3.81	3.32	3.11
nnUnet	31.19	88.85	89.98	81.14	86.46	87.21	83.13	89.26	87.58	84.77	4.35	2.55	2.44	4.76	2.90	2.35	3.20	2.55	2.89
TransBTS	31.65	89.03	88.19	81.34	86.16	85.54	79.94	88.43	85.78	78.58	6.80	3.97	3.30	7.72	4.33	3.08	6.63	3.61	3.03
SegResNet	18.80	90.71	92.33	84.22	88.51	90.16	85.40	89.59	89.42	82.67	2.62	2.11	1.90	3.03	2.14	2.08	3.17	2.64	2.47
SwinUNETR	62.19	90.26	88.39	79.99	88.08	84.36	78.29	89.29	86.74	78.88	3.16	2.10	1.88	3.94	2.64	2.12	3.24	2.65	2.74
SegMamba	66.85	89.36	89.12	85.31	87.59	86.68	84.17	90.48	87.54	86.67	3.96	2.84	2.69	4.45	3.12	2.31	3.97	2.55	2.12
LightM-Unet	5.02	92.44	95.59	88.27	89.11	89.01	82.99	89.99	87.27	80.45	1.59	1.43	1.34	3.40	3.00	2.89	3.47	2.48	2.34
SuperLightUnet	2.97	92.08	92.90	85.17	88.51	88.09	84.61	90.57	89.23	86.33	1.93	1.68	1.57	3.51	2.87	2.23	3.40	2.53	2.11
Top-k=1 M^{4}Fuse-T	0.29	86.94	84.43	75.52	85.41	82.49	76.81	87.01	85.97	81.73	5.97	4.30	3.82	6.00	4.02	3.24	4.08	3.18	2.56
Top-k=1 M^{4}Fuse-S	0.63	88.88	89.07	80.67	87.17	84.62	83.01	88.82	88.23	86.72	4.72	3.30	2.95	4.48	3.04	2.36	4.25	2.70	2.12
Top-k=1 M^{4}Fuse-B	1.11	89.85	93.60	85.81	87.00	88.02	83.01	89.14	89.74	87.49	3.76	2.33	2.18	3.99	2.89	2.04	3.66	2.42	1.90
Top-k=1 M^{4}Fuse-L	2.45	89.62	90.39	83.47	87.98	89.63	82.82	89.33	90.17	85.89	3.56	2.59	2.39	4.30	2.94	2.16	3.83	2.57	1.88

1.2. Five-Fold Cross-Validation

The following Tables S1.2 and S1.3 present an extended analysis of the metric results for each fold of the **BraTS2019** dataset, as reported in **Table 1**, under identical input resolution and number of experts.

1.3. Comparison of MoE Encoder Adaptation Number (Top-k)

From the M^{4}Fuse model architecture and global experimental configuration, we observe that under identical data conditions, the model extracts richer feature information from higher-resolution data, significantly enhancing segmentation accuracy. This explains why MRI and CT data modalities are more widely used than PET and other modalities. Overcoming low resolution while still achieving effective large-scale data processing remains a core challenge in medical imaging. Additionally, the expert mechanism is positioned within the encoder section of our model architecture, playing a crucial role in feature extraction. Therefore, we conducted extensive experiments across various datasets, considering factors such as different types, modalities, and structures, to validate the generalization capability of our encoder and the effectiveness of our model structure. Table S2 below details the optimal number of experts and their performance on the **BraTS2019** dataset. Moving forward, we plan to apply this approach to more datasets and tasks, pursuing lightweight implementation and further expansion.

Table S1.2. Five-Fold Cross-Validation on the **BraTS2019** Dataset using *Params (M)* and *Dice (%)* Metric: Results with Input Resolution (In Res = 128×128×128) and Input Resolution Sequential length (**IRS = 2.09M**).

Model		Params	Subset1			Subset2			Subset3			Subset4			Subset5		
			WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
3D Unet		12.34	85.32	75.83	62.53	81.83	76.06	73.36	82.52	59.74	66.23	87.77	81.56	70.51	86.19	77.16	64.86
nnUnet		31.19	86.54	79.21	75.44	87.10	81.26	77.21	87.22	80.77	77.00	87.49	78.33	76.88	88.81	79.00	77.43
TransBTS		31.65	88.99	79.26	76.91	88.01	78.51	75.44	87.02	79.01	76.12	88.46	82.16	78.43	85.12	77.21	74.21
SegResNet		18.80	90.02	84.39	73.54	89.74	82.48	78.52	87.48	82.27	75.28	89.56	84.94	77.15	86.76	76.11	66.93
SwinUNETR		62.19	89.27	80.29	72.26	88.98	82.04	78.11	87.11	81.23	76.67	89.02	81.98	77.88	88.43	78.12	74.21
SegMamba		66.85	89.27	81.62	71.21	89.29	80.84	79.52	89.46	82.64	78.91	88.70	86.23	74.25	88.82	79.91	74.67
LightM-Unet		5.02	87.60	75.71	67.14	89.80	81.41	78.07	88.53	80.11	72.70	88.69	85.22	72.56	88.90	78.57	71.71
SuperLightUnet		2.97	88.09	79.51	70.30	89.38	82.30	81.12	88.27	80.83	78.50	89.97	84.32	75.50	87.02	76.26	66.19
Top-k=2	M^{4}Fuse-T	0.32	88.10	79.31	66.20	88.69	81.97	78.56	86.96	80.96	76.70	88.73	83.52	72.41	89.10	79.27	70.91
Top-k=2	M^{4}Fuse-S	0.70	88.80	81.96	69.95	88.50	80.90	79.14	88.72	81.27	78.02	88.28	86.12	74.51	89.00	80.36	70.30
Top-k=2	M^{4}Fuse-B	1.23	89.50	82.62	71.42	89.17	81.43	80.75	89.38	82.17	79.11	89.31	86.62	74.00	89.21	80.63	70.52
Top-k=2	M^{4}Fuse-L	2.72	88.61	80.55	72.14	89.81	83.80	81.71	88.79	80.78	76.89	89.09	83.94	74.57	88.16	79.65	72.24

Table S1.3. Five-Fold Cross-Validation on the **BraTS2019** Dataset using *Params (M)* and *HD95 (mm)* Metric: Results with Input Resolution (In Res = 128×128×128) and Input Resolution Sequential length (**IRS = 2.09M**).

Model		Params	Subset1			Subset2			Subset3			Subset4			Subset5		
			WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
3D Unet		12.34	7.32	9.38	10.26	11.36	6.93	6.96	13.95	28.78	27.96	3.12	7.78	8.93	8.32	15.07	13.05
nnUnet		31.19	5.87	4.76	6.44	4.99	3.80	4.33	5.07	5.12	4.74	4.32	6.27	5.83	4.10	5.89	5.11
TransBTS		31.65	4.29	5.10	4.21	4.27	4.96	3.71	5.12	5.77	4.33	4.20	3.99	3.41	5.56	5.10	4.32
SegResNet		18.80	4.45	6.46	6.85	3.33	3.61	2.43	4.49	3.99	4.37	2.60	3.32	3.09	5.06	10.01	11.18
SwinUNETR		62.19	1.89	4.55	4.45	3.26	4.21	4.44	4.39	5.77	5.85	4.79	4.00	6.02	5.67	6.12	5.11
SegMamba		66.85	3.93	4.96	6.37	3.49	4.61	2.62	4.47	4.87	3.15	4.50	4.05	6.10	4.99	9.75	10.61
LightM-Unet		5.02	6.12	5.27	6.71	4.00	5.53	2.17	6.45	6.35	5.92	4.30	3.78	3.41	2.54	4.95	5.15
SuperLightUnet		2.97	7.12	8.95	9.55	4.03	3.67	3.05	4.14	5.37	3.84	2.68	3.11	2.80	5.73	10.68	7.42
Top-k=2	M^{4}Fuse-T	0.32	2.70	4.71	5.41	3.59	4.46	2.86	4.98	6.45	3.12	3.36	4.00	4.49	3.90	4.81	4.50
Top-k=2	M^{4}Fuse-S	0.70	6.04	7.26	7.38	5.01	4.25	2.74	6.93	6.70	3.33	4.55	2.96	3.59	5.41	5.52	4.93
Top-k=2	M^{4}Fuse-B	1.23	4.24	5.12	5.45	4.37	4.75	2.56	4.56	3.66	3.29	4.39	2.95	4.81	3.16	6.73	7.65
Top-k=2	M^{4}Fuse-L	2.72	2.51	5.26	5.50	3.21	3.04	1.99	4.38	4.94	4.30	2.22	3.06	3.01	4.20	6.23	6.73

Table S2.1. Comparison of Different Top-k(s) Based on the **BraTS2019** Dataset (**IRS=2.09M**).

#	*Top-k	*Params (M)*	*FLOPs (G)*	*Dice (%)*	*HD95 (mm)*	*TM (M)*	*IM (M)*	*TT (ms)*	*IT(ms)*
0	Structures (HGG/LGG-Top-k=2)	1.23	191.97	82.38	4.51	5163	4797	166	68
1	Modalities (MRI-Top-k=1)	1.11	191.97	79.26	7.30	5163	4797	229	55
2	Types (F/T1/T1ce/T2-Top-k=4)	1.48	186.53	56.42	19.92	5162	4775	312	129
3	else (Multi-dataset and multi-task)	-	-	-	-	-	-	-	-

Table S2.2. Aligned efficiency benchmarking (same protocol).

Method	Params(M)↓	GFLOPs↓	Mem(GB)↓	Lat p50(ms)↓	Lat p90(ms)↓
LightM-UNet	5.02	92.44	25.45	224.99	225.41
SuperLightUNet	2.97	47.06	9.79	51.11	57.78
M^{4}Fuse-B	1.23	191.97	5.33	74.56	74.87

1.4. Quantitative Evaluations

To comprehensively evaluate the overall performance of the model, the following Table presents the quantization results of the lightweight model M^{4}Fuse-B on the **BraTS2019** dataset. This section serves as an extension of **Table 2**.

Table S3. Quantitative Evaluation on the **BraTS2019** Dataset with Input Resolution of 128×128×128 (**IRS=2.09M**), Including Metrics: Parameter Count, GFLOPs (Giga Floating-Point Operations Per Second), Dice Coefficient, HD95 (95th Percentile Hausdorff Distance), Training Memory (TM), Inference Memory (IM), Training Time (TT), and Inference Time (IT).

	Model	*Params (M)*	*FLOPs (G)*	*Dice (%)*	*HD95 (mm)*	*TM (M)*	*IM (M)*	*TT (ms)*	*IT (ms)*
	3D Unet	12.34	497.16	75.42	9.48	6655	5936	222	74
	nnUnet	31.19	192.41	81.31	5.10	14740	3814	168	45
	TransBTS	31.65	284.36	80.99	4.55	5555	5486	192	50
	SegResNet	18.80	587.00	81.69	4.83	5636	4939	270	82
	SwinUNETR	62.19	774.98	81.57	4.70	30475	13606	503	162
	SegMamba	66.85	-	82.53	5.23	-	-	-	-
	LightM-Unet	5.02	92.44	80.44	4.84	26509	25760	684	236
	SuperLightUnet	2.97	47.06	81.16	5.47	17080	9635	194	59
Top-k=2	M^{4}Fuse-T	0.32	51.45	80.75	4.21	3188	2747	98	41
Top-k=2	M^{4}Fuse-S	0.70	110.38	81.72	5.28	4179	3777	135	56
Top-k=2	M^{4}Fuse-B	1.23	191.97	82.38	4.51	5163	4797	166	68
Top-k=2	M^{4}Fuse-L	2.72	423.19	82.04	4.03	7140	6849	249	105

2. Quantitative Complexity with Proofs

Preliminaries. Let $x \in \mathbb{R}^{C\times D\times H\times W}$ and $L := DHW$. A U-shaped pyramid has scales $s = 1, \ldots, 5$ with voxel counts $L_s = L/2^{3(s-1)}$ under $2\times$ down-sampling per axis. Channel widths at scale s are C_s. We count multiply-adds as 2 FLOPs. Bias terms do not change the asymptotics.

2.1. Baseline building blocks (closed forms and proofs)

Lemma 2.1 (3D convolution)**.** *A $k \times k \times k$ 3D convolution mapping $C_{\text{in}} \to C_{\text{out}}$ over L_s voxels has*

$$P_{\text{conv}} = k^3 C_{\text{in}} C_{\text{out}}$$

$$\mathcal{C}_{\text{conv}} = 2k^3 C_{\text{in}} C_{\text{out}} L_s$$

$$\mathcal{M}_{\text{conv}} = \Theta(C_{\text{out}} L_s)$$

Proof. Each output channel uses a $k^3 C_{\text{in}}$ kernel. Summing over C_{out} gives $k^3 C_{\text{in}} C_{\text{out}}$ parameters. At each voxel, convolution performs $k^3 C_{\text{in}}$ multiplies and the same number of additions per output channel, hence $2k^3 C_{\text{in}} C_{\text{out}}$ FLOPs per voxel and $2k^3 C_{\text{in}} C_{\text{out}} L_s$ over all voxels. Activations scale with the number of output features, i.e., $\Theta(C_{\text{out}} L_s)$. □ □

Lemma 2.2 (MHSA at scale s)**.** *Let token length $n = L_s$ and width C_s. Multi-head self-attention (with a constant number of heads) satisfies*

$$P_{\text{attn}} = \Theta(C_s^2)$$

$$\mathcal{C}_{\text{attn}} = \Theta(n^2 C_s) + \Theta(n C_s^2)$$

$$\mathcal{M}_{\text{attn}} = \Theta(n^2) + \Theta(n C_s)$$

Proof. Linear projections Q, K, V each cost C_s^2 parameters, as does the output projection: $\Theta(C_s^2)$ in total. FLOPs: projections cost $\Theta(nC_s^2)$. Score matrix $QK^\top$ costs $\Theta(n^2 C_s)$, softmax is $\Theta(n^2)$, and AV (attention-times-value) costs $\Theta(n^2 C_s)$. Dominant terms give $\mathcal{C}_{\text{attn}} = \Theta(n^2 C_s) + \Theta(nC_s^2)$. Storing the $n \times n$ attention plus token features gives $\Theta(n^2) + \Theta(nC_s)$ memory. □ □

Lemma 2.3 (Single-group SSM (Mamba-like) scan)**.** *For width C, state size d, local convolution span k,*

$$P_{\text{ssm}} = \Theta(Cd) + \Theta(d^2) + \Theta(kd)$$

$$\mathcal{C}_{\text{ssm}} = \Theta(LC) + \Theta(Ld) + \Theta(LC_{\text{out}})$$

$$\mathcal{M}_{\text{ssm}} = \Theta(LC_{\text{out}})$$

Proof. A linear state-space update $h_{t+1} = Ah_t + Bx_t$ and readout $y_t = Ch_t$ uses $B \in \mathbb{R}^{d\times C}$ and $C \in \mathbb{R}^{C_{\text{out}}\times d}$, contributing $\Theta(Cd) + \Theta(C_{\text{out}}d)$ parameters; A is parameterized in practice as diagonal/low-rank plus a small convolutional filter (span k), so $\Theta(d^2)$ (or $\Theta(d)$) for A and $\Theta(kd)$ for the filter suffice. The selective/associative scan compiles to per-time-step $O(C + d + C_{\text{out}})$ arithmetic; summing over L steps yields the stated FLOPs and memory. □ □

2.2. Our modules with formal savings

PetaloMixer (grouped and weight-shared). Channels are split into g equal groups of width C/g. A *single* SSM instance $\mathcal{M}$ is shared among groups, followed by one linear projection $W \in \mathbb{R}^{C\times C_{\text{out}}}$ to mix groups.

Proposition 2.4 (Exact parameter reduction by sharing)**.** *Let $P_{\text{ssm}}(w, d, k)$ denote the parameter count of one SSM with width w. A grouped-but-nonshared design uses $g \cdot P_{\text{ssm}}(C/g, d, k)$ parameters, whereas PetaloMixer uses*

$$P_{\text{petalo}} = P_{\text{ssm}}(C/g, d, k) + CC_{\text{out}}$$

Therefore the dominant SSM term is reduced by a factor $1/g$ and the only added cost is a single projection CC_{out}.

Proof. Non-sharing replicates the SSM weights g times; sharing instantiates them once. The post-group projection is applied exactly once, independent of g. □ □

CSBridge (cross-scale dual-stage gating). Per scale s, spatial gating uses a 7^3 conv on two pooled maps (Avg/Max) to one mask; channel gating uses a linear $W_s \in \mathbb{R}^{C_\Sigma\times C_s}$ with bias b_s.

Lemma 2.5 (Bridge complexity is negligible)**.**

$$P_{\text{bridge}} = \sum_{s=1}^{5} \Big(2 \cdot 7^3 + C_\Sigma C_s + C_s\Big)$$

$$\mathcal{C}_{\text{bridge}} = \sum_{s=1}^{5} \Theta(7^3 L_s) + \Theta(C_\Sigma C_s)$$

Moreover, for typical 3^3 conv blocks with width C_s, $2 \cdot 7^3 \ll 3^3 C_s^2$ and $\Theta(7^3 L_s) \ll \Theta(3^3 C_s^2 L_s)$, hence the bridge cost is a vanishing fraction of one decoder conv block.

Proof. By Lemma 2.1 with $C_{\text{in}} = 2$, $C_{\text{out}} = 1$, $k = 7$, spatial cost is constant per voxel; channel gating is a matrix-vector product per scale independent of L_s. Comparing to a 3^3 conv with $C_{\text{in}} = C_{\text{out}} = C_s$ gives the stated dominance. □ □

PEU (sample-level domain experts). At layers $\ell \in \{e4, e5, b\}$, PEU has one shared branch (K_ℓ^{sh} params) and M experts (K_ℓ params each). Top-1 routing activates only one expert per sample per ℓ.

Proposition 2.6 (Linear parameter law and constant per-sample compute)**.** *The total parameters satisfy*

$$P_{\text{peu}} = \sum_{\ell} (K_\ell^{\text{sh}} + MK_\ell) \;=\; P_{\text{base}} + M \cdot K$$

$$K := \sum_{\ell} K_\ell$$

and the expected per-sample FLOPs at each ℓ equal those of the shared branch plus one *expert, i.e., independent of M.*

Proof. Summing the branches yields the linear form in M. Top-1 routing evaluates precisely one expert per sample, hence compute does not scale with M. □ □

2.3. Whole-model accounting with upper bounds

Let $\mathcal{S}_{\text{mix}} = \{e4, e5, b, d1, d2, d3\}$ and $\mathcal{S}_{\text{br}} = \{e1, \dots, e5\}$. With light encoder convs and one PetaloMixer per decoder stage,

$$P_{\text{M}^4\text{Fuse}} = P^{\text{light}}_{\text{enc,conv}} + \sum_{s \in \mathcal{S}_{\text{mix}}} \Big(P_{\text{ssm}}(C_s/g, d, k) + C_s C^{\text{out}}_s \Big) + P_{\text{bridge}} + P_{\text{peu}}$$

$$\mathcal{C}_{\text{M}^4\text{Fuse}} = \mathcal{C}^{\text{light}}_{\text{enc,conv}} + \sum_{s \in \mathcal{S}_{\text{mix}}} \Theta(L_s C_s) + \Theta\Big(\sum_{s=1}^{5} 7^3 L_s \Big) + \mathbb{E}[\mathcal{C}_{\text{peu}}]$$

$$\mathcal{M}_{\text{M}^4\text{Fuse}} = \mathcal{M}^{\text{light}}_{\text{enc,conv}} + \Theta\Big(\sum_{s \in \mathcal{S}_{\text{mix}}} L_s C_s \Big)$$

Using $L_{e4} = L/2^9, L_{e5} = L/2^{12}, L_b = L/2^{15}$, the mixer-plus-PEU costs at high-semantic layers are bounded by $\Theta(L\bar{C}/2^9)$ for a typical width $\bar{C}$.

2.4. Dominance results against conventional designs

Proposition 2.7 (PetaloMixer vs. MHSA at the same scale)**.** *For token length $n = L_s$ and width C_s,*

$$\frac{\mathcal{C}_{\text{petalo}}(s)}{\mathcal{C}_{\text{attn}}(s)} = \mathcal{O}\left(\frac{nC_s}{n^2 C_s + nC_s^2} \right) = \mathcal{O}(n^{-1})$$

$$\frac{\mathcal{M}_{\text{petalo}}(s)}{\mathcal{M}_{\text{attn}}(s)} = \mathcal{O}(n^{-1})$$

Proof. By Lemma 2.3, $\mathcal{C}_{\text{petalo}}(s) = \Theta(nC_s)$ and $\mathcal{M}_{\text{petalo}}(s) = \Theta(nC_s)$. By Lemma 2.2, $\mathcal{C}_{\text{attn}}(s) = \Theta(n^2 C_s) + \Theta(nC_s^2)$ and $\mathcal{M}_{\text{attn}}(s) = \Theta(n^2) + \Theta(nC_s)$. Taking ratios yields $\mathcal{O}(n^{-1})$ in both FLOPs and memory since $n \geq C_s$ or $n \gg 1$ at volumetric scales. □ □

Proposition 2.8 (Replacing decoder widening by one PetaloMixer per stage)**.** *Let a baseline double decoder widths $C_{d_k} \mapsto 2C_{d_k}$ at $k \in \{1, 2, 3\}$ with two 3^3 convs per stage. Then*

$$\Delta P^{\text{base}}_{\text{dec}} = \Theta\Big(\sum_k 3^3 C^2_{d_k} \Big)$$

$$\Delta \mathcal{C}^{\text{base}}_{\text{dec}} = \Theta\Big(\sum_k 3^3 C^2_{d_k} L_{d_k} \Big)$$

Using one PetaloMixer instead gives

$$\Delta P^{\text{ours}}_{\text{dec}} = \Theta(C_{d_k} d) + C_{d_k} C^{\text{out}}_{d_k}$$

$$\Delta \mathcal{C}^{\text{ours}}_{\text{dec}} = \Theta(L_{d_k} C_{d_k})$$

If $d \ll C_{d_k}$ and $C^{\text{out}}_{d_k} \approx C_{d_k}$,

$$\frac{\Delta P^{\text{ours}}_{\text{dec}}}{\Delta P^{\text{base}}_{\text{dec}}} = \mathcal{O}\Big(\tfrac{d}{27\, C_{d_k}} \Big) \ll 1$$

$$\frac{\Delta \mathcal{C}^{\text{ours}}_{\text{dec}}}{\Delta \mathcal{C}^{\text{base}}_{\text{dec}}} = \mathcal{O}\Big(\tfrac{1}{27\, C_{d_k}} \Big) \ll 1$$

Proof. Apply Lemma 2.1 to the widened baseline and Lemma 2.3 to PetaloMixer. Compare leading terms; the constants 27 come from 3^3. □ □

Proposition 2.9 (Shared vs. non-shared grouped SSM)**.** *With g groups of width C/g,*

$$\frac{P_{\text{petalo}} - CC_{\text{out}}}{g \cdot P_{\text{ssm}}(C/g, d, k)} = \frac{1}{g}$$

$$\mathcal{C}_{\text{petalo}} = \Theta(LC) + \Theta(LC_{\text{out}})$$

i.e., exact $1/g$ parameter reduction on the SSM part, unchanged linear-time scan.

Proof. Directly from Proposition 2.4 and Lemma 2.3. □ □

2.5. Input-size regime and absolute gains (formal)

Lemma 2.10 (Scaling from 128^3 to $64\times128\times128$)**.** *Let $L_{128} = 128^3$ and $L_{64} = 64 \cdot 128 \cdot 128 = \frac{1}{2}L_{128}$. Then for MHSA, $\mathcal{C}_{\text{attn}} \propto L_s^2$ so FLOPs reduce by $\approx 4\times$ when halving one axis, while the memory term $\Theta(L_s^2)$ remains quadratic. For PetaloMixer, both FLOPs and memory scale linearly in L_s, thus reduce by exactly $2\times$.*

Proof. Substitute $n = L_s$ into Lemma 2.2 and Lemma 2.3, then apply $L_{64}/L_{128} = 1/2$. □ □

2.6. Master bound and Pareto implication

Theorem 2.11 (Linear-time, linear-parameter regime of M^{4}Fuse)**.** *Under the placements $\mathcal{S}_{\text{mix}} = \{e4, e5, b, d1, d2, d3\}$ and $\mathcal{S}_{\text{br}} = \{e1, \ldots, e5\}$, and top-1 PEU at $\{e4, e5, b\}$, the dominant costs satisfy*

$$P_{M^4Fuse} = \Theta\Big(\sum_{s\in\mathcal{S}_{\text{mix}}} C_s d\Big) + \Theta\Big(\sum_{s\in\mathcal{S}_{\text{mix}}} C_s C_s^{\text{out}}\Big) + P_{\text{bridge}} + P_{\text{base}} + M \cdot K$$

$$\mathcal{C}_{M^4Fuse} = \Theta\Big(\sum_{s\in\mathcal{S}_{\text{mix}}} L_s C_s\Big) + \Theta\Big(\sum_{s=1}^{5} 7^3 L_s\Big) + \mathbb{E}[\mathcal{C}_{\text{peu}}]$$

$$\mathcal{M}_{M^4Fuse} = \Theta\Big(\sum_{s\in\mathcal{S}_{\text{mix}}} L_s C_s\Big)$$

In particular, the model avoids any $\Theta(L_s^2 C_s)$ or $\Theta(C_s^2 L_s)$ term. Hence under both 128^3 and $64\times128\times128$ regimes, the accuracy–parameter trade-off is Pareto-optimal against (i) MHSA-based decoders and (ii) symmetric wide 3D-CNN decoders.

Proof. Summing Lemma 2.3 over mixer placements and Lemma 2.5 over scales yields the stated linear forms in L_s and C_s. Proposition 2.6 gives linear parameters in M and constant per-sample expert compute. Absence of attention blocks removes $\Theta(L_s^2 C_s)$; not widening decoders removes $\Theta(C_s^2 L_s)$. Therefore the master bounds hold and imply the stated Pareto property relative to Proposition 2.7 and Proposition 2.8. □ □

3. More Discussions

Potential Impacts and future work. Based on the model structure and training process of M^{4}Fuse, two key characteristics can be identified: the input of low-resolution data and the introduction of an expert mechanism. Resolution is often the core distinction between different modalities in medical imaging. MRI and CT are widely used in clinical practice and intelligent post-processing due to their ability to provide high-resolution 3D detailed features, while low-resolution modalities such as PET are usually used in combination with CT and MRI (PET-CT/PET-MRI). Addressing the low-resolution issue is therefore a critical task in medical imaging. Under the premise of limited hardware and other technical conditions, it is highly necessary to design more advanced neural networks to enhance adaptability to low-resolution data. In the encoding stage, M^{4}Fuse incorporates an expert mechanism to maximize the utilization of diverse data structures, modalities, and types for accurate low-resolution feature extraction, achieving excellent performance. In future work, we will continue to explore diverse data and different tasks to conduct generalization tests on lightweight models, aiming to maximize the level of intelligent processing for low-resolution images through optimized lightweight performance.

Limitations. Currently, M^{4}Fuse is primarily deployed for brain tumor image segmentation under a supervised learning paradigm. Given the inherent characteristics of the open-source BraTS dataset and other relevant benchmarks, over-reliance on labeled data alone is vastly insufficient, leading to suboptimal data utilization that hinders effective model training. In future work, we will extend our research to semi-supervised and unsupervised learning frameworks, while broadening the model's applicability to the intelligent processing of medical images across diverse anatomical regions.