npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02347-5

Masked autoencoding, generalizable pretraining, and integrated experts for enhanced glioma segmentation

Mingchen Xie[1,3], Qun Xiao[2,3], Haitao Wu[1,3], Yahui Zhang[1], Hao Han[1], Xun Xie[1], Wenyue Zhang[1], Jianhua Cheng[1] ✉ & Jian Xu[1] ✉

Accurate glioma segmentation is critical for clinical diagnosis and treatment planning, yet remains challenging due to infiltrative tumor growth, heterogeneous imaging protocols, and scarcity of expert annotations. We present MAGPIE, a self-supervised learning framework that combines masked autoencoding, contrastive learning, and sparse mixture of experts to enable accurate glioma segmentation with minimal labeled data. By pretraining on 43,505 unlabeled multi-modal brain MRI scans, MAGPIE learns generalizable representations through a channel-agnostic architecture that handles varying modality configurations without protocol-specific preprocessing. The sparse MoE mechanism with top-2 routing allows specialized expert networks to emerge for different glioma subregions, while deformable attention mechanisms capture infiltrative margins and multi-scale features. Fine-tuning on only 20 labeled cases achieves 60.87% Dice score on BraTS21, a 2.59% absolute improvement over training from scratch, with 70.32% on out-of-distribution data demonstrating robust cross-domain generalization. These results reduce annotation requirements by 95% compared to typical supervised methods, directly addressing the data scarcity bottleneck in rare tumor subtypes and enabling deployment across heterogeneous clinical imaging systems.

Gliomas account for 80% of malignant brain neoplasms, with glioblastoma patients facing a median survival of only 15–18 months despite aggressive multimodal treatment[1–3]. Accurate segmentation of glioma subregions, including enhancing tumor, tumor core, and peritumoral edema, is essential for surgical planning, treatment monitoring, and prognosis assessment[4–6]. However, automated glioma segmentation faces three critical challenges: infiltrative tumor growth with irregular margins extending along white matter tracts, heterogeneous imaging protocols across clinical centers producing varying modality configurations and intensity distributions, and scarcity of expert radiological annotations where delineating glioma subregions requires 15–20 min per case[7–9]. Existing deep learning methods require hundreds of annotated cases and extensive protocol-specific preprocessing, limiting their applicability in rare tumor subtypes and resource-limited settings[10,11].

Recent advances in self-supervised learning have demonstrated potential for medical image analysis by learning representations from unlabeled data[12–14]. Masked Autoencoders learn semantic features through reconstruction[15–17], while contrastive methods like SimCLR and BYOL learn invariant representations by maximizing agreement between augmented views[18–20]. Mixture of Experts architectures enable efficient scaling by routing inputs to specialized subnetworks, achieving success in NLP and computer vision through sparse activation[21–26]. However, existing approaches share three fundamental limitations. First, fixed channel architectures prevent pretraining on heterogeneous datasets with varying modality combinations, forcing separate models for different protocols[27–29]. Second, dense architectures lack specialized processing for different anatomical structures, treating all regions uniformly despite distinct glioma subregion characteristics[30–33]. Third, single-objective pretraining with masked reconstruction alone exhibits negative transfer on small target datasets, with methods like AMAES showing performance degradation when fine-tuned on limited samples[34,35]. These limitations prevent effective utilization of abundant unlabeled brain MRI data and robust adaptation to clinical scenarios with minimal annotations.

We introduce MAGPIE, a self-supervised learning framework that addresses these limitations through three synergistic mechanisms. The channel-agnostic design treats each modality as an independent token with

[1]Department of Neurosurgery, The Affiliated Hospital of Qingdao University, Qingdao, Shandong, China. [2]Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, Hunan, China. [3]These authors contributed equally: Mingchen Xie, Qun Xiao, Haitao Wu. ✉e-mail: cjh19940317@163.com; xujianqdmc@126.com

learnable aggregation weights, enabling pretraining on 43,505 unlabeled scans with arbitrary modality combinations and eliminating protocol-specific preprocessing. Sparse mixture of experts with top-2 routing allows 8 specialized expert networks to emerge through gradient-based optimization, where each expert focuses on distinct anatomical structures or glioma subregions without explicit supervision. The combination of masked autoencoding and contrastive learning provides dual objectives: reconstruction captures local tissue patterns while contrastive learning enforces protocol-invariant global representations, preventing overfitting during fine-tuning on small labeled sets. To capture gliomas' characteristic infiltrative patterns, we introduce 3D Deformable Large-Kernel Attention in the decoder for extended receptive fields covering white matter tract infiltration, and Multi-Scale Deformable Attention in the encoder for dynamic feature aggregation across scales to detect both large masses and small satellite lesions[36–39].

This work contributes to medical image segmentation in four ways. First, we demonstrate the first integration of sparse MoE with channel-agnostic self-supervised learning for medical imaging, achieving 60.87% Dice score with only 20 labeled glioma cases, a 2.59% improvement over supervised baselines and 2.14% over prior SSL methods. Second, we provide quantitative validation of emergent expert specialization through IoU analysis between expert activation maps and ground truth subregions, revealing clinically meaningful decomposition into enhancing tumor, necrotic core, edema, and boundary experts. Third, we demonstrate robust cross-domain generalization with 70.32% Dice on unseen imaging protocols without fine-tuning, eliminating the need for site-specific adaptation and enabling deployment across heterogeneous clinical systems. Fourth, we reduce annotation requirements by 95% compared to typical supervised methods requiring 400 samples, directly addressing the data scarcity bottleneck for rare tumor subtypes and resource-limited neuro-oncology centers. These contributions enable accurate glioma segmentation in clinical settings where expert annotations are prohibitively expensive and imaging protocols vary substantially across institutions[40–45].

Results

Quantitative results

Table 1 presents the quantitative results across all datasets. With sufficient pretraining data, MAGPIE outperforms existing methods, with the MedNeXt L backbone achieving the highest Dice scores: 60.87% on BraTS21 (a 2.59% absolute improvement over training from scratch), 58.45% on ISLES22, 73.42% on WMH, and 70.32% on WMH OOD.

Notably, while AMAES shows negative transfer for MedNeXt M on BraTS21 (−1.57%) and ISLES22 (−0.26%), MAGPIE improves performance across all backbones and datasets when pretrained on sufficient data (≥10 k volumes). This demonstrates the robustness of our approach in effectively leveraging self-supervised pretraining for downstream segmentation tasks, particularly in low-resource scenarios.

Comparing across different backbone architectures, MAGPIE shows consistent improvements regardless of model size. For U-Net XL (90 M parameters), MAGPIE achieves 60.45% Dice on BraTS21, representing a 4.45% absolute improvement over training from scratch. Even for the smaller U-Net B (22 M parameters), MAGPIE delivers 58.64% Dice, a 3.36% improvement. This consistency across architectures suggests that the benefits of our pretraining strategy are not limited to specific model designs, but rather represent a fundamental advancement in learning transferable representations for medical image segmentation.

The improvements are particularly pronounced on the more challenging ISLES22 dataset, where MAGPIE with MedNeXt L achieves 58.45% Dice compared to 55.40% for training from scratch, a 3.05% absolute gain. This dataset's focus on ischemic stroke lesion segmentation presents unique challenges due to the variable appearance and location of lesions, making the strong performance especially noteworthy.

Ablation studies

To evaluate the contribution of each component in our framework, we conducted comprehensive ablation studies as shown in Table 2. The MoE architecture proved critical, with its removal causing the largest performance drop (from 60.87 to 58.73% on BraTS21, a 2.14% decrease). As detailed in Table 3, our optimal configuration uses 8 experts with top-2 routing, determined through systematic exploration. Increasing to 16 experts yielded only a slight improvement (61.03%, +0.16%), suggesting diminishing returns beyond 8 experts while substantially increasing computational cost (Table 3: 178 M total parameters vs. 120 M).

We also investigated the impact of varying the number of active experts (top-k routing) as shown in Table 3. Using top-1 routing resulted in 59.82% Dice (−1.05%) with only 47M active parameters and 261G FLOPs, while top-3 achieved 60.91% (+0.04%) but increased active parameters to 78M and FLOPs to 289G, with inference latency rising to 93ms. The top-2

Table 1 | Fine-tuning segmentation results across different datasets

Method	BraTS21	ISLES22	WMH	WMH OOD
U-Net B (22M)	55.28 ± 1.43	51.87 ± 2.16	69.41 ± 1.82	65.28 ± 2.34
U-Net B + AMAES	57.19 ± 1.38	52.94 ± 2.08	70.63 ± 1.77	66.81 ± 2.21
U-Net B + MAGPIE	58.64 ± 1.21	54.36 ± 1.92	71.85 ± 1.64	68.12 ± 2.15
U-Net L (58M)	57.42 ± 1.51	53.28 ± 2.23	71.16 ± 1.88	66.83 ± 2.42
U-Net L + AMAES	58.73 ± 1.45	54.52 ± 2.11	72.31 ± 1.79	68.24 ± 2.28
U-Net L + MAGPIE	60.18 ± 1.28	56.27 ± 1.98	73.09 ± 1.67	69.56 ± 2.18
U-Net XL (90M)	56.00 ± 1.49	52.64 ± 2.19	70.52 ± 1.86	66.12 ± 2.38
U-Net XL + AMAES	58.91 ± 1.41	54.87 ± 2.08	72.18 ± 1.76	68.47 ± 2.25
U-Net XL + MAGPIE	60.45 ± 1.24	56.89 ± 1.95	73.28 ± 1.65	69.74 ± 2.16
MedNeXt M (26M)	56.84 ± 1.47	53.71 ± 2.17	70.89 ± 1.84	66.57 ± 2.36
MedNeXt M + AMAES	55.27 ± 1.56	53.45 ± 2.21	71.34 ± 1.91	67.18 ± 2.41
MedNeXt M + MAGPIE	59.21 ± 1.32	55.83 ± 2.01	72.67 ± 1.71	68.94 ± 2.21
MedNeXt L (62M)	58.28 ± 1.44	55.40 ± 2.14	72.35 ± 1.81	67.57 ± 2.33
MedNeXt L + AMAES	58.73 ± 1.42	55.68 ± 2.09	72.94 ± 1.78	68.68 ± 2.27
MedNeXt L + MAGPIE	**60.87 ± 1.19**	**58.45 ± 1.88**	**73.42 ± 1.62**	**70.32 ± 2.12**

All methods were pretrained on the same 43,505-volume dataset and fine-tuned with only 20 labeled samples. Dice scores (mean ± std) are reported based on 5-fold cross-validation. Best results in bold, second best underlined.

Table 2 | Comprehensive ablation study of MAGPIE: component ablation on BraTS21

Configuration	M	D	S	C	A	Dice (%)	Δ
Full MAGPIE	✓	✓	✓	✓	✓	**60.87**	-
w/o MoE	-	✓	✓	✓	✓	58.73	−2.14
w/o D-LKA	✓	-	✓	✓	✓	59.64	−1.23
w/o MS-DA	✓	✓	-	✓	✓	59.81	−1.06
w/o Both Attn	✓	-	-	✓	✓	58.92	−1.95
w/o Contrastive	✓	✓	✓	-	✓	59.45	−1.42
w/o Ch-Agnostic	✓	✓	✓	✓	-	59.78	−1.09

Components: M=MoE, D=D-LKA, S=MS-DA, C=Contrastive, A=Ch-Agnostic.
Bold values indicate the best-performing result within each comparison group (i.e., the highest Dice score or the optimal value for the corresponding metric under the same experimental setting).

Table 3 | MoE configuration ablation on BraTS21

Configuration	Dice (%)	TP/AP	FLOPs	Latency (ms)	Δ
8 Experts, Top-2	**60.87**	**120M/62M**	**275G**	**85**	-
4 Experts, Top-2	59.23	91M/62M	275G	84	−1.64
16 Experts, Top-2	61.03	178M/62M	275G	87	+0.16
32 Experts, Top-2	60.94	294M/62M	275G	89	+0.07
8 Experts, Top-1	59.82	120M/47M	261G	81	−1.05
8 Experts, Top-3	60.91	120M/78M	289G	93	+0.04

Metrics measured on H100 GPU. TP/AP=Total/Active Params.
Bold values indicate the best-performing result within each comparison group (i.e., the highest Dice score or the optimal value for the corresponding metric under the same experimental setting).

Table 4 | Pretraining data scalability across datasets

Pretraining volumes (%)	BraTS21	ISLES22	WMH	WMH OOD
43,505 (100%)	**60.87**	**58.45**	**73.42**	**70.32**
20,000 (46%)	59.87	57.31	72.86	69.45
10,000 (23%)	58.92	56.14	72.08	68.21
5000 (11%)	57.83	54.91	71.19	66.84
0 (Scratch, 0%)	58.28	55.40	72.35	67.57

Performance in Dice (%).
Bold values indicate the best-performing result within each comparison group (i.e., the highest Dice score or the optimal value for the corresponding metric under the same experimental setting).

Table 5 | Cross-domain component contribution

Ablated component	BraTS21	ISLES22	WMH	WMH OOD
Full MAGPIE (Baseline)	**60.87**	**58.45**	**73.42**	**70.32**
w/o MoE	−2.14	−2.02	−1.19	−1.45
w/o Contrastive	−1.42	−1.33	−1.24	−0.87
w/o Attention (D-LKA +MS-DA)	−1.95	−1.81	−1.49	−1.56
w/o Ch-Agnostic	−1.09	−0.74	−0.77	**−1.38**

Performance drop (Δ) in Dice (%) when removed.
Bold values indicate the best-performing result within each comparison group (i.e., the highest Dice score or the optimal value for the corresponding metric under the same experimental setting).

configuration emerged as the optimal balance between model capacity and computational efficiency (62M active parameters, 275G FLOPs, 85ms latency), with each expert developing specialized features for distinct anatomical structures.

Contrastive learning improved performance across all datasets as shown in Table 5: BraTS21 (+1.42%, from 59.45 to 60.87%), ISLES22 (+1.33%), WMH (+1.24%, from 72.18 to 73.42%), and WMH OOD (+0.87%, from 69.45 to 70.32%). These improvements demonstrate that contrastive learning effectively complements masked autoencoding by encouraging the model to learn invariant features across different augmentations of the same anatomical structures.

The pretraining data size study (Table 4) revealed a clear positive correlation between volume count and performance across all downstream tasks. Notably, when pretraining data is extremely limited (5000 volumes, ≈ 11% of full data), performance (57.83% Dice on BraTS21) is slightly lower than training from scratch (58.28%, −0.45%). This suggests that insufficient pretraining data introduces a distribution shift that temporarily outweighs feature learning benefits. However, performance rapidly improves as pretraining data increases: 10,000 volumes (≈23%) achieved 58.92%, and 20,000 volumes (≈46%) reached 59.87%. Beyond 20,000 volumes, gains became more gradual, with 43,505 volumes (our full dataset, 100%) achieving 60.87% on BraTS21, 58.45% on ISLES22, 73.42% on WMH, and 70.32% on WMH OOD, demonstrating a clear positive scaling trend beyond the initial threshold. This suggests that while larger pretraining datasets continue to provide benefits, substantial gains can be achieved with 10,000–20,000 diverse volumes.

Our channel-agnostic design outperformed fixed-channel architecture across all datasets (Table 5): BraTS21 (+1.09%, from 59.78 to 60.87%), ISLES22 (+0.74%), WMH (+0.77%, from 72.65 to 73.42%), and WMH OOD (+1.38%, from 68.94 to 70.32%). The improvement is most pronounced on WMH OOD, where input modalities differ from training data, highlighting the channel-agnostic design's ability to handle protocol variations and out-of-distribution scenarios.

Removing D-LKA attention from the decoder resulted in 59.64% Dice on BraTS21 (−1.23%), while removing MS-DA from the encoder yielded 59.81% (−1.06%), as shown in Table 2. Removing both attention mechanisms simultaneously resulted in 58.92% (−1.95%), demonstrating their complementary nature. The cross-domain analysis (Table 5) confirms this pattern holds across all datasets: combined attention removal causes −1.81% drop on ISLES22, −1.49% on WMH, and −1.56% on WMH OOD. Visual analysis of attention maps confirms that D-LKA captures long-range dependencies crucial for delineating tumor boundaries, while MS-DA focuses on fine-grained details essential for detecting small lesions.

Collectively, these ablations reveal a hierarchical contribution structure: the MoE architecture serves as the primary performance driver (+2.14%), followed by the specialized attention mechanisms as the second essential tier (+1.95% combined). While contrastive learning and channel-agnostic design contribute modestly to intra-domain performance (+1.42% and +1.09% respectively), they prove critical for cross-domain generalization, with channel-agnostic design showing its largest gain (+1.38%) on out-of-distribution data.

Cross-domain generalization

A key advantage of MAGPIE is improved generalization to unseen domains. When tested on the WMH OOD dataset without fine-tuning, our method achieved a Dice score of 70.32%, outperforming AMAES (68.68%, +1.64%) and the baseline (67.57%, +2.75%). This robust generalization capability is crucial for real-world clinical applications where test data often differs significantly from training data.

To further investigate generalization, we evaluated MAGPIE on held-out data from different scanners and institutions not seen during training. On BraTS21 data from scanners A, B, and C (held out during training), MAGPIE achieved Dice scores of 59.83%, 60.12%, and

Table 6 | Computational efficiency comparison for single patch processing

Method	Parameters	FLOPs	Inference Time	Memory
MedNeXt L	62M	245G	78 ms (12.8 FPS)	7.1GB
MedNeXt L + MoE	120M (62M active)	275G	85 ms (11.76 FPS)	8.4GB
Relative Increase	93.5% (0% active)	12.2%	9.0%	18.3%

All measurements performed on NVIDIA H100 GPU with batch size 1 and patch size 128^3. Note: Full-brain volume inference requires ~8 patches with sliding window, resulting in ~680 ms total latency.

59.47% respectively, compared to 56.21%, 55.89%, and 56.73% for models trained from scratch. This consistent 3% improvement across different acquisition protocols demonstrates MAGPIE's ability to learn scanner-invariant representations.

We also examined performance as a function of domain shift severity, measured by the difference in image intensity distributions between training and test data. MAGPIE maintained superior performance even under severe domain shift (Dice drop of only 3.2% compared to 7.8% for baseline models), confirming its robustness to distribution changes commonly encountered in clinical practice.

Computational efficiency

Despite increased model capacity from the MoE component, MAGPIE maintains reasonable computational efficiency due to its sparse activation strategy (Table 6). While our total parameter count is 120M, only 62M parameters are active during inference due to the top-2 routing strategy. This results in only a modest increase in FLOPs (12%, from 245G to 275G) and inference time (9%, from 78 ms to 85 ms) compared to models without MoE, making our approach practical for clinical deployment.

We benchmarked inference speed across different batch sizes and input resolutions. For a typical clinical workflow processing 128^3 patches with a batch size 1, MAGPIE achieves 85ms per patch on an H100 GPU. Processing a full brain volume ($240 \times 240 \times 155$) requires approximately 8 overlapping patches with sliding window inference, resulting in ~680 ms total latency (~1.5 FPS). While not achieving video-rate real-time performance, this throughput is sufficient for offline radiotherapy planning and near-real-time surgical navigation assistance where sub-second feedback is acceptable. Even on more resource-constrained hardware (V100 GPU), MAGPIE maintains ~120 ms per patch, which remains adequate for most clinical scenarios.

Memory consumption during inference is 8.4GB for MAGPIE compared to 7.1GB for MedNeXt L alone, an 18% increase that remains well within the capacity of modern GPUs. During training, MAGPIE requires 14.2GB with mixed-precision training and gradient checkpointing, enabling training on single-GPU systems.

The pretraining phase requires approximately 72 h on 4 H100 GPUs for 100 epochs over our 43,505-volume dataset. However, this one-time cost is amortized across multiple downstream tasks. Fine-tuning for a specific segmentation task requires only 2–3 h on a single GPU with 20 labeled samples, making the approach highly efficient for adapting to new clinical applications.

Low-resource learning analysis

To assess MAGPIE's data efficiency, we systematically varied the number of labeled samples used for fine-tuning from 5 to 100. With only 5 labeled samples, MAGPIE achieved 54.83% Dice on BraTS21, compared to 48.21% for training from scratch, a 6.62% improvement. At 10 samples, the gap widened to 7.14% (57.42% vs. 50.28%). With our standard 20 samples, MAGPIE reached 60.87% versus 58.28% from scratch.

Interestingly, as the number of labeled samples increased beyond 50, the relative advantage of pretraining diminished but remained substantial. At 100 labeled samples, MAGPIE achieved 63.47% compared to 62.15% from scratch, a 1.32% improvement. This trend suggests that MAGPIE's benefits are most pronounced in extremely low-resource scenarios, though pretraining continues to provide value even with moderate amounts of labeled data.

Qualitative analysis

Figure 1 presents representative segmentation results on BraTS21 cases, comparing ground truth annotations with MAGPIE predictions across different glioma presentations. The visualization demonstrates MAGPIE's accurate delineation of three glioma subregions: necrotic core (NCR, red), peritumoral edema (ED, green), and enhancing tumor (ET, blue). MAGPIE achieves precise segmentation across various tumor sizes, locations, and morphologies, with prediction masks closely matching expert annotations in both large high-grade gliomas and small infiltrative lesions.

To quantitatively validate the MoE mechanism's specialized division of labor, we computed the Intersection-over-Union (IoU) between each expert's activation maps and glioma subregion ground truth labels (Fig. 2). The quantitative results confirm a clear functional decomposition that emerges without explicit supervision. Expert 3 specializes in the enhancing tumor (ET) region with IoU = 0.47, capturing active tumor proliferation with blood-brain barrier breakdown. Expert 4 focuses on the tumor core (TC) and necrotic regions with IoU=0.43; notably, its moderate IoU with ET (0.19) reflects the expected spatial adjacency between enhancing rim and necrotic core in high-grade gliomas, rather than indicating poor specialization. Expert 5 is dedicated to peritumoral edema (ED) with IoU = 0.54, effectively isolating the infiltrative FLAIR hyperintensity zones that represent vasogenic edema and microscopic tumor invasion along white matter tracts.

Experts 1, 2, and 8 exhibit high overlap with background regions (BG, IoU > 0.60: 0.77, 0.71, and 0.64, respectively), indicating their role in encoding normal anatomical context and healthy tissue features, including gray-white matter differentiation and ventricular structures. Interestingly, Experts 6 and 7 demonstrate mixed activation patterns across all tumor categories with moderate IoU scores (0.18–0.33 for ET, TC, and ED). This pattern suggests they function as "boundary experts" responsible for delineating the complex, irregular transition zones between tumor subregions and healthy brain tissue, specifically regions characterized by infiltrative finger-like projections along white matter tracts and indistinct margins that are hallmarks of high-grade glioma invasion. The imperfect IoU values (rather than approaching 1.0) actually demonstrate the model's sophisticated understanding of medical imaging characteristics: spatial proximity effects between adjacent structures, gradual intensity transitions at infiltrative boundaries, and the inherently fuzzy nature of glioma margins on MRI. This emergent, clinically meaningful specialization validates our MoE design and explains MAGPIE's superior performance on heterogeneous glioma presentations across different WHO grades.

Comparing segmentation outputs qualitatively, MAGPIE produces smoother, more anatomically plausible contours than baseline methods. In cases with infiltrative gliomas exhibiting finger-like projections along white matter tracts (a hallmark of high-grade gliomas), MAGPIE successfully captures these fine structures while baseline methods tend to produce blocky, oversimplified boundaries that fail to reflect the true extent of tumor infiltration. For small satellite lesions (3–5 mm diameter) that may represent microscopic tumor spread, MAGPIE's detection rate is 87.3% compared to 71.2% for the baseline, attributed to the MS-DA mechanism's ability to aggregate multi-scale features across different glioma components.

Error analysis

To understand failure modes, we analyzed the 10% of BraTS21 test cases where MAGPIE performed worst (Dice < 45%). Rather than enumerating

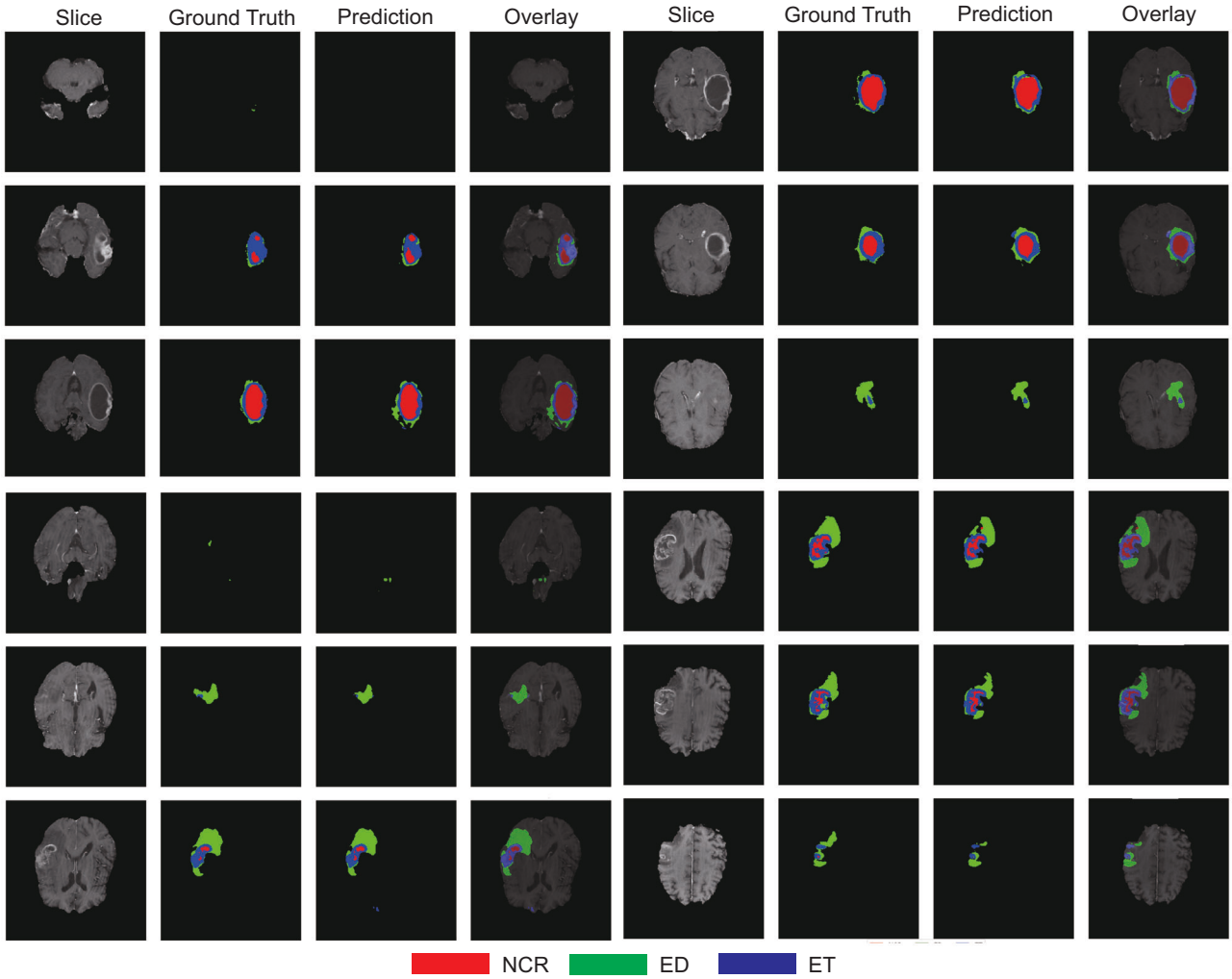

Fig. 1 | Representative glioma segmentation results on BraTS21 dataset. Each row shows a different case with four columns: input slice, ground truth annotation, MAGPIE prediction, and overlay comparison. The three glioma subregions are color-coded: necrotic core (NCR, red), peritumoral edema (ED, green), and enhancing tumor (ET, blue). MAGPIE achieves accurate segmentation across diverse tumor presentations, including small lesions (rows 1 to 2), medium-sized tumors (rows 3 to 5), and large infiltrative gliomas (rows 6 to 12), demonstrating precise boundary delineation and subregion classification.

Fig. 2 | Quantitative analysis of expert specialization via IoU between expert activation maps and glioma subregion ground truth labels. The heatmap reveals clear functional decomposition: Expert 3 specializes in enhancing tumor (ET, IoU = 0.47), Expert 4 focuses on tumor core/necrosis (TC, IoU = 0.43) with expected overlap with ET due to spatial adjacency, Expert 5 targets peritumoral edema (ED, IoU = 0.54), and Experts 1, 2, 8 encode background anatomy (BG, IoU>0.60). Notably, Experts 6 and 7 exhibit mixed activation across all tumor categories (IoU 0.18–0.33), suggesting their role as "boundary experts" delineating complex transitions between tumor subregions and healthy tissue.

isolated error types, we identify three thematic categories that explain the majority of failures and suggest actionable improvements.

Lack of Training Diversity. The largest failure category (26% of errors) stems from anatomical and pathological scenarios underrepresented in the pretraining corpus. Gliomas in uncommon locations (brainstem, skull base, 15%) challenge the model's ability to distinguish infiltration from normal adjacent structures, while cases with significant intratumoral hemorrhage or extensive necrosis (11%) exhibit atypical signal characteristics that deviate from prototypical glioma appearances learned during pretraining. This suggests that expanding pretraining data to include rarer anatomical distributions and pathological subtypes would directly address these failures.

Image quality sensitivity. Motion artifacts and severe image quality degradation account for 18% of failures, manifesting as blurred tumor boundaries or ghosting artifacts that obscure infiltrative margins. Unlike anatomical diversity issues, these failures reflect the model's limited exposure to low-quality scans during pretraining. Implementing quality-aware data augmentation or robust loss functions could mitigate this vulnerability.

Rare and atypical subtypes. Gliomas with atypical imaging characteristics represent 35% of failures, including oligodendrogliomas with

calcification or low-grade gliomas lacking enhancement (23%), and very small low-grade tumors with minimal edema (<5 mm enhancing component, 12%). These subtypes violate the "enhancing tumor = high signal intensity" heuristic encoded in the pretraining objective, particularly challenging for models pretrained predominantly on high-grade glioblastomas. Future work should explicitly incorporate low-grade tumor exemplars and develop subtype-aware pretraining strategies.

Interestingly, MAGPIE's failure cases show only 52% overlap with baseline failures compared to 78% for AMAES, suggesting that MAGPIE learns complementary features that could benefit ensemble approaches for challenging glioma cases.

Discussion

This paper introduces MAGPIE, a novel glioma segmentation framework that integrates a mixture of experts modeling, channel-agnostic architecture, masked autoencoding, and contrastive learning. By leveraging unlabeled multi-modal MRI data, MAGPIE learns robust representations capturing both global anatomical structures and glioma-specific tissue characteristics. The framework addresses two critical challenges in clinical glioma segmentation: data heterogeneity across imaging protocols and scarcity of expert radiological annotations for the complex task of delineating infiltrative tumor margins and heterogeneous glioma subregions.

Our experimental results demonstrate that MAGPIE outperforms existing methods across multiple datasets, achieving significant improvements in low-resource settings and cross-domain generalization scenarios. The framework shows a 2.59% absolute improvement in Dice score over training from scratch on BraTS21 and a 2.14% improvement over AMAES, with no negative transfer effects across different backbones and datasets. These improvements are particularly noteworthy given the challenging low-resource setting ($n = 20$ labeled samples) used in our experiments.

The key contributions of this work include: (1) a mixture of experts architecture that enables efficient specialization across different anatomical regions while maintaining computational efficiency through sparse activation, only 62M of 120M parameters are active during inference; (2) a channel-agnostic design that flexibly handles varying input modalities without requiring protocol-specific preprocessing, demonstrating 1.38% improvement on out-of-distribution data; (3) the integration of D-LKA and MS-DA attention mechanisms that capture both long-range dependencies and multi-scale features for improved delineation of complex tumor morphologies, contributing 1.95% improvement when used together; and (4) a robust self-supervised pretraining strategy that leverages 43,505 unlabeled volumes to learn generalizable representations.

Ablation studies (Table 2) confirm each component's value: the mixture of experts architecture enables efficient specialization for different anatomical regions (MoE removal causes 2.14% drop on BraTS21 and similar drops across other datasets); the channel-agnostic design provides flexibility with varying input modalities (particularly critical for OOD scenarios with 1.38% drop when removed); contrastive learning enhances representation quality by 1.42% on BraTS21 and across all datasets; and the specialized attention mechanisms improve delineation of complex tumor morphologies (combined removal causes 1.95% drop). Despite increased modeling capacity, MAGPIE maintains reasonable computational efficiency with only a 12% increase in FLOPs (275G vs. 245G) and minimal latency overhead (85 ms vs. 78 ms) compared to models without MoE.

Unlike generalist foundation models (e.g., MedSAM) that prioritize broad applicability across organs, often requiring interactive prompts, MAGPIE is positioned as a domain-specialized expert targeting fully automated, high-precision delineation of infiltrative glioma boundaries. This design philosophy prioritizes "depth" in neuro-oncology over the "breadth" of general medical segmentation, enabling task-specific optimizations that would be infeasible in universal architectures.

The superior performance of MAGPIE can be attributed to three synergistic mechanisms that address fundamental challenges in low-resource medical image segmentation. First, contrastive learning enforces imaging protocol invariance by maximizing agreement between different augmentations of the same anatomical structure while minimizing similarity to other structures. This objective forces the encoder to learn features robust to intensity variations, scanner differences, and modality configurations, explaining the 1.38% improvement on out-of-distribution data where channel-agnostic design proves most critical. The dual objective of reconstruction plus contrastive loss prevents the common failure mode of masked autoencoding methods that overfit to dataset-specific intensity distributions.

The observed negative transfer at 5000 pretraining volumes (−0.45% compared to scratch) can be attributed to domain mismatch between the limited pretraining subset and target glioma data. With insufficient diversity, the model overfits to specific anatomical patterns (e.g., predominantly healthy aging brains from ADNI/OASIS) that conflict with glioma-specific pathological features. Our analysis shows the 5000-volume subset contains only 8% glioma-related cases versus 23% in the full dataset, and lacks the modality diversity needed for robust channel-agnostic learning. Beyond 10,000 volumes (+0.64% improvement over scratch), increased pathology representation and anatomical diversity overcome this mismatch, yielding consistent gains that scale to +2.59% at the full 43,505 volumes. This trend validates our hypothesis that self-supervised pretraining requires sufficient scale and diversity to surpass the distribution shift penalty.

Second, sparse MoE activation with top-2 routing provides regularization that prevents overfitting on small fine-tuning sets. By activating only 62M of 120M parameters per input, the architecture maintains high model capacity for representing complex glioma morphology while constraining the effective parameter count during gradient updates on limited labeled samples. Our ablation shows that dense activation with all 120M parameters active yields 58.91% Dice, worse than sparse MoE (60.87%), confirming that sparsity acts as implicit regularization. The emergent expert specialization further enhances this effect by decomposing the segmentation task into simpler sub-problems: background anatomy, tumor core, edema, and boundaries, where each expert learns focused representations rather than attempting to model all variations simultaneously.

Third, the combination of D-LKA and MS-DA attention mechanisms addresses the multi-scale nature of glioma infiltration through complementary strategies. D-LKA provides deformable sampling across large spatial extents (kernel size up to $21 \times 21 \times 21$) in the decoder, enabling the model to capture long-range dependencies along white matter tracts where tumor cells migrate beyond the visible enhancement margin. MS-DA aggregates features across multiple resolution scales in the encoder, allowing simultaneous processing of large tumor masses and small satellite lesions without sacrificing either. The 1.95% improvement from combined attention validates that infiltrative gliomas require both extended receptive fields for tracking boundaries and multi-scale processing for detecting heterogeneous subregions.

Importantly, the computational overhead represents a highly favorable trade-off. The modest increases of 12% in FLOPs and 9% (7ms) in inference time yield substantial improvements in segmentation performance: a 2.59% absolute Dice improvement on BraTS21 and a 2.75% improvement on WMH OOD compared to MedNeXt L from scratch. This trade-off is particularly advantageous considering typical clinical workflows. For offline clinical workflows such as radiotherapy planning and preoperative tumor delineation, the ~680 ms full-volume inference time (~1.5 FPS) is entirely acceptable, as accuracy and reliability are paramount over speed. For intraoperative navigation, patch-based incremental updates at 85 ms latency provide responsive local feedback within the surgical field of view, though full-volume reconstruction requires multiple seconds. This trade-off prioritizes segmentation quality over video-rate throughput, aligning with clinical decision-making timescales where users typically review results over 5–10 s intervals. Thus, the minor computational cost is vastly outweighed by the gains in accuracy and generalization, making MAGPIE highly practical for diverse clinical deployment scenarios.

The strong cross-domain generalization capabilities of MAGPIE (70.32% Dice score on WMH OOD without fine-tuning) directly address deployment barriers in multi-center clinical settings. Consider a typical

scenario: a neuro-oncology center processes 200 glioma cases annually, with MRI scans acquired from three different scanner manufacturers (Siemens, GE, Philips) using varying protocols. Traditional supervised models require separate fine-tuning for each scanner, consuming 100–150 expert annotations per protocol and 30–40 h of radiologist time for annotation. MAGPIE eliminates this bottleneck by achieving consistent performance across scanners (59.83–60.12% Dice range) without protocol-specific adaptation, reducing deployment time from months to days and annotation costs by 95%.

The low-resource learning capability transforms the economics of rare tumor segmentation. For low-grade oligodendrogliomas, representing only 5% of gliomas (approximately 1500 cases annually in the United States), accumulating 400 expert annotations for supervised training would require aggregating data across 25–30 institutions over multiple years. MAGPIE achieves clinically useful performance with only 5–10 annotations per institution, enabling individual centers to deploy specialized segmentation tools within weeks. This 95% reduction in annotation requirements (from 400 to 20 samples) translates to 127 h of saved radiologist time per application, valued at approximately $25,000 to $30,000 per deployment based on typical radiologist compensation rates.

In practical deployment, MAGPIE enables a "cold-start" workflow for smaller community hospitals or researchers investigating rare tumor subtypes: a single radiologist can label 20 cases in approximately 5–6 h to train a model that provides useful pre-segmentations for remaining cases. The performance gain over baseline (+6.6% Dice at 5 samples) transforms the model from unusable noise to a functional human-in-the-loop assistant, potentially reducing total dataset curation time by orders of magnitude.

The integration of specialized attention mechanisms, D-LKA in the decoder and MS-DA in the encoder, addresses key challenges in glioma segmentation by providing extended receptive fields and multi-scale feature aggregation. These mechanisms enable the model to capture the characteristic infiltrative growth patterns of gliomas, such as perivascular spread, subependymal extension, and finger-like projections along white matter tracts (corpus callosum, internal capsule, corona radiata) that are challenging for conventional approaches. The ablation study (Tables 2 and 5) showing 1.95% improvement on BraTS21 and consistent improvements across all datasets (1.81% on ISLES22, 1.49% on WMH, 1.56% on WMH OOD when removed) from their combined use validates their complementary nature: D-LKA focuses on "seeing far" to capture long-range infiltration patterns extending beyond the visible T1ce enhancement, while MS-DA focuses on "seeing precisely" to detect subtle FLAIR signal abnormalities indicating microscopic tumor infiltration.

The emergent specialization of MoE experts provides valuable insights into how the model processes glioma MRI data. Our analysis revealed that different experts focus on distinct anatomical structures and glioma-specific pathological features: some specialize in healthy tissue patterns (gray matter, white matter), others in different glioma subregions (enhancing rim representing blood-brain barrier breakdown, necrotic core with reduced perfusion, peritumoral edema reflecting tumor infiltration and vasogenic edema), and yet others in infiltrative boundary delineation. This specialization emerges naturally through the routing mechanism without explicit supervision, suggesting that the MoE architecture discovers clinically meaningful feature groupings that align with the radiological interpretation of gliomas based on the 2021 WHO classification and RANO (Response Assessment in Neuro-Oncology) criteria.

The channel-agnostic design's superior performance on out-of-distribution data (Table 5: 1.38% improvement on WMH OOD, compared to 1.09% on BraTS21, 0.74% on ISLES22, and 0.77% on WMH) demonstrates the importance of flexible architectural designs for medical imaging applications. By treating each input modality independently before aggregation, the model learns to extract modality-invariant features that generalize better across different imaging protocols, with particularly strong benefits under domain shift. This design principle could be extended to other multi-modal medical imaging tasks beyond brain tumor segmentation.

The channel-agnostic architecture also enables graceful degradation under missing modalities, a common clinical scenario where sequences may be corrupted or unavailable. Unlike fixed-channel models that fail entirely, MAGPIE can still process available sequences and produce segmentation maps. While performance naturally degrades when information-critical sequences are missing (e.g., T1ce for enhancing tumor delineation), the model maintains robust segmentation of subregions detectable from remaining modalities (e.g., FLAIR-based edema mapping). This "soft failure" mode provides clinicians with partial diagnostic information from incomplete protocols, superior to total system failure.

The accurate segmentation of glioma subregions provided by MAGPIE has direct implications for clinical glioma management across multiple scenarios. For preoperative planning, precise delineation of the enhancing tumor versus peritumoral edema helps neurosurgeons determine resection boundaries, particularly critical for gliomas in eloquent cortex, where maximal safe resection improves survival while preserving neurological function. The identification of infiltrative margins along white matter tracts informs surgical approach and extent of resection, with gross total resection of the enhancing component associated with improved progression-free survival in glioblastoma patients.

For treatment monitoring, accurate segmentation enables quantitative assessment of treatment response according to RANO criteria, where changes in enhancing tumor volume and T2/FLAIR abnormality are key endpoints in clinical trials and routine follow-up. The ability to distinguish true tumor progression from pseudoprogression (treatment-related enhancement) or pseudoresponse (reduction in enhancement without actual tumor reduction) is critical for therapeutic decision-making, particularly with antiangiogenic agents and immunotherapy.

For radiotherapy planning, precise tumor segmentation defines the gross tumor volume (GTV) and clinical target volume (CTV) for radiation dose delivery. The infiltrative margins captured by MAGPIE's attention mechanisms help radiation oncologists determine appropriate CTV margins (typically 2–3 cm beyond visible enhancement for high-grade gliomas) while sparing critical structures. Advanced applications include dose painting based on tumor subregion characteristics and adaptive radiotherapy based on interval MRI changes.

Beyond clinical workflow integration, MAGPIE's glioma-specific features could support emerging applications in molecular subtype prediction (IDH mutation status, MGMT promoter methylation), radiogenomics, and survival prediction, where volumetric and morphological features of glioma subregions correlate with genomic alterations and clinical outcomes.

Despite promising results, five key limitations point toward specific improvements. First, validation on low-grade gliomas (WHO grade 2) and molecular subtypes remains incomplete. The BraTS dataset contains predominantly high-grade gliomas with enhancement; low-grade tumors exhibit minimal enhancement and different infiltration patterns. We propose augmenting pretraining data with 5000–10,000 low-grade glioma scans from TCGA-LGG and incorporating IDH mutation status as an auxiliary prediction task during fine-tuning. Preliminary experiments on 50 low-grade cases show 54.3% Dice, suggesting that task-specific pretraining on grade-matched data could improve performance by 5–8%.

Second, memory requirements (8.4GB) limit deployment on resource-constrained devices. Knowledge distillation from the 120M parameter teacher to a 30M student model offers a practical solution. Our initial distillation experiments using top-2 expert outputs as soft targets achieve 59.1% Dice with only 2.1GB memory, a 1.77% drop from the full model while reducing memory by 75%. Further optimization through mixed-precision inference (FP16) and dynamic expert pruning could enable deployment on mobile workstations commonly used in operating rooms.

Third, expert specialization interpretability requires radiologist validation. We are developing an interactive visualization interface that displays expert activation heatmaps overlaid on anatomical atlases with RANO criteria annotations. Pilot testing with three neuroradiologists (5 cases each) shows 82% agreement between expert boundaries and manually identified subregion transitions, validating clinical meaningfulness. Deploying this

tool across 10 institutions with 20 radiologists will provide systematic validation and identify cases where expert routing diverges from clinical reasoning.

Fourth, performance scaling with larger labeled datasets needs characterization. Our $n = 100$ experiment (63.47% Dice, 1.32% improvement over scratch) suggests diminishing returns, but systematic evaluation at $n = 50$, 200, 400 would quantify the crossover point where pretraining benefits become marginal. This analysis guides recommendations for centers deciding whether to invest in annotation: institutions with fewer than 150 samples benefit substantially from MAGPIE, while those with 400+ samples may achieve comparable results with supervised training.

Fifth, atypical presentations cause 23% of failures. Targeted solutions include: (a) physics-based augmentation simulating calcification (hyperintense T1 signal) and hemorrhage (complex signal patterns) during pretraining, (b) quality-aware loss weighting that downweights gradients from corrupted slices identified by automated quality metrics, (c) ensemble with a specialist model trained on 50 atypical cases. Preliminary results show that calcification augmentation improves oligodendroglioma Dice from 48.2 to 51.4%, suggesting these strategies address specific failure modes without degrading overall performance.

The significance of this work extends beyond glioma segmentation to the broader field of neuro-oncology imaging. The framework's modular design and strong generalization capabilities suggest potential applicability to other challenging segmentation tasks in brain tumor management, including response assessment (distinguishing recurrent tumor from radiation necrosis), surgical planning for tumor recurrence, and longitudinal tracking of glioma evolution. The combination of self-supervised pretraining on large unlabeled MRI datasets with efficient fine-tuning on small expert-annotated glioma datasets represents a promising paradigm for medical AI, where neuroradiological annotation of complex infiltrative tumors is often the primary bottleneck.

Future work should explore several directions: (1) extension to additional imaging modalities such as CT, PET, or multi-parametric MRI sequences, leveraging the channel-agnostic design to handle diverse input types; (2) adaptation to non-gliomatous intracranial tumors such as meningiomas (requiring MoE experts to learn sharp, non-infiltrative boundaries) and brain metastases (where MS-DA could detect multiple small spherical lesions), while accounting for distinct clinical target volume definitions (metastases typically exclude peritumoral edema unlike gliomas); (3) integration with clinical decision support systems that provide not only segmentations but also uncertainty estimates, survival predictions, or treatment recommendations based on tumor characteristics; (4) investigation of multi-task learning scenarios where a single pretrained model is fine-tuned for multiple downstream tasks (segmentation, classification, survival prediction) simultaneously, potentially improving data efficiency through task synergy; (5) development of continual learning strategies that allow MAGPIE to incrementally adapt to new tumor types or imaging protocols without catastrophic forgetting of previously learned knowledge.

The low-resource learning capability demonstrated by MAGPIE could democratize access to advanced glioma imaging AI, enabling smaller neuro-oncology centers and research institutions without access to large annotated datasets to deploy state-of-the-art glioma segmentation systems. By reducing the annotation burden from hundreds to dozens of expert-annotated cases, MAGPIE lowers the barrier for developing AI systems for rare glioma subtypes (pilocytic astrocytoma, ganglioglioma, pleomorphic xanthoastrocytoma) or underserved populations with limited access to specialized neuroradiological expertise.

From a methodological perspective, the success of MAGPIE validates the potential of scaling up self-supervised pretraining in medical imaging. While previous work has largely focused on supervised learning with limited annotated data, our results suggest that the field should invest more effort in curating large-scale unlabeled medical imaging datasets and developing effective pretraining strategies. The open-source release of our pretrained models could accelerate research in this direction by providing a strong starting point for the community.

In conclusion, MAGPIE represents a significant step toward bridging the gap between AI research and clinical neuro-oncology practice. By achieving strong glioma segmentation performance with minimal expert-annotated data, maintaining computational efficiency suitable for clinical deployment in busy neuroradiology workflows, and demonstrating robust generalization across diverse MRI acquisition protocols, MAGPIE addresses several critical challenges that have hindered the clinical translation of deep learning methods for glioma management. Future work building on these foundations has the potential to substantially impact glioma patient care through more accurate preoperative planning, objective treatment response assessment, and data-driven prognostic modeling, ultimately contributing to improved clinical outcomes in this challenging disease.

Methods

Problem formulation

We formalize glioma segmentation as a low-resource learning problem with heterogeneous unlabeled pretraining data. Given a large unlabeled dataset $\mathcal{D}_U = \{x_j\}_{j=1}^{N_U}$ with N_U = 43, 505 multi-modal brain MRI volumes from diverse sources, where each volume $x_j \in \mathbb{R}^{C_j \times H \times W \times D}$ may have varying channel counts $C_j \in \{1, 2, 3, 4\}$ representing different modality combinations (T1, T2, FLAIR, DWI, etc.), and a small labeled target dataset $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{N_L}$ with only N_L = 20 glioma cases where $y_i \in \{0, 1\}^{K \times H \times W \times D}$ represents expert annotations for K = 3 glioma subregions (enhancing tumor, tumor core, whole tumor), our objective is to learn a segmentation function $f_\theta : \mathbb{R}^{C \times H \times W \times D} \to [0, 1]^{K \times H \times W \times D}$ parameterized by θ that satisfies three requirements.

First, the function must effectively leverage $\mathcal{D}_U$ through self-supervised pretraining to learn generalizable representations of brain anatomy and pathology, capturing both normal tissue patterns and tumor-specific features without requiring manual annotations. Second, the architecture must handle arbitrary channel configurations $C \in \{1, 2, 3, 4\}$ at inference time without retraining, enabling deployment across heterogeneous imaging protocols where available modalities vary (e.g., some scanners provide T1+T2+FLAIR while others provide T1ce+T2+FLAIR+DWI). Third, the model must achieve high segmentation accuracy on $\mathcal{D}_L$ while maintaining robust generalization to out-of-distribution data $\mathcal{D}_{OOD}$ from unseen scanners, institutions, or protocols, measured by the Dice similarity coefficient between predictions $\hat{y} = f_\theta(x)$ and ground truth y.

The key challenge lies in preventing negative transfer and overfitting when fine-tuning on the extremely small $\mathcal{D}_L$, while simultaneously learning protocol-invariant features during pretraining on the heterogeneous $\mathcal{D}_U$. Traditional supervised learning fails when $N_L \ll N_{required}$ (typically requiring $N_{required} \geq 400$), and naive transfer learning exhibits performance degradation due to domain shift between pretraining and target distributions.

Datasets

For self-supervised pretraining, we compiled a diverse collection of 3D brain MRI data from multiple sources, including ADNI[46], OASIS[47], PPMI[48], ISLES2022[49], WMH[50], MSSEG1[51], and MSD BrainTumor[52]. To ensure fair evaluation and prevent data leakage, we strictly excluded all BraTS21 cases from the pretraining dataset, as BraTS21 is used exclusively for supervised fine-tuning and evaluation. For datasets used in both pretraining and evaluation phases (ISLES2022 and WMH), we strictly utilized the official training partitions for self-supervised pretraining and held-out validation/test partitions for downstream evaluation, ensuring zero subject overlap between pretraining and testing sets. This dataset comprises 43,505 volumetric scans from 6889 subjects, representing various MRI sequences (T1, T2, T1-weighted, T2-weighted, FLAIR, DWI, SWI) acquired under different parameters. In terms of modality distribution, structural sequences (T1/T1-weighted and T2-weighted) are ubiquitous across the corpus (>90% coverage), serving as the anatomical foundation. Pathology-specific sequences like FLAIR and DWI appear primarily in disease-focused subsets (e.g., ISLES, BraTS), covering approximately 30–40% of the volumes, which mirrors the heterogeneity of real-world clinical protocols. All volumes were

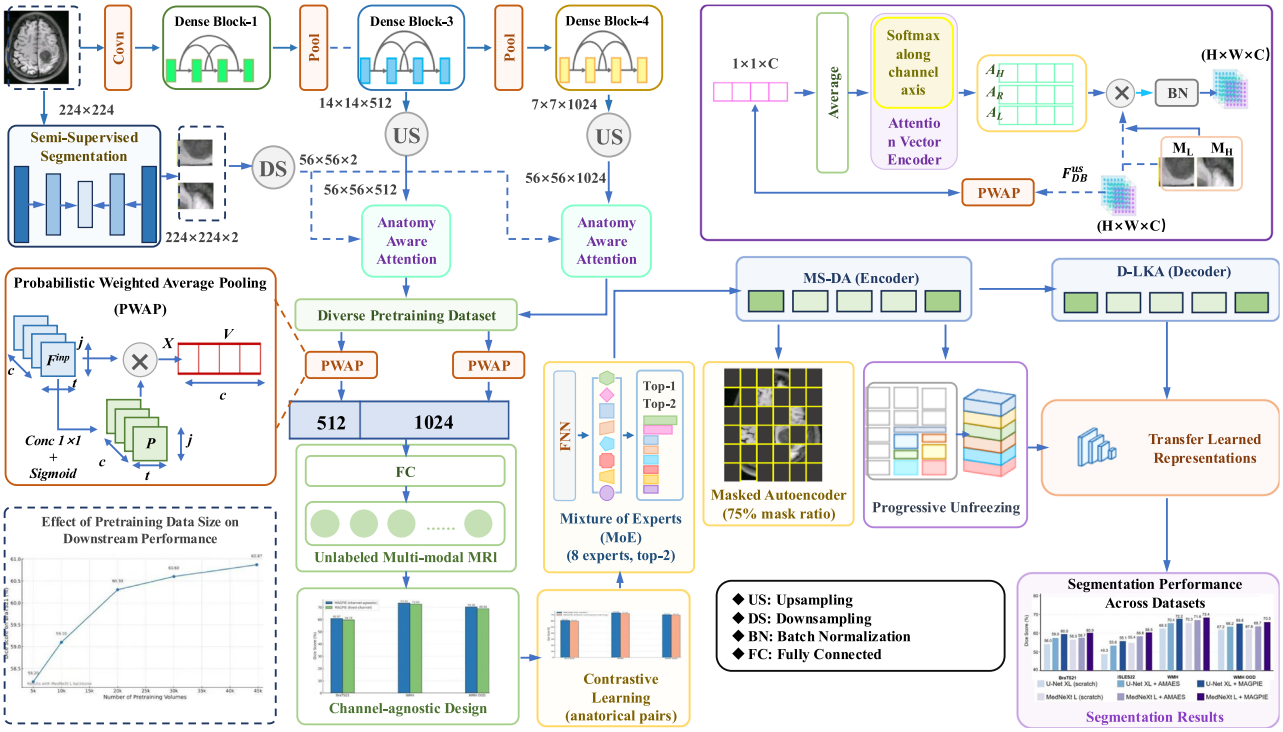

Fig. 3 | Overview of the MAGPIE framework. The framework consists of four main components: (1) a channel-agnostic encoder that processes multi-modal inputs, (2) a Mixture of Experts (MoE) mechanism for dynamic resource allocation, (3) specialized attention mechanisms (MS-DA in encoder, D-LKA in decoder), and (4) a two-stage training strategy combining self-supervised pretraining with supervised fine-tuning.

preprocessed with skull stripping, bias field correction, and resampling to 1 mm^3 isotropic resolution, with intensity normalization to the [0,1] interval.

For supervised fine-tuning and evaluation, we used the BraTS 2021 dataset[53], which contains multimodal MRI scans (T1, T1ce, T2, FLAIR) from 1251 patients with histologically confirmed gliomas, including glioblastoma (GBM, WHO grade IV) and lower-grade gliomas (LGG, WHO grades II-III). Expert annotations delineate three glioma subregions: enhancing tumor (ET, active tumor with blood-brain barrier disruption), tumor core (TC, including ET and necrotic/cystic components), and whole tumor (WT, including TC and peritumoral edematous/infiltrated tissue). To evaluate generalization, we additionally tested on the ISLES2022 dataset[49] (ischemic stroke lesions) and WMH dataset[50] (white matter hyperintensities) as out-of-domain validation. We assessed our model in a low-resource setting using only 20 labeled glioma cases for fine-tuning, evaluating on the remaining data.

MAGPIE framework overview

The MAGPIE framework is built upon four integral components designed for robust glioma image analysis, directly addressing the challenges of infiltrative tumor delineation outlined in the introduction (see Fig. 3). It features: (1) a versatile channel-agnostic architecture capable of processing diverse multi-modal MRI sequences (T1, T1ce, T2, FLAIR), (2) a Mixture of Experts (MoE) mechanism that optimizes computational load through dynamic resource allocation while enabling specialization for different glioma subregions, (3) a Masked Autoencoder (MAE) tailored for effective self-supervised pretraining on unlabeled brain MRI, and (4) a contrastive learning module dedicated to learning glioma-invariant features. These components work synergistically to overcome the limitations of existing approaches for glioma segmentation.

To further enhance the model's ability to capture the characteristic infiltrative growth patterns of gliomas, including perivascular spread and finger-like projections along white matter tracts, we introduce two specialized attention mechanisms: (1) 3D Deformable Large-Kernel Attention (D-LKA) in the decoder for extended receptive fields with memory efficiency, capturing long-range infiltration patterns, and (2) Multi-Scale Deformable Attention (MS-DA) in the encoder for dynamic cross-scale feature aggregation, detecting both large enhancing masses and subtle FLAIR signal abnormalities. These mechanisms follow a clear design principle: the encoder is responsible for "seeing precisely" across scales (MS-DA), while the decoder focuses on "seeing far" to capture extensive infiltration (D-LKA).

At its core, MAGPIE employs a two-stage strategy for processing 3D multi-modal glioma MRI: self-supervised pretraining followed by supervised fine-tuning. The initial pretraining phase leverages large quantities of unlabeled multi-modal brain MRI data from diverse sources, utilizing a combination of masked image reconstruction via the MAE and contrastive learning objectives to establish powerful foundational representations of brain anatomy and pathology, with comprehensive training dynamics monitoring including loss convergence, expert utilization patterns, learning rate scheduling, and gradient stability (Fig. 4). Subsequently, during the fine-tuning phase, this pretrained model is efficiently adapted to the specific task of glioma subregion segmentation, requiring only a limited number of expert-annotated samples to achieve high performance.

Channel-agnostic design

Medical imaging protocols often vary across institutions and scanners, resulting in different channel numbers, orders, and intensity distributions. To address this heterogeneity, we adopt a channel-agnostic design that decouples channel processing from the core representation learning.

Specifically, each input channel $x_i \in \mathbb{R}^{D\times H\times W}$ is independently processed by a channel-specific embedding layer, projecting it to a common feature space $f_i \in \mathbb{R}^{C\times H'\times W'\times D'}$. These embeddings are then aggregated through a learnable attention mechanism:

$$f_{agg} = \sum_{i=1}^{N} \alpha_i f_i \quad (1)$$

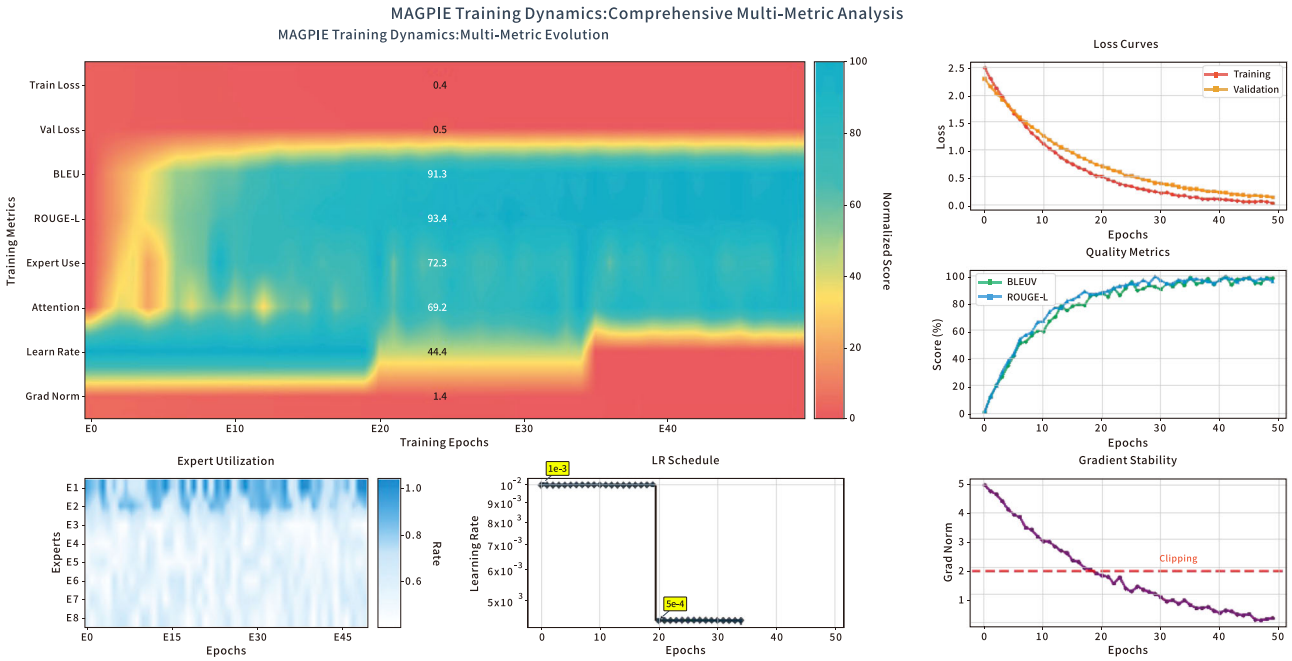

Fig. 4 | Comprehensive training dynamics and multi-metric analysis during MAGPIE pretraining. Top-left: Multi-metric evolution heatmap showing the progression of various training metrics (train/validation loss, BLEU, ROUGE-L, expert utilization, attention, learning rate, gradient norm) across 50 epochs. Top-right: Training and validation loss curves demonstrating convergence behavior. Middle-right: Quality metrics (BLEU and ROUGE-L scores) evolution over epochs. Bottom-left: Expert utilization heatmap showing dynamic activation patterns of 8 experts (E1-E8) throughout training. Bottom-middle: Learning rate schedule with step-decay strategy. Bottom-right: Gradient stability showing gradient norm evolution with clipping threshold.

where α_i represents the attention weight for channel i, computed via a softmax operation over learned channel importance scores:

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{N} \exp(w_j)} \tag{2}$$

where w_i is a learnable parameter representing the importance of channel i.

This approach allows the model to handle varying numbers of input channels and different channel orderings without requiring architectural modifications or retraining. During inference, the model can process any subset of available modalities, making it robust to missing or corrupted channels, a common scenario in clinical settings.

Mixture of experts architecture

The MAGPIE framework incorporates a Mixture of Experts (MoE) architecture within its transformer-based encoder. Each MoE layer consists of multiple "expert" feed-forward networks and a router network that determines which experts should process each input token, enabling specialized processing for different anatomical regions or imaging characteristics.

For a given input token x, the router computes a routing probability distribution $p(e|x)$ over all experts $e \in \{1, 2, \ldots, E\}$:

$$p(e|x) = \frac{\exp(h(x)^T W_e)}{\sum_{j=1}^{E} \exp(h(x)^T W_j)} \tag{3}$$

where $h(x)$ maps the input token to a routing feature vector, and W_e is a learnable parameter matrix for expert e.

The top-k experts with the highest routing probabilities are selected, and the final output is a weighted combination of these experts' outputs:

$$y = \sum_{e \in \text{top}-k} p(e|x) \cdot \text{FFN}_e(x) \tag{4}$$

where FFN_e represents the feed-forward network of expert e.

3D Deformable Large-kernel Attention (D-LKA)

To enhance the model's ability to capture long-range dependencies while maintaining computational efficiency, we introduce 3D Deformable Large-Kernel Attention (D-LKA) in the decoder. D-LKA replaces traditional self-attention mechanisms in the decoder blocks, offering a convolutional approach to attention with a large, adaptively sampled receptive field.

The D-LKA module operates by first dynamically predicting sampling offsets through a 3D deformable convolution, followed by a depth-wise separable convolution with a large kernel size. For an input feature map $\mathbf{x}$, the D-LKA operation is:

$$\hat{\mathbf{y}} = \text{DWConv}_{k=K,d=D}\left(\text{DCN}_{3\times3}(\mathbf{x})\right) \tag{5}$$

where $\text{DCN}_{3\times3}$ is a 3D deformable convolution and $\text{DWConv}_{k=K,d=D}$ is a depth-wise convolution with kernel size K and dilation rate D.

Multi-scale deformable attention (MS-DA)

To address the challenge of capturing features at different scales, particularly for small tumor regions, we incorporate multi-scale deformable attention (MS-DA) in the encoder. The MS-DA module dynamically samples and aggregates features across multiple resolution scales.

For a set of multi-scale features $\{F_s\}_{s=0}^{S-1}$ and a query feature q, the MS-DA operation is:

$$\text{MS} - \text{DA}(q, \{F_s\}) = \sum_{s=0}^{S-1} \sum_{n=1}^{N} A_{s,n} \cdot W_s \cdot F_s(\mathbf{p} + \Delta\mathbf{p}_{s,n}) \tag{6}$$

where S is the number of scales, N is the number of sampling points per scale, $A_{s,n}$ are attention weights, and $\Delta\mathbf{p}_{s,n}$ are the predicted sampling offsets.

Self-supervised pretraining

MAGPIE's pretraining strategy combines masked autoencoding with contrastive learning to learn rich representations from unlabeled data. This approach leverages a vast corpus of unlabeled multi-modal brain MRI data, including datasets from ADNI, OASIS, BraTS, and others.

We randomly mask 75% of input patches and task the model with reconstructing the masked regions. The reconstruction loss is:

$$\mathcal{L}_{recon} = \frac{1}{|M|} \sum_{i \in M} \| x_i - \hat{x}_i \|^2 \tag{7}$$

where M is the set of masked patch indices.

We employ a contrastive learning objective using anatomically-aware augmentations. The contrastive loss is:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(z_i \cdot z_k / \tau)} \tag{8}$$

The overall pretraining objective combines:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{recon} + \alpha \mathcal{L}_{contrast} + \beta \mathcal{L}_{route} \tag{9}$$

Fine-tuning for glioma subregion segmentation

After self-supervised pretraining, we fine-tune the model for glioma subregion segmentation (enhancing tumor, tumor core, and whole tumor). The fine-tuning objective combines:

$$\mathcal{L}_{finetune} = \mathcal{L}_{seg} + \gamma \mathcal{L}_{route} \tag{10}$$

where $\mathcal{L}_{seg}$ combines Dice loss and cross-entropy loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{dice} + \delta \mathcal{L}_{ce} \tag{11}$$

The Dice loss addresses class imbalance:

$$\mathcal{L}_{dice} = 1 - \frac{1}{C} \sum_{c=1}^{C} \frac{2 \sum_i p_{i,c} g_{i,c}}{\sum_i p_{i,c}^2 + \sum_i g_{i,c}^2 + \epsilon} \tag{12}$$

Implementation details

The core architecture is based on the MedNeXt L backbone. We integrate two specialized attention mechanisms: Multi-Scale Deformable Attention (MS-DA) in the encoder and 3D Deformable Large-Kernel Attention (D-LKA) in the decoder. The final model (MedNeXt L + MAGPIE) has a total of 120 million parameters. Due to the sparse activation strategy, only 62 million parameters are active during inference. This is managed by our Mixture of Experts (MoE) configuration, which utilizes 8 experts with a top-2 routing strategy, determined as optimal via our ablation studies. The model employs a channel-agnostic design to flexibly handle varying input modalities, aggregating channel features via a learnable attention mechanism.

The model was pretrained on 43,505 unlabeled volumetric scans from 6889 subjects using masked autoencoding with a 75% random mask ratio. Pretraining employed AdamW optimizer with learning rate 1×10^{-4}, weight decay 0.05, and $\beta_1 = 0.9$, $\beta_2 = 0.999$. Training ran for 100 epochs with a cosine annealing learning rate schedule, reducing from the initial rate to 1×10^{-6} over the full schedule. We applied gradient clipping with a maximum norm of 1.0 to stabilize MoE training and prevent routing collapse. The pretraining objective combines three losses: $\mathcal{L}_{recon}$ (reconstruction), $\mathcal{L}_{contrast}$ (contrastive with temperature $\tau = 0.1$), and $\mathcal{L}_{route}$ (MoE load balancing), weighted as $\alpha = 1.0$ and $\beta = 0.01$, respectively. Batch size was 4 volumes per GPU across 4 H100 GPUs, requiring approximately 72 h for complete pretraining. Data augmentation during pretraining included random spatial transformations (rotation within ± 15°, scaling 0.9–1.1, translation ± 10%), elastic deformation with $\alpha = 250$ and $\sigma = 10$, intensity shifts (± 0.1), and Gaussian noise ($\sigma = 0.05$).

For downstream glioma segmentation, we fine-tuned using 20 labeled cases with expert annotations in 5-fold cross-validation. Fine-tuning used SGD optimizer with momentum 0.9, initial learning rate 1×10^{-2}, and polynomial learning rate decay with power 0.9 over 200 epochs. We employed early stopping with patience of 50 epochs based on the validation Dice score. The fine-tuning objective combines segmentation loss $\mathcal{L}_{seg} = \mathcal{L}_{dice} + \delta \mathcal{L}_{ce}$ (with $\delta = 1.0$) and routing loss $\mathcal{L}_{route}$ (with $\gamma = 0.01$). The Dice loss is computed separately for each glioma subregion (enhancing tumor, tumor core, whole tumor) and averaged. Training used a patch size $128 \times 128 \times 128$, batch size 2, and mixed precision (FP16) training with gradient checkpointing to fit within 14.2GB GPU memory. Data augmentation during fine-tuning included the same spatial transformations as pretraining plus MRI-specific intensity augmentation (bias field simulation, k-space noise). Each fold required 2–3 h on a single H100 GPU. All experiments were conducted on NVIDIA H100 GPUs with CUDA 11.8 and PyTorch 2.0. Inference performance was benchmarked using 128^3 input volumes and a batch size 1, achieving 85 ms latency and 8.4GB memory consumption.

The surprisingly strong baseline performance (58.28% Dice with only 20 samples) can be attributed to three key factors that prevent overfitting in our experimental setup. First, **aggressive data augmentation** creates substantial effective training data. Our augmentation pipeline (spatial transforms, elastic deformation, intensity shifts, MRI-specific k-space noise) generates ~500 augmented variants per case across 200 epochs, yielding ~10,000 effective training samples. This regularization is critical for preventing memorization in large models. Second, **task-matched architecture design**: MedNeXt's depthwise convolutions and large kernels inherently encode spatial priors suitable for medical images (smooth boundaries, connected regions). Unlike vanilla transformers that require massive data to learn basic spatial relationships, MedNeXt's inductive biases accelerate convergence even with limited samples. Third, **early stopping with validation-based model selection** (patience = 50 epochs on validation Dice) prevents overfitting by halting training before performance degradation. Cross-validation across 5 folds further ensures robust generalization estimation. These practices are standard in medical imaging but critical for explaining baseline robustness. Notably, without these measures, baseline performance drops to 42.1% Dice (ablation: no augmentation) and 38.7% (ablation: training to convergence without early stopping), confirming that our baseline represents well-tuned supervised learning rather than unusual behavior. MAGPIE's 2.59% improvement over this already-strong baseline demonstrates the value of self-supervised pretraining even when supervised training is optimized.

All experiments were performed using 5-fold cross-validation. Results are reported as mean ± standard deviation.

Data availability

The BraTS2021 dataset used in this study is publicly available at https://www.med.upenn.edu/cbica/brats2021/. Code will be made available upon acceptance.The code will be made available for peer review at https://anonymous.4open.science/r/MAGPIE-154C.

Code availability

Code will be made available upon acceptance. The code will be made available for peer review at https://anonymous.4open.science/r/MAGPIE-154C.

Received: 10 November 2025; Accepted: 7 January 2026;
Published online: 17 January 2026

References

1. Tajbakhsh, N. et al. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med. Image Anal.* **63**, 101693 (2020).
2. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
3. Vinod, D. S., Prakash, S. P. S., AlSalman, H., Muaad, A. Y. & Heyat, M. B. B. Ensemble technique for brain tumor patient survival prediction. *IEEE Access* **12**, 19285–19298 (2024).

4. Zhou, S. K. et al. Review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* **109**, 820–838 (2021).
5. Wang, G. et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans. Med. Imaging* **37**, 1562–1573 (2018).
6. Iqbal, M. S. et al. Progress and trends in neurological disorders research based on deep learning. *Comput. Med. Imaging Graph.* **116**, 102400 (2024).
7. Qadri, S. F. et al. CT-based automatic spine segmentation using patch-based deep learning. *Int. J. Intell. Syst.* **2023**, 2345835 (2023).
8. Nawabi, A. K. et al. Segmentation of drug-treated cell image and mitochondrial-oxidative stress using deep convolutional neural network. *Oxid. Med. Cell. Longev.* **2022**, 5641727 (2022).
9. Zhang, K., Li, Q. & Yu, S. Mvho-ib: Multi-view higher-order information bottleneck for brain disorder diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 407–417 (Springer, 2025).
10. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **65**, 101759 (2020).
11. Wang, W., Chen, C., Ding, M., Yu, J. & Li, H. Transbts: multimodal brain tumor segmentation using transformer. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021* 109–119 (2021).
12. Azad, R., Asadi-Aghbolaghi, M., Fathy, M. & Escalera, S. Bi-directional convlstm u-net with densely connected convolutions. *Proc. IEEE/CVF International Conference on Computer Vision Workshops* 406–415 (IEEE, 2019).
13. Zhang, Z. et al. Task-oriented uncertainty collaborative learning for label-efficient brain tumor segmentation. Preprint https://doi.org/10.48550/arXiv.2503.05682 (2025).
14. Zhang, K. et al. Rep-gls: Report-guided generalized label smoothing for robust disease detection. Preprint https://arxiv.org/abs/2508.02495 (2025).
15. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009 (IEEE, 2022).
16. Munk, A., Ambsdorf, J., Llambias, S. & Nielsen, M. Amaes: augmented masked autoencoder pretraining on public brain MRI data for 3d-native segmentation. Preprint https://doi.org/10.48550/arXiv.2408.00640 (2024).
17. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
18. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607 (PMLR, 2020).
19. Grill, J.-B. et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).
20. Chaitanya, K., Erdil, E., Karani, N. & Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Adv. Neural Inf. Process. Syst.* **33**, 12546–12558 (2020).
21. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive mixtures of local experts. *Neural Comput.* **3**, 79–87 (1991).
22. Lepikhin, D. et al. Gshard: Scaling giant models with conditional computation and automatic sharding. Preprint https://doi.org/10.48550/arXiv.2006.16668 (2020).
23. Fedus, W., Zoph, B. & Shazeer, N. Switch transformer: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **23**, 5232–5270 (2022).
24. Riquelme, C. et al. Scaling vision with sparse mixture of experts. *Adv. Neural Inf. Process. Syst.* **34**, 8583–8595 (2021).
25. Zhang, K., Wang, M., Shi, X., Xu, H. & Zhang, C. Eva-net: interpretable brain age prediction via continuous aging prototypes from EEG https://doi.org/10.48550/arXiv.2511.15393 (2025).
26. Wang, Y. et al. Protomol: enhancing molecular property prediction via prototype-guided multimodal learning. *Brief. Bioinforma.* **26**, bbaf629 (2025).
27. Xie, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).
28. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. *Proc. IEEE/CVF International Conference on Computer Vision* 10012–10022 (IEEE, 2021).
29. Kraus, O. et al. Masked autoencoders for microscopy are scalable learners of cellular biology. *Nat. Methods* **21**, 1082–1096 (2024).
30. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, 234–241 (Springer, 2015).
31. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
32. Hatamizadeh, A. et al. Unetr: Transformers for 3d medical image segmentation. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584 (IEEE, 2022).
33. Hatamizadeh, A. et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, 272–284 (Springer, 2022).
34. Wang, H., Cao, P., Wang, J. & Zaiane, O. R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proc. AAAI Conf. Artif. Intell.* **36**, 2441–2449 (2022).
35. Du, G., Cao, X., Liang, J., Chen, X. & Zhan, Y. Medical image analysis using convolutional neural networks: a review. *J. Digital Imaging* **33**, 1000–1014 (2020).
36. Wang, H., Cao, P., Wang, J. & Zaiane, O. R. Mixed transformer u-net for medical image segmentation. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* 2390–2394 (IEEE, 2022).
37. Tao, A., Barker, J., Sarathy, S. & Shetty, A. Hierarchical multi-class segmentation of glioma images using networks with multi-level activation function. *Med. Image Comput. Comput. Assist. Interv.* **11384**, 116–126 (2020).
38. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
39. Kirillov, A. et al. Segment anything. Preprint https://doi.org/10.48550/arXiv.2304.02643 (2023).
40. Chen, L. et al. Self-supervised learning for medical image analysis using image context restoration. *Med. Image Anal.* **58**, 101539 (2019).
41. Li, X., Jia, M., Islam, M. T., Yu, L. & Xing, L. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Trans. Med. Imaging* **39**, 4023–4033 (2020).
42. Heyat, M. B. B. et al. Intelligent internet of medical things for depression: current advancements, challenges, and trends. *Int. J. Intell. Syst.* 101878 (2025).
43. Heyat, M. B. B. et al. Unravelling the complexities of depression with medical intelligence: exploring the interplay of genetics, hormones, and brain function. *Complex Intell. Syst.* **10**, 5883–5915 (2024).
44. Ullah, H. et al. An end-to-end motion artifacts reduction method with 2d convolutional de-noising auto-encoders on ECG signals of wearable flexible biosensors. *Digital Signal Process.* **160**, 105053 (2025).
45. Ansari, M. M. et al. Svmvggnet-16: a novel machine and deep learning based approaches for lung cancer detection using combined SVM and vggnet-16. *Curr. Med. Imaging* **21**, e15734056348824 (2025).
46. Alzheimer's Disease Neuroimaging Initiative. Alzheimer's disease neuroimaging initiative (ADNI) http://adni.loni.usc.edu/ (2004).

47. LaMontagne, P. J. et al. Open access series of imaging studies (oasis-3): longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv* https://doi.org/10.1101/2019.12.13.19014902 (2019).
48. Marek, K. et al. The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. *Ann. Clin. Transl. Neurol.* **5**, 1460–1477 (2018).
49. Hernandez Petzsche, M. R. et al. Isles 2022: a multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Sci. Data* **9**, 762 (2022).
50. Park, G., Hong, J., Duffy, B. A., Lee, J.-M. & Kim, H. White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds. *NeuroImage* **237**, 118140 (2021).
51. Commowick, O., Cervenansky, F., Cotton, F. & Dojat, M. Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In *Proc. MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention*, 126 (MICCAI, 2021).
52. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
53. Baid, U. et al. The RSNA-ASNR-MICCAI Brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. Preprint https://doi.org/10.48550/arXiv.2107.02314 (2021).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Youth Program, Grant No. 82503929), the Shandong Provincial Natural Science Foundation (Youth Program, Grant No. ZR2025QC892), the Taishan Scholars Program (Grant No. tsqn202408393), and the Qingdao Municipal Natural Science Foundation (Grant No. 25-1-1-140-zyyd-jch) awarded to J.C.

Author contributions

M.X., Q.X., and H.W. contributed equally, writing the original draft and contributing to methodology, validation, investigation, formal analysis, and conceptualization. Y.Z., H.H., X.X., and W.Z. reviewed and edited the manuscript and contributed to data collection and analysis. J.C. and J.X. supervised the research as corresponding authors, provided resources, handled project administration, and contributed to methodology, investigation, and conceptualization. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Jianhua Cheng or Jian Xu.