Bridging Heterogeneous Medical Datasets via Mixture-of-Specialists Adapters for Unified Medical Image Classification

Shixing Huang

University of Sydney, Sydney, NSW, Australia
shixingm3@gmail.com

Abstract. Building a single classifier that unifies many fragmented medical imaging datasets is challenged by their intrinsic heterogeneity, which spans diverse clinical tasks and acquisition formats, including both 2D and 3D modalities. Although recent foundation models seek to bridge these domains, they typically rely on costly full-model pre-training. Moreover, naïvely aggregating heterogeneous datasets often leads to negative transfer caused by feature entanglement, where conflicting visual priors (*e.g.*, texture-dominant pathology versus shape-driven radiology patterns) compete within a shared representation space. In this paper, we propose MOSAIC (**M**ixture-**o**f-**S**pecialists **A**dapter for **I**maging **C**lassification), a parameter-efficient framework that unifies 18 medical datasets spanning six imaging modalities using a backbone pre-trained on natural images (ImageNet) and kept frozen during adaptation. MOSAIC introduces a Mixture-of-Specialists (MoS) mechanism with explicit modality-aware routing, which structurally decouples texture-dominant, shape-driven, and volumetric feature subspaces. By separating modality-sensitive representations while retaining shared general knowledge, this design effectively mitigates cross-domain interference arising from multi-modal aggregation. Extensive experiments demonstrate that MOSAIC achieves an average accuracy of 84.16%, matching or surpassing single-task specialist models while requiring only 7.40M trainable parameters. External validation on previously unseen 2D and 3D domains further confirms its generalization capability. Overall, our findings suggest that explicit feature decoupling within a frozen backbone provides a scalable and efficient pathway toward unified multi-modal medical image analysis. Our code and supplementary results are available at: https://github.com/BurkeBelle/MOSAIC-MedMNIST.

Keywords: Medical Image Analysis · Multi-modal Learning · Parameter-Efficient Fine-Tuning · Mixture-of-Specialists · Vision Transformer

1 Introduction

The success of deep learning in computer vision has been largely driven by large-scale, unified datasets such as ImageNet [4]. In contrast, medical imaging data remains highly fragmented across institutions, modalities, and clinical tasks,

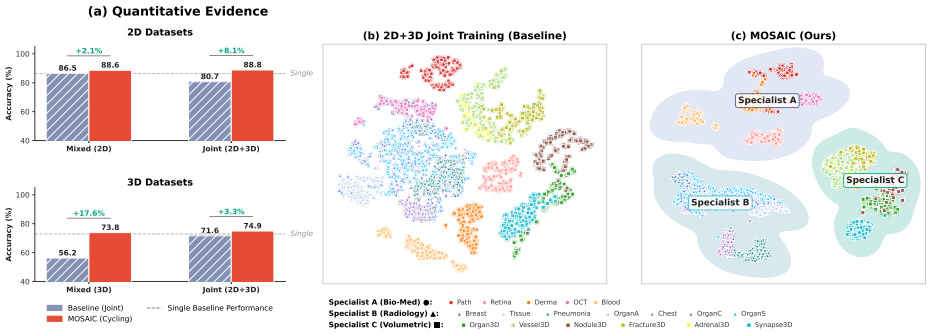

Fig. 1. Motivation: resolving cross-modal interference via dimensionality- and appearance-aware specialization. (a) Quantitative evidence: single-dataset training (dashed line) defines the upper-bound performance reference. Naïve joint 2D-3D training induces negative transfer (blue), with pronounced degradation on 3D tasks (56.2%). In contrast, MOSAIC (red) restores performance and consistently achieves positive transfer across domains. (b-c) t-SNE visualization: (b) The joint baseline shows pronounced feature entanglement; (c) MOSAIC routes inputs to dedicated Specialists, yielding clearer subspace separation.

owing to privacy constraints and the substantial cost of expert annotation [16]. To mitigate this fragmentation, recent efforts have aggregated heterogeneous datasets to develop unified medical AI models [12]. For diagnostic classification, domain-specific benchmarks include RadImageNet [15] for 2D imaging and Med3D [3] for volumetric data, while frameworks such as UniMiSS [22] and MedCoSS [24] explore joint 2D-3D representation learning through large-scale pretraining. Despite their progress, these approaches typically require full-model optimization and extensive computational resources, limiting scalability in resource-constrained clinical environments and misaligned with Green AI principles [18].

To this end, we investigate a complementary and parameter-efficient alternative: *can a frozen general-purpose vision model be efficiently adapted to heterogeneous medical imaging tasks while explicitly mitigating cross-modal interference during adaptation?* In preliminary experiments, we identify a critical challenge: naïvely mixing heterogeneous 2D and 3D datasets leads to substantial performance degradation, even falling below single-task baselines (Table 1, Fig. 1(a)). We attribute this phenomenon to feature entanglement [25]: when modalities characterized by distinct visual priors (*e.g.*, texture-dominant pathology vs. shape-driven radiology) are jointly optimized, their representations interfere within a shared embedding space, as illustrated in Fig. 1(b).

Inspired by parameter-efficient adapters [7,2] and modular mixture-of-experts (MoE) architectures [19,17], we propose MOSAIC (**M**ixture-**o**f-**S**pecialists **A**dapter for **I**maging **C**lassification), a unified framework for 2D-3D medical image analysis. MOSAIC operates on a frozen ImageNet-pretrained Vision Transformer (ViT) backbone [5], employing modality-aware tokenizers to project heterogeneous inputs into a shared token space. To address feature entanglement, we introduce

a Mixture-of-Specialists (MoS) adapter with explicit modality routing based on prior knowledge. Unlike learned gating mechanisms, which are susceptible to routing imbalance on relatively small medical datasets, our explicit routing design structurally decouples modality-sensitive subspaces, yielding the clear representation partitioning shown in Fig. 1(c), which we further quantify via CKA and weight-space analysis. In addition, we incorporate a teacher-student consistency framework [14] to stabilize optimization during joint multi-task training. Our contributions are twofold: (1) We present MOSAIC, a parameter-efficient framework that bridges 2D and 3D medical imaging by training only 7.40M parameters while specializing representations through a Mixture-of-Specialists adapter with explicit routing; (2) We conduct comprehensive evaluations on 18 MedMNIST datasets and six external benchmarks, demonstrating robust generalization that matches or surpasses single-task specialist models.

2 Method

To address the intrinsic heterogeneity of medical imaging, we propose MOSAIC, a unified framework illustrated in Fig. 2. The architecture mitigates heterogeneity at three complementary levels: (1) Spatial-Dimensional Alignment: *modality-aware tokenizers* map 2D and 3D inputs into a shared token space; (2) Feature Representation: *Mixture-of-Specialists (MoS) adapters* disentangle modality-dependent semantics; (3) Label Space: a *cyclic teacher-student framework* harmonizes disjoint task objectives during joint training.

Modality-Aware Tokenization. A fundamental challenge for unified models is mapping heterogeneous inputs into a common latent space, as 2D images and 3D volumes possess incompatible spatial dimensionality. To bridge this gap, we adopt modality-aware tokenization (Fig. 2(a)).

For 2D inputs, we retain the standard ViT tokenizer [5] with its ImageNet-pretrained weights, partitioning images into 16×16 patches ($N_{2D} = 196$ tokens). For 3D volumes, we employ a learnable volumetric tokenizer inspired by VideoMAE [21]: a volume $\mathbf{x}_{3D} \in \mathbb{R}^{64\times64\times64}$ is partitioned into $8 \times 8 \times 8$ patches, yielding $N_{3D} = 512$ tokens. The smaller 3D patch size preserves dense anatomical information along the depth axis [21,6]. Despite differing sequence lengths, both modalities are projected into an identical feature space $\mathbb{R}^{N\times C}$ with learnable modality-specific positional embeddings before entering the frozen ViT backbone.

Mixture-of-Specialists Adapter. To mitigate feature entanglement (Sec. 1), we propose the MoS adapter, which partitions adaptation into structurally decoupled subspaces (Fig. 2(b)). Each specialist is a lightweight bottleneck adapter inserted in parallel with the frozen MLP layers [2]. Given the intermediate feature $\mathbf{x}'_\ell$ at layer ℓ, the routing function $R(\cdot)$ deterministically assigns the input to specialist $k \in \{A, B, C\}$ based on its modality tag, and the output is:

$$\mathbf{x}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{x}'_\ell)) + s \cdot \mathbf{W}^k_{\mathrm{up}}\, \sigma\big(\mathbf{W}^k_{\mathrm{down}}\, \mathrm{LN}(\mathbf{x}'_\ell)\big) + \mathbf{x}'_\ell,$$

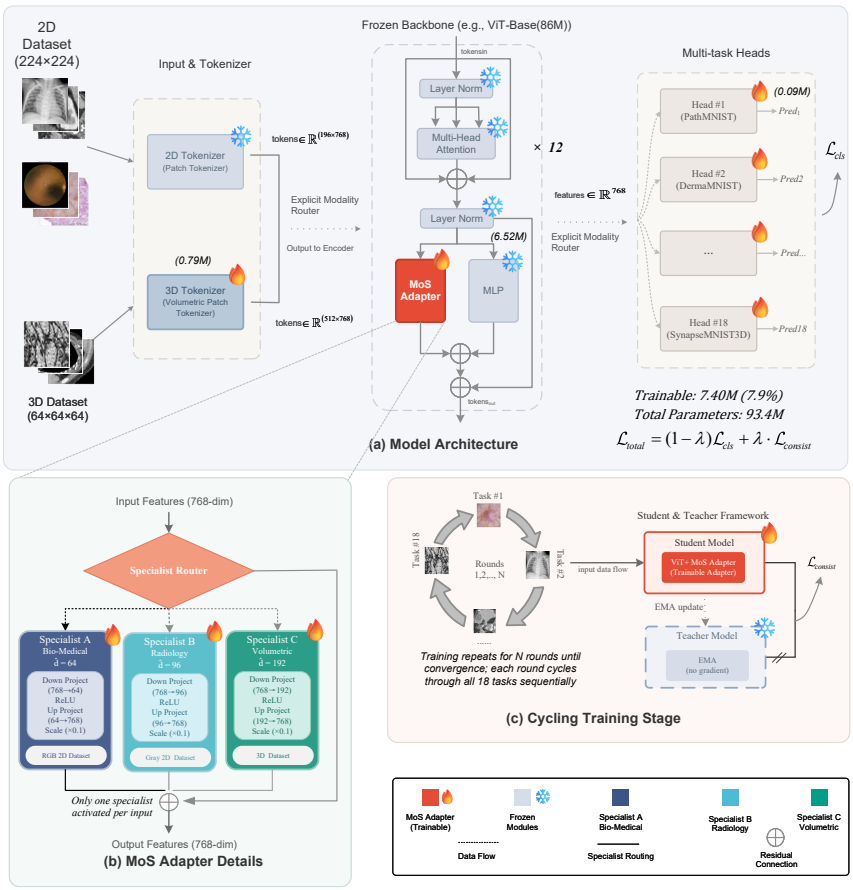

Fig. 2. Overview of MOSAIC. (a) Modality-aware tokenizers unify heterogeneous 2D–3D inputs into a shared token space, enabling a frozen ViT to extract universal visual primitives via parallel MoS adapters. (b) Explicit modality routing assigns inputs to structurally decoupled specialists with asymmetric capacities $\hat{d} = 64, 96, 192$ reflecting the intrinsic dimensionality of texture-dominant, shape-driven, and volumetric tasks. (c) A cyclic teacher-student framework stabilizes sequential optimization, where an EMA-updated teacher serves as a knowledge anchor against catastrophic forgetting.

where $\mathbf{W}^k_{\text{down}} \in \mathbb{R}^{C \times d_k}$ and $\mathbf{W}^k_{\text{up}} \in \mathbb{R}^{d_k \times C}$ are specialist-specific projections, $\sigma(\cdot)$ is ReLU, and $s = 0.1$ scales adapter outputs to prevent distorting frozen representations.

MoS differs from standard MoE in two aspects. First, routing is explicitly defined by input format (RGB, grayscale, volumetric) rather than learned gating, which is prone to imbalance on small medical datasets [19]; this deterministic design also ensures that the active specialist for any input is fully traceable, supporting clinical interpretability. We define three specialists: A (Bio-Medical,

Table 1. Performance on 18 MedMNIST datasets spanning six modalities. Naïve joint training suffers from cross-modal interference, degrading average ACC to 77.66%. In contrast, MOSAIC mitigates this negative transfer via explicit specialization, matching or surpassing single-task specialists on 13 of 18 datasets with a single unified model. Bold denotes the best performance per dataset.

Modality	Dataset	Official (ResNet-18)		Single-task (ViT-B)		Joint baseline (ViT-B)		MOSAIC (Ours)	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
2D Datasets									
Microscopy	PathMNIST	90.9	98.9	93.29±2.39	99.27±0.15	91.40±0.66	99.27±0.05	**94.71±0.57**†	**99.31±0.18**
	BloodMNIST	96.3	99.8	98.26±0.76	99.92±0.02	96.56±0.40	99.86±0.01	**99.05±0.17**†	**99.95±0.01**†
	TissueMNIST	68.1	93.3	74.16±0.22	94.56±0.12	51.80±0.89	83.61±0.31	**75.40±0.40**†	**95.76±0.10**†
Optics	DermaMNIST	75.4	92.0	84.74±1.27	97.26±0.24	78.77±2.78	95.13±0.29	84.67±0.35	97.23±0.08†
	RetinaMNIST	49.3	71.0	56.92±1.01	81.41±0.34	79.92±2.79	75.88±0.56	**82.67±1.01**	**82.79±0.17**†
	OCTMNIST	76.3	95.8	86.67±1.72	98.85±0.51	56.13±1.58	93.36±0.25	**90.80±0.78**†	**99.17±0.16**†
X-ray	ChestMNIST	94.7	77.3	94.69±0.14	80.49±1.31	94.72±0.01	72.64±0.30	**94.82±0.01**†	**81.17±0.27**†
	PneumoniaMNIST	86.4	95.6	89.69±3.07	97.91±0.41	86.75±2.16	95.97±0.44	89.21±3.22	**98.80±0.23**†
Ultrasound	BreastMNIST	83.3	89.1	86.54±1.89	92.35±1.73	81.41±2.31	86.64±0.95	84.62±1.89†	91.56±0.90†
CT	OrganAMNIST	95.1	99.8	95.74±0.46	99.71±0.09	90.41±0.26	99.40±0.02	**95.92±0.41**†	**99.89±0.01**†
	OrganCMNIST	92.0	99.4	92.70±1.01	99.60±0.08	86.78±0.21	98.92±0.06	**92.76±0.24**†	**99.68±0.02**†
	OrganSMNIST	77.8	97.4	82.72±0.33	97.92±0.07	73.88±1.00	96.96±0.14	80.69±0.41†	97.55±0.06†
2D Average		82.13	92.45	86.34±0.43	94.94±0.22	80.71±0.63	91.47±0.16	**88.78±0.28**	**95.24±0.09**
3D Datasets									
CT	OrganMNIST3D	90.7	99.6	68.63±2.03	93.90±0.51	60.16±4.30	93.10±0.76	**83.88±0.99**†	**98.90±0.18**†
	NoduleMNIST3D	84.4	86.3	86.24±1.06	86.90±0.23	80.00±0.93	71.26±1.32	81.08±0.30	80.84±1.20†
	AdrenalMNIST3D	72.1	82.7	77.74±2.09	73.55±2.40	76.84±0.39	69.30±5.27	**78.41±1.14**	**80.32±1.66**†
	FractureMNIST3D	50.8	71.2	44.31±4.06	62.37±0.35	52.22±0.96	64.16±1.73	45.00±1.77	61.50±0.91
MRI	VesselMNIST3D	87.7	87.4	87.35±1.08	71.38±1.62	88.65±0.14	72.44±0.52	88.39±0.25	**79.44±0.64**†
Microscopy	SynapseMNIST3D	74.5	82.0	73.01±0.00	53.19±0.56	71.50±2.62	53.91±2.15	72.82±1.17	**69.48±2.77**†
3D Average		76.70	84.87	72.88±0.59	73.55±0.25	71.56±1.11	70.70±1.07	**74.93±0.27**	**78.41±0.91**
Overall		80.32	89.92	81.86±0.38	87.81±0.07	77.66±0.66	84.55±0.35	**84.16±0.21**	**89.63±0.24**

† Results are statistically significant ($p < 0.05$) compared with the joint baseline (paired t-test, $n=3$).

$d = 64$) for texture-dominant RGB images; B (Radiology, $d = 96$) for shape-driven grayscale anatomy; and C (Volumetric, $d = 192$) for 3D spatial dependencies. Second, capacities are asymmetric, reflecting the intrinsic dimensionality hypothesis [8,1]: volumetric data requires larger bottlenecks to encode dense spatial semantics.

Cyclic Teacher-Student Optimization. Training on 18 datasets with disjoint label spaces and extreme data imbalance (*e.g.*, from ≈10k to >700k samples) poses two challenges: gradient dominance from large-scale tasks and catastrophic forgetting during sequential updates. We address both through a cyclic teacher-student framework [13,14] (Fig. 2(c)). The model iterates sequentially through datasets D_k, each assigned an independent linear head H_k. To stabilize this process, a teacher network updated via exponential moving average (EMA) [20] serves as a knowledge anchor, and the total objective combines task-specific supervision with a consistency loss that enforces alignment between the student's predictions on strongly augmented views and the teacher's stable outputs (*i.e.*, $\mathcal{L}_{\text{total}} = (1-\lambda)\,\mathcal{L}_{\text{task}} + \lambda\,\mathcal{L}_{\text{consist}}$) [14].

3 Experiments and Results

We evaluate MOSAIC using a ViT-Base backbone on 18 MedMNIST datasets spanning six diverse imaging modalities, and further validate on five external 2D

benchmarks [15] and one external 3D benchmark [3]. Comparisons include domain-specific baselines: RadImageNet [15], Med3D [3], and UniMiSS [22]. Performance is reported as mean ± std over three independent runs (MedMNIST), 10 runs (2D external), and five runs (3D external), using ACC and AUC.

Specialist Disentanglement. Beyond the qualitative separation in Fig. 1(c), we quantify whether the specialists learn decoupled representations. At the feature level, a CKA analysis on CLS-token features across the 18 datasets shows that Specialist C attains $\approx 4.6\times$ higher intra-group CKA than the joint baseline (0.079 vs. 0.017), indicating that deterministic routing consolidates modality-specific representations. At the weight level, the cosine similarity between specialist projection vectors is near zero (A–B: 0.008, A–C and B–C: -0.002), indicating largely independent transformations, and this separation increases with depth. The modest CKA magnitudes are expected, as adapter outputs are scaled by s=0.1 to preserve the frozen backbone [2]. Together, these feature- and weight-level analyses indicate that the decoupling extends beyond the t-SNE visualization.

Internal Validation on 18 MedMNIST Datasets. Table 1 confirms that MOSAIC achieves the highest overall performance (84.16% ACC, 89.63% AUC) across 18 datasets using a single unified model. It significantly outperforms the joint baseline overall (paired t-test across 18 datasets; $p = 0.016$ for ACC, $p < 0.001$ for AUC). Our explicit modality routing not only recovers the severe degradation caused by cross-modal feature entanglement in the joint baseline (77.66% ACC), but also surpasses the performance of isolated Single-task specialists (81.86% ACC). This demonstrates that heterogeneous data aggregation yields positive transfer rather than negative interference when modality-specific subspaces are explicitly specialized.

Within the 2D domain, the most significant gain appears in TissueMNIST, where performance recovers from 51.80% ACC (joint baseline) to 75.40% ACC. We attribute the baseline's collapse to feature entanglement between texture-dominant bio-medical and shape-driven radiological tasks [11,23]. MOSAIC resolves this by routing conflicting data to structurally decoupled specialists (Bio-Medical $\rightarrow$ A, Radiology $\rightarrow$ B), achieving a clear separation visualized in Fig. 1(c).

The 3D domain introduces an additional challenge: extreme data imbalance ($\approx$708k vs. $\approx$10k). Consistent with the data-intensive nature of Transformers [5], the joint baseline suffers from severe degradation on small-scale volumes (OrganMNIST3D ACC drops to 60.16%) due to 2D gradient dominance. MOSAIC reverses this by allocating a dedicated subspace (Specialist C), recovering ACC by $+$23.72% (to 83.88%). This design allows data-scarce 3D tasks to leverage generic visual primitives from the frozen backbone while adapters handle dimension-specific alignment [3], effectively converting 2D dominance from a source of interference into transferable knowledge. We note that the 3D branch is trained from scratch, a constraint shared by all compared Transformer baselines (Table 1).

External Validation on Unseen Domains. On 2D external benchmarks (Fig. 3(a)), MOSAIC significantly outperforms all RadImageNet variants [15] and UniMiSS on all five tasks in AUC (paired t-test, all $p < 0.01$), validating

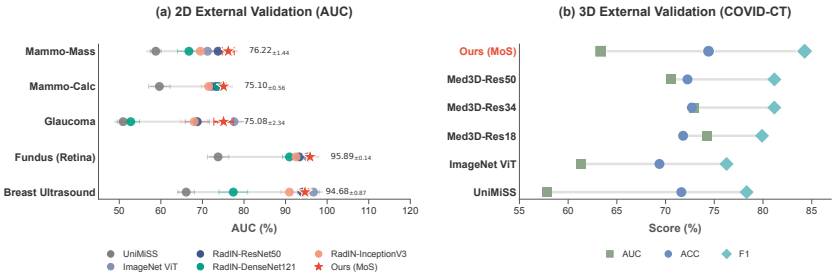

Fig. 3. External validation on unseen 2D (a) and 3D (b) domains. (a) AUC comparison across five unseen 2D benchmarks; MOSAIC (mean ± std over 10 runs) achieves the highest AUC on three of five tasks, while the ImageNet ViT (full fine-tuning) leads on BUS and Glaucoma, reflecting the inherent trade-off between parameter efficiency (7.9%) and full model adaptation. (b) COVID-CT evaluation on an unseen 3D domain: MOSAIC achieves the highest mean ACC and mean F1 over five runs among all methods. While CNN-based baselines retain higher mean AUC, this pattern is consistent across all Transformer models (ImageNet ViT, UniMiSS), suggesting an architectural inductive bias rather than a method-specific limitation.

effective transfer to both RGB (Specialist A) and grayscale (Specialist B) domains. ImageNet ViT (full fine-tuning) achieves higher AUC on BUS ($p = 0.004$) and Glaucoma ($p = 0.049$), reflecting the inherent trade-off between parameter efficiency (7.9%) and full model adaptation.

On 3D external validation (COVID-CT, Fig. 3(b)), MOSAIC achieves the highest ACC (74.41%) and F1 (84.27%) among all compared methods, demonstrating competitive classification performance with domain-specific Med3D [3] variants despite training only 7.40M parameters. However, CNN-based Med3D retains substantially higher AUC (74.29% vs. 63.34%)—a gap consistent across all Transformer-based methods (ImageNet ViT: 61.28%, UniMiSS: 57.84%), suggesting a systematic advantage of 3D domain-specific pretraining and CNN locality bias over 2D-to-3D Transformer transfer.

Parameter Efficiency. MOSAIC freezes the 86M ViT-Base backbone and trains only 7.40M parameters, comprising the volumetric tokenizer and embeddings (0.79M), MoS adapters (6.52M), and task-specific heads (0.09M). This structural consolidation serves 18 tasks from a single 93.4M checkpoint—a 16× storage reduction relative to fully fine-tuning 18 separate models (1.55B parameters). The margin narrows against parameter-efficient baselines, but MOSAIC still consolidates all tasks into one shared model, adding only a lightweight head (0.005M) per new task. Unlike full-parameter fine-tuning [22,24], MOSAIC achieves competitive performance (84.16% ACC) while preserving the backbone's universal visual primitives [7]. Crucially, this efficiency reflects how the trainable budget is organized, not its size. As Table 2 shows, at a matched budget MOSAIC still outperforms LoRA-Joint at r=192, while raising LoRA-Joint's rank from 8 to 192 lifts accuracy by less than one point. The bottleneck for joint adaptation is thus cross-task interference rather than capacity, which explicit specialization resolves

Table 2. Matched-budget comparison. *Independent*: one model per dataset (18 models); *Joint*: single shared model. LoRA-Joint at r=192 matches MOSAIC's trainable budget. Bold marks the best joint result.

Method	Training	#Models	Trainable	ACC (%)	AUC (%)
LoRA [8]	Independent	18	5.40M	$85.50_{\pm 0.05}$	$88.75_{\pm 0.34}$
VPT [9]	Independent	18	1.80M	$83.83_{\pm 0.45}$	$86.73_{\pm 1.22}$
VPT [9]	Joint	1	0.18M	$78.01_{\pm 0.24}$	$85.45_{\pm 0.48}$
LoRA (r=8) [8]	Joint	1	0.38M	$80.02_{\pm 0.42}$	$86.69_{\pm 0.60}$
LoRA (r=192) [8]	Joint	1	7.08M	$81.01_{\pm 0.25}$	$87.89_{\pm 0.73}$
MOSAIC (Ours)	Joint	1	7.40M	$\mathbf{84.16_{\pm 0.21}}$	$\mathbf{89.63_{\pm 0.24}}$

Table 3. Component-wise ablation validating each design choice. N: number of specialists. Removing the MoS adapter drops accuracy below the joint baseline (72.42% vs. 77.66%), confirming the necessity of explicit specialization. Bold marks the best result.

Method	Adapter		Cycling	Teacher	Freeze	ACC (%)		
	MoS	N				Overall	2D	3D
Joint baseline	✗	–	✗	✗	✓	$77.66_{\pm 0.66}$	$80.71_{\pm 0.63}$	$71.56_{\pm 1.11}$
w/o MoS adapter	✗	–	✓	✓	✓	$72.42_{\pm 3.71}$	$73.32_{\pm 4.32}$	$70.62_{\pm 2.52}$
w/o 3D Specialist	✓	2	✓	✓	✓	$83.49_{\pm 0.30}$	$88.06_{\pm 0.33}$	$74.33_{\pm 0.43}$
w/o Teacher	✓	3	✓	✗	✓	$83.22_{\pm 0.51}$	$88.17_{\pm 0.58}$	$73.32_{\pm 0.91}$
w/o Freeze	✓	3	✓	✓	✗	$81.36_{\pm 2.07}$	$86.07_{\pm 1.96}$	$71.94_{\pm 2.55}$
MOSAIC (Full)	✓	3	✓	✓	✓	$\mathbf{84.16}_{\pm 0.21}$	$\mathbf{88.78}_{\pm 0.28}$	$\mathbf{74.93}_{\pm 0.27}$

without additional parameters. Our asymmetric capacity allocation (Specialists A: 1.19M, B: 1.78M, C: 3.55M) reflects the intrinsic dimensionality of the data: 3D volumetric tasks necessitate higher capacity for spatial dependencies compared to 2D texture or shape cues, a strategy validated by our ablation studies.

Ablation Study. Table 3 validates the necessity of each MOSAIC component. Removing the MoS Adapter results in a substantial drop to 72.42%, even underperforming the joint baseline (77.66%). This confirms that cyclic training alone [13] cannot mitigate the cross-modal gradient interference [25] inherent in 2D-3D heterogeneity. The superiority of the 3-Specialist topology (84.16%) over the 2-Specialist version (83.49%) further justifies our expert design: merging 3D with Radiology degrades volumetric performance to 74.33%, whereas isolating 3D into Specialist C restores it to 74.93%, a non-trivial gain given the extreme data scarcity (≈10k samples), confirming that these modalities occupy distinct representational subspaces.

Freezing the backbone is also critical: unfreezing reduces ACC to 81.36% with substantially higher variance (±2.07%), consistent with feature distortion in data-scarce regimes [10].

4 Conclusion and Future Work

This paper presents MOSAIC, a parameter-efficient unified framework that mitigates cross-modal interference through explicit routing within a frozen backbone. Extensive validation on 18 MedMNIST datasets and six external benchmarks demonstrates that a single unified model (93.4M parameters) can rival 18 independently trained specialists (1.55B parameters), uniquely achieving 2D-3D unification within a single frozen backbone—a capability absent in domain-specific (RadImageNet, Med3D) or resource-intensive approaches (UniMiSS).

Our design reflects two deliberate trade-offs to address data scarcity and stability. First, while relying on a frozen ImageNet backbone restricts task-specific semantic adaptation, it is a necessary constraint to prevent feature distortion in data-scarce medical regimes, ensuring that robust visual primitives are preserved. Second, we employ coarse-grained explicit modality routing instead of learnable gating. While this introduces localized within-group heterogeneity (*e.g.*, in ultrasound), it prioritizes structural stability to avert the routing imbalance common in soft mixture-of-experts. Future work will address these limitations by exploring stable, adaptive routing mechanisms to achieve finer-grained decoupling without sacrificing convergence stability.

Acknowledgments. I would like to thank Prof. Weidong Cai, Dr. Zihao Tang, and other faculty members for their invaluable guidance throughout this research. I am also grateful to the anonymous reviewers for their constructive suggestions.

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

References

1. Aghajanyan, A., Saraf, L., Gupta, S.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: ACL (2021)
2. Chen, S., et al.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: NeurIPS (2022)
3. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: WACV. pp. 574–584 (2022)
7. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799 (2019)
8. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)

9. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV. pp. 709–727 (2022)
10. Kumar, A., et al.: Fine-tuning can distort pre-trained features and underperform out-of-distribution. In: ICLR (2022)
11. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
12. Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open ai foundation model applied to chest radiography. Nature **643**, 488–500 (2025)
13. Ma, D., et al.: Foundation ark: Accruing and reusing knowledge for superior and robust performance. In: MICCAI (2023)
14. Ma, D., Pang, J., Velan, S.S., Gotway, M.B., Liang, J.: Ark+: Supervised training a single high-performance ai foundation model from many differently labeled datasets—no label consolidation required. Medical Image Analysis p. 103828 (2026)
15. Mei, X., et al.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
16. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
17. Pan, Y., Ge, Y., Lu, B., et al.: Uni-moe: Scaling unified multimodal llms with mixture of experts. arXiv preprint arXiv:2405.11273 (2024)
18. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green ai. Communications of the ACM **63**(12), 54–63 (2020)
19. Shazeer, N., et al.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: ICLR (2017)
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models. In: NeurIPS (2017)
21. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS. vol. 35, pp. 10078–10093 (2022)
22. Xie, Y., Zhang, J., Liao, Z., Xia, Y., Shen, C.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: ECCV. pp. 558–575 (2022)
23. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
24. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11114–11124 (2024)
25. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: NeurIPS. vol. 33, pp. 5824–5836 (2020)