ReLiF-3D: Prior-Guided Semi-Supervised 3D MRI Segmentation via Robust Bias-Consistent Paired Views

Kunal Jangid, Tanmay Basu, and Vinod Kurmi

Department of Data Science and Engineering,
Indian Institute of Science Education & Research (IISER), Bhopal, India
{kunal24, tanmay, vinodkk}@iiserb.ac.in

Abstract. Foundation models provide strong anatomical priors that can improve semi-supervised 3D MRI segmentation under limited annotation. However, many approaches still rely on unstable teacher-student updates and heuristic pseudo-labeling, which can accumulate errors when learning from unlabeled volumes. In MRI, scanner-dependent intensity inhomogeneity and distribution shifts further degrade robustness, especially in extreme low-label regimes with heterogeneous lesions. We propose ReLiF-3D, a foundation-guided semi-supervised framework that trains a lightweight 3D U-Net under supervision from a frozen volumetric foundation model (SAM-Med3D) prior, avoiding EMA teacher drift while retaining a stable spatial guide. Our main contribution is the Smooth Orthogonal Bias Field (SOBF), a paired-view generator that simulates realistic MRI bias through orthogonal multiplicative and additive smooth fields with a progressive strength schedule, producing anatomically consistent yet scanner-shifted views for reliable consistency learning. To stabilize unlabeled optimization, ReLiF-3D applies confidence-gated voxel co-regularization to enforce consistency only on reliable agreements and lesion-aware representation alignment guided by the SAM-Med3D prior. Experiments on public BraTS and Left Atrium datasets with 1-5 labeled volumes demonstrate significant improvements in overlap and boundary metrics over SAM-assisted SSL baselines and state-of-the-art methods. Project page is available at https://visdomlab.github.io/ReLiF_3D

Keywords: Semi-supervised learning · 3D MRI segmentation · Foundation models · Bias field Augmentation · Robustness.

1 Introduction

Precise 3D medical image segmentation (MIS) is essential for quantitative diagnosis and treatment planning, yet dense voxel-level annotation remains costly, time-consuming, and constrained by privacy and clinical workflows [8, 13]. Semi-supervised learning (SSL) mitigates annotation burden by leveraging unlabeled scans via consistency regularization and pseudo-labeling [14, 7, 11]. However, robust SSL in clinical Magnetic Resonance Imaging (MRI) remains challenging.

Noisy pseudo-labels and drifting teacher targets can amplify errors on unlabeled data, while acquisition-dependent intensity variability induces distribution shifts that degrade generalization [24]. Recent SAM-guided SSL methods (e.g., SemiSAM [27], SemiSAM+[26]) integrate the foundation model as a pseudo-label guidance within teacher-student pipelines rather than as a fixed supervisory priors [30]. However, when pseudo-labels are unreliable (e.g., under domain shift), exponential moving average (EMA) can suffer from confirmation bias and target drift, reinforcing early mistakes on unlabeled data [21, 6]. In MRI, scanner-dependent intensity inhomogeneity further amplifies this issue.

A key source of MRI variability is scanner and protocol-dependent intensity inhomogeneity (bias fields). Bias fields arise from coil sensitivity variations, which cause non-uniform illumination or shading artifacts [10], and magnetic field inhomogeneity introduces low-frequency intensity drifts and signal shifts [17, 16]. These artifacts distort tissue contrast and alter lesion appearance, thereby complicating segmentation generalization across scanners and protocols. Under extreme label scarcity, SSL can overfit to such intensity cues rather than anatomical structure, resulting in (i) limited robustness to bias-induced perturbations [17, 10, 16], (ii) over-confident unsupervised updates that propagate incorrect pseudo-labels [24, 14], and (iii) weak lesion representations for small or low-contrast targets [23]. Existing correction methods (e.g., N3 [12], N4 bias correction [15], [4]) remain limited in cross-site adaptability, motivating the need for learning frameworks inherently robust to scanner-induced intensity variations.

We propose **R**obust **L**esion-aware and **F**oundation-guided semi-supervised **3D** segmentation (**ReLiF-3D**) for robust 3D MRI segmentation under extreme label scarcity. ReLiF-3D couples a frozen SAM-Med3D [18] foundation model with training a lightweight 3D U-Net [5] as the task-adaptive segmenter. Keeping SAM-Med3D frozen prevents EMA drift and provides a stable spatial and lesion-aware supervisory prior for both labeled and unlabeled data. To explicitly address MRI-specific acquisition variability, we propose **Smooth Orthogonal Bias Field** (SOBF), a paired-view generator that simulates scanner-induced intensity inhomogeneity while preserving anatomical structure. SOBF synthesizes a smooth multiplicative gain and additive offset, and enforces *orthogonality* between them to yield complementary, non-redundant perturbations with a progressive strength schedule for stable training. Unlike conventional augmentation-driven SSL, SOBF is grounded in MRI bias-field physics, providing task-relevant perturbations rather than generic intensity jitter [7, 11]. To further prevent error propagation on unlabeled data, ReLiF-3D incorporates lightweight reliability constraints. **Confidence-Gated Co-Regularization** (CGCR) enforces voxel-wise consistency only where clean and bias-augmented predictions agree with high confidence, reducing the influence of uncertain regions. **Lesion-Aware Representation Consistency** (LARC) aligns foreground embeddings across paired views using lesion priors derived from the frozen foundation model, focusing representation learning on lesion-relevant regions. A KL-divergence further aligns semantic consistency between the 3D U-Net predictions and foundation outputs across different views. Our contributions are summarized as:

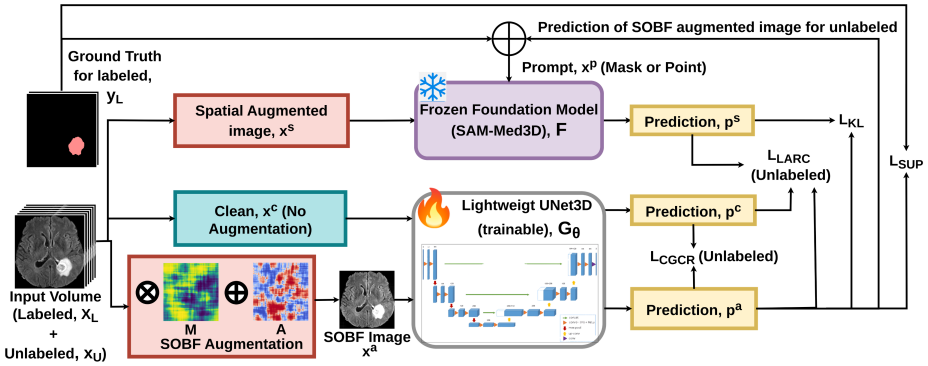


Fig. 1. Overview of **ReLiF-3D**. Frozen SAM-Med3D guides a 3D U-Net with stable priors; SOBF creates bias-shifted paired views (Sec. 2.1); while CGCR (Sec. 2.2) and LARC (Sec. 2.3) enforce reliable consistency for robust learning with limited labels.

1. We present **ReLiF-3D**, a foundation-guided SSL framework leveraging a *frozen* SAM-Med3D prior to avoid EMA drift under extreme label scarcity.
2. We introduce **SOBF**, an MRI-specific paired-view generator based on *orthogonal* multiplicative-additive smooth bias fields, improving robustness to scanner-dependent intensity inhomogeneity.
3. We integrate reliability constraints and foundation-aligned supervision to stabilize unlabeled optimization and demonstrate consistent gains on public MRI benchmarks under 1-5 labeled volumes.

2 Methodology

Let the training dataset be $\mathcal{D}$ with labeled set $\mathcal{D}_L = \{(x_l^i, y_l^i)\}_{i=1}^S$ and unlabeled set $\mathcal{D}_U = \{x_u^j\}_{j=1}^T$, where x_l^i and x_u^j are the input images with S labeled (l) and T unlabeled (u) volumes, respectively, y_l^i denotes the ground truth of labeled set, and $S \ll T$ indicates the extremely low-label regime. Our goal is to learn a compact segmenter (3D U-Net [5]) using both $\mathcal{D}_L$ and $\mathcal{D}_U$, guided by a frozen volumetric foundation model F (SAM-Med3D [18]), where θ denotes learnable parameters of 3D U-Net (G_θ). The frozen foundation model provides stable anatomical priors to guide optimization of the task-specific segmenter, without EMA teacher updates. Concretely, F remains fixed throughout training, while only θ is updated. Fig. 1 illustrates the pipeline. For each input volume x (labeled or unlabeled), we form three views as (i) a spatially augmented view x^s processed by F to produce p^s, (ii) a clean view x^c processed by G_θ to produce p^c, and (iii) a bias-augmented view x^a produced by SOBF (Sec. 2.1) and processed by G_θ to produce p^a. The foundation model receives a prompt input $x^p = \text{concat}(y_l^i, p^a)$, where we concatenate ground truth (y_l^i) for labeled volumes and prediction (p^a) from frozen F for unlabeled volumes. During inference, only G_θ is used.

2.1 Smooth Orthogonal Bias Field (SOBF) Paired-View Generator

MRI scans often suffer from scanner-dependent bias fields and intensity inhomogeneities that distort tissue contrast and reduce model generalization [17, 10, 16]. **Smooth Orthogonal Bias Field (SOBF)** augmentation, a data-driven and physics-consistent *paired-view generator* for simulating scanner bias, with the goal of adapting the network to real scanner variability while retaining anatomical semantics. SOBF constructs two complementary smooth fields, a multiplicative gain field M (low-frequency scaling) and an additive offset field A (smooth intensity shift). To avoid redundant perturbations, we explicitly enforce *orthogonality* between the two fields, yielding diverse but anatomically consistent intensity variations. We first sample random noise fields $n_m, n_a \sim \mathcal{N}(0,1)$ and smooth them using efficient 3D box filtering as $\hat{m} = Norm(\mathcal{G}_{k_m}(n_m)), \quad \hat{a} = Norm(\mathcal{G}_{k_a}(n_a))$, where $\mathcal{G}_{k_m}$ and $\mathcal{G}_{k_a}$ are 3D box filters with kernel sizes k_m and k_a, and $Norm(\cdot)$ applies zero-mean and unit-variance normalization over spatial dimensions. Typically, $k_m > k_a$ so that M varies more smoothly than A, consistent with MRI bias characteristics [17]. To ensure independent perturbation components, we project $\hat{a}$ onto the subspace orthogonal to $\hat{m}$ as

$$a_\perp = \hat{a} - \frac{\langle \hat{a}, \hat{m} \rangle}{\|\hat{m}\|^2 + \varepsilon} \hat{m} \tag{1}$$

so that $(a_\perp, \hat{m}) \approx 0$ per sample. We then bound both fields using *tanh* to ensure smooth and realistic intensity perturbations as $M = 1 + \alpha_m tanh(\hat{m}), \quad A = \alpha_a tanh(a_\perp)$, where α_m and α_a control the perturbation strength, gradually increased during training following a curriculum schedule to ensure stability, and generate the SOBF view as

$$x^a = x \cdot M + A \tag{2}$$

Finally, x^a is clipped to the 1st-99th percentile of the original intensity range to remove outliers while preserving contrast. By training on paired clean/bias views (x^c, x^a), the model is encouraged to rely on structural and lesion cues rather than scanner-specific intensity patterns.

2.2 Confidence-Gated Co-Regularization (CGCR)

Consistency regularization over all voxels can propagate errors from uncertain predictions, especially in low-contrast or unlabeled regions. We apply **CGCR**, which enforces consistency only on voxels where the clean and SOBF views agree *and* are confident. Let $p^c, p^a \in [0,1]^{C \times D \times H \times W}$ be the softmax outputs of G_θ for x^c and x^a. For every voxel i, $v_c = \arg\max_j p^c_{i,j}$, $v_a = \arg\max_j p^a_{i,j}$, $s_c = \max_j p^c_{i,j}$, and $s_a = \max_j p^a_{i,j}$

A voxel is considered reliable if both views predict the same class and both confidences exceed a threshold β_t:

$$G_i = \mathbf{1}[v_c = v_a] \cdot \mathbf{1}[s_c \geq \beta_t] \cdot \mathbf{1}[s_a \geq \beta_t] \tag{3}$$

We linearly increase β_t during training so that early learning uses broader supervision while later stages focus on high-confidence regions. The CGCR loss is a masked MSE over reliable voxels:

$$\mathcal{L}_{\mathrm{CGCR}} = \frac{\sum_i G_i \, \|p_i^c - p_i^a\|_2^2}{\sum_i G_i + \varepsilon} \tag{4}$$

2.3 Lesion-Aware Representation Consistency (LARC)

Lesions are often small and sparse; background-dominated objectives can suppress lesion features. LARC aligns lesion-specific representations between clean and SOBF views using the frozen foundation prior to focus on foreground regions. Let $f_\theta(\cdot)$ denote the feature map from the encoder G_θ, and g_ϕ be a lightweight projection head. For an unlabeled batch, we extract f^c and $f^a \in \mathbb{R}^{B\times C\times D\times H\times W}$. Using SAM-Med3D output $p^s \in [0,1]^{B\times 1\times D\times H\times W}$, we extract a foreground confidence mask as $E = (p^s)^\gamma$ for confident lesion areas. We obtain lesion embeddings via masked global average pooling defined in Eq. 5. Then processed through shared projection head as $z^c = g_\phi(e^c)$ and $z^a = g_\phi(e^a)$ followed by ℓ_2-normalized. We align positive pairs with a stop-gradient ($sg(.)$) on the clean branch as defined in Eq. 6. We adopt a positive-only alignment strategy without explicit negative sampling. In medical segmentation, background voxels dominate spatially and are highly heterogeneous; treating them as negatives can introduce noisy gradients and suppress small lesion representations. Focusing alignment on lesion-relevant foreground embeddings improves representation stability.

$$e^c = \frac{\sum(E \cdot f^c)}{\sum E + \varepsilon}, \quad e^a = \frac{\sum(E \cdot f^a)}{\sum E + \varepsilon} \tag{5}$$

$$\mathcal{L}_{\mathrm{LARC}} = \frac{1}{B}\sum_{b=1}^{B}\Big(1 - \cos\big(z_b^a, \mathrm{sg}(z_b^c)\big)\Big). \tag{6}$$

Progressive Scheduling. Supervised Loss. With very few labeled volumes, early optimization can be unstable. We therefore interpolate between Dice and cross-entropy (CE) supervision as training progresses as $\lambda_{\mathrm{pixel}} = C_t/C_{\mathrm{total}}$, $\lambda_{\mathrm{overlap}} = 1 - \lambda_{\mathrm{pixel}}$. So, $\mathcal{L}_{\mathrm{sup}} = \lambda_{\mathrm{overlap}}\mathcal{L}_{\mathrm{Dice}} + \lambda_{\mathrm{pixel}}\mathcal{L}_{\mathrm{CE}}$, this schedule stabilizes training by emphasizing region overlap early and voxel-wise refinement later.

Bias-Strength Scheduling. We progressively increase SOBF strength over training using $s(t) = \min(1.0, C_t/C_{\mathrm{total}})$, $\alpha_m = \alpha_{\mathrm{m,max}}\big(0.5 + 0.5\,s(t)\big)$, and $\alpha_a = \alpha_{\mathrm{a,max}}\big(0.5 + 0.5\,s(t)\big)$, increasing perturbations from 0.5 to 1.0 [19]. Mild bias early stabilizes learning of anatomical structure, while stronger bias later improves robustness to scanner variability.

Foundation Alignment. We align G_θ with a frozen F prior using a Kullback-Leibler (KL) divergence. Let $\mathbf{R}_i^s = \mathrm{softmax}(\mathbf{p}^s/\tau)$, $\mathbf{Q}_i^a = \mathrm{softmax}(\mathbf{p}^a/\tau)$, then $\mathcal{L}_{\mathrm{KL}} = \sum_i R_i^s \log\left(R_i^s/Q_i^a\right)$, where τ is a temperature parameter and $R^s, Q^a \in [0,1]^{C\times D\times H\times W}$ are per-voxel class probabilities. The total loss is $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{sup}} +$

$\mathcal{L}_{\text{CGCR}} + \omega_{\text{larc}}\mathcal{L}_{\text{LARC}} + \omega_{\text{kl}}\mathcal{L}_{\text{KL}}$. We use a ramp-up coefficient ω_{kl} [9], and a warm-up schedule for representation alignment: $\omega_{\text{larc}} = \lambda_{\text{larc}} \cdot \min(1, C_t/T_{\text{warmup}})$, which delays LARC to avoid early over-regularization.

3 Experiments and Results

Datasets. BraTS 2019 [2] contains multi-modal brain MRI (T1, T1ce, T2, FLAIR) with tumor annotations. Following common practice, we use the FLAIR modality and segment *whole tumor*. The dataset is split into 250/25/60 volumes for training/validation/testing. **Left Atrium (LA)** [20] provides 100 3D gadolinium-enhanced MR images (GE-MRI) with left atrium masks, and uses the standard split of 75/5/20 volumes for training/validation/testing. **Low-label protocol.** To simulate extreme annotation scarcity, we train with $S \in \{1, 2, 3, 5\}$ labeled volumes, and treat all remaining training volumes as unlabeled.

Implementation and Evaluation Metrics. All models are implemented in PyTorch and follow the SemiSAM+ [26] training protocol for consistent baselines. We optimize using SGD with initial learning rate 0.01, momentum 0.9, weight decay 10^{-4}, and follows a polynomial decay schedule as $\text{lr} = \text{lr}_0\,(1 - C_t/C_{\text{total}})^{0.9}$, with $C_{\text{total}} = 30{,}000$ iterations. We use a batch size of 2 and train on cropped volumes of size $128 \times 128 \times 128$. Standard spatial augmentations (random cropping, flipping, and rotation) are applied to reduce overfitting. At test time, we generate full-volume predictions using a sliding-window inference strategy. The values of parameters k_m=41, k_a=21; $\alpha_{\text{m,max}} = 0.25$, $\alpha_{\text{a,max}} = 0.12$; τ=2; $\gamma = 1.5$; $\lambda_{larc} = 0.08$; and T_{warmup}=2000 are chosen empirically. The Dice score (DSC), Intersection over Union (IoU), 95 Hausdorff Distance (95HD), and Average Surface Distance (ASD) are adopted as our evaluation metrics.

Comparison Study. Table 1 summarizes results on BraTS and LA under extreme low-label settings ($S \in \{1, 2, 3, 5\}$). We compare against MT [14], UA-MT [24], DAN [28], DTC [9], DyCON [1], and SKCDF [29], all using the same SAM-Med3D prior for fair comparison. ReLiF-3D achieves the best overlap and boundary accuracy across low-label regimes. In the most challenging **1-label** setting, it reaches **76.41% Dice** on BraTS and **76.84% Dice** on LA. Lower 95HD/ASD further indicates improved surface fidelity, which is critical for clinically actionable contours, supporting less manual correction and more reliable geometry for treatment planning and quantitative follow-up. These improvements address the robustness to scanner-induced intensity shifts, stabilization of unlabeled optimization, and strengthened lesion-focused representations.

Comparison with SOTA. Table 2 compares ReLiF-3D with SAM-based adaptation and SAM-assisted SSL on LA. ReLiF-3D consistently outperforms these baselines in the low-label regime, achieving **76.84% Dice** with **1** labeled case (vs H-SAM 67.73% and SemiSAM+ 48.91%) and achieves consistent improvement for $S \in \{2, 3, 4, 5\}$. This demonstrates that effective use of a frozen foundation prior with 3D U-Net to mitigate teacher drift under extreme label scarcity.

Robustness and qualitative analysis. SOBF generates anatomy-preserving, scanner-like low-frequency intensity drift and focuses on anatomy regions (Fig-

Table 1. Comparison on BraTS and LA under extremely limited labels (mean ± SD). † indicates ReLiF-3D significantly outperforms SemiSAM+ (Wilcoxon, $p < 0.05$).

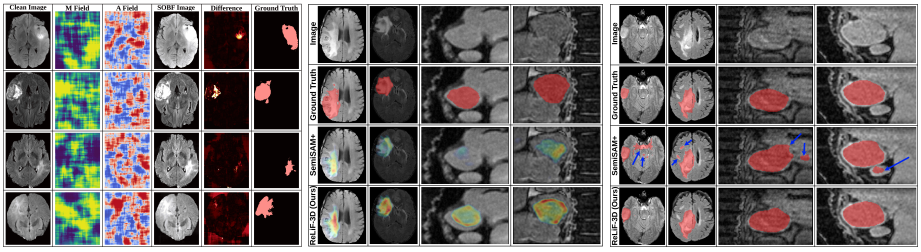
No. of Labeled Samples	Method	Brain Tumor (BraTS)				Left Atrium (LA)			
		Dice (↑)	Jaccard (↑)	95HD (↓)	ASD (↓)	Dice (↑)	Jaccard (↑)	95HD (↓)	ASD (↓)
1 labeled	FS Baseline	42.82 ± 18.13	29.01 ± 15.43	63.55 ± 17.71	30.43 ± 11.88	37.97 ± 21.80	25.83 ± 17.98	39.50 ± 13.57	16.46 ± 6.72
	SSL MT / SAM-Med3D	68.53 ± 22.22	56.22 ± 24.39	38.81 ± 29.05	16.66 ± 15.93	48.91 ± 22.79	35.46 ± 20.77	38.49 ± 15.40	14.61 ± 7.12
	SSL UA-MT / SAM-Med3D	71.75 ± 17.78	58.71 ± 20.12	34.24 ± 27.51	12.91 ± 14.17	50.10 ± 22.83	36.56 ± 20.87	37.02 ± 14.27	14.27 ± 7.16
	SSL DAN / SAM-Med3D	70.32 ± 20.89	57.93 ± 23.19	34.18 ± 26.93	13.78 ± 14.57	62.26 ± 14.34	46.82 ± 15.67	31.88 ± 10.48	11.85 ± 5.74
	SSL DTC / SAM-Med3D	67.29 ± 23.84	55.19 ± 25.19	30.82 ± 25.37	11.90 ± 13.67	53.52 ± 23.05	39.84 ± 21.25	35.19 ± 17.46	13.25 ± 8.61
	SSL DyCON / SAM-Med3D	68.46 ± 17.13	54.52 ± 19.19	26.38 ± 26.40	7.84 ± 12.51	45.46 ± 17.48	31.07 ± 14.71	36.13 ± 6.87	14.47 ± 5.25
	SSL SKCDF / SAM-Med3D	70.16±20.58	57.46±21.85	31.98±27.67	10.73±12.38	50.48 ± 16.85	35.67 ± 17.00	43.17 ± 6.87	17.07 ± 4.21
	ReLiF-3D (Ours)	**76.41 ± 13.92**†	**63.72 ± 16.88**†	**15.87 ± 19.72**†	**4.83 ± 10.29**†	**76.84 ± 9.17**†	**63.28 ± 11.88**†	**18.98 ± 10.54**†	**5.69 ± 3.63**†
2 labeled	FS Baseline	53.70 ± 18.77	38.98 ± 17.89	77.37 ± 25.85	33.57 ± 15.33	59.99 ± 17.43	45.02 ± 17.56	38.44 ± 16.41	14.11 ± 7.33
	SSL MT / SAM-Med3D	72.59 ± 18.57	60.04 ± 21.22	37.57 ± 28.71	14.32 ± 15.04	73.09 ± 12.27	59.01 ± 14.64	34.05 ± 16.99	10.83 ± 6.23
	SSL UA-MT / SAM-Med3D	72.26 ± 19.80	60.04 ± 22.64	33.14 ± 30.04	13.43 ± 15.46	74.06 ± 12.33	60.26 ± 14.91	30.75 ± 16.75	9.81 ± 6.27
	SSL DAN / SAM-Med3D	72.54 ± 19.78	60.34 ± 22.39	22.47 ± 26.26	8.48 ± 13.84	72.28 ± 14.53	58.46 ± 16.61	35.29 ± 17.66	11.42 ± 6.92
	SSL DTC / SAM-Med3D	70.96 ± 20.49	58.48 ± 22.27	33.26 ± 30.08	13.93 ± 15.53	73.76 ± 14.38	60.34 ± 16.91	29.56 ± 17.40	9.72 ± 7.06
	SSL DyCON / SAM-Med3D	70.61 ± 16.66	57.00 ± 19.05	24.40 ± 25.81	5.86 ± 11.10	52.18 ± 16.67	36.98 ± 14.98	33.25 ± 8.63	11.44 ± 4.57
	SSL SKCDF / SAM-Med3D	76.65 ± 21.44	66.15 ± 23.35	17.02 ± 21.76	5.00 ± 10.40	71.48 ± 15.15	57.54 ± 16.39	19.89 ± 8.20	6.55 ± 3.38
	ReLiF-3D (Ours)	**78.26 ± 13.58**†	**66.18 ± 17.12**†	**14.41 ± 19.21**†	**4.16 ± 9.12**†	**79.14 ± 10.61**†	**66.63 ± 13.29**†	**19.83 ± 11.04**†	**5.70 ± 3.91**†
3 labeled	FS Baseline	56.06 ± 24.68	42.81 ± 22.83	47.67 ± 20.56	17.52 ± 12.94	62.33 ± 16.10	47.19 ± 16.45	36.59 ± 16.52	12.16 ± 6.39
	SSL MT / SAM-Med3D	73.92 ± 21.26	62.51 ± 23.28	19.77 ± 21.89	6.13 ± 9.76	75.66 ± 15.38	62.36 ± 17.93	22.25 ± 17.51	6.59 ± 6.72
	SSL UA-MT / SAM-Med3D	73.86 ± 21.17	62.38 ± 23.07	21.46 ± 24.61	7.18 ± 12.15	75.97 ± 13.69	63.04 ± 16.29	24.78 ± 16.74	7.82 ± 6.12
	SSL DAN / SAM-Med3D	74.52 ± 19.35	62.68 ± 21.57	21.22 ± 23.96	6.56 ± 12.69	73.67 ± 16.46	60.72 ± 18.66	21.83 ± 15.65	7.72 ± 6.43
	SSL DTC / SAM-Med3D	75.30 ± 19.24	63.76 ± 22.08	19.83 ± 23.01	5.27 ± 7.39	73.16 ± 17.36	60.38 ± 19.83	21.22 ± 15.64	6.76 ± 6.57
	SSL DyCON / SAM-Med3D	68.25 ± 19.51	54.93 ± 21.35	27.80 ± 28.09	9.04 ± 15.05	45.29 ± 15.10	30.54 ± 13.02	36.29 ± 6.16	14.47 ± 4.00
	SSL SKCDF / SAM-Med3D	73.50 ± 25.96	63.45 ± 26.43	16.13 ± 17.91	3.22 ± 4.76	74.33 ± 13.13	60.67 ± 14.63	27.66 ± 10.96	8.89 ± 4.02
	ReLiF-3D (Ours)	**76.33 ± 18.28**	**64.78 ± 20.86**	**16.41 ± 19.37**	**3.75 ± 5.51**†	**79.50 ± 11.71**†	**67.36 ± 14.42**†	**17.64 ± 12.95**†	**5.24 ± 4.62**†
5 labeled	FS Baseline	62.35 ± 26.60	50.16 ± 25.54	36.68 ± 31.61	11.83 ± 17.33	63.39 ± 18.47	48.91 ± 18.64	36.10 ± 17.95	11.91 ± 7.01
	SSL MT / SAM-Med3D	76.00 ± 15.86	63.72 ± 19.18	25.82 ± 26.15	8.89 ± 12.64	77.02 ± 13.18	64.33 ± 15.94	23.30 ± 16.05	7.48 ± 6.00
	SSL UA-MT / SAM-Med3D	74.05 ± 21.11	62.61 ± 23.02	22.61 ± 22.30	6.55 ± 8.36	74.92 ± 13.63	61.65 ± 16.19	24.06 ± 17.26	7.82 ± 6.35
	SSL DAN / SAM-Med3D	74.97 ± 17.30	62.56 ± 20.09	27.88 ± 24.86	10.24 ± 13.67	75.25 ± 13.07	61.93 ± 15.49	24.24 ± 15.29	7.55 ± 5.32
	SSL DTC / SAM-Med3D	76.53 ± 17.11	64.82 ± 20.61	20.44 ± 22.84	6.52 ± 10.09	79.52 ± 11.33	67.33 ± 14.17	19.03 ± 14.14	5.54 ± 4.94
	SSL DyCON / SAM-Med3D	68.36 ± 18.60	54.79 ± 20.48	23.83 ± 26.47	6.97 ± 10.90	41.34 ± 17.19	27.57 ± 14.00	35.19 ± 5.90	14.67 ± 3.83
	SSL SKCDF / SAM-Med3D	74.40 ± 25.43	64.42 ± 25.92	15.74 ± 17.28	3.01 ± 3.70	74.48 ± 17.30	61.99 ± 19.37	18.10 ± 10.49	6.01 ± 4.00
	ReLiF-3D (Ours)	**77.69 ± 17.77**	**66.46 ± 20.45**	**14.30 ± 16.98**	**3.29 ± 4.51**†	**80.22 ± 10.89**†	**68.17 ± 13.24**†	**17.52 ± 12.09**	**4.95 ± 4.34**†
All labeled	FS Upperbound	87.29 ± 9.61	78.55 ± 13.08	7.77 ± 9.78	1.73 ± 1.87	88.18 ± 5.30	79.23 ± 7.71	11.34 ± 11.85	2.90 ± 3.27

ure 2a). Under synthetic intensity perturbations on BraTS (2 labeled), ReLiF-3D degrades but less than SemiSAM+ (MT), improving Dice by +**5.21**% and reducing 95HD by **12.69** mm at strong shift (amp=0.15, noise=0.05) (Fig. 3a). Qualitatively, Grad-CAM activations are more compact and lesion-aligned for ReLiF-3D (Fig. 2b), and predicted masks show fewer false positives/negatives with sharper boundaries (Fig. 2c). This confirms that bias-consistent paired views effectively mitigate scanner-dependent intensity shifts discussed in Sec. 1.

Ablation Study. Table 2 analyzes proposed components on BraTS with 2 labeled volumes. SOBF alone is strong (**76.86** Dice / **14.63** 95HD), validating bias-consistent paired views. Adding reliability constraints provides complementary gains: CGCR stabilizes unlabeled learning by enforcing consistency only on confident agreements, while LARC strengthens lesion-focused invariance under bias-shifted views. The full model achieves the best performance (**78.26** Dice / **14.41** 95HD), indicating synergy between robust paired views and reliability-aware learning. We further ablate training choices. Removing progressive supervision interpolation in $\mathcal{L}_{\text{sup}}$ or the progressive SOBF schedule slightly degrades performance, supporting stable optimization in low-label regimes. Replacing the mask prompt with a point prompt reduces accuracy, indicating that richer spatial prompting achieves more informative priors. Removing $\mathcal{L}_{\text{KL}}$ decreases Dice, while removing **CGCR** causes the largest drop (notably 95HD), highlighting importance of reliability gating on unlabeled data. Applying **LARC** from the start (no warm-up) reduces performance, supporting delayed representation alignment when predictions are still unstable. This validates the necessity of reliability gating to prevent confirmation bias and error propagation under scarce supervision.

Table 2. SOTA comparison and ablation analysis. The left block compares ReLiF-3D with SAM-based and SSL baselines on LA using 1-3 labeled volumes. The middle block reports module ablations of SOBF, CGCR, and LARC across 2 labeled BraTS volumes. The right block reports training and prompt ablations on BraTS with 2 labeled volumes. The three blocks correspond to distinct experimental settings and should be interpreted independently.

SOTA Comparison on LA				**Module Ablation**					**Loss / Prompt Ablation**		
Method	1 Label	2 Label	3 Label	**SOBF**	**CGCR**	**LARC**	**Dice**	**95HD**	**Variant**	**Dice**	**95HD**
H-SAM [3]	67.73 ± 7.47	75.86 ± 6.02	78.36 ± 6.50	✓			76.86	14.63	No prog. in $\mathcal{L}_{\text{sup}}$	77.91	14.48
Auto-SAM-Med3D [26]	61.07 ± 12.30	68.95 ± 8.40	72.92 ± 7.84		✓		76.68	15.14	No prog. in SOBF	77.87	14.56
UGMCL [25]	44.08 ± 23.17	66.04 ± 14.64	69.05 ± 16.35			✓	74.84	19.39	Point prompt	77.95	13.92
ACMT [22]	44.17 ± 21.82	71.74 ± 12.64	73.45 ± 13.41	✓	✓		76.69	15.15	w/o $\mathcal{L}_{\text{KL}}$	76.86	14.74
MT [14]	40.64 ± 23.82	61.84 ± 15.89	64.35 ± 16.33	✓		✓	77.56	14.86	w/o $\mathcal{L}_{\text{CGCR}}$	74.20	28.94
MT / SAM-Med3D [26]	48.91 ± 22.79	73.09 ± 12.27	75.66 ± 15.38		✓	✓	76.07	18.91	$\mathcal{L}_{\text{LARC}}$ from start (no warm-up)	76.27	15.06
ReLiF-3D (Ours)	**76.84 ± 9.17**	**79.14 ± 10.61**	**79.50 ± 11.71**	✓	✓	✓	**78.26**	**14.41**	**ReLiF-3D (full)**	**78.26**	**14.41**



(a) (b) (c)

Fig. 2. Exemplar 2D results (a) SOBF generates smooth, anatomy-preserving intensity drift. (b) ReLiF-3D yields more lesion-aligned Grad-CAM than SemiSAM+. (c) improves boundary fidelity and reduces missed regions.

SOBF parameter sensitivity. Fig. 3b studies SOBF design choices on BraTS (2 labeled volumes). Using only multiplicative ($\alpha_{a,\max} = 0$) or only additive bias ($\alpha_{m,\max} = 0$) reduces performance, indicating that the combined orthogonal fields better capture MRI variability. Moderate strengths ($\alpha_{m,\max} = 0.25$, $\alpha_{a,\max} = 0.12$) achieve the best trade-off, whereas weaker perturbations reduce diversity and stronger perturbations can distort lesion appearance. Kernel sizes also matter: $(k_m, k_a) = (41, 21)$ is most stable, while smaller kernels introduce overly local artifacts and larger kernels over-smooth the perturbations.

4 Conclusion

ReLiF-3D demonstrates that foundation models can be effectively combined for MIS, enabling the use of large amounts of unlabeled data while relying on only a very limited number of labeled samples. The core component, **SOBF**, generates bias-consistent paired views via orthogonal multiplicative-additive smooth fields, improving robustness to scanner-dependent intensity inhomogeneity. Together with reliability constraints (**CGCR**) and lesion-focused representation

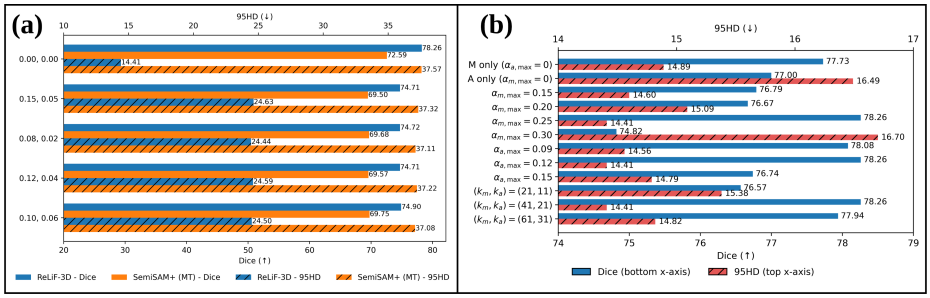


Fig. 3. (a) Robustness under intensity perturbations. (b) SOBF parameter sensitivity.

alignment (**LARC**), ReLiF-3D stabilizes learning from unlabeled data and improves boundary quality. Extensive experiments on BraTS and LA under 1-5 labeled volumes show consistent gains over SAM-assisted SSL baselines.

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

References

1. Assefa, M., Naseer, M., Ganapathi, I.I., Ali, S.S., Seghier, M.L., Werghi, N.: Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30850–30860 (2025)
2. Bakas, S.S.: Brats MICCAI Brain tumor dataset (2020). https://doi.org/10.21227/hdtd-5j88
3. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3511–3522 (2024)
4. Chuang, K.H., Wu, P.H., Li, Z., Fan, K.H., Weng, J.C.: Deep learning network for integrated coil inhomogeneity correction and brain extraction of mixed mri data. Scientific reports **12**(1), 8578 (2022)
5. Çiçek, ¨O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
6. Ding, Q., Yin, J., Zhang, D., Gao, J.: Combating confirmation bias: a unified pseudo-labeling framework for entity alignment: Q. ding et al. Data Mining and Knowledge Discovery **39**(5), 51 (2025)
7. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
8. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

9. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
10. Manson, E., Inkoom, S., Mumuni, A.: Impact of magnetic field inhomogeneity on the quality of magnetic resonance images and compensation techniques: a review. rep med imaging 15: 43–56 (2022)
11. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12674–12684 (2020)
12. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE transactions on medical imaging **17**(1), 87–97 (2002)
13. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical image analysis **63**, 101693 (2020)
14. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
15. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE transactions on medical imaging **29**(6), 1310–1320 (2010)
16. Venkatesh, V., Sharma, N., Singh, M.: Intensity inhomogeneity correction of mri images using inhomonet. Computerized Medical Imaging and Graphics **84**, 101748 (2020)
17. Vovk, U., Pernus, F., Likar, B.: A review of methods for correction of intensity inhomogeneity in mri. IEEE transactions on medical imaging **26**(3), 405–421 (2007)
18. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision. pp. 51–67. Springer (2024)
19. Wang, Y., Yue, Y., Lu, R., Liu, T., Zhong, Z., Song, S., Huang, G.: Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5852–5864 (2023)
20. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)
21. Xu, R., Yu, Y., Cui, H., Kan, X., Zhu, Y., Ho, J., Zhang, C., Yang, C.: Neighborhood-regularized self-training for learning with few labels. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 10611–10619 (2023)
22. Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., Tong, R.K.y.: Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Medical Image Analysis **88**, 102880 (2023)
23. You, C., Dai, W., Min, Y., Liu, F., Clifton, D., Zhou, S.K., Staib, L., Duncan, J.: Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. Advances in neural information processing systems **36**, 9984–10021 (2023)

24. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 605–613. Springer (2019)
25. Zhang, Y., Jiao, R., Liao, Q., Li, D., Zhang, J.: Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artificial Intelligence in Medicine **138**, 102476 (2023)
26. Zhang, Y., Lv, B., Xue, L., Zhang, W., Liu, Y., Fu, Y., Cheng, Y., Qi, Y.: Semisam+: rethinking semi-supervised medical image segmentation in the era of foundation models. Medical Image Analysis p. 103733 (2025)
27. Zhang, Y., Yang, J., Liu, Y., Cheng, Y., Qi, Y.: Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3982–3986. IEEE (2024)
28. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International conference on medical image computing and computer-assisted intervention. pp. 408–416. Springer (2017)
29. Zhang, Z., Yin, G., Zhang, B., Liu, W., Zhou, X., Wang, W.: A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25940–25949 (2025)
30. Zhao, H., Li, H., Ma, L.L., Sun, D.: Segment anything model meets semi-supervised medical image segmentation: A novel perspective. Advances in Neural Information Processing Systems **38**, 170128–170148 (2026)