Learning Invariant Anatomical Representations from Multi-Sequence MRI for Brain Segmentation

Jianwen Liang[1], Pujin Cheng[1,2], and Xiaoying Tang[1,3⋆ (✉)]

[1] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
[2] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China
[3] Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China

Abstract. Self-supervised learning (SSL) has emerged as a powerful paradigm in medical image analysis, but existing SSL frameworks for multi-sequence MRI often fail to exploit inherent cross-sequence structural representations or remain agnostic to pathological heterogeneity. In this paper, we propose Text-guided Invariant Anatomical Learning (TeInL), a novel SSL framework for multi-sequence brain MRI. TeInL decouples image representations into sequence-invariant anatomical features and sequence-specific style features, enforces structural consistency by text-image alignment, learns anatomical consistency, and performs cross-sequence reconstruction. We integrate anatomical phenotypes from brain atlases and pathological signatures from unsupervised lesion segmentation into our SSL framework via image-text alignment. Aligning anatomy and pathology-aware textual descriptions with visual features enables the model to capture fine-grained correspondences between complex structural phenotypes and semantic representations. Extensive experiments across six benchmark datasets covering multi-granularity brain parcellations, stroke lesions, and brain tumors demonstrate that TeInL consistently outperforms state-of-the-art SSL methods and reduces annotation efforts by at least 40%. Code and pre-trained models will be available at https://github.com/liangjianwen01/MICCAI2026_TeInL.

Keywords: Self-supervised learning · Multi-sequence MRI · Brain segmentation · Contrastive language-image pre-training.

1 Introduction

Accurate brain segmentation from magnetic resonance imaging (MRI) requiring expert voxel-level annotations, which are prohibitively expensive and prone to inter-observer variability. Self-Supervised Learning (SSL) has emerged as a compelling paradigm to mitigate the dependency on large-scale annotated datasets.

⋆ Corresponding Author (tangxy@sustech.edu.cn).

While generic SSL strategies have shown promise in medical imaging [23], they often treat multi-sequence data as independent samples, failing to exploit structural correlations across sequences. Clinically, although multiple MRI sequences are desired [2], diagnostic priorities often lead to a focus on specific sequences for high-resolution imaging. For instance, routine neurological screening may prioritize T1-weighted (T1w), whereas tumor-related diagnostics emphasize T1-weighted contrast-enhanced (T1ce). Furthermore, retrospective researches provide abundant multi-sequence MRI data, characterized by significant sequence inconsistency, such as, the four-sequence for brain tumor patients in BraTS [2] and the dual-sequence for healthy cohorts in HCP [18]. Existing multi-modal methods predominantly rely on fusion strategies that mandate the availability of a complete set of sequences during training and inference. Therefore, there is a critical need for a unified SSL framework capable of leveraging diverse and potentially incomplete and inconsistent multi-sequence MRI data.

On the other hand, advances have underscored the value of injecting medical priors into SSL. MDM [13] learns brain deformation to enhance representational capability. Meanwhile, contrastive language-image pre-training (CLIP) has inspired innovations in medical imaging. MCLIM [11] leverages Large Language Models (LLMs) to generate anatomical descriptions from brain atlases, and then incorporates those atlas priors through vision-language alignment. However, these methods are primarily tailored for single-sequence and rely strictly on healthy templates, remaining agnostic to the structural correlations across multi-sequence MRI and the presence of pathological anomalies. To exploit the richness of cross-sequence data, BrainMVP [15] introduces a multi-sequence paradigm that aggregates features via learnable sequence-specific templates. Nevertheless, its static templates struggle to accommodate high pathological heterogeneity, exhibiting limited sensitivity to fine-grained abnormalities. Consequently, there remains a critical gap for a unified SSL framework that can simultaneously capture cross-sequence anatomical invariants and pathology-aware variations.

In such context, we propose Text-guided Invariant Anatomical Learning (TeInL), a novel SSL framework for multi-sequence structural brain MRI. TeInL leverages multi-sequence anatomical invariants for intrinsic supervision while integrating rich anatomical priors from brain atlases and pathological insights from unsupervised lesion segmentation. These multi-source priors are injected into TeInL via granular textual descriptions. While these sources may lack voxel-level precision, they offer sufficient localization fidelity to provide robust semantic guidance. As illustrated in Fig. 1, TeInL consists of three core components: (1) a disentangled representation learning scheme that decouples anatomy from style, enforcing the capture of invariant anatomical patterns through cross-sequence reconstruction and structural consistency; (2) a semantic alignment module that establishes fine-grained correspondences between visual anatomical features and text embeddings; and (3) a style-alignment mechanism that maps sequence-specific style prototypes to their corresponding textual descriptions. TeInL achieves state-of-the-art (SOTA) performance across six benchmark datasets, while significantly reducing the annotation efforts.

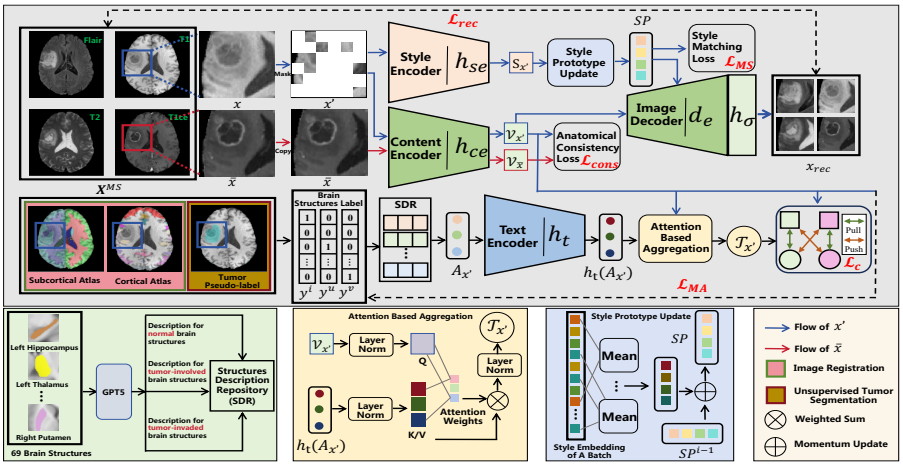

Fig. 1. We illustrate the detailed of the structures description repository (SDR), and four pre-text tasks: image reconstruction, anatomical consistency learning, image-text alignment, anatomy and style matching.

2 Methodology

2.1 Structures Description Repository

To address the scarcity of paired image-text data in neuroimaging, we construct a Structures Description Repository (SDR) by synthesizing textual characterizations from brain atlases that represent expert consensus on neuroanatomical parcellation. Given that our pre-training cohort includes both healthy individuals and tumor patients, relying exclusively on healthy atlas descriptions would be clinically incongruent. Thus, we extend the semantic space to characterize each brain structure in three pathological states: **healthy**, **mildly tumor-invaded**, and **severely tumor-invaded**. We use GPT-5 [16] to transform raw atlas-based parcellations into structured anatomical knowledge. Specifically, for each anatomical structure A defined in the Harvard-Oxford cortical and subcortical atlases [7], we prompt GPT-5 to generate detailed attributes, including its shape and anatomical location. These attributes are then populated into a predefined template: "This is a human brain MRI patch including A that is ⟨status of A⟩, its shape is ⟨shape description⟩, it ⟨anatomical location⟩." This process yields a library of structured, pathology-aware textual representations that facilitate fine-grained vision-language alignment.

2.2 Disentangled Representation Learning

Let $\mathbf{X}^{MS} \in \mathbb{R}^{s\times H\times W\times D}$ denote the multi-sequence brain MRI with s ($s \geq 2$) sequences, rigidly registered to the MNI152 standard space, and (H, W, D) denote the spatial dimensions. During pre-training, a single sequence image is

randomly selected and then cropped into a volumetric patch $x \in \mathbb{R}^{1\times h\times w\times d}$. We apply a random masking strategy with a 75% ratio to x, yielding a masked view x'. A dual-branch encoder is then employed to process x': the content encoder extracts sequence-invariant anatomical features $\mathcal{V}_{x'} = h_{ce}(x')$, while the style encoder captures sequence-specific style features $\mathcal{S}_{x'} = h_{se}(x')$.

Dynamic Style Prototypes. To model the intensity distribution of each sequence, we maintain a set of learnable style prototypes $\mathcal{SP} = \{p^k\}_{k=1}^{s}$, where p^k represents the style embedding for the k-th sequence. As illustrated in Fig. 1, these prototypes are randomly initialized and updated during pre-training via a momentum-based strategy. updated via a momentum strategy. For a given sequence k in a mini-batch $\mathcal{B}$, we first compute the centroid of the style features:

$$\hat{p}^k = \frac{1}{N_k} \sum_{x' \in \mathcal{B}_k} h_{se}(x'), \tag{1}$$

where N_k is the number of samples of sequence k. The prototype p^k is then updated as $p^k \leftarrow m \cdot p^k + (1-m) \cdot \hat{p}^k$, with the momentum coefficient $m = 0.99$.

Cross-Sequence Reconstruction. To ensure that the encoded anatomical features are sequence-agnostic, we introduce cross-sequence reconstruction as a pre-text task. We use Adaptive Instance Normalization (AdaIN) to inject style prototypes $\mathcal{SP}$ into the decoder. By conditioning the anatomical features $\mathcal{V}_{x'}$ on $\mathcal{SP}$, the network is tasked with reconstructing co-located patches across all s sequences, $x_{rec} = h_\sigma(d_e(\mathcal{V}_{x'}, \mathcal{SP}))$, $x_{rec} \in \mathbb{R}^{s\times h\times w\times d}$, where h_σ is the intensity head. The reconstruction loss is defined by the $L2$ norm:

$$\mathcal{L}_{rec} = \sum_{k=1}^{s} ||x_{rec}^k - x^k||^2. \tag{2}$$

Here, x^k denotes the original (unmasked) patch of the k-th sequence. This mechanism forces $\mathcal{V}_{x'}$ to retain only structural information, as the reconstruction of any sequence's intensity profile must rely exclusively on the guidance provided by the style prototypes.

Anatomical Consistency Learning. To further ensure that the content encoder extracts sequence-invariant representations, we introduce an anatomical consistency learning task. For x' from sequence k, we crop a spatially aligned counterpart $\bar{x}$ from a randomly selected different sequence j ($j \neq k$) of the same subject. Since the multi-sequence images are rigidly co-registered, x' and $\bar{x}$ share identical anatomical structures. We obtain the anatomical features of $\bar{x}$: $\mathcal{V}_{\bar{x}} = h_{ce}(\bar{x})$. To enforce consistency in the latent space and prevent representation collapse, we adopt a symmetric consistency loss inspired by SimSiam [6]. Specifically, a predictor head G is used to map the anatomical features of one sequence to the embedding space of another, and cosine similarity $\mathcal{D}$ is used to measure the distance between the two embeddings. The anatomical consistency loss $\mathcal{L}_{cons}$ is defined as:

$$\mathcal{L}_{cons}(\mathcal{V}_{x'}, \mathcal{V}_{\bar{x}}) = -\frac{1}{2}\left[\mathcal{D}(G(\mathcal{V}_{x'}), \text{stopgrad}(\mathcal{V}_{\bar{x}})) + \mathcal{D}(G(\mathcal{V}_{\bar{x}}), \text{stopgrad}(\mathcal{V}_{x'}))\right]. \tag{3}$$

2.3 Image-Language Alignment Learning

Text Embedding. Since $\mathbf{X}^{MS}$ is registered to the MNI152 space, we apply the same spatial cropping to the brain atlases as performed for x' to identify the constituent anatomical structures. Meanwhile, we employ LF-SynthSeg [20], an unsupervised algorithm, to generate tumor pseudo-labels for brain tumor cohorts. The LF-SynthSeg pseudo labels are used only during pre-training to provide pathology-aware textual guidance, ensuring tumor-containing samples are described in a clinically consistent manner. To ensure reliability, samples with a pseudo-label volume below 500 voxels are excluded. For each structure $A_{x'}$ within patch x', we calculate the infiltration ratio r based on the pseudo-label. The pathological status is categorized into three levels: *healthy* (y^i if $r = 0$), *mildly invaded* (y^u if $0 < r \leq 50\%$), and *severely invaded* (y^v if $r > 50\%$). Depending on the categorized status, we retrieve the corresponding structured textual description for $A_{x'}$ from the SDR. These descriptions are then encoded using the pre-trained text encoder h_t from BiomedCLIP [21]. Given that a single patch often encompasses multiple structures, we implement an attention-based integration mechanism to aggregate multi-structural information into a unified semantic representation $\mathcal{T}_{x'}$. Specifically, $\mathcal{V}_{x'}$ is utilized as the Query, while the text embeddings of $A_{x'}$ serve as the Key and Values.

Anatomy-Text Alignment. We align the anatomical features with the texture features by the contrastive objective:

$$\mathcal{L}_c^{T\leftarrow I}(\mathcal{T}_{x'}, \mathcal{V}_{x'}) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{exp(\mathcal{T}_{x'}^i \cdot {\mathcal{V}_{x'}^i}^T/\tau)}{\sum_{j=1}^{B} exp(\mathcal{T}_{x'}^i \cdot {\mathcal{V}_{x'}^j}^T/\tau)}, \tag{4}$$

$$\mathcal{L}_c = \mathcal{L}_c^{T\leftarrow I} + \mathcal{L}_c^{I\leftarrow T}, \tag{5}$$

where B is the batch size, τ is the temperature parameter and $\mathcal{L}_c^{I\leftarrow T}$ is the image-to-text counterpart of $\mathcal{L}_c^{T\leftarrow I}$. Before computing the contrastive loss, both $\mathcal{T}_{x'}$ and $\mathcal{V}_{x'}$ are normalized using the $L2$ norm. By utilizing masked image patches, the model is compelled to infer missing spatial structures to achieve effective cross-modal alignment. To further strengthen the fine-grained correlation between visual features and the characteristics and pathological status of brain anatomy, we introduce a pathology-conditioned anatomy matching loss $\mathcal{L}_{MA}$. Specifically, we define a multi-label ground truth $y = [y^i, y^u, y^v] \in \{0, 1\}^{3L}$, where L is the number of brain structures in atlases, indicating all structures in three pathological states. The matching loss is formulated using binary cross-entropy:

$$\mathcal{L}_{MA}(y, y') = -\frac{1}{3L}\sum_{l=1}^{3L} \left[y_l \log y'_l + (1 - y_l) \log(1 - y'_l)\right], \tag{6}$$

where $y'_l = \sigma(\mathcal{D}(\mathcal{V}_{x'}, h_t(A_l)))$ represents the predicted probability, indicated by the similarity between the visual anatomical feature and the text embedding of a structure-status pair $A_l \in$ SDR. Here, σ denotes the sigmoid activation function. Similarly, we employ a style matching loss $\mathcal{L}_{MS}$, with the same form

Table 1. Quantitative DSC ($mean_{std}$) comparisons with SOTA SSL methods on brain segmentation. Datasets A, B, C, D, E, and F are respective Mindboggle-101, CADNI, JHU, BraTS2021, BraTS2021-MEN, and ATLAS2, with the number of segmentation regions in "()". The best results are highlighted in **bold**, and the second-best results are underlined. "*" indicates $p < 0.05$ in a Wilcoxon matched-pairs signed rank test when comparing the best and the second-best methods.

Methods	Brain Structure				Brain Tumor			Stroke Lesion	Average
	A (101)	B (30)	C (289)	Average	D (3)	E (3)	Average	F (1)	
Scratch	$76.71_{2.95}$	$86.86_{5.28}$	$80.24_{7.86}$	81.27	$75.98_{21.11}$	$70.40_{25.78}$	73.19	$52.13_{29.79}$	73.72
MG (2021) [23]	$80.07_{1.68}$	$86.08_{7.06}$	$77.06_{11.83}$	81.07	$76.01_{21.42}$	$70.46_{25.92}$	73.23	$51.91_{29.59}$	73.59
Tang et al. (2022) [17]	$80.58_{2.44}$	$87.09_{5.97}$	$80.92_{5.68}$	82.86	$79.01_{20.83}$	$74.92_{22.98}$	76.96	$54.59_{31.37}$	76.18
PCRLv2 (2023) [22]	$80.41_{1.82}$	$86.91_{5.16}$	$80.63_{6.74}$	82.65	$80.38_{19.51}$	$73.40_{24.30}$	76.89	$55.50_{29.76}$	76.20
MDM (2025) [13]	$\underline{80.70}_{2.52}$	$\underline{87.25}_{4.83}$	$\mathbf{81.17}_{5.58}$	$\underline{83.04}$	$\underline{80.88}_{21.19}$	$72.58_{24.77}$	76.73	$55.54_{28.71}$	76.35
BrainMVP (2025) [15]	$80.60_{1.75}$	$86.93_{4.48}$	$80.90_{5.62}$	82.81	$80.95_{20.11}$	$\underline{76.26}_{22.03}$	$\underline{78.60}$	$\underline{55.91}_{29.61}$	$\underline{76.92}$
TeInL (Ours)	$\mathbf{81.30}^{*}_{1.61}$	$\mathbf{87.34}^{*}_{4.79}$	$\underline{81.16}_{5.87}$	$\mathbf{83.26}^{*}$	$\mathbf{81.74}^{*}_{20.78}$	$\mathbf{77.25}^{*}_{23.56}$	$\mathbf{79.49}^{*}$	$\mathbf{57.25}^{*}_{28.58}$	$\mathbf{77.67}^{*}$

as the anatomy-text matching loss (Eq. 6), to align the style prototypes p^k with their respective textual descriptions generated by GPT-5. The entire framework is pre-trained end-to-end by minimizing the multi-task loss:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{cons} + \mathcal{L}_c + \mathcal{L}_{MA} + \mathcal{L}_{MS}. \tag{7}$$

For downstream segmentation tasks, h_{ce} and d_e are transferred for initialization and the intensity head h_σ is replaced with a segmentation head h_ω.

3 Experiments and Results

3.1 Datasets and Implementation

For self-supervised pre-training, we curate a large-scale, heterogeneous dataset comprising 4,404 multi-sequence MRI samples (totaling 14,815 volumes). Specifically, we incorporate glioma-related cohorts (BraTS2018 [3], BraTS2020 [14], BraTS2021 [2], UPENN-GBM [4], and UCSF-PDGM [5]) covering T1w, T2w, T1ce, and FLAIR sequences, together with healthy brain datasets (HCP [18] and IXI) providing T1w, T2w, and PD contrasts. All volumes are rigidly registered to the MNI152 template via ANTs [1]. To demonstrate the generalizability of TeInL, we evaluate the framework across three scenarios: (1) multi-granularity whole-brain parcellation on Mindboggle-101 [9], CANDI [8], and JHU [19] (spanning 30 to 289 structures); (2) stroke lesion segmentation on ATLAS2 [12]; and (3) brain tumor sub-region segmentation on BraTS2021 [2] and BraTS2023-MEN [10] using only T1ce. Each dataset is partitioned into training and testing sets with a 60%/40% ratio. The intensity values of all images are scaled to the range $[0, 1]$ based on the 1st and 99th percentiles, followed by zero-mean and unit-variance normalization. Foreground image patches of $96 \times 96 \times 96$ are randomly cropped and augmented with rotation, gamma correction, and scaling. The network architecture is implemented using 3D CNN for both the encoder and the decoder. All experiments are conducted with NVIDIA RTX A6000 GPUs. The temperature

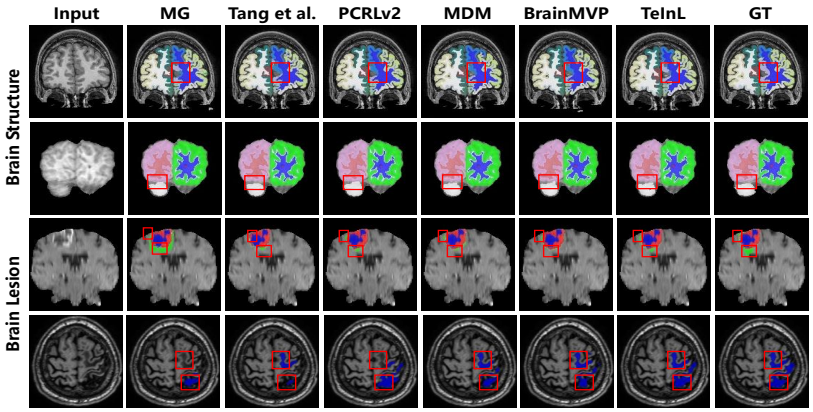

Fig. 2. Qualitative results from TeInL and comparative methods on representative images. Regions with visual improvements are highlighted with bounding boxes.

coefficient τ is set as 0.05. The AdamW optimizer with a cosine learning rate scheduler is employed to optimize the pre-training and fine-tuning objectives. For pre-training, we train the model for 100 epochs with 10 warm-up epochs. For fine-tuning, we train the model for 25,000 iterations with 400 warm-up iterations for all downstream tasks. The batch size is set to 48 and 2, and the initial learning rate is 1×10^{-4} and 4×10^{-4} for pre-training and fine-tuning.

3.2 Evaluation Results

Comparisons with SOTA. For a fair comparison, all evaluated methods are pre-trained from scratch with the same network architecture and dataset. We quantify the segmentation performance using the Dice similarity coefficient (DSC) in all experiments. We denote model training from scratch as 'scratch' in subsequent sections. As summarized in Table 1, our TeInL obtains the highest average DSC of 83.26% on brain structure segmentation. Across the three brain structure datasets, TeInL demonstrates significant superiority over SOTA methods on two datasets, except that MDM achieves a marginal advantage of 0.01% over TeInL on the JHU dataset without statistical significance ($p > 0.05$). For segmentation of brain tumor and stroke lesion, TeInL also reaches the best performance with the mean DSC of 79.49% and 57.25% respectively, achieving significant improvements of 0.89% and 1.34% over the previous best methods. Collectively, TeInL obtains the best average DSC of 77.67% across all six datasets, demonstrating its robustness. Qualitative comparisons in Fig. 2 show that our method exhibits superior discrimination capability in segmenting boundaries of the cerebellum and the occipital lobe over comparative approaches while achieving notably higher accuracy in lesion segmentation.

Ablation Studies. We evaluate the contribution of each component in TeInL. As shown in Table 2, the proposed method significantly outperforms

Table 2. Quantitative DSC ($mean_{std}$) from ablation analysis for the loss terms. Datasets A, B, C, D, E, and F are respective Mindboggle-101, CADNI, JHU, BraTS2021, BraTS2021-MEN, and ATLAS2, with the number of segmentation regions in "()". "*" indicates $p < 0.05$ in a Wilcoxon matched-pairs signed rank test comparing the DSC of the pertinent component before and after application."†" indicates the use of text descriptions restricted to healthy anatomical structures and $\mathcal{L}_{text} = \mathcal{L}_c + \mathcal{L}_{MA} + \mathcal{L}_{MS}$.

$\mathcal{L}_{rec}$	$\mathcal{L}_{cons}$	$\mathcal{L}^{\dagger}_{text}$	$\mathcal{L}_{text}$	Brain Structure				Brain Tumor			Stroke Lesion	Average
				A (101)	B (30)	C (289)	Average	D (3)	E (3)	Average	F (1)	
Scratch				$76.71_{2.95}$	$86.86_{5.28}$	$80.24_{7.86}$	81.27	$75.98_{21.11}$	$70.40_{25.78}$	73.19	$52.13_{29.79}$	73.72
✓				$80.15^{*}_{2.84}$	$86.96^{*}_{5.78}$	$80.92^{*}_{6.40}$	82.67^{*}	$80.81^{*}_{20.13}$	$72.09_{26.37}$	76.45^{*}	$52.26_{30.73}$	75.53^{*}
✓	✓			$80.62_{1.67}$	$\underline{87.23}^{*}_{4.48}$	$\underline{81.09}_{6.13}$	$\underline{82.98}$	$\underline{81.70}^{*}_{19.22}$	$\underline{73.24}^{*}_{24.30}$	$\underline{77.47}^{*}$	$\underline{55.05}^{*}_{28.83}$	$\underline{76.48}^{*}$
✓	✓	✓		$\underline{80.63}_{2.32}$	$87.11_{5.99}$	$80.57_{6.70}$	82.77	$80.61_{19.56}$	$71.83_{23.98}$	76.22	$53.16_{29.58}$	75.65
✓	✓		✓	$\mathbf{81.30}^{*}_{1.61}$	$\mathbf{87.34}^{*}_{4.79}$	$\mathbf{81.16}_{5.87}$	$\mathbf{83.26}^{*}$	$\mathbf{81.74}^{*}_{20.78}$	$\mathbf{77.25}^{*}_{23.56}$	$\mathbf{79.49}^{*}$	$\mathbf{57.25}^{*}_{28.58}$	$\mathbf{77.67}^{*}$

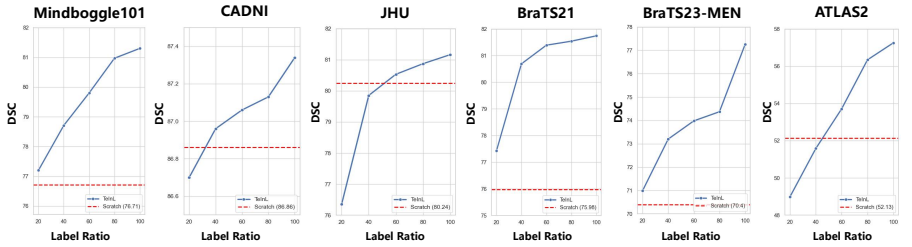

Fig. 3. Segmentation results of MCLIM with different annotation ratios. The red dotted line represents the performance of scratch with 100% training data.

scratch, improving the average DSC by 3.95% (the 1^{st} line *vs.* the *last* line). Visual pre-text tasks ($\mathcal{L}_{rec}$ and $\mathcal{L}_{cons}$) establish a strong baseline of 76.48%. A critical observation is that incorporating text descriptions restricted to healthy anatomy ($\mathcal{L}^{\dagger}_{text}$) leads to a performance drop compared to the vision-only baseline (76.48% → 75.65%, the 3^{rd} line *vs.* the 4^{th} line), particularly in brain tumor segmentation where the DSC decreases from 77.47% to 76.22%. This suggests that inaccurate semantic priors (i.e., describing lesions as healthy tissue) conflict with visual features. In contrast, integrating our pathology-aware textual descriptions ($\mathcal{L}_{text}$) achieves the best performance in all tasks, boosting the average DSC to 77.67%. On the other hand, as shown in Fig. 3, TeInL demonstrates superior data efficiency, reducing required annotation efforts by at least 40% compared to scratch.

4 Conclusion

In this work, we present TeInL, a SSL framework for multi-sequence MRI. By explicitly disentangling anatomical invariants from sequence-specific styles and injecting multi-source priors via image-text alignment, TeInL effectively learns generalized representations from multi-sequence MRI data. A key insight from

our study is the critical role of pathological awareness in textual guidance to learn robust representations. Extensive validation across six benchmarks confirms that TeInL not only establishes new SOTA performance for brain segmentation but also reduces annotation costs by over 40%. Future work will explore extending this paradigm to other medical imaging settings and more downstream tasks.

Acknowledgment. This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (T2422012); the Guangdong Basic and Applied Basic Research (2024B1515020088); the High Level of Special Funds (G030230001, G03034K003); the Guangdong Key Research and Development Program (2025B1111080001); the SUSTech Fang Keng Faculty Award.

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

References

1. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ants). Insight j **2**(365), 1–35 (2009)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
4. Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J., Flores Santamaria, N., Fathi Kazerooni, A., Pati, S., et al.: Multi-parametric magnetic resonance imaging (mpmri) scans for de novo glioblastoma (gbm) patients from the university of pennsylvania health system (upenn-gbm). The Cancer Imaging Archive **10** (2021)
5. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P.: The university of california san francisco preoperative diffuse glioma mri dataset. Radiology: Artificial Intelligence **4**(6), e220058 (2022)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
7. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al.: An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage **31**(3), 968–980 (2006)
8. Kennedy, D.N., Haselgrove, C., Hodge, S.M., Rane, P.S., Makris, N., Frazier, J.A.: Candishare: a resource for pediatric neuroimaging data. Neuroinformatics **10**, 319–322 (2012)

9. Klein, A., Hirsch, J.: Mindboggle: a scatterbrained approach to automate brain labeling. NeuroImage **24**(2), 261–280 (2005)
10. LaBella, D., Adewole, M., Alonso-Basanta, M., Altes, T., Anwar, S.M., Baid, U., Bergquist, T., Bhalerao, R., Chen, S., Chung, V., et al.: The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. arXiv preprint arXiv:2305.07642 (2023)
11. Liang, J., Lyu, J., Yuan, Y., Tang, X.: Masked contrastive language-image modeling for brain segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 310–319. Springer (2025)
12. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
13. Lyu, J., Bartlett, P.F., Nasrallah, F.A., Tang, X.: Masked deformation modeling for volumetric brain mri self-supervised pre-training. IEEE Transactions on Medical Imaging (2024)
14. Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Murugesan, G.K., et al.: Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. The journal of machine learning for biomedical imaging **2022**, https–www (2022)
15. Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5164–5174 (2025)
16. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
17. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
18. Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., et al.: The human connectome project: a data acquisition perspective. Neuroimage **62**(4), 2222–2231 (2012)
19. Wu, D., Ma, T., Ceritoglu, C., Li, Y., Chotiyanonta, J., Hou, Z., Hsu, J., Xu, X., Brown, T., Miller, M.I., et al.: Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on t1-weighted mri. Neuroimage **125**, 120–130 (2016)
20. Xu, P., Lyu, J., Lin, L., Cheng, P., Tang, X.: Lf-synthseg: Label-free brain tissue-assisted tumor synthesis and segmentation. IEEE Journal of Biomedical and Health Informatics (2024)
21. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
22. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: Pcrlv2: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. arXiv preprint arXiv:2301.00772 (2023)
23. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)