TO CONTROL FDR, CONSTRAIN FNR: A REVERSED CONFORMAL RISK TRAINING FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

In safety-critical image segmentation, controlling the False Discovery Rate (FDR) is essential. However, existing post-hoc Conformal Risk Control (CRC) for non-monotonic metrics like FDR yields overly conservative predictions, severely degrading model utility. While Conformal Risk Training (CRT) could theoretically mitigate this by embedding calibration into the training loop, it is fundamentally inapplicable to FDR. The inherent non-monotonicity of FDR violates the prerequisites for implicit differentiation, rendering the standard bi-level optimization intractable. To avoid direct differentiation through this non-monotone calibration problem during training, we propose Reversed Conformal Risk Training (R-CRT). R-CRT employs a task reversal mechanism: it minimizes a differentiable surrogate of FDR while enforcing the strictly monotonic False Negative Rate (FNR) as a proxy constraint, thereby enabling implicit differentiation through a monotone proxy threshold-selection problem. Extensive experiments on medical datasets demonstrate that R-CRT successfully trains highly effective models that significantly outperform baseline methods while maintaining highly reliable empirical FDR control, grounded in the theoretical framework of Conservative CRC.

1 INTRODUCTION

Image segmentation is a cornerstone of computer vision, serving as a critical perception module in domains such as medical diagnostics, autonomous navigation, and remote sensing. While deep learning has driven remarkable advances in average-case segmentation accuracy, deploying these models in safety-critical environments demands robust, instance-level uncertainty quantification. Conventional metrics, such as the Dice coefficient or Intersection over Union (IoU), merely reflect aggregate performance and fail to provide the rigorous reliability guarantees required when algorithmic errors carry severe real-world consequences.

Conformal Prediction (CP) (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023) has emerged as a principled paradigm to address this need, offering distribution-free uncertainty quantification with strict finite-sample guarantees. Recent advances in Conformal Risk Control (CRC) (Angelopoulos et al., 2024) have generalized CP to bound complex, user-specified performance metrics, such as the False Negative Rate (FNR). However, standard CRC fundamentally assumes the loss function is monotonic with respect to the prediction set size. In many high-stakes applications, controlling false positives is equally paramount to prevent catastrophic outcomes—ranging from unnecessary invasive biopsies in healthcare to dangerous emergency braking in autonomous driving. This necessitates controlling the False Discovery Rate (FDR), an inherently *non-monotonic* metric. While recent pioneering work (Angelopoulos, 2026; Aldirawi et al., 2026) has extended post-hoc CRC to handle non-monotonic losses, it introduces a severe practical dilemma: applying post-hoc calibration to a fixed, pre-trained model often yields overly conservative thresholds. To strictly satisfy the FDR constraint, the model is forced to predict nearly empty masks, trivially achieving a low FDR but suffering from severe under-detection (i.e., an unacceptably high FNR), rendering the system practically useless.

A natural remedy to this conservatism is Conformal Risk Training (CRT) (Yeh et al., 2025), which embeds the calibration procedure directly into the training loop via bi-level optimization. By making the model "risk-aware" during training, CRT preserves predictive utility while satisfying risk constraints. Unfortunately, applying CRT to FDR encounters a fundamental mathematical

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

bottleneck. CRT relies on implicit differentiation through the inner threshold-selection step, which strictly requires the constraint risk to be monotonically non-increasing. Because the empirical FDR fluctuates non-monotonically as the threshold varies, the inner root-finding problem becomes ill-posed, leading to intractable optimization and gradient instability (as illustrated in Figure 1). To overcome this fundamental limitation, we propose a novel end-to-end training framework that

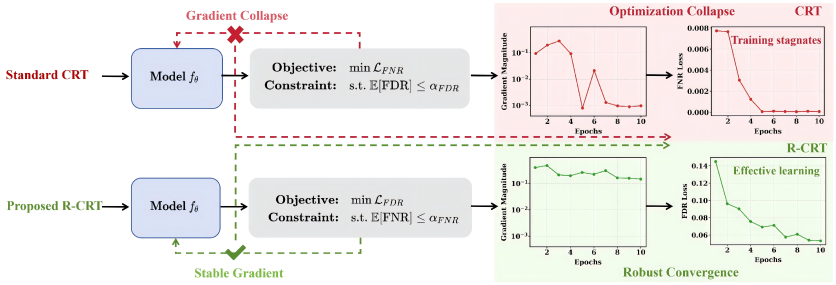

Figure 1: Comparison of Standard CRT and our proposed R-CRT. Standard CRT suffers from gradient collapse due to the non-monotonic FDR constraint, whereas R-CRT ensures stable optimization via a task reversal mechanism.

circumvents the non-monotonicity bottleneck. To comprehensively evaluate the effectiveness of our proposed approach, we conduct extensive experiments across three diverse medical image segmentation datasets. Empirical results demonstrate that our method successfully prevents the severe failure modes inherent in standard post-hoc calibration. By explicitly optimizing the underlying model representations to be risk-aware, R-CRT significantly minimizes the FNR while providing highly reliable, stability-adjusted empirical FDR control. In summary, our main contributions are four-fold:

- **Formulation of a Novel and Critical Problem:** To the best of our knowledge, we are the first to formulate and investigate the problem of conformal risk training under non-monotonic constraints. While prior research has predominantly optimized aggregate metrics (e.g., IoU) that are insufficient for safety-critical applications, we tackle the FDR constraint—a metric that fundamentally aligns with the asymmetric risk demands of real-world clinical and autonomous systems.
- **Reversed Conformal Risk Training (R-CRT):** We propose an innovative bi-level optimization framework. By integrating a task reversal mechanism with implicit differentiation, R-CRT transforms the intractable non-monotonic constraint into a well-posed proxy problem, enabling the stable end-to-end training of risk-aware models.
- **Dynamic Risk Annealing Strategy:** We design a curriculum-based annealing scheme to optimize the proxy constraint. This strategy effectively prevents premature convergence and smoothly guides the optimization trajectory, optimally balancing the inherent trade-off between the False Discovery Rate and the False Negative Rate.
- **Comprehensive Empirical Validation:** We conduct extensive experiments across diverse medical image segmentation datasets. The results demonstrate that, compared to traditional post-hoc calibration baselines, our R-CRT method significantly enhances practical model utility while strictly satisfying user-specified FDR constraints throughout.

2 Related Work

2.1 Conformal Segmentation

Existing research on conformal segmentation (Gade et al., 2024; Brunekreef et al., 2024; Elyassirad et al., 2025; Chen et al., 2025; Huang et al., 2026) mainly falls into two categories: one leverages (Mossina & Friedrich, 2025a;b; Gaillochet et al., 2026; Cheung et al., 2026) the spatial correlation of images to achieve more effective conformal segmentation, while the other (Teneggi et al., 2023; 2025; Luo & Zhou, 2025; Blot et al., 2025; Bereska et al., 2025; Luo et al., 2026) improves the calibration step to enhance conditional coverage. Additionally, some studies (Lambert et al., 2024;

Tan et al., 2025) have explored CRC for 3D medical image segmentation. For instance, Liu et al. (2025) investigated conformal segmentation in hyperspectral images, and Mossina et al. (2024) extended CRC to multi-dimensional semantic segmentation tasks. However, all these methods are dedicated to controlling the FNR.

While controlling FNR ensures true targets are detected, many safety-critical applications are equally sensitive to false discoveries—such as unnecessary invasive biopsies or dangerous emergency braking. Despite its immense practical value, controlling the FDR in conformal segmentation remains largely unexplored. This gap primarily stems from a fundamental mathematical challenge: the inherent non-monotonicity of the FDR metric, which renders standard risk-bounding optimization techniques inapplicable. Our work specifically targets this unexplored domain, aiming to achieve a rigorous and optimal balance between FDR and FNR.

2.2 Conformal Training

Traditional CP and CRC are typically applied as post-hoc procedures following model training. This decoupled "train-then-control" paradigm suffers from a notable limitation: the model remains agnostic to the downstream calibration process during optimization. Consequently, enforcing strict risk or coverage constraints often leads to a significant degradation in average-case performance. To address this issue, conformal training has emerged, whose core idea is to embed the conformal calibration process directly into model training, thereby enabling end-to-end joint optimization.

Conformal training (Stutz et al., 2022) and related works (Einbinder et al., 2022; Correia et al., 2024; Noorani et al., 2025; Bao et al., 2025) introduce methods to differentiably incorporate CP during model training by treating a portion of each mini-batch as a pseudo-calibration set. Recently, Yeh et al. (2025) extended this end-to-end optimization paradigm by proposing conformal risk training (CRT). CRT generalizes the optimization objective from standard classification coverage to the broader class of Optimized Certainty Equivalent (OCE) risk control. By formulating the derivation of the risk-controlling parameter as a bi-level optimization problem, CRT enables the exact gradient computation with respect to the risk-controlling parameter.

3 Problem Formulation

3.1 Risk-Constrained Segmentation Objective

For image segmentation, we consider an input image X_i with its ground truth binary mask $Y_i \in \{0,1\}^N$, which delineates a region of interest. Our primary objective is to construct a prediction set $\hat{C}(X_i) \subset \{1,2,...,N\}$ that controls the FDR in expectation. Specifically, we aim to ensure:

$$\mathbb{E}\left[\frac{\sum_{j=1}^{N} \hat{C}(X_i)[j] \cdot (1 - Y_i[j])}{\max\left(\sum_{j=1}^{N} \hat{C}(X_i)[j], \epsilon\right)}\right] \leq \alpha, \tag{1}$$

where $\alpha \in (0,1)$ is a user-specified risk level. Here, $\sum_{j=1}^{N} \hat{C}(X_i)[j]$ denotes the total number of pixels included in the prediction set. We also define ϵ as a small positive constant (e.g., 10^{-6}) used to prevent division by zero in cases where the prediction set is empty.

However, controlling the FDR alone is insufficient, as predicting all zeros trivially achieves a zero FDR but causes severe under-detection. Therefore, under the strict FDR constraint, we must simultaneously maximize true detections, which is equivalent to minimizing the FNR:

$$\text{FNR}(X_i, \widehat{C}) = 1 - \frac{\sum_{j=1}^{N} \widehat{C}(X_i)[j] \cdot Y_i[j]}{\max\left(\sum_{j=1}^{N} Y_i[j], \epsilon\right)}. \tag{2}$$

Formally, let $R_{\text{FDR}}(\widehat{C})$ denote the expected FDR risk defined in Equation 1. The overarching goal of our task can be formulated as the following risk-constrained optimization problem:

$$\min_{\widehat{C}} \quad \mathbb{E}\left[\text{FNR}(X_i, \widehat{C}(X_i))\right], \quad \text{s.t.} \quad R_{\text{FDR}}(\widehat{C}) \leq \alpha. \tag{3}$$

3.2 FINITE-SAMPLE GUARANTEES VIA CRC-C

Given a base segmentation model outputting a probability map $p(X_i) = (p(X_i), ..., p_N(X_i))$, we define the prediction set at threshold $\tau \in [0, 1]$ as $\widehat{C}(X_i, \tau) = \{j : \hat{p}_j(X_i) \geq \tau\}$. Standard CRC (Angelopoulos et al., 2024) provides finite-sample risk guarantees under the assumption that the loss function is monotonically non-increasing with respect to τ. However, ℓ_{FDR} is inherently non-monotonic, as varying τ simultaneously alters both the numerator and the denominator.

To establish rigorous theoretical guarantees for non-monotonic losses, we employ the Conservative Conformal Risk Control (CRC-C) framework (Angelopoulos, 2026), which bounds the risk via algorithmic stability. Let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ be a calibration dataset that is not seen during model training. We define our calibration algorithm $\mathcal{A}$ to select the most permissive (i.e., smallest) threshold that controls the empirical risk $\hat{R}_{\text{FDR}}$ at level α:

$$\mathcal{A}(\mathcal{D}_{\text{cal}}) = \inf \left\{ \tau \in [0, 1] : \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{(X_i, Y_i) \in \mathcal{D}_{\text{cal}}} \ell_{\text{FDR}}(X_i, Y_i; \tau) \leq \alpha \right\}. \tag{4}$$

The validity of CRC-C relies on the β-stability of $\mathcal{A}$. Formally, $\mathcal{A}$ is β-stable with respect to a reference algorithm $\mathcal{A}^*$ if its expected leave-one-out loss deviates from the expected full-data loss of $\mathcal{A}^*$ by at most $\beta \geq 0$. This stability directly translates to a distribution-free risk guarantee on unseen data.

Theorem 1 (Finite-Sample FDR Control, adapted from Angelopoulos (2026)). *Let (X_{test}, Y_{test}) be a new test sample. Assume $\mathcal{A}$ is symmetric and β-stable with respect to $\mathcal{A}^*$, and that the combined sequence $\mathcal{D}_{cal} \cup \{(X_{test}, Y_{test})\}$ is exchangeable. If the reference algorithm satisfies $\mathbb{E}[\ell_{FDR}(X_{test}, Y_{test}; \mathcal{A}^*(\mathcal{D}_{cal} \cup \{(X_{test}, Y_{test})\}))] \leq \alpha - \beta$, then the calibrated threshold $\hat{\tau} = \mathcal{A}(\mathcal{D}_{cal})$ guarantees:*

$$\mathbb{E}\left[\ell_{\text{FDR}}(X_{\text{test}}, Y_{\text{test}}; \hat{\tau})\right] \leq \alpha. \tag{5}$$

Remark 1 (Practical Implementation.) *In practice, a valid deterministic upper bound on the true stability parameter β is typically inaccessible. Instead, we estimate it via bootstrap resampling on $\mathcal{D}_{\text{cal}}$, yielding $\hat{\beta}$, and target an adjusted level $\alpha' = \alpha - \hat{\beta}$. As noted by Angelopoulos (2026), while the bootstrap provides a consistent estimate in well-behaved settings, it does not constitute a strict deterministic upper bound. Consequently, while our method is grounded in the finite-sample theory of Theorem 1, the practical bootstrap implementation provides an empirical stability-adjusted FDR bound rather than an unconditional finite-sample guarantee. Notably, in image segmentation, empirical regularity renders this stability penalty negligible (e.g., $\mathcal{O}(10^{-5})$), allowing our practical implementation to achieve robust empirical control without sacrificing statistical power.*

3.3 CONFORMAL RISK TRAINING AND THE NON-MONOTONICITY BOTTLENECK

While post-hoc calibration (e.g., CRC-C) provides rigorous guarantees, applying it to fixed, pre-trained models often yields overly conservative thresholds, severely degrading task utility (i.e., unacceptably high FNR). Conformal Risk Training (CRT) (Yeh et al., 2025) addresses this by jointly optimizing the model parameters θ and the threshold τ via a bi-level optimization framework:

$$\min_{\theta} \quad \mathbb{E}\left[\ell_{\text{task}}(\theta, \tau(\theta))\right] \quad \text{s.t.} \quad \tau(\theta) = \inf \left\{ \tau \in [0, 1] : \hat{R}_{\text{constraint}}(\theta, \tau) \leq \alpha \right\}. \tag{6}$$

By differentiating through the inner threshold-selection step, CRT trains the model to be "risk-aware," encouraging feature representations that naturally satisfy the risk constraint without sacrificing average-case performance.

The Non-Monotonicity Bottleneck. A fundamental mathematical prerequisite of standard CRT is that the constraint risk $\hat{R}_{\text{constraint}}$ must be strictly monotonic with respect to τ. Monotonicity ensures a unique solution for the inner problem and stable implicit gradients $\partial\tau/\partial\theta$. However, our target constraint, FDR, is inherently non-monotonic. Directly applying CRT to bound FDR leads to an ill-posed inner optimization and gradient instability.

4 REVERSED CONFORMAL RISK TRAINING

As established in Section 3, CRT requires a known monotonic relationship between the constraint risk and the threshold with respect to the decision threshold τ. When the target constraint is the FDR, the optimization becomes intractable due to the inherent non-monotonicity of FDR. To circumvent this fundamental bottleneck, we propose **Reversed Conformal Risk Training (R-CRT)**, a framework that reformulates the learning objective via task reversal, utilizes implicit differentiation for stable gradient estimation, and employs a dynamic annealing schedule to discover the optimal risk trade-off.

4.1 TASK REVERSAL VIA MONOTONICITY

Let f_θ denote the segmentation network parameterized by θ. The standard objective in CRC is to minimize the FNR subject to a user-specified FDR constraint α_{FDR}. Because the empirical FDR fluctuates non-monotonically as the threshold τ varies, finding the optimal τ during the forward pass of training is ill-posed and numerically unstable.

Conversely, the empirical FNR is monotone non-decreasing with respect to τ. Recognizing this structural property, we propose to *reverse* the constrained optimization problem. Instead of constraining FDR, we constrain the FNR to a proxy level α_{FNR} and minimize the surrogate FDR. To motivate the reversed formulation, we first consider a pooled proxy constraint that selects a single threshold for the entire pseudo-calibration batch:

$$\min_{\theta} \quad \mathcal{L}_{\text{FDR}}(\theta, \tau^*(\theta)) \quad \text{s.t.} \quad \tau^*(\theta) = \sup\left\{\tau \in [0,1] : \text{FNR}(\tau;\theta) \le \alpha_{\text{FNR}}\right\}, \tag{7}$$

This pooled formulation can be sensitive to the composition of a mini-batch: a small number of atypical samples may dominate the batch-level FNR constraint and lead to an unstable or overly conservative shared threshold. Therefore, the implementation used in our experiments first computes an individual proxy threshold for each pseudo-calibration sample and then aggregates these thresholds into a single differentiable batch threshold.

Let $P_{\text{cal},i} = \sigma(f_\theta(X^i_{\text{cal}}))$ denote the predicted probability map of pseudo-calibration sample i, where σ is the sigmoid function. For each sample $i \in \mathcal{B}_{\text{cal}}$, we compute

$$\lambda_i^*(\theta) = \sup\left\{\tau \in [0,1] : \widetilde{\text{FNR}}(\tau; P^i_{\text{cal}}, Y^i_{\text{cal}}) \le \alpha_{\text{FNR}}\right\}. \tag{8}$$

The training threshold is then obtained by soft-quantile aggregation:

$$\tilde{\tau}^*(\theta) = \text{SoftQuantile}\left(\{\lambda_i^*(\theta)\}_{i\in\mathcal{B}_{\text{cal}}}, 1-q\right). \tag{9}$$

The outer objective is evaluated on $\mathcal{B}_{\text{pred}}$ using $\tilde{\tau}^*(\theta)$.

Here, $\mathcal{L}_{\text{FDR}}$ is a differentiable surrogate for FDR. The pooled FNR is monotone non-decreasing in τ, so binary search returns the largest feasible threshold. The per-sample formulation retains this property for each λ_i^* while reducing the sensitivity of threshold selection to heterogeneous mini-batches.

4.2 IMPLICIT DIFFERENTIATION FOR THRESHOLD OPTIMIZATION

To enable end-to-end training, gradients must flow through the per-sample thresholds $\{\lambda_i^*(\theta)\}$, their soft-quantile aggregation $\tilde{\tau}^*(\theta)$, and the outer FDR objective. We introduce a temperature-scaled sigmoid function $\sigma_T(z) = (1+\exp(-z/T))^{-1}$ to construct smooth, differentiable approximations of the risk metrics, denoted as $\widetilde{\text{FNR}}_T$ and $\widetilde{\text{FDR}}_T$. For each pseudo-calibration sample i, the smoothed constraint defines the root-finding problem $g_i(\theta, \lambda_i^*) = \widetilde{\text{FNR}}_T(\lambda_i^*; P_{\text{cal},i}, Y^i_{\text{cal}}) - \alpha_{\text{FNR}} = 0$. By invoking the Implicit Function Theorem (IFT), we compute a stabilized implicit gradient of the threshold with respect to the model outputs without unrolling the binary search steps:

$$\frac{\partial \lambda_i^*}{\partial \theta} = -\left(\frac{\partial g_i(\theta,\lambda_i^*)}{\partial \lambda_i^*}\right)^{-1} \frac{\partial g_i(\theta,\lambda_i^*)}{\partial \theta}. \tag{10}$$

The gradients of $\{\lambda_i^*\}$ are propagated through the soft-quantile aggregation $\tilde{\tau}^*$, and then through the outer FDR objective.

Accordingly, the gradient of the outer objective includes both the direct dependence on θ and the indirect dependence through $\{\lambda_i^*\}$ and $\tilde{\tau}^*$:

$$\nabla_\theta \mathcal{L}_{\text{FDR}} = \frac{\partial \mathcal{L}_{\text{FDR}}}{\partial \theta} + \frac{\partial \mathcal{L}_{\text{FDR}}}{\partial \tilde{\tau}^*} \sum_{i \in \mathcal{B}_{\text{cal}}} \frac{\partial \tilde{\tau}^*}{\partial \lambda_i^*} \frac{\partial \lambda_i^*}{\partial \theta}. \tag{11}$$

4.3 Dynamic Risk Annealing

A fundamental challenge of the reversed formulation is that the optimal proxy constraint α_{FNR} corresponding to the user's true desired α_{FDR} is unknown *a priori*. Furthermore, enforcing a strict α_{FNR} constraint from the initialization phase often induces severe gradient shock, leading to degenerate feature representations. To address this, we treat the proxy constraint as a dynamic target

Algorithm 1 Reversed Conformal Risk Training (R-CRT) Framework

Require: Pre-trained segmentation model f_θ, Training data $\mathcal{D}_{\text{train}}$, Calibration data $\mathcal{D}_{\text{cal}}$, Target FDR level α_{FDR}, Total epochs E, Initial proxy constraint α_{start}, Final proxy constraint α_{end}, quantile-complement q.

Ensure: Fine-tuned model parameters θ^*, Calibrated threshold $\hat{\tau}$.

▷ ***Phase 1: Reversed Conformal Risk Training***

for epoch $t = 1, \ldots, E$ **do**
  Determine dynamic proxy FNR constraint via annealing schedule (Eq. 12).
  for each mini-batch $\mathcal{B} = (X, Y) \in \mathcal{D}_{\text{train}}$ **do**
    Partition the mini-batch: $(\mathcal{B}_{\text{cal}}, \mathcal{B}_{\text{pred}}) \leftarrow \mathcal{B}$.
    Obtain probability maps: $P_{\text{cal}} = \sigma(f_\theta(X_{\text{cal}}))$ and $P_{\text{pred}} = \sigma(f_\theta(X_{\text{pred}}))$.
    For each $i \in \mathcal{B}_{\text{cal}}$, find an individual proxy threshold:
    $$\lambda_i^* \leftarrow \sup\left\{\lambda \in [0,1] : \widetilde{\text{FNR}}(\lambda; P_{\text{cal},i}, Y_{\text{cal}}^i) \le \alpha^{(t)}\right\}.$$
    Aggregate the individual thresholds:$\tilde{\tau}^* \leftarrow \text{SoftQuantile}\left(\{\lambda_i^*\}_{i \in \mathcal{B}_{\text{cal}}}, 1-q\right)$.
    Calculate the implicit gradients of $\{\lambda_i^*\}$ using the IFT in Eq. 10.
    Evaluate the surrogate FDR loss on $\mathcal{B}_{\text{pred}}$ using $\tilde{\tau}^*$.
    Backpropagate through the threshold aggregation and update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{FDR}}$.
  end for
end for
$\theta^* \leftarrow \theta$
Freeze the trained parameters θ^* before calibration.

▷ ***Phase 2: Post-hoc Calibration via CRC-C***

Estimate stability parameter $\hat{\beta}$ via bootstrap resampling on $\mathcal{D}_{\text{cal}}$.
Calculate adjusted risk level $\alpha' \leftarrow \alpha_{\text{FDR}} - \hat{\beta}$.
Determine the final calibrated threshold $\hat{\tau}$ on $\mathcal{D}_{\text{cal}}$:
$$\hat{\tau} \leftarrow \inf\left\{\tau \in [0,1] : \tfrac{1}{|\mathcal{D}_{\text{cal}}|} \textstyle\sum_{(X_i,Y_i) \in \mathcal{D}_{\text{cal}}} \ell_{\text{FDR}}(X_i, Y_i; \tau) \le \alpha'\right\}.$$
return $\theta^*, \hat{\tau}$.

$\alpha^{(t)}$ that evolves over training epochs $t \in \{1, \ldots, E\}$. We propose a **Dynamic Risk Annealing** strategy, initializing the constraint at a relaxed level α_{start} and gradually decaying it to a stricter terminal level α_{end} via a scheduling function $\mathcal{S}$:

$$\alpha^{(t)} = \mathcal{S}(t; \alpha_{\text{start}}, \alpha_{\text{end}}, E). \tag{12}$$

This annealing mechanism acts as a curriculum: early in training, the relaxed constraint allows the model to learn general discriminative features without being overly penalized. As training progresses, the tightening constraint is intended to smoothly guide the optimization trajectory along the Pareto frontier of the FDR-FNR trade-off, thereby avoiding premature convergence to suboptimal local minima.

Remark 2 (Decoupled Post-hoc Calibration.) *Although R-CRT effectively shapes the output distribution to minimize empirical FDR, it is fundamentally a representation learning strategy that*

lacks distribution-free statistical guarantees for the target α_{FDR}. To bridge this gap, we decouple risk-aware training from statistical calibration. By freezing the optimized parameters θ^* and applying the Conservative Conformal Risk Control (CRC-C) procedure on a held-out calibration set $\mathcal{D}_{cal}$ (Section 3), we obtain a final threshold $\hat{\tau}$. This decoupled approach leverages the CRC-C framework to provide highly reliable, stability-adjusted control over the non-monotonic FDR at the user-specified level α_{FDR}, while preserving the significantly minimized FNR achieved during training.

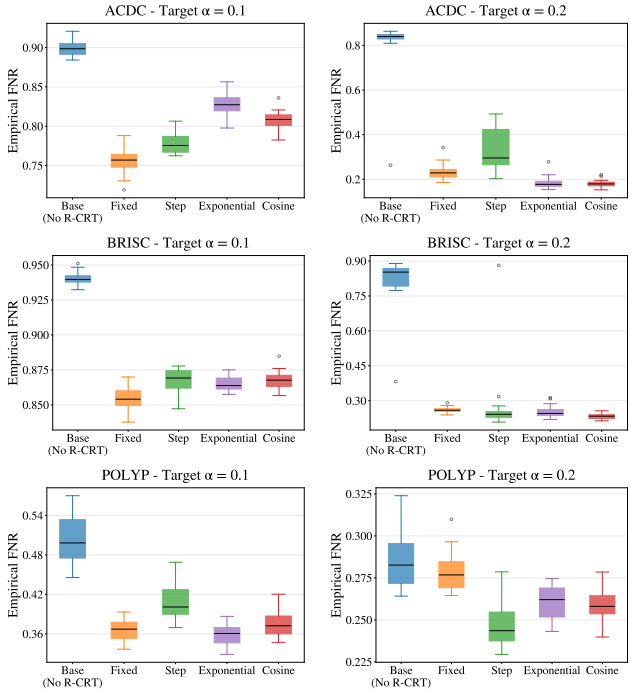

Figure 2: Comparison of Empirical False Negative Rate (FNR) across different datasets.

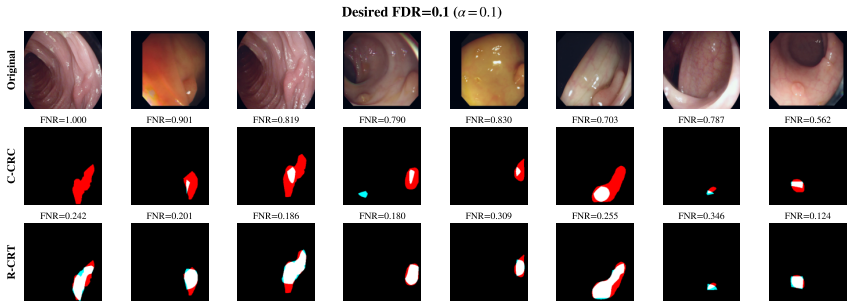

Figure 3: Qualitative comparison of CRC-C and R-CRT prediction sets at a desired FDR of $\alpha = 0.1$. The top row shows the original polyp images, the middle row displays the CRC-C prediction sets, and the bottom row presents the R-CRT prediction sets. White pixels represent true positives, red indicates false negatives, and cyan indicates false positives. The FNR is provided below each prediction. Compared to CRC-C, our R-CRT approach demonstrates more consistent coverage and significantly lower FNR across diverse scenarios, while maintaining a realistic distribution of false negatives, highlighting the effectiveness of our conditional risk control approach.

5 Experiments

In this section, we evaluate four variants of our proposed R-CRT framework against a vanilla CRC-C baseline. In our primary evaluations, we employ PSPNet (Zhao et al., 2017) as the foundational

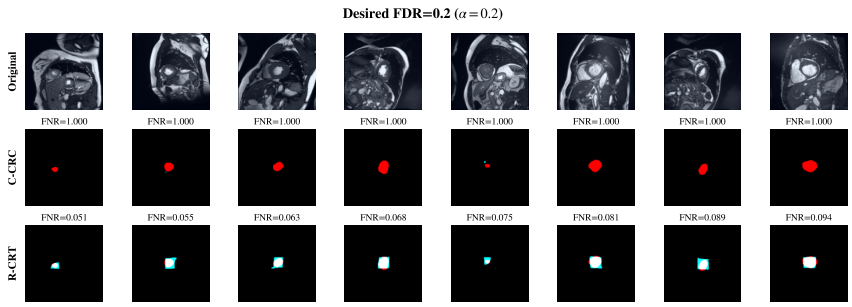

Figure 4: Qualitative comparison of CRC-C and R-CRT prediction sets at a desired FDR of $\alpha = 0.2$.

Table 1: Comparison of Base models and different R-CRT annealing strategies across various segmentation backbones on the POLYP dataset. Results are reported as Mean ± Std over 20 runs. For FNR, the relative improvement (↓) compared to the respective Base model is shown in parentheses. Best results in each group are highlighted in bold.

Backbone	Target α	Method	FDR	FNR ↓	IoU ↑	Dice ↑
UNet	0.1	Base (No R-CRT)	0.027 ± 0.047	0.894 ± 0.183	0.101 ± 0.173	0.118 ± 0.202
		Fixed	0.099 ± 0.024	0.603 ± 0.042 (↓ 32.5%)	0.381 ± 0.035	0.463 ± 0.036
		Step	0.100 ± 0.017	**0.596 ± 0.031 (↓ 33.3%)**	**0.388 ± 0.025**	**0.469 ± 0.026**
		Exponential	0.094 ± 0.018	0.602 ± 0.030 (↓ 32.7%)	0.382 ± 0.025	0.463 ± 0.026
		Cosine	0.099 ± 0.016	0.601 ± 0.024 (↓ 32.8%)	0.382 ± 0.020	0.466 ± 0.020
	0.2	Base (No R-CRT)	0.196 ± 0.019	0.435 ± 0.011	0.481 ± 0.010	0.557 ± 0.010
		Fixed	0.197 ± 0.023	**0.383 ± 0.019 (↓ 12.0%)**	**0.526 ± 0.012**	**0.605 ± 0.012**
		Step	0.190 ± 0.022	0.387 ± 0.014 (↓ 11.0%)	0.525 ± 0.013	0.602 ± 0.013
		Exponential	0.191 ± 0.026	0.386 ± 0.017 (↓ 11.2%)	0.520 ± 0.010	0.600 ± 0.011
		Cosine	0.195 ± 0.021	0.383 ± 0.014 (↓ 12.0%)	0.524 ± 0.011	0.603 ± 0.011
DeepLabV3+	0.1	Base (No R-CRT)	0.098 ± 0.016	0.336 ± 0.018	0.624 ± 0.013	0.695 ± 0.012
		Fixed	0.096 ± 0.014	0.316 ± 0.013 (↓ 5.9%)	0.638 ± 0.011	0.709 ± 0.010
		Step	0.094 ± 0.008	0.321 ± 0.010 (↓ 4.6%)	0.637 ± 0.010	0.707 ± 0.010
		Exponential	0.096 ± 0.017	0.314 ± 0.016 (↓ 6.6%)	0.641 ± 0.010	0.712 ± 0.010
		Cosine	0.100 ± 0.014	**0.311 ± 0.016 (↓ 7.6%)**	**0.644 ± 0.012**	**0.715 ± 0.011**
	0.2	Base (No R-CRT)	0.195 ± 0.017	0.240 ± 0.011	0.647 ± 0.010	0.724 ± 0.010
		Fixed	0.192 ± 0.017	**0.222 ± 0.012 (↓ 7.8%)**	**0.662 ± 0.010**	**0.739 ± 0.010**
		Step	0.202 ± 0.015	0.225 ± 0.014 (↓ 6.5%)	0.655 ± 0.011	0.732 ± 0.011
		Exponential	0.194 ± 0.015	0.231 ± 0.010 (↓ 4.0%)	0.653 ± 0.011	0.730 ± 0.010
		Cosine	0.190 ± 0.019	0.230 ± 0.009 (↓ 4.3%)	0.652 ± 0.011	0.729 ± 0.010
HRNet	0.1	Base (No R-CRT)	0.001 ± 0.001	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Fixed	0.095 ± 0.018	0.785 ± 0.013 (↓ 21.5%)	0.208 ± 0.012	0.274 ± 0.013
		Step	0.092 ± 0.014	0.775 ± 0.013 (↓ 22.5%)	0.216 ± 0.012	0.280 ± 0.015
		Exponential	0.103 ± 0.017	**0.753 ± 0.014 (↓ 24.7%)**	**0.234 ± 0.013**	**0.302 ± 0.015**
		Cosine	0.098 ± 0.017	0.765 ± 0.013 (↓ 23.5%)	0.225 ± 0.012	0.290 ± 0.014
	0.2	Base (No R-CRT)	0.155 ± 0.009	0.759 ± 0.012	0.226 ± 0.011	0.290 ± 0.013
		Fixed	0.202 ± 0.031	0.686 ± 0.010 (↓ 9.7%)	0.288 ± 0.008	0.363 ± 0.009
		Step	0.205 ± 0.015	0.665 ± 0.014 (↓ 12.4%)	0.302 ± 0.012	0.376 ± 0.014
		Exponential	0.204 ± 0.018	**0.662 ± 0.016 (↓ 12.8%)**	**0.303 ± 0.013**	**0.377 ± 0.014**
		Cosine	0.207 ± 0.016	0.664 ± 0.016 (↓ 12.6%)	0.300 ± 0.013	0.373 ± 0.015

backbone. To demonstrate the effectiveness of our approach, comprehensive experiments are conducted across three diverse medical image segmentation datasets, including Polyp (Jha et al., 2019; Bernal et al., 2017; Vázquez et al., 2017; Tajbakhsh et al., 2015; Silva et al., 2014), ACDC Bernard et al. (2018), and BRISC Fateh et al. (2026). For rigorous evaluation, the target risk levels are set to $\alpha \in \{0.1, 0.2\}$, and all results are reported across 20 random data splits to ensure statistical robustness. In Appendices A E, we elaborate on the parameter settings and conduct ablation experiments.

To further illustrate the practical advantages of our proposed method, we present qualitative comparisons between the vanilla CRC-C baseline and our R-CRT framework in Figures 3-4 and appendix Figures 6. As observed in the middle rows across all figures, the baseline CRC-C method exhibits a severe failure mode. To strictly satisfy the desired FDR constraint, the post-hoc calibration on an unoptimized model is forced to apply excessively conservative thresholds. This results in highly shrunken prediction sets dominated by red pixels. Particularly in the ACDC (Figure 4) and BRISC

Table 2: Comparison of Base model and different R-CRT annealing strategies across various datasets and Target α. Results are reported as Mean $\pm$ Std over 20 runs. For FNR, the relative improvement ($\downarrow$) or degradation ($\uparrow$) compared to the Base model is shown in parentheses. Best results in each group are highlighted in bold.

Dataset	Target α	Method	FDR	FNR $\downarrow$	IoU $\uparrow$	Dice $\uparrow$
ACDC	0.1	Base (No R-CRT)	0.095 ± 0.020	0.898 ± 0.009	0.058 ± 0.004	0.095 ± 0.006
		Fixed	0.093 ± 0.016	**0.756 ± 0.016 ($\downarrow$ 15.8%)**	**0.194 ± 0.014**	**0.267 ± 0.017**
		Step	0.087 ± 0.015	0.779 ± 0.013 ($\downarrow$ 13.3%)	0.173 ± 0.013	0.245 ± 0.016
		Exponential	0.085 ± 0.016	0.828 ± 0.014 ($\downarrow$ 7.8%)	0.124 ± 0.010	0.185 ± 0.013
		Cosine	0.094 ± 0.021	0.808 ± 0.012 ($\downarrow$ 10.1%)	0.146 ± 0.009	0.213 ± 0.013
	0.2	Base (No R-CRT)	0.203 ± 0.023	0.810 ± 0.126	0.136 ± 0.096	0.201 ± 0.107
		Fixed	0.200 ± 0.025	0.236 ± 0.035 ($\downarrow$ 70.9%)	0.578 ± 0.015	0.682 ± 0.015
		Step	0.189 ± 0.042	0.351 ± 0.150 ($\downarrow$ 56.7%)	0.513 ± 0.115	0.613 ± 0.127
		Exponential	0.200 ± 0.015	0.185 ± 0.027 ($\downarrow$ 77.2%)	**0.607 ± 0.013**	**0.705 ± 0.014**
		Cosine	0.199 ± 0.020	**0.181 ± 0.016 ($\downarrow$ 77.7%)**	0.601 ± 0.014	0.699 ± 0.014
BRISC	0.1	Base (No R-CRT)	0.087 ± 0.013	0.941 ± 0.005	0.059 ± 0.005	0.099 ± 0.007
		Fixed	0.102 ± 0.021	**0.855 ± 0.008 ($\downarrow$ 9.1%)**	**0.145 ± 0.007**	**0.206 ± 0.010**
		Step	0.097 ± 0.014	0.867 ± 0.009 ($\downarrow$ 7.9%)	0.133 ± 0.009	0.194 ± 0.012
		Exponential	0.093 ± 0.018	0.865 ± 0.005 ($\downarrow$ 8.1%)	0.135 ± 0.005	0.196 ± 0.007
		Cosine	0.090 ± 0.016	0.867 ± 0.007 ($\downarrow$ 7.8%)	0.132 ± 0.007	0.192 ± 0.009
	0.2	Base (No R-CRT)	0.205 ± 0.023	0.815 ± 0.106	0.178 ± 0.087	0.259 ± 0.101
		Fixed	0.199 ± 0.012	0.261 ± 0.011 ($\downarrow$ 68.0%)	0.609 ± 0.009	0.721 ± 0.010
		Step	0.204 ± 0.017	0.275 ± 0.141 ($\downarrow$ 66.3%)	0.597 ± 0.110	0.711 ± 0.120
		Exponential	0.198 ± 0.010	0.253 ± 0.026 ($\downarrow$ 69.0%)	0.619 ± 0.012	0.733 ± 0.010
		Cosine	0.201 ± 0.010	**0.231 ± 0.011 ($\downarrow$ 71.6%)**	**0.625 ± 0.006**	**0.735 ± 0.007**
POLYP	0.1	Base (No R-CRT)	0.095 ± 0.016	0.503 ± 0.035	0.475 ± 0.028	0.578 ± 0.027
		Fixed	0.092 ± 0.012	0.365 ± 0.017 ($\downarrow$ 27.5%)	0.578 ± 0.013	0.660 ± 0.013
		Step	0.097 ± 0.015	0.407 ± 0.026 ($\downarrow$ 19.0%)	0.551 ± 0.016	0.633 ± 0.016
		Exponential	0.099 ± 0.013	**0.359 ± 0.017 ($\downarrow$ 28.7%)**	**0.584 ± 0.010**	0.664 ± 0.010
		Cosine	0.096 ± 0.010	0.376 ± 0.020 ($\downarrow$ 25.3%)	0.581 ± 0.014	**0.668 ± 0.013**
	0.2	Base (No R-CRT)	0.194 ± 0.019	0.285 ± 0.016	**0.606 ± 0.010**	**0.692 ± 0.010**
		Fixed	0.196 ± 0.017	0.279 ± 0.011 ($\downarrow$ 2.0%)	0.572 ± 0.012	0.659 ± 0.011
		Step	0.197 ± 0.020	**0.247 ± 0.013 ($\downarrow$ 13.4%)**	0.596 ± 0.014	0.680 ± 0.013
		Exponential	0.196 ± 0.018	0.261 ± 0.010 ($\downarrow$ 8.4%)	0.587 ± 0.012	0.671 ± 0.011
		Cosine	0.195 ± 0.016	0.259 ± 0.009 ($\downarrow$ 9.1%)	0.594 ± 0.011	0.679 ± 0.010

(Figure 6) datasets, the FNR of the baseline frequently hits 1.000, meaning the model collapses into degenerate, trivial predictions to avoid false discoveries.

In stark contrast, the bottom rows demonstrate that our R-CRT approach generates highly accurate and clinically meaningful segmentation masks. By explicitly incorporating risk-aware optimization during the representation learning phase, R-CRT successfully captures the target anatomical structures, as evidenced by the predominantly white regions. The FNR is drastically reduced.

Generalizability Across Architectures: To demonstrate that our framework is architecture-agnostic, we evaluated three widely used segmentation backbones—UNet (Ronneberger et al., 2015), DeepLabV3+ (Chen et al., 2018), and HRNet (Wang et al., 2020)—on the POLYP dataset. As shown in Table 1, our annealing strategies consistently reduce FNR and improve IoU across all models while strictly satisfying the target FDR constraints. These results confirm that our approach robustly enhances risk-aware performance independent of the underlying network.

6 Conclusion

This paper introduces R-CRT to address the non-monotonicity of FDR control in safety-critical segmentation. By reformulating the objective to constrain the monotonic FNR, R-CRT effectively overcomes the severe conservatism inherent in standard post-hoc calibration, yielding highly discriminative models that maintain robust empirical FDR control, supported by the theoretical foundations of algorithmic stability. Although R-CRT provides a practical solution, conformal risk control for non-monotonic metrics remains an open challenge. Our future work will continue to investigate the theoretical properties and optimization dynamics of marginal coverage under such complex metrics, aiming to develop more generalized and tighter risk-controlling frameworks.

AI USE STATEMENT

The authors confirm that generative AI tools were used exclusively for grammatical refinement and polishing of the manuscript's prose. No AI tools were involved in core mathematical derivations, experimental design, or scientific ideation, and all AI-assisted text was thoroughly reviewed and verified by the human authors.

ETHICS STATEMENT

This work proposes a R-CRT framework for reliable False Discovery Rate control in safety-critical image segmentation. All experiments rely entirely on publicly available benchmark datasets without involving new human subjects, rendering IRB approval unnecessary. While R-CRT enhances clinical utility by bounding risk, it serves strictly as a decision-support tool and should not be deployed as a fully autonomous agent for high-stakes decisions.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility, we provide comprehensive theoretical, algorithmic, and empirical details. Theorem 1 establishes the finite-sample risk guarantees, Section 4 details the R-CRT pipeline alongside the reference implementation in Listing 1, and Appendices document the hyperparameter setups, dataset splits, and hardware specifications. Source codes and checkpoints will be released upon acceptance.

REFERENCES

Tareq Aldirawi, Yun Li, and Wenge Guo. Non-monotonicity in conformal risk control. *arXiv preprint arXiv:2604.01502*, 2026.

Anastasios N Angelopoulos. Conformal risk control for non-monotonic losses. *arXiv preprint arXiv:2602.20151*, 2026.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Jie Bao, Chuangyin Dang, Rui Luo, Hanwei Zhang, and Zhixin Zhou. Enhancing adversarial robustness with conformal prediction: A framework for guaranteed model reliability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Oty1LQrnFc.

Jacqueline Isabel Bereska, Hamed Karimi, and Reza Samavi. SACP: Spatially-adaptive conformal prediction in uncertainty quantification of medical image segmentation. In *Medical Imaging with Deep Learning*, 2025. URL https://openreview.net/forum?id=uQaPr1wU1W.

Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sanchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Vincent Blot, Anastasios Nikolas Angelopoulos, Michael Jordan, and Nicolas JB Brunel. Automatically adaptive conformal risk control. In *International Conference on Artificial Intelligence and Statistics*, pp. 19–27. PMLR, 2025.

Joren Brunekreef, Eric Marcus, Ray Sheombarsing, Jan-Jakob Sonke, and Jonas Teuwen. Kandinsky conformal prediction: efficient calibration of image segmentation algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4135–4143, 2024.

Danhui Chen, Ziquan Liu, Chuxi Yang, Dan Wang, Yan Yan, Yi Xu, and Xiangyang Ji. Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24045–24055, 2025.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Matt Y Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. Efficient conformal volumetry for template-based segmentation. *arXiv preprint arXiv:2603.00798*, 2026.

Alvaro Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gKLgY3m9zj.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in neural information processing systems*, 35:22380–22395, 2022.

Danial Elyassirad, Benyamin Gheiji, Mahsa Vatanparast, Amir Mahmoud Ahmadzadeh, and Shahriar Faghani. Conseg: Voxelwise uncertainty quantification for glioma segmentation using conformal prediction. *American Journal of Neuroradiology*, 46(12):2553–2560, 2025.

Amirreza Fateh, Yasin Rezvani, Sara Moayedi, Sadjad Rezvani, Fatemeh Fateh, Mansoor Fateh, and Vahid Abolghasemi. Brisc: Annotated dataset for brain tumor segmentation and classification. *Scientific Data*, 2026.

Marius Gade, Kevin Mekhaphan Nguyen, Sol Gedde, and Alvaro Fernandez-Quilez. Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based mri prostate segmentation. *Insights into Imaging*, 15(1):286, 2024.

Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. Anatomically-aware conformal prediction for medical image segmentation with random walks. *arXiv preprint arXiv:2601.18997*, 2026.

Jiayi Huang, Amirmohammad Farzaneh, and Osvaldo Simeone. Optimized certainty equivalent risk-controlling prediction sets. *arXiv preprint arXiv:2602.13660*, 2026.

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*, pp. 451–462. Springer, 2019.

Benjamin Lambert, Florence Forbes, Senan Doyle, and Michel Dojat. Robust conformal volume estimation in 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 633–643. Springer, 2024.

Kangdao Liu, Tianhao Sun, Hao Zeng, Yongshan Zhang, Chi-Man Pun, and Chi-Man Vong. Spatial-aware conformal prediction for trustworthy hyperspectral image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Rui Luo and Zhixin Zhou. Conditional conformal risk adaptation. *arXiv preprint arXiv:2504.07611*, 2025.

Rui Luo, Jie Bao, Xiaoyi Su, Wen Jung Li, and Suqun Cao. Enhancing image-conditional coverage in segmentation: Adaptive thresholding via differentiable miscoverage loss. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Gd2AiWes1J.

Luca Mossina and Corentin Friedrich. Conformal prediction for image segmentation using morphological prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 78–88. Springer, 2025a.

Luca Mossina and Corentin Friedrich. Controlling false positives in image segmentation via conformal prediction. *arXiv preprint arXiv:2511.15406*, 2025b.

Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3574–3584, 2024.

Sima Noorani, Orlando Romero, Nicolo Dal Fabbro, Hamed Hassani, and George J. Pappas. Conformal risk minimization with variance reduction. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=y2hYL7y8T9.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

Binyu Tan, Zhiyuan Wang, Jinhao Duan, Kaidi Xu, Heng Tao Shen, Xiaoshuang Shi, and Fumin Shen. Conformal lesion segmentation for 3d medical images. *arXiv preprint arXiv:2510.17897*, 2025.

Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, pp. 33940–33960. PMLR, 2023.

Jacopo Teneggi, J Webster Stayman, and Jeremias Sulam. Conformal risk control for semantic uncertainty quantification in computed tomography. *arXiv preprint arXiv:2503.00136*, 2025.

David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uBEiZd2P0K.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

A Experimental Settings and Hyperparameter Details

We utilized six 24GB NVIDIA RTX 4090 GPUs for parallel model training. The base segmentation network is optimized using the Adam optimizer with an initial learning rate of 5×10^{-5} and a total training duration of $E = 10$ epochs, with a mini-batch size of 600.

For temperature scaling in smooth risk approximations, the inner False Negative Rate (FNR) temperature is set to $\tau_{\text{FNR}} = 0.005$, and the outer False Discovery Rate (FDR) temperature is set to $\tau_{\text{FDR}} = 0.01$. For the dynamic risk annealing schedule, the terminal risk constraint is $\alpha_{\text{end}} = 0.2$, and the initial constraint is $\alpha_{\text{start}} = 0.3$ (except for ACDC and BRISC at target $\alpha = 0.1$, where we set $\alpha_{\text{start}} = 0.95$ and $\alpha_{\text{end}} = 0.9$).

Table 3: Training time (seconds) and number of images for different datasets across annealing methods

Dataset	Images	Cosine	Exponential	Step	Fixed
BRISC	3933	118.79	111.82	109.42	111.13
ACDC	1344	85.72	85.25	87.19	84.45
POLYP	1450	107.01	108.72	109.88	109.43

Table 4: Calibration and Testing Time (Seconds) across Datasets and Methods (Mean $\pm$ Std over 20 runs)

Dataset	Method	Calib Images	Test Images	Calibration Time (s)	Testing Time (s)
BRISC	Baseline	215	430	4.6819 ± 0.0470	0.1053 ± 0.0117
	Cosine	215	430	4.6736 ± 0.0204	0.1025 ± 0.0003
	Exponential	215	430	4.6649 ± 0.0097	0.1025 ± 0.0006
	Step	215	430	4.6682 ± 0.0079	0.1028 ± 0.0008
	Fixed	215	430	4.6840 ± 0.0379	0.1027 ± 0.0010
ACDC	Baseline	142	284	3.0747 ± 0.0108	0.0681 ± 0.0004
	Cosine	142	284	3.0652 ± 0.0052	0.0688 ± 0.0005
	Exponential	142	284	3.0641 ± 0.0056	0.0688 ± 0.0004
	Step	142	284	3.0739 ± 0.0285	0.0689 ± 0.0002
	Fixed	142	284	3.0677 ± 0.0042	0.0689 ± 0.0003
POLYP	Baseline	200	399	4.3750 ± 0.0179	0.0954 ± 0.0005
	Cosine	200	399	4.3651 ± 0.0062	0.0955 ± 0.0006
	Exponential	200	399	4.3636 ± 0.0082	0.0951 ± 0.0001
	Step	200	399	4.3667 ± 0.0087	0.0952 ± 0.0003
	Fixed	200	399	4.3587 ± 0.0040	0.0951 ± 0.0002

B Computational Efficiency and Dataset Statistics

Tables 3 and 4 present the dataset distributions, training durations, calibration times, and inference latencies. As indicated in Table 3, training time is primarily dictated by dataset scale, and introducing our annealing strategies introduces virtually zero extra computational overhead. Furthermore, Table 4 demonstrates that our trained models maintain identical calibration and inference speeds to the vanilla baseline while achieving substantial utility gains.

C Additional Datasets Evaluation (FIRE and CAR)

To further verify the generalization of our method, we evaluate R-CRT on two additional datasets: FIRE and CAR. As reported in Table 5, our dynamic annealing strategies consistently outperform the base models, achieving significant FNR reductions while maintaining accurate FDR control.

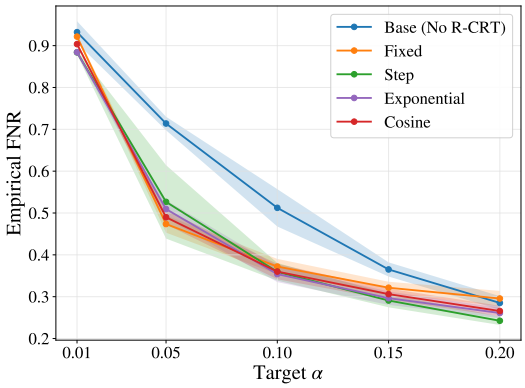

Figure 5: Empirical FNR at varying target risk levels (α) on the POLYP dataset. The proposed R-CRT models are trained under a target risk constraint of $\alpha = 0.1$. The solid lines represent the mean FNR across multiple independent runs, while the shaded regions denote the corresponding standard deviations. It can be observed that, after applying R-CRT, the proposed methods (except fixed α) significantly reduce the empirical FNR compared to the baseline across all evaluated target α values.

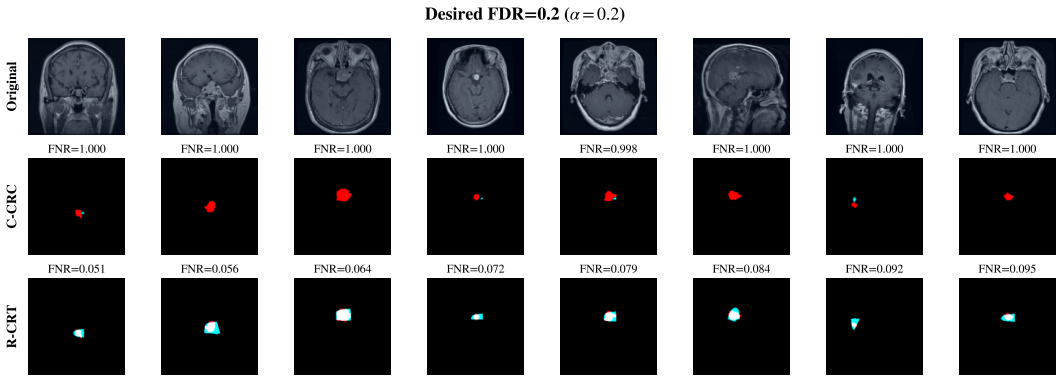

Figure 6: Qualitative comparison of CRC-C and R-CRT prediction sets at a desired FDR of $\alpha = 0.2$.

D Hyperparameter and Temperature Sensitivity Analysis

Tables 6 and 7 provide comprehensive ablation studies concerning initial threshold bounds (α_{start}) and temperature configurations ($\tau_{\text{FNR}}, \tau_{\text{FDR}}$).

We conduct ablation experiments comparing our per-sample soft-quantile computation against the single-threshold calculation adopted in the original CRT work. As shown in Table 8, the single-threshold scheme performs better when the base model exhibits extremely high FNR (near collapse), since the global threshold prevents excessive smoothing. In contrast, our method excels in typical scenarios where the base model possesses basic discriminative capacity. Both approaches are built upon the shared R-CRT framework.

E Additional Controlled Baseline Experiments and Discussion

To further validate the design choices and necessity of our R-CRT framework, we have added two sets of controlled baseline experiments. Starting from the identical pre-trained base model, we evaluate:

1. **Base (+10 Epochs):** The base model is trained with 10 additional epochs using the standard segmentation loss, strictly matching R-CRT's training budget.

Table 5: Comparison of Base models and different R-CRT annealing strategies across various datasets and Target α. Results are reported as Mean $\pm$ Std over 20 runs. For FNR, the relative improvement ($\uparrow$) or degradation ($\downarrow$) compared to the Base models is shown in parentheses. Best results in each group are highlighted in bold.

Dataset	Target α	Method	FDR	FNR $\downarrow$	IoU $\uparrow$	Dice $\uparrow$
FIRE	0.1	Base (No R-CRT)	0.095 ± 0.018	0.910 ± 0.006	0.090 ± 0.006	0.142 ± 0.008
		Fixed	0.093 ± 0.014	0.716 ± 0.005 ($\downarrow$ 21.3%)	0.280 ± 0.005	0.407 ± 0.006
		Step	0.092 ± 0.014	**0.628 ± 0.006 ($\downarrow$ 31.0%)**	**0.364 ± 0.006**	**0.502 ± 0.007**
		Exponential	0.098 ± 0.011	0.635 ± 0.006 ($\downarrow$ 30.2%)	0.357 ± 0.005	0.495 ± 0.007
		Cosine	0.097 ± 0.012	0.668 ± 0.006 ($\downarrow$ 26.5%)	0.325 ± 0.005	0.462 ± 0.007
	0.2	Base (No R-CRT)	0.198 ± 0.016	0.777 ± 0.024	0.218 ± 0.023	0.314 ± 0.029
		Fixed	0.200 ± 0.011	**0.250 ± 0.011 ($\downarrow$ 67.8%)**	**0.632 ± 0.003**	**0.757 ± 0.003**
		Step	0.202 ± 0.009	0.304 ± 0.022 ($\downarrow$ 60.9%)	0.595 ± 0.010	0.727 ± 0.009
		Exponential	0.201 ± 0.012	0.411 ± 0.080 ($\downarrow$ 47.1%)	0.518 ± 0.054	0.653 ± 0.053
		Cosine	0.200 ± 0.007	0.285 ± 0.014 ($\downarrow$ 63.3%)	0.609 ± 0.006	0.737 ± 0.005
CAR	0.1	Base (No R-CRT)	0.092 ± 0.022	0.911 ± 0.019	0.045 ± 0.014	0.078 ± 0.023
		Fixed	0.093 ± 0.020	0.687 ± 0.019 ($\downarrow$ 24.7%)	0.270 ± 0.019	0.385 ± 0.024
		Step	0.100 ± 0.019	0.673 ± 0.027 ($\downarrow$ 26.2%)	0.278 ± 0.022	0.393 ± 0.028
		Exponential	0.095 ± 0.016	**0.648 ± 0.015 ($\downarrow$ 28.9%)**	**0.302 ± 0.015**	**0.423 ± 0.018**
		Cosine	0.091 ± 0.020	0.652 ± 0.030 ($\downarrow$ 28.5%)	0.298 ± 0.029	0.422 ± 0.036
	0.2	Base (No R-CRT)	0.198 ± 0.020	0.662 ± 0.052	0.279 ± 0.043	0.398 ± 0.052
		Fixed	0.200 ± 0.018	**0.316 ± 0.018 ($\downarrow$ 52.2%)**	**0.542 ± 0.011**	**0.656 ± 0.012**
		Step	0.205 ± 0.019	0.352 ± 0.035 ($\downarrow$ 46.9%)	0.518 ± 0.021	0.638 ± 0.020
		Exponential	0.198 ± 0.024	0.374 ± 0.050 ($\downarrow$ 43.4%)	0.508 ± 0.030	0.629 ± 0.028
		Cosine	0.198 ± 0.024	0.343 ± 0.047 ($\downarrow$ 48.1%)	0.525 ± 0.025	0.644 ± 0.024

2. **Direct Loss Fine-tuning:** 10 epochs optimizing a surrogate weighted loss:

$$\mathcal{L}_{\text{direct}} = w_{\text{FDR}}\widetilde{\text{FDR}} + w_{\text{FNR}}\widetilde{\text{FNR}}, \tag{13}$$

tested under three weight settings: $(w_{\text{FDR}}, w_{\text{FNR}}) \in \{(1,0), (0.5,0.5), (0,1)\}$.

Based on the quantitative comparisons (Table 9), we derive two critical insights:

- **Extended training is insufficient:** As demonstrated by the "Base (+10 Epochs)" results, simply extending the training duration with standard segmentation losses yields only marginal or even negative utility gains, failing to address the fundamental conservatism of post-hoc calibration.
- **Direct loss is unstable and prone to collapse:** The core objective of conformal risk control is a *constrained optimization problem* (minimizing FNR subject to an FDR constraint at a specific target risk level α). Direct loss optimization fails because it applies a fixed, unconditional penalty that is completely unaware of the target α. Consequently, especially when the base model's initial predictions are relatively poor, the soft-FDR penalty heavily penalizes any false positives. The model takes the "lazy" route by predicting empty masks, causing the predicted FDR to approach 0 and the FNR to approach 1 (complete model collapse).

These results highlight R-CRT's absolute necessity. Unlike unstable direct losses, R-CRT's bi-level formulation explicitly couples representation learning with α-aware threshold selection. This aligns training with the final calibration objective, successfully avoiding trivial empty-mask solutions and consistently improving practical task utility.

Table 6: Hyperparameter analysis of different annealing strategies (Target $\alpha = 0.1$). Results are reported as Mean (Std) over 20 runs. Best results in each group are highlighted in bold.

Gap	α_{start}	Method	FDR	FNR	IoU
0.1	0.90	Fixed	0.085 (0.017)	**0.515 (0.031)**	0.447 (0.021)
		Step	0.093 (0.012)	0.539 (0.019)	0.438 (0.013)
		Exponential	0.089 (0.014)	0.546 (0.018)	0.435 (0.015)
		Cosine	0.078 (0.006)	0.526 (0.010)	**0.449 (0.009)**
	0.80	Fixed	0.096 (0.010)	0.603 (0.015)	0.383 (0.012)
		Step	0.100 (0.011)	0.510 (0.019)	0.465 (0.012)
		Exponential	0.095 (0.011)	0.469 (0.015)	0.487 (0.011)
		Cosine	0.094 (0.012)	**0.460 (0.015)**	**0.496 (0.012)**
	0.70	Fixed	0.097 (0.013)	0.481 (0.019)	0.481 (0.013)
		Step	0.095 (0.013)	0.487 (0.013)	0.485 (0.010)
		Exponential	0.099 (0.012)	**0.425 (0.013)**	**0.528 (0.010)**
		Cosine	0.097 (0.018)	0.473 (0.020)	0.482 (0.011)
	0.60	Fixed	0.090 (0.010)	0.443 (0.020)	0.497 (0.014)
		Step	0.095 (0.016)	0.455 (0.012)	0.508 (0.009)
		Exponential	0.100 (0.016)	0.455 (0.019)	0.500 (0.015)
		Cosine	0.097 (0.012)	**0.413 (0.017)**	**0.536 (0.013)**
	0.50	Fixed	0.096 (0.009)	0.420 (0.012)	0.520 (0.011)
		Step	0.098 (0.011)	0.400 (0.017)	0.536 (0.011)
		Exponential	0.102 (0.013)	**0.386 (0.015)**	**0.549 (0.008)**
		Cosine	0.098 (0.015)	0.400 (0.014)	0.541 (0.008)
	0.40	Fixed	0.102 (0.011)	0.403 (0.014)	0.527 (0.015)
		Step	0.096 (0.011)	**0.356 (0.019)**	**0.578 (0.013)**
		Exponential	0.097 (0.012)	0.372 (0.012)	0.559 (0.007)
		Cosine	0.098 (0.013)	0.385 (0.014)	0.553 (0.012)
	0.30	Fixed	0.092 (0.012)	0.365 (0.017)	0.578 (0.013)
		Step	0.097 (0.015)	0.407 (0.026)	0.551 (0.016)
		Exponential	0.099 (0.013)	**0.359 (0.017)**	**0.584 (0.010)**
		Cosine	0.096 (0.010)	0.376 (0.020)	0.581 (0.014)
	0.20	Fixed	0.101 (0.012)	**0.369 (0.019)**	**0.587 (0.013)**
		Step	0.101 (0.010)	0.368 (0.017)	0.582 (0.012)
		Exponential	0.100 (0.015)	0.446 (0.043)	0.524 (0.034)
		Cosine	0.096 (0.016)	0.418 (0.039)	0.545 (0.028)
0.2	0.90	Fixed	0.099 (0.017)	0.587 (0.026)	0.394 (0.019)
		Step	0.097 (0.015)	0.461 (0.017)	0.499 (0.012)
		Exponential	0.097 (0.010)	0.495 (0.015)	0.478 (0.011)
		Cosine	0.085 (0.013)	**0.452 (0.024)**	**0.509 (0.018)**
	0.70	Fixed	0.099 (0.016)	0.469 (0.016)	0.491 (0.010)
		Step	0.091 (0.017)	0.421 (0.019)	0.529 (0.010)
		Exponential	0.097 (0.016)	**0.411 (0.018)**	**0.544 (0.011)**
		Cosine	0.094 (0.010)	0.423 (0.016)	0.516 (0.013)
	0.50	Fixed	0.096 (0.011)	0.411 (0.016)	0.526 (0.012)
		Step	0.096 (0.019)	0.389 (0.030)	0.565 (0.020)
		Exponential	0.098 (0.015)	0.376 (0.018)	0.558 (0.012)
		Cosine	0.103 (0.010)	**0.354 (0.015)**	**0.573 (0.010)**
	0.30	Fixed	0.097 (0.013)	**0.375 (0.013)**	**0.561 (0.009)**
		Step	0.092 (0.023)	0.450 (0.044)	0.522 (0.034)
		Exponential	0.101 (0.015)	0.403 (0.029)	0.557 (0.019)
		Cosine	0.093 (0.016)	0.416 (0.026)	0.550 (0.018)

Table 7: Hyperparameter analysis of different annealing strategies (Target $\alpha = 0.1$). Results are reported as Mean (Std) over 20 runs. Best results in each configuration are highlighted in bold.

Setting	Value	Method	FDR	FNR	IoU
Varying τ_{fnr}	0.001	Fixed	0.097 (0.014)	0.377 (0.014)	0.566 (0.010)
		Step	0.097 (0.013)	0.378 (0.020)	0.573 (0.012)
		Exponential	0.100 (0.010)	**0.343 (0.011)**	**0.590 (0.009)**
		Cosine	0.097 (0.012)	0.353 (0.018)	0.590 (0.013)
	0.005	Fixed	0.092 (0.012)	0.365 (0.017)	0.578 (0.013)
		Step	0.097 (0.015)	0.407 (0.026)	0.551 (0.016)
		Exponential	0.099 (0.013)	**0.359 (0.017)**	**0.584 (0.010)**
		Cosine	0.096 (0.010)	0.376 (0.020)	0.581 (0.014)
	0.01	Fixed	0.099 (0.011)	0.376 (0.020)	0.557 (0.013)
		Step	0.099 (0.009)	0.351 (0.015)	**0.592 (0.013)**
		Exponential	0.103 (0.013)	0.372 (0.025)	0.580 (0.017)
		Cosine	0.102 (0.014)	**0.342 (0.018)**	0.591 (0.013)
	0.05	Fixed	0.096 (0.015)	0.403 (0.018)	0.538 (0.015)
		Step	0.101 (0.012)	**0.366 (0.014)**	0.574 (0.011)
		Exponential	0.077 (0.006)	0.392 (0.011)	0.560 (0.010)
		Cosine	0.095 (0.009)	0.367 (0.020)	**0.575 (0.015)**
	0.1	Fixed	0.074 (0.005)	0.435 (0.008)	0.519 (0.008)
		Step	0.045 (0.008)	0.760 (0.007)	0.238 (0.007)
		Exponential	0.091 (0.011)	**0.381 (0.017)**	0.560 (0.012)
		Cosine	0.096 (0.013)	0.386 (0.018)	**0.563 (0.014)**
	1	Fixed	0.082 (0.029)	0.504 (0.131)	0.459 (0.102)
		Step	0.389 (0.009)	**0.249 (0.013)**	0.410 (0.007)
		Exponential	0.369 (0.008)	0.260 (0.013)	0.410 (0.007)
		Cosine	0.184 (0.006)	0.329 (0.012)	**0.533 (0.011)**
Varying τ_{fdr}	0.001	Fixed	0.099 (0.015)	0.380 (0.017)	0.554 (0.010)
		Step	0.098 (0.014)	0.386 (0.026)	0.573 (0.018)
		Exponential	0.102 (0.013)	0.359 (0.017)	**0.587 (0.011)**
		Cosine	0.096 (0.013)	**0.353 (0.018)**	0.586 (0.013)
	0.005	Fixed	0.099 (0.011)	0.380 (0.018)	0.556 (0.014)
		Step	0.096 (0.013)	0.373 (0.016)	0.582 (0.010)
		Exponential	0.102 (0.012)	**0.343 (0.015)**	**0.585 (0.013)**
		Cosine	0.095 (0.014)	0.389 (0.025)	0.569 (0.016)
	0.01	Fixed	0.092 (0.012)	0.365 (0.017)	0.578 (0.013)
		Step	0.097 (0.015)	0.407 (0.026)	0.551 (0.016)
		Exponential	0.099 (0.013)	**0.359 (0.017)**	**0.584 (0.010)**
		Cosine	0.096 (0.010)	0.376 (0.020)	0.581 (0.014)
	0.05	Fixed	0.098 (0.009)	0.386 (0.015)	0.546 (0.013)
		Step	0.096 (0.013)	0.406 (0.021)	0.556 (0.015)
		Exponential	0.103 (0.012)	0.366 (0.016)	0.580 (0.011)
		Cosine	0.101 (0.014)	**0.334 (0.017)**	**0.595 (0.009)**
	0.1	Fixed	0.103 (0.016)	**0.374 (0.016)**	**0.567 (0.013)**
		Step	0.062 (0.039)	0.595 (0.157)	0.389 (0.139)
		Exponential	0.095 (0.022)	0.417 (0.078)	0.541 (0.067)
		Cosine	0.101 (0.015)	0.397 (0.029)	0.558 (0.019)
	1	Fixed	0.097 (0.006)	0.449 (0.013)	0.520 (0.012)
		Step	0.099 (0.004)	0.454 (0.006)	0.516 (0.007)
		Exponential	0.109 (0.007)	**0.441 (0.010)**	**0.526 (0.009)**
		Cosine	0.106 (0.007)	0.449 (0.013)	0.520 (0.012)

Table 8: Comparison of Base model, Pooled Single-Threshold annealing strategies, and the best Per-Sample Soft-Quantile variant across various datasets and Target α. Results are reported as Mean $\pm$ Std over 20 runs. For FNR, the relative improvement ($\downarrow$) or degradation ($\uparrow$) compared to the Base model is shown in parentheses. Best results in each group are highlighted in bold.

Dataset	Target α	Method	FDR	FNR $\downarrow$	IoU $\uparrow$	Dice $\uparrow$
ACDC	0.1	Base (No R-CRT)	0.095 ± 0.020	0.898 ± 0.009	0.058 ± 0.004	0.095 ± 0.006
		Fixed	0.102 ± 0.022	0.661 ± 0.029 ($\downarrow$ 26.4%)	0.274 ± 0.021	0.351 ± 0.024
		Step	0.094 ± 0.014	**0.551 ± 0.042 ($\downarrow$ 38.6%)**	**0.346 ± 0.025**	**0.416 ± 0.029**
		Exponential	0.092 ± 0.010	0.588 ± 0.032 ($\downarrow$ 34.5%)	0.317 ± 0.023	0.395 ± 0.027
		Cosine	0.100 ± 0.019	0.686 ± 0.029 ($\downarrow$ 23.6%)	0.249 ± 0.023	0.323 ± 0.028
		Soft-Quantile (Best)	*0.093 ± 0.016*	*0.756 ± 0.016 ($\downarrow$ 15.8%)*	*0.194 ± 0.014*	*0.267 ± 0.017*
	0.2	Base (No R-CRT)	0.203 ± 0.023	0.810 ± 0.126	0.136 ± 0.096	0.201 ± 0.107
		Fixed	0.200 ± 0.019	0.272 ± 0.022 ($\downarrow$ 66.4%)	0.523 ± 0.013	0.616 ± 0.014
		Step	0.196 ± 0.023	0.320 ± 0.015 ($\downarrow$ 60.5%)	0.532 ± 0.015	0.620 ± 0.016
		Exponential	0.194 ± 0.018	0.240 ± 0.020 ($\downarrow$ 70.4%)	0.573 ± 0.013	0.665 ± 0.014
		Cosine	0.202 ± 0.020	0.264 ± 0.020 ($\downarrow$ 67.4%)	0.578 ± 0.015	0.670 ± 0.016
		Soft-Quantile (Best)	*0.199 ± 0.020*	***0.181 ± 0.016 ($\downarrow$ 77.7%)***	***0.601 ± 0.014***	***0.699 ± 0.014***
BRISC	0.1	Base (No R-CRT)	0.087 ± 0.013	0.941 ± 0.005	0.059 ± 0.005	0.099 ± 0.007
		Fixed	0.024 ± 0.003	0.856 ± 0.010 ($\downarrow$ 9.0%)	0.140 ± 0.009	0.178 ± 0.011
		Step	0.089 ± 0.014	0.798 ± 0.008 ($\downarrow$ 15.2%)	0.198 ± 0.007	0.265 ± 0.009
		Exponential	0.100 ± 0.017	0.787 ± 0.010 ($\downarrow$ 16.4%)	0.208 ± 0.010	0.273 ± 0.013
		Cosine	0.097 ± 0.014	**0.749 ± 0.010 ($\downarrow$ 20.4%)**	**0.244 ± 0.010**	**0.309 ± 0.012**
		Soft-Quantile (Best)	*0.102 ± 0.021*	*0.855 ± 0.008 ($\downarrow$ 9.1%)*	*0.145 ± 0.007*	*0.206 ± 0.010*
	0.2	Base (No R-CRT)	0.205 ± 0.023	0.815 ± 0.106	0.178 ± 0.087	0.259 ± 0.101
		Fixed	0.205 ± 0.013	0.330 ± 0.009 ($\downarrow$ 59.5%)	0.569 ± 0.009	0.669 ± 0.009
		Step	0.199 ± 0.013	0.337 ± 0.031 ($\downarrow$ 58.7%)	0.581 ± 0.017	0.687 ± 0.018
		Exponential	0.197 ± 0.010	0.285 ± 0.010 ($\downarrow$ 65.0%)	0.612 ± 0.009	0.714 ± 0.009
		Cosine	0.196 ± 0.019	0.289 ± 0.016 ($\downarrow$ 64.5%)	0.611 ± 0.008	0.718 ± 0.008
		Soft-Quantile (Best)	*0.201 ± 0.010*	***0.231 ± 0.011 ($\downarrow$ 71.6%)***	***0.625 ± 0.006***	***0.735 ± 0.007***
POLYP	0.1	Base (No R-CRT)	0.095 ± 0.016	0.503 ± 0.035	0.475 ± 0.028	0.578 ± 0.027
		Fixed	0.096 ± 0.011	0.424 ± 0.012 ($\downarrow$ 15.7%)	0.537 ± 0.011	0.631 ± 0.012
		Step	0.094 ± 0.010	0.385 ± 0.016 ($\downarrow$ 23.5%)	0.567 ± 0.013	0.647 ± 0.013
		Exponential	0.098 ± 0.016	0.391 ± 0.019 ($\downarrow$ 22.3%)	0.561 ± 0.012	0.641 ± 0.012
		Cosine	0.097 ± 0.011	0.375 ± 0.016 ($\downarrow$ 25.4%)	0.568 ± 0.011	0.647 ± 0.012
		Soft-Quantile (Best)	*0.099 ± 0.013*	***0.359 ± 0.017 ($\downarrow$ 28.7%)***	***0.584 ± 0.010***	***0.664 ± 0.010***
	0.2	Base (No R-CRT)	0.194 ± 0.019	0.285 ± 0.016	**0.606 ± 0.010**	**0.692 ± 0.010**
		Fixed	0.183 ± 0.021	0.319 ± 0.017 ($\uparrow$ 11.9%)	0.575 ± 0.010	0.666 ± 0.011
		Step	0.192 ± 0.013	0.278 ± 0.013 ($\downarrow$ 2.5%)	0.592 ± 0.011	0.677 ± 0.011
		Exponential	0.188 ± 0.022	0.302 ± 0.011 ($\uparrow$ 6.0%)	0.583 ± 0.011	0.666 ± 0.010
		Cosine	0.180 ± 0.018	0.288 ± 0.011 ($\uparrow$ 1.1%)	0.583 ± 0.012	0.668 ± 0.011
		Soft-Quantile (Best)	*0.197 ± 0.020*	***0.247 ± 0.013 ($\downarrow$ 13.4%)***	*0.596 ± 0.014*	*0.680 ± 0.013*

Table 9: Comparison of baselines, different loss weight combinations (FDR vs FNR), and our best method across various datasets and Target α. Results are reported as Mean $\pm$ Std over 20 runs. Best results in each group are highlighted in bold.

Dataset	Target α	Method / Weights	FDR	FNR $\downarrow$	IoU $\uparrow$	Dice $\uparrow$
ACDC	0.1	Base (No R-CRT)	0.095 ± 0.020	0.898 ± 0.009	0.058 ± 0.004	0.095 ± 0.006
		Base (Same Epochs)	0.091 ± 0.015	0.896 ± 0.009	0.060 ± 0.003	0.100 ± 0.005
		Weights: 1.0 / 0.0	0.103 ± 0.020	0.956 ± 0.009	0.000 ± 0.000	0.000 ± 0.000
		Weights: 0.5 / 0.5	0.034 ± 0.052	0.946 ± 0.016	0.007 ± 0.011	0.012 ± 0.019
		Weights: 0.0 / 1.0	0.076 ± 0.027	0.932 ± 0.014	0.018 ± 0.008	0.032 ± 0.014
		Ours (Best)	0.093 ± 0.016	$\mathbf{0.756 \pm 0.016}$	$\mathbf{0.194 \pm 0.014}$	$\mathbf{0.267 \pm 0.017}$
	0.2	Base (No R-CRT)	0.203 ± 0.023	0.810 ± 0.126	0.136 ± 0.096	0.201 ± 0.107
		Base (Same Epochs)	0.198 ± 0.036	0.841 ± 0.018	0.112 ± 0.015	0.176 ± 0.021
		Weights: 1.0 / 0.0	0.186 ± 0.034	0.921 ± 0.100	0.022 ± 0.065	0.033 ± 0.099
		Weights: 0.5 / 0.5	0.098 ± 0.015	0.930 ± 0.008	0.023 ± 0.002	0.041 ± 0.003
		Weights: 0.0 / 1.0	0.172 ± 0.014	0.885 ± 0.007	0.057 ± 0.004	0.094 ± 0.006
		Ours (Best)	0.200 ± 0.015	$\mathbf{0.185 \pm 0.027}$	$\mathbf{0.607 \pm 0.013}$	$\mathbf{0.705 \pm 0.014}$
BRISC	0.1	Base (No R-CRT)	0.087 ± 0.013	0.941 ± 0.005	0.059 ± 0.005	0.099 ± 0.007
		Base (Same Epochs)	0.089 ± 0.012	0.932 ± 0.004	0.068 ± 0.004	0.111 ± 0.006
		Weights: 1.0 / 0.0	0.044 ± 0.006	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Weights: 0.5 / 0.5	0.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Weights: 0.0 / 1.0	0.068 ± 0.031	0.977 ± 0.013	0.022 ± 0.013	0.040 ± 0.022
		Ours (Best)	0.102 ± 0.021	$\mathbf{0.855 \pm 0.008}$	$\mathbf{0.145 \pm 0.007}$	$\mathbf{0.206 \pm 0.010}$
	0.2	Base (No R-CRT)	0.205 ± 0.023	0.815 ± 0.106	0.178 ± 0.087	0.259 ± 0.101
		Base (Same Epochs)	0.200 ± 0.017	0.794 ± 0.033	0.203 ± 0.031	0.293 ± 0.039
		Weights: 1.0 / 0.0	0.046 ± 0.007	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Weights: 0.5 / 0.5	0.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Weights: 0.0 / 1.0	0.177 ± 0.026	0.910 ± 0.013	0.086 ± 0.012	0.142 ± 0.019
		Ours (Best)	0.201 ± 0.010	$\mathbf{0.231 \pm 0.011}$	$\mathbf{0.625 \pm 0.006}$	$\mathbf{0.735 \pm 0.007}$
POLYP	0.1	Base (No R-CRT)	0.095 ± 0.016	0.503 ± 0.035	0.475 ± 0.028	0.578 ± 0.027
		Base (Same Epochs)	0.096 ± 0.012	0.498 ± 0.032	0.481 ± 0.026	0.582 ± 0.025
		Weights: 1.0 / 0.0	0.096 ± 0.013	0.751 ± 0.023	0.233 ± 0.020	0.332 ± 0.023
		Weights: 0.5 / 0.5	0.094 ± 0.015	0.455 ± 0.016	0.515 ± 0.011	0.610 ± 0.010
		Weights: 0.0 / 1.0	0.083 ± 0.009	0.865 ± 0.008	0.124 ± 0.007	0.185 ± 0.010
		Ours (Best)	0.099 ± 0.013	$\mathbf{0.359 \pm 0.017}$	$\mathbf{0.584 \pm 0.010}$	$\mathbf{0.664 \pm 0.010}$
	0.2	Base (No R-CRT)	0.194 ± 0.019	0.285 ± 0.016	0.606 ± 0.010	0.692 ± 0.010
		Base (Same Epochs)	0.192 ± 0.013	0.279 ± 0.013	$\mathbf{0.608 \pm 0.010}$	$\mathbf{0.698 \pm 0.010}$
		Weights: 1.0 / 0.0	0.183 ± 0.017	0.583 ± 0.013	0.355 ± 0.008	0.465 ± 0.010
		Weights: 0.5 / 0.5	0.193 ± 0.011	0.357 ± 0.011	0.525 ± 0.012	0.631 ± 0.012
		Weights: 0.0 / 1.0	0.176 ± 0.010	0.657 ± 0.008	0.287 ± 0.006	0.381 ± 0.008
		Ours (Best)	0.197 ± 0.020	$\mathbf{0.247 \pm 0.013}$	0.596 ± 0.014	0.680 ± 0.013

```python
import torch
import torch.nn.functional as F
import math

# ==========================================
# 1. Core Mathematical Utilities
# ==========================================
def soft_fnr_batch(preds, gts, lam, tau=0.01):
    preds_bin = torch.sigmoid((preds - lam) / tau)
    return (gts * (1.0 - preds_bin)).sum(dim=1) / (gts.sum(dim=1) + 1e-8)

def soft_fdr_global(preds, gts, lam, tau=0.01):
    preds_bin = torch.sigmoid((preds - lam) / tau)
    # fp + tp is mathematically equivalent to preds_bin.sum()
    return ((1.0 - gts) * preds_bin).sum() / (preds_bin.sum() + 1.0)

def soft_quantile(lam_batch, q, sigma=0.05):
    lam_sorted, _ = torch.sort(lam_batch)
    N = lam_sorted.size(0)
    idx = torch.arange(N, dtype=torch.float32, device=lam_batch.device)
    # The Gaussian weighting + normalization is exactly a softmax operation
    weights = F.softmax(-0.5 * ((idx - q * (N - 1)) / (sigma * N + 1e-6)) ** 2, dim=0)
    return (lam_sorted * weights).sum()

# ==========================================
# 2. Implicit Function Theorem (IFT) Autograd
# ==========================================
class LambdaFindIFT(torch.autograd.Function):
    @staticmethod
    def forward(ctx, preds, gts, alpha, tau):
        low, high = torch.zeros_like(preds[:, :1]), torch.ones_like(preds[:, :1])
        for _ in range(20):
            mid = (low + high) / 2.0
            mask = soft_fnr_batch(preds, gts, mid, tau).unsqueeze(1) > alpha
            high, low = torch.where(mask, mid, high), torch.where(~mask, mid, low)

        ctx.save_for_backward(preds, gts, mid)
        ctx.alpha, ctx.tau = alpha, tau
        return mid.squeeze(1)

    @staticmethod
    def backward(ctx, grad_out):
        preds, gts, lam = ctx.saved_tensors
        with torch.enable_grad():
            p, l = preds.detach().requires_grad_(), lam.detach().requires_grad_()
            g_sum = (soft_fnr_batch(p, gts, l, ctx.tau) - ctx.alpha).sum()
            # Compute both gradients simultaneously
            dg_dlam, dg_dp = torch.autograd.grad(g_sum, [l, p])

        grad_preds = -grad_out.unsqueeze(1) * (dg_dp / dg_dlam.clamp(min=1e-6))
        return grad_preds, None, None, None

# ==========================================
# 3. CRT Trainer and Annealing Logic
# ==========================================
def crt_trainer(loader, model, optimizer, alpha, opt):
    model.train()
    for imgs, gts in loader:
        if len(imgs) < 2: continue
        optimizer.zero_grad()

        half = len(imgs) // 2
        probs = torch.sigmoid(model(imgs))
        cal_p, pred_p = probs[:half].view(half, -1), probs[half:].flatten()
        cal_g, pred_g = gts[:half].view(half, -1), gts[half:].flatten()

        lam_batch = LambdaFindIFT.apply(cal_p, cal_g, alpha, opt.tau_fnr)
        lam_global = soft_quantile(lam_batch, 1.0 - opt.quantile)

        loss = soft_fdr_global(pred_p, pred_g, lam_global, opt.tau_fdr)
        if loss.item() > 0:
            loss.backward()
            torch.nn.utils.clip_grad_norm_(model.parameters(), opt.clip)
            optimizer.step()

def train_with_annealing(loader, model, optimizer, opt):
    for ep in range(opt.crt_epoch):
        p = ep / max(1, opt.crt_epoch - 1)
        a, a0 = opt.alpha, opt.alpha_start

        # Concise annealing mapping
        alpha = {
            'cosine': a + 0.5 * (a0 - a) * (1 + math.cos(math.pi * p)),
            'exponential': a + (a0 - a) * math.exp(-opt.decay_rate * p),
            'step': max(a, a0 - (ep // math.ceil(opt.crt_epoch / max(1, opt.steps))) * ((a0 - a) / max(1, opt.steps - 1)))
        }.get(opt.anneal_method, a0) # Default to fixed (a0)

        crt_trainer(loader, model, optimizer, alpha, opt)
```

Listing 1: Refined CRT Framework