FReMCI: Frequency-Reliable Memory Cross Intervention for Semi-Supervised Ultrasound Image Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Semi-supervised ultrasound image segmentation is clinically attractive because dense annotations are costly, but probe-dependent appearance, speckle noise, low contrast, and ambiguous boundaries make unlabeled supervision unreliable. Existing methods can propagate errors through pseudo labels, spatially mixed regions, and weak foreground features. We present FReMCI, a frequency-reliable memory cross intervention framework that controls cross-sample transfer at frequency, region, and feature-memory levels. Phase-Preserved Low-Frequency Amplitude Sharing (PLAS) extends Fourier amplitude interpolation to paired labeled–unlabeled images while retaining each image's phase and high-frequency amplitude. Reliability-Guided Semantic Region Intervention (ReSRI) combines confidence, entropy, and class margin from an EMA teacher with teacher–student agreement to construct reliability-weighted pseudo targets and foreground-centered bidirectional region exchanges with seam-suppressed supervision. Memory Cross Injection (MCI) stores labeled and reliable unlabeled foreground features and injects opposite-source context into the student decoder; teacher predictions remain independent of the memory state. Inference uses the single student segmentation network. Under the reported fixed, single-run protocol on five ultrasound tasks from four dataset sources, FReMCI ranks highest in average Dice across all three labeled-data protocols ($1/10$, $3/10$, and $1/2$), outperforming the strongest competing averages by up to 3.87 points.

1 Introduction

Ultrasound (US) imaging is widely used for breast lesion, thyroid nodule, cardiac chamber, muscle fascicle, and aponeurosis assessment because it is portable, inexpensive, and radiation-free. Accurate segmentation of anatomical or pathological structures, often built on encoder–decoder designs such as U-Net (Ronneberger et al., 2015), can support measurement, diagnosis, treatment planning, and longitudinal follow-up. However, pixel-level US annotation is costly and variable: expert readers must interpret low-contrast boundaries, acoustic shadows, speckle noise, and probe-dependent appearance. This makes semi-supervised learning (SSL), which trains segmentation models from a small labeled set and many unlabeled images, a natural fit for US segmentation.

Modern SSL segmentation methods usually rely on consistency regularization, pseudo labeling, cross-model teaching, spatial perturbation, or region mixing. Mean Teacher-style models produce temporally smoothed targets (Tarvainen & Valpola, 2017; Yu et al., 2019); FixMatch and UniMatch strengthen weak-to-strong consistency (Sohn et al., 2020; Yang et al., 2023); cross pseudo supervision uses mutually trained predictors (Chen et al., 2021); and medical image methods further incorporate uncertainty, multi-decoder perturbations, and prototype consistency (Wu et al., 2022; Wang & Li, 2023; Lu et al., 2023; Assefa et al., 2025). These strategies have significantly improved label-efficient segmentation, but ultrasound exposes several common weaknesses in their training signals.

First, low-frequency image statistics and tissue–background appearance can differ substantially across scans because of scanner gain and other acquisition settings, even when their anatomical content is similar. Unrestricted cross-image transfer can therefore entangle acquisition style with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

anatomy. Second, pseudo labels are often unreliable around low-contrast or weakly defined boundaries. Speckle noise, shadowing, and small lesions can produce fragmented foreground support; region-mixing strategies can further introduce hard copy-paste seams that do not correspond to tissue boundaries. Treating these pixels as ordinary supervision reinforces both prediction errors and augmentation artifacts. Third, ultrasound foreground structures may occupy only a small fraction of an image and yield weak or sparse decoder activations. Cross-sample feature matching restricted to the current mini-batch may consequently provide too little foreground evidence or evidence biased toward the batch's acquisition style. These problems can reinforce one another: appearance shifts destabilize pseudo labels, unreliable regions contaminate spatial intervention, and weak foreground evidence limits semantic transfer.

These observations motivate selective cross-sample information transfer under ultrasound uncertainty. FReMCI implements this principle at the frequency, region, and feature-memory levels rather than treating the three components as unrelated augmentations. Phase-Preserved Low-Frequency Amplitude Sharing (PLAS) symmetrically interpolates low-frequency amplitude between paired labeled and unlabeled images while retaining their individual phase and high-frequency content, providing a controlled appearance perturbation without exchanging spatial geometry. Reliability-Guided Semantic Region Intervention (ReSRI) uses an EMA teacher and teacher–student agreement to construct foreground-centered exchanges from labeled or reliability-filtered pseudo foreground, while unreliable pixels and artificial seam bands receive zero loss weight. Memory Cross Injection (MCI) stores labeled and reliable unlabeled foreground features and retrieves opposite-source context for the student decoder. The EMA teacher operates without memory, so its pseudo targets are computed independently of the memory state. A conceptual problem–solution summary is provided in Appendix I. The complete training workflow is summarized in Figure 1. The upper panel constructs the reliable and mixed supervision signals, whereas the lower panel shows how the resulting inputs are processed by the student with cross-sample foreground memory.

Our contributions are threefold. First, we develop PLAS by extending phase-preserving Fourier amplitude interpolation to a paired labeled–unlabeled perturbation placed before reliability-guided region intervention. Second, we design ReSRI, which reuses one pixel-reliability map for weighted pseudo supervision, foreground-centered bidirectional region construction, and seam-excluded region loss. Third, we propose MCI, which retrieves opposite-source foreground context from separate memories of labeled and reliable unlabeled features and injects it into student decoder features; the complete framework is evaluated against eighteen representative methods under three label ratios on five tasks from four ultrasound dataset sources.

2 Related Work

2.1 Reliable Semi-Supervision and Region Mixing

Semi-supervised segmentation learns from a small labeled set and many unlabeled images through consistency or pseudo supervision. Mean Teacher (Tarvainen & Valpola, 2017), FixMatch (Sohn et al., 2020), UniMatch (Yang et al., 2023), and cross pseudo supervision (Chen et al., 2021) are representative formulations. Medical variants add uncertainty estimation, cross teaching, prototype consistency, and region mixing (Yu et al., 2019; Wu et al., 2022; Wang & Li, 2023; Lu et al., 2023; Bai et al., 2023). Recent methods adapt the perturbation or supervision strength to the model's learning state and the reliability of individual regions (Lei et al., 2025; Shen et al., 2026), while BCP performs bidirectional copy-paste for medical SSL (Bai et al., 2023). ReSRI uses a foreground-centered box rather than a class mask, and its reliability map controls the pseudo-target weight, unlabeled box eligibility, and memory admission; a separate seam band removes direct loss supervision at the synthetic boundary.

2.2 Frequency-Domain Augmentation

Fourier-domain methods exploit the complementary roles of amplitude and phase: amplitude is associated with appearance or style, while phase preserves content and structural cues (Yi et al., 2024; Lee et al., 2024). Recent medical-image methods use Fourier reconstruction or semantic augmentation to improve domain generalization (Zhao et al., 2024; Oh et al., 2026). PLAS does not claim the underlying amplitude–phase decomposition as new. Its distinction is operational: it

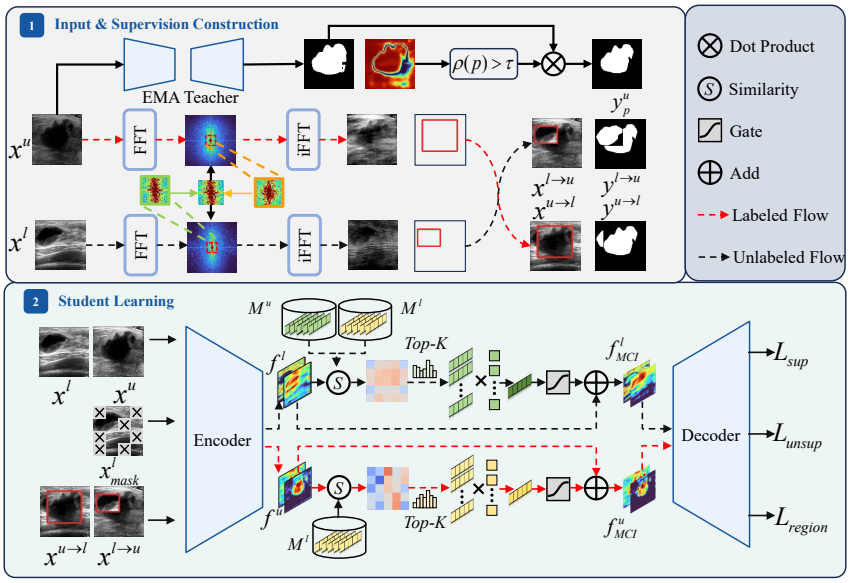


Figure 1: FReMCI overview. PLAS supplies frequency-perturbed views to ReSRI, which constructs reliability-guided bidirectional region supervision with seam exclusion. MCI cross-injects labeled and reliable unlabeled memories only in the paired student feature path; the EMA teacher and mixed-region forward pass are memory-free.

forms one symmetric convex low-frequency amplitude from a labeled–unlabeled pair, retains each image's remaining amplitude and phase, and supplies the paired views specifically to the subsequent reliability-guided region branch.

2.3 Memory-Based Dense Representation

Memory and prototype mechanisms provide context beyond a mini-batch. Recent medical SSL methods improve representations through prototype consistency or cross-image semantic correlation (Lu et al., 2023; Gao et al., 2023; Ding et al., 2025). MCI differs in both memory source and use: it keeps labeled and reliability-filtered unlabeled foreground queues separate, retrieves top-k entries only from the opposite source group, and injects the aggregated context residually into student decoder features without introducing a contrastive objective.

3 Methods

3.1 Overview

Let the labeled set be $D_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and the unlabeled set be $D_u = \{x_i^u\}_{i=1}^{N_u}$, where $N_l \ll N_u$. The student network f_θ is a single encoder-decoder segmentation model that predicts a dense logit map for each image, and an EMA teacher $f_{\bar{\theta}}$ provides stable pseudo labels for unlabeled images. FReMCI contains three training-time modules: PLAS perturbs ultrasound frequency style, ReSRI constructs reliability-guided semantic region supervision, and MCI injects cross-sample foreground memory into the student decoder.

Figure 1 gives the corresponding data flow. In the upper panel, the memory-free EMA teacher predicts the weak unlabeled view and, together with the paired student prediction, produces the reliability-weighted pseudo target. In parallel, PLAS operates on the labeled image and strongly augmented unlabeled view before ReSRI constructs labeled-to-unlabeled and unlabeled-to-labeled region exchanges. In the student path, MCI is applied to paired labeled–unlabeled decoder features used for supervised and pseudo-supervised learning. The ReSRI mixed-region batch is processed by the same student encoder–decoder without memory injection and is optimized by the region loss. All

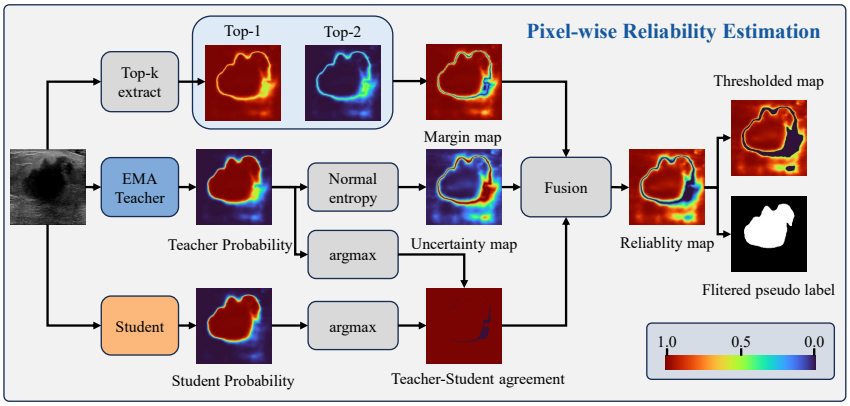


Figure 2: ReSRI reliability estimation. Teacher confidence, entropy certainty, top-two margin, and teacher–student agreement form ρ; thresholding filters pseudo labels and ignores unreliable pixels.

three proposed modules are training-only. Figure 2 previews the pixel-level reliability computation used by ReSRI and detailed in Section 3.3.

For a mini-batch $\mathcal{B} = \{(x^l, y^l), x_w^u, x_s^u\}$ with weak and strong unlabeled views, the student first predicts the paired labeled and weak unlabeled images, with MCI enabled after memory warm-up. The teacher predicts the weak unlabeled view without memory. The student is then optimized on a masked unlabeled view with MCI and on a separate, memory-free batch of reliability-guided mixed regions generated after PLAS. The supervised loss on labeled samples is

$$L_{\text{sup}} = L_{\text{ce}}(f_\theta(x^l), y^l) + L_{\text{dice}}(f_\theta(x^l), y^l). \tag{1}$$

The teacher parameters are updated as

$$\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta. \tag{2}$$

The final objective combines supervised learning, reliability-weighted pseudo supervision, and seam-suppressed region supervision. The EMA teacher, PLAS, ReSRI, and MCI are used only during training.

3.2 PLAS: Phase-Preserved Low-Frequency Amplitude Sharing

US appearance varies with scanner gain, probe pressure, tissue acoustic properties, and reconstruction settings. Motivated by recent Fourier reconstruction and augmentation methods (Zhao et al., 2024; Oh et al., 2026; Yi et al., 2024), PLAS perturbs low-frequency amplitude while retaining phase components associated with spatial structure (Lee et al., 2024). For a labeled image x^l and a strongly augmented unlabeled image x_s^u, we compute their two-dimensional Fourier transforms:

$$\mathcal{F}(x) = A(x)\exp(jP(x)), \tag{3}$$

where A and P denote amplitude and phase. The subscript in A_s denotes the shared amplitude. A centered binary mask $M_r \in \{0,1\}^{H\times W}$ selects a low-frequency square with ratio r. Rather than hard-swapping spectra, PLAS forms a convex shared low-frequency amplitude:

$$A_s = \alpha_A M_r A(x^l) + (1-\alpha_A) M_r A(x_s^u). \tag{4}$$

For each image $x \in \{x^l, x_s^u\}$, only the selected low-frequency amplitude is replaced, while the image-specific high-frequency amplitude and phase are retained:

$$\hat{x} = \mathcal{F}^{-1}(((1-M_r)A(x) + A_s)\exp(jP(x))). \tag{5}$$

The final PLAS view is interpolated in image space:

$$\tilde{x} = x + \alpha_x(\hat{x} - x). \tag{6}$$

Here α_A controls the labeled–unlabeled amplitude blend and α_x controls the residual interpolation back to the original image; both are set to 0.5. This design exposes paired samples to partially shared low-frequency statistics while retaining each sample's phase and remaining spectrum. PLAS is applied before region intervention and removed at inference; the present experiments do not by themselves establish cross-scanner robustness.

3.3 ReSRI: Reliability-Guided Semantic Region Intervention

Spatial intervention is useful only when the mixed target is trustworthy. The EMA teacher predicts unlabeled logits $z^t = f_{\bar{\theta}}(x_w^u)$ and probabilities $p^t = \mathrm{softmax}(z^t)$, while the student produces $z^s = f_\theta(x_w^u)$ and $p^s = \mathrm{softmax}(z^s)$. The teacher probabilities define the pseudo label $\hat{y}^u = \arg\max_c p_c^t$. For each pixel, FReMCI estimates reliability from teacher confidence q, normalized entropy h, top-two probability margin m, and teacher–student agreement a:

$$\rho = q(1-h)(0.5+0.5m)a. \tag{7}$$

Here $q = \max_c p_c^t$, $h = -\sum_c p_c^t \log p_c^t / \log C$, $m = p_{(1)}^t - p_{(2)}^t$, and $a = \mathbf{1}[\arg\max_c p_c^s = \hat{y}^u]$ compares the student and teacher labels. A preliminary floor $\tau_0 = 0.35$ sets smaller values of ρ to zero, and the region threshold is $\tau = 0.45$. In binary segmentation, q, h, and m are deterministic summaries of the same teacher distribution rather than independent uncertainty estimators; their product is used as a conservative nonlinear confidence shaping, while a contributes the separate teacher–student consistency test. We therefore use ρ as an operational training weight and do not interpret it as a calibrated probability of pseudo-label correctness. The reliable pseudo target used by the region branch is

$$y_\rho^u(p) = \begin{cases} \hat{y}^u(p), & \rho(p) \geq \tau, \\ \text{ignore}, & \rho(p) < \tau. \end{cases} \tag{8}$$

Figure 2 visualizes this computation on a fixed UDIAT example. The top-1 and top-2 maps are extracted exclusively from the EMA teacher probability p^t; their difference gives the teacher's class-separation margin m. The student probability p^s is used only in the agreement branch, which compares the teacher and student argmax labels and produces a. Thus, the teacher confidence, entropy, margin, and pseudo label are not mixed with student probabilities. The four evidence terms are fused pixel by pixel to produce the continuous reliability map ρ. Thresholding at τ preserves reliable foreground supervision while assigning low-reliability boundary pixels to the ignore set. ReSRI therefore filters the pseudo target used for region construction while retaining ρ for reliability-weighted optimization.

FReMCI then constructs foreground-centered rectangular regions from reliable evidence. For labeled samples, $M_l = \mathrm{Box}(\mathrm{Dilate}((y^l > 0) \wedge (y^l \neq \text{ignore})))$. For unlabeled samples, $M_u = \mathrm{Box}(\mathrm{Dilate}((\hat{y}^u > 0) \wedge (\rho \geq \tau)))$. The box operator returns the tight foreground bounding box with 15% padding, so the intervention deliberately retains contiguous local context and may include surrounding background; reliability determines whether an unlabeled foreground can anchor the box, not whether every pixel inside is foreground. If no reliable unlabeled foreground exists, a random box preserves a defined tensor path, but it does not create random labels: targets still come from y_ρ^u, and ignored pixels retain zero weight.

FReMCI performs bidirectional intervention. The labeled-to-unlabeled image is

$$x^{l\to u} = M_l \tilde{x}^l + (1 - M_l)\tilde{x}^u, \tag{9}$$

with target and pixel weight

$$y^{l\to u} = M_l y^l + (1 - M_l) y_\rho^u, \quad w^{l\to u} = M_l + (1 - M_l)\rho. \tag{10}$$

The unlabeled-to-labeled image is

$$x^{u\to l} = M_u \tilde{x}^u + (1 - M_u)\tilde{x}^l, \tag{11}$$

with

$$y^{u\to l} = M_u y_\rho^u + (1 - M_u) y^l, \quad w^{u\to l} = M_u \rho + (1 - M_u). \tag{12}$$

Ground-truth pixels have unit weight; pseudo-labeled pixels are weighted by ρ; unreliable pixels are ignored. To avoid supervising non-anatomical paste boundaries, FReMCI computes a seam band for each intervention mask:

$$S(M) = \mathrm{Dilate}_b(M) - \mathrm{Erode}_b(M), \tag{13}$$

and sets region-loss weights to zero on $S(M)$. This operation preserves valid supervision away from the intervention boundary and removes direct region-loss supervision on the artificial seam.

3.4 MCI: Memory Cross Injection

Ultrasound foreground structures are often small, weakly contrasted, and under-represented in each mini-batch. FReMCI therefore maintains a foreground memory that accumulates normalized decoder features from labeled foreground pixels and reliable unlabeled foreground pixels. The memory is implemented as two first-in-first-out feature queues, $\mathcal{M}_l$ and $\mathcal{M}_u$. Given decoder feature $g_\theta(x)$, the labeled memory is updated with normalized features sampled from $(y^l > 0) \wedge (y^l \neq \text{ignore})$, while the unlabeled memory is updated only from pixels satisfying $(\hat{y}^u > 0) \wedge (\rho \geq \tau_m)$:

$$\mathcal{M}_l \leftarrow \text{Enqueue}(\mathcal{M}_l, \text{L2Norm}(g_\theta(x^l)_{(y^l>0)\wedge(y^l\neq\text{ignore})})), \tag{14}$$

$$\mathcal{M}_u \leftarrow \text{Enqueue}(\mathcal{M}_u, \text{L2Norm}(g_\theta(x^u_w)_{\hat{y}^u>0,\rho\geq\tau_m})). \tag{15}$$

Old entries are removed when the memory exceeds its capacity. The update is performed without gradients after the optimization step, so features inserted at the current iteration can only be retrieved by later iterations. Algorithm 2 details the reliable foreground selection, sampling, normalization, and FIFO operations.

During student decoding, FReMCI injects memory context into decoder features through similarity-based cross injection. For a decoder query feature g_i, the query projection Q_i, memory key K_j, and memory value V_j are used for similarity-based retrieval:

$$s_{ij} = Q_i^\top K_j, \quad \beta_{ij} = \frac{\exp(s_{ij}/T)}{\sum_{j'\in\mathcal{N}_k(i)} \exp(s_{ij'}/T)}, \tag{16}$$

where $\mathcal{N}_k(i)$ denotes the top-k retrieved memory entries and $s_{i(1)}$ denotes the largest similarity among them. The memory context is

$$c_i = \sum_{j\in\mathcal{N}_k(i)} \beta_{ij} V_j, \tag{17}$$

and the injected feature is

$$g'_i = g_i + \omega\, \gamma_i c_i, \quad \gamma_i = \sigma((s_{i(1)} - \delta)/T). \tag{18}$$

Labeled memory is injected into unlabeled features with weight $\omega_{l\to u}$, while reliable unlabeled memory is injected into labeled features with weight $\omega_{u\to l}$. We use $T = 0.1$, $\delta = 0.1$, $\omega_{l\to u} = 1.0$, and $\omega_{u\to l} = 0.3$. The EMA teacher and ReSRI mixed-region batch never receive the memory state. Consequently, pseudo targets are generated by the plain teacher network, whereas memory context affects only the paired student feature path.

3.5 Loss Function and Inference

To exploit unlabeled images, the student predicts the masked view $\bar{x}^u$, while the EMA teacher provides pseudo labels from the corresponding weak view. The unlabeled objective is a reliability-weighted CE-Dice loss:

$$L_u = L_{\text{wce}}(f_\theta(\bar{x}^u), \hat{y}^u, \rho) + L_{\text{wdice}}(f_\theta(\bar{x}^u), \hat{y}^u, \rho). \tag{19}$$

The region loss L_{region} applies the same weighted CE-Dice form to the concatenated bidirectional intervention images, using unit weights for ground-truth pixels, reliability weights for pseudo-labeled pixels, and zero weights for unreliable or seam pixels. The full training objective is

$$L = L_{\text{sup}} + \lambda_u(e) L_u + \lambda_r L_{\text{region}}. \tag{20}$$

Here $\lambda_u(e)$ follows a 40-epoch sigmoid ramp to 1.0 after warm-up, and $\lambda_r = 1.0$. The foreground memory modifies student decoder features during training but introduces no additional loss term. Algorithms 1 and 2 in Appendix C summarize the complete optimization and memory-update procedures. At inference, the EMA teacher, PLAS, reliability computation, ReSRI, seam weighting, and foreground memory are discarded. Prediction uses only the student segmentation network.

Table 1: Results under $1/10$ labels (%). Best values are bold; second-best values are underlined.

Method	UDIAT		Aneil		Aryan		BUSI		DDTI		Avg.	
	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.
UAMT (MICCAI 2019)	45.25	36.26	69.34	54.09	78.48	65.01	65.24	55.09	63.61	50.56	64.38	52.20
CPS (CVPR 2021)	46.25	37.09	69.87	54.77	78.49	65.03	64.55	55.43	61.48	48.10	64.13	52.08
MC-Net V1 (MICCAI 2021)	54.26	44.11	69.37	54.35	78.00	64.41	65.39	55.94	56.05	42.60	64.61	52.28
URPC (MedIA 2022)	50.85	41.33	69.16	54.27	77.61	63.94	66.70	56.96	63.54	51.03	65.57	53.51
MC-Net V2 (MedIA 2022)	55.26	45.34	69.90	54.77	77.35	63.66	65.69	55.66	56.48	42.61	64.94	52.41
BCP (CVPR 2023)	62.30	49.82	72.92	58.07	79.17	65.84	67.02	57.05	63.27	49.42	68.94	56.04
UniMatch (CVPR 2023)	69.00	57.45	74.36	59.71	79.81	66.70	66.75	56.81	67.08	54.15	71.40	58.96
CAML (MICCAI 2023)	64.16	52.99	70.54	55.69	78.61	65.32	66.50	56.27	62.66	49.65	68.49	55.98
UPCoL (MICCAI 2023)	53.45	41.15	67.50	51.86	71.17	56.19	63.55	51.58	43.22	29.11	59.78	45.98
ABD (CVPR 2024)	65.26	54.94	67.67	52.11	78.68	65.28	63.55	54.04	63.45	50.03	67.72	55.28
CorrMatch (CVPR 2024)	63.59	51.42	71.39	56.44	**80.69**	**67.86**	70.69	60.86	66.20	52.87	70.51	57.89
PH-Net (CVPR 2024)	65.75	52.98	74.78	60.16	80.04	66.95	71.20	61.32	56.68	41.54	69.69	56.59
UnCo (TMI 2025)	68.58	58.86	70.29	55.27	79.14	65.97	68.78	58.49	64.15	52.14	70.19	58.15
SKCDF (CVPR 2025)	67.24	55.05	70.94	56.04	78.88	65.54	67.15	55.12	57.45	42.61	68.33	54.87
CSC-PA (CVPR 2025)	67.45	55.22	72.09	57.52	79.54	66.36	71.77	62.26	69.50	56.99	72.07	59.67
BCSI-2D (AAAI 2026)	59.93	49.31	68.68	53.30	74.25	59.74	62.46	53.69	55.60	41.59	64.18	51.53
AdaMix (MedIA 2026)	55.35	45.42	66.41	51.17	52.01	36.68	69.57	60.37	59.57	46.81	60.58	48.09
CHAP (MedIA 2026)	65.31	54.61	72.73	57.94	78.65	65.26	66.04	55.89	67.00	53.72	69.95	57.48
FReMCI (Ours)	**74.47**	**63.89**	**76.60**	**62.58**	80.24	67.27	**73.41**	**64.53**	**74.99**	**63.57**	**75.94**	**64.37**

4 EXPERIMENTS

4.1 DATASETS AND PROTOCOLS

We evaluate FReMCI on five ultrasound segmentation tasks from four dataset sources, covering breast lesions, thyroid nodules, and lower-leg musculoskeletal structures. The two rows derived from FALLMUD[1] are treated as separate aponeurosis tasks, not independent dataset sources. Sources, targets, and overall sample counts are summarized in Table 5. Experiments use the $1/10$, $3/10$, and $1/2$ labeled protocols. Dice and Jaccard, reported as percentages, are the primary metrics; higher values are better.

4.2 IMPLEMENTATION DETAILS

All images are resized to 256×256. FReMCI uses a ResNet-50–DeepLabV3+ network trained for 200 epochs with SGD. The principal PLAS, ReSRI, and MCI settings are $r = 0.125$, $\tau = 0.45$, and top-$k = 16$, respectively. Complete optimization, memory, and region-intervention settings are reported in Appendix D. We compare against eighteen semi-supervised segmentation methods under the same input resolution and labeled ratios; method details and citations are provided in Appendix B.

4.3 QUANTITATIVE RESULTS

Tables 1–3 report fixed-protocol, single-run point estimates. Accordingly, the following comparisons describe the observed table values and should not be read as estimates of seed-to-seed significance or cross-scanner robustness.

Under the $1/10$ protocol (Table 1), FReMCI records the highest five-task average Dice/Jaccard, $75.94/64.37$. Relative to the strongest competing average, CSC-PA at $72.07/59.67$, the observed margins are $3.87/4.70$ points. FReMCI has the highest point estimate on four tasks, while CorrMatch is highest on Aryan.

Under $3/10$ labels, FReMCI records an average Dice/Jaccard of $78.62/67.30$ (Table 2), with observed margins of $2.52/2.75$ points over CorrMatch at $76.10/64.55$. Its point estimates are highest across all five tasks in both metrics in this run.

With half of the training images labeled, FReMCI records the highest average Dice/Jaccard, $78.49/67.36$ (Table 3), with observed margins of $0.88/1.13$ points over CorrMatch at $77.61/66.23$.

[1] https://kalisteo.cea.fr/index.php/fallmud/

Table 2: Results under $3/10$ labels (%). Best values are bold; second-best values are underlined.

Method	UDIAT		Aneil		Aryan		BUSI		DDTI		Avg.	
	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.
UAMT (MICCAI 2019)	62.53	52.97	75.43	61.15	80.89	68.21	66.60	57.71	71.16	59.39	71.32	59.89
CPS (CVPR 2021)	63.52	53.62	74.74	60.38	80.72	67.98	68.48	59.35	72.51	60.68	71.99	60.40
MC-Net V1 (MICCAI 2021)	63.89	53.62	76.62	62.63	80.55	67.80	66.98	56.47	69.72	57.52	71.55	59.61
URPC (MedIA 2022)	62.18	52.28	75.25	60.96	80.66	67.95	68.45	59.39	70.52	58.54	71.41	59.82
MC-Net V2 (MedIA 2022)	65.06	55.17	75.97	61.84	80.45	67.65	68.18	58.68	66.48	53.37	71.23	59.34
BCP (CVPR 2023)	63.37	49.77	73.00	58.22	80.33	67.37	66.30	55.74	65.99	52.68	69.80	56.76
UniMatch (CVPR 2023)	70.40	59.98	76.65	62.59	80.61	67.72	66.40	57.06	67.32	54.36	72.28	60.34
CAML (MICCAI 2023)	66.16	56.48	76.72	62.78	81.02	68.39	68.58	59.19	71.77	59.96	72.85	61.36
UPCoL (MICCAI 2023)	61.23	48.42	71.49	56.62	77.12	63.25	68.29	57.13	53.51	39.06	66.33	52.90
ABD (CVPR 2024)	66.65	56.96	73.82	59.20	81.30	68.73	66.23	56.27	70.84	58.44	71.77	59.92
CorrMatch (CVPR 2024)	77.47	67.27	75.46	61.31	81.07	68.35	72.04	62.92	74.46	62.92	76.10	64.55
PH-Net (CVPR 2024)	68.23	56.80	77.60	63.78	81.35	68.77	73.07	63.71	71.95	60.46	74.44	62.70
UnCo (TMI 2025)	73.00	61.96	76.06	62.03	81.11	68.53	67.95	59.18	70.02	58.11	73.63	61.96
SKCDF (CVPR 2025)	62.58	50.10	73.37	58.62	80.88	68.19	68.64	57.75	58.73	44.23	68.84	55.78
CSC-PA (CVPR 2025)	76.23	65.75	76.15	62.04	80.99	68.31	74.56	64.99	72.54	60.65	76.09	64.35
BCSI-2D (AAAI 2026)	68.38	57.58	74.29	59.78	78.32	64.80	64.50	55.28	63.92	50.54	69.88	57.60
AdaMix (MedIA 2026)	66.75	56.76	75.48	61.17	64.24	47.73	65.51	58.12	70.03	57.84	68.40	56.32
CHAP (MedIA 2026)	71.51	59.66	76.99	63.03	81.19	68.60	69.33	59.41	70.77	57.57	73.96	61.65
FReMCI (Ours)	**81.89**	**71.57**	**78.08**	**64.44**	**81.51**	**69.00**	**75.10**	**66.29**	**76.51**	**65.18**	**78.62**	**67.30**

Table 3: Results under $1/2$ labels (%). Best values are bold; second-best values are underlined.

Method	UDIAT		Aneil		Aryan		BUSI		DDTI		Avg.	
	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.	Dice	Jac.
UAMT (MICCAI 2019)	67.71	57.44	76.28	62.21	81.61	69.25	69.71	61.12	72.42	60.61	73.55	62.13
CPS (CVPR 2021)	67.03	57.73	76.49	62.47	81.71	69.38	70.13	60.36	71.59	59.92	73.39	61.97
MC-Net V1 (MICCAI 2021)	67.86	58.02	76.77	62.88	81.74	69.43	70.05	60.63	69.40	57.49	73.16	61.69
URPC (MedIA 2022)	67.95	57.30	76.27	62.15	81.35	68.85	69.22	61.05	71.45	59.09	73.25	61.69
MC-Net V2 (MedIA 2022)	67.87	56.10	77.45	63.63	81.59	69.20	69.33	60.15	66.34	53.33	72.52	60.48
BCP (CVPR 2023)	64.88	51.99	72.46	57.50	81.18	68.58	67.14	57.60	61.99	48.50	69.53	56.83
UniMatch (CVPR 2023)	73.32	61.75	77.19	63.23	81.42	68.93	71.59	61.12	72.11	59.94	75.13	62.99
CAML (MICCAI 2023)	70.92	59.66	77.38	63.56	81.98	69.72	71.24	61.52	71.80	60.19	74.66	62.93
UPCoL (MICCAI 2023)	63.28	50.30	73.75	59.06	78.51	64.99	70.48	59.46	57.69	43.39	68.74	55.44
ABD (CVPR 2024)	68.41	57.11	75.48	61.26	**82.39**	**70.26**	70.86	60.77	73.27	61.03	74.08	62.09
CorrMatch (CVPR 2024)	**78.60**	**68.17**	76.55	62.71	81.61	69.11	74.62	65.85	76.69	65.31	77.61	66.23
PH-Net (CVPR 2024)	71.34	59.90	77.59	63.83	82.09	69.80	75.35	**67.04**	69.22	56.26	75.12	63.37
UnCo (TMI 2025)	71.23	60.26	76.46	62.42	81.78	69.44	70.43	61.53	69.76	57.50	73.93	62.23
SKCDF (CVPR 2025)	67.33	55.05	74.33	59.73	81.44	68.91	66.65	55.10	58.97	44.07	69.74	56.57
CSC-PA (CVPR 2025)	75.86	65.13	77.26	63.39	81.41	68.85	75.57	66.16	75.37	63.84	77.09	65.47
BCSI-2D (AAAI 2026)	68.90	58.07	75.85	61.62	80.93	68.30	66.07	56.86	66.26	53.91	71.60	59.75
AdaMix (MedIA 2026)	69.82	59.78	76.52	62.41	53.57	37.09	68.38	59.97	71.14	59.45	67.89	55.74
CHAP (MedIA 2026)	74.18	63.48	76.62	62.74	82.12	69.92	71.44	61.45	71.24	58.48	75.12	63.21
FReMCI (Ours)	78.05	68.16	**78.57**	**64.99**	82.12	69.88	**75.94**	**67.04**	**77.75**	**66.72**	**78.49**	**67.36**

It is highest on Aneil, BUSI, and DDTI, second on UDIAT, and below ABD on Aryan Dice. Dataset- and baseline-specific discussion is provided in Appendix B.

4.4 Qualitative Results

We visualize the most label-scarce $1/10$ protocol, where unreliable pseudo labels and boundary errors have the largest effect. Figure 3 compares eleven representative baselines and FReMCI on fixed cases from all five tasks. Green, red, and yellow denote ground-truth-only, prediction-only, and intersecting regions, respectively; bright-blue text reports per-image Dice.

In these displayed cases, FReMCI shows compact support, continuous thin structures, and limited boundary leakage, with the highest displayed per-image Dice in all five examples. These selected examples illustrate spatial error patterns and are not used to estimate dataset-level performance.

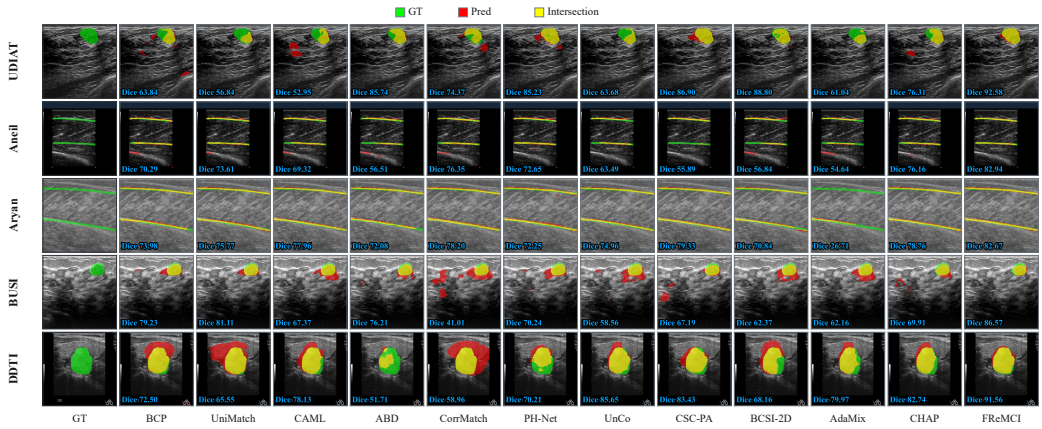


Figure 3: Qualitative comparison under $1/10$ labels on five datasets. Green/red/yellow denote ground-truth-only/prediction-only/intersection; blue text reports Dice.

Table 4: BUSI $1/10$ structural ablation and parameter sensitivity (%). Best settings are bold.

(a) Structural components

Configuration	PLAS	ReSRI	MCI	Dice	Jac.
Base				64.57	54.51
+ReSRI		√		67.82	57.63
+MCI			√	66.91	56.62
W/O MCI	√	√		69.54	59.62
W/O PLAS		√	√	70.26	60.51
FReMCI	√	√	√	**73.41**	**64.53**

(b) Parameter sensitivity

PLAS ratio r			MCI retrieval k		
r	Dice	Jac.	k	Dice	Jac.
$1/10$	72.83	63.74	8	72.74	63.70
$1/8$	**73.41**	**64.53**	**16**	**73.41**	**64.53**
$1/5$	73.08	64.15	32	73.12	64.18
$1/4$	73.10	63.89	64	73.15	64.11
$3/10$	71.92	62.93	128	71.88	63.20

4.5 Ablation Studies

Structural components. We evaluate PLAS, ReSRI, and MCI on BUSI under the $1/10$ protocol. All variants share the EMA reliability estimator and reliability-weighted masked pseudo supervision. Since PLAS is defined within the ReSRI region branch, we report the six dependency-aware configurations in Table 4.

In this BUSI $1/10$ run, ReSRI and MCI alone are $3.25/3.12$ and $2.34/2.11$ Dice/Jaccard points above the Base. Adding PLAS to ReSRI and MCI to ReSRI yields $69.54/59.62$ and $70.26/60.51$, while the full configuration reaches $73.41/64.53$ ($+8.84/+10.02$ over the Base). Removing MCI or PLAS lowers the corresponding point estimate by 3.87 or 3.15 Dice points. These component trends motivate the combined design, but the single-task ablation does not isolate every sub-operation or establish statistical interaction among modules. Structural-ablation heatmaps are provided in Appendix G.

Parameter sensitivity. Table 4(b) supports the PLAS low-frequency ratio $r = 0.125$ and MCI retrieval size $k = 16$; ReSRI uses $\tau = 0.45$. The corresponding sensitivity curves and detailed discussion are reported in Appendix H.

5 Conclusion

We presented FReMCI, a frequency-reliable memory cross intervention framework that controls cross-sample transfer at frequency, region, and feature-memory levels. PLAS provides a phase-preserving low-frequency perturbation before region exchange, ReSRI constructs reliability-weighted pseudo and seam-excluded region supervision, and MCI injects opposite-source foreground context into the student decoder. Under the reported fixed, single-run protocol on five tasks from four sources, FReMCI records the highest average Dice and Jaccard among the evaluated methods at all three label ratios, with the largest observed margin at $1/10$ labels.

AI USE STATEMENT

Generative AI tools were used for limited assistance with figure asset organization, format conversion, language polishing, and code/document quality checks. All scientific claims, equations, experimental values, figures, and final manuscript content were reviewed and verified by the authors, who take responsibility for the submitted work.

REPRODUCIBILITY STATEMENT

The method equations, training pseudocode, optimizer schedule, memory-update procedure, and module settings are provided in the main paper and Appendix D. Unless explicitly varied in Table 4, the manuscript uses $r = 0.125$, $\tau = 0.45$, and $k = 16$. The reported benchmark values are fixed-protocol point estimates; repeated-seed confidence intervals and training-cost measurements are not available and are not implied by the reported results.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

Maregu Assefa, Muzammal Naseer, Iyyakutti Iyappan Ganapathi, Syed Sadaf Ali, Mohamed L Seghier, and Naoufel Werghi. Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30850–30860, 2025.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11514–11524, 2023.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021.

Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4070–4080, 2024.

Zhenhui Ding, Guilian Chen, Qin Zhang, Huisi Wu, and Jing Qin. CSC-PA: Cross-image semantic correlation via prototype attentions for single-network semi-supervised breast tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15632–15641, 2025.

Shengbo Gao, Ziji Zhang, Jiechao Ma, Zihao Li, and Shu Zhang. Correlation-aware mutual learning for semi-supervised medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023*, pp. 98–108. Springer Nature Switzerland, 2023.

Kaiwen Huang, Yizhe Zhang, Yi Zhou, Tianyang Xu, and Tao Zhou. Bidirectional channel-selective semantic interaction for semi-supervised medical segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5040–5048, 2026.

Siyao Jiang, Huisi Wu, Junyang Chen, Qin Zhang, and Jing Qin. PH-Net: Semi-supervised breast lesion segmentation via patch-wise hardness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11418–11427, 2024.

Ingyun Lee, Wooju Lee, and Hyun Myung. Domain generalization with vital phase augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):2892–2900, 2024.

Tao Lei, Yi Wang, Xingwu Wang, Xuan Wang, Bin Hu, and Asoke K. Nandi. Unified feature consistency of under-performing pixels and valid regions for semi-supervised medical image segmentation. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 9(2):169–181, 2025.

Wenjing Lu, Jiahao Lei, Peng Qiu, Rui Sheng, Jinhua Zhou, Xinwu Lu, and Yang Yang. Upcol: Uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 662–672. Springer, 2023.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022.

Kwanseok Oh, Eunjin Jeon, Da-Woon Heo, Yooseung Shin, and Heung-Il Suk. FIESTA: Fourier-based semantic augmentation with uncertainty guidance for enhanced domain generalizability in medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 37 (3):1063–1077, 2026.

Lina Pedraza, Carlos Vargas, Fabian Narvaez, Oscar Duran, Eduardo Munoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, pp. 188–193, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Zhiqiang Shen, Peng Cao, Junming Su, Jinzhu Yang, and Osmar R Zaiane. Adaptive mix for semi-supervised medical image segmentation. *Medical Image Analysis*, 108:103857, 2026.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3097–3107, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

Haonan Wang and Xiaomeng Li. Towards generic semi-supervised framework for volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 36:1833–1848, 2023.

Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 297–306. Springer, 2021.

Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7236–7246, 2023.

Moi Hoon Yap, Gemma Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Andrew K. Davison, and Robert Marti. Breast ultrasound region of interest detection and lesion localisation. *Artificial Intelligence in Medicine*, 107:101880, 2020.

Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8159–8168. Association for Computing Machinery, 2024.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. 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.

Xiang Zeng, Shengwu Xiong, Jinming Xu, Guangxing Du, and Yi Rong. Uncertainty co-estimator for improving semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(9):3870–3881, 2025.

Zheng Zhang, Guanchun Yin, Bo Zhang, Wu Liu, Xiuzhuang Zhou, and Wendong Wang. A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25940–25949, 2025.

Haoyu Zhao, Wenhui Dong, Rui Yu, Zhou Zhao, Bo Du, and Yongchao Xu. MoreStyle: Relax low-frequency constraint of Fourier-based image reconstruction in generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 434–444. Springer, 2024.

Si-Ping Zhou, Zhi-Fang Gong, Kai-Ni Wang, Ping Zhou, Yang Chen, and Guang-Quan Zhou. CHAP: Channel-spatial hierarchical adversarial perturbation for semi-supervised medical image segmentation. *Medical Image Analysis*, 109:103918, 2026.

A Dataset Details

Our benchmark contains five ultrasound segmentation tasks from four dataset sources, covering breast lesions, thyroid nodules, and aponeurosis structures. BUSI, DDTI, and UDIAT are public datasets cited by their original releases, while the two FALLMUD rows are processed aponeurosis subsets from the same source and are evaluated as separate tasks. Table 5 reports the overall number of paired image–label files in each dataset. Experiments use the $1/10$, $3/10$, and $1/2$ labeled protocols. All images are resized to 256×256 before training.

The reported average assigns equal weight to the five task rows; it is not an average over four statistically independent dataset sources. BUSI and UDIAT contain breast lesions, whereas DDTI contains thyroid nodules.

Table 5: Dataset summary.

Dataset	Target structure	Source	Total
FALLMUD-Aneil	Aponeurosis	FALLMUD	309
FALLMUD-Aryan	Aponeurosis	FALLMUD	504
Breast-BUSI	Breast lesion	BUSI (Al-Dhabyani et al., 2020)	647
ThyroidNodule-DDTI	Thyroid nodule	DDTI (Pedraza et al., 2015)	637
Breast-UDIAT	Breast lesion	UDIAT (Yap et al., 2020)	163

B Comparison Methods and Detailed Quantitative Discussion

We compare FReMCI with eighteen representative semi-supervised methods: ABD (Chi et al., 2024), AdaMix (Shen et al., 2026), BCP (Bai et al., 2023), BCSI-2D (Huang et al., 2026), CAML (Gao et al., 2023), CHAP (Zhou et al., 2026), CorrMatch (Sun et al., 2024), MC-Net V1 (Wu et al., 2021), MC-Net V2 (Wu et al., 2022), PH-Net (Jiang et al., 2024), SKCDF (Zhang et al., 2025), CPS (Chen et al., 2021), UAMT (Yu et al., 2019), URPC (Luo et al., 2022), UnCo (Zeng et al., 2025), UniMatch (Yang et al., 2023), UPCoL (Lu et al., 2023), and CSC-PA (Ding et al., 2025). These methods cover displacement and region mixing, adversarial perturbation, correlation matching, mutual consistency, patch-wise hardness modeling, semantic decoupling, uncertainty-aware self-ensembling, prototype learning, and correlation-based SSL.

The methods retain their original architectures and training mechanisms, while the input resolution and label ratios are kept fixed. Dataset-wise columns reveal task-specific behavior, and the average columns summarize performance across the five ultrasound tasks.

Half-label protocol. At the five-task average level, FReMCI records 78.49 Dice versus 77.61 for CorrMatch, an observed margin of 0.88 points. On BUSI, FReMCI reaches 75.94, compared with 75.57 for the strongest per-task competitor CSC-PA and 74.62 for CorrMatch. The method is below CorrMatch on UDIAT and below ABD on Aryan, while it has the highest point estimate on Aneil, BUSI, and DDTI. These values are single-run point estimates; repeated-seed dispersion is not available in the current comparison.

C Complete Training Procedures

Algorithms 1 and 2 provide the complete optimization and reliable foreground memory-update procedures, respectively. In Algorithm 2, s is the local labeled/unlabeled source index, while C and S denote the queue capacity and per-image sampling limit for that procedure.

Algorithm 1 FReMCI Training

Require: D_l, D_u; student f_θ; EMA teacher $f_{\bar{\theta}}$; warm-up E_w; MCI start E_m; minimum queue length n_q
Ensure: Optimized student parameters θ

Initialize $\bar{\theta} \leftarrow \theta$ and $\mathcal{M}_l, \mathcal{M}_u \leftarrow \varnothing$
for each epoch e and paired mini-batch (x^l, y^l, x_w^u, x_s^u) **do**
  if $e < E_w$ **then**
    $(z^l, g^l) \leftarrow f_\theta(x^l; \varnothing)$
    Update θ with $L_{\sup}(z^l, y^l)$; update $\bar{\theta}$ by EMA
    $(\mathcal{M}_l, \mathcal{M}_u) \leftarrow \textsc{MemoryUpdate}(\mathcal{M}_l, \mathcal{M}_u; g^l, \varnothing, y^l, \varnothing, \varnothing)$
  else
    $\mathcal{Q} \leftarrow \{\mathcal{M}_s : e \geq E_m \wedge |\mathcal{M}_s| \geq n_q,\ s \in \{l, u\}\}$
    $(z^l, z^s, g^l, g^u) \leftarrow f_\theta([x^l, x_w^u]; \mathcal{Q})$
    Update θ with $L_{\sup}(z^l, y^l)$; update $\bar{\theta}$ by EMA
    $z^t \leftarrow f_{\bar{\theta}}(x_w^u; \varnothing)$
    $(\hat{y}^u, \rho) \leftarrow \textsc{ReliablePseudoLabel}(z^t, z^s)$
    Construct the uncertainty-masked view $\bar{x}^u$ from x_w^u
    $(\tilde{x}^l, \tilde{x}^u) \leftarrow \mathrm{PLAS}(x^l, x_s^u)$
    $(x^r, y^r, w^r) \leftarrow \textsc{ReSRI}(\tilde{x}^l, y^l, \tilde{x}^u, \hat{y}^u, \rho)$
    $\bar{z}^u \leftarrow [f_\theta([x^l, \bar{x}^u]; \mathcal{Q})]_u$; use the corresponding labeled output for $L_{\sup}$
    $z^r \leftarrow f_\theta(x^r; \varnothing)$
    $L \leftarrow L_{\sup} + \lambda_u(e) L_u(\bar{z}^u, \hat{y}^u, \rho)$
    $+\lambda_r L_{\mathrm{region}}(z^r, y^r, w^r)$
    Update θ with L; update $\bar{\theta}$ by EMA
    $(\mathcal{M}_l, \mathcal{M}_u) \leftarrow \textsc{MemoryUpdate}(\mathcal{M}_l, \mathcal{M}_u; g^l, g^u, y^l, \hat{y}^u, \rho)$
  end if
end for

Algorithm 2 Reliable Foreground Memory Update

Require: Decoder features g^l, g^u; $y^l, \hat{y}^u, \rho$; memories $\mathcal{M}_l, \mathcal{M}_u$
Require: Reliability threshold τ_m; capacity C; sample limit S; minimum foreground count $n_{\min}$
Ensure: Updated FIFO memories $\mathcal{M}_l, \mathcal{M}_u$

for $s \in \{l, u\}$ **do**
  if $g^s = \varnothing$ **then**
    continue
  else if $s = l$ **then**
    $B_s \leftarrow (y^l > 0) \wedge (y^l \neq \text{ignore})$
  else
    $B_s \leftarrow (\hat{y}^u > 0) \wedge (\rho \geq \tau_m)$
  end if
  Resize B_s to the spatial size of g^s by nearest neighbor
  for each image i in the mini-batch **do**
    $I \leftarrow \{p : B_{s,i}(p) = 1\}$
    if $|I| < n_{\min}$ **then**
      continue
    end if
    $I \leftarrow$ a random subset of I with at most S indices
    $V \leftarrow \{\mathrm{L2Norm}(g_i^s(p)) : p \in I\}$
    $\mathcal{M}_s \leftarrow \mathrm{Concat}(\mathcal{M}_s, \mathrm{StopGrad}(V))$
    if $|\mathcal{M}_s| > C$ **then**
      Retain only the most recent C entries in $\mathcal{M}_s$
    end if
  end for
end for
return $(\mathcal{M}_l, \mathcal{M}_u)$

D Implementation and Reproducibility Details

All images are resized to 256×256. FReMCI uses a ResNet-50 encoder and a DeepLabV3+ decoder, and is optimized by SGD with learning rate 0.003, momentum 0.9, weight decay 5×10^{-4}, and a polynomial schedule with power 0.9. We train for 200 epochs with 10 epochs of supervised warm-

up, a batch size of four per labeled and unlabeled stream, and an EMA decay of 0.99. The supervised objective is equally weighted CE and Dice. The unlabeled coefficient follows a 40-epoch sigmoid ramp to 1.0, and the region-loss coefficient is 1.0.

For PLAS, $r = 0.125$ and the two explicitly separated coefficients satisfy $\alpha_A = \alpha_x = 0.5$. ReSRI first zeros reliability below $\tau_0 = 0.35$ and uses $\tau = 0.45$ for reliable region construction; mask dilation is two pixels, box padding is 15%, and the seam band is four pixels wide. MCI starts after warm-up and uses two FIFO memories of size 1024, at most 32 sampled foreground features per image, top-$k = 16$, and temperature $T = 0.1$. The memory-update threshold is $\tau_m = 0.45$. An image updates a memory only when at least eight valid foreground pixels are available, and each direction is activated after its source memory contains at least 16 entries. The relevance-gate offset is $\delta = 0.1$, with $\omega_{l\to u} = 1.0$ and $\omega_{u\to l} = 0.3$. Unless a parameter is explicitly varied in Table 4, these manuscript settings define the reported method.

After warm-up, each paired mini-batch performs two optimizer and EMA updates: a supervised update from the paired labeled–weak-unlabeled forward pass, followed by a joint supervised, reliability-weighted unlabeled, and region-loss update. MCI is active in the paired and masked labeled–unlabeled student passes after its source queue is ready. The bidirectional ReSRI mixed batch is forwarded separately without memory. Thus, the method has additional training-time computation; training-time FLOPs, wall-clock time, and peak memory were not measured and are not claimed here. All comparison methods are evaluated at the same input resolution and label ratios.

During warm-up, the student learns from labeled targets and the teacher tracks the student without memory injection. After warm-up, reliable unlabeled features can enter the unlabeled memory, but only after passing the foreground and reliability tests in Algorithm 2. The minimum queue length delays retrieval until a direction has enough entries to form a meaningful candidate set. FIFO replacement bounds the memory state and makes the mechanism responsive to recent training features rather than an unbounded history.

The reliability threshold plays two related but distinct roles. In the pseudo-supervision branch, ρ is a continuous pixel weight, so moderately reliable pixels can contribute with reduced influence. In the foreground memory branch, the threshold defines which unlabeled pixels are eligible for feature storage. This distinction prevents a low-weight pseudo pixel from automatically becoming a long-lived memory entry. Labeled foreground features use the annotation mask directly, whereas unlabeled entries require both a foreground pseudo label and sufficient reliability.

MCI is restricted to the paired student path. The memory-free EMA teacher supplies $\hat{y}^u$ and ρ before reliable unlabeled features are admitted to the queue, so the pseudo-label computation is independent of the memory state. Labeled and reliable-unlabeled features remain in separate queues because their foreground masks have different sources of supervision. The current ablation does not separately test a memory-conditioned teacher.

E Feature Representation Analysis

Figure 4 visualizes all four encoder stages for three BUSI fold-1 test images under the $1/10$ labeled protocol: one benign case (`benign0188`) and two malignant cases (`malignant0153` and `malignant0162`). Ground truth is used only after encoder inference to identify foreground and background locations; it is never supplied to the encoder. For each image–stage pair, we sample equal numbers of background and lesion locations, up to 500 per class, and fit t-SNE independently after feature normalization and PCA. The later stages preserve coherent lesion/background organizations in several panels, although the degree of separation varies across images and is not strictly monotonic across stages.

Within each panel, a compact green cluster indicates that sampled lesion locations have similar projected features within that panel; overlap with orange points indicates that the corresponding stage still shares feature support between lesion and background. Later encoder stages often encode more task-specific context, but the spatial resolution decreases and the number of available samples changes. In addition, t-SNE preserves local neighborhoods while distorting global distances. Because each panel is fitted independently, absolute coordinates and apparent inter-cluster distances are not comparable across stages.

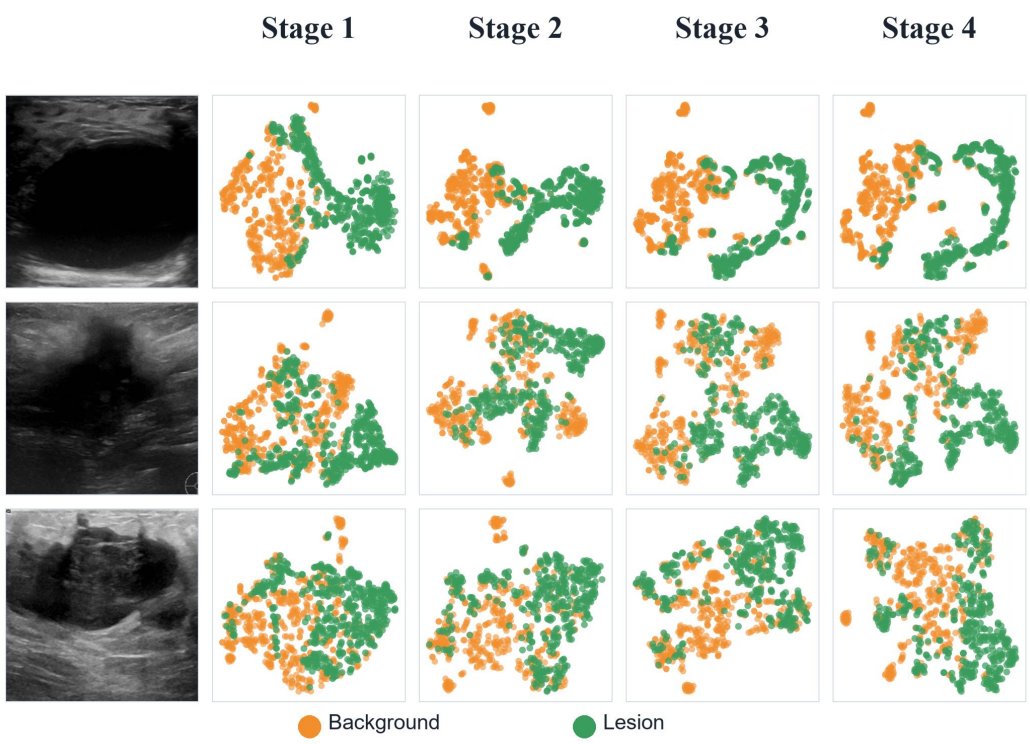


Figure 4: Stage-wise t-SNE visualization of encoder features from FReMCI on three BUSI fold-1 test images under the $1/10$ labeled protocol. The first column shows the input ultrasound images, followed by Encoder Stages 1–4. Orange and green denote background and lesion features, respectively. Each panel uses balanced sampling with up to 500 locations per class and an independent t-SNE fit; therefore, absolute coordinates are not comparable across panels.

F Extended Qualitative Comparisons

Figures 5 and 6 extend the main-paper comparison to the $3/10$ and $1/2$ protocols, using the method order and color convention of Figure 3. Each protocol shows one fixed case from each dataset.

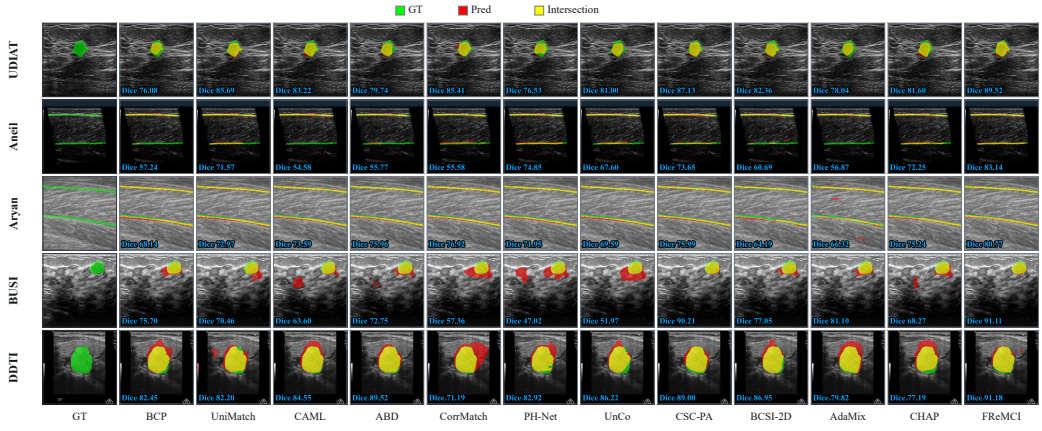


Figure 5: Qualitative comparison under the $3/10$ labeled protocol. Rows correspond to UDIAT, Aneil, Aryan, BUSI, and DDTI; columns show the ground-truth overlay, eleven representative baselines, and FReMCI in the same order and style as Figure 3.

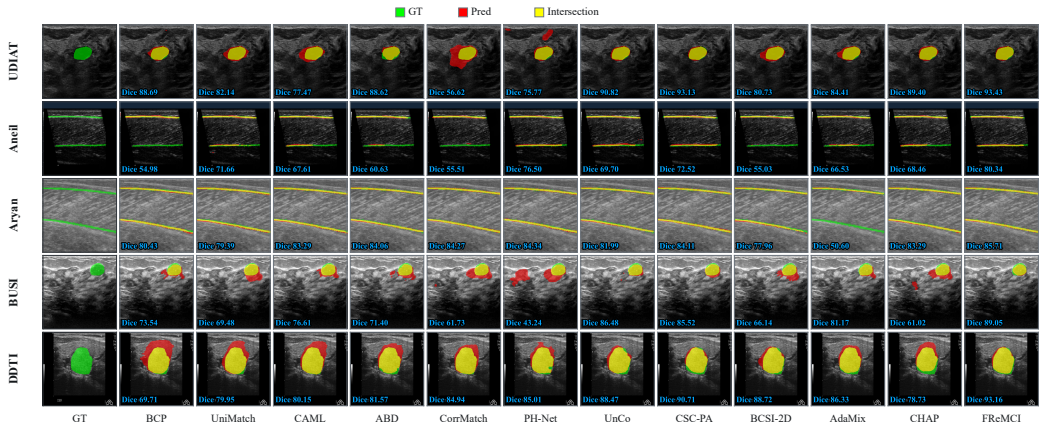


Figure 6: Qualitative comparison under the $1/2$ labeled protocol. Rows correspond to UDIAT, Aneil, Aryan, BUSI, and DDTI; columns show the ground-truth overlay, eleven representative baselines, and FReMCI in the same order and style as Figure 3.

Across both protocols, FReMCI shows compact foreground support and coherent boundaries for lesion, nodule, and musculoskeletal targets. Reduced red regions correspond to fewer false positives and reduced green regions to fewer missed foreground pixels. These fixed examples illustrate the spatial error patterns associated with the dataset-level results in Tables 2 and 3.

G Structural-Ablation Visualizations

Figure 7 visualizes the final foreground-probability distributions of the six structural variants for three BUSI test cases. The fixed selection rule chooses one case from each ground-truth lesion-area tertile, requires FReMCI to attain the best per-image Dice at or above its test-set 70th percentile, and requires every variant to exceed 75% Dice. The ordering of the selected cases follows the dataset-level trend in Table 4. Across the three cases, ReSRI reduces diffuse background responses, while PLAS and MCI affect the continuity and concentration of foreground support; the complete model provides the most coherent confidence distribution.

ReSRI changes where region supervision is applied by excluding unreliable pseudo foreground and artificial seams. PLAS modifies the appearance presented to the region branch, whereas MCI supplies foreground context accumulated across samples. Heatmap concentration must be considered together with the 0.5 contour and the displayed Dice, because a confident response can still be spatially inaccurate.

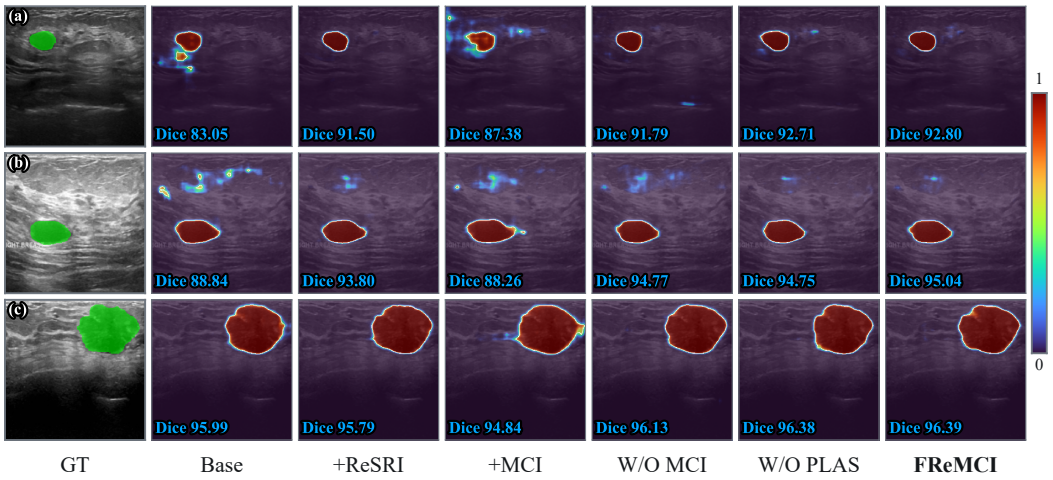


Figure 7: Structural-ablation foreground-probability heatmaps for three BUSI cases under the $1/10$ labeled protocol. Rows are ordered by increasing ground-truth lesion area. The first column is the green ground-truth overlay, followed by Base, +ReSRI, +MCI, W/O MCI, W/O PLAS, and FReMCI. All method panels share the same foreground softmax range from 0 to 1, shown by the vertical color bar at the right with endpoint labels 1 and 0; the white contour marks the prediction boundary at probability 0.5. Dice values are shown in bright blue at the lower-left of each method panel.

H Parameter Sensitivity

We study the low-frequency ratio r in PLAS and the retrieval size k in MCI on BUSI under the $1/10$ protocol. All other settings remain fixed. Numerical results are reported in Table 4(b), while Figure 8 visualizes the trends.

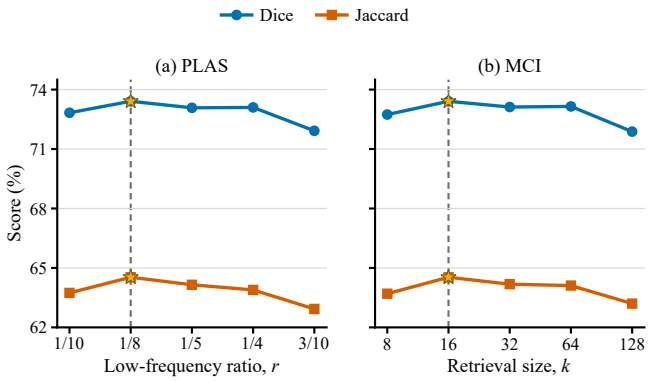


Figure 8: Sensitivity to the PLAS low-frequency ratio and the MCI memory retrieval size on BUSI under the $1/10$ protocol. Stars and dashed lines indicate the selected settings.

A very small low-frequency region provides limited style perturbation, whereas a large region modifies excessive spectral content and weakens geometry-preserving supervision. The best balance is obtained at $r = 0.125$. For MCI, retrieving only eight entries provides insufficient cross-sample context, while large retrieval sets increasingly include less relevant foreground features. The selected $k = 16$ gives the best Dice and Jaccard. Performance remains close around both selected values, indicating that FReMCI is not overly sensitive to moderate parameter changes.

Increasing r strengthens appearance perturbation but eventually includes frequency components that are less specific to global style. For MCI, a small k limits context diversity, whereas a large k reduces retrieval specificity. We use $r = 0.125$ and $k = 16$ throughout the benchmark; substantial changes in resolution, foreground size, or memory capacity may require retuning.

I Conceptual Motivation

Figure 9 connects the three ultrasound-specific difficulties identified in the introduction to the corresponding transfer controls. Cross-sample learning can supply useful supervision and foreground context, but appearance statistics, pseudo-labeled regions, and retrieved features do not have equal reliability.

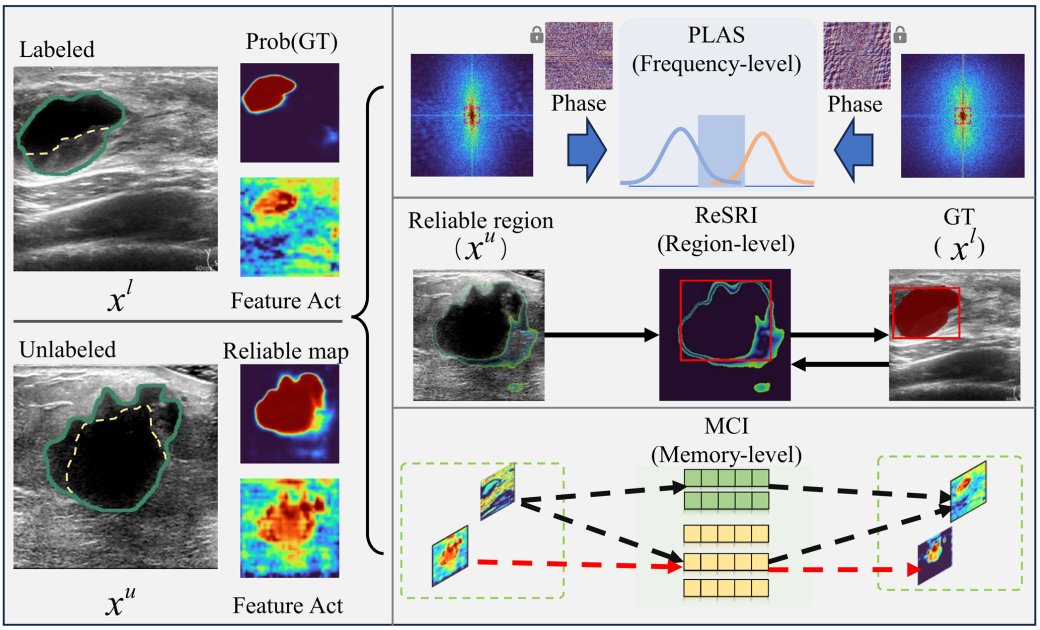


Figure 9: Conceptual motivation of FReMCI. Paired labeled and unlabeled ultrasound images can differ in low-frequency appearance, contain unreliable foreground regions, and exhibit weak or sparse foreground activations. FReMCI therefore controls cross-sample information transfer at three complementary levels: PLAS shares only low-frequency magnitude while preserving image-specific phase; ReSRI gates semantic region exchange with ground-truth or reliable unlabeled evidence; and MCI exchanges selected foreground context through labeled and reliable-unlabeled memories.

PLAS operates before region exchange and perturbs only the selected low-frequency amplitude, leaving phase and the remaining spectrum sample-specific. It exposes the region branch to partially shared appearance statistics without copying spatial geometry. ReSRI then uses teacher confidence, entropy, class margin, and teacher–student agreement to weight pseudo supervision and to select the unlabeled foreground eligible for region exchange. Its seam mask excludes the artificial paste boundary from the region loss.

MCI addresses the limited foreground evidence available in a mini-batch. It stores labeled and reliability-filtered unlabeled foreground features in separate memories and retrieves top-k entries from the opposite source group. The memory expands the candidate set across iterations, while reliability filtering and semantic retrieval restrict the transferred context.

These controls act at different points in the training path. PLAS cannot detect an incorrect pseudo label, ReSRI cannot supply missing cross-sample feature context, and MCI cannot remove a paste seam. Together they restrict appearance, spatial supervision, and feature transfer without treating the labeled and unlabeled streams as interchangeable. The three uncertainty sources may also interact:

appearance variation can destabilize pseudo labels, which can then affect region intervention and memory updates.

J LIMITATIONS

The current study is restricted to binary 2-D ultrasound segmentation, and the reported comparisons are point estimates without repeated-seed confidence intervals. The five-task average also gives separate weight to two tasks derived from one FALLMUD source. Consequently, the present results do not establish robustness to random initialization, multi-class anatomy, scanner shifts, or volumetric acquisition. The composite reliability score is used as a conservative training weight, but the present study does not provide calibration curves or term-wise ablations, and the structural ablation does not isolate seam exclusion, each injection direction, or a memory-conditioned teacher. FReMCI also adds FFT, mixed-region, and memory-retrieval operations during training; their wall-clock and memory overhead remains to be quantified. Extending FReMCI to 3-D ultrasound, MRI, or CT is not a direct replacement of the current 2-D operations. In particular, PLAS assumes that a compact low-frequency region primarily represents appearance variation while phase preserves spatial structure. This interpretation may be less reliable in 3-D MRI and CT, where low-frequency coefficients can encode volumetric anatomy, scanner-dependent intensity fields, reconstruction kernels, and acquisition artifacts. Exchanging a fixed cuboidal low-frequency region could therefore alter clinically relevant structures rather than only imaging style. A volumetric extension would require modality-aware frequency selection and an explicit assessment of anatomical preservation. ReSRI would likewise require 3-D connected regions and seam suppression across adjacent slices, while MCI would incur substantially higher memory and retrieval costs for dense volumetric features. Slice-wise processing could reduce this cost but would not model through-plane continuity. These issues should be evaluated with modality-specific 3-D benchmarks, repeated runs, and cross-scanner testing before extending the conclusions beyond the 2-D ultrasound setting studied here.