

Harmonized Feature Conditioning and Frequency-Prompt Personalization for Multi-Rater Medical Segmentation

Sanaz Karimijafarbigloo¹ Armin Khosravi² Alireza Kheyrikhah³
 Reza Azad¹ Mauricio Reyes^{4,5} Dorit Merhof^{1,6}

¹Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

²Sharif University of Technology, Tehran, Iran

³Iran University of Science and Technology, Tehran, Iran

⁴ARTORG Center for Biomedical Research, University of Bern, Bern, Switzerland

⁵Department of Radiation Oncology, University Hospital Bern, University of Bern, Bern, Switzerland

⁶Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

Abstract

Multi-rater medical image segmentation captures the inherent ambiguity of clinical interpretation, where diagnostic boundaries vary across experts and imaging devices. Existing approaches often reduce this diversity to consensus labels or treat rater differences as noise, resulting in overconfident and poorly calibrated models. We propose a harmonized probabilistic framework that disentangles acquisition artifacts from genuine annotator variability through adaptive feature conditioning and frequency-domain personalization. A lightweight Harmonizer Network implicitly models scanner-specific artifacts and performs dynamic feature modulation to standardize latent representations, ensuring that uncertainty reflects anatomy rather than noise. To represent rater-specific styles, we introduce a novel High-Frequency Prompt Modules that operate in the spectral domain to encode annotator-dependent boundary precision and textural sensitivity. These prompts adaptively modulate harmonized features to produce personalized yet anatomically consistent segmentations. Furthermore, a Generalized Energy Distance (GED)-based regularization aligns the generative distribution with empirical annotation variability, promoting diversity where experts disagree and consensus where they converge. Experiments on LIDC-IDRI and NPC-170 show SOTA aggregated and individualized segmentation, with notable GED reductions and improved Dice scores, especially on noisy cases. Beyond accuracy, the model exhibits clinically meaningful uncertainty. Confidence rises in agreement regions and declines in ambiguous areas, supporting its use as a reliable and interpretable tool for multi-expert clinical workflows. [GitHub code](#)

1. Introduction

Multi-rater medical image segmentation utilizes annotations from multiple clinical experts to delineate anatomical structures, acknowledging the inherent ambiguity and inter-observer variability in medical imaging [13]. This variability renders the notion of a single, definitive ground truth untenable and stems from multiple factors: the subtlety and ambiguity of imaging features, differences in annotators' training and clinical judgment, and the intrinsic uncertainty associated with early-stage or borderline pathologies [25]. Inter-rater variability is a significant challenge in medical image segmentation, particularly when delineating structures such as lung nodules, prostate tumors, or retinal layers, where interpretations often diverge among specialists (e.g., radiologists, oncologists, or ophthalmologists) [12]. Traditional preprocessing strategies, such as majority voting or annotation averaging, are ill-suited to capture the nuances of interobserver variability in medical imaging. These aggregation-driven approaches discard critical information about annotator disagreement and underlying uncertainty, often biasing models toward overly confident or misleading labels [5]. Recognizing this limitation, multi-rater medical image segmentation has emerged as a paradigm that explicitly leverages, rather than suppresses, expert disagreement. In contrast to conventional supervised learning, these methods utilize multiple annotations to infer probabilistic label fusion, model annotator-specific behavior, or generate a distribution of plausible segmentations [10], thereby better reflecting the inherent uncertainty in clinical data and enhancing model interpretability, robustness, and clinical trustworthiness [2, 29]. STAPLE [26], an early statistical approach, estimated a latent ground truth but relied on restrictive parametric assumptions. The advent of deep learning has since

enabled a shift toward expressive probabilistic frameworks: the Probabilistic U-Net [11] introduced sample-level uncertainty modeling via latent distributions, paving the way for subsequent advances such as expert-specific disentanglement (D-LEMA [16]), relational annotator modeling [9], and prompt-guided personalization (D-Persona [28]).

Despite these advances, most existing multi-rater segmentation frameworks still operate primarily in the spatial domain, making them vulnerable to scanner-specific noise, acquisition artifacts, and inconsistent annotation quality. Such perturbations can propagate through the latent space, conflating clinically meaningful uncertainty with irrelevant variability and hindering cross-scanner generalization. Previous efforts like D-Persona [28] and DiffOSeg [31] introduced annotator-specific latent codes for personalized predictions, yet they remain limited to feature-level modulation and do not explicitly disentangle noise effects from diagnostic variability. Moreover, recent analyses of latent probabilistic models [21] show that naive latent distributions often become sparse or under-regularized, reducing their ability to represent structured variation across scanners. To address these issues, we propose a lightweight Harmonizer Network that adaptively conditions features across scales. Hence, it learns data-driven feature modulation to adapt to scanner variations, reducing noise and yielding a latent space that captures true anatomy.

Building on this harmonized foundation, our innovative framework integrates personalization directly within the frequency domain, where stylistic differences among annotators, such as boundary precision, texture sensitivity, and lesion delineation style, manifest as high-frequency cues. To capture these variations, we introduce High-Frequency Prompt Modules that decompose features via wavelet transforms and adaptively modulate the harmonized latent representation to reflect each expert’s diagnostic tendencies. This spectral personalization allows the model to synthesize expert-specific segmentations without retraining or replicating the backbone, maintaining both efficiency and interpretability. In parallel, a GED-based regularization aligns the learned predictive distribution with empirical annotation variability, reinforcing consensus where experts agree and promoting diversity where they diverge. Together, these components form a unified probabilistic framework that disentangles scanner-induced artifacts from genuine annotator variability, providing a lightweight yet effective solution for harmonized and personalized multi-rater segmentation.

2. Related work

2.1. Label Fusion

Label fusion techniques in multi-rater medical image segmentation aggregate diverse expert annotations into a single consensus ground truth, often referred to as meta-

segmentation, to mitigate inter-observer variability and provide a reliable reference for training deterministic models. Early methods, like majority voting or STAPLE [26], probabilistically weight annotations based on estimated rater performance to fuse labels, assuming an underlying true segmentation exists but is obscured by individual biases.

Deep learning has introduced more advanced strategies for label fusion in multi-rater medical image segmentation. Soft Label Fusion [15] preserves inter-rater variability by averaging annotations into probabilistic soft labels, reducing overconfidence near uncertain boundaries and enhancing calibration relative to deterministic label fusion methods. D-LEMA [16] advances this by using voxel-wise reliability maps to adaptively weight annotations across spatial dimensions, enabling more robust consensus in regions of local disagreement. Subsequent methods incorporate task-specific context; for instance, DiFF [27] presents a diagnosis-aware framework that weights annotator labels based on their relevance to diagnostic outcomes, jointly modeling segmentation and disease classification to produce a fused representation that is both consistent and clinically pertinent. Further developments explicitly model uncertainty. DiffOSeg [31], a two-stage diffusion-based architecture, establishes a robust population-level fusion followed by the generation of personalized expert-style segmentations using adaptive prompts. This integration of agreement-centric and preference-oriented learning reflects the evolution from simple static averaging to dynamic, expert-informed representations.

Despite their utility, label fusion methods often overlook the inherent diversity of expert opinions, risking an over-aggregated representation that discards valuable uncertainty. They also assume a definitive meta-segmentation, which may fail in ambiguous cases where consensus doesn’t equal truth. These methods relate to diversity-preserving approaches by supplying fused labels for generative training, yet they emphasize a single reliable outcome over multiple plausible hypotheses.

2.2. Diversity-preserving

Diversity-preserving methods model the conditional distribution of plausible labels given an input image to generate multiple diverse yet realistic segmentation samples, capturing annotation ambiguities without collapsing to a single consensus. Kohl et al. [10] pioneered the Probabilistic U-Net, a one-stage CVAE-integrated framework with a low-dimensional latent space where variants are encoded; it uses a prior network for image-conditioned distribution estimation and a training-only posterior for learning, producing unlimited samples via stochastic sampling. However, its isotropic Gaussian latent distribution hinders modeling of complex, multimodal annotation variations.

Subsequent works enhance hierarchical structures to

propagate stochasticity across resolutions for improved sample quality. Baumgartner et al. [2] propose PHiSeg, a one-stage hierarchical CVAE with separate latent variables per resolution level in a U-Net-like architecture, ensuring stochasticity through all skips; it generates unlimited samples matching ground-truth distributions on prostate MRI and LIDC, outperforming Probabilistic U-Net in GED and NCC metrics. Similarly, the Hierarchical Probabilistic U-Net [17] decomposes latents coarsely to finely, modeling multi-scale ambiguities for instance segmentation tasks like neuronal cells; it achieved superior GED and HM-IoU on the LIDC dataset compared to PHiSeg. Diffusion-based models advance diversity through stochastic denoising processes. CIMD et al. [18], a one-stage diffusion framework that models ambiguity through latent space diffusion without additional priors, produces diverse masks across CT, ultrasound, and MRI modalities; it introduces the CI score and outperforms VAE baselines in D_max and collective insight. Although these approaches implicitly model diverse annotation styles across multiple raters, they fail to capture true personalized segmentation behavior tailored to individual experts, remaining centered on consensus-level variability rather than expert-specific adaptation.

2.3. Personalized adaptation

Personalized and expert-aware modeling approaches aim to produce segmentation predictions tailored to the distinct annotation styles and preferences of individual experts. Early methods such as Who Said [6] and Disentangle [32] modeled annotator-specific biases using confusion matrices and probabilistic noise models, disentangling true anatomical structures from annotation noise to enable robust consensus estimation while retaining individual annotator characteristics. Based on probabilistic frameworks, BayesianNN [8] introduced a shared encoder with rater-specific Bayesian decoders that feature dedicated attention mechanisms. This enabled efficient modeling of annotator-specific styles, outperforming classical probabilistic baselines such as PHiSeg.

Personalization evolved toward structured, interpretable designs. TAX [4] enabled transparency via annotator-specific kernels and a prototype bank, while That Label’s Got Style [30] disentangled style from uncertainty using one-hot style conditioning. Prompt-based methods unified this direction: PU-Net [22] introduced rater-specific and aggregated prompts, treating annotators as domains; D-Persona [28] enhanced diversity through a constrained latent space queried by expert prompts. Most recently, DifIOseg [31] integrated agreement modeling and personalization in a diffusion framework, with probabilistic aggregation in Stage I and adaptive prompting in Stage II, replicating expert styles without architectural redundancy.

2.4. Motivation

Despite notable progress, existing multi-rater segmentation frameworks remain limited by their sensitivity to scanner-dependent noise and their inability to distinguish data-level ambiguity from genuine inter-rater variability. Scanner artifacts, acquisition heterogeneity, and inconsistent annotation quality often entangle within the latent space, obscuring clinically meaningful uncertainty and degrading generalization. Meanwhile, rater-specific nuances in boundary precision or texture perception demand adaptive modeling rather than static fusion. Our framework addresses these intertwined challenges through a unified design centered on the Harmonizer Network, which mitigates acquisition-induced artifacts while preserving annotator diversity via frequency-aware conditioning. This dual-level harmonization lets the model coherently represent both ambiguity sources, stabilizing features and capturing rater-specific behavior.

3. Method

As illustrated in Figure 1, our unified probabilistic framework aims to learn the conditional distribution $p(y | x)$ of plausible segmentation masks given an input medical image x , while explicitly disentangling two intertwined sources of ambiguity: data-level noise, originating from scanners and acquisition heterogeneity, and rater-level variability, driven by subjective diagnostic interpretation. To address these challenges, we introduce two cooperating modules within a single architecture: a Noise Harmonizer that standardizes latent representations across acquisition domains, and a Personalization Module that adapts the harmonized features in the frequency domain to capture rater-specific nuances. Both operate under a shared probabilistic backbone, ensuring coherent uncertainty modeling across spatial and spectral dimensions.

We build our base model upon the probabilistic formulation of the Probabilistic U-Net [10] approach, which models the conditional distribution of segmentation masks y given an image x through a low-dimensional latent variable z :

$$p_{\theta}(y | x) = \int p_{\theta}(y | x, z) p_{\theta}(z | x) dz, \quad (1)$$

where $p_{\theta}(z | x)$ is the learned prior and $p_{\theta}(y | x, z)$ is the decoder likelihood.

Given an image $x \in \mathbb{R}^{H \times W}$ and its n corresponding rater annotations $\mathcal{A} = \{A^{(1)}, \dots, A^{(n)}\}$, the encoder network F_{θ}^{enc} extracts spatial features $f = F_{\theta}^{\text{enc}}(x)$, while the prior $F_{\theta}^{\text{prior}}$ and posterior F_{θ}^{post} networks predict Gaussian parameters $\{\mu_{\text{prior}}, \sigma_{\text{prior}}\}$ and $\{\mu_{\text{post}}, \sigma_{\text{post}}\}$, respectively. Latent codes z are sampled via the reparameterization trick $z = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, and concatenated with feature maps f before being fed to a decoder F_{θ}^{dec} to produce segmentation hypotheses $\hat{y}$ [10].

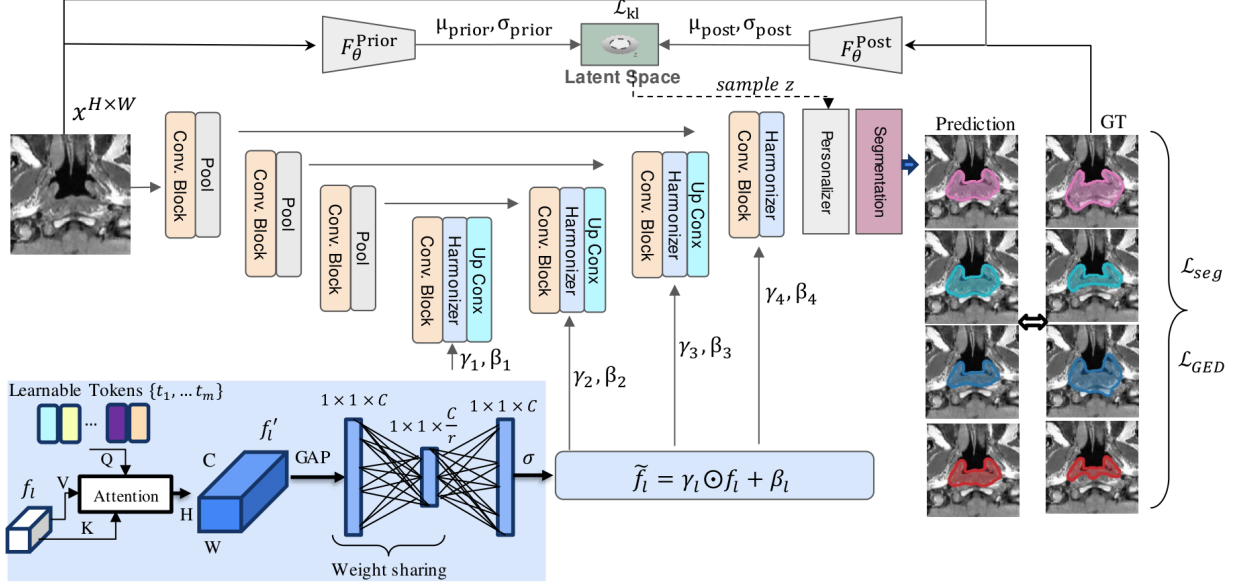


Figure 1. Illustration of the proposed Harmonizer Network for multi-rater medical image segmentation.

3.1. Noise Harmonizer

The Noise Harmonizer $\mathcal{H}_\phi^{(n)}$ mitigates scanner- and acquisition-induced artifacts that distort latent features and confound the learning of true anatomical variability. Given decoder features f_l at layer l , the harmonizer predicts spatially broadcasted modulation parameters (γ_l, β_l) conditioned on the features and a learnable bank of artifact tokens: $t = \{t_1, \dots, t_M\}$ representing characteristic noise patterns:

$$\begin{aligned} [\gamma_l, \beta_l] &= \mathcal{H}_\phi^{(n)}(f_l, t) = W_2(\sigma(W_1(GAP(f'_l)))) \\ f'_l &= \text{Softmax}\left(\frac{Q_j K_j^\top}{\sqrt{D_h}}\right) V_j, \end{aligned} \quad (2)$$

where K and V are generated from f_l , while Q is obtained from learnable tokens. $GAP(\cdot)$ denotes global average pooling, and σ denotes a non-linearity. The harmonized output:

$$\tilde{f}_l = \gamma_l \odot f_l + \beta_l, \quad (3)$$

with $\odot$ denoting element-wise multiplication. This affine transformation acts as a data-driven normalization mechanism that dynamically suppresses intensity drifts, motion artifacts, and domain biases without requiring explicit knowledge of noise distribution. Weight sharing across layers imposes a cross-scale regularization, ensuring consistent denoising behavior throughout the network.

The learned parameters (γ_l, β_l) implicitly encode the acquisition condition, guiding the network produce anatomically stable latent codes z that are invariant to scanner variability, yet expressive to capture structural uncertainty.

3.2. Rater-Level Adaptation

While the Noise Harmonizer standardizes data-level features, the Personalization Module $\mathcal{H}_\phi^{(p)}$ adapts the latent representation to reflect rater-specific tendencies in the frequency domain. This module assumes that subtle stylistic variations, such as differences in edge delineation, texture sensitivity, or lesion extent, manifest predominantly in the high-frequency spectrum.

Given feature map $X \in \mathbb{R}^{H \times W \times D}$, we begin by reducing its dimensionality through a linear projection $\tilde{X} = XW_d$, where $W_d \in \mathbb{R}^{D \times \frac{D}{4}}$. The transformed feature $\tilde{X}$ is then decomposed in the spectral domain using the Discrete Wavelet Transform (DWT) with a Haar basis, resulting in four sub-bands that separate low- and high-frequency information, $\tilde{X} = [X_{LL}, X_{LH}, X_{HL}, X_{HH}] \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{D}{4}}$, where X_{LL} represents the structural outline, while X_{LH} , X_{HL} , and X_{HH} contain detailed texture and edge information that are often interpreted differently by individual raters. These high-frequency maps are concatenated into a unified tensor $X_H = [X_{LH}, X_{HL}, X_{HH}] \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times D'}$, ($D' = \frac{3D}{4}$), which serves as input to the Rater-Aware Prompt Projection block.

Personalized Module: The personalized module (shown in Figure 2) is designed to encode annotator-specific preferences in the spectral domain and to propagate these latent representations into the decoding process. Within this block, N prompt components $P_c \in \mathbb{R}^{N \times \frac{H}{2} \times \frac{W}{2} \times D'}$ are used to represent potential annotation biases. A learnable personalized weight c_i modulates these components, enabling the model to generate person-specific tokens. The model infers an adaptive weighting vector $w \in \mathbb{R}^{D'}$ from X_H to com-

bine these codes into a context-conditioned prompt:

$$\begin{aligned} \mathbf{w} &= \text{Softmax}(\text{PWC}(X_H)), \\ P &= \text{Conv}_{3 \times 3} \left(\sum_{c=1}^{D'} \mathbf{w}(c_i(P_c)) \right). \end{aligned} \quad (4)$$

This prompt encapsulates rater preference information in the spectral domain, providing adaptive modulation over high-frequency features.

The prompt and spectral features interact through an attention mechanism that aligns image textures with the inferred rater biases, producing refined frequency responses:

$$X'_H = \text{Conv}_{1 \times 1}(\text{Attention}(X_H, P)). \quad (5)$$

For the attention mechanism, we employ the Large Kernel Attention module [7]. Subsequently, a 3×3 convolution is applied to the recalibrated high-frequency components X'_H together with the low-frequency map X_L to maintain local spatial coherence, producing a compact contextual feature representation denoted X_d . Next, we construct a memory of latent priors by drawing M_z samples from the fixed prior distribution $\mathcal{D}_{\text{prior}}$, forming a prior bank $\mathbf{Z}_{\text{prior}}^{\text{bank}} \in \mathbb{R}^{D \times 1 \times M_z}$. The localized feature X_d serves as the query in a cross-attention operation, while the elements of $\mathbf{Z}_{\text{prior}}^{\text{bank}}$ act as keys and values, allowing contextual cues to be aligned with the learned latent prototypes. Subsequently, an inverse discrete wavelet transform (IDWT) is performed over the concatenation of the refined high-frequency feature X'_H and the preserved low-frequency signal X_L , resynthesizing a full-spectrum representation. The resulting feature, fused with the cross-attention output, passes through a linear projection that produces a rater-adaptive latent vector z' encoding the annotator's structural bias. This personalized latent code is finally integrated with the shared backbone features, yielding individualized yet anatomically consistent segmentations.

Through this design, the model benefits from enhanced sensitivity to subtle variations in texture and edge information, attributes that often underlie inter-rater disagreement. The jointly learned latent prompt codes therefore provide a compact representation of annotation bias, enabling the network to modulate high-frequency features according to rater-specific tendencies while maintaining anatomical fidelity and segmentation stability.

Personalized Decoding: Given z'_i , the decoder produces a personalized segmentation $\hat{y}_i = F_{\theta}^{\text{dec}}(f, z'_i, i)$ for rater r_i . During training, $\hat{y}_i$ is supervised by the corresponding rater annotation $A^{(i)}$, with the shared backbone and prior frozen:

$$\mathcal{L}_{\text{Seg}} = \sum_{i=1}^n \mathbb{E}_{z \sim p_{\theta}(z|x)} [\mathcal{L}_{\text{seg}}(\hat{y}_i, A^{(i)})]. \quad (6)$$

Because $\mathcal{P}_i$ is lightweight, personalization can be achieved efficiently even with limited rater-specific data, making the

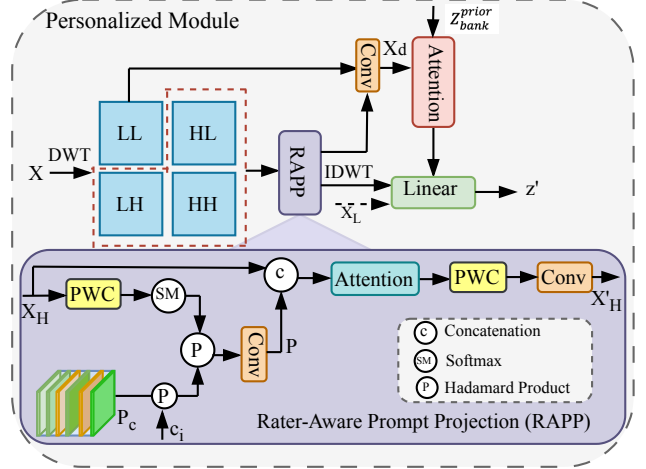


Figure 2. Proposed frequency recalibration method for personalized segmentation.

framework suitable for semi-supervised or few-shot personalization scenarios.

3.3. Diversity-Preserving

To model the distributional nature of expert disagreement, we formalize segmentation as a matching problem between two conditional distributions: the model-induced distribution $\mathcal{P}(y | x)$ and the empirical annotation distribution $\mathcal{A}(y | x)$. An ideal probabilistic model should minimize the discrepancy between these distributions while preserving variability among its own predictions. Formally, this discrepancy can be expressed through the GED [3, 10], which provides a symmetric, unbiased measure of distance between distributions. Using a segmentation metric $d(\cdot, \cdot)$ (we employ $d = 1 - \text{IoU}$), the empirical loss becomes:

$$\begin{aligned} \mathcal{L}_{\text{GED}} &= \frac{2}{KN} \sum_{k=1}^K \sum_{i=1}^N d(P_k, A_i) \\ &\quad - \frac{2}{K(K-1)} \sum_{1 \leq k < k' \leq K} d(P_k, P_{k'}), \end{aligned} \quad (7)$$

where $\{P_k\}$ are K segmentation samples drawn from the model and $\{A_i\}$ are N expert annotations. The first term aligns the model's predictive distribution with the observed annotation manifold (fidelity), while the second enforces diversity across generated samples, preventing degeneracy toward a single consensus mask. This formulation regularizes the model to generate anatomically plausible, uncertainty-aware segmentations aligned with expert variability.

3.4. Probabilistic Objective

The overall objective combines probabilistic reconstruction loss, KL divergence regularization, a harmonization

penalty, and a GED-based distribution alignment constraint:

$$\begin{aligned}\mathcal{L}_{\text{total}} = & \mathbb{E}_{z \sim q_{\theta}(z|x, A^{(k)})} [\mathcal{L}_{\text{seg}}(F_{\theta}^{\text{dec}}(f', z'), A^{(k)})] \\ & + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}(q_{\theta}(z|x, A^{(k)}) \| p_{\theta}(z|x)) \\ & + \lambda_{\text{harm}} \sum_{l=1}^L (\|\gamma_l - 1\|_2^2 + \|\beta_l\|_2^2) + \lambda_{\text{GED}} \mathcal{L}_{\text{GED}}(y, \mathcal{A}).\end{aligned}\tag{8}$$

Here, $\mathcal{L}_{\text{seg}}$ combines Dice and cross-entropy losses, while $\mathcal{L}_{\text{GED}}$ enforces statistical alignment between predicted and empirical annotation distributions.

4. Experiments and Results

4.1. Dataset

LIDC-IDRI: We evaluate our method on the LIDC-IDRI dataset [1], a standard benchmark for studying inter-observer variability in thoracic CT segmentation. The dataset provides diagnostic CT scans annotated by up to four radiologists through a two-stage process: an independent (blinded) segmentation of nodules ≥ 3 mm, followed by a consensus (unblinded) review. This process introduces substantial diversity in boundary delineations and nodule extent. Following common practice [23, 32], we extract 1,609 axial slices from 214 patients, resample all scans to 0.5 mm isotropic resolution, and crop them into 128×128 nodule-centered patches. A four-fold cross-validation at the patient level is employed for fair evaluation.

NPC-170: We further validate our framework on the NPC-170 dataset [28], which contains multi-modal MRI scans of 170 nasopharyngeal carcinoma (NPC) patients annotated by four experienced radiation oncologists. Each subject includes T1-weighted, T2-weighted, and contrast-enhanced T1 (T1c) modalities, with independent delineations of the gross tumor volume (GTVp). The dataset is divided into 100/20/50 patients for training, validation, and testing, corresponding to 6,134, 1,126, and 3,058 annotated slices, respectively. We follow [24] to ensure fair comparison and consistent evaluation across studies.

4.2. Implementation Details and Training Strategy

Our framework is built upon the probabilistic U-Net backbone [10], extended with the proposed Noise Harmonizer, Personalization Module, and the GED loss for distributional alignment. All default hyperparameter values, including learning rate schedules, optimizer configuration, and loss weighting, follow the original probabilistic U-Net implementation unless otherwise stated. To ensure stable latent representation learning and clear separation between data-level and rater-level variability, training is performed in two sequential phases, while maintaining a unified architecture.

Phase 1 — Latent Space and Noise Harmonization. In the first phase, the rater-specific frequency adaptation head is excluded, and the probabilistic backbone is trained jointly

with the Noise Harmonizer. This phase focuses on learning artifact-invariant and anatomically consistent latent features before introducing rater-dependent modulation. Training is conducted for 100 epochs using the Adam optimizer with an initial learning rate of $1e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-5} . The latent space dimensionality is set to $D = 6$, and a memory buffer of size $M = 100$ is maintained to stabilize the conditional prior distribution.

Phase 2 — Frequency-based Personalization. After the first phase converges, backbone parameters (encoder, decoder, and Noise Harmonizer) are frozen, and only the Personalization Module is trained for 150 epochs. This phase fine-tunes the frequency-domain adaptation to align spectral features with rater-specific annotation patterns while keeping the shared probabilistic representation fixed. A reduced learning rate of $5e-5$ is used to ensure smooth adaptation.

Implementation Details: All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. The full model contains 30.31 M parameters (baseline 30.11 M, $\mathcal{H}_{\phi}^{(n)}$ 0.14 M and $\mathcal{H}_{\phi}^{(p)}$ 0.07 M parameters), including both harmonization and personalization modules, and requires around 0.42 s per forward pass during inference. The network is implemented in PyTorch with mixed-precision training for efficiency, using a fixed seed in all experiments.

4.3. Evaluation Metrics

To assess performance in multi-rater segmentation, we use metrics capturing set-level diversity and expert-specific fidelity. We follow established evaluation practices from prior work on probabilistic and personalized segmentation [9, 10, 23], and use the same metrics as those employed in recent personalized segmentation studies [12, 28]. To evaluate the distributional alignment between predicted and expert segmentation distributions, we use the GED [3], a statistical metric quantifying both sample quality and internal diversity. However, as GED ignores voxel-level agreement, we complement it with the soft-thresholded Dice score [9, 23], measuring overlap between soft predictions and soft annotations across multiple thresholds. This enables a fine-grained view of voxel-wise calibration and improves robustness to specific operating points.

To assess whether the predicted set captures expert variability, we use two coverage metrics. Dice_{max} measures upper-bound coverage by selecting the best prediction for each annotation, while $\text{Dice}_{\text{match}}$ enforces stricter one-to-one matching via optimal bipartite assignment. Together, they distinguish true diversity from redundancy. Finally, personalized agreement is measured by the Dice score between each expert annotation and its corresponding individualized prediction, denoted $\text{Dice}_{A(i)}$. Averaging across experts gives $\text{Dice}_{\text{mean}}$, reflecting how well the model captures annotator-specific styles rather than collapsing to a global consensus.

Table 1. Comparison of distribution fitting and sampling diversity performance on LIDC-IDRI and NPC-170 datasets.

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 [10] (#10)	0.2181	88.79	88.60	88.43	0.3585	81.24	84.25	80.18
Prob. U-Net [10] (#30)	0.2169	88.79	88.80	88.73	0.3536	81.22	84.19	80.14
Prob. U-Net [10] (#50)	0.2168	88.80	88.87	88.81	0.3528	81.19	84.19	80.13
D-Persona [28] (#10)	0.1461	90.24	90.75	90.51	0.2314	83.21	83.03	80.18
D-Persona [28] (#30)	0.1375	90.42	91.23	91.16	0.2090	83.74	82.78	80.98
D-Persona [28] (#50)	0.1358	90.45	91.37	91.33	0.1978	84.01	82.79	81.69
Harmonizer Network (#10)	0.1164	91.54	92.28	91.70	0.1960	84.27	82.21	81.18
Harmonizer Network (#30)	0.1063	91.79	92.27	91.84	0.1764	84.82	82.23	82.36
Harmonizer Network (#50)	0.1048	91.81	92.28	91.94	0.1758	84.83	82.26	82.65

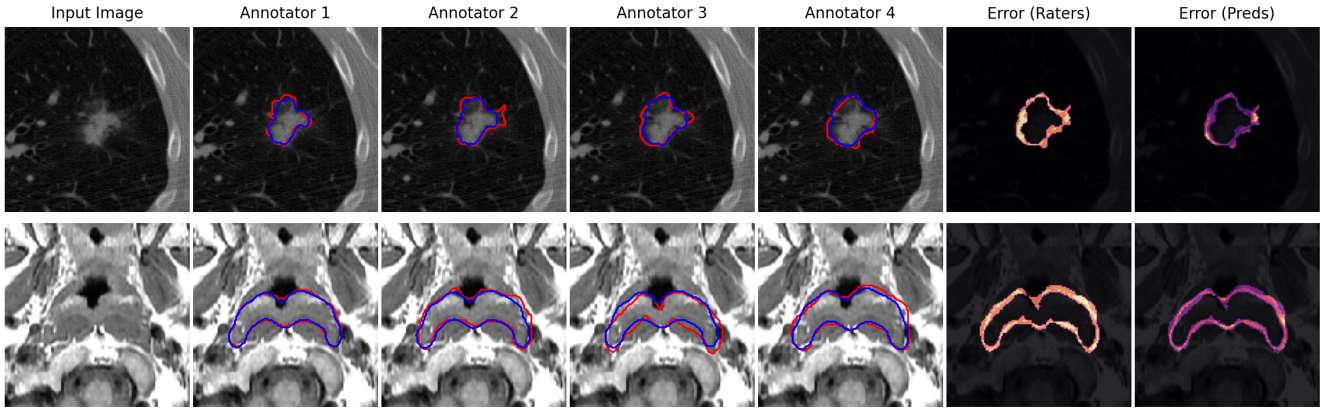


Figure 3. Visualization results on the LIDC (first row) and NPC-170 (second row) dataset with multi-rater annotations and the corresponding error map. The red boundary indicates the ground truth, while the blue boundary represents the predicted mask.

4.4. Results

Distributional fit and diversity of the sample set: As summarized in Table 1, the Harmonizer prior in Phase I achieves the strongest distributional alignment across both LIDC-IDRI and NPC-170 datasets. While the Probabilistic U-Net [10] improves marginally with more samples K , its latent priors remain poorly calibrated, often generating redundant or anatomically inconsistent hypotheses. In contrast, our phase I framework, equipped solely with the Noise Harmonizer and the GED-based distributional constraint, produces a latent manifold that is cleaner, more compact, and better aligned with the empirical annotation distribution. As K increases from 10 to 50, GED steadily decreases and $Dice_{soft}$ improves, indicating that the model systematically expands its coverage of plausible annotations without over-dispersing predictions. Compared with D-Persona [28], which learns a spatially modulated latent prior without explicit noise disentanglement, our method first suppresses scanner- and acquisition-dependent artifacts before modeling inter-rater variability. This separation enables a more faithful mapping between anatomical un-

certainty and sample diversity, yielding consistently lower GED (0.105 vs. 0.136 on LIDC and 0.175 vs. 0.197 on NPC-170) and higher $Dice_{soft}$. In essence, the Harmonizer prior learns where uncertainty matters, focusing diversity on clinically ambiguous boundaries while preserving deterministic predictions in unambiguous regions.

Personalization segmentation: When augmented with rater-aware frequency prompts in phase II, the Harmonizer Network delivers further improvements over all personalized baselines (Tables 2 and 3). On LIDC-IDRI, our model achieves the best personalized bounds ($Dicemax = 92.65\%$, $Dicematch = 90.0\%$) and balanced per-rater performance, outperforming D-Persona by roughly +1.61 pp in Dice mean. On NPC-170, a more challenging multi-modal dataset, our method surpasses both transformer-based TAB [14] and probabilistic Pionono [19], reaching 81.63% mean Dice despite pronounced inter-rater divergence. Unlike D-Persona, which conditions expert prompts on spatial features susceptible to residual acquisition noise, our frequency-prompt module operates on harmonized spectral representations, allowing it to isolate

Table 2. Comparison of the proposed Harmonizer Network against individually trained U-Net models (single-expert annotations, top), and personalized segmentation frameworks (bottom) on LIDC-IDRI dataset.

Method	Diversity Performance		Personalized Bounds (%)		Personalized Segmentation Performance (%)				
	$GED \downarrow$	$Dice_{soft} \uparrow$ (%)	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$
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
Geometric-Structural [24]	-	-	-	-	88.53	87.98	88.33	88.16	88.25
Prob. U-Net [10]	0.2168	88.80	88.87	88.81	-	-	-	-	-
CM-Global [20]	0.2432	88.53	87.51	87.51	86.13	88.76	88.99	86.18	87.51
CM-Pixel [32]	0.2407	88.64	87.72	87.72	85.99	88.81	89.31	86.77	87.72
TAB [14]	0.2322	86.35	87.11	86.08	85.00	86.35	86.77	85.77	85.97
Pionono [19]	0.1502	90.00	90.10	88.97	87.94	89.11	89.55	88.76	88.84
D-Persona [28]	0.1444	90.31	90.38	89.17	88.54	89.50	90.03	88.60	89.17
Harmonizer Network	0.1419	91.35	92.65	90.00	89.87	90.07	91.06	88.56	90.78

Table 3. Comparison of the proposed Harmonizer Network against individually trained U-Net models (single-expert annotations, top), and personalized segmentation frameworks (bottom) on NPC-170 dataset.

Method	Diversity Performance		Personalized Bounds (%)		Personalized Segmentation Performance (%)				
	$GED \downarrow$	$Dice_{soft} \uparrow$ (%)	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$
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
Geometric-Structural [24]	-	-	-	-	82.37	78.62	77.64	78.35	79.24
Prob. U-Net [10]	0.3528	81.19	84.19	80.13	-	-	-	-	-
CM-Global [20]	0.3609	81.31	85.39	80.13	84.78	77.60	77.99	80.15	80.13
CM-Pixel [32]	0.4523	78.09	81.20	76.77	80.47	74.79	74.06	77.77	76.77
TAB [14]	0.2900	81.25	81.60	80.02	83.10	78.10	79.50	76.50	79.30
Pionono [19]	0.33348	81.66	85.00	80.32	85.10	77.59	78.27	80.31	80.32
D-Persona[28]	0.2970	82.30	81.60	80.50	84.20	79.70	80.50	77.20	80.40
Harmonizer Network	0.2685	83.10	84.46	81.63	85.79	79.30	80.40	81.03	81.63

stylistic boundary cues (e.g., sharpness, textural sensitivity) without disrupting structural consistency. This separation yields personalized masks that are both anatomically aligned and rater-specific: the gap between $Dice_{max}$ and $Dice_{match}$ remains narrow, indicating that each individualized prediction is genuinely tailored rather than a random sample fitting by chance. Qualitative visualization (3) further shows that our personalized outputs adaptively refine boundary thickness and contrast according to each annotator’s style, validating that the proposed frequency-aware conditioning achieves a more faithful and interpretable form of expert personalization.

Moreover, the top blocks in Tables 2 and 3 report four independently trained U-Nets, each tuned to a single annotator. While each model peaks on its own rater (bold entries), it underperforms on others and lacks distributional coverage. D-Persona, by contrast, produces (i) a calibrated set of hypotheses that aligns with the empirical annotation distribution and (ii) personalized predictions that respect individual styles without training one network per reader.

Ablation Study and Acknowledgments

Ablation is provided in the supplementary material. This work was funded by the German Research Foundation (DFG) – project number 455548460.

5. Conclusion

We introduced the Harmonizer Network, a novel probabilistic multi-rater medical image segmentation framework. To address literature gap, the first stage employs a Noise Harmonizer with GED-based alignment to model artifact-invariant latent distributions, while the second stage integrates frequency-domain prompts to capture annotator-specific variations. Evaluations on LIDC-IDRI and NPC-170 show consistent improvements over probabilistic and personalized methods in distributional fit and personalized accuracy. The model concentrates uncertainty in ambiguous regions, maintains balanced per-rater adaptation, and remains robust to noise and blur (see supplementary), improving segmentation reliability and interpretability.

References

- [1] Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, 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] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötter, Urs J Muehlethaler, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.
- [3] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.
- [4] Yuan-Chia Cheng, Zu-Yun Shiao, Fu-En Yang, and Yu-Chiang Frank Wang. Tax: Tendency-and-assignment explainer for semantic segmentation with multi-annotators. *arXiv preprint arXiv:2302.09561*, 2023.
- [5] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.
- [6] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.
- [7] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational visual media*, 9(4):733–752, 2023.
- [8] Qingqiao Hu, Hao Wang, Jing Luo, Yunhao Luo, Zhiheng Zhang, Jan S Kirschke, Benedikt Wiestler, Bjoern Menze, Jianguo Zhang, and Hongwei Bran Li. Inter-rater uncertainty quantification in medical image segmentation via rater-specific bayesian neural networks. *arXiv preprint arXiv:2306.16556*, 2023.
- [9] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, 2021.
- [10] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.
- [11] Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019.
- [12] Suruchi Kumari and Pravendra Singh. Annotation ambiguity aware semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10404–10413, 2025.
- [13] Hongwei Bran Li, Fernando Navarro, Ivan Ezhov, Amirhossein Bayat, Dhritiman Das, Florian Kofler, Suprosanna Shit, Diana Waldmannstetter, Johannes C Paetzold, Xiaobin Hu, et al. Qubiq: Uncertainty quantification for biomedical image segmentation challenge. *arXiv preprint arXiv:2405.18435*, 2024.
- [14] Zehui Liao, Shishuai Hu, Yutong Xie, and Yong Xia. Transformer-based annotation bias-aware medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 24–34, 2023.
- [15] João Lourenço-Silva and Arlindo L Oliveira. Using soft labels to model uncertainty in medical image segmentation. In *International MICCAI brainlesion workshop*, pages 585–596. Springer, 2021.
- [16] Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghasan Hamarneh. D-lemma: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1837–1846, 2021.
- [17] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawłowski, Bernardo Marques, Konstantinos Kamnitsas, Mark Van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33:12756–12767, 2020.
- [18] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11536–11546, 2023.
- [19] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In *ICCV*, pages 21097–21106, 2023.
- [20] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *CVPR*, pages 11244–11253, 2019.
- [21] MM Amaan Valiuddin, Christiaan GA Viviers, Ruud JG Van Sloun, Peter HN De With, and Fons van der Sommen. Investigating and improving latent density segmentation models for aleatoric uncertainty quantification in medical imaging. *IEEE Transactions on Medical Imaging*, 2024.
- [22] Jinhong Wang, Yi Cheng, Jintai Chen, Hongxia Xu, Danny Chen, and Jian Wu. Multi-rater prompting for ambiguous medical image segmentation. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2024*, 2024.
- [23] Lin Wang, Xiufen Ye, Lie Ju, Wanji He, Donghao Zhang, Xin Wang, Yelin Huang, Wei Feng, Kaimin Song, and Zongyuan Ge. Medical matting: Medical image segmentation with uncertainty from the matting perspective. *Computers in Biology and Medicine*, 158:106714, 2023.

- [24] Tao Wang, Zhenxuan Zhang, Yuanbo Zhou, Xinlin Zhang, Yuanbin Chen, Tao Tan, Guang Yang, and Tong Tong. From noisy labels to intrinsic structure: A geometric-structural dual-guided framework for noise-robust medical image segmentation. *arXiv preprint arXiv:2509.02419*, 2025.
- [25] Yifeng Wang, Luyang Luo, Mingxiang Wu, Qiong Wang, and Hao Chen. Learning robust medical image segmentation from multi-source annotations. *Medical Image Analysis*, 101:103489, 2025.
- [26] Simon K Warfield, Kelly H Zou, and William M Wells. 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.
- [27] Junde Wu, Yu Zhang, Huihui Fang, Lixin Duan, Mingkui Tan, Weihua Yang, Chunhui Wang, Huiying Liu, Yueming Jin, and Yanwu Xu. Calibrate the inter-observer segmentation uncertainty via diagnosis-first principle. *IEEE Transactions on Medical Imaging*, 43(9):3331–3342, 2024.
- [28] Yicheng Wu, Xiangde Luo, Zhe Xu, Xiaoqing Guo, Lie Ju, Zongyuan Ge, Wenjun Liao, and Jianfei Cai. Diversified and personalized multi-rater medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11470–11479, 2024.
- [29] Andre Ye, Quan Ze Chen, and Amy Zhang. Confidence contours: Uncertainty-aware annotation for medical semantic segmentation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 186–197, 2023.
- [30] Kilian Zepf, Eike Petersen, Jes Frellsen, and Aasa Feragen. That label’s got style: Handling label style bias for uncertain image segmentation. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.
- [31] Han Zhang, Xiangde Luo, Yong Chen, and Kang Li. Diffoseg: Omni medical image segmentation via multi-expert collaboration diffusion model. *arXiv preprint arXiv:2507.13087*, 2025.
- [32] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. *Advances in Neural Information Processing Systems*, 33:15750–15762, 2020.