Probabilistic Multi-Rater Segmentation via Style-Aware Boundary Conditioning

Sanaz Karimijafarbigloo[1], Amirhossein Kazerouni[2,3,4], Alireza Kheirkhah[5], Sergey Protserov[2], Reza Azad[1], Michael Brudno[2,3,4], Babak Taati[2,3,4] Wojciech Samek[6,7], and Dorit Merhof[1]

[1]University of Regensburg, [2]University of Toronto, [3]Vector Institute, [4]University Health Network, [5]Iran University of Science and Technology, [6]Technische Universität Berlin [7]Fraunhofer Heinrich Hertz Institute
sanaz.karimijafarbigloo@ur.de

Abstract. Multi-rater medical image segmentation reflects the inherent ambiguity of clinical interpretation, where experts often agree on the semantic extent of a target but differ in fine-grained boundary precision and textural cues. Existing approaches frequently collapse this variability into consensus labels or treat inter-rater differences as noise, leading to overconfident predictions and limited control over personalized outputs. In this work, we view rater variability as structured and preference-driven, modeling it through a style factor that captures individual expert delineation of ambiguous regions while preserving a shared content representation for anatomical consistency. We introduce a unified probabilistic framework that disentangles shared content from rater-specific style and employs lightweight rater conditioning to generate personalized yet anatomically consistent segmentations without decoder duplication. To better align learning with actual expert disagreement, we propose a novel disagreement-aware boundary loss that focuses supervision within a controllable band around regions of annotation divergence. This loss heightens sensitivity to contour-level mismatches where inter-rater ambiguity is greatest, while remaining stable in regions of agreement, thereby tying uncertainty estimation and personalization to clinically meaningful boundary disagreement. Experiments on LIDC-IDRI, Kvasir-SEG, and NPC-170 show better aggregated and per-rater performance than strong multi-rater baselines, while generating diverse, plausible segmentations. https://github.com/sanazkarimi/style-aware

Keywords: Medical Image Segmentation · Multi-rater · Uncertainty

1 Introduction

Multi-rater medical image segmentation leverages annotations from multiple experts to address inherent inter-observer variability and ambiguity in delineating anatomical or pathological structures in modalities such as magnetic resonance imaging (MRI), computed tomography (CT), or ultrasound (US) [10]. This variability arises from subtle imaging features, differences in clinical expertise, and uncertainty in borderline cases, making a single ground truth inadequate [8].

Traditional consensus methods such as majority voting, soft averaging, or statistical fusion (e.g., STAPLE [21]) often discard valuable information about disagreement, biasing models toward overly confident predictions [2]. By explicitly modeling this diversity, multi-rater approaches improve uncertainty quantification, robustness, and clinical reliability [7].

Early statistical methods such as STAPLE [21] estimated latent ground truth from noisy labels, while consensus methods (e.g., majority vote, soft averaging) simplified supervision but discarded inter-rater variability [2]. Early deep models used MC Dropout [4], ensembles [9], or multi-headed networks [15] to represent uncertainty, yet often blurred epistemic vs. aleatoric effects and did not match expert label distributions [14]. Probabilistic U-Net [7] advanced this line by introducing latent variables to model sample-level ambiguity and generate diverse, plausible segmentations. Recent work has moved toward personalized, generative segmentation. D-LEMA [13] explicitly disentangles expert styles, while relational models learn annotator-specific behaviors [6]. Prompt-guided methods such as D-Persona [23] and PU-Net [18] unify diversity and personalization via latent queries and conditioning. Diffusion-based approaches (e.g., DiffOSeg [25], CaliDiff [19]) further enable robust, expert-aligned sampling through probabilistic aggregation. More recent probabilistic frameworks, including ProSeg [12] and MrPrism [22], focus on calibrated uncertainty/self-calibration and extend to annotation-scarce, semi-supervised regimes [8].

Despite strong performance, most personalization methods struggle to capture fine-grained rater-specific boundary behavior. For example, D-Persona [23] first learns a shared representation of annotation diversity, then freezes the backbone and adapts each rater using a lightweight projection head. This design improves stability and preserves overall anatomical structure, but it also limits personalization: once the shared representation is fixed, any subtle boundary cues that were smoothed out during the first stage cannot be recovered in the second stage. As a result, rater-specific outputs are produced mainly by reinterpreting the same frozen features, which can reduce sensitivity to small contour shifts and local boundary details where disagreement is most common. In contrast, we disentangle a shared content representation that preserves anatomical consistency from a rater-specific style factor that encodes each expert's boundary preference in ambiguous regions. Concretely, we personalize within the network by conditioning intermediate feature maps on the rater's style, so the representation itself can shift toward that rater's delineation. Rather than freezing a shared latent space and learning only rater-specific heads (or duplicating decoders), we introduce **Style-Conditioned Personalization (SCP)** to capture the style code from intermediate encoder features, enabling fine-grained, rater-dependent boundary refinement in the decoding process. To ensure supervision targets the locations where disagreement is clinically meaningful, we further propose a **Disagreement-Aware Boundary Loss (DABL)** that focuses learning within a controllable band around regions of inter-rater divergence, emphasizing contour-level mismatches while remaining stable in areas of agreement. **SCP**

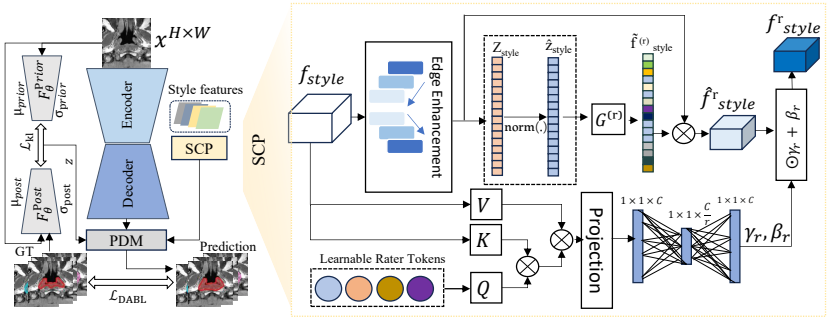

Fig. 1: Illustration of the style-conditioned personalization with a disagreement-aware boundary loss for robust, rater-specific segmentation.

and **DABL** generate personalized, anatomically consistent segmentations with uncertainty concentrated in disagreement regions and low where experts agree.

2 Method

As illustrated in Figure 1, our unified probabilistic framework aims to learn the conditional distribution $p(y \mid x)$ of plausible segmentation masks given an input medical image x, while explicitly accounting for structured inter-rater variability arising from differences in boundary interpretation and fine-grained delineation behavior. In contrast to approaches that encode rater-specificity solely at the latent level, we argue that effective personalization requires conditioning intermediate feature representations to recover subtle cues that may be suppressed during shared representation learning. To this end, we introduce two key components: (i) a style-conditioned feature modulation mechanism that enables rater-specific refinement without duplicating the backbone, and (ii) a disagreement-aware boundary loss that concentrates supervision in regions of annotation divergence. Both components are integrated under a shared probabilistic formulation, ensuring coherent uncertainty modeling and stable training.

2.1 Probabilistic Backbone

We build our framework upon the Probabilistic U-Net [7], which models the conditional distribution of segmentation masks using a latent variable z: $p_\theta(y \mid x) = \int p_\theta(y \mid x, z)\, p_\theta(z \mid x)\, \mathrm{d}z$. Given an input image $x \in \mathbb{R}^{H \times W}$ and a set of rater annotations $\mathcal{A} = \{A^{(1)}, \ldots, A^{(R)}\}$, a shared encoder F_θ^{enc} extracts feature maps $f = F_\theta^{\text{enc}}(x)$. A prior network F_θ^{prior} predicts Gaussian parameters $(\mu_{\text{prior}}, \sigma_{\text{prior}})$ conditioned on x, while a posterior network F_θ^{post} infers $(\mu_{\text{post}}, \sigma_{\text{post}})$ conditioned on $(x, A^{(r)})$. Latent codes are sampled via the reparameterization trick $z = \mu + \sigma \odot \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, and concatenated with f before decoding to produce segmentation hypotheses.

2.2 Style-Conditioned Personalization (SCP)

To enable rater-specific refinement within the decoding process, we introduce a style-conditioned feature modulation mechanism that operates on intermediate encoder representations. Let $f_{\text{style}} \in \mathbb{R}^{C\times H\times W}$ denote style features obtained from the concatenation of selected encoder feature maps. First, to emphasize fine-grained boundary information, we extract multi-scale edge cues from f_{style}. Specifically, we apply downsampling (D) and upsampling (U) at two scales:

$$f_{u1} = U\big(D(f_{\text{style}}, s_1), s_0\big), f_{u2} = U\big(D(f_{\text{style}}, s_2), s_0\big), s_i = (0.25, 0.5, 1.0) \quad (1)$$

The edge component is computed as $f_{\text{edge}} = |f_{u1} - f_{u2}|$, and injected back into the style features using learnable channel-wise weights $\lambda \in \mathbb{R}^{C_l}$: $f'_{\text{style}} = f_{\text{style}} + \lambda \odot f_{\text{edge}}$. Then, a global style descriptor is obtained via spatial aggregation: $z_{\text{style}} = \text{avg}(f'_{\text{style}}) \in \mathbb{R}^{C_l}$, followed by channel-wise normalization: $\hat{z}_{\text{style}} = \frac{z_{\text{style}} - \mu(z_{\text{style}})}{\sigma(z_{\text{style}}) + \epsilon}$, where $\mu(\cdot)$ and $\sigma(\cdot)$ denote channel-wise mean and standard deviation.

In addition, each rater r is associated with a learnable style token s_r. A lightweight MLP maps s_r to a scale parameter c_r, yielding a rater-dependent excitation function:

$$g^{(r)} = \exp\left(-\frac{\hat{z}^2_{\text{style}}}{2c_r^2}\right). \quad (2)$$

This excitation modulates the enhanced style features channel-wise: $\hat{f}^{(r)}_{\text{style}} = g^{(r)} \odot f'_{\text{style}}$. To further adapt features during decoding, rater embeddings are mapped to FiLM parameters: $[\gamma_l^{(r)}, \beta_l^{(r)}] = \phi_l(s_r)$, and applied as: $f^{(r)}_{\text{style}} = \gamma^{(r)} \odot \hat{f}^{(r)}_{\text{style}} + \beta^{(r)}$. Finally, the **Personalized Decoding Module (PDM)** takes a latent sample z_i and the rater style feature $f^{(r)}_{\text{style}}$, and decodes a personalized segmentation via

$$\hat{y}_i = F_\theta^{\text{dec}}\big(f, z_i, f^{(r)}_{\text{style}}\big). \quad (3)$$

During training, $\hat{y}_i$ is supervised with the corresponding rater's annotations while the backbone features f and latent prior are shared across raters; finally, conditioning on $f^{(r)}_{\text{style}}$ enables rater-specific outputs without decoder duplication.

2.3 Disagreement-Aware Boundary Loss (DABL)

In multi-rater segmentation, annotation discrepancies are predominantly localized near object boundaries, where local texture ambiguity and subjective contour interpretation lead to systematic disagreement among annotators. To explicitly account for this phenomenon, we introduce **DABL** that concentrates supervision within regions of inter-rater disagreement while preserving stable region-based learning elsewhere. Given R binary ground-truth masks $\{G^{(r)}\}_{r=1}^{R}$, we first identify pixels where annotators disagree by computing the union and intersection masks

$$G_U(x) = \max_r G^{(r)}(x), \qquad G_I(x) = \min_r G^{(r)}(x), \quad (4)$$

and defining the disagreement region as $D(x) = G_U(x) - G_I(x)$. This region captures pixels for which at least one annotator differs from the others and typically corresponds to ambiguous boundary locations. To extend supervision to a local neighborhood around these ambiguous regions, we dilate D with a radius τ, obtaining a boundary band D_τ, where τ controls the spatial extent of boundary emphasis. We convert the boundary band into a spatial weight map $w(x) = 1 + \lambda D_\tau(x)$, where λ controls the relative importance assigned to boundary regions compared to interior regions. Using this weight map, we define weighted soft intersection and union between a predicted probabilistic segmentation $P^{(r)}$ and the corresponding annotator mask $G^{(r)}$ as

$$I_w^{(r)} = \sum_x w(x)\, P^{(r)}(x)\, G^{(r)}(x),\ U_w^{(r)} = \sum_x w(x)\, \big(P^{(r)}(x) + G^{(r)}(x) - P^{(r)}(x)G^{(r)}(x)\big). \tag{5}$$

The Disagreement-Aware Boundary Loss for annotator r is then defined as a boundary-focused Difference-over-Union objective,

$$\mathcal{L}_{\mathrm{DABL}}^{(r)} = \frac{U_w^{(r)} - I_w^{(r)}}{U_w^{(r)} - \alpha I_w^{(r)} + \varepsilon}, \tag{6}$$

where α balances the relative contribution of overlap and mismatch, and ε is a small constant for numerical stability. In the multi-rater setting, the final loss is obtained by averaging over annotators, $\mathcal{L}_{\mathrm{DABL}} = \frac{1}{R}\sum_{r=1}^{R} \mathcal{L}_{\mathrm{DABL}}^{(r)}$. By explicitly focusing optimization on regions of inter-rater disagreement, the proposed loss encourages precise yet flexible boundary placement, yielding anatomically consistent predictions while satisfying annotator-specific boundary preferences.

2.4 Probabilistic Objective

The final training objective is:

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \mathbb{E}_{z\sim q_\theta(z|x,A^{(r)})}\big[\mathcal{L}_{\mathrm{seg}}(\hat{y}^{(r)}, A^{(r)})\big] \\ + \lambda_{\mathrm{KL}}\, \mathrm{KL}\big(q_\theta(z \mid x, A^{(r)}) \,\|\, p_\theta(z \mid x)\big) + \lambda_{\mathrm{DABL}}\, \mathcal{L}_{\mathrm{DABL}}.\end{aligned} \tag{7}$$

Here, $\mathcal{L}_{\mathrm{seg}}$ denotes the standard Dice–CE loss. This objective jointly enforces probabilistic consistency, rater-specific refinement, and boundary-aligned supervision. We train the model in two phases. In Phase I, we optimize only the probabilistic backbone for 100 epochs to learn the encoder, prior, and latent space. In Phase II, we freeze these components and train the SCP and decoder for 150 epochs. Training uses Adam with an initial learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 1×10^{-5}, we set $\lambda_{\mathrm{KL}} = 1$ and $\lambda_{\mathrm{DABL}} = 0.5$. This staged optimization preserves a stable shared probabilistic latent space before introducing rater-conditioned modulation, as fully end-to-end optimization may drift the shared representation toward annotator-specific biases and weaken disentanglement between anatomy and personalization. Moreover, SCP directly modulates intermediate feature representations through FiLM conditioning, enabling localized boundary refinement while preserving globally consistent anatomy.

Table 1: Comparison of distribution fitting and sampling diversity performance on LIDC–IDRI and NPC–170 datasets. # denotes the number of samples.

Method	LIDC–IDRI Dataset				NPC–170 Dataset			
	$GED \downarrow$	$Dice_{soft} \uparrow$	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$GED \downarrow$	$Dice_{soft} \uparrow$	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$
Prob. U-Net [7] (#10)	0.2181	88.79	88.60	88.43	0.3585	81.24	**84.25**	80.18
D-Persona [23] (#10)	0.1461	90.24	90.75	90.51	0.2314	83.21	83.03	80.18
Proposed Method (#10)	**0.1145**	**91.56**	**92.33**	**91.42**	**0.2001**	**83.95**	81.65	**80.30**
Prob. U-Net [7] (#30)	0.2169	88.79	88.80	88.73	0.3536	81.22	**84.19**	80.14
D-Persona [23] (#30)	0.1375	90.42	91.23	91.16	0.2090	83.74	82.78	80.98
Proposed Method (#30)	**0.1053**	**91.76**	**92.31**	**91.89**	**0.1885**	**84.37**	81.69	**81.71**
Prob. U-Net [7] (#50)	0.2168	88.80	88.87	88.81	0.3528	81.19	**84.19**	80.13
D-Persona [23] (#50)	0.1358	90.45	91.37	91.33	0.1978	84.01	82.79	81.69
Proposed Method (#50)	**0.1035**	**91.80**	**92.29**	**91.98**	**0.1808**	**84.51**	81.90	**82.10**

3 Experimental Results

Datasets: Experiments are conducted on three multi-rater benchmarks: **LIDC-IDRI** [1] contains thoracic CT scans annotated by up to four radiologists with substantial inter-observer variability. Following prior work [26], 1,609 axial slices from 214 patients are extracted, resampled to 0.5 mm isotropic resolution, and cropped into 128×128 nodule-centered patches. Four-fold patient-level cross-validation is used. **NPC-170** [23] consists of multi-modal MRI scans (T1, T2, T1c) from 170 nasopharyngeal carcinoma patients, each annotated independently by four radiation oncologists. The dataset is split into 100/20/50 patients for training/validation/testing, following [20]. The **Kvasir-SEG** dataset [5] contains 1,000 gastrointestinal endoscopy images, we follow the [20] setting.

Evaluation Metrics: We assess diversity and expert-specific fidelity using GED, $\text{Dice}_{\text{soft}}$, coverage metrics ($\text{Dice}_{\text{max}}$, $\text{Dice}_{\text{match}}$), and per-annotator agreement (Dice_A, $\text{Dice}_{\text{mean}}$), following prior works [7,23,5].

Quality and Diversity of the Predicted Distribution: As shown in Table 1, the proposed Network achieves the strongest distributional alignment on both datasets. As the number of samples K increases from 10 to 50, GED steadily decreases and $\text{Dice}_{\text{soft}}$ improves, indicating that the model systematically expands its coverage of plausible annotations without over-dispersing predictions. The Probabilistic U-Net [7] improves only marginally with larger K (GED drops from 0.2181 to 0.2168 on LIDC), suggesting that its latent prior remains poorly calibrated and tends to generate redundant hypotheses. D-Persona [23] learns a spatially modulated prior but without explicit style disentanglement; our method separates shared anatomical content from rater-specific boundary style before modeling inter-rater variability, enabling a more faithful mapping between anatomical uncertainty and sample diversity. This separation yields consistently lower GED (0.1035 vs. 0.1358 on LIDC and 0.1808 vs. 0.1978 on NPC at K=50) and higher $\text{Dice}_{\text{soft}}$ (91.80 vs. 90.45 and 84.51 vs. 84.01, respectively). In essence, our method's prior focuses diversity on clinically ambiguous boundaries while preserving deterministic predictions in unambiguous regions.

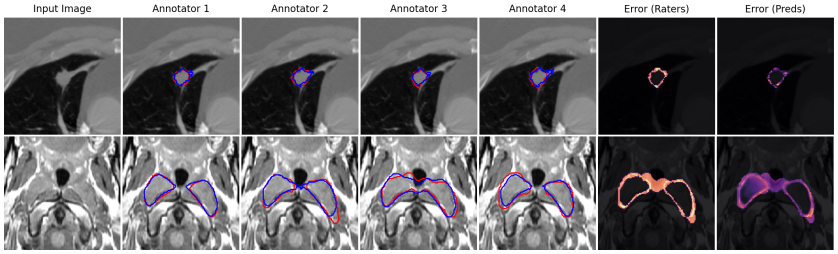

Fig. 2: Visualization results on LIDC and NPC-170, where red contours denote ground truth annotations and blue contours indicate model predictions.

Table 2: Our model compared to single-expert U-Nets (top) and personalized frameworks (bottom) on LIDC-IDRI and NPC-170.

Dataset	Method	Diversity Performance		Personalized Bounds (%)		Personalized Segmentation Performance (%)				
		$GED \downarrow$	$Dice_{soft} \uparrow$ (%)	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$
LIDC-IDRI	U-Net (A_1)	0.3062	86.59	N/A		87.80	87.47	85.49	80.67	85.36
	U-Net (A_2)	0.2459	88.43			87.16	89.08	88.59	85.15	87.50
	U-Net (A_3)	0.2436	88.20			85.29	88.48	89.40	87.20	87.59
	U-Net (A_4)	0.2962	85.83			80.80	85.48	88.22	88.90	85.85
	Prob. U-Net [7]	0.2168	88.80	88.87	88.81	-	-	-	-	
	CM-Global [17]	0.2432	88.53	87.51	87.51	86.13	88.76	88.99	86.18	87.51
	CM-Pixel [26]	0.2407	88.64	87.72	87.72	85.99	88.81	89.31	86.77	87.72
	TAB [11]	0.2322	86.35	87.11	86.08	85.00	86.35	86.77	85.77	85.97
	Pionono [16]	0.1502	90.00	90.10	88.97	87.94	89.11	89.55	88.76	88.84
	D-Persona [23]	0.1444	90.31	90.38	89.17	88.54	89.50	90.03	88.60	89.17
	GSD-Net [20]	-	-	-	-	88.53	87.98	88.33	88.16	88.25
	AmbiSSL [8]	0.1620	89.86	90.03	89.84	-	-	-	-	-
	Proposed Method	**0.1213**	**91.22**	**92.37**	**90.23**	**89.26**	**89.98**	**91.06**	**90.64**	**90.23**
NPC-170	U-Net (A_1)	0.4322	78.65	N/A		87.02	73.68	73.15	77.05	77.85
	U-Net (A_2)	0.4520	76.45			77.70	79.25	74.10	74.05	76.35
	U-Net (A_3)	0.4515	76.75			76.35	76.05	79.15	74.85	76.70
	U-Net (A_4)	0.4600	77.80			80.45	76.10	77.20	79.85	78.90
	Prob. U-Net [7]	0.3528	81.19	84.19	80.13	-	-	-	-	-
	CM-Global [17]	0.3609	81.31	**85.39**	80.13	84.78	77.60	77.99	80.15	80.13
	CM-Pixel [26]	0.4523	78.09	81.20	76.77	80.47	74.79	74.06	77.77	76.77
	TAB [11]	0.2900	81.25	81.60	80.02	83.10	78.10	79.50	76.50	79.30
	Pionono [16]	0.3334	81.66	85.00	80.32	**85.10**	77.59	78.27	80.31	80.32
	D-Persona[23]	0.2970	82.30	81.60	80.50	84.20	**79.70**	**80.50**	77.20	80.40
	GSD-Net [20]	-	-	-	-	82.37	78.62	77.64	78.35	79.24
	Proposed Method	**0.2635**	**82.78**	84.27	**80.97**	84.69	78.54	79.82	**80.58**	**80.97**

Rater-Specific Segmentation: With style-conditioned personalization (Phase II), the proposed method delivers consistent improvements over all personalized baselines (Table 2). On LIDC-IDRI, our model achieves the best personalized bounds ($\text{Dice}_{\text{max}} = 92.37\%$, $\text{Dice}_{\text{match}} = 90.23\%$) and the highest balanced per-rater performance, outperforming D-Persona [23] by +1.06 pp in $\text{Dice}_{\text{mean}}$ (90.23% vs. 89.17%). On NPC-170, a more challenging multi-modal dataset with pronounced inter-rater disagreement, our method surpasses both transformer-based TAB [11] and probabilistic Pionono [16], reaching 80.97% mean Dice. Unlike D-Persona, which freezes a shared representation and adapts each rater via a lightweight projection head, our SCP module conditions intermediate feature maps on rater-specific style tokens through FiLM modulation, allowing it to capture fine-grained boundary cues without disrupting structural consistency. The narrow gap between Dice_{max} and $\text{Dice}_{\text{match}}$ on both datasets indicates that

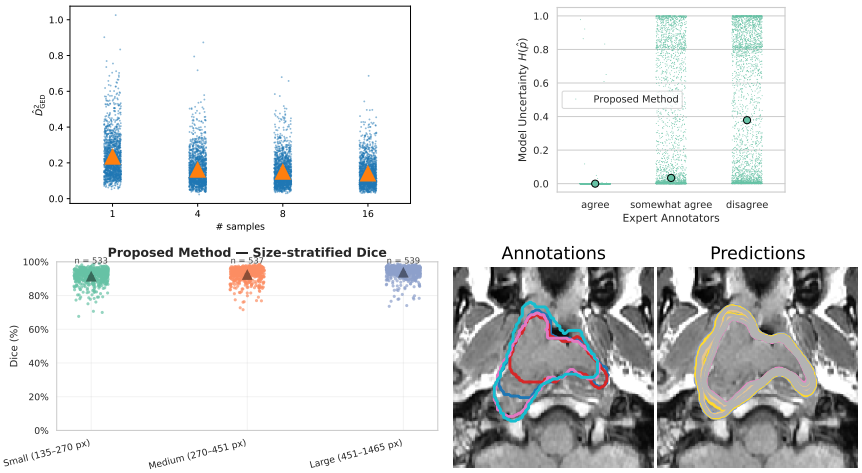

Fig. 3: (a) Diversity vs. sample count, (b) model uncertainty, (c) size-stratified Dice, and (d) qualitative limitations on the LIDC dataset.

each individualized prediction is genuinely tailored rather than a random sample fitting by chance. Qualitative examples are provided in Fig. 2.

Efficiency/Performance Ablations: Our method remains lightweight: the full model has 30.22 M parameters, where the Probabilistic U-Net backbone has 30.11 M and the SCP module adds only 0.11 M; DABL adds no learnable parameters. The overhead over the backbone is $< 0.37\%$, showing that personalization is achieved via lightweight modulation rather than heavy architectural expansion. Our method has 3.6 GFLOPs. In ablations, removing SCP decreases Dice by ~ 1.0 point, while removing DABL reduces Dice by ~ 0.24 on the LIDC.

Further Ablations: Figure 3 illustrates (a) GED loss decreases as the sample count increases, improving alignment with the rater distribution. (b): relationship between model uncertainty and rater uncertainty, showing that predictive entropy increases monotonically with inter-rater disagreement, indicating well-calibrated and human-aligned uncertainty estimates. It also (c): reports size-stratified Dice scores, demonstrating consistent segmentation performance across small, medium, and large lesions, with no degradation for small or ambiguous structures. Eventually (d): a failure case where the model struggles to learn a minority small-boundary annotation when most raters label a larger boundary. On Kvasir-SEG (Table 3), we evaluate under three simulated noisy-label splits: S_R (random noise, σ=0.2), S_E (extreme noise, σ=0.8), and S_{DE} (diverse expert noise), while testing on clean expert annotations. Our method achieves the best Dice across all splits and outperforms the strongest baseline GSD-Net [20] by +5.17, +3.62, and +0.16, respectively. Furthermore, Table 3 shows per-rater calibration. Moreover, on the LIDC dataset, across all annotators, ECE metric remains very low (3×10^{-3}–5×10^{-3}) and Brier scores stay below 6×10^{-3}, indicating consistently well-calibrated and reliable probabilistic predictions.

Table 3: **Left:** Performance on the Kvasir dataset using Dice score. **Right:** Per-rater calibration on LIDC (Personalized), where lower is better.

Method	Kvasir dataset		
	S_R	S_E	S_{DE}
RMD [3]	68.34±2.18	71.57±1.09	66.90±1.75
CDR [24]	67.87±3.51	70.58±1.65	63.51±1.67
IDMPS [27]	77.52±1.21	74.16±0.81	69.47±0.81
D-Persona [23]	84.69±0.19	81.77±0.16	78.93±0.18
GSD-Net [20]	80.04±0.42	79.39±0.33	79.97±0.53
Our Method	**85.21±0.22**	**83.01±0.17**	**80.13±0.19**

Rater	ECE ↓	Brier ↓
A1	0.003276	0.003409
A2	0.003130	0.003511
A3	0.003314	0.003545
A4	0.005133	0.005440
Mean	0.00370	0.00398

4 Conclusion

We propose a lightweight probabilistic multi-rater segmentation framework that captures rater-specific boundary preferences via style-conditioned personalization and a disagreement-aware loss. Across benchmarks, it improves accuracy and calibration over strong baselines, with minor parameter overhead and efficient inference, while remaining robust under inter-rater disagreement.

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

Acknowledgments: This work was funded by the German Research Foundation (DFG), project 455548460, and DeSBi [KI-FOR 5363], project 459422098.

References

1. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
2. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis **54**, 280–296 (2019). https://doi.org/10.1016/j.media.2019.03.009
3. Fang, C., Wang, Q., Cheng, L., Gao, Z., Pan, C., Cao, Z., Zheng, Z., Zhang, D.: Reliable mutual distillation for medical image segmentation under imperfect annotations. IEEE Transactions on Medical Imaging **42**(6), 1720–1734 (2023)
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR (2016)
5. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
6. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12341–12350 (2021)

7. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
8. Kumari, S., Singh, P.: Annotation ambiguity aware semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10404–10413 (2025)
9. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30, pp. 6405–6416. Curran Associates, Inc. (2017)
10. Li, H.B., Navarro, F., Ezhov, I., Bayat, A., Das, D., Kofler, F., Shit, S., Waldmannstetter, D., Paetzold, J.C., Hu, X., et al.: Qubiq: Uncertainty quantification for biomedical image segmentation challenge. arXiv preprint arXiv:2405.18435 (2024)
11. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Transformer-based annotation bias-aware medical image segmentation. pp. 24–34 (2023)
12. Liu, K., Gao, S., Fu, Y., Gao, S., Shen, C.: Probabilistic modeling of multi-rater medical image segmentation for diversity and personalization. arXiv preprint arXiv:2512.00748 (2025)
13. Mirikharaji, Z., Abhishek, K., Izadi, S., Hamarneh, G.: D-lema: Deep learning ensembles from multiple annotations – application to skin lesion segmentation. In: Proc. IEEE/CVF Computer Vision and Pattern Recognition Workshops (CVPRW) – ISIC (2021)
14. Roshanzamir, P., Rivaz, H., Ahn, J., Mirza, H., Naghdi, N., Anstruther, M., Battié, M.C., Fortin, M., Xiao, Y.: How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. pp. 74–83. Springer (2023)
15. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 3591–3600 (2017). https://doi.org/10.1109/ICCV.2017.385
16. Schmidt, A., Morales-Álvarez, P., Molina, R.: Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. pp. 21097–21106 (2023)
17. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. pp. 11244–11253 (2019)
18. Wang, J., Cheng, Y., Chen, J., Xu, H., Chen, D., Wu, J.: Multi-rater prompting for ambiguous medical image segmentation. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2024 (2024)
19. Wang, J., Wang, J., Ma, J., Chen, B., Chen, Z., Zheng, Y.: Calidiff: Multi-rater annotation calibrating diffusion probabilistic model towards medical image segmentation. Medical Image Analysis p. 103812 (2025)
20. Wang, T., Zhang, Z., Zhou, Y., Zhang, X., Chen, Y., Tan, T., Yang, G., Tong, T.: From noisy labels to intrinsic structure: A geometric-structural dual-guided framework for noise-robust medical image segmentation. arXiv preprint arXiv:2509.02419 (2025)
21. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)

22. Wu, J., Fang, H., Zhu, J., Zhang, Y., Li, X., Liu, Y., Liu, H., Jin, Y., Huang, W., Liu, Q., Chen, C., Liu, Y., Duan, L., Xu, Y., Xiao, L., Yang, W., Liu, Y.: Multi-rater prism: Learning self-calibrated medical image segmentation from multiple raters. Science Bulletin **69**(18), 2906–2919 (2024). https://doi.org/10.1016/j.scib.2024.06.037
23. Wu, Y., Luo, X., Xu, Z., Guo, X., Ju, L., Ge, Z., Liao, W., Cai, J.: Diversified and personalized multi-rater medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11470–11479 (2024)
24. Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., Chang, Y.: Robust early-learning: Hindering the memorization of noisy labels. In: International conference on learning representations (2020)
25. Zhang, H., Luo, X., Chen, Y., Li, K.: Diffoseg: Omni medical image segmentation via multi-expert collaboration diffusion model. arXiv preprint arXiv:2507.13087 (2025)
26. Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.: Disentangling human error from ground truth in segmentation of medical images. Advances in Neural Information Processing Systems **33**, 15750–15762 (2020)
27. Zhao, X., Li, Z., Luo, X., Li, P., Huang, P., Zhu, J., Liu, Y., Zhu, J., Yang, M., Chang, S., et al.: Ultrasound nodule segmentation using asymmetric learning with simple clinical annotation. IEEE Transactions on Circuits and Systems for Video Technology **34**(10), 9010–9023 (2024)