HyWaveDiff: Reliability-Guided Diffusion for Semi-Supervised Medical Image Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Reliable pseudo-labeling remains a central challenge in semi-supervised medical image segmentation, particularly near low-contrast and irregular anatomical boundaries. We propose **HyWaveDiff**, a reliability-guided mask-diffusion framework that converts stochastic diffusion variation into actionable supervision for unlabeled data. An EMA teacher generates multiple predictions and feature embeddings, from which Multi-Cue Reliability Refinement (MRR) integrates prediction confidence, stochastic consistency, geometry-dependent feature dispersion, and image-derived boundary evidence. The resulting reliability is used to correct soft pseudo-labels and assign continuous pixel-wise supervision weights. A Poincaré-ball dispersion models feature disagreement, while multi-scale Haar-wavelet guidance supplies frequency-specific structural information to the denoiser. Reliability-adaptive optimization and frequency-oriented consistency further regulate unlabeled learning. With only three labeled training cases per dataset, HyWaveDiff achieves Dice scores of 82.38% on PROMISE12 MRI and 88.33% on Task09 Spleen CT, obtaining the highest observed Dice among the evaluated semi-supervised methods. On Task09 Spleen, it also achieves the lowest HD95 and ASD among the evaluated semi-supervised methods. These results support reliability-guided use of stochastic diffusion predictions for low-annotation medical image segmentation.

1 Introduction

Medical image segmentation requires detailed annotations that are costly to obtain. Semi-supervised medical image segmentation addresses this limitation by learning from a small labeled set and additional unlabeled images. Teacher–student approaches generate training targets using an exponential moving average (EMA) teacher (Tarvainen & Valpola, 2017), while uncertainty-aware methods reduce the influence of unreliable predictions (Yu et al., 2019). However, erroneous pseudo-labels near low-contrast or irregular boundaries can still be reinforced during student training.

Image-conditioned diffusion models provide multiple segmentation predictions through stochastic denoising (Wolleb et al., 2022; Wu et al., 2024). Their sample variability offers information beyond the ensemble mean, but using this information requires a reliability mechanism that connects disagreement to training targets and supervision weights. Prediction confidence, feature disagreement, and image-derived boundary responses describe different aspects of this problem.

We construct pseudo-label reliability from stochastic feature dispersion and image-derived structural cues. Teacher feature embeddings are mapped to a Poincaré ball, and their dispersion is measured around a center obtained by mapping the tangent-space mean. This provides a geometry-dependent disagreement signal motivated by hyperbolic representation learning (Ganea et al., 2018; Atigh et al., 2022). Haar-wavelet responses additionally expose coarse appearance and directional intensity transitions (Liu et al., 2018). The usefulness of this construction is assessed through component ablations and comparisons with alternative uncertainty estimators.

We propose **HyWaveDiff**, an image-conditioned mask-diffusion framework built on a VMamba U-Net (Liu et al., 2024b). A multi-scale wavelet structural guidance mechanism injects coarse anatomical structure and directional boundary cues into the denoising hierarchy. For unlabeled images, Multi-Cue Reliability Refinement (MRR) combines teacher confidence, stochastic prediction

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.

consistency, hyperbolic feature dispersion, and boundary responses to produce refined soft pseudo-labels and continuous pixel-wise weights. Frequency-oriented consistency (FOC) supervises mask structure, while reliability-aware semi-supervised learning (RA-SSL) regulates the overall unlabeled contribution. Together, these cues define a unified reliability pipeline for refining pseudo-labels and weighting unlabeled supervision.

Experiments on PROMISE12 MRI (Litjens et al., 2014) and Task09 Spleen CT (Antonelli et al., 2022) evaluate the framework with three labeled training cases per dataset. HyWaveDiff achieves Dice scores of 82.38% and 88.33%, respectively, leading the evaluated semi-supervised methods on this metric. Our contributions are summarized as follows:

- We introduce **HyWaveDiff**, a reliability-guided diffusion framework that unifies stochastic prediction consistency, geometry-dependent feature dispersion, and image-derived boundary evidence into pseudo-label reliability estimation, enabling both soft-label correction and continuous pixel-wise supervision.
- We develop **Multi-Cue Reliability Refinement (MRR)**, a concrete reliability-fusion mechanism that jointly refines pseudo-labels and estimates their supervision weights. We further evaluate hyperbolic dispersion against Euclidean dispersion in fixed-teacher diagnostics, and against Euclidean dispersion and predictive entropy in downstream estimator-replacement experiments.
- We validate HyWaveDiff in a challenging three-labeled-case setting on PROMISE12 MRI and Task09 Spleen CT. Component ablations and estimator comparisons examine the contributions of the major reliability and structural design choices.

2 Related Work

Semi-supervised medical image segmentation. Mean Teacher and uncertainty-aware teacher–student learning use consistency and reliability to exploit unlabeled data (Tarvainen & Valpola, 2017; Yu et al., 2019). BCP, ABD, and β-FFT improve learning through copy-paste, confidence-guided displacement, and frequency-domain interpolation, respectively (Bai et al., 2023; Chi et al., 2024; Hu et al., 2025). Semi-Mamba-UNet and DCMamba investigate state-space representations and complementary training signals (Ma & Wang, 2024; Li et al., 2026). Our focus is the construction of soft targets and reliability weights from stochastic diffusion predictions and image structure.

Diffusion-based segmentation. Stochastic diffusion sampling supports segmentation ensembles and image-conditioned mask generation (Wolleb et al., 2022; Wu et al., 2024). DiffRect uses diffusion for pseudo-label rectification, while MCSD uses diffusion features and consistency learning for segmentation (Liu et al., 2024a; Chen et al., 2025). HyWaveDiff uses both the predictions and associated feature embeddings of a diffusion teacher to guide unlabeled supervision.

Hyperbolic and frequency-domain modeling. Hyperbolic neural networks and image segmentation extend representation learning to negatively curved spaces (Ganea et al., 2018; Atigh et al., 2022). Fréchet statistics characterize manifold-valued observations (Pennec, 2006), and Co-Manifold Learning combines Euclidean and hyperbolic representations for semi-supervised segmentation (Peiris et al., 2025). In the frequency domain, MWCNN incorporates wavelets into multi-scale processing, and FRCNet introduces frequency and regional consistency (Liu et al., 2018; He et al., 2024). HyWaveDiff instead focuses on estimating pseudo-label reliability from stochastic diffusion behavior and image structure, and uses this reliability to jointly refine training targets and weight unlabeled supervision.

3 Method

3.1 Problem Formulation and Overview

Let $\mathcal{D}_l = \{(I_i, Y_i)\}_{i=1}^{N_l}$ and $\mathcal{D}_u = \{I_j\}_{j=1}^{N_u}$ denote the labeled and unlabeled training sets, respectively, where $N_l \ll N_u$. We represent a segmentation mask in the continuous diffusion space as

$$X_0 = 2\ \mathrm{onehot}(Y) - 1, \qquad X_0 \in \mathbb{R}^{C\times H\times W}, \tag{1}$$

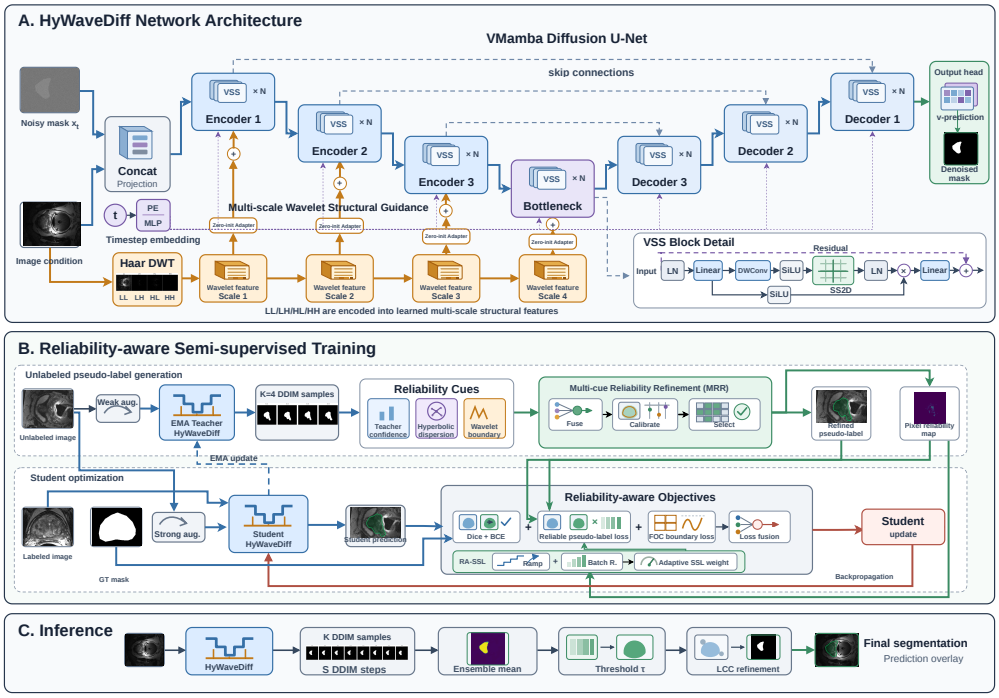


Figure 1: Overview of HyWaveDiff. (a) An image-conditioned, v-parameterized VMamba diffusion U-Net receives the image and a noisy mask. Multi-scale Haar-wavelet features enter the denoiser through zero-initialized adapters. (b) The EMA teacher generates K stochastic DDIM predictions. MRR combines prediction- and feature-based reliability cues with image-derived boundary responses to obtain refined soft pseudo-labels and continuous pixel-wise weights. RA-SSL regulates their contribution to student optimization. The selection operation denotes soft reliability weighting rather than hard pixel rejection. (c) Inference averages stochastic predictions, applies a foreground threshold, and retains the largest connected foreground component after three-dimensional reconstruction.

where C is the number of classes and $H \times W$ is the spatial resolution. Both benchmarks considered in this work involve binary foreground–background segmentation.

As illustrated in Figure 1, HyWaveDiff treats stochastic diffusion predictions as reliability evidence for unlabeled learning. An EMA teacher generates multiple segmentation predictions and their associated feature embeddings. MRR converts prediction confidence, stochastic consistency, geometry-dependent feature dispersion, and image-derived boundary evidence into refined soft pseudo-labels and continuous pixel-wise reliability weights. A multi-scale wavelet structural guidance branch supplies frequency-specific anatomical information to the diffusion denoiser, while RA-SSL regulates the overall contribution of unlabeled supervision.

3.2 Conditional Mask Diffusion and Wavelet Guidance

Mask diffusion and v-prediction. We apply diffusion to the mask representation while retaining the image I as the conditioning input. Given the cumulative noise schedule $\bar{\alpha}_t$, the noisy mask, prediction target, and reconstructed clean mask are

$$\begin{aligned} X_t &= \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \\ v &= \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t} X_0, \\ \hat{X}_0 &= \sqrt{\bar{\alpha}_t} X_t - \sqrt{1-\bar{\alpha}_t}\, v_\theta(I, X_t, t). \end{aligned} \tag{2}$$

We use v-prediction following Salimans & Ho (2022) and DDIM reverse sampling (Song et al., 2021).

The denoiser follows a U-shaped encoder–decoder architecture built from visual state-space blocks (Liu et al., 2024b). It concatenates the image condition with the noisy mask, incorporates the diffusion timestep embedding, and uses multi-directional selective scans to capture long-range spatial dependencies.

Multi-scale Wavelet Structural Guidance. To expose anatomical structure at different frequency bands, we decompose the input image with a two-dimensional Haar transform into one low-frequency and three high-frequency subbands:

$$\{I_{LL}, I_{LH}, I_{HL}, I_{HH}\} = \mathrm{DWT}(I). \tag{3}$$

The LL subband preserves coarse anatomical structure, whereas LH, HL, and HH emphasize directional intensity transitions and boundary details. We jointly encode these subbands into a multi-scale structural feature pyramid $\{C_l^{\mathrm{wav}}\}_{l=1}^{L}$ that is aligned with the encoder and bottleneck representations. At scale l, the structural feature is injected as

$$\widetilde{F}_l = F_l + \mathcal{Z}_l(C_l^{\mathrm{wav}}), \tag{4}$$

where $\mathcal{Z}_l$ denotes a residual adapter that is zero-initialized following Zhang et al. (2023). Consequently, the guidance branch initially contributes no residual and progressively learns frequency-specific structural corrections during training.

3.3 Hyperbolic Reliability and Multi-Cue Refinement

Stochastic EMA teacher. For a weakly augmented unlabeled image I_w^u, the EMA teacher generates K predictions using independently initialized DDIM trajectories:

$$(P_k^u, E_k^u) = T_{\bar{\theta}}(I_w^u; \epsilon_k), \qquad k = 1, \ldots, K, \tag{5}$$

where P_k^u denotes the predicted class probabilities, E_k^u denotes the associated feature embeddings, and ϵ_k is an independent Gaussian initialization. We compute the ensemble probability $\bar{P}^u = K^{-1}\sum_{k=1}^{K} P_k^u$ and normalize it across the class dimension.

Approximate hyperbolic dispersion. We estimate feature dispersion in a Poincaré ball with curvature $-c$, where $c > 0$ and $c = 1$ in our experiments. Instead of iteratively computing an exact Fréchet mean, we construct an approximate center by averaging the embeddings in the tangent space and mapping the result to the ball. Specifically, with $\bar{E}^u = K^{-1}\sum_{k=1}^{K} E_k^u$,

$$Z_k^u = \mathrm{Exp}_0^c(E_k^u), \qquad \mu_H^u = \mathrm{Exp}_0^c(\bar{E}^u), \tag{6}$$

where

$$\mathrm{Exp}_0^c(v) = \tanh(\sqrt{c}\|v\|)\frac{v}{\sqrt{c}\|v\|}, \tag{7}$$

with the continuous extension $\mathrm{Exp}_0^c(0) = 0$. For two points inside the ball, the distance is

$$d_c(z, \mu) = \frac{1}{\sqrt{c}}\,\mathrm{arcosh}\left(1 + \frac{2c\|z-\mu\|^2}{(1-c\|z\|^2)(1-c\|\mu\|^2)}\right). \tag{8}$$

At each spatial location x, we compute the mean squared distance to the approximate center and normalize the result within each image:

$$\begin{aligned} U_{H,\mathrm{raw}}^u(x) &= \frac{1}{K}\sum_{k=1}^{K} d_c^2(Z_k^u(x), \mu_H^u(x)), \\ U_H^u &= \mathcal{N}_{0.95}(U_{H,\mathrm{raw}}^u), \\ \mathcal{N}_{0.95}(X) &= \mathrm{clip}\left(\frac{X}{Q_{0.95}(X)+\varepsilon}, 0, 1\right), \end{aligned} \tag{9}$$

where $Q_{0.95}(X)$ is the 95th percentile over spatial locations and $\varepsilon > 0$ is a numerical stabilizer. The resulting map measures stochastic feature disagreement and serves as a reliability cue for pseudo-label refinement.

Table 1: Roles of the cues in pseudo-label refinement and weighting. These roles describe the formulation and do not establish statistical independence between cues.

Cue	Quantity	Role
Prediction confidence	R_P^u, C^u	Confidence factors in reliability weighting.
Stochastic consistency	V_K^u	Controls R_N^u through probability variance.
Hyperbolic dispersion	U_H^u	Modulates Haar calibration and reliability weights.
Haar response	E_W^u	Determines the direction of the first boundary correction.
Sobel response	E_S^u	Corrects the soft target within B_R^u.

Geometric rationale. The distance above depends on both feature separation and position within the ball. For a fixed coordinate displacement, the factors $1 - c\|z\|^2$ and $1 - c\|\mu\|^2$ increase the measured distance as points approach the boundary (Ganea et al., 2018). This motivates testing whether geometry-dependent dispersion provides a useful reliability cue for stochastic teacher features. Near the origin, $d_c(z, \mu) \approx 2\|z - \mu\|$, so hyperbolic and Euclidean dispersion can induce similar rankings. We therefore evaluate the estimator through both segmentation after training and fixed-teacher diagnostics, without assuming that hyperbolic distance universally ranks pseudo-label errors better.

Multi-Cue Reliability Refinement. MRR converts stochastic teacher predictions into a refined soft pseudo-label and a continuous reliability map. Starting from $p_0 = \bar{P}_{\mathrm{fg}}^u$, it applies Haar- and Sobel-guided boundary corrections:

$$\begin{aligned} p_1 &= \mathrm{clip}\left[p_0 + 0.20 B_f^u G_U^u (G_W^u - p_0),\, 0,\, 1\right], \\ \widetilde{p}^u &= \mathrm{clip}\left[p_1 + 0.20 B_R^u (G_S^u - p_1),\, 0,\, 1\right]. \end{aligned} \tag{10}$$

Here, B_f^u and B_R^u are the prediction-derived boundary bands before and after the first calibration. The sigmoid gates G_U^u, G_W^u, and G_S^u are computed from hyperbolic dispersion, the Haar response, and the Sobel response, respectively. The first update combines feature disagreement with wavelet boundary information; the second incorporates Sobel evidence into the calibrated target. Both updates retain continuous foreground probabilities. Gate definitions and boundary-band construction are given in Appendix B.

We combine calibrated confidence, stochastic consistency, and hyperbolic reliability:

$$\begin{aligned} R_P^u &= 2|p_1 - 0.5|, & R_N^u &= \exp(-V_K^u / 0.03), \\ R_H^u &= \exp[-\mathcal{N}_{\mathrm{mm}}(U_H^u)/0.20], & W_H^u &= \mathrm{clip}(1 - U_H^u, 0.10, 1), \\ R^u &= \mathrm{clip}(R_P^u R_N^u R_H^u W_H^u, 0.10, 1). \end{aligned} \tag{11}$$

Here, V_K^u is the variance of the K foreground probability predictions, and $\mathcal{N}_{\mathrm{mm}}$ denotes image-wise min–max normalization. The expression for R^u is algebraically equivalent to the implemented weighting rule under the shared lower bound of 0.10 and $W_H^u \le 1$.

A subsequent Sobel-guided boundary correction produces $\widetilde{p}^u$, giving the soft target $\widetilde{P}^u = [1 - \widetilde{p}^u, \widetilde{p}^u]$. Let $\hat{y}^u(x) = \arg\max_c \widetilde{P}_c^u(x)$ and $C^u(x) = \max_c \widetilde{P}_c^u(x)$. The final reliability is

$$W^u(x) = \mathrm{clip}\left[R^u(x)\sigma\left(\frac{C^u(x) - \tau_{\hat{y}^u(x)}}{0.05}\right) M_{\hat{y}^u(x)},\, 0.10,\, 1\right], \tag{12}$$

where $(\tau_0, \tau_1) = (0.90, 0.80)$ and $(M_0, M_1) = (0.25, 1.00)$ for background and foreground. The confidence thresholds define smooth gates. The positive weight floor retains supervision at low-reliability locations, so selection corresponds to continuous weighting rather than hard rejection.

Roles of the reliability cues. The refinement follows $p_0 \to p_1 \to \widetilde{p}^u$, while the reliability branch produces W^u. Table 1 distinguishes cues that modify the soft target from those that control its supervision weight. In particular, the Sobel boundary band affects probability correction; it does not provide an additional reliability floor beyond the final clipping operation.

3.4 Reliability-Aware Objective and Optimization

Supervised objective. For labeled samples, the training objective combines diffusion prediction, segmentation supervision, and frequency-oriented consistency:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_v + \mathcal{L}_{\text{rec}} + 0.5\mathcal{L}_{\text{aux}} + 0.05\mathcal{L}_{\text{con}} + 0.05\mathcal{L}^s_{\text{FOC}}. \tag{13}$$

Here, $\mathcal{L}_v$ supervises the v-prediction target, $\mathcal{L}_{\text{rec}}$ supervises the reconstructed mask, and $\mathcal{L}_{\text{aux}}$ and $\mathcal{L}_{\text{con}}$ denote the auxiliary segmentation and consistency terms, respectively. The term $\mathcal{L}^s_{\text{FOC}}$ applies frequency-oriented consistency to labeled predictions.

Frequency-oriented consistency. The frequency-oriented consistency (FOC) loss compares low-frequency mask structure and high-frequency response magnitudes. For a prediction P and target Y,

$$\begin{aligned} D_L(P,Y) &= \|\mathcal{W}_{LL}(P) - \mathcal{W}_{LL}(Y)\|_1, \\ D_H(P,Y) &= \sum_{b\in\{LH,HL,HH\}} \big\|\,|\mathcal{W}_b(P)| - |\mathcal{W}_b(Y)|\,\big\|_1, \\ \mathcal{L}_{\text{FOC}}(P,Y) &= 0.5D_L(P,Y) + 2.0D_H(P,Y). \end{aligned} \tag{14}$$

This objective supplies supervision on both coarse structure and directional high-frequency content.

Unlabeled objective. The student predicts P^u from a strongly augmented unlabeled image. We spatially align the teacher's refined pseudo-label and reliability map with the student prediction, retaining the notation $(\widetilde{P}^u, W^u)$ for the aligned quantities. The unlabeled objective is

$$\begin{aligned} \mathcal{L}_{\text{pseudo}} &= \mathcal{L}_{\text{wDice}}(P^u, \widetilde{P}^u; W^u) + \mathcal{L}_{\text{wBCE}}(P^u, \widetilde{P}^u; W^u), \\ \mathcal{L}_{\text{unsup}} &= \mathcal{L}_{\text{pseudo}} + 0.02\mathcal{L}_{\text{FOC}}(P^u, \widetilde{P}^u), \end{aligned} \tag{15}$$

where $\mathcal{L}_{\text{wDice}}$ and $\mathcal{L}_{\text{wBCE}}$ denote reliability-weighted Dice and binary cross-entropy losses.

Reliability-aware loss scheduling. RA-SSL regulates the overall unlabeled contribution using a Gaussian ramp-up and an EMA-smoothed reliability estimate. Let n denote the optimization iteration, distinct from the diffusion timestep t. The spatial mean reliability of each unlabeled image is

$$r_n = \frac{1}{HW}\sum_x W^u_n(x), \qquad \bar{r}_n = 0.98\bar{r}_{n-1} + 0.02r_n, \tag{16}$$

where the image-wise means are averaged over the unlabeled mini-batch before updating $\bar{r}_n$.

For s epochs elapsed since SSL activation, we define

$$\begin{aligned} q_s &= \text{clip}(s/100, 0, 1), \\ \lambda_{\text{sched}}(s) &= 0.30\exp[-5(1-q_s)^2], \\ G_{\text{RA}}(n) &= 0.50 + 0.50\sigma\left(\frac{\bar{r}_n - 0.16}{0.04}\right), \\ \lambda^{\text{eff}}_u(s,n) &= \lambda_{\text{sched}}(s)G_{\text{RA}}(n), \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{sup}} + \lambda^{\text{eff}}_u(s,n)\mathcal{L}_{\text{unsup}}. \end{aligned} \tag{17}$$

The scheduled coefficient gradually introduces unlabeled supervision, while the reliability gate adjusts its strength within a bounded range. Pixel-wise weighting and loss-level scheduling thus act at complementary levels of the training objective.

Training and inference. Training proceeds from supervised warm-up to semi-supervised optimization with an EMA teacher. Sampling and evaluation settings are summarized in Section 4.1, with further details in Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and evaluation. We evaluate on PROMISE12 prostate MRI (Litjens et al., 2014) and Task09 Spleen CT (Antonelli et al., 2022). The patient-level training/validation/test splits are 36/4/10 and 30/4/7, respectively. Only three training cases or volumes are labeled, leaving 33 and 27 unlabeled. All compared semi-supervised baselines were rerun under the same patient-level partitions and evaluation protocol used for HyWaveDiff.

Dice, Jaccard, HD95, and ASD are computed after three-dimensional reconstruction of slice-wise predictions. Surface distances use physical voxel spacing adjusted for preprocessing and are reported in millimeters.

Implementation. HyWaveDiff uses 224×224 inputs and two-stage training: 180 supervised warm-up epochs followed by 150 semi-supervised epochs, with unlabeled supervision activated from epoch 6. The EMA teacher uses $K = 4$ stochastic predictions and five DDIM steps during semi-supervised training. Final inference averages $K = 8$ predictions with 20 DDIM steps, applies a foreground threshold of 0.40, and retains the largest connected foreground component after three-dimensional reconstruction. Checkpoints are selected using validation Dice. Optimizer settings, preprocessing, and Stage-1 validation details are given in Appendix C.

4.2 COMPARISON WITH EXISTING METHODS

We compare HyWaveDiff with β-FFT (Hu et al., 2025), ABD (Chi et al., 2024), BCP (Bai et al., 2023), Semi-Mamba-UNet (Ma & Wang, 2024), DiffRect (Liu et al., 2024a), and MCSD-full (Chen et al., 2025).

Table 2 summarizes the quantitative results on the held-out test sets. The L/U column reports the numbers of labeled and unlabeled training cases or volumes. SupOnly uses only the three labeled training cases or volumes, whereas full supervision uses annotations for the complete training pool. These settings serve as supervision references and are excluded from the ranking of semi-supervised methods.

Results on PROMISE12. HyWaveDiff achieves the highest Dice (0.8238) among the evaluated semi-supervised methods. Its Jaccard (0.7022) is close to that of β-FFT (0.7059). Relative to β-FFT, the Dice gain is 0.23 percentage points, while the Jaccard values remain close under the case-wise evaluation.

Compared with SupOnly, incorporating the 33 unlabeled cases increases Dice from 0.5577 to 0.8238 and Jaccard from 0.3912 to 0.7022.

These results indicate improved foreground overlap under the evaluated three-labeled-case setting. Surface-distance rankings differ: BCP achieves the lowest HD95 and β-FFT the lowest ASD.

Results on Task09 Spleen. HyWaveDiff achieves the highest Dice (0.8833), together with the lowest HD95 (15.2580 mm) and ASD (2.9786 mm) among the evaluated semi-supervised methods. Compared with β-FFT, Dice increases by 0.08 percentage points, while HD95 and ASD decrease by 6.7578 mm and 3.1050 mm, respectively. Using the 27 unlabeled volumes also increases Dice from the SupOnly reference of 0.8639 to 0.8833. Thus, the evaluated Spleen setting combines competitive region overlap with improved surface distances relative to the compared semi-supervised methods.

4.3 ABLATION AND ESTIMATOR COMPARISON

Table 3 summarizes component removal and uncertainty-estimator comparisons on PROMISE12. The full model uses hyperbolic dispersion. The no-dispersion setting is identical to removing hyperbolic dispersion and is therefore reported once.

Component contributions. The complete framework achieves the best reported mean across all four metrics in Table 3. Removing hyperbolic dispersion reduces Dice from 0.8238 to 0.7237, while

Table 2: Quantitative comparison on the held-out test sets. L/U denotes the numbers of labeled and unlabeled training cases or volumes. Entries show means and standard deviations. Bold and underlined values indicate the best and second-best means among semi-supervised methods, respectively. SupOnly and fully supervised models are supervision references and are excluded from ranking. HD95 and ASD are reported in millimeters.

Method	L/U	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
PROMISE12 MRI					
SupOnly	3/0	0.5577 ± 0.0821	0.3912 ± 0.0789	14.6601 ± 1.8015	5.8998 ± 0.8824
β-FFT	3/33	$\underline{0.8215 \pm 0.0852}$	$\mathbf{0.7059 \pm 0.1227}$	$\underline{5.0496 \pm 3.3318}$	$\mathbf{1.3912 \pm 0.5661}$
ABD	3/33	0.8153 ± 0.0948	0.6969 ± 0.1211	7.6530 ± 9.8224	2.8559 ± 3.3992
BCP	3/33	0.7959 ± 0.1090	0.6733 ± 0.1373	$\mathbf{3.9011 \pm 2.1867}$	2.0241 ± 1.5936
Semi-Mamba-UNet	3/33	0.7269 ± 0.1344	0.5866 ± 0.1481	6.0947 ± 4.3881	$\underline{2.0171 \pm 0.9835}$
DiffRect	3/33	0.7289 ± 0.1105	0.5846 ± 0.1287	9.5692 ± 15.3828	2.8752 ± 2.2795
MCSD-full	3/33	0.7875 ± 0.0686	0.6546 ± 0.0891	7.4111 ± 2.2769	2.2100 ± 0.7607
HyWaveDiff	3/33	$\mathbf{0.8238 \pm 0.0414}$	$\underline{0.7022 \pm 0.0597}$	7.9813 ± 4.6437	2.3776 ± 0.8502
Full supervision	36/0	0.8680 ± 0.0035	0.7668 ± 0.0055	5.8323 ± 1.1747	1.6474 ± 0.1601
Task09 Spleen CT					
SupOnly	3/0	0.8639 ± 0.0388	0.7625 ± 0.0601	14.0182 ± 9.3935	2.9577 ± 1.7722
β-FFT	3/27	$\underline{0.8825 \pm 0.0500}$	$\mathbf{0.7933 \pm 0.0801}$	22.0158 ± 32.5974	6.0836 ± 8.2070
ABD	3/27	0.8491 ± 0.0336	0.7393 ± 0.0500	60.5288 ± 43.7981	17.0348 ± 13.5561
BCP	3/27	0.7861 ± 0.1300	0.6640 ± 0.1539	51.4537 ± 31.9207	13.2961 ± 7.6919
Semi-Mamba-UNet	3/27	0.8778 ± 0.0793	0.7903 ± 0.1143	$\underline{16.5778 \pm 27.8570}$	$\underline{3.2104 \pm 4.0457}$
DiffRect	3/27	0.7461 ± 0.2291	0.6483 ± 0.2914	86.6765 ± 19.4203	30.8917 ± 7.9154
MCSD-full	3/27	0.7360 ± 0.1108	0.5933 ± 0.1265	22.3704 ± 10.8881	5.0794 ± 2.8534
HyWaveDiff	3/27	$\mathbf{0.8833 \pm 0.0377}$	$\underline{0.7931 \pm 0.0616}$	$\mathbf{15.2580 \pm 10.1959}$	$\mathbf{2.9786 \pm 2.0502}$
Full supervision	30/0	0.9342 ± 0.0175	0.8770 ± 0.0310	5.2465 ± 2.4962	1.1236 ± 0.3849

Table 3: Component ablation and uncertainty-estimator comparison on PROMISE12 with three labeled and 33 unlabeled cases. "w/o" denotes removal from the complete framework. The final two rows replace the uncertainty estimator. Entries show means and standard deviations. Bold indicates the best mean in each column. HD95 and ASD are in millimeters.

Setting	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
HyWaveDiff (full)	$\mathbf{0.8238 \pm 0.0414}$	$\mathbf{0.7022 \pm 0.0597}$	$\mathbf{7.9813 \pm 4.6437}$	$\mathbf{2.3776 \pm 0.8502}$
w/o RA-SSL	0.8037 ± 0.0489	0.6746 ± 0.0683	11.6158 ± 8.0210	2.8628 ± 1.3271
w/o hyperbolic disp.	0.7237 ± 0.2600	0.6082 ± 0.2268	11.5409 ± 9.3030	4.4818 ± 4.8361
w/o MRR	0.7818 ± 0.0533	0.6449 ± 0.0718	10.9054 ± 7.0048	3.1525 ± 1.4083
w/o FOC	0.7935 ± 0.0489	0.6604 ± 0.0672	8.4951 ± 3.3575	2.4904 ± 0.6733
w/o wavelet guidance	0.7198 ± 0.2554	0.6010 ± 0.2170	9.7970 ± 6.8915	3.9563 ± 4.3489
Euclidean feature	0.7368 ± 0.2620	0.6255 ± 0.2273	12.6721 ± 8.5778	4.3166 ± 5.0003
Predictive entropy	0.7079 ± 0.2578	0.5880 ± 0.2255	10.9856 ± 7.5232	4.4544 ± 5.0183

removing wavelet structural guidance reduces it to 0.7198. These decreases of 10.01 and 10.40 percentage points support retaining both components in the evaluated framework. Removing MRR reduces Dice by 4.20 percentage points, consistent with the benefit of refining teacher targets and their supervision weights. Removing RA-SSL or FOC reduces Dice by 2.01 or 3.03 percentage points, respectively. These comparisons assess the contributions of target refinement, loss scheduling, and frequency-oriented supervision within the full training pipeline.

Comparison of uncertainty estimators. The hyperbolic configuration obtains Dice 0.8238, compared with 0.7368 for Euclidean feature dispersion and 0.7079 for predictive entropy. The corresponding Dice differences are 8.70 and 11.59 percentage points. It also achieves higher Jaccard and lower HD95 and ASD than these alternatives. Relative to Euclidean dispersion, HD95 decreases from 12.6721 to 7.9813 mm and ASD from 4.3166 to 2.3776 mm. These replacement experiments show that the choice of dispersion estimator materially affects downstream segmentation performance when embedded in the complete reliability-guided training pipeline. They do not, however,

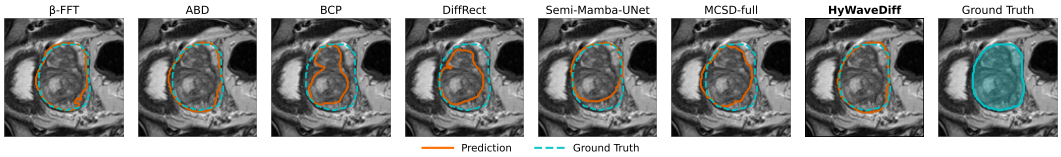
establish that hyperbolic dispersion is intrinsically superior for static uncertainty ranking. The no-dispersion result is the component-removal entry in the same table and is not counted as a separate experiment.

4.4 Qualitative Analysis

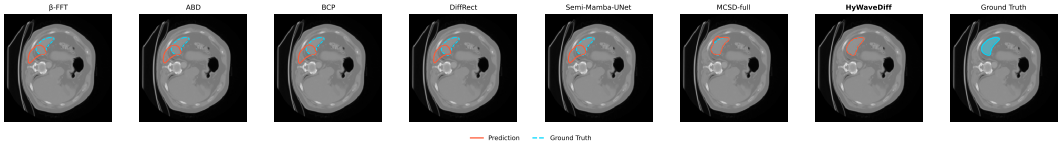
Figure 2 compares segmentation predictions on PROMISE12 and Task09 Spleen. Each row shows the same image slice for all methods, allowing direct inspection of the predicted contours against the reference annotation.

In the displayed PROMISE12 example, several competing methods exhibit local under-segmentation or deviations around the irregular prostate boundary, whereas HyWaveDiff follows the reference contour more closely in these regions. The Spleen example illustrates differences in boundary contraction and displacement, with the HyWaveDiff prediction showing close alignment with the reference contour.

These selected examples complement the complete test-set results in Table 2. They illustrate local segmentation behavior without implying that the same ordering holds for every test case.



(a) PROMISE12: prostate MRI.



(b) Task09 Spleen: spleen CT.

Figure 2: Qualitative segmentation comparisons on PROMISE12 MRI and Task09 Spleen CT. Orange solid contours indicate model predictions, while cyan dashed contours indicate ground-truth boundaries. All methods within each row are visualized on the same image slice.

4.5 Uncertainty Diagnostics

Post-hoc diagnostics use a fixed supervised warm-up teacher and shared predictions generated with $K = 8$ and 20 DDIM steps (Appendix A). Hyperbolic and Euclidean dispersion produce closely aligned risk–coverage and selective Dice curves. These static rankings should be distinguished from the segmentation outcomes after training in Table 3. The diagnostic protocol, ground-truth-assisted evaluation regions, and visualizations are detailed in the appendix.

5 Conclusion

We presented HyWaveDiff, a reliability-guided diffusion framework that converts stochastic teacher disagreement and image-derived structural evidence into refined pseudo-labels and continuous supervision weights for semi-supervised medical image segmentation. With three labeled training cases per dataset, HyWaveDiff achieves the highest Dice among the evaluated semi-supervised methods on PROMISE12 and Task09 Spleen. Ablations and estimator-replacement experiments support the design within the evaluated settings, while fixed-teacher diagnostics show that the downstream benefit is not explained by static uncertainty ranking alone. Future work will explore efficient volumetric and multi-class extensions.

AI USE STATEMENT

In this work, generative AI tools were used selectively to assist with research ideation, literature retrieval and discovery, drafting limited portions of the manuscript, and language polishing. The overall research methodology, experimental design, implementation, data analysis, and scientific conclusions were developed and validated by the authors. All AI-assisted content was critically reviewed, revised where necessary, and independently verified against the relevant experiments and original sources. The authors take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. Section 3 provides the formulation of HyWaveDiff, including the stochastic EMA teacher, hyperbolic feature dispersion, wavelet-guided refinement, Multi-Cue Reliability Refinement (MRR), and reliability-aware optimization. Section 4.1 reports the patient-level dataset partitions, evaluation protocol, training and inference settings, and checkpoint selection procedure. Additional details on data preprocessing, optimization, validation, and implementation settings are provided in Appendix C. The fixed-teacher uncertainty diagnostics, including the sampling configuration, evaluation regions, and uncertainty analyses, are documented in Appendix A. All compared semi-supervised baselines were rerun under the same patient-level partitions and evaluation protocol used for HyWaveDiff,and the held-out test sets were excluded from training, checkpoint selection, and hyperparameter tuning.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, et al. The medical segmentation decathlon. *Nature Communications*, 13:4128, 2022. doi: 10.1038/s41467-022-30695-9. URL https://www.nature.com/articles/s41467-022-30695-9.

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4443–4452, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00441. URL https://ieeexplore.ieee.org/document/9880422/.

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. doi: 10.1109/CVPR52729.2023.01108. URL https://openaccess.thecvf.com/content/CVPR2023/html/Bai_Bidirectional_Copy-Paste_for_Semi-Supervised_Medical_Image_Segmentation_CVPR_2023_paper.html.

Yunzhu Chen, Yang Liu, Manti Lu, Liyao Fu, and Feng Yang. Multi-consistency for semi-supervised medical image segmentation via diffusion models. *Pattern Recognition*, 161:111216, May 2025. doi: 10.1016/j.patcog.2024.111216. URL https://www.sciencedirect.com/science/article/pii/S0031320324009671.

Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive Bidirectional Displacement for Semi-Supervised Medical Image Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4070–4080, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.00390. URL https://ieeexplore.ieee.org/document/10658193/.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5345–5355, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html.

Along He, Tao Li, Yanlin Wu, Ke Zou, and Huazhu Fu. FRCNet: Frequency and Region Consistency for Semi-supervised Medical Image Segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical*

Image Computing and Computer Assisted Intervention – MICCAI 2024, volume 15008, pp. 305–315. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72110-6 978-3-031-72111-3. doi: 10.1007/978-3-031-72111-3_29. URL https://link.springer.com/chapter/10.1007/978-3-031-72111-3_29. Series Title: Lecture Notes in Computer Science.

Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya Wen, Meng Wu, Cong Hu, Bingliang Hu, and Quan Wang. β-FFT: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30839–30849, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hu_beta-FFT_Nonlinear_Interpolation_and_Differentiated_Training_Strategies_for_Semi-Supervised_Medical_CVPR_2025_paper.html.

Shumeng Li, Jian Zhang, Lei Qi, Luping Zhou, Yinghuan Shi, and Yang Gao. Diversity-enhanced Collaborative Mamba for Semi-supervised Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 45(2):500–514, February 2026. doi: 10.1109/TMI.2025.3601450. URL https://ieeexplore.ieee.org/document/11133471/.

Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip Eddie Edwards, Bianca Maan, Ferdinand Van Der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, February 2014. ISSN 13618415. doi: 10.1016/j.media.2013.12.002. URL https://www.sciencedirect.com/science/article/pii/S1361841513001734.

Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 886–895, June 2018. doi: 10.1109/CVPRW.2018.00121. URL https://openaccess.thecvf.com/content_cvpr_2018_workshops/w13/html/Liu_Multi-Level_Wavelet-CNN_for_CVPR_2018_paper.html.

Xinyu Liu, Wuyang Li, and Yixuan Yuan. DiffRect: Latent Diffusion Label Rectification for Semi-supervised Medical Image Segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012, pp. 56–66. Springer Nature Switzerland, Cham, 2024a. ISBN 978-3-031-72389-6 978-3-031-72390-2. doi: 10.1007/978-3-031-72390-2_6. URL https://link.springer.com/chapter/10.1007/978-3-031-72390-2_6. Series Title: Lecture Notes in Computer Science.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual State Space Model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063, December 2024b. doi: 10.52202/079017-3273. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/baa2da9ae4bfed26520bb61d259a3653-Abstract-Conference.html.

Chao Ma and Ziyang Wang. Semi-mamba-unet: Pixel-level contrastive and cross-supervised visual mamba-based UNet for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 300:112203, 2024. doi: 10.1016/j.knosys.2024.112203. URL https://www.sciencedirect.com/science/article/pii/S0950705124008372.

Himashi Peiris, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. Co-Manifold learning for semi-supervised medical image segmentation. *Neurocomputing*, 639:130208, July 2025. ISSN 09252312. doi: 10.1016/j.neucom.2025.130208. URL https://www.sciencedirect.com/science/article/pii/S092523122500880X.

Xavier Pennec. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, July 2006. ISSN 0924-9907, 1573-7683. doi: 10.1007/s10851-006-6228-4. URL https://link.springer.com/article/10.1007/s10851-006-6228-4.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1195–1204, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion Models for Implicit Image Segmentation Ensembles. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pp. 1336–1348. PMLR, 2022. URL https://proceedings.mlr.press/v172/wolleb22a.html.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 1623–1639. PMLR, 2024. URL https://proceedings.mlr.press/v227/wu24a.html.

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11765 of *Lecture Notes in Computer Science*, pp. 605–613. Springer International Publishing, July 2019. doi: 10.1007/978-3-030-32245-8_67. URL https://doi.org/10.1007/978-3-030-32245-8_67.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00355. URL https://ieeexplore.ieee.org/document/10377881/.

APPENDIX

A ADDITIONAL UNCERTAINTY DIAGNOSTICS

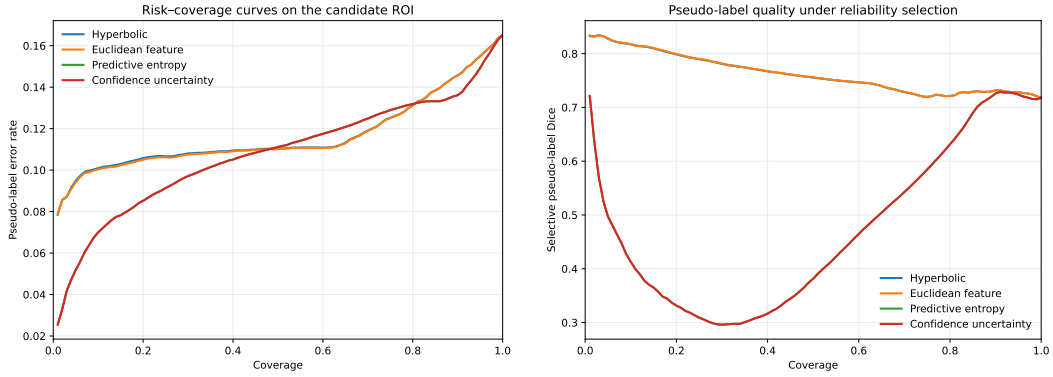
Diagnostic protocol. The post-hoc diagnostics use a fixed supervised warm-up teacher on the PROMISE12 test set. All estimators use the same cached stochastic predictions, generated with $K = 8$ samples and 20 DDIM reverse steps. Raw pseudo-labels are obtained by thresholding the mean foreground probability at 0.40.

This setting differs from training-time teacher sampling ($K = 4$, five DDIM steps) and from evaluation of the final Stage-2 model. The diagnostic figures therefore describe raw warm-up pseudo-label behavior rather than the final segmentation results in Table 2.

Risk–coverage and selective Dice. The candidate region is a two-dimensional, ten-pixel dilation of the union of the ground truth and raw pseudo-label. This ground-truth-assisted region is used only for post-hoc evaluation.

Candidate pixels are ranked in ascending order of uncertainty. At each coverage, the corresponding fraction of lowest-uncertainty pixels is retained. Risk is the pseudo-label error rate on the retained pixels. Selective Dice measures foreground overlap on the retained support, excluding unselected pixels. When both retained foreground sets are empty, selective Dice is defined as one. Curves are computed per case and then averaged.

Figure 3 compares the pixel-level selection behavior of hyperbolic and Euclidean feature dispersion under the fixed-teacher diagnostic protocol. The two estimators produce closely aligned risk–coverage and selective-Dice curves, while Table 3 evaluates their effects when integrated into the complete training pipeline. In that downstream comparison, the hyperbolic configuration yields the strongest segmentation performance among the evaluated estimators.



(a) Risk–coverage. (b) Selective Dice.

Figure 3: Uncertainty-based selection using a fixed supervised warm-up teacher on PROMISE12. Coverage is the retained fraction of candidate pixels. Lower risk and higher selective Dice are preferable. The candidate region is defined using the ground truth and raw pseudo-label for post-hoc evaluation only.

Spatial uncertainty behavior. Figure 4 compares uncertainty maps with raw pseudo-label errors. Figure 5 provides the corresponding full-field views.

The examples are the lowest-scoring case and the lower-median case when the ten test cases are ranked by raw pseudo-label Dice. For each case, the displayed slice has the largest ground-truth foreground area.

The enlarged visualization uses the bounding box of the union of the ground truth and raw pseudo-label, expanded by ten pixels. This rectangular crop is distinct from the dilated candidate mask used for the quantitative curves.

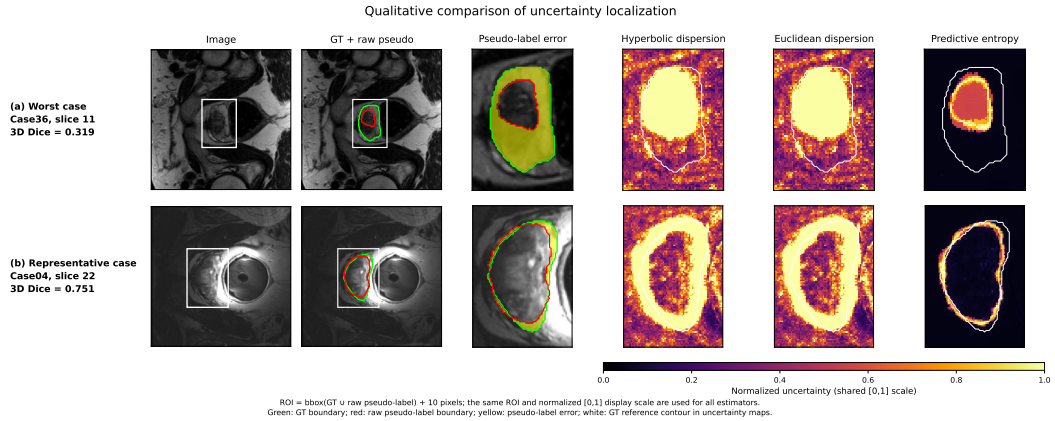


Figure 4: Qualitative comparison of uncertainty localization. The rows show the lowest-scoring and lower-median cases under raw pseudo-label Dice. Green and red contours denote ground-truth and raw pseudo-label boundaries, respectively. Yellow regions show pseudo-label errors, and white contours indicate ground-truth boundaries in uncertainty maps. The crop is the bounding box of the union of the ground truth and raw pseudo-label, expanded by ten pixels.

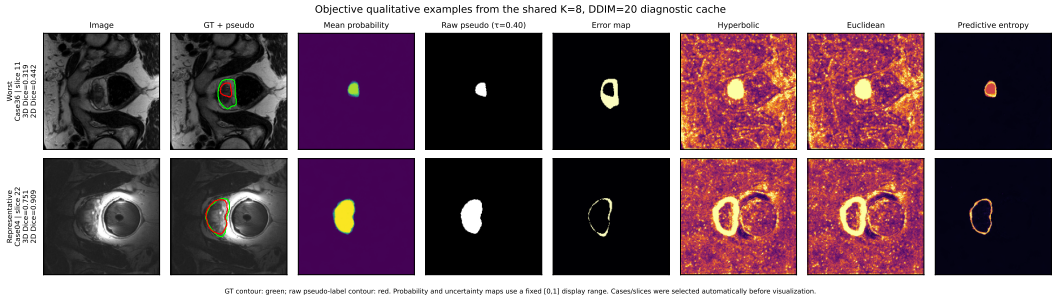


Figure 5: Full-field diagnostic examples from the shared $K = 8$, 20-step DDIM prediction cache. The panels show the image, reference and pseudo-label contours, mean foreground probability, raw pseudo-label, error map, and uncertainty estimates. These examples characterize the fixed warm-up teacher and are not final Stage-2 segmentation predictions.

Uncertainty maps are displayed on normalized $[0, 1]$ scales. A common display range does not imply identical numerical calibration across estimators.

B Boundary Calibration and Reliability Refinement

Haar boundary response. The image-derived boundary response is

$$E_W^u = \mathcal{N}_{0.95}\left(\mathrm{Up}\left[\frac{1}{3C_I}\sum_{j=1}^{C_I}\sum_{b\in\{LH,HL,HH\}}\left|\mathcal{W}_b(I_{w,j}^u)\right|\right]\right), \tag{18}$$

where C_I is the number of image channels, $\mathcal{W}_b$ extracts Haar subband b, and Up restores the original spatial resolution. The normalization $\mathcal{N}_{0.95}$ is defined in Eq. (9).

Hyperbolic–wavelet calibration. Let $p_0 = \bar{P}_{\mathrm{fg}}^u$ be the ensemble foreground probability. The foreground boundary band B_f^u is constructed from the teacher prediction using morphological dilation and erosion with a 5×5 kernel. The gates are

$$G_U^u = \sigma\left(\frac{U_H^u - 0.45}{0.10}\right), \qquad G_W^u = \sigma\left(\frac{E_W^u - 0.40}{0.10}\right), \tag{19}$$

where σ denotes the sigmoid function. The calibrated probability is

$$\begin{aligned} p_1 = \mathrm{clip}\big[&p_0 + 0.20 B_f^u G_U^u G_W^u (1 - p_0) \\ &- 0.20 B_f^u G_U^u (1 - G_W^u) p_0,\ 0,\ 1\big]. \end{aligned} \tag{20}$$

Within the boundary band, the wavelet response determines the correction direction, while hyperbolic dispersion modulates its strength. The initial hyperbolic reliability weight is

$$W_H^u = \mathrm{clip}(1 - U_H^u, 0.10, 1). \tag{21}$$

Reliability construction. MRR combines calibrated prediction confidence, stochastic consistency, and hyperbolic reliability:

$$\begin{aligned} R_P^u &= 2|p_1 - 0.5|, \\ V_K^u &= \frac{1}{K} \sum_{k=1}^{K} (P_{k,\mathrm{fg}}^u - \bar{P}_{\mathrm{fg}}^u)^2, \\ R_N^u &= \exp(-V_K^u / 0.03), \\ R_H^u &= \exp[-\mathcal{N}_{\mathrm{mm}}(U_H^u)/0.20], \\ R_0^u &= R_P^u R_N^u R_H^u. \end{aligned} \tag{22}$$

Image-wise min–max normalization is

$$\mathcal{N}_{\mathrm{mm}}(X) = \frac{X - \min_x X(x)}{\max_x X(x) - \min_x X(x) + \varepsilon}. \tag{23}$$

The resulting reliability is

$$R^u = \mathrm{clip}(R_0^u W_H^u, 0.10, 1). \tag{24}$$

This expression is algebraically equivalent to the implemented weighting rule under the shared lower bound of 0.10 and $W_H^u \le 1$.

Equivalence of the clipped reliability rule. Let $r = R_0^u$, $w = W_H^u$, $b = B_R^u$, and $m = 0.10$. For $r \ge 0$, $0 \le w \le 1$, and $0 \le b \le 1$,

$$\begin{aligned} &\mathrm{clip}\left([r(1-b) + \max(r, m)b]w,\ m,\ 1\right) \\ &\quad = \mathrm{clip}(rw,\ m,\ 1). \end{aligned} \tag{25}$$

If $r \ge m$, the two expressions agree before clipping. If $r < m$, the bracketed term is at most m. Multiplication by $w \le 1$ leaves both expressions at most m, so the final clipping returns m. Thus, this boundary-dependent intermediate floor does not independently alter the final reliability. The boundary band remains active in the subsequent Sobel-guided probability correction.

Sobel-guided probability correction. For the subsequent correction, define

$$B_R^u = \mathrm{Band}_5(\mathbb{I}[p_1 > 0.5]), \tag{26}$$

where Band_5 constructs a boundary band using morphological dilation and erosion with a 5×5 kernel. The Sobel response and gate are

$$\begin{aligned} E_S^u &= \mathcal{N}_{\mathrm{mm}}\left(\sqrt{(S_x * I_w^u)^2 + (S_y * I_w^u)^2 + \varepsilon}\right), \\ G_S^u &= \sigma\left(\frac{E_S^u - 0.40}{0.10}\right), \end{aligned} \tag{27}$$

where S_x and S_y are the horizontal and vertical Sobel kernels. The refined foreground probability is

$$\begin{aligned} \widetilde{p}^u = \mathrm{clip}\big[&p_1 + 0.20 B_R^u \big(G_S^u (1 - p_1) \\ &- (1 - G_S^u) p_1\big),\ 0,\ 1\big]. \end{aligned} \tag{28}$$

The refined soft pseudo-label is $\widetilde{P}^u = [1 - \widetilde{p}^u, \widetilde{p}^u]$.

Class-aware weighting. Let $\hat{y}^u(x) = \arg\max_c \widetilde{P}^u_c(x)$ and $C^u(x) = \max_c \widetilde{P}^u_c(x)$. The final reliability weight is

$$W^u(x) = \mathrm{clip}\left[R^u(x)\sigma\left(\frac{C^u(x) - \tau_{\hat{y}^u(x)}}{0.05}\right) M_{\hat{y}^u(x)},\ 0.10,\ 1\right], \tag{29}$$

with $(\tau_0, \tau_1) = (0.90, 0.80)$ and $(M_0, M_1) = (0.25, 1.00)$. Indices 0 and 1 denote background and foreground. MRR returns the refined soft target $\widetilde{P}^u$ and continuous weight map W^u. The thresholds define smooth confidence gates, and the positive floor retains supervision at low-reliability locations.

C Experimental Details

Datasets and partitions. PROMISE12 contains 50 prostate MRI cases (Litjens et al., 2014). The patient-level training, validation, and test partitions contain 36, four, and ten cases, respectively. Three training cases are labeled and the other 33 are unlabeled, corresponding to approximately 8.3% of the training pool being labeled.

Task09 Spleen contains 41 annotated CT volumes (Antonelli et al., 2022). We use 30 training, four validation, and seven test volumes. Three training volumes are labeled and 27 are unlabeled, giving a 10% labeled setting. After two-dimensional preprocessing, the labeled training, unlabeled training, validation, and test subsets contain 226, 2,506, 371, and 547 slices, respectively.

The compared methods use the same patient-level partitions. Validation data are used for checkpoint selection; test data are excluded from training and model selection.

Evaluation. Slice-wise predictions are ordered by case and slice index and reconstructed into three-dimensional volumes. Dice and Jaccard assess volumetric overlap; HD95 and ASD assess surface distance. Physical voxel spacing is propagated through preprocessing: in-plane spacing is adjusted according to the resizing ratio, while through-plane spacing is preserved. Surface distances use the resulting spacing and are reported in millimeters. Tables report means and standard deviations.

Implementation. Inputs are resized to 224×224, and the base feature dimension is 80. AdamW uses a learning rate of 5×10^{-5}, weight decay of 10^{-4}, three learning-rate warm-up epochs, and cosine decay. Labeled and unlabeled batch sizes are both two, with gradient accumulation over four iterations. The EMA teacher decay is 0.995.

Supervised warm-up. Stage 1 trains the complete architecture on labeled data for 180 epochs, with the unlabeled loss disabled. Validation is performed every five epochs using four stochastic predictions, 20 DDIM steps, a foreground threshold of 0.50, and three-dimensional largest-connected-component post-processing. The checkpoint with the highest validation whole-organ Dice initializes the Stage-2 student. The EMA teacher is initialized as an identical copy.

Semi-supervised training. Stage 2 runs for 150 epochs. The unlabeled branch is activated after the first validation cycle and contributes from epoch 6 onward. The teacher generates $K = 4$ stochastic predictions with five DDIM steps per unlabeled image. The unsupervised coefficient follows a 100-epoch Gaussian ramp-up and is modulated by RA-SSL.

Final inference. We average $K = 8$ stochastic predictions generated with 20 DDIM steps and apply a foreground threshold of 0.40 on both datasets. Following three-dimensional reconstruction, the largest connected foreground component is retained. Model checkpoints are selected using validation Dice.