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

Sanaz Karimijafarbigloo[1] Armin Khosravi[2] Alireza Kheyrkhah[3]
Reza Azad[1] Mauricio Reyes[4,5] Dorit Merhof[1,6]

[1]Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
[2]Sharif University of Technology, Tehran, Iran
[3]Iran University of Science and Technology, Tehran, Iran
[4]ARTORG Center for Biomedical Research, University of Bern, Bern, Switzerland
[5]Department of Radiation Oncology, University Hospital Bern, University of Bern, Bern, Switzerland
[6]Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

Overview

To further support our experimental results, we provide additional empirical evidence and extended analyses validating the proposed Harmonizer Network. This supplementary material includes detailed ablations on hyperparameters, GED behavior, relationships between uncertainty and correctness, alignment of uncertainty with rater disagreement, per-rater calibration, lesion-size robustness, and the effects of noise. These studies demonstrate that the observed performance gains arise from principled model design rather than dataset-specific tuning. Additional visualizations on the LIDC-IDRI, NPC-170, and Kvasir datasets illustrate consistent boundary quality, realistic latent-space diversity, and faithful personalization of individual rater styles. Robustness experiments under severe noise confirm the stability of the harmonized latent space, while frequency-domain visualizations reveal how high-frequency prompts enhance boundary-specific details. Finally, cross-dataset evaluations and efficiency analyses verify that the method generalizes well, remains computationally lightweight, and delivers superior performance even under challenging noisy-label conditions. GitHub code

1. Selection of Hyper-parameters

To ensure a fair and reproducible evaluation, all hyperparameters governing the composite loss in Eq. (8) of the main paper were systematically tuned through one-fold validation on the LIDC–IDRI dataset and subsequently fixed across all experiments and datasets. Specifically, we optimized the balance among the segmentation loss ($\mathcal{L}_{\text{seg}}$), KL divergence term, harmonization penalty, and GED-based regularizer. The weighting coefficients λ_{KL}, λ_{harm}, and λ_{GED} were empirically selected to achieve stable convergence and the best trade-off between reconstruction fidelity, latent regularity, and distributional alignment.

During tuning, we performed a grid search around $\lambda_{\text{KL}} \in [1,5]\times10^{-3}$, $\lambda_{\text{harm}} \in [1,5]\times10^{-4}$, and $\lambda_{\text{GED}} \in [0.5, 2.0]$, evaluating each configuration by the validation GED and soft-Dice scores. The optimal configuration ($\lambda_{\text{KL}} = 2\times10^{-3}$, $\lambda_{\text{harm}} = 3\times10^{-4}$, $\lambda_{\text{GED}} = 1.0$) was adopted for all subsequent experiments, including NPC–170, without further modification. This strategy provided consistent convergence behaviour, stable uncertainty calibration, and ensured that improvements observed in cross-dataset evaluations stemmed from the model design rather than dataset-specific hyperparameter tuning.

2. GED vs. Sample Count

To examine how well the proposed model approximates the empirical annotation distribution as the number of generated samples increases, we analyze the GED as a function of the number of prior samples K, following the standard evaluation protocol introduced in Probabilistic U–Net [7]. For each test image and $K \in \{1, 4, 8, 16, 32\}$, we draw K stochastic segmentations predictions to compute GED between the predicted and annotated sets using the soft-Dice distance from GED loss, and report the mean $\pm$ standard deviation across folds. As expected, GED decreases monotonically with increasing K, reflecting improved coverage of annotation variability, and stabilizes once additional samples contribute little new diversity. Incorporating the GED regularizer during training leads to a consistent downward shift in the GED curve and earlier saturation, showing that the learned prior allocates its variability primarily to ambiguous regions while remaining deterministic elsewhere.

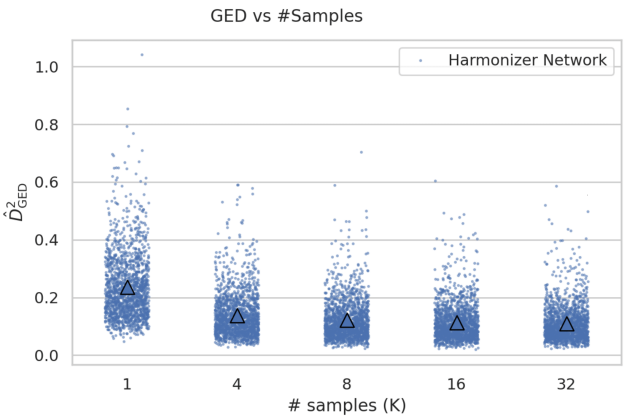


Figure 1. Impact of the proposed GED loss on LIDC. GED is reported as a function of the number of generated samples K (lower is better). Models trained with the GED loss demonstrate consistently lower discrepancy and reach saturation with fewer samples, indicating more efficient coverage of the underlying rater distribution.

Formally, optimizing $\mathcal{L}_{\text{GED}} = 2\,\mathbb{E}[d(P, A)] - \mathbb{E}[d(P, P')]$, with $d(\cdot, \cdot)$ denoting the soft-Dice distance, simultaneously reduces cross-set discrepancies and maintains controlled intra-set diversity, yielding a more calibrated posterior. In practice, GED computation scales linearly with K for $d(P, A)$ and quadratically for $d(P, P')$; we therefore cap K at 32 and subsample pairs for $d(P, P')$ without affecting observed trends. As shown in Figure 1, the Harmonizer prior exhibits consistently lower GED values and faster convergence than D–Persona [14], confirming its superior distributional alignment and sample efficiency in capturing clinically meaningful annotation diversity.

3. Uncertainty vs. Prediction Correctness

We assess whether the model's pixelwise predictive uncertainty is indicative of prediction correctness, following the evaluation protocol from [2]. For each test image, we generate $K = 16$ stochastic segmentations from the learned prior and compute the mean per-pixel foreground probability $\bar{p}(x)$ across samples. Uncertainty is quantified as normalized binary entropy ($H(\hat{p})$)of the predictive mean:

$$H(\hat{p})(x) = -\,\bar{p}(x)\log_2 \bar{p}(x) - (1-\bar{p}(x))\log_2(1-\bar{p}(x)),$$

which yields values in the range $[0, 1]$.

To establish ground truth in the presence of multiple annotators, we average per-pixel rater masks to form a consensus map, and retain only those pixels with unanimous agreement (i.e., consensus label exactly 0 or 1), filtering out ambiguous regions unless noted otherwise. Thresholding $\bar{p}(x)$ at 0.5, we classify each retained pixel into one of four correctness categories: True Positive (TP), False Positive (FP), False Negative (FN), or True Negative (TN).

Figure 2 presents a scatter plot of entropy values grouped by TP/FP/FN/TN classification. Each dot represents an individual pixel, and per-group medians are highlighted with circled markers. As expected for a well-calibrated model, TN and TP pixels exhibit near-zero uncertainty, while FP and FN pixels cluster around high entropy, with medians approaching 1.0. This sharp separation between correct and incorrect predictions confirms that the model is confident when accurate, and uncertain when wrong, an essential property for reliable downstream usage.

Interestingly, a small subset of true positives shows elevated entropy. These typically lie near lesion boundaries or within low-contrast regions, where even correct predictions tend to be probabilistic, yielding $\bar{p}(x) \approx 0.5$ and thus high entropy. In contrast, TN pixels, corresponding to background, consistently show minimal uncertainty, which can be attributed to their spatial homogeneity, larger area coverage, and stronger rater agreement.

This pattern persists even in the presence of noisy boundaries: although some entropy values are elevated near object edges due to genuine inter-rater disagreement, the group-wise medians remain cleanly separated. This suggests that uncertainty is not uniformly inflated but is instead tightly coupled to the model's confidence in its predictions.

In practical terms, this behavior enables automatic pixel- or region-level triage. High-uncertainty areas align with prediction errors (FP/FN), making them candidates for clinician review or further processing, while low-uncertainty regions (TP/TN) indicate reliable predictions. The observed asymmetry, where some TPs have high uncertainty but TNs rarely do, is consistent with the nature of the task: foreground lesions are more variable and subject to subjective interpretation, while background regions are more stable.

Hence, the model's entropy-based uncertainty estimates are both diverse and diagnostic. They not only reflect inter-rater ambiguity, but also serve as a practical signal for identifying erroneous predictions, reinforcing their interpretability and utility in multi-rater medical segmentation workflows.

4. Size-Stratified Robustness

To examine whether our model maintains consistent performance across lesions of different sizes, we perform a size-stratified evaluation, motivated by the well-known challenge that small lesions are typically harder to segment due to their limited spatial extent, while large lesions may introduce structural variability. A model that overfits to a dominant lesion size in the training distribution could fail to generalize across the full spectrum of cases seen in practice. This ablation explicitly tests for such scale-related biases.

For each test image, we compute a consensus mask by averaging expert annotations and thresholding at 0.5 to obtain a binary foreground region. Lesion size is then defined

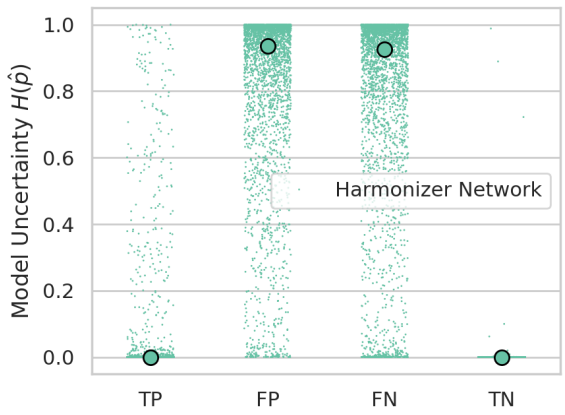


Figure 2. Pixelwise uncertainty by correctness (TP/FP/FN/TN). Dots are individual pixels; medians are circled.

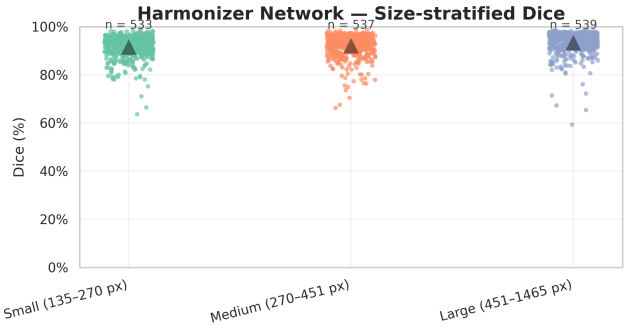


Figure 3. **Size-stratified Dice.** Each dot represents a single test case, and large triangles mark the mean Dice per bin. *n* indicates the number of cases per bin. Similar means and variances across bins confirm that model performance is consistent across lesion sizes.

as the pixel count of the resulting foreground mask. Using size quantiles computed on the test set, we partition all lesions into three equally populated bins: *Small* (135–270 px), *Medium* (270–451 px), and *Large* (451–1465 px), with ranges shown on the x-axis of Figure 3.

Within each bin, we compute the Dice similarity between the model's predicted mask and the consensus annotation. Figure 3 plots all per-case Dice scores as individual points (with horizontal jitter for visibility), while large colored triangles represent the mean Dice within each bin. The total number of test cases in each group is annotated above the plots.

The results show high and consistent Dice scores across all three bins, with overlapping distributions and tightly clustered means. Importantly, the model performs robustly on small lesions, which are often more ambiguous due to faint boundaries, low contrast, and lower inter-rater agreement. The fact that performance does not drop for these more challenging cases suggests that our method is not biased toward larger or more prominent regions, and is capable of reliably segmenting lesions regardless of their scale.

This ablation confirms that our model demonstrates strong scale-invariance in segmentation accuracy, maintaining stable performance from small to large lesions. This robustness is critical for clinical applicability, where lesion size varies widely between patients and imaging scenarios.

5. Model Uncertainty vs. Rater Uncertainty

A key question in uncertainty modeling is whether the model's predictive uncertainty arises from genuine ambiguity in expert annotations, or whether it simply reflects noise or internal instability. To address this, we examine how model uncertainty correlates with rater uncertainty, as measured by inter-expert agreement on the test set. The goal is to verify whether high model entropy coincides with regions where annotators themselves disagree, suggesting that uncertainty is interpretable and human-aligned rather than random.

For each test image, we sample K stochastic predictions from the learned prior and compute pixelwise model uncertainty using the entropy formulation defined in Eq. 3, where $\bar{p}(x)$ denotes the mean per-pixel foreground probability, with entropy values normalized to the range $[0, 1]$.

We then compute per-pixel rater uncertainty from the mean annotation probability p_A, obtained by averaging binary masks from all experts. Based on p_A, each pixel is assigned to one of three rater agreement regimes: agree ($p_A \approx 0$ or 1), somewhat agree (intermediate values away from 0.5), and disagree ($p_A \approx 0.5$), using small tolerance margins around the endpoints and midpoint.

Figure 4 shows the distribution of model uncertainty $H(\hat{p})$ across these categories. Points represent individual pixels (subsampled for clarity), and group medians are marked by large circles. The results reveal a clear monotonic trend:

- In regions where annotators agree, the model expresses low uncertainty, with entropy collapsing near zero.
- In regions of partial agreement, uncertainty rises moderately, indicating cautious predictions.
- In regions where annotators disagree, model uncertainty is highest, reflecting maximum ambiguity.

This behavior demonstrates that the model is not only well-calibrated but also sensitive to human disagreement. Its uncertainty estimates increase precisely in regions where expert interpretation diverges, rather than being uniformly elevated or erratic. This alignment between model and rater uncertainty is critical for practical deployment: it enhances interpretability, facilitates automatic triage of ambiguous re-

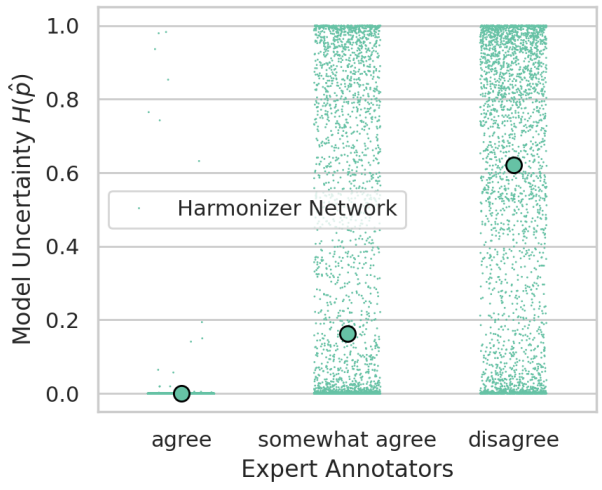


Figure 4. **Model uncertainty vs. rater agreement.** Pixelwise uncertainty $H(\hat{p})$ grouped by rater consensus: agree, somewhat agree, and disagree. Each dot represents a sampled pixel; large circles denote medians. Uncertainty increases smoothly with annotation ambiguity, indicating that the model's confidence is aligned with inter-rater variability.

gions, and reinforces trust in the model's outputs.

Moreover, these results complement distributional evaluation metrics such as GED by showing that the model's uncertainty is not just diverse, but also context-aware. It can differentiate between confident and uncertain predictions in a way that reflects human annotation variability, making it a meaningful tool for decision support and quality control in multi-rater clinical environments.

6. Per-Rater Calibration: ECE and Brier Score

While Dice and GED quantify overlap-based segmentation accuracy, an equally important property of multi-rater probabilistic models is the *calibration* of their predicted probabilities. A well-calibrated model ensures that predicted confidences, such as 0.3 or 0.8, accurately reflect the empirical frequency of foreground pixels. To evaluate this property, we compute the Expected Calibration Error (ECE) [4] and Brier score [10] for each personalized rater head on the LIDC dataset.

For each rater A_r, the personalized head produces a foreground probability map $p_r(x) \in [0,1]$, which is evaluated against that rater's binary annotation $y_r(x) \in \{0,1\}$. All pixels from all cross-validation folds are pooled together and grouped into $K = 10$ equal-width bins $\{B_k\}_{k=1}^{K}$ spanning the probability range $[0,1]$. For each bin B_k, we compute the mean predicted confidence $\mathrm{conf}(B_k)$ and the corresponding empirical foreground frequency $\mathrm{acc}(B_k)$.

The Expected Calibration Error aggregates deviations between confidence and accuracy as

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|, \tag{1}$$

where N denotes the total number of evaluated pixels. An ECE value close to zero indicates that the model's predicted probabilities faithfully correspond to empirical frequencies. Complementarily, the Brier score measures the mean squared error between predicted probabilities and binary labels:

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} \left(p_i - y_i\right)^2, \tag{2}$$

where lower values correspond to more accurate and better-calibrated probabilistic estimates.

Table 1. Per-rater calibration on LIDC (Personalized). Lower is better for both metrics.

Rater	ECE ↓	Brier ↓
A1	0.003238	0.003440
A2	0.003166	0.003449
A3	0.003260	0.003592
A4	0.005175	0.005360
Mean	0.00371	0.00396

Table 1 reports per-rater calibration metrics, while Figure 5 visualizes both ECE and Brier score jointly. All raters achieve extremely low ECE values (3×10^{-3}–5×10^{-3}) and Brier scores below 6×10^{-3}, demonstrating strong probabilistic reliability across annotators. Raters A1–A3 exhibit nearly identical calibration behavior, indicating that the model captures their annotation styles consistently. In contrast, Rater A4 shows slightly higher error, which aligns with this rater being the most challenging (e.g., exhibiting sharper boundaries or more aggressive contouring), yet the absolute differences remain very small.

Overall, these results confirm that our Harmonized Personalization framework not only reproduces individual rater styles but also yields well-calibrated and trustworthy probability estimation.

7. Noise Effects

To evaluate robustness under image degradation, we simulate common acquisition artifacts directly on the LIDC–IDRI test set during inference. Three categories of perturbations are introduced: (i) additive Gaussian noise at three intensity levels ($\sigma_1 = 0.1$, $\sigma_2 = 0.15$, $\sigma_3 = 0.25$); (ii) Gaussian blurring with kernel size 115 and increasing standard deviations ($\sigma_1 = 1.0$, $\sigma_2 = 1.5$, $\sigma_3 = 1.75$); and

Table 2. Robustness comparison of probabilistic and personalized segmentation methods under synthetic degradations on the LIDC–IDRI dataset. Identical noise configurations are used across models. K denotes the Gaussian blur kernel size. $|\Delta|$ indicates Dice drop relative to the clean image baseline.

Method	Gaussian Noise						Gaussian Blur ($K = 115$)						Gamma Jittering					
	$\sigma = 0.10$		$\sigma = 0.15$		$\sigma = 0.25$		$\sigma = 1.0$		$\sigma = 1.5$		$\sigma = 1.75$		$\gamma = 0.75$		$\gamma = 1.5$		$\gamma = 1.75$	
	DSC↑	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$	**DSC↑**	$\lvert\Delta\rvert\downarrow$
Prob. U–Net [7]	85.43	3.66	81.76	7.32	73.22	15.87	87.75	1.33	86.65	2.43	85.86	3.23	86.48	2.60	79.92	9.17	79.03	10.06
D–Persona [14]	85.74	3.43	81.80	7.37	71.11	18.06	88.54	0.63	87.65	1.51	86.91	2.25	87.66	1.51	81.85	7.31	81.53	7.64
Harmonizer (ours)	**89.37**	**1.44**	**87.92**	**2.88**	**84.27**	**6.53**	**90.60**	**0.22**	**90.25**	**0.56**	**89.76**	**1.04**	**90.14**	**0.67**	**86.75**	**4.06**	**86.23**	**4.58**

(iii) photometric distortions via gamma jittering ($\gamma_1 = 0.75$, $\gamma_2 = 1.5$, $\gamma_3 = 1.75$), as summarized in Table 2. Using identical noise settings across methods ensures a fair comparison. We evaluate the Harmonizer Network against the probabilistic U-Net [7] and the D-Persona [14] to assess degradation resilience. Across all noise regimes, our method exhibits markedly smaller performance drops in both GED and Dice metrics. Under the strongest Gaussian noise (σ=0.25), the Harmonizer retains more than 95% of its clean–data Dice, whereas D–Persona shows a steeper decline. Similar stability is observed under blurring and jitter perturbations, confirming that our harmonization mechanism effectively suppresses low–frequency degradation and preserves structural coherence in the latent space.

Figure 6 presents a representative example comparing the four expert annotations (red contours), the model prediction on the unperturbed slice (Clean), and predictions obtained under progressively stronger perturbations. Blue/purple contours denote the model's outputs, while the rightmost column shows the fixed Gamma map that captures inter-rater uncertainty. Across Gaussian noise conditions (G0.100–G0.200), the model demonstrates remarkable resilience. Even under severe pixel-level corruption, the predicted boundaries remain closely aligned with the raters' annotations and preserve the overall nodule geometry. The contours do not collapse or drift, instead reflecting a stable consensus shape that balances variations among raters. For instance, when Rater 1 delineates sharper edges while Rater 3 marks smoother boundaries, the model yields an intermediate contour that remains anatomically plausible. This shows that harmonization guides the model toward rater-consistent latent structure rather than superficial textures, effectively suppressing high-frequency noise.

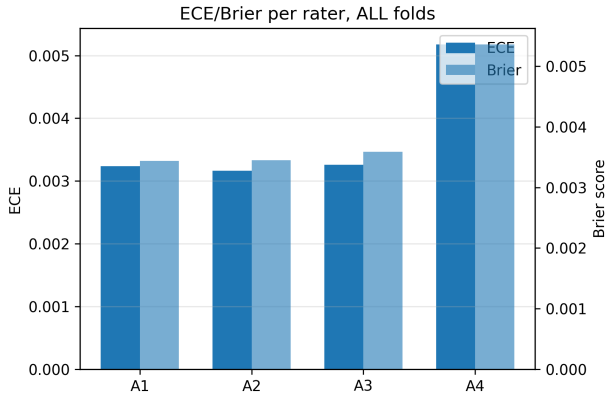


Figure 5. ECE and Brier score per rater, aggregated across all LIDC folds. Bar heights correspond to the numerical values reported in Table 1.

Under intensity jitter (J0.300–J0.500), the model again shows strong invariance. Global brightness and contrast shifts do not affect segmentation alignment; even at higher jitter levels, delineations remain consistent and faithful to the raters. In ambiguous regions, predictions gently interpolate between annotators rather than fluctuating, suggesting that the model prioritizes structural and contextual cues over raw pixel intensities, critical for robustness to scanner variability and acquisition differences.

The Gaussian blur perturbations (B7–B11) reveal the most visually challenging condition. As blur increases, edges become less defined and rater discrepancies more visible: Rater 1 and Rater 4 tend to preserve sharper margins, while Rater 2 and Rater 3 provide smoother contours. The model remains stable across these variations, producing an averaged contour that reflects consensus structure. Under extreme blur (B11–2.50), where the lesion boundary nearly disappears, the prediction contracts slightly toward the core region that all raters agree upon. This behavior is semantically meaningful, rather than failing arbitrarily, the model defaults to the most confident portion of the lesion.

As the Gamma map stays nearly constant across noise types, it serves as a stable reference for uncertainty. Regions that shift under strong blur align with high-uncertainty areas in the Gamma map, while consistent regions remain unaffected. This shows that the model's sensitivity reflects human disagreement rather than noise artifacts.

8. Frequency-Domain Visualization

The main purpose of the frequency adapter is to strengthen high-frequency information, enabling the model to better exploit boundary-related and textural cues during personalization. To verify that the module performs this enhancement, we visualize the spectral response of the decoder's final feature map using a 1D FFT along the central spa-

tial axis, following the procedure in [1]. Figure 7 compares the baseline model with our frequency-calibrated one. Without the adapter, the spectrum is dominated by low-frequency components, indicating limited usage of fine details. After applying the adapter, the high-frequency energy increases substantially, while the low-frequency structure remains stable. This confirms that the module effectively amplifies informative high-frequency signals, enriching fine-scale structure in the feature space and enabling sharper, more personalized predictions.

9. Latent Space Diversity

A fundamental property of probabilistic segmentation models is their ability to represent multiple plausible outputs for a given image. To assess whether our model's latent space captures genuine inter-rater diversity rather than random variability, we analyze its generative behavior by sampling multiple predictions from the learned prior distribution. Figure 8 and Figure 9 present representative examples of sampled segmentation masks on the LIDC and NPC-170 datasets, respectively. Each row corresponds to a single input image, while the columns depict diverse segmentations generated by different latent codes.

These examples demonstrate that the latent space learned by our method is both structured and diverse. Each sample remains anatomically valid while exhibiting subtle yet meaningful variations, particularly along boundary regions where expert annotations typically diverge. This diversity indicates that the model does not collapse to a single deterministic solution but instead preserves multiple plausible hypotheses consistent with the empirical annotation distribution. The frequency-domain prompts introduced in our framework play a key role in this property by modulating harmonized features within a rater-aware spectral space. Consequently, the model maintains controlled variability concentrated in clinically ambiguous regions rather than distributing it arbitrarily across the image.

To quantify latent diversity, we also evaluate the distributional coverage using GED as a function of the number of generated samples on the NPC-170 dataset (Figure 10). We observe a consistent reduction in GED, with a simultaneous increase in soft-Dice as the number of samples grows, confirming that the model efficiently spans the annotation manifold. Importantly, the improvements saturate beyond approximately 30 samples, suggesting that the latent space is compact yet expressive enough to capture the full range of expert annotations without redundancy. These findings provide strong evidence that the learned latent representation encodes realistic structural uncertainty and enables faithful multi-rater modeling through controlled stochastic sampling.

10. Extended Results on Kvasir Dataset

We further evaluate our approach on the noisy medical image segmentation dataset Kvasir [6], which presents varying annotation challenges. As shown in Table 3, our method achieves the best overall performance on the Kvasir dataset.

Dataset and Experimental Setup

The Kvasir-SEG dataset [6] contains 1,000 gastrointestinal endoscopy images, each paired with an expert-annotated polyp segmentation mask. The images exhibit substantial variability in polyp size, morphology, and visual appearance, and are acquired under heterogeneous illumination conditions. Following the noise-injection protocol of [11], we simulate annotation noise and inter-rater disagreement by generating three additional mask variants for each training image: S_R (Simulated Random noise, $\sigma = 0.2$), S_E (Simulated Extreme noise, $\sigma = 0.8$), and S_{DE} (Simulated Diverse Expert noise). This yields four masks per training sample. The test set uses only clean expert annotations to ensure unbiased evaluation. In our experiments, we follow [11] and resize images to 256×256 and normalize them to $[0, 1]$.

Table 3. Performance comparison on Kvasir dataset

Method	Kvasir dataset		
	S_R	S_E	S_{DE}
RCE Loss [12]	73.51±1.58	73.68±1.57	66.17±1.97
RMD [3]	68.34±2.18	71.57±1.09	66.90±1.75
ADELE [9]	60.97±14.78	67.10±11.42	60.62±13.04
CDR [15]	67.87±3.51	70.58±1.65	63.51±1.67
Co-Teaching [5]	74.26±1.71	75.57±1.12	74.03±1.48
IDMPS [16]	77.52±1.21	74.16±0.81	69.47±0.81
JoCoR [13]	67.98±3.23	71.43±1.37	65.62±1.94
SP-Guide [8]	69.24±2.09	62.97±1.65	61.74±1.56
GSD-Net [11]	80.04±0.42	79.39±0.33	**79.97±0.53**
D-Persona [14]	84.69±0.19	81.77±0.16	78.93±0.18
Harmonizer Network	**85.13±0.17**	**82.96±0.15**	78.89±0.18

Results

Table 3 presents a comprehensive comparison of our Harmonizer Network against baseline methods on the Kvasir dataset across three noise conditions. Our Harmonizer Network achieves the highest overall performance across all noise conditions, with Dice scores of 85.13%, 82.96%, and 78.89% for S_R, S_E, and S_{DE}, respectively. The probabilistic method D-Persona [14] ranks second with scores of 84.69%, 81.77%, and 78.93% for S_R, S_E, and S_{DE}, while the GSD-Net [11] achieves third place with consistent performance around 80% across all conditions. Compared to D-Persona, our method shows modest but consistent improvements of +0.44% on S_R, +1.19% on S_E, and

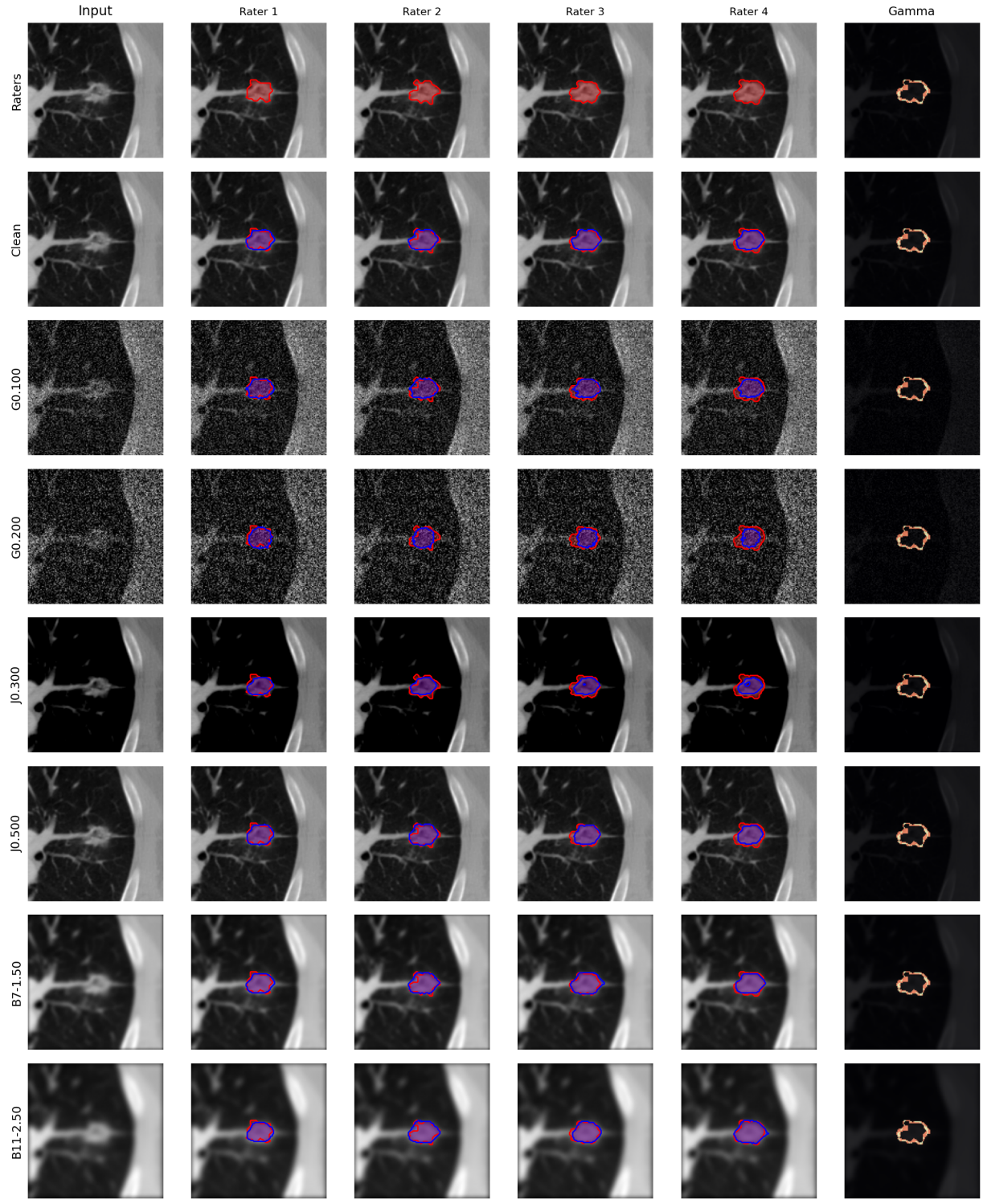
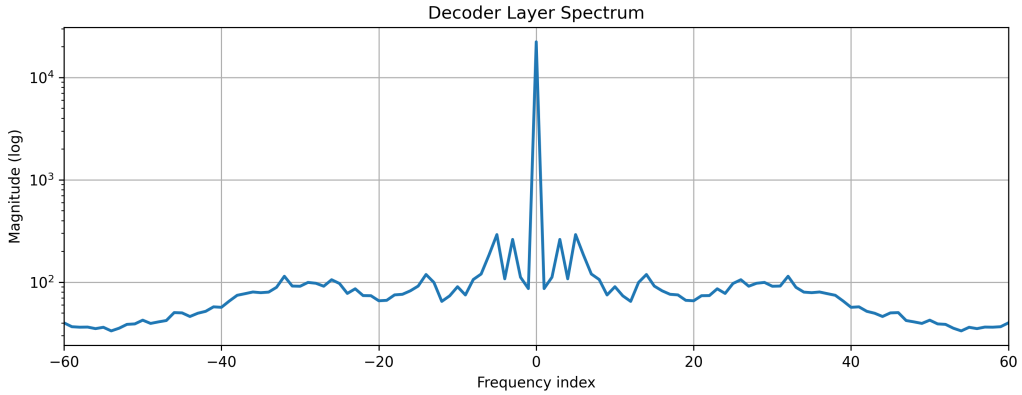
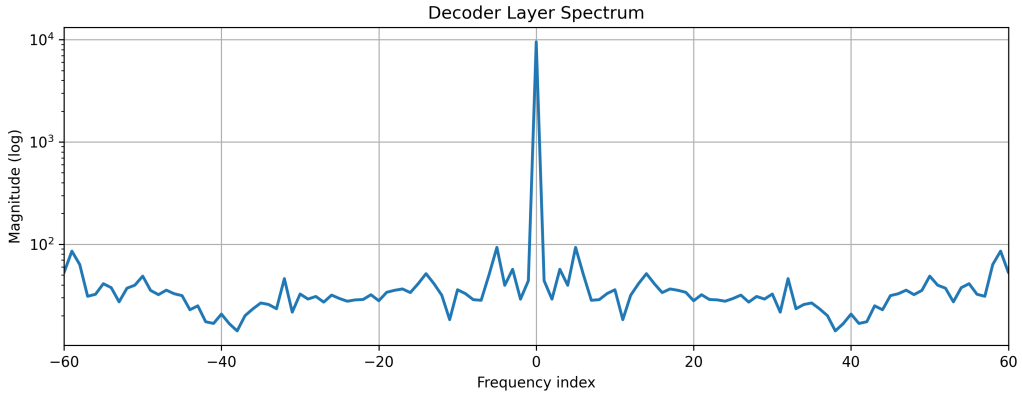


Figure 6. **Qualitative robustness analysis under noise perturbations.** Row labels indicate the applied corruption type and magnitude: **G·** denotes Gaussian noise with standard deviation values; **J·** indicates brightness/contrast jitter; and **B·** represents Gaussian blur, where the first and second rows use kernel sizes of 7 and 11, respectively. Red contours show ground-truth rater segmentations, blue/purple contours show model predictions, and the rightmost column displays the **Gamma map**, visualizing inter-rater uncertainty. The model preserves structure under all perturbations and degrades meaningfully in regions of high rater ambiguity.



(a) Baseline decoder (without frequency module).



(b) With frequency adapter (ours).

Figure 7. Spectral response comparison between the baseline decoder and our frequency-calibrated decoder. The plots show the 1D FFT magnitude (log-scale) along the central spatial axis. The adapter increases high-frequency activation while preserving stable low-frequency structure.

-0.04% on S_{DE}. While D-Persona slightly outperforms our method on diverse expert noise, our approach shows superior robustness to extreme noise conditions, suggesting better handling of severe annotation errors through the harmonization strategy. Our method outperforms GSD-Net by +5.09% and +3.57% on random and extreme noise conditions, respectively, while trailing by -1.08% on diverse expert noise. The larger gains on random and extreme noise indicate that our approach more effectively filters noisy labels compared to traditional co-teaching strategies. Co-Teaching [5] achieves 74.26–75.57%, while JoCoR [13] reaches only 65.62–71.43%, both of which fall significantly behind our method. Loss-based approaches show mixed results, with RCE Loss [12] achieving moderate performance (66.17–73.68%), while RMD [3] and CDR [15] perform similarly (63.51–71.57%). These methods lag behind our approach by 11–19 percentage points, highlighting the limitations of purely loss-based noise handling without explicit multi-annotator modeling. IDMPS [16] shows strong performance on S_R (77.52%) but degrades on S_E (74.16%) and S_{DE} (69.47%). SP-Guide [8], performs poorly on extreme noise (62.97%), suggesting that hand-crafted priors struggle with severe annotation errors. Our method demonstrates superior stability with low standard deviations (±0.15–0.18) compared to most baselines, particularly ADELE (±11–15%), CDR (±1.65–3.51%), and JoCoR (±1.37–3.23%), indicating that our harmonization approach provides more consistent training dynamics across different noise realizations. Additionally, we provide visual comparisons between our model's outputs and other baseline methods in Figure 11, qualitatively illustrating the superior segmentation quality and boundary accuracy achieved by our approach across different noise conditions.

11. Model Complexity and Computational Efficiency

Although the proposed framework integrates both harmonization and personalization mechanisms, it remains lightweight and computationally efficient. The full model contains 30.31 M parameters (baseline 30.11 M, $\mathcal{H}_\phi^{(n)}$

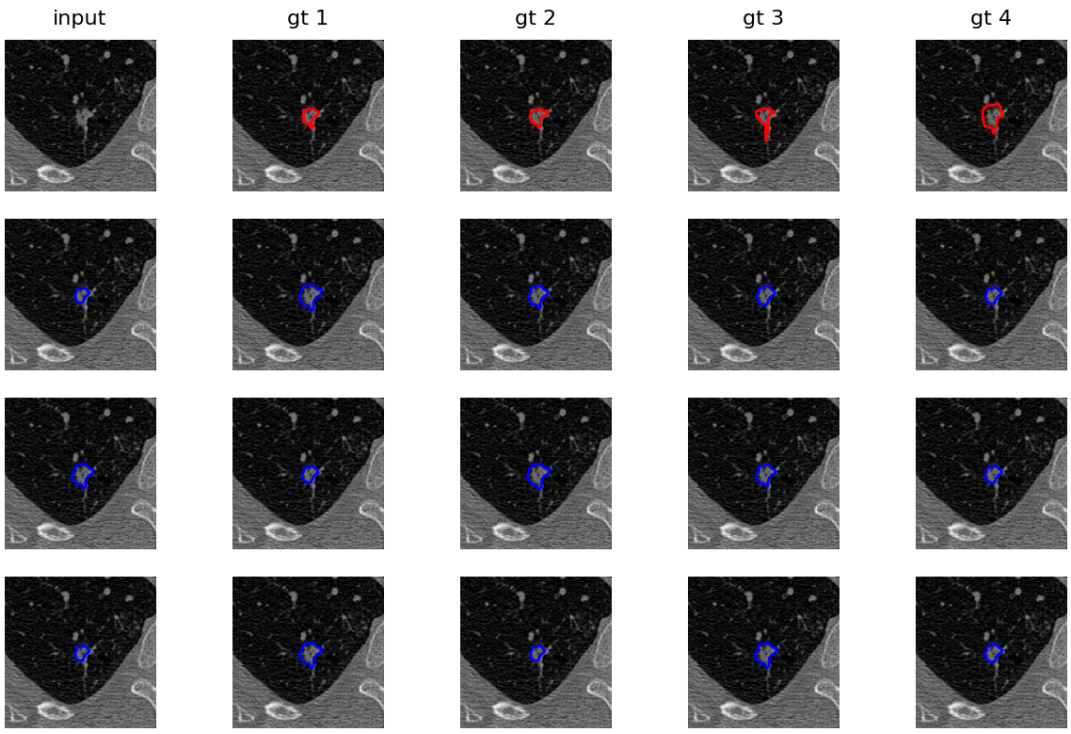


Figure 8. Diverse segmentation hypotheses generated by sampling from the latent space on the LIDC dataset. Each column represents a stochastic sample, illustrating structured diversity along ambiguous boundaries.

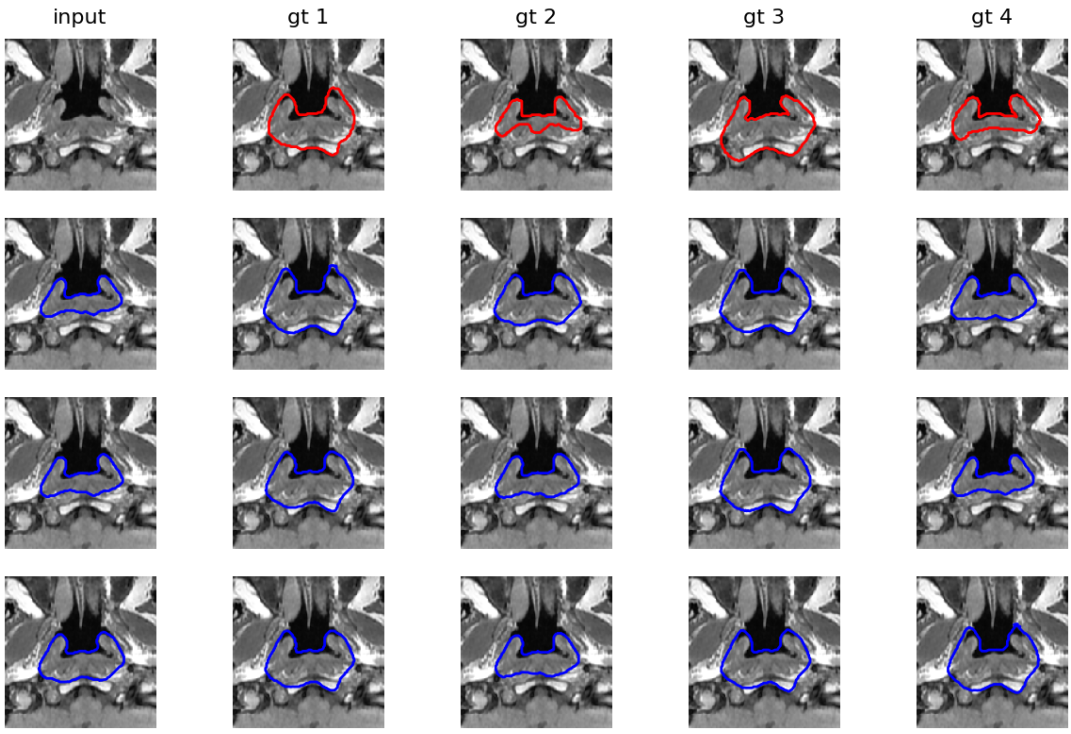


Figure 9. Diverse segmentation hypotheses generated by sampling from the latent space on the NPC-170 dataset. Each column represents a stochastic sample, illustrating structured diversity along ambiguous boundaries.

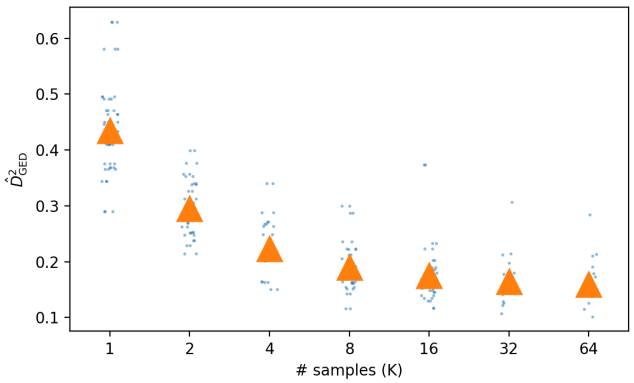


Figure 10. Impact of the proposed GED loss on distributional coverage. Plot shows GED as a function of the number of generated samples K (lower is better). Models trained with the GED loss achieve consistently lower discrepancy and reach convergence with fewer samples, indicating more efficient and faithful coverage of the underlying annotation distribution.

0.14 M, and $\mathcal{H}_\phi^{(p)}$ 0.07 M), including both harmonization and personalization modules, and requires around 0.42 s per forward pass during inference. We estimate the computational complexity using multiply–accumulate operations (MACs) on 128×128 input images. The full Harmonizer Network requires approximately 2.4 GFLOPs per forward pass. This marginal cost demonstrates that the added modules perform lightweight modulation rather than heavy computational processing. During inference, the model runs at approximately 2.3 frames per second on a single NVIDIA RTX 3090 GPU, which includes latent sampling and personalization. Training employs mixed-precision arithmetic, further reducing memory consumption and accelerating convergence. In fact our design remains significantly efficient while offering higher fidelity and interpretability. The harmonizer's shared weights across layers and the low-dimensional prompt tokens ensure scalability without exponential growth in computational cost. This makes the model suitable for deployment in real-world medical imaging settings, where computational resources and latency constraints are critical considerations. The lightweight design underscores that harmonized probabilistic modeling, along with frequency-domain personalization, can be achieved efficiently without sacrificing performance or uncertainty quality.

11.1. Two-stage training/inference pseudo-code

This section summarizes the overall optimization and inference pipeline of the proposed framework. The pseudo-code is organized into two consecutive training phases to clearly separate the learning of data-level harmonization from the learning of rater-specific personalization. In the first phase, the encoder E, decoder D, and Noise Harmonizer H are jointly optimized so that the network can reconstruct segmentations while learning artifact-reduced and stable feature representations. This stage focuses on capturing the underlying anatomical content and the global ambiguity of the task without yet specializing to individual annotators. In the second phase, these backbone components are frozen and only the Personalizer P is trained. This design ensures that rater-specific adaptation is learned on top of already harmonized features, preventing scanner noise or acquisition bias from being mixed with annotator-dependent variation. During inference, the same harmonized representation is first extracted, and then the personalizer conditions the latent code for a selected rater to generate the corresponding individualized segmentation.

Pseudocode

Phase 1 (train E, D, H):
$f = E(x)$, $z \sim q(z|x, y)$, $\hat{f} = H(f)$, $\hat{y} = D(\hat{f}, z)$;
optimize $L_{\text{ELBO}} + \lambda_{\text{GED}} L_{\text{GED}}$. Freeze E, D, H.

Phase 2 (train P only):
$\hat{f} = H(E(x))$; for each rater r: $z \sim p(z|x)$, $z' = \mathcal{P}(z, \hat{f}, r)$, $\hat{y}^r = D(\hat{f}, z')$;
optimize $\sum_{r=1}^{n} \mathbb{E}_{z \sim p(z|x)}[L_{\text{seg}}(\hat{y}^r, y^r)]$ (Eq. 6).
[z: prior latent, z': rater-adaptive latent]

Inference: $\hat{f} = H(E(x))$, $z \sim p(z|x)$, $z' = \mathcal{P}(z, \hat{f}, r)$, $\hat{y}^r = D(\hat{f}, z')$.

The pseudo-code highlights the functional role of each module in the proposed training strategy. In Phase 1, the latent variable is sampled from the posterior distribution $q(z|x, y)$, allowing the model to learn a structured representation of segmentation uncertainty while the harmonizer refines encoder features into a more stable and artifact-suppressed form. The joint objective combines the ELBO loss with the GED term so that the model not only reconstructs plausible outputs but also captures distributional diversity across annotations. Once this representation is learned, E, D, and H are frozen. In Phase 2, the prior latent $z \sim p(z|x)$ is adapted by the personalizer into a rater-aware latent code z', which injects annotator-specific information without altering the shared anatomical backbone. This makes the second stage focused and stable, since only the personalization branch is optimized. At inference time, the framework follows the same decomposition: harmonized features are first extracted, then the latent sample is personalized for a chosen rater, and finally the decoder generates the corresponding segmentation. In this way, the model preserves both common anatomical structure and individualized annotation style in a controlled manner.

12. Additional Highlights

Importance of Each Module

The proposed framework integrates two purpose-built modules, the Noise Harmonizer and the Frequency-Prompt Personalizer, each addressing a fundamentally different challenge in the multi-rater segmentation setting. Importantly,

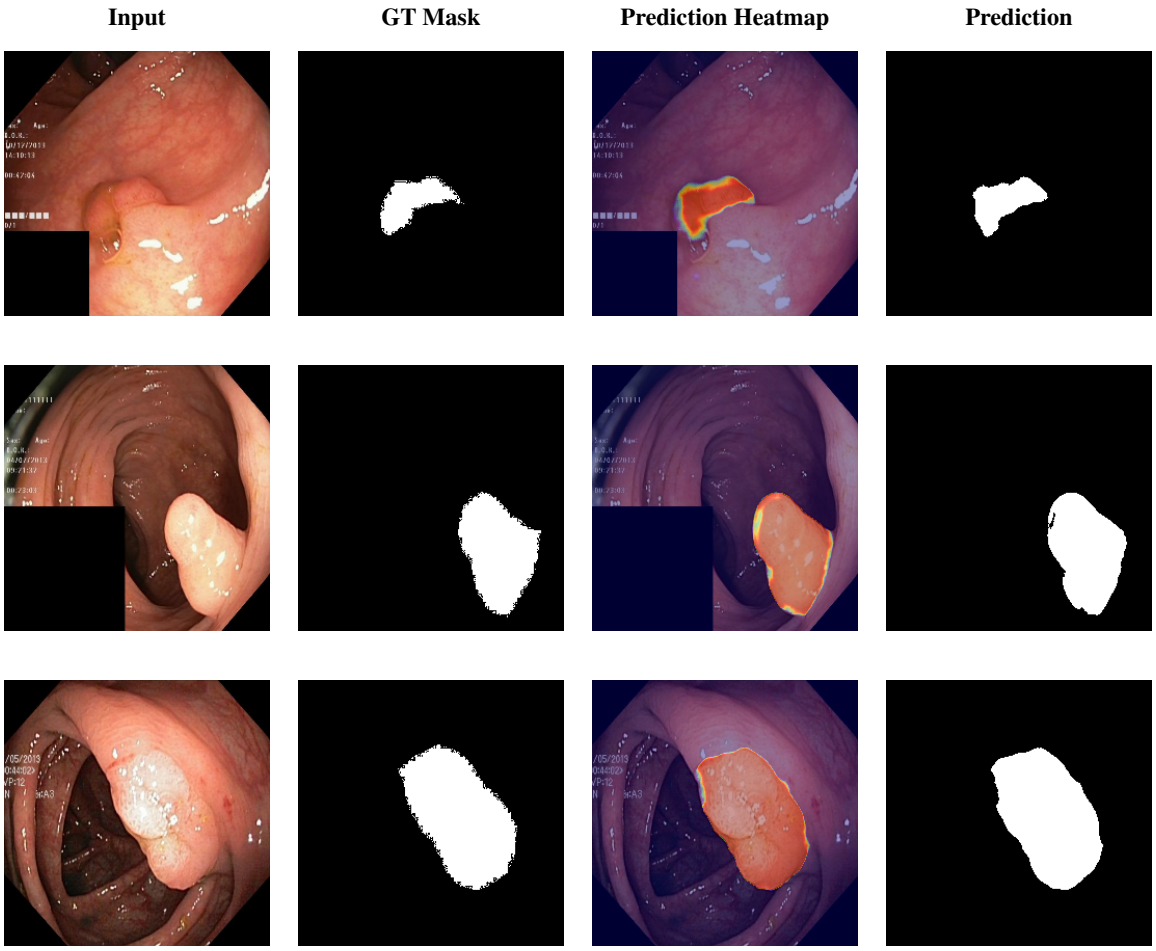


Figure 11. Visual comparison of segmentation results on the Kvasir dataset, showing input images, ground truth masks, prediction heatmaps, and final predictions for three representative samples.

both components are designed not to artificially inflate accuracy on clean datasets, but to explicitly resolve the core sources of disagreement and inconsistency inherent in multi-annotator data.

The Noise Harmonizer regulates high-frequency corruption and scanner-induced variability, ensuring that the latent space remains stable even under significant perturbations. As demonstrated in the robustness experiments on Gaussian noise, blur, and jitter (Section 7), this module substantially improves resilience to degradation. Under the strongest corruption level ($\sigma = 0.25$), the Harmonizer retains over 95% of its clean-data Dice, outperforming D-Persona and Probabilistic U-Net by large margins. Conversely, removing this module on clean LIDC/NPC-170 samples yields only marginal performance changes, consistent with its intended role as a stabilizer rather than a direct accuracy booster.

The Frequency-Prompt Personalizer provides rater-specific adaptation by learning localized spectral modulations that reproduce annotators' stylistic tendencies without overwhelming the shared anatomical manifold. Section 5 shows that its effect is spatially precise, primarily enhancing high-frequency energy around ambiguous boundaries. Compared with the projection heads of D-Persona, the proposed personalizer consistently yields lower GED, higher soft-Dice, and stronger rater-aligned calibration (Sections 2 and 6). Across NPC-170 and LIDC, these improvements translate into observable gains in segmentation fidelity and more coherent personalized outputs.

Together, the two modules offer complementary gains: one stabilizes the latent representation under real-world noise, while the other enriches stylistic personalization.

Comparison with Consensus-Based Methods

Consensus-based or label-fusion methods operate by collapsing multiple annotations into a single target, often through majority voting, STAPLE, or random sampling. As highlighted in prior work (e.g., D-Persona), such approaches inevitably discard meaningful rater-specific vari-

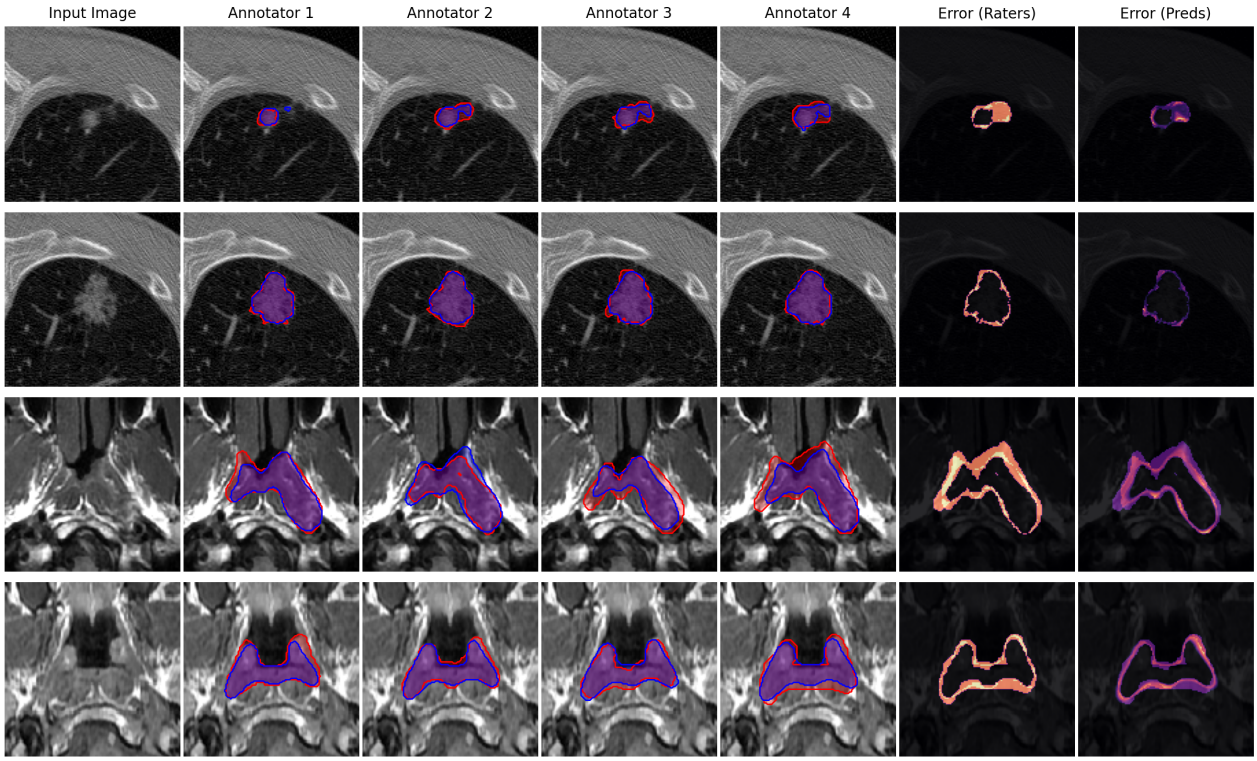


Figure 12. Visualization results on the LIDC (first-second rows) and NPC-170 (third-fourth rows) dataset with multi-rater annotations and the corresponding error map. The red boundary indicates the ground truth, while the blue boundary represents the predicted mask.

ations and tend to underperform in settings with genuine annotation disagreement. Our empirical findings are consistent with this trend [14].

Additional Visualization

To further demonstrate the qualitative behavior of our model, we provide additional visualizations of segmentation results on both the LIDC-IDRI and NPC-170 datasets in Figure 12. These visual samples highlight how the model effectively adapts to heterogeneous annotator styles while preserving structural consistency across multiple raters. In both datasets, we observe that the proposed Harmonizer Network produces highly coherent boundaries that align well with the ground-truth masks, even when annotations diverge. The personalized predictions successfully replicate subtle stylistic tendencies of individual annotators, such as sharper contour delineation or smoother regional filling, demonstrating the model's ability to internalize each rater's segmentation characteristics.

The accompanying error maps visualize regions of disagreement between predicted and annotated masks. Red boundaries indicate ground-truth annotations, whereas blue lines denote predicted segmentations. Areas with overlapping red-blue contours correspond to high agreement, while isolated colored edges represent zones of residual ambiguity. These differences often coincide with clinically uncertain regions, low-contrast lesion edges, weak tissue boundaries, or ambiguous nodular margins where experts themselves disagree. The proposed model allocates uncertainty precisely in such areas, confirming that its confidence modulation and personalized adaptation reflect genuine inter-rater ambiguity rather than random noise. These results further demonstrate that the method maintains both personalization and structural consistency across multiple annotation distributions, effectively capturing the underlying anatomical manifold while respecting individual rater styles.

13. Acquisition-Domain Shifts

To directly address acquisition domain shift, we conducted a new experiment using the manufacturer metadata provided in the LIDC-IDRI dataset. We partitioned the LIDC-IDRI dataset based on scanner manufacturer to create a realistic cross-domain evaluation scenario: *Training/Validation Set*: All samples from GE MEDICAL SYSTEMS, TOSHIBA, and Philips manufacturers *Test Set*: All samples acquired exclusively from SIEMENS scanners

Table 4 demonstrates that our method exhibits superior robustness to real acquisition-domain shifts compared to D-

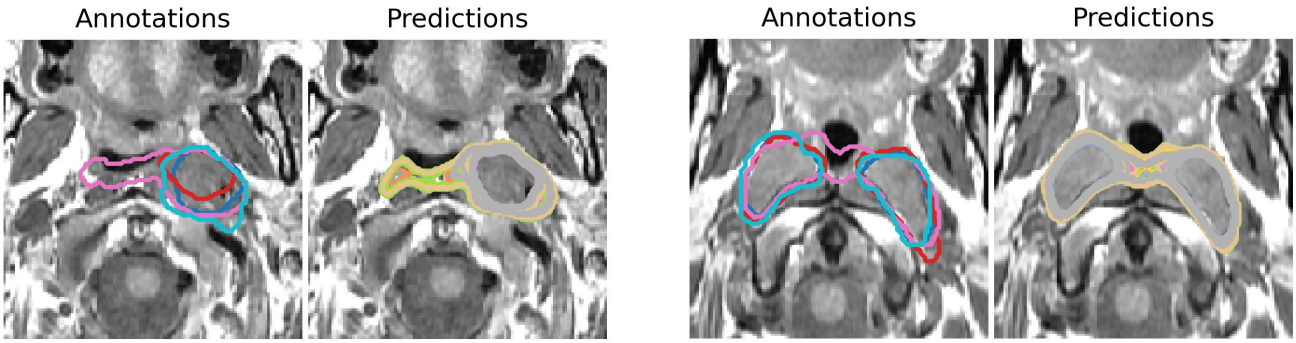


Figure 13. **Representative failure cases.** *Left:* One annotator marks a much larger region than others; the model detects similar patterns but is constrained by rater-specific regularization, yielding a truncated prediction. *Right:* Two symmetric structures lead to inconsistent rater choices. The model captures both hypotheses but cannot fully disambiguate them. Residual errors thus stem from annotation inconsistency rather than model instability.

Table 4. Personalized segmentation performance on LIDC–IDRI under different train/test scaners configuration.

Method	Train/Test All → All		Train/Test All (except Siemens) → Siemens	
	DSC ↑	|Δ| ↓	DSC ↑	|Δ| ↓
D-Persona	89.17	-	83.02	6.15
Harmonizer Network	**90.78**	-	**85.30**	**5.48**

persona. This performance gap along with synthetic perturbation experiments provides direct empirical evidence that our Noise Harmonizer is not merely acting as a general feature regularizer, but is specifically modeling and mitigating artifacts. Moreover, the Noise Harmonizer uses weight sharing across all decoder layers, which enforces consistent artifact-suppression behavior across multiple spatial resolutions. Acquisition artifacts such as intensity drift, reconstruction noise, and motion blur tend to exhibit scale-consistent patterns, whereas rater variability is typically localized around object boundaries and varies spatially across cases. By learning shared, scale-consistent modulation parameters (γ_l, β_l), the harmonizer is biased toward capturing global, low-to-mid frequency perturbations that are characteristic of acquisition artifacts. The Personalization Module trained in Phase 2 then operates on these pre-harmonized features and explicitly models high-frequency, boundary-level variations associated with individual rater styles.

14. Limitations

Despite its overall robustness, the framework exhibits several limitations that are closely tied to the complexity and inconsistency of multi-rater annotation behavior. As illustrated in Figure 13, we identify two representative failure scenarios.

First, in the left sample, one annotator delineates a noticeably larger region than all others, often because they perceived subtle texture cues that other experts ignored. In these cases, our model partially detects those underlying patterns as well, hence several personalized heads show activations in the same extended region. However, the frequency-prompt personalization module simultaneously constrains each prediction to match the stylistic tendencies of the corresponding annotator. This dual mechanism, pattern detection versus stylistic constraint, can create a bottleneck: the model recognizes a plausible region but is restricted from fully expressing it if it contradicts the annotator's typical behavior. As a result, the personalized mask may appear truncated relative to the annotator's unusually large annotation, producing a small drop in per-rater quantitative metrics.

Second, we observe cases where the anatomical structure exhibits two overlapping or symmetric patterns, both of which could be interpreted as valid targets. Some annotators segment both regions, while others label only one. This mixture of plausible but inconsistent interpretations makes it inherently difficult for the model to disambiguate the "intended" target for each annotator. While the harmonized latent space captures both spatial hypotheses, the personalized decoding sometimes struggles to resolve the ambiguity when annotator styles diverge only subtly. This further reflects the dependency on annotation consistency: when multiple interpretations are equally defensible, perfect personalization becomes inherently ill-posed.

References

[1] Reza Azad, Amirhossein Kazerouni, Babak Azad, Ehsan Khodapanah Aghdam, Yury Velichko, Ulas Bagci, and Dorit Merhof. Laplacian-former: Overcoming the limitations of vision transformers in local texture detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 736–746. Springer, 2023. 6

[2] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *Information Processing in Medical Imaging (IPMI)*, pages 715–726. Springer, 2021. 2

[3] Chaowei Fang, Qian Wang, Lechao Cheng, Zhifan Gao, Chengwei Pan, Zhen Cao, Zhaohui Zheng, and Dingwen Zhang. Reliable mutual distillation for medical image segmentation under imperfect annotations. *IEEE Transactions on Medical Imaging*, 42(6):1720–1734, 2023. 6, 8

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017. 4

[5] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 6, 8

[6] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 6

[7] 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. 1, 5

[8] Shuailin Li, Zhitong Gao, and Xuming He. Superpixel-guided iterative learning from noisy labels for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–535. Springer, 2021. 6, 8

[9] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2606–2616, 2022. 6

[10] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010. 4

[11] 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. 6

[12] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019. 6, 8

[13] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13726–13735, 2020. 6, 8

[14] 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. 2, 5, 6, 12

[15] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020. 6, 8

[16] Xingyue Zhao, Zhongyu Li, Xiangde Luo, Peiqi Li, Peng Huang, Jianwei Zhu, Yang Liu, Jihua Zhu, Meng Yang, Shi Chang, 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. 6, 8