UA-CP: UNCERTAINTY-AWARE CONFORMAL PREDICTION FOR RELIABLE ENDOSCOPIC LESION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable AI-assisted endoscopic lesion detection requires both accurate predictions and explicit uncertainty characterization, yet subtle appearances, ambiguous boundaries, and imbalanced category distributions challenge classification and localization. We propose UA-CP, a two-stage framework combining detector-adaptive Prior-Aware Training (PAT) with Dual Conformal Calibration (DCC). PAT incorporates category-frequency priors into classification objectives for proposal-based and dense detectors or assignment objectives for query-based detectors. DCC then constructs conformal category prediction sets using RAPS and simultaneous localization regions using conformalized quantile regression. Under exchangeability, these outputs provide finite-sample marginal coverage guarantees, with a joint lower bound obtained by the union bound. Across three complementary endoscopic benchmarks and four detector configurations, PAT generally improves detection performance, while DCC characterizes architecture-dependent trade-offs between uncertainty validity and efficiency. Results demonstrate that UA-CP generally improves detection utility while providing valid and informative category and localization uncertainty across diverse endoscopic detection settings.

1 INTRODUCTION

The integration of artificial intelligence into computer-assisted endoscopic diagnosis has shown strong potential for gastrointestinal (GI) disease screening by enabling automatic detection of subtle lesions and supporting clinical decision-making Jiang et al. (2023); Zhang et al. (2023). Reliable lesion detection is clinically important, as missed lesions may delay diagnosis, while excessive false alerts can increase clinicians' inspection burden Holzwanger et al. (2021); Liu et al. (2020). Despite strong benchmark performance, most deep learning-based detectors produce deterministic detections without characterizing uncertainty or the burden of false alerts, limiting their suitability for safety-critical clinical deployment Jungo & Reyes (2019); Kurz et al. (2022).

Unlike natural-scene object detection, endoscopic lesion detection presents reliability challenges across acquisition settings. Subtle variations in mucosal texture, color, contrast, and morphology can lead to missed detections or confusion between lesion categories, particularly when rare categories are underrepresented in training data Feng et al. (2021); Wu et al. (2024). Furthermore, diffuse or irregular lesion boundaries introduce uncertainty in spatial localization (Fig. 1).

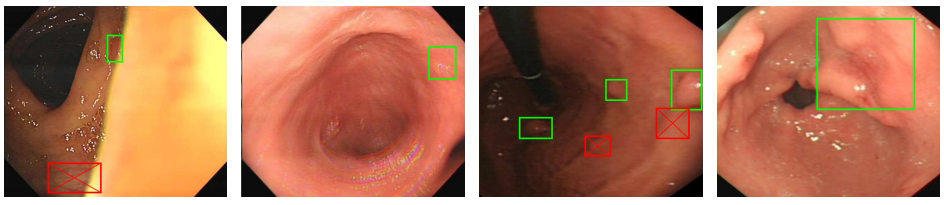

Figure 1: **Lesion appearance challenges:** In colonoscopy (a) and gastroscopy (c) images, lesions (□) may exhibit subtle visual differences from surrounding mucosa(□). **Lesion boundary challenges:** In colonoscopy (b) and gastroscopy (d) images, lesion boundaries may be diffuse or poorly delineated.

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.

These challenges motivate a reliability formulation that jointly considers detection utility and uncertainty characterization as complementary dimensions. Detection utility concerns whether lesions are detected while limiting false-alert burden Yamada et al. (2021); Ahmad et al. (2022), whereas uncertainty characterization quantifies ambiguity in lesion category and spatial localization. Standard detection pipelines, however, provide point predictions without calibrated measures of plausible categories or spatial extent Huseljic et al. (2025). Conformal prediction (CP) provides a model-agnostic, distribution-free framework for constructing prediction sets with finite-sample coverage guarantees under exchangeability Andéol et al. (2025). Existing CP-based object-detection methods are primarily developed for natural images with relatively well-defined object boundaries and calibrate the outputs of a fixed detector post hoc Timans et al. (2024). By leaving detector learning unchanged, these methods cannot directly address detector-level deficiencies arising from visual ambiguity, annotation noise, and class imbalance in endoscopy, which can limit both detection utility and uncertainty efficiency Cherubin (2023); Huang et al. (2023).

Motivated by the need to improve lesion detection while providing valid and compact uncertainty estimates, we propose Uncertainty-Aware Conformal Prediction (UA-CP), a framework that integrates detector-adaptive Prior-Aware Training (PAT) with Dual Conformal Calibration (DCC) for reliable endoscopic lesion detection. UA-CP consists of two complementary stages. First, PAT incorporates category-frequency information into an architecture-appropriate component of detector learning to improve detection utility under class imbalance. Second, DCC calibrates the intermediate detections to construct conformal category prediction sets and simultaneous localization regions, explicitly characterizing category and localization uncertainty. Unlike purely post-hoc CP approaches Su et al. (2024); Stutz et al. (2023); Yang & Pavone (2023), UA-CP couples training-time adaptation with conformal calibration to jointly consider detection utility and uncertainty characterization. We further assess detection utility, uncertainty validity, and uncertainty efficiency as distinct dimensions of reliability, providing an auditable evaluation framework for AI-assisted endoscopic lesion detection.

Our main contributions are as follows:

- We formulate reliable endoscopic lesion detection as jointly considering detection utility and uncertainty characterization, and assess uncertainty in terms of both validity and efficiency.
- We propose UA-CP, integrating detector-adaptive Prior-Aware Training (PAT) with Dual Conformal Calibration (DCC) to improve detection under category imbalance while constructing calibrated category prediction sets and localization regions.
- We evaluate UA-CP on three complementary frame-level endoscopic benchmarks and multiple detection paradigms, characterizing the trade-offs among detection utility, uncertainty validity, and efficiency for category and localization uncertainty.

2 Related Work

Uncertainty Quantification (UQ) in Object Detection Deep learning has significantly advanced endoscopic lesion detection, evolving from anchor-based detectors Ren et al. (2016); Redmon et al. (2016) to anchor-free and transformer-based architectures Carion et al. (2020); Tian et al. (2019); Zhang et al. (2022). Despite these advances, existing detectors typically produce point estimates of lesion category and location without explicitly characterizing their uncertainty. This limitation is especially critical in safety-sensitive medical imaging, where overconfident predictions may lead to unreliable clinical decisions Leibig et al. (2017); Liz-López et al. (2025). To adress this limitation, uncertainty estimation methods such as Bayesian neural networks Blundell et al. (2015); Gal & Ghahramani (2016) and deep ensembles Lakshminarayanan et al. (2017) approximate predictive uncertainty via stochastic inference or model averaging. Evidential Deep Learning (EDL) Agbelese et al. (2025) has been proposed to model both aleatoric and epistemic uncertainty in a single forward pass but does not provide formal statistical coverage guarantees.

Challenges in Endoscopy Lesion Detection Endoscopy-specific detectors have been developed for both colonoscopy Jiang et al. (2023); Wu et al. (2021); Xu et al. (2025) and gastroscopy Zhang et al. (2023; 2025), primarily targeting real-time and low-contrast settings. However, endoscopic lesion detection remains challenging due to heterogeneous lesion appearances, category imbalance, and inter-class visual similarity Tan et al. (2021; 2024); Yaman et al. (2023), which can complicate category

discrimination. Meanwhile, diffuse or poorly delineated lesion boundaries introduce ambiguity in spatial localization Agbelese et al. (2025). Inadequate characterization of such uncertainty can limit the reliability of automated systems in clinical settings López et al. (2025); Zou et al. (2023).

Conformal Prediction for Object Detection Conformal prediction (CP) Vovk (2015) provides a distribution-free framework for uncertainty quantification with finite-sample coverage guarantees. In standard settings, Regularized Adaptive Prediction Sets (RAPS) Angelopoulos et al. (2020) construct calibrated prediction sets for classification, while conformalized quantile regression (CQR) Romano et al. (2019) provides CP intervals for regression. Recent works have extended CP to object detection, quantifying uncertainty in class predictions, object localization, or their downstream propagation Copley et al. (2024); Timans et al. (2024); Su et al. (2024); Yang & Pavone (2023). However, existing approaches primarily focus on natural-image and autonomous-perception settings, leaving conformal uncertainty for endoscopic lesion detection underexplored. This motivates tighter integration of uncertainty-aware learning to improve reliability under challenging endoscopic conditions.

3 PRELIMINARIES

Conformal prediction Conformal prediction (CP) is a distribution-free framework for constructing prediction sets with finite-sample coverage guarantees. Let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{N}$ denote a held-out calibration set, where the calibration samples and the test sample (X_{N+1}, Y_{N+1}) are assumed to be exchangeable. Given a nonconformity score function $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where larger values indicate poorer agreement between an input–label pair and the calibration data, we compute calibration scores $s_i = S(X_i, Y_i)$ for $i = 1, \ldots, N$.

For a target miscoverage level $\alpha \in (0, 1)$, define the conformal threshold as

$$q_\alpha = \text{Quantile}_{\lceil (N+1)(1-\alpha) \rceil / N} \left(\{s_i\}_{i=1}^{N} \right). \tag{1}$$

The conformal prediction set for a new input X_{N+1} is then given by

$$\hat{\mathcal{C}}(X_{N+1}) = \{y \in \mathcal{Y} : S(X_{N+1}, y) \leq q_\alpha\}. \tag{2}$$

Under exchangeability, this construction satisfies the marginal finite-sample coverage guarantee

$$\mathbb{P}\left(Y_{N+1} \in \hat{\mathcal{C}}(X_{N+1})\right) \geq 1 - \alpha. \tag{3}$$

Importantly, this guarantee does not require assumptions on the underlying data distribution or the predictive model Papadopoulos (2008); Vovk et al. (2022).

Conformalized quantile regression Standard CP provides valid marginal coverage but may yield inefficient prediction sets under heterogeneous noise. Conformalized Quantile Regression (CQR) addresses this by conformalizing conditional quantile estimates Romano et al. (2019). Let $\hat{q}_\tau(x)$ denote an estimator of the conditional τ-quantile of $Y|X = x$. For a desired miscoverage level α, a nominal quantile-regression interval is first constructed as

$$\left[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)\right]. \tag{4}$$

To obtain finite-sample validity, CQR calibrates this interval using the nonconformity score

$$s_i = \max\left\{\hat{q}_{\alpha/2}(X_i) - Y_i,\, Y_i - \hat{q}_{1-\alpha/2}(X_i)\right\}. \tag{5}$$

Let q_α^{CQR} be the conformal quantile of these scores, computed in the same way as above. The calibrated CQR prediction interval for a test input X_{N+1} is

$$\hat{\mathcal{C}}_{\text{CQR}}(X_{N+1}) = \left[\hat{q}_{\alpha/2}(X_{N+1}) - q_\alpha^{\text{CQR}}, \hat{q}_{1-\alpha/2}(X_{N+1}) + q_\alpha^{\text{CQR}}\right]. \tag{6}$$

This conformal correction preserves the finite-sample marginal coverage guarantee while allowing the interval width to adapt to input-dependent uncertainty.

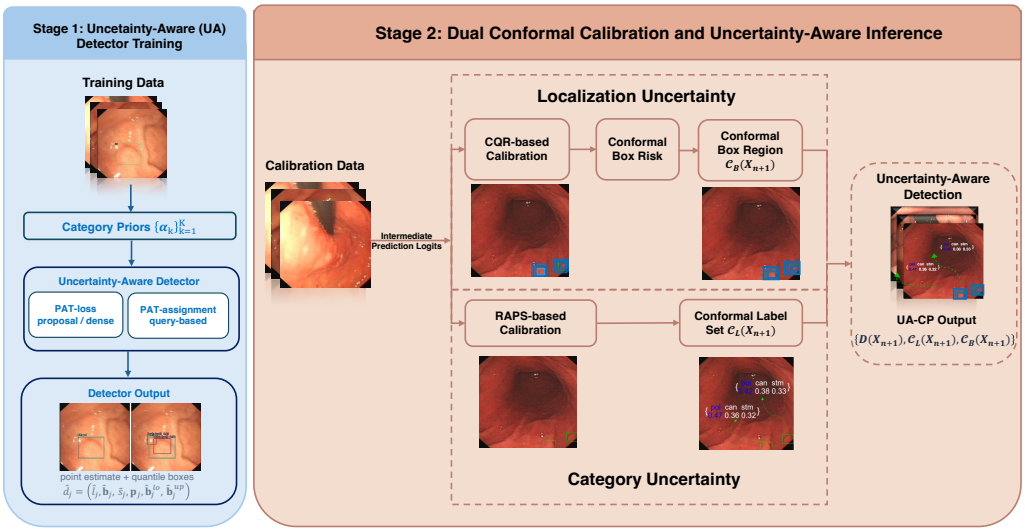

Figure 2: **Overview of UA-CP** UA-CP consists of two stages. **Stage 1** performs uncertainty-aware (UA) detector training with architecture-adaptive PAT. **Stage 2** applies dual conformal calibration (DCC): RAPS constructs conformal label sets for category uncertainty, while CQR constructs conformal box regions for localization uncertainty. The final output combines point detections with calibrated category and localization uncertainty.

4 METHODOLOGY

4.1 PROBLEM FORMULATION

Given an endoscopic dataset $\mathcal{D}$, we divide it into three mutually disjoint partitions,

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}, \qquad \mathcal{D}_a \cap \mathcal{D}_b = \emptyset, \text{ for } a \neq b$$

We use $\mathcal{D}_{\text{train}}$ for detector training, $\mathcal{D}_{\text{cal}}$ for conformal calibration, and $\mathcal{D}_{\text{test}}$ for evaluating the detector and the proposed framework. Conformal validity assumes exchangeability between matched calibration and validation detections. Dataset-specific split construction is reported in Appendix C.

Let label space $\mathcal{K} = \{1, \ldots, K\}$, for an input $X_{N+1} \in \mathcal{D}_{\text{cal}}$, let

$$\mathcal{G}(X_{N+1}) = \{(l_i, \mathbf{b}_i)\}_{i=1}^{N_G},$$

where $l_i \in \mathcal{K}$ is the lesion category and $\mathbf{b}_i = (x_1, y_1, x_2, y_2) \in \mathbb{R}^4$ is the corresponding bounding box. After training, the detector returns

$$\widehat{D}(x) = \left\{\widehat{d}_j\right\}_{j=1}^{N_D(x)}, \qquad \widehat{d}_j = \left(\widehat{l}_j, \widehat{\mathbf{b}}_j, \widetilde{s}_j, \mathbf{p}_j, \widehat{\mathbf{b}}_j^{lo}, \widehat{\mathbf{b}}_j^{up}\right), \tag{7}$$

where $\widehat{l}_j \in \mathcal{K}, \widehat{\mathbf{b}}_j \in \mathbb{R}^4$ is the predicted category and box, respectively; $\mathbf{p}_j = (p_{j,1}, \ldots, p_{j,K}) \in \Delta^{K-1}$, $\Delta^{K-1} = \left\{\mathbf{p} \in \mathbb{R}^K : p_k \geq 0, \sum_{k=1}^K p_k = 1\right\}$; $\tilde{s}_j = p_{j,\widehat{l}_j}$ denotes the predicted detector score; $\widehat{\mathbf{b}}_j^{lo}, \widehat{\mathbf{b}}_j^{up} \in \mathbb{R}^4$ are the lower- and upper-quantile box predictions.

UA-CP preserves each point detection and augments it with calibrated detection uncertainty:

$$\widehat{\mathcal{O}}(X_{N+1}) = \{\hat{o}_j\}_{j=1}^{N_D(X_{N+1})}, \qquad \hat{o}_j = \left(\widehat{d}_j, \widehat{\mathcal{C}}_{L,j}(x), \widehat{\mathcal{C}}_{B,j}(x)\right).$$

where $\widehat{\mathcal{C}}_{L,j}(x) \in \mathcal{K}$ ithe conformal class-label set, and $\widehat{\mathcal{C}}_{B,j}(X) = \prod_{d=1}^4 \left[\widehat{L}_{d,j}(X), \widehat{U}_{d,j}(X)\right]$ is the conformal bounding-box interval.

4.2 UNCERTAINTY-AWARE LEARNING

Class imbalance can weaken predictions for rare lesion categories by biasing confidence estimates Dong et al. (2025). We introduce *Prior-Aware Training* (PAT), a detector-adaptive strategy that incorporates category-frequency priors into either the classification objective (PAT-loss) or the assignment objective (PAT-assignment). Specifically, let n_k denote the number of training samples for category k, and define the category-dependent prior weight as $\alpha_k = 1/\sqrt{|n_k|}$. Depending on the detector architecture, the prior weights $\{\alpha_k\}_{k=1}^{K}$ are incorporated according to how predictions are associated with ground-truth objects.

Loss-level PAT. For detectors with explicit classification heads, PAT augments the detector's native classification objective with an error-adaptive, prior-weighted group penalty. Let $W = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}^{D\times K}$ denote the foreground classifier parameters, where $\mathbf{w}_k \in \mathbb{R}^D$ is the learnable parameter vector associated with category k. We define

$$\mathcal{L}_{PAT}^{\text{loss}} = -\frac{1}{N_{\text{cls}}}\sum_{i=1}^{N_{\text{cls}}} \log p_{i,l_i} + \lambda_{\text{loss}} \sum_{k=1}^{K} \alpha_k \left(1 - \mathbb{E}_{\{l_i=k\}}\left[p_{i,l_i}\right]\right) \|\mathbf{w}_k\|_2 \tag{8}$$

where p_{i,l_i} denotes the predicted probability of the ground-truth category, N_{cls} denotes the number of classification units, and $\mathbb{E}_{\{i:\, l_i=k\}}[p_{i,k}]$ is the mean confidence for category k. The prior weight α_k emphasizes rare categories, while λ_{loss} controls PAT contribution.

Assignment-level PAT. For query-based detectors, one-to-one query–target assignment couples classification and localization. Under class imbalance, the standard matching objective may favor frequent classes. We therefore incorporate category-frequency information into the Hungarian matching cost. For a query-target pair (i, k), the prior-aware matching term is

$$\text{Cost}_{PAT-A}(i,k) = -\gamma \sum_{k=1}^{K} p_{i,l_i} \log(p_{i,l_i} + \varepsilon) + \lambda_{\text{cost}} \alpha_k e_{i,k}, \qquad e_{i,k} = 1 - p_{i,k} \tag{9}$$

The first term measures predictive uncertainty, while the error term $e_{q,i}$ measures the query's lack of confidence in the ground-truth category. γ and λ_{cost} control the contributions of predictive uncertainty and prior-weighted classification error, respectively, and $\varepsilon > 0$ ensures numerical stability. The resulting Hungarian matching cost is:

$$\text{Cost}_{\text{match}}^{PAT}(i,k) = \text{Cost}_{cls}(i,k) + \text{Cost}_{bbox}(i,k) + \text{Cost}_{IoU}(i,k) + \text{Cost}_{PAT-A}(i,k) \tag{10}$$

4.3 DUAL CONFORMAL CALIBRATION

Given the intermediate detections for a held-out input image $X_{N+1} \in \mathbb{R}^{H\times W\times D}$, with height H, width W and D channels, we apply *Dual Conformal Calibration* (DCC) to quantify category and localization uncertainty through a unified prediction interface. Following the matching procedure in Appendix B.1, let $\mathcal{M}(X)$ denote the matched detection pairs for image X. The matched calibration collection is

$$\mathcal{M}_{\text{cal}} = \bigcup_{X\in\mathcal{D}_{\text{cal}}} \mathcal{M}(X), \qquad N_D = |\mathcal{M}_{\text{cal}}|\,.$$

Category calibration. For each detection j, DCC quantifies category uncertainty with RAPS Angelopoulos et al. (2020). Let $\hat{\pi}_j(k) = p_{j,k}$ denote the predicted probability of category $k \in \mathcal{K}$, $\rho_j(k)$ denote the cumulative probability mass of categories ranked above k, and $o_j(k)$ denote the rank of category k. The RAPS score is

$$S_j^L(k; u_j) = \rho_j(k) + \hat{\pi}_j(k)\cdot u_j + \lambda_R(o_j(k) - k_{\text{reg}})^+ \qquad u_j \sim \text{Unif}(0,1) \tag{11}$$

where operator $(\cdot)^+$ denotes the positive part, while λ_R and k_{reg} control the rank regularization. For each matched calibration detection r with ground-truth category $l_r \in \mathcal{K}$, the nonconformity score is $S_r^L(l_r; u_r)$. Given the target label-miscoverage level α_L, the conformal threshold is

$$\widehat{q}_L = \text{Quantile}_{\frac{\lceil (N_D+1)(1-\alpha_L)\rceil}{N_D}}\left(\left\{S_r^L(l_r,u_r) : l_r = k\right\}_{r=1}^{N_D}\right). \tag{12}$$

For a detection j in $X_{N+1} \in \mathcal{D}_{\text{test}}$, the conformal label set is

$$\hat{\mathcal{C}}_L(X_{N+1}) = \left\{k \in \mathcal{K} : S_j^L(k; u_j) \le \widehat{q}_L\right\} \qquad P(l_{N+1} \in \hat{\mathcal{C}}_L(X_{N+1})) \ge 1 - \alpha_L. \tag{13}$$

Under exchangeability of matched calibration and test detections, $\widehat{\mathcal{C}}_L(X)$ provides $1 - \alpha_L$ category coverage while leaving the detector prediction $\widehat{l}_j$ unchanged.

Localization calibration To quantify localization uncertainty, each matched detection r includes lower- and upper-quantile box predictions $\widehat{\mathbf{b}}_i^{lo}, \widehat{\mathbf{b}}_i^{up} \in \mathbb{R}^4$, with corresponding ground-truth box $\mathbf{b}_r \in \mathbb{R}^4$. We define CQR residual

$$S_{r,d}^B = \max\left\{\widehat{b}_{r,d}^{lo} - b_{r,d},\ b_{r,d} - \widehat{b}_{r,d}^{up}\right\}, \qquad d \in \{1, \ldots, 4\} \tag{14}$$

To control simultaneous four-coordinates miscoverage at level α_B, the conformal correction is

$$\hat{q}_{B,d} = \text{Quantile}_{\frac{\lceil (N_D+1)(1-\alpha_B/4) \rceil}{N_D}}\left(\{S_{r,d}^B\}_{r=1}^{N_D}\right) \tag{15}$$

The conformal bounding-box region for detection j is

$$\widehat{\mathcal{C}}_{B,j}(X_{N+1}) = \prod_{d=1}^{4}\left[\widehat{b}_{d,j}^{lo} - \hat{q}_{B,d}^{j},\ \widehat{b}_{d,j}^{up} + \hat{q}_{B,d}^{j}\right] = \prod_{d=1}^{4}\left[\widehat{L}_{d,j}, \widehat{U}_{d,j}\right]. \tag{16}$$

Under exchangeability of matched calibration and validation detections, the conformal box region provides finite-sample coverage at level $1 - \alpha_B$ for simultaneous containment of all four ground-truth box coordinates. Details are provided in Appendix B.3.

Detection-level output and coverage. Combining the detector predictions with the calibrated category $\widehat{\mathcal{C}}_L(X_{N+1})$ and localization $\widehat{\mathcal{C}}_B(X_{N+1})$ outputs, the resulting UA-CP output is

$$\widehat{\mathcal{O}}(X_{N+1}) = \left\{\widehat{D}(X_{N+1}), \widehat{\mathcal{C}}_L(X_{N+1}), \widehat{\mathcal{C}}_B(X_{N+1})\right\} \tag{17}$$

Under exchangeability of matched calibration and validation detections, the union bound gives the following theoretical joint coverage guarantee for a matched validation detection

$$P(l_{N+1} \in \widehat{\mathcal{C}}_{L,j}(X_{N+1}),\ \mathbf{b}_{N+1} \in \widehat{\mathcal{C}}_{B,j}(X_{N+1})) \ge 1 - \alpha_L - \alpha_B \tag{18}$$

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate UA-CP on three endoscopic lesion detection datasets spanning complementary category and localization uncertainty settings: LGLDDZhang et al. (2023), Kvasir-CapsuleBorgli et al. (2020) and LDPolypVideoMa et al. (2021). LGLDD, a long-tailed gastroscopy dataset, contains 10,083 gastroscopic images with 12,292 lesion boxes across four categories: polyp (*pol*), ulcer (*uls*), cancer (*can*), and submucosal tumor (*stm*). Kvasir-Capsule contains 4,224 capsule-endoscopy images across nine severely imbalanced categories, while LDPolypVideo provides a single-class setting with 160 colonoscopy clips and 33,884 polyp-positive frames. Predictions are made independently at the image/frame level without temporal modeling. Dataset and split details are provided in Appendix C.

Implementation Details. We evaluate UA-CP on four detector configurations spanning three paradigms: proposal-based Faster R-CNN Ren et al. (2016) with ResNet-50 and ResNeXt-101 backbones, dense one-stage FCOS Tian et al. (2019), and query-based DINO Zhang et al. (2022). All backbones are initialized from ImageNet-pretrained weights, and models are trained on 4× NVIDIA RTX 2080 GPUs using Adam with a learning rate of 1×10^{-4}. For the primary uncertainty evaluation, we set $\alpha_L = \alpha_B = 0.05$, yielding a theoretical joint coverage coverage lower bound of $1 - \alpha_L - \alpha_B = 0.9$. For RAPS, we use $k_{\text{reg}} = 2$ and $\lambda_R = 0.01$. Coverage–efficiency trade-offs are examined in Appendix D.2, with further implementation details in Appendix B.

5.2 Evaluation Metrics

We evaluate the UA-CP from two conplementary dimension: detection utility, detection validity, and uncertainty effectiveness.

Dectection Utility We evaluate detection performance using COCO-style average precision (*AP*), *recall* (R), *F1*-score, and false positives per image ($\mathrm{FP/image}$). Let $\mathcal{T} = \{0.50, 0.55, \ldots, 0.95\}$ denote the evaluated IoU thresholds. Then

$$AP = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathrm{AP}_\tau, \qquad \mathrm{AP}_\tau = \sum_{n=0}^{m-1} (r_{n+1} - r_n) \rho_{\mathrm{inter},\tau}(r_{n+1}), \tag{19}$$

where $\rho_{inter}(r_{n+1}) = \max_{\tilde{r}:\tilde{r} \geq r_{n+1}} \rho(\tilde{r})$. Here, $\rho(r_n)$ denotes the precision at recall $\tilde{r}$. The remaining detection metrics are

$$R = \frac{TP}{TP + FN}, \qquad F1 = \frac{2 \times TP}{2TP + FP + FN} \qquad \mathrm{FP/image} = \frac{N_{\text{pred}} - N_{\text{TP}}}{N_{\text{images}}} \tag{20}$$

Here, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. $\mathrm{FP/image}$ measures the average number of unmatched predictions per image.

Detection Validity Based on matched validation collection $\mathcal{M}_{\text{test}}$, We further evaluate the detection validity to measure whether UA-CP provides valid category and localization coverage.

$$Cov_L = \frac{1}{|\mathcal{M}_{\text{test}}|} \sum_{(X,i,j) \in \mathcal{M}_{\text{test}}} \mathbf{1}\left(l_k \in \hat{\mathcal{C}}_{L,j}(X_)\right), \tag{21}$$

where $\mathbf{1}$ is the indicator function. Localization coverage requires simultaneous containment of all four ground-truth box coordinates:

$$Cov_B = \frac{1}{|\mathcal{M}_{\text{test}}|} \sum_{(X,i,j) \in \mathcal{M}_{\text{test}}} \prod_{d=1}^{4} \mathbf{1}\left[\widehat{L}_{d,j}(X) \leq b_{d,i} \leq \widehat{U}_{d,j}(X)\right] \tag{22}$$

A nominal lower bounds are $1 - \alpha_L$ for category coverage, $1 - \alpha_B$ for localization coverage, and $1 - \alpha_L - \alpha_B$ for joint coverage by union bound.

Uncertainty Efficiency. We further assess uncertainty efficiency using the set size (*Size*) and mean prediction interval width (*MPIW*)Andéol et al. (2024):

$$MPIW = \frac{1}{4|\mathcal{D}_{\text{test}}|} \sum_{j=1}^{|\mathcal{D}_{\text{test}}|} \sum_{k=1}^{4} \left|\widehat{U}_d(X_j) - \widehat{L}_d(X_j)\right|, \qquad \text{Size} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{j=1}^{|\mathcal{D}_{\text{test}}|} |\widehat{\mathcal{C}}_L(X_j)| \tag{23}$$

Smaller label sets indicate and narrower intervals indicate more informative uncertainty outputs, but efficiency is compared only between methods attaining comparable coverage.

5.3 Quantitative Results

Comparison with Detection Baseline We compare each detector baseline with its PAT-trained counterpart under identical data and inference protocols across the three benchmarks. YOLOv5 Jocher et al. (2020) serves as an additional detection reference.

As demonstrated in Table 1, UA-CP generally improves detection performance, although gains vary across architectures and datasets. On LGLDD, Faster R-CNN-ResNeXt-101 yields the largest AP gain (**37.4** $\rightarrow$ **42.8**) and reduces FP/image (**2.77** $\rightarrow$ **0.40**), with lower recall (**89.2** $\rightarrow$ **81.3**). On Kvasir-Capsule, DINO-R50 achieves the highest AP (**63.5**), while F1 and FP/image gains are less consistent. On LDPolypVideo, UA training improves recall across all architectures; DINO-R50 increases recall (**52.8** $\rightarrow$ **65.9**) while reducing FP/image (**25.43** $\rightarrow$ **5.64**). Additional per-dataset and per-category results are provided in Appendix D.1.

Table 1: **Quantitative Results for detection utility:** Detection utility across the three endoscopic benchmarks. We report AP(%), AP50(%), worst-category $\mathrm{AP}_{\mathrm{Worst}}$ (%), recall(%), F1(%), and FP/image. $\mathrm{AP}_{\mathrm{Worst}}$ denotes the minimum category-wise AP, with the evaluated category fixed within each baseline–UA-CP pair but potentially differing across detectors.

Dataset	Method	AP↑	AP50	$\mathrm{AP}_{\mathrm{Worst}}$	Recall↑	F1↑	FP/image↓
LGLDD	Faster RCNN-R50	32.7	70.2	23.4	79.1	68.2	0.63
	Faster RCNN-Rxt 101	37.4	74.4	24.0	<u>89.2</u>	42.1	2.77
	DINO-R50	38.6	<u>78.5</u>	27.1	87.4	46.2	2.26
	FCOS	29.5	66.8	19.8	80.8	61.9	0.95
	YOLOv5	37.2	74.2	27.1	**89.5**	50.5	1.96
LGLDD	Faster RCNN-R50	36.1	73.9	25.7	78.8	<u>74.1</u>	**0.40**
UA-CP (Ours)	Faster RCNN-Rxt101	**42.8**	76.9	29.6	81.3	**75.5**	**0.40**
	FCOS	31.6	69.5	18.4	81.3	62.8	0.92
	DINO-R50	<u>39.9</u>	**80.7**	30.7	85.6	69.6	0.72
Kvasir-Capsule	Faster RCNN-R50	38.7	80.1	19.0	89.8	27.8	4.57
	Faster RCNN-Rxt 101	57.5	91.5	34.1	94.5	61.3	1.14
	DINO-R50	<u>60.9</u>	92.8	37.7	94.1	86.2	0.24
	FCOS	58.5	<u>93.9</u>	42.8	95.5	**90.0**	<u>0.17</u>
	YOLOv5	55.8	93.3	45.1	**96.2**	57.6	1.38
Kvasir-Capsule	Faster RCNN-R50	44.4	82.2	22.8	89.3	35.6	3.13
UA-CP (Ours)	Faster RCNN-Rxt 101	60.1	92.7	35.3	92.9	88.9	**0.16**
	DINO-R50	**63.5**	93.6	40.3	<u>95.7</u>	<u>89.0</u>	0.19
	FCOS	58.6	**95.5**	43.9	95.2	86.8	0.24
LDPolypVedio	Faster RCNN-R50	25.9	51.3	–	29.7	37.6	2.80
	Faster RCNN-Rxt 101	24.4	<u>59.3</u>	–	52.4	32.2	**0.56**
	DINO-R50	25.8	54.7	–	52.8	27.3	25.43
	FCOS	22.1	40.5	–	27.1	24.2	<u>0.59</u>
	YOLOv5	25.9	53.4	–	53.0	50.1	0.61
LDPolypVideo	Faster RCNN-R50	26.1	55.1	–	56.9	**62.5**	5.91
UA-CP (Ours)	Faster RCNN-Rxt 101	**28.1**	**69.1**	–	57.9	<u>61.9</u>	1.65
	DINO-R50	<u>27.4</u>	56.3	–	**65.9**	51.1	5.64
	FCOS	21.7	51.8	–	<u>58.0</u>	49.3	0.77

Table 2: **Category and localization uncertainty on LGLDD.** Comparison of DCC with detector-specific conformal baselines across architectures. Results are reported as mean ± standard deviation across repeated calibration runs.

Detector	Method	Cov_L.	Cov_B.	*Join Cov.*	*Size.*	*MPIW.*
FasterRCNN	Box-Std	0.98±0.000	0.97±0.005	0.95±0.005	1.00±0.000	46.22±0.49
R50	DCC	0.98±0.000	0.82±0.012	0.80±0.011	1.04±0.002	45.90±1.15
FasterRCNN	Box-Std	0.96±0.000	0.94±0.006	0.90±0.006	1.00±0.000	30.64±1.04
Rxt101-UA	DCC	0.98±0.001	0.95±0.005	0.94±0.005	1.03±0.001	30.58±0.80
DINO	Class-thr	0.91±0.003	0.76±0.013	0.71±0.013	1.14±0.006	42.13±1.61
	DCC	0.98±0.000	0.85±0.012	0.84±0.011	2.16±0.011	61.63±1.39
Fcos	Box-Std	0.93±0.001	0.89±0.008	0.84±0.007	1.00±0.0002	80.55±2.64
	DCC	0.97±0.001	0.90±0.008	0.88±0.007	1.12±0.004	84.14±2.80

Category and localization uncertainty evaluation At $\alpha_L = \alpha_B = 0.05$, corresponding to coverage targets of 0.95 and a conservative joint target of 0.90, we compare *DCC* against the detector-specific Box-Std baseline for Faster R-CNNTimans et al. (2024) and conformal class thresholding (ClassThr) for DINO and FCOS Sadinle et al. (2019). As shown in Table 2, UA-CP generally maintains high label coverage while improving joint coverage over the corresponding baseline on LGLDD. Faster R-CNN-ResNeXt-101 attains **0.94** joint coverage with a compact MPIW of **30.58**. Results on LDPolypVideo and Kvasir-Capsule are provided in Appendix D.

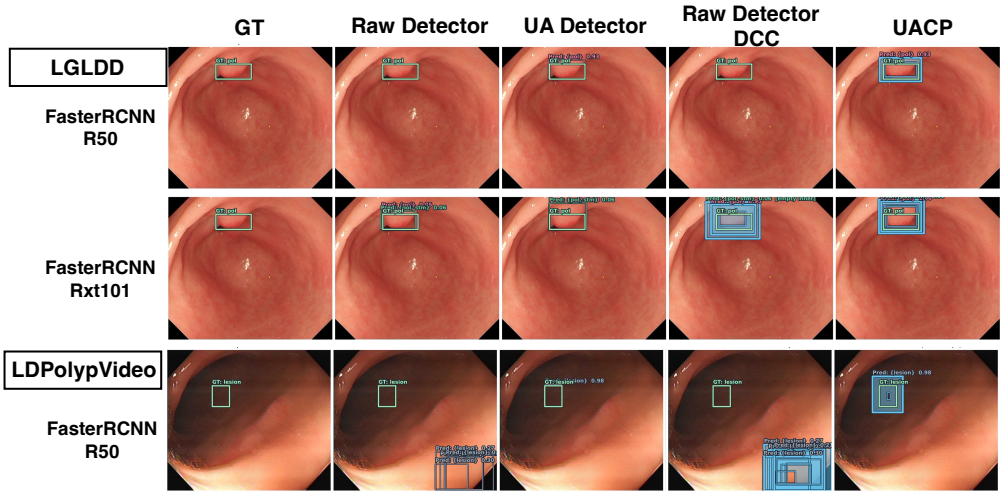

Figure 3: **Qualitative Results of UA-CP.** From left to right, each row shows the ground truth, raw detector output, UA-trained detector output, DCC applied to the raw detections, and the UA-CP output. **L1**: Faster R-CNN (R50) on LGLDD; **L2**: Faster R-CNN (RxT101) on LGLDD; and **L3**: Faster R-CNN (R50) on LDPolypVideo. Colored boxes indicate lesion categories, while dark-blue boxes denote conformal localization regions.

5.4 Qualitative Results

Figure 3 illustrates the progression from raw detector predictions to UA-CP outputs. UA improves candidate quality, enabling DCC characterize the remaining localization uncertainty through conformal regions. Additional challenging cases and uncertainty patterns are provided in Appendix E.

5.5 Ablation Study

The ablation results (Table 3) demonstrate that UA consistently improves AP and more compact localization uncertainty regions while maintaining stable label coverage.

Table 3: Abalation Study on LGLDD.

Detector	**UA**	**AP**	Cov_L.	Cov_B.	*Join Cov.*	*Size*	*MPIW*
FasterRCNN	–	32.7	0.97±0.001	0.74±0.009	0.72±0.009	1.06±0.003	64.35±2.08
R50	√	36.1	0.98±0.000	0.82±0.012	0.80±0.012	1.04±0.002	45.90±1.15
FasterRCNN	–	37.4	0.97±0.001	0.93±0.004	0.90±0.004	1.12±0.003	103.93±2.34
Rxt101	√	42.8	0.9700±0.001	0.90±0.008	0.88±0.007	1.12±0.004	84.14±2.80
DINO	–	38.6	0.98±0.001	0.82±0.007	0.82±0.007	1.73±0.030	80.40±5.79
	√	39.9	0.98±0.001	0.85±0.012	0.84±0.011	2.16±0.011	61.63±1.39
Fcos	–	29.5	0.97±0.001	0.93±0.004	0.90±0.004	1.12±0.003	103.93±2.34
	√	31.6	0.97±0.001	0.90±0.008	0.88±0.007	1.12±0.004	84.14±2.80

6 Conclusion and Discussion

In this paper, we present UA-CP, a unified uncertainty-aware conformal prediction framework for reliable endoscopic lesion detection. By integrating detector-adaptive prior-aware training with dual conformal calibration, UA-CP generally improves detection utility while providing calibrated category and localization uncertainty across diverse detector architectures and endoscopic benchmarks. For future research, incorporating calibrated detection uncertainty into downstream decision models could bridge prediction-level reliability with clinical actions and operational outcomes in endoscopic lesion detection.

AI USE STATEMENT

We used generative AI tools to improve the clarity and fluency of the writing and to assist with code implementation and debugging. The authors reviewed and revised all AI-assisted text and manually inspected, tested, and verified all AI-assisted code. We did not use these tools to generate research ideas, scientific claims, experimental results, data, or citations. We take full responsibility for the final manuscript, code, and accompanying artifacts.

REPRODUCIBILITY STATEMENT

To facilitate reproduction, we provide an anonymous repository containing the code for our framework. The repository has been checked to ensure that it contains no identifying author information.

REFERENCES

Damola Agbelese, Krishna Chaitanya, Pushpak Pati, Chaitanya Parmar, Pooya Mobadersany, Shreyas Fadnavis, Lindsey Surace, Shadi Yarandi, Louis R Ghanem, Molly Lucas, et al. Megan: Mixture of experts for robust uncertainty estimation in endoscopy videos. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 3–13. Springer, 2025.

Omer F Ahmad, Juana González-Bueno Puyal, Patrick Brandao, Rawen Kader, Faisal Abbasi, Mohamed Hussein, Rehan J Haidry, Daniel Toth, Peter Mountney, Ed Seward, et al. Performance of artificial intelligence for detection of subtle and advanced colorectal neoplasia. *Digestive Endoscopy*, 34(4):862–869, 2022.

Léo Andéol, Thomas Fel, Florence de Grancey, and Luca Mossina. Conformal prediction for trustworthy detection of railway signals. *AI and Ethics*, 4(1):157–161, 2024.

Léo Andéol, Luca Mossina, Adrien Mazoyer, and Sébastien Gerchinovitz. Conformal Object Detection by Sequential Risk Control, 2025. working paper or preprint.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, 2015.

Hanna Borgli, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang-Nguyen, Dag Johansen, Carsten Griwodz, Håkon K. Stensland, Enrique Garcia-Ceja, Peter T. Schmidt, Hugo L. Hammer, Michael A. Riegler, Pål Halvorsen, and Thomas de Lange. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Giovanni Cherubin. How do the performance of a conformal predictor and its underlying algorithm relate? In *Conformal and Probabilistic Prediction with Applications*, pp. 546–548. PMLR, 2023.

Vicky Copley, Greg Finlay, and Ben Hiett. The uncertain object: Application of conformal prediction to aerial and satellite images. In *The 13th Symposium on Conformal and Probabilistic Prediction with Applications*, pp. 73–89, 2024.

Jinzong Dong, Zhaohui Jiang, Dong Pan, Zhiwen Chen, Qingyi Guan, Hongbin Zhang, Gui Gui, and Weihua Gui. A survey on confidence calibration of deep learning-based classification models under class imbalance data. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9): 15664–15684, 2025.

Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *Int. Conf. Comput. Vis.*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 2016.

Erik A Holzwanger, Mohammad Bilal, Jeremy R Glissen Brown, Shailendra Singh, Aymeric Becq, Kenneth Ernest-Suarez, and Tyler M Berzin. Benchmarking definitions of false-positive alerts during computer-aided polyp detection in colonoscopy. *Endoscopy*, 53(09):937–940, 2021.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*, 2023.

Denis Huseljic, Marek Herde, Paul Hahn, Mehmet Muejde, and Bernhard Sick. Systematic evaluation of uncertainty calibration in pretrained object detectors. *International Journal of Computer Vision*, 133(3):1033–1047, 2025.

Yuncheng Jiang, Zixun Zhang, Yiwen Hu, Guanbin Li, Xiang Wan, Song Wu, Shuguang Cui, Silin Huang, and Zhen Li. Ecc-polypdet: Enhanced centernet with contrastive learning for automatic polyp detection. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4785–4796, 2023.

Glenn Jocher, Alex Stoken, Jirka Borovec, Liu Changyu, Adam Hogan, Laurentiu Diaconu, Jake Poznanski, Lijun Yu, Prashant Rai, Russ Ferriday, et al. ultralytics/yolov5: v2. 0. *Zenodo*, 2020.

Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 48–56. Springer, 2019.

Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof Von Kalle, and Titus Josef Brinker. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8): e36427, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

Peixi Liu, Pu Wang, Jeremy R Glissen Brown, Tyler M Berzin, Guanyu Zhou, Weihui Liu, Xun Xiao, Ziyang Chen, Zhihong Zhang, Chao Zhou, et al. The single-monitor trial: an embedded cade system increased adenoma detection during colonoscopy: a prospective randomized study. *Therapeutic advances in gastroenterology*, 13:1756284820979165, 2020.

Helena Liz-López, Áurea Anguera de Sojo-Hernández, Sergio D'Antonio-Maceiras, Miguel Angel Díaz-Martínez, and David Camacho. Deep learning innovations in the detection of lung cancer: Advances, trends, and open challenges. *Cognitive Computation*, 17(2):67, 2025.

L López, Shaza Elsharief, Dhiyaa Al Jorf, Firas Darwish, Congbo Ma, and Farah E Shamout. Uncertainty quantification for machine learning in healthcare: a survey. *arXiv preprint arXiv:2505.02874*, 2025.

Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *International conference on medical image computing and computer-assisted intervention*, pp. 387–396, 2021.

F Marques and C Paulo. Universal distribution of the empirical coverage in split conformal prediction. *Statistics & Probability Letters*, 219(C), 2025.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *arXiv preprint arXiv:2307.09302*, 2023.

Sanbao Su, Songyang Han, Yiming Li, Zhili Zhang, Chen Feng, Caiwen Ding, and Fei Miao. Collaborative multi-object tracking with conformal uncertainty propagation. *IEEE Robotics and Automation Letters*, 9(4):3323–3330, 2024.

Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

Qiaozhi Tan, Long Bai, Guankun Wang, Mobarakol Islam, and Hongliang Ren. Endoood: Uncertainty-aware out-of-distribution detection in capsule endoscopy diagnosis. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, pp. 9627–9636, 2019.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. In *European Conference on Computer Vision*, pp. 363–398. Springer, 2024.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74 (1):9–28, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Cham, 2022. doi: 10.1007/978-3-031-06649-8.

Lingyun Wu, Zhiqiang Hu, Yuanfeng Ji, Ping Luo, and Shaoting Zhang. Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 302–312. Springer, 2021.

Zheng Wu, Kehua Guo, Entao Luo, Tian Wang, Shoujin Wang, Yi Yang, Xiangyuan Zhu, and Rui Ding. Medical long-tailed learning for imbalanced data: Bibliometric analysis. *Computer Methods and Programs in Biomedicine*, 247:108106, 2024.

Chao Xu, Zhiwei Fan, Yaoyao Ma, Yuling Huang, Jing Wang, Yishen Xu, and Di Wu. Ce-yolo: A channel-efficient yolo algorithm for colorectal polyp detection. *Biomedical Signal Processing and Control*, 104:107529, 2025.

Masayoshi Yamada, Yutaka Saito, Shigemi Yamada, Hiroko Kondo, and Ryuji Hamamoto. Detection of flat colorectal neoplasia by artificial intelligence: a systematic review. *Best Practice & Research Clinical Gastroenterology*, 52:101745, 2021.

Burhaneddin Yaman, Tanvir Mahmud, and Chun-Hao Liu. Instance-aware repeat factor sampling for long-tailed object detection. In *NeurIPS*, 2023.

Heng Yang and Marco Pavone. Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8947–8958, 2023.

Kôsaku Yosida. *Functional analysis*. Springer Science & Business Media, 2012.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. arXiv preprint arXiv:2203.03605.

Xuanye Zhang, Kaige Yin, Siqi Liu, Zhijie Feng, Xiaoguang Han, Guanbin Li, and Xiang Wan. Self-and semi-supervised learning for gastroscopic lesion detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 83–93, 2023.

Xuanye Zhang, Xiaoqing Hu, Guanbin Li, Si-Qi Liu, Xiang Wan, and Yuanhuan Xiong. Pattern-anchored adaptive prototype learning for gastroscopic lesion detection and beyond. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 436–445. Springer, 2025.

Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 1(1):100003, 2023.

A THEORETICAL PRELIMINARIES

A.1 MARGINAL COVERAGE

Marginal coverage provides the fundamental reliability guarantee for distribution-free uncertainty estimation. Within our Uncertainty-Aware Conformal Prediction (*UA-CP*) framework, instead of producing deterministic point estimates, we employ Regularized Adaptive Prediction Sets (*RAPS*) Angelopoulos et al. (2020) to generate a prediction set, $\hat{\mathbf{L}}$, that rigorously quantifies detection uncertainty. To further address localization uncertainty, we construct relaxed prediction regions using *Conformalized Quantile Regression* (*CQR*) Romano et al. (2019), which provides strict finite-sample risk guarantees. By integrating these complementary conformal guarantees, the proposed *UA-CP* framework offers a statistically robust and clinically interpretable solution for endoscopic lesion detection. Formal proofs establishing marginal coverage for both detection and localization are presented below.

Proposition A.1 (*RAPS* Coverage Guarantee)**.** *Suppose the matched calibration detections and a matched test detection are exchangeable. Let $\widehat{\mathcal{C}}_L(X_{N+1})$ be the RAPS prediction set defined in Eq. (13), with conformal threshold $\widehat{q}_L$ defined in Eq. (12). Then, for target miscoverage level $\alpha_L \in (0,1)$,*

$$\mathbb{P}\left(l_{N+1} \in \widehat{\mathcal{C}}_L(X_{N+1})\right) \geq 1 - \alpha_L. \tag{24}$$

Proof. For each matched calibration detection i, define the RAPS nonconformity score

$$S_i^L = S_i^L(l_i; u_i) = \rho_i(l_i) + \widehat{\pi}_i(l_i)u_i + \lambda_R\big(o_i(l_i) - k_{\text{reg}}\big)^+.$$

Let

$$\widehat{q}_L = \text{Quantile}_{\frac{\lceil (N_D+1)(1-\alpha_L)\rceil}{N_D}}\left(\{S_i^L\}_{i=1}^{N_D}\right)$$

be the conformal threshold computed from the matched calibration detections.

For a matched test detection, define

$$S_{N_D+1}^L = S_{N_D+1}^L(l_{N_D+1}; u_{N_D+1}).$$

By construction of the RAPS prediction set in Eq. (13),

$$l_{N_D+1} \in \widehat{\mathcal{C}}_L(X_{N_D+1}) \quad \Longleftrightarrow \quad S_{N_D+1}^L \leq \widehat{q}_L.$$

Under exchangeability of the matched calibration and test detections, the augmented scores

$$S_1^L, \ldots, S_{N_D}^L, S_{N_D+1}^L$$

are exchangeable. Therefore, by the standard split-conformal rank argument,

$$\mathbb{P}\big(S_{N_D+1}^L > \widehat{q}_L\big) \leq \alpha_L.$$

Consequently,

$$\mathbb{P}\Big(l_{N_D+1} \in \widehat{\mathcal{C}}_L(X_{N_D+1})\Big) = \mathbb{P}\big(S_{N_D+1}^L \leq \widehat{q}_L\big) \geq 1 - \alpha_L.$$

□

Theorem A.2. *Let $(X_i, Y_i)_{i=1}^{N+1}$ be exchangeable, where the first N samples form the calibration set. Define the CQR nonconformity scores*

$$e_i = \max\left\{\widehat{q}_{\alpha/2}(X_i) - Y_i,\, Y_i - \widehat{q}_{1-\alpha/2}(X_i)\right\}, \qquad i = 1, \ldots, N.$$

Let

$$k = \lceil (N+1)(1-\alpha) \rceil$$

and let $\widehat{q} = e_{(k)}$ denote the k-th order statistic of the calibration scores, with $\widehat{q} = +\infty$ if $k > N$. Then the CQR interval

$$\widehat{\mathcal{C}}(X_{N+1}) = \left[\widehat{q}_{\alpha/2}(X_{N+1}) - \widehat{q},\, \widehat{q}_{1-\alpha/2}(X_{N+1}) + \widehat{q}\right]$$

satisfies

$$\mathbb{P}\Big(Y_{N+1} \in \widehat{\mathcal{C}}(X_{N+1})\Big) \geq 1 - \alpha.$$

If the augmented scores $e_1, \ldots, e_{N+1}$ are almost surely distinct, then

$$\mathbb{P}\Big(Y_{N+1} \in \widehat{\mathcal{C}}(X_{N+1})\Big) = \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1},$$

and hence

$$1 - \alpha \leq \mathbb{P}\Big(Y_{N+1} \in \widehat{\mathcal{C}}(X_{N+1})\Big) < 1 - \alpha + \frac{1}{N+1}.$$

Proof. Define the test nonconformity score

$$e_{N+1} = \max\left\{\widehat{q}_{\alpha/2}(X_{N+1}) - Y_{N+1},\, Y_{N+1} - \widehat{q}_{1-\alpha/2}(X_{N+1})\right\}.$$

By construction,

$$Y_{N+1} \in \widehat{\mathcal{C}}(X_{N+1}) \quad \Longleftrightarrow \quad e_{N+1} \leq \widehat{q}.$$

Because $(X_i, Y_i)_{i=1}^{N+1}$ are exchangeable and the fitted quantile functions are fixed independently of the calibration and test samples, the augmented scores

$$e_1, \ldots, e_N, e_{N+1}$$

are exchangeable.

Let R_{N+1} denote the rank of e_{N+1} among these $N+1$ scores, with ties handled conservatively. By exchangeability,

$$\mathbb{P}(R_{N+1} \leq k) \geq \frac{k}{N+1}.$$

Moreover,

$$R_{N+1} \leq k \quad \Longrightarrow \quad e_{N+1} \leq e_{(k)} = \widehat{q}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Big(Y_{N+1} \in \widehat{\mathcal{C}}(X_{N+1})\Big) &= \mathbb{P}(e_{N+1} \leq \widehat{q}) \\
&\geq \frac{k}{N+1} \\
&= \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \\
&\geq 1 - \alpha.
\end{aligned}$$

If the augmented scores are almost surely distinct, the rank R_{N+1} is uniformly distributed over $\{1, \ldots, N+1\}$. Consequently,

$$\mathbb{P}(e_{N+1} \leq \widehat{q}) = \frac{k}{N+1} = \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1}.$$

Since $\lceil x \rceil < x + 1$,

$$1 - \alpha \leq \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} < 1 - \alpha + \frac{1}{N+1},$$

□

Proposition A.3 (*Simultaneous Localization Coverage*)**.** *Fix a trained detector and matching rule. Suppose that, for each coordinate $d \in \{1, \ldots, 4\}$, the matched calibration and validation residuals are exchangeable. If each coordinate is calibrated at miscoverage level $\alpha_B/4$, then the conformal box region satisfies*

$$P\left(\mathbf{b}_{N+1} \in \widehat{\mathcal{C}}_{B,j}(X_{N+1})\right) \geq 1 - \alpha_B. \tag{25}$$

Proof. For coordinate d, the split-conformal quantile

$$\widehat{q}_{B,d} = \text{Quantile}_{\frac{\lceil (N_D+1)(1-\alpha_B/4) \rceil}{N_D}} \left(\{ S^B_{r,d} \}_{r=1}^{N_D} \right) \tag{26}$$

provides the coordinate-wise coverage guarantee

$$P\Big(b_{N+1,d} \notin \Big[\widehat{L}_{d,j}, \widehat{U}_{d,j} \Big] \Big) \leq \frac{\alpha_B}{4}. \tag{27}$$

Let

$$E_d = \Big\{ b_{N+1,d} \in \Big[\widehat{L}_{d,j}, \widehat{U}_{d,j} \Big] \Big\}$$

denote the coverage event for coordinate d. Failure of simultaneous box coverage occurs whenever at least one coordinate is not covered. Therefore, by the union bound,

$$P\Big(\mathbf{b}_{N+1} \notin \widehat{\mathcal{C}}_{B,j}(X_{N+1}) \Big) = P\left(\bigcup_{d=1}^{4} E_d^c \right) \tag{28}$$

$$\leq \sum_{d=1}^{4} P(E_d^c) \tag{29}$$

$$\leq \sum_{d=1}^{4} \frac{\alpha_B}{4} = \alpha_B. \tag{30}$$

Taking complements yields

$$P\left(\mathbf{b}_{N+1} \in \prod_{d=1}^{4} \Big[\widehat{L}_{d,j}, \widehat{U}_{d,j} \Big] \right) \geq 1 - \alpha_B. \tag{31}$$

□

Proposition A.4 (joint coverage guarantee)**.** *Suppose the conformal label set and conformal box region satisfy*

$$P\Big(l_{N+1} \notin \widehat{\mathcal{C}}_{L,j}(X_{N+1}) \Big) \leq \alpha_L, \tag{32}$$

$$P\Big(\mathbf{b}_{N+1} \notin \widehat{\mathcal{C}}_{B,j}(X_{N+1}) \Big) \leq \alpha_B. \tag{33}$$

Then

$$P\Big(l_{N+1} \in \widehat{\mathcal{C}}_{L,j}(X_{N+1}),\ \mathbf{b}_{N+1} \in \widehat{\mathcal{C}}_{B,j}(X_{N+1}) \Big) \geq 1 - \alpha_L - \alpha_B. \tag{34}$$

Proof. When $K \in \widehat{\mathcal{C}}_L(X)$, the maximum correction over the predicted label set includes the correction associated with the true category K. Define the label-coverage event as

$$A_L = \Big\{ K \in \widehat{\mathcal{C}}_L(X) \Big\},$$

and the simultaneous localization-coverage event for the true category as

$$A_B^K = \Big\{ B_d \in [\widehat{L}_d^K(X), \widehat{U}_d^K(X)],\ \forall d \in \{1, \ldots, 4\} \Big\}.$$

The joint event $A_L \cap A_B^K$ can therefore fail only if either the label-coverage event or the true-category localization-coverage event fails. By the union bound,

$$\begin{aligned} \mathbb{P}\big((A_L \cap A_B^K)^c \big) &= \mathbb{P}\big(A_L^c \cup (A_B^K)^c \big) \\ &\leq \mathbb{P}(A_L^c) + \mathbb{P}((A_B^K)^c) \\ &\leq \alpha_L + \alpha_B. \end{aligned}$$

Hence,

$$\mathbb{P}\big(A_L \cap A_B^K \big) \geq 1 - \alpha_L - \alpha_B.$$

□

A.2 TASK-ADAPTIVE PRIOR-AWARE GROUP REGULARIZATION

To facilitate robust uncertainty-aware classification, we introduce a task-adaptive **Prior-Aware Group Regularization** (*PAT*) framework governed by two objective functions: $\mathcal{L}_{Penalty}$ for two-stage detectors and $Cost_{prior}$ for one-stage detectors. These objectives jointly form a composite operator that penalizes uncertainty while incorporating prior structural knowledge.

To ensure reliable optimization, it is necessary to verify the mathematical properties of this composite objective. In this section, we present formal proofs of **Theorem A.5** and **Theorem A.6**, establishing that the optimization procedure for *PAT* is well-defined, stable, and convergent.

Theorem A.5. *Consider the PAT classifier-head objective under fixed input features. Consider the objective function*

$$\mathcal{L}_{\mathrm{PAT}}(w) = -\frac{1}{N}\sum_{i=1}^{N}\log p_{i,l_i} + \lambda\sum_{k=1}^{K}\alpha_k\left(1-\mathbb{E}_{\{l_i=k\}}\left[p_{i,l_i}\right]\right)\|w_k\|_2 = f(w)+g(w),$$

where $f(w)$ denotes the cross-entropy loss and $g(w)$ represents the task-adaptive group penalty. Then, for any $x \in \mathbb{R}^D$ and $\gamma > 0$, the proximal mapping $\mathrm{prox}_{\gamma\mathcal{L}}(x)$ exists and is unique. The proximal operator is defined as

$$\mathrm{prox}_{\gamma\mathcal{L}}(x) = \arg\min_{w}\left(\mathcal{L}_{\mathrm{PAT}}(w) + \frac{1}{2\gamma}\|w-x\|_2^2\right). \tag{35}$$

Proof. We first show that $f(w)$ is a convex, proper, and lower semi-continuous (LSC) function.

In neural network classifiers, p_{i,l_i} is typically the output of a softmax layer:

$$p_{i,l_i} = \frac{\exp(w_{l_i}^\top x_i)}{\sum_j \exp(w_j^\top x_i)}. \tag{36}$$

The negative log-softmax $-\log p_{i,l_i}$ is convex with respect to the weights w. Since $f(w)$ is an average of convex functions, it follows that $f(w)$ is convex. Moreover, $f(w)$ is finite for all $w \in \mathbb{R}^D$ and continuous on its domain. Hence, $f(w)$ is proper and LSC.

Next, consider the weighted regularization term

$$g(w) = \lambda\sum_{k=1}^{K}\alpha_k\left(1-\mathbb{E}_{\{l_i=k\}}\left[p_{i,l_i}\right]\right)\|w_k\|_2. \tag{37}$$

Here $\lambda > 0$ and $\|w_k\|_2$ denotes the ℓ_2 norm. The ℓ_2 norm is convex and continuous. The coefficient $\alpha_k(1-\mathbb{E}[p_{i,l_i}])$ is strictly positive because $\alpha_k = \frac{1}{\sqrt{|n_k|}} > 0$, where n_k denotes the number of samples in class k, and $p_{i,l_i} \in (0,1)$. Consequently, $g(w)$ is a nonnegative weighted sum of convex functions, and is therefore convex, finite, and continuous.

To establish the existence of the proximal mapping, we invoke the *Moreau–Yosida theorem* Moreau (1965); Yosida (2012), which states that a unique proximal solution exists if the objective function is proper, LSC, and convex.

Since

$$\mathcal{L}_{\mathrm{PAT}}(w) = f(w) + g(w)$$

is the sum of two proper, LSC, and convex functions, it follows that $\mathcal{L}_{Penalty}(w)$ is also proper, LSC, and convex. Therefore, for any $\gamma > 0$, the proximal operator

$$\mathrm{prox}_{\gamma\mathcal{L}}(x) = \arg\min_{w}\left(\mathcal{L}_{\mathrm{PAT}}(w) + \frac{1}{2\gamma}\|w-x\|_2^2\right) \tag{38}$$

is well defined, and the corresponding proximal optimization problem admits a unique minimizer. □

Theorem A.6. *Consider the prior-aware cost function*

$$Cost_{\mathrm{PAT}}(p, e) = -\gamma \sum_{k=1}^{K} p_{i,l_i} \log(p_{i,l_i} + \varepsilon) + \lambda \alpha_k e_{i,k} = f(p) + g(e), \tag{39}$$

where $f(p)$ represents entropy-regularized probability estimates and $g(e)$ denotes the linear error penalty. Then, for any $x \in \mathbb{R}$ and $\gamma > 0$, the proximal mapping $\mathrm{prox}_{\gamma\, Cost_{\mathrm{PAT}}}(x)$ exists and is unique. The proximal operator is defined as

$$\mathrm{prox}_{\gamma\, Cost_{\mathrm{PAT}}}(x) = \arg\min_{u} \left(Cost_{\mathrm{PAT}}(u) + \frac{1}{2\gamma} \|u - x\|_2^2 \right). \tag{40}$$

Proof. We prove that the prior-aware cost function admits a unique proximal solution by verifying that it is convex, proper, and lower semi-continuous (LSC).

First, we establish convexity with respect to the optimization variables. Consider the negative entropy bias term

$$f(p_{i,l_i}) = p_{i,l_i} \log(p_{i,l_i} + \varepsilon),$$

where $p_{i,l_i} \geq 0$ and $\varepsilon > 0$ is a small constant introduced for numerical stability. The second derivative of f with respect to p_{i,l_i} is

$$f''(p_{i,l_i}) = \frac{1}{p_{i,l_i} + \varepsilon} + \frac{\varepsilon}{(p_{i,l_i} + \varepsilon)^2} > 0. \tag{41}$$

Hence, $f(p_{i,l_i})$ is strictly convex. Consequently, the summation

$$\sum_{k=1}^{K} p_{i,l_i} \log(p_{i,l_i} + \varepsilon)$$

is convex. The scaling factor $-\gamma$ only changes the orientation of the function but does not affect convexity in the probability space.

Next, consider the linear penalty term

$$g(e) = \lambda \alpha_k e_{i,k},$$

where $\alpha_k = \frac{1}{\sqrt{|n_k|}}$, $e_{i,k} = 1 - p_{i,l_i}$, and n_k denotes the number of samples in class k. Since $g(e)$ is linear with respect to e, it is convex.

Therefore, the prior-aware cost function

$$Cost_{\mathrm{PAT}}(p, e) = f(p) + g(e)$$

is the sum of convex functions and is thus convex.

Next, we verify that the function is proper. Since $p \in [0, 1]$, K is finite, and $\varepsilon > 0$ prevents $\log(p_{i,l_i} + \varepsilon)$ from diverging to $-\infty$, both $f(p)$ and $g(e)$ remain bounded on their domains. Moreover, $Cost_{prior}(p, e)$ is finite for at least one point in its domain. Therefore,

$$-\infty < Cost_{\mathrm{PAT}}(p, e) < +\infty$$

for all feasible points, implying that $Cost_{prior}$ is a proper function.

Finally, we show that $Cost_{\mathrm{PAT}}(p, e)$ is lower semi-continuous. The function $f(p)$ consists of continuous terms of the form $p_{i,l_i} \log(p_{i,l_i} + \varepsilon)$, and $g(e)$ is linear. Since both components are continuous, $Cost_{\mathrm{PAT}}(p, e)$ is continuous on its domain. Because every continuous function is LSC, $Cost_{\mathrm{PAT}}(p, e)$ is LSC and closed on $\mathbb{R}^K$.

According to the *Moreau-Yosida theorem* Moreau (1965); Yosida (2012), if a function is convex, proper, and LSC, then its proximal operator is well defined. Thus, the proximal subproblem

$$\mathrm{prox}_{\gamma Cost_{\mathrm{PAT}}}(x) = \arg\min_{u} \left(Cost_{\mathrm{PAT}}(u) + \frac{1}{2\gamma} \|u - x\|_2^2 \right) \tag{42}$$

is strongly convex and therefore admits a unique minimizer for any $\gamma > 0$. This guarantees that the optimization step for the proposed prior-aware cost is mathematically well defined, stable, and convergent. □

Algorithm 1 *RAPS* Conformal Calibration

Input: label-miscoverage level α_L; calibration preidictions $\{(\mathbf{p}, \hat{l})\}$, where $\mathbf{p} \in \Delta^{K-1}$ and $\hat{l} \in \mathcal{K}$; k_{reg}; λ; boolean *rand*.
procedure *RAPS*$(\alpha_L, s, I, l, \lambda_R)$
for $i \in \{1, \ldots, N_{\text{cal}}\}$ **do**
  Sort $\mathbf{p}_i$ in descending order:

$$s_{i,1} \geq \cdots \geq s_{i,K}, \qquad s_{i,m} = p_{i,I_{i,m}},$$

  where $(I_{i,1}, \ldots, I_{i,K})$ is the corresponding permutation of $\mathcal{K}$.
  $L_i \leftarrow j$ such that $I_{i,j} = l_i$
  if rand **then**
    $u_i \sim \text{Unif}(0, 1)$
  end if
  Compute the RAPS nonconformity score:

$$S_i^L \leftarrow \sum_{m=1}^{L_i - 1} s_{i,m} + u_i s_{i,L_r} + \lambda_R \big(L_i - k_{\text{reg}}\big)^+$$

end for
$k_L \leftarrow \lceil (1 - \alpha_L)(1 + n) \rceil$
if $k_L \leq N_{\text{cal}}$ **then**
  $\widehat{q}_L \leftarrow k_L$-th smallest value in $\{S_r^L\}_{r=1}^{N_{\text{cal}}}$.
else
  $\widehat{q}_L \leftarrow +\infty$.
end if

return $\widehat{q}_L$.
Output: The generalized quantile, $\hat{q}_L$ ▷The value in Eq.(12)

B Implementation Details

B.1 Matching Protocol

For each image, let $\mathcal{G}(X) = \{(l_i, \mathbf{b}_i)\}_{i=1}^{N_G}$ denote the ground-truth lesions and let

$$\widehat{\mathcal{D}}(x) = \{\hat{d}_j\}_{j=1}^{N_D(X)}$$

denote the intermediate detector outputs. We define the pairwise IoU matrix

$$A_{ij} = \text{IoU}(\hat{\mathbf{b}}_i, \hat{\mathbf{b}}_j) = \frac{\text{Area}(\hat{\mathbf{b}}_i \cap \hat{\mathbf{b}}_j)}{\text{Area}(\hat{\mathbf{b}}_i \cup \hat{\mathbf{b}}_j)}. \tag{43}$$

We obtain a class-agnostic one-to-one assignment by maximizing the total IoU. Let $\text{Match}(i)$ denote the detection assigned to ground-truth lesion i. We retain only assigned pairs whose IoU exceeds the matching threshold:

$$\mathcal{M}(X) = \{(i, j) : j = \text{Match}(i),\ A_{ij} \geq \tau_{\text{match}}\} \tag{44}$$

B.2 RAPS Category Calibration

RAPS quantifies category uncertainty for each matched detection.

Calibration. Algorithm 1 computes regularized nonconformity scores from ranked category probabilities and calibrates the threshold $\hat{q}_L$ at label-miscoverage level α_L, corresponding to Eq. (13).

Prediction. Algorithm 2 constructs the prediction set $\hat{\mathcal{C}}_L(X)$ by including ranked categories whose RAPS scores do not exceed the calibrated threshold q_L.

Algorithm 2 *RAPS* Prediction Sets

Input: detector f_θ with lower- and upper-quantile box heads; matching threshold τ_{match}; box-miscoverage level α_B.
procedure *RAPS*$(\alpha_L, s, I, \hat{q}_L, k_{\text{reg}}, \lambda_R, \textit{rand})$
$\widehat{\mathcal{C}}_L(X) \leftarrow |\{j \in \mathcal{K} : \sum_{i=1}^{j} s_i + \lambda_R(j - k_{\text{reg}})^+ \le \hat{q}_L\}| + 1$
if rand **then**
  $U \sim \text{Unif}(0,1)$
  $\widehat{\mathcal{C}}_L(X) \leftarrow \widehat{\mathcal{C}}_L(X) - \mathbb{I}\left\{\frac{\sum_{i=1}^{L} s_i + \lambda(L-k_{\text{reg}})^+ - \hat{q}_d}{s_L + \lambda\mathbb{I}(L>k_{\text{reg}})} \le U\right\}$
end if
return $\widehat{\mathcal{C}}_L(X) = \{1, \ldots, I_L\}$ ▷ The $\widehat{\mathcal{C}}_L(X$ most likely categories
Output: The $1-\alpha_L$ confidence set, $\widehat{\mathcal{C}}_L$ ▷ The set in Eq. (13)

Table 4: **Summary of the evaluated endoscopic datasets.** Dataset characteristics and lesion settings.

Dataset	Evaluated subset	Lesion scenario and challenge
LGLDD	Upper-GI gastroscopy; 10,083 images, 12,292 boxes, and 4 categories	Long-tailed category frequencies; small or low-contrast lesions; multiple lesions; and irregular or diffuse boundaries
Kvasir-Capsule	Capsule endoscopy; 4,224 spatially annotated images and nine categories	Severe category imbalance; small findings on textured mucosa; semantic ambiguity; and uncertain boundaries
LDPolypVideo	Lower-GI video colonoscopy; 160 clips and 33,884 polyp-positive frames	Small polyps, variable appearance and viewpoint, and uncertain lesion extent

B.3 Box-CQR localization calibration.

To provide rigorous coverage guarantees for object localization, we adopt *Conformal Quantile Regression(CQR)*. As detailed in Algorithm 3, CQR calibrates coordinate-wise corrections $\hat{q}_{B,d}$ to construct simultaneous conformal box regions.

C Dataset Statistics

C.1 Dataset Introduction

We evaluate UA-CP on three complementary endoscopic lesion-detection datasets, each representing a distinct combination of category imbalance, semantic ambiguity, lesion scale, and boundary uncertainty.

The four LGLDD categories are polyp (*pol*), ulcer (*uls*), cancer (*can*), and submucosal tumor (*stm*). The nine-category Kvasir-Capsule subset comprises Angiectasia (*Ang*), Blood-fresh (*bf*), Blood-hematin (*bh*), Erosion (*Ero*), Erythema (*Ery*), Foreign-body (*Fb*), Lymphangiectasia (*Lym*), Polyp (*Pol*), and Ulcer (*Ulc*). LDPolypVideo is a single-category polyp-detection benchmark; consequently, it primarily evaluates detection and localization rather than category uncertainty. All experiments operate on individual images or extracted frames, without temporal aggregation or tracking.

C.2 Data Partitioning

We adopt a common 3:1:1 train–calibration–test allocation:

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}, .$$

Each partition has a distinct role. $\mathcal{D}_{\text{train}}$ is used exclusively for detector optimization and uncertainty-aware learning. After fixing model parameters, $\mathcal{D}_{\text{cal}}$ is used to estimate the conformal corrections for RAPS CQR conformal thresholds, while $\mathcal{D}_{\text{test}}$ is reserved evaluating detection utility, uncertainty validity, and uncertainty efficiency.

Algorithm 3 BBox Uncertainty and Conformal Calibration

Input: Detector f_θ with lower- and upper- quantile box heads trained on $\mathcal{D}_{\text{train}}$; Calibration dataset $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=N}^{N+|\mathcal{D}_{cal}|}$; test image X_{N+1}; box-miscoverage level α_B.

Ensure: Coordinate-wise conformal corrections $\{\widehat{q}_{B,d}\}_{d=1}^{4}$ and conformal box regions $\{\widehat{\mathcal{C}}_{B,|}(X_{N+1})\}_{j=1}^{N_D(X_{N+1})}$

Initialize: Quantile levels $\tau_l \leftarrow \alpha_B/2$ (Lower), $\tau_u \leftarrow 1 - \alpha_B/2$ (Upper)

$\mathcal{R}_B \leftarrow \varnothing, \quad N_{\text{cal}} \leftarrow 0.$

for each $(X_{N+1}, Y_{N+1}) \in \mathcal{D}_{\text{cal}}$ **do**

Obtain detector predictions

$$\widehat{D}(X) = \left\{ \left(\widehat{l}_j, \widehat{\mathbf{b}}_j, \widetilde{s}_j, \mathbf{p}_j, \widehat{\mathbf{b}}_j^{lo}, \widehat{\mathbf{b}}_j^{up} \right) \right\}_{j=1}^{N_D(X_{N+1})}.$$

Construct the class-agnostic one-to-one matched set

$$\mathcal{M}(X) = \left\{ (i,j) : \text{IoU}(\mathbf{b}_i, \widehat{\mathbf{b}}_j) \geq \tau_{\text{match}} \right\}.$$

for each $(i,j) \in \mathcal{M}(X)$ **do**

for $d = 1, \ldots, 4$ **do**

$$S_{i,j,d}^{B} \leftarrow \max \left\{ \widehat{b}_{j,d}^{lo} - b_{i,d},\ b_{i,d} - \widehat{b}_{j,d}^{up} \right\}.$$

end for

Compute the joint box nonconformity score:

$$S_{i,d}^{B} \leftarrow \max_{d \in \{1,\ldots,4\}} S_{i,j,d}^{B}.$$

$\mathcal{R}_{B,d} \leftarrow \mathcal{R}_{B,d} \cup \{S_{i,d}^{B}\}_{d=1}^{4}.$

$N_{\text{cal}} \leftarrow N_{\text{cal}} + 1.$

end for

end for

$k_B \leftarrow \lceil (N_{\text{cal}} + 1)(1 - \alpha_B) \rceil.$

for $d = 1, \ldots, 4$ **do**

if $k_B \leq N_{\text{cal}}$ **then**

$\widehat{q}_B \leftarrow k_B$-th smallest value in $\mathcal{R}_B$.

else

$\widehat{q}_B \leftarrow +\infty.$

end if

end for

Prediction

for $j = 1, \ldots, N_D(X_{N+1})$ **do**

Construct the conformal box region

$$\widehat{\mathcal{C}}_{B,j}(X_{N+1}) \leftarrow \prod_{d=1}^{4} \left[\widehat{b}_{j,d}^{lo} - \widehat{q}_{B,d},\ \widehat{b}_{j,d}^{up} + \widehat{q}_{B,d} \right].$$

end for

Output: Conformal Box Prediction set $\{\widehat{\mathcal{C}}_{B,j}(X_{N+1})\}_{j=1}^{N_D(X_{N+1})}$.

All experiments are conducted at the image/frame level without temporal modeling. Conformal guarantees assume exchangeability of matched calibration and validation detections and are restricted to this evaluation setting.

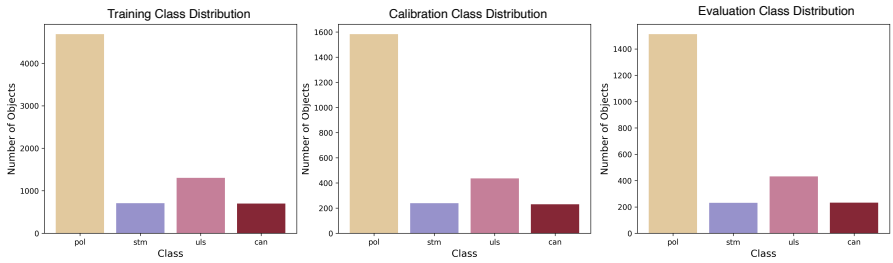

Figure 4: **Class distribution of the LGLDD dataset.** The *pol* category heavily dominates as the head class, while other clinically critical lesions (*stm, uls, and can*) are significantly underrepresented.

C.3 Category Imbalance and Prior Stability

Medical detection datasets can exhibit long-tailed category distributions, with substantial variation in sample frequency across lesion categories. LGLDD provides a representative example of such category imbalance across its training, calibration, and evaluation partitions. As illustrated in Figure 4, the *pol* category accounts for the majority of instances, while other clinically important categories (*stm*, *uls*, and *can*) are substantially underrepresented. This imbalance motivates incorporating training-set category frequencies into PAT and evaluating whether aggregate detection performance conceals failures in less frequent categories.

C.4 Effect of Prior-Aware Training

We assessed the stability of the category priors in *PAT* by analyzing class distributions across the LGLDD training, calibration, and validation sets. A fundamental assumption for effective group regularization is that the empirical category priors estimated from the training set remain consistent and representative of the clinical distribution encountered during inference.

As illustrated in Figure 5, the empirical category rates exhibit high stability as the sample size increases. Specifically, as the cumulative number of instances grows, the fluctuation in category probability diminishes, reaching a steady state once the sample size exceeds approximately *500* instances. This observation indicates that the estimated priors are not artifacts of small-sample bias but instead provide a reliable approximation of the underlying dataset distribution, thereby supporting the robustness of *PAT*.

D Additional Quantitative Experiments

D.1 Complete Detection Utility Results

This section reports complete detection-utility and model-complexity results across all three benchmarks. Because DCC leaves point detections unchanged, differences in detection utility and model complexity reflect PAT.

Detection utility across benchmarks. We report aggregate detection-utility results for LGLDD, Kvasir-Capsule, and LDPolypVideo in Table 5, 6, and 8, respectively. Category-level results for Kvasir-Capsule are provided in Table 7.

PAT generally improves detection utility across datasets and architectures.The largest AP gains are obtained by Faster R-CNN–ResNeXt-101 on LGLDD ($37.4 \rightarrow 42.8$) and Faster R-CNN–R50 on Kvasir-Capsule ($38.7 \rightarrow 44.4$). On LDPolypVideo, AP50 and recall improve across all four architectures, with Faster R-CNN-ResNeXt-101 reaching the highest AP50 of 69.1 and DINO–R50 the highest recall of 65.9. The magnitude of improvement varies across categories and operating metrics, reflecting differences across detector architectures.

Computational Complexity.

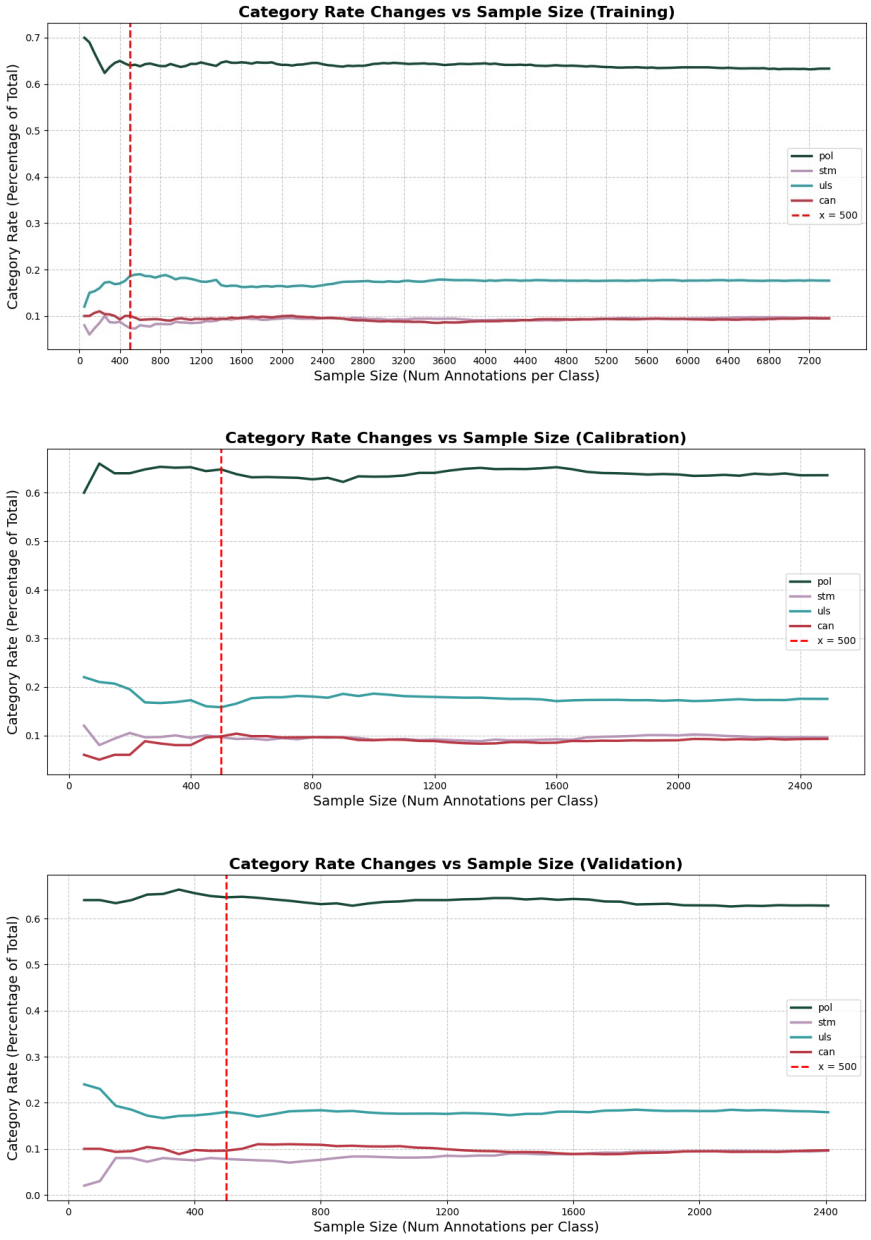

Figure 5: **Statistical Stability of Category Priors.** The convergence of empirical category probabilities across the **training (top)**, **calibration (middle)**, and **validation (bottom)** subsets of the LGLDD dataset.

To assess whether the gains of UA-CP require substantial additional computational capacity, we compare the computational complexity of detectors across LGLDD (Table 9), Kvasir-Capsule (Table 10), and LDPolypVideo (Table 11). We report the number of trainable parameters (*Params*) and giga floating-point operations (*GFLOPs*) at the dataset-specific input resolution (*Resolusion*).

As shown in Table 9- 11, introduces negligible additional model capacity. Parameter counts remain unchanged for most detector-dataset pairs, with only marginal increases in a few cases (e.g., 57.37M to 57.38M for Faster R-CNN-R50 on LDPolypVideo and 47.44M to 47.55M for DINO on LGLDD). On Kvasir-Capsule and LDPolypVideo, baseline and UA-CP variants also have identical GFLOPs at

Table 5: **Detection performance on LGLDD:** We report AP (%), AP50 (%), AP75 (%), and per-category AP (%) for polyp (pol), submucosal tumor (stm), ulcer (uls), and cancer (can). Values in parentheses denote AP improvements over the corresponding detector baseline.

Method	AP	AP50	AP75	pol	stm	uls	can
CenterNet	29.3	57.2	25.4	41.6	36.0	27.3	12.1
YOLOv3	30.2	70.2	19.7	39.5	39.5	22.2	19.5
YOLOv5	38.2	75.4	<u>34.5</u>	49.3	52.4	28.6	22.7
Faster RCNN-R50	32.7	70.2	26.5	39.0	42.2	22.3	27.2
Faster RCNN-Rxt101	37.4	74.4	33.6	43.0	48.8	26.5	28.9
DINO-R50	38.6	<u>78.5</u>	32.9	44.7	51.1	28.9	29.8
FCOS	29.5	66.8	21.1	39.3	37.9	19.0	21.6
PAAPLZhang et al. (2025)	37.8	75.7	33.6	45.1	50.2	26.6	29.4
Faster RCNN-R50+UACP(Ours)	36.1(+3.4)	73.9	30.3	42.9	48.7	25.7	27.2
Faster RCNN-Rxt101+UACP(Ours)	**42.8**(+5.4)	76.9	**34.8**	45.0	51.9	29.6	31.1
DINO+UACP(Ours)	<u>39.9(+1.3)</u>	**80.7**	33.7	45.7	51.4	30.4	30.8
FCOS+UACP(Ours)	31.6(+2.1)	69.5	23.9	41.4	42.9	18.4	23.8

Table 6: **Detection utility on Kvasir-Capsule:** We report AP(%), AP50(%), worst-category AP(%), $\mathrm{AP}_{\mathrm{Worst}}$, Recall(%), F1(%), and FP/image. $\mathrm{AP}_{\mathrm{Worst}}$ denotes the minimum category-wise AP for each detector.

Method	AP↑	AP50	$\mathrm{AP}_{\mathrm{Worst}}$	Recall↑	F1↑	FP/image↓
CenterNet	57.9	93.6	55.2	97.0	25.2	5.41
Faster RCNN-R50	38.7	80.1	19.0	89.8	27.8	4.57
Faster RCNN-Rxt 101	57.5	91.5	34.1	94.5	61.3	1.14
DINO-R50	<u>60.9</u>	92.8	37.7	94.1	86.2	0.24
FCOS	58.5	<u>93.9</u>	42.8	95.5	**90.0**	<u>0.17</u>
YOLOv5	55.8	93.3	45.1	**96.2**	57.6	1.38
Faster RCNN-R50+UACP	44.4(+6.7)	82.2	22.8	89.3	35.6	3.13
Faster RCNN-Rxt 101+UACP	60.1(+2.6)	92.7	35.3	92.9	88.9	**0.16**
DINO-R50+UACP	**63.5**(+2.6)	93.6	40.3	<u>95.7</u>	<u>89.0</u>	0.19
FCOS+UACP	58.6(+0.1)	**95.5**	43.9	95.2	86.8	0.24

the same input resolution. These results indicate that UA-CP provides uncertainty characterization without relying on increased model capacity or substantial additional inference computation.

Table 7: **Per-category detection performance on Kvasir-Capsule.** We report category-wise AP (%) across the nine lesion categories for each detector and its UA-CP counterpart.

Method	Ang	Bf	Bh	Ero	Ery	Fb	LymPol	Ulc
CenterNet	47.6	72.7	50.0	48.3	43.4	64.7	65.272.0	57.4
YOLOv5	52.8	52.3	50.0	51.0	45.1	61.8	67.259.4	62.2
Faster RCNN-R50	39.9	37.9	30.0	18.8	24.0	38.5	46.276.8	35.8
Faster RCNN-Rxt101	46.3	64.9	80.0	33.9	39.9	64.6	63.375.7	48.5
FCOS	49.1	69.4	60.0	42.8	44.9	63.9	64.374.5	57.7
DINO-R50	51.0	69.5	100.0	41.8	37.7	62.3	60.470.2	55.5
Faster RCNN-R50+UACP	40.1	41.9	50.0	22.6	31.4	46.6	48.377.5	39.6
Faster RCNN-Rxt101+UACP	47.6	68.0	80.0	35.2	43.0	65.5	63.887.1	46.9
FCOS+UACP	47.6	67.2	60.0	41.0	50.3	57.7	64.272.8	58.7
DINO+UACP	50.0	71.6	100.0	43.3	40.3	62.0	61.682.3	60.1

Table 8: **Detection Performance on LD-PolypVideo Datasets:** We report AP50 (%), recall (%), and F1 (%). Values in parentheses indicate AP50 improvements over the corresponding detector baseline.

Method	AP50(%)	Recall(%)	F1(%)
CenterNet	55.5	31.1	39.9
YOLOv3	47.6	52.2	46.4
YOLOv5	53.4	53.0	50.1
Faster RCNN-R50	51.3	29.7	37.6
Faster RCNN-Rxt 101	<u>59.3</u>	52.4	32.2
FCOS	40.5	27.1	24.2
DINO-R50	54.7	52.8	27.3
Faster RCNN-R50+UA-CP	55.1(+3.8)	56.9	**62.5**
Faster RCNN-Rxt101+UA-CP	**69.1(+9.8)**	57.9	<u>61.9</u>
FCOS+UACP	51.8(+11.3)	<u>58.0</u>	49.3
DINO-R50+UACP	56.3(+1.6)	**65.9**	51.1

Table 9: **Computational Complexity on LGLDD.**: We report input resolution (H×W), number of parameters (M), and GFLOPs for each detector and its UA-CP counterpart.

Method	Resolusion (H×W)	Params (M)	GLOPs
YOLOv5	640 × 640	7.03	7.94
Faster RCNN-R50	800 × 832	57.40	163
Faster RCNN-Rxt101	800 × 832	76.02	220
FCOS	800 × 832	32.12	188
DINO-R50	800 × 832	47.44	137
Faster RCNN-R50+UACP	800 × 832	57.40	149
Faster RCNN-Rxt101+UACP	800 × 832	76.02	220
FCOS+UACP	800 × 832	32.12	137
DINO+UACP	800 × 832	47.55	197

D.2 Validate-Efficiency Analysis.

We examine whether the validity and efficiency of UA-CP are sensitive to the choice of calibration samples, nominal coverage levels, and RAPS hyperparameters.

Calibration-subset protocol. We construct 10 calibration replicates by sampling 75% of the calibration images Marques & Paulo (2025). The same ten accepted subsets are reused across all coverage and hyperparameter cells, while the evaluation set remains fixed. For each configuration, we report mean±standard deviation across the 10 calibration replicates.

Table 10: **Computational Complexity on Kvasir-Capsule.** We report input resolution (H×W), number of parameters (M), and GFLOPs for each detector and its UA-CP counterpart.

Method	Resolusion (H×W)	Params (M)	GLOPs
CenterNet	1024 × 1024	32.13	206
YOLOv5	640 × 640	7.04	7.96
Faster RCNN-R50	800 × 800	41.39	134
Faster RCNN-Rxt101	800 × 800	76.07	201
FCOS	800 × 800	32.13	123
DINO-R50	800 × 800	47.40	169
Faster RCNN-R50+UACP	800 × 800	41.39	134
Faster RCNN-Rxt101+UACP	800 × 800	76.07	201
FCOS+UACP	800 × 800	32.13	123
DINO+UACP	800 × 800	47.40	169

Table 11: **Computational Complexity on LDPolypVideo.** We report input resolution (H×W), number of parameters (M), and GFLOPs for each detector and its UA-CP counterpart.

Method	Resolusion (H×W)	Params (M)	GLOPs
YOLOv5	640 × 640	7.02	7.93
Faster RCNN-R50	896 × 1024	57.37	234
Faster RCNN-Rxt101	896 × 1024	75.99	305
FCOS	800 × 960	32.12	148
DINO-R50	800 × 960	47.39	202
Faster RCNN-R50+UACP	896 × 1024	57.38	234
Faster RCNN-Rxt101+UACP	896 × 1024	75.99	305
FCOS+UACP	800 × 960	32.12	148
DINO+UACP	800 × 960	47.39	202

Coverage-level sensitivity. The primary analysis independently varies the category and localization coverage targets over $\{0.90, 0.95, 0.99\}$. For each target pair, we report conditional label coverage, simultaneous box coverage, and empirical joint coverage. The theoretical lower bound on joint coverage follows from the union bound,

$$\max\{0,\ 1 - \alpha_{\text{label}} - \alpha_{\text{box}}\}\,,$$

This theoretical lower bound need *not equal* empirical joint coverage, which may vary due to finite-sample variability or deviations from exchangeability.

Figures 6 and 7 report the baseline and UA-CP results on LGLDD, while Figures 8 and 9 provide the correponding results on Kvasir-Capsule. Higher coverage targets generally increase both empirical coverage and uncertainty size. Given variation in localization coverage across architectures and datasets, we compare efficiency only among configurations with comparable empirical coverage.

Because LDPolypVideo contains a single foreground category, we vary only the box coverage target. Figure 10 shows that higher targets generally increase both box coverage and MPIW. Across operating points, higher coverage is accompanied by wider intervals, reflecting the coverage–efficiency trade-off in localization uncertainty.

RAPS-parameter sensitivity. Based on the protocolAngelopoulos et al. (2020), we evaluate sensitivity to the regularization strength $\lambda \in \{0, 0.001, 0.01, 0.1\}$ and onset rank k_{reg}. Figures 11 and 12 report the resulting mean label-set size and MPIW. Increasing k_{reg} generally weakens the rank penalty and can produce larger label sets, although the effect varies across detector architectures. Among configurations satisfying the validity requirement, we use $k_{\text{reg}} = 2$ and $\lambda = 0.01$ throughout the multi-class experiments. This common setting provides moderate regularization while maintaining stable uncertainty efficiency across the evaluated datasets and architectures.

Additional uncertainty validaity and Efficiency. We provide additional uncertainty results on Kvasir-Capsule andvLDPolypVideo, complementing the LGLDD analysis in Table 2. Table 12 reports category and localization uncertainty under the multi-category setting, where DCC maintains high category coverage, while localization and joint coverage vary across detector architectures. Table 13 reports localization uncertainty in the single-category setting, focusing on box coverage and interval width.

E Qualitative Experiments

This section complements the quantitative evaluation by organizing examples by detection challenge to illustrate how UA and DCC operate at different stages of UA-CP across detector architectures.

Protocol and visual grammar. We evaluate LGLDD, Kvasir-Capsule, and LDPolypVideo across four detector configurations: Faster R-CNN with R50 and RxT101 backbones, DINO, and FCOS. Based on the preceding analysis, we set $\alpha_L = \alpha_B = 0.05$, $k_{\text{reg}} = 2$, $\lambda_r = 0.01$. Since LDPolypVideo contains a single foreground class, only localization uncertainty is evaluated.

Table 12: **Category and localization uncertainty on Kvasir-Capsule.** Comparison of DCC with detector-specific conformal baselines across architectures. Results are reported as mean ± standard deviation across repeated calibration runs.

Detector	**Method**	*Cov$_L$.*	*Cov$_B$.*	*Join Cov.*	*Size*	*MPIW.*
FasterRCNN R50	Box-Std	0.95±0.004	0.96±0.007	0.90±0.009	1.17±0.031	78.51±4.05
	DCC	0.98±0.003	0.96±0.007	0.94±0.007	1.50±0.028	78.51±4.05
FasterRCNN Rxt101-UA	Box-Std	0.98±0.000	0.97±0.005	0.95±0.005	1.06±0.008	54.99±2.78
	DCC	0.99±0.000	0.97±0.005	0.96±0.005	1.00±0.000	54.99±2.78
DINO	Class-thr	0.9778±0.0000	0.9542±0.0024	0.94±0.002	2.72±0.115	50.75±3.03
	DCC	0.9926±0.0000	0.95±0.002	0.95±0.002	1.00±0.001	50.75±3.03
Fcos	Class-thr	0.99±0.000	0.97±0.008	0.95±0.0077	1.046±0.001	60.48±6.18
	DCC	1.00±0.001	0.97±0.008	0.96±0.008	1.000±0.000	60.48±6.18

Table 13: **Localization uncertainty on LDPolypVideo.** Comparison of DCC with detector-specific conformal baselines across architectures. Results are reported as mean ± standard deviation across repeated calibration runs.

Detector	**Method**	*Cov$_B$.*	*Join Cov.*	*MPIW.*
FasterRCNN R50	Box-Std	0.89±0.003	0.89±0.003	70.08±1.05
	DCC	0.95 ± 0.002	0.95 ± 0.002	69.4401 ± 0.81
FasterRCNN Rxt101-UA	Box-Std	0.88±0.005	0.88±0.005	62.46±0.90
	DCC	0.87 ± 0.004	0.87 ± 0.004	62.3417 ± 0.89
DINO	Class-thr	0.85±0.003	0.85±0.003	50.19±0.49
	DCC	0.84 ± 0.003	0.84 ± 0.003	47.8521 ± 0.44
Fcos	Box-Std	0.88±0.004	0.88±0.004	66.89±0.74
	DCC	0.88 ± 0.004	0.88 ± 0.004	66.89 ± 0.74

Rows correspond to detector architectures, while columns show the input with ground truth, raw detections, UA-enhanced detections, conformalized raw detections, and the complete UA-CP output. denotes the ground-truth frame, and denotes the blue inner and outer conformal boundaries, with the shaded outer-minus-inner band representing localization uncertainty.

Small and low-contast lesions. Small lesions and lesions with low mucosal contrast are susceptible to confidence filtering or competition from more salient off-target detections. Figures 13–16 illustrate this challenge across capsule endoscopy, gastroscopy, and colonoscopy. UA more consistently retains lesion-aligned detections, while DCC represents the remaining localization uncertainty as calibrated plausible regions rather than segmentation masks.

Irregular and diffuse boundaries.

Irregular, diffuse, and peripheral boundaries Endoscopic lesions with diffuse boundaries, irregular shapes, or limited context near image borders primarily challenge localization. Figures 18–22 show that, despite plausible point boxes, uncertainty remains in the lesion extent. DCC explicitly represents this residual ambiguity through the inner and outer conformal boundaries, with narrower bands indicating more localized uncertainty.

Category Ambiguity and Prediction Sets For ambiguous lesion appearances, category and localization uncertainty are interpreted jointly: UA retains a lesion-aligned detection, while DCC represents the remaining ambiguity through plausible category sets and spatial regions.

Multiple lesions, missed detections, and candidate clutter Figures 26 and 27 illustrate the complementary roles of UA and DCC in challenging multi-candidate scenes: UA improves candidate retention, while DCC quantifies uncertainty for the resulting detections rather than correcting off-target predictions.

Failure analysis. The qualitative examples also expose distinct failure modes across the UA-CP pipeline. *Candidate failure* occurs when no lesion-aligned detection is retained, in which case subsequent conformal calibration cannot recover the missed lesion. *Localization undercoverage* occurs when the ground-truth box is not fully contained within the outer conformal region, whereas *localization inefficiency* corresponds to spatially broad or image-clipped regions. For the multi-class LGLDD and Kvasir-Capsule datasets, *category-set failure* additionally occurs when the ground-truth category is excluded from the conformal prediction set. These examples illustrate representative failure modes in the pipeline and highlight directions for improving candidate retention, localization quality, and multi-lesion recovery.

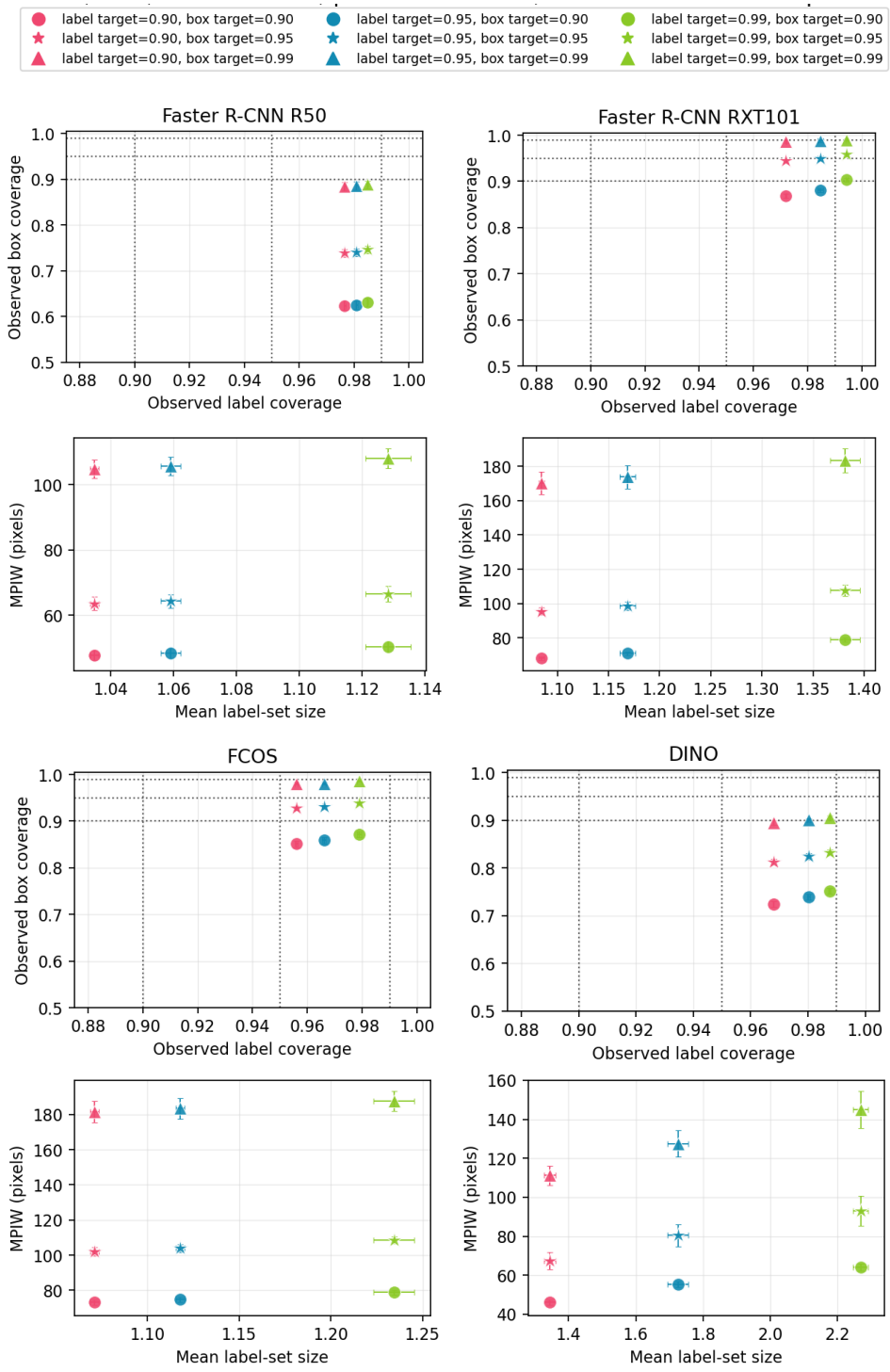

Figure 6: **Coverage–efficiency trade-off for baseline detectors on LGLDD.** Across varying category and box coverage targets, we report observed joint coverage behavior (top) and uncertainty efficiency measured by mean label-set size and MPIW (bottom) for four detector configurations. Error bars indicate variability across repeated calibration splits.

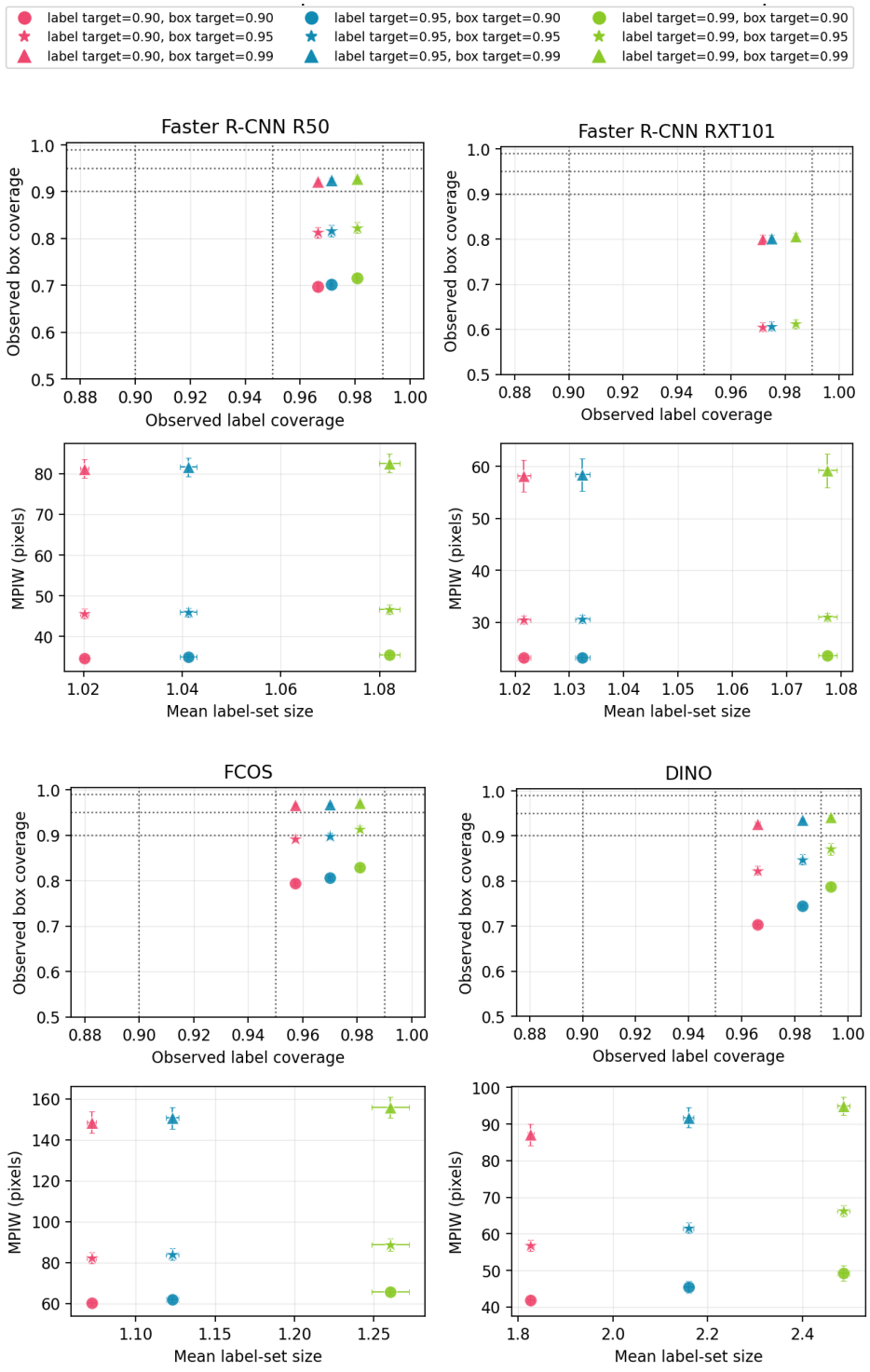

Figure 7: **Coverage–efficiency trade-off for UA-CP detectors on LGLDD.** Across varying category and box coverage targets, we report observed joint coverage behavior (top) and uncertainty efficiency measured by mean label-set size and MPIW (bottom) for four detector configurations. Error bars indicate variability across repeated calibration splits.

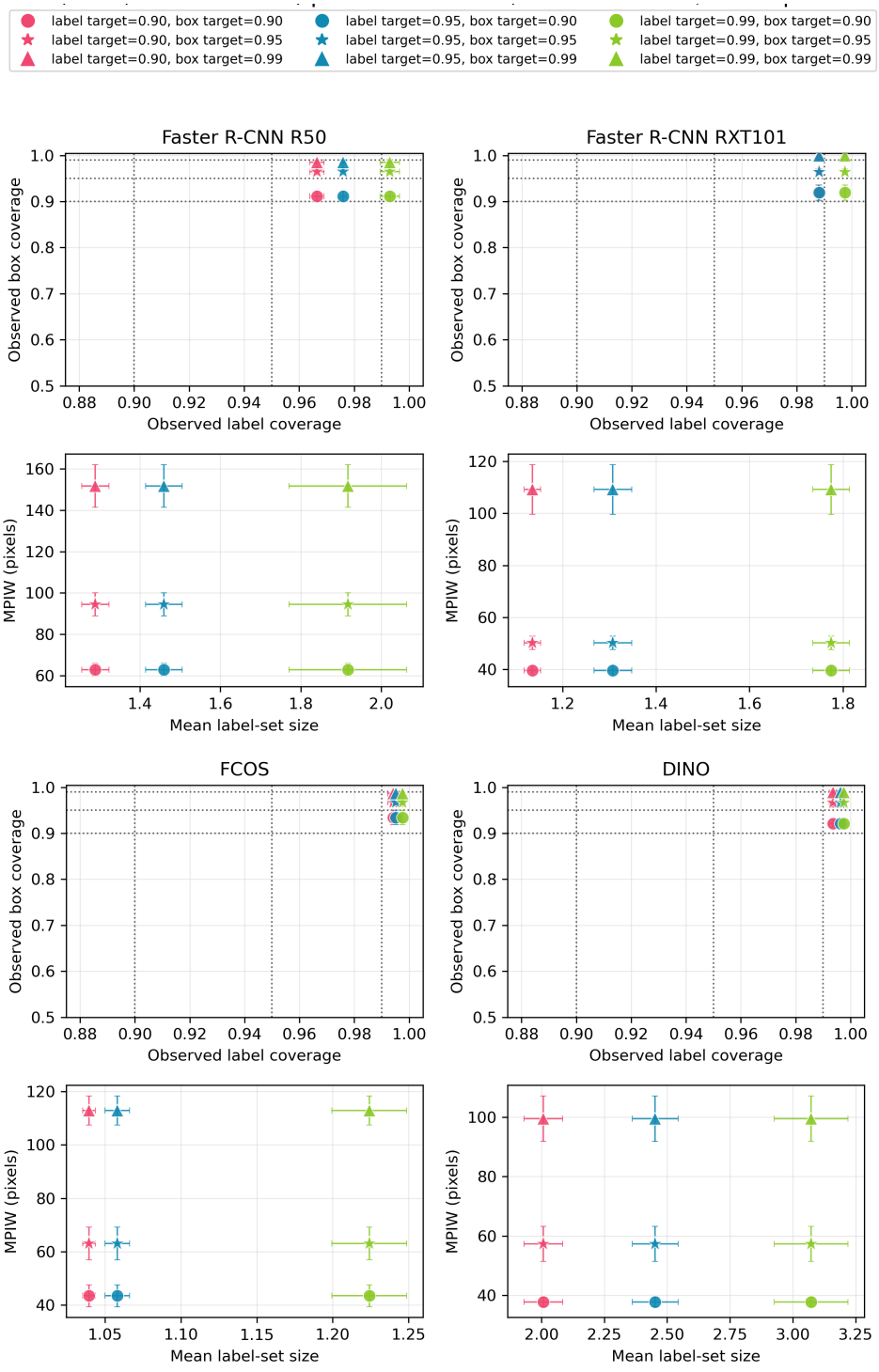

Figure 8: **Coverage–efficiency trade-off for baseline detectors on Kvasir-Capsule.** Across varying category and box coverage targets, we report observed joint coverage behavior (top) and uncertainty efficiency measured by mean label-set size and MPIW (bottom) for four detector configurations. Error bars indicate variability across repeated calibration splits.

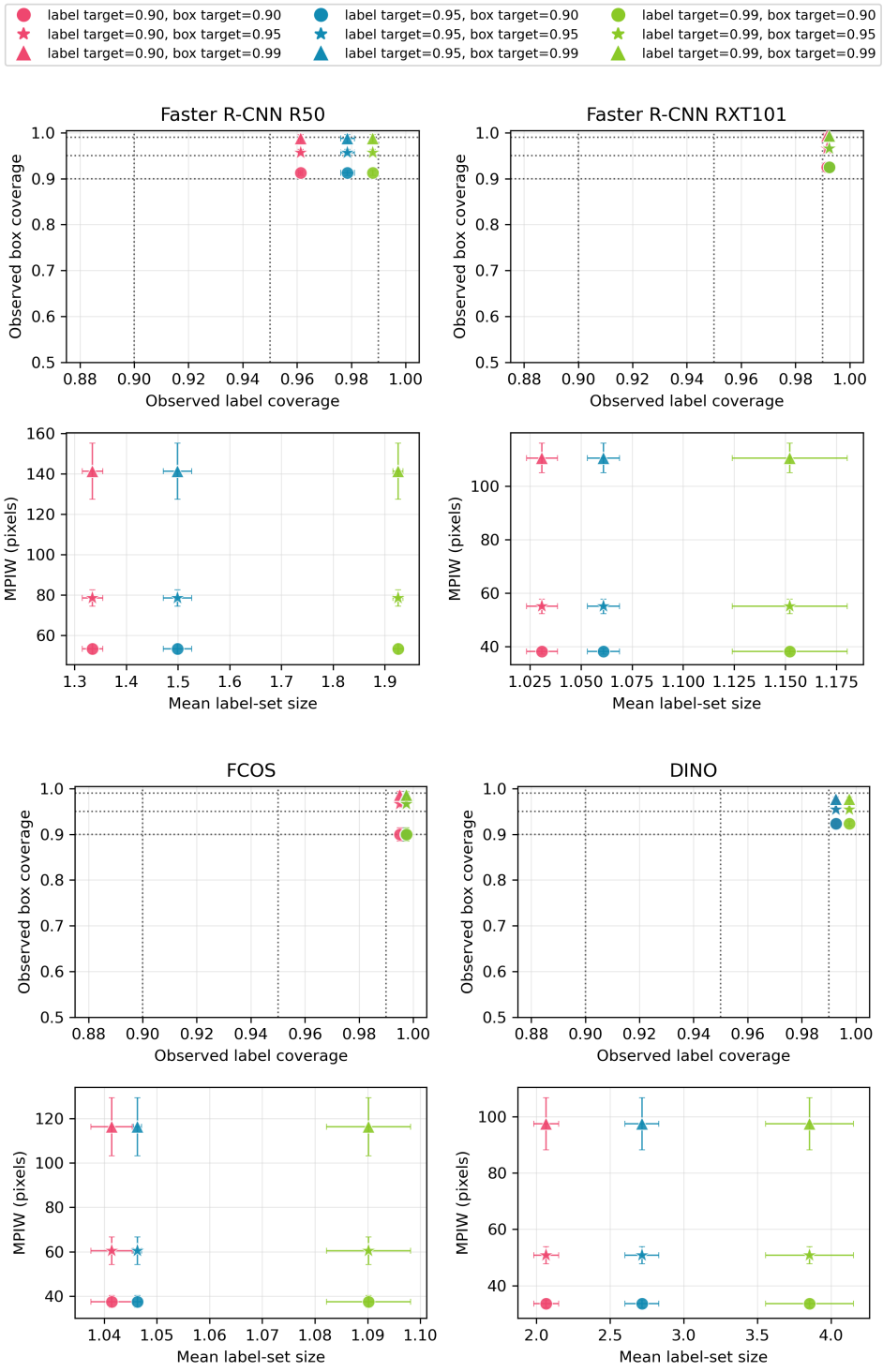

Figure 9: **Coverage–efficiency trade-off for UA-CP detectors on Kvasir-Capsule.** Across varying category and box coverage targets, we report observed joint coverage behavior (top) and uncertainty efficiency measured by mean label-set size and MPIW (bottom) for four detector configurations. Error bars indicate variability across repeated calibration splits.

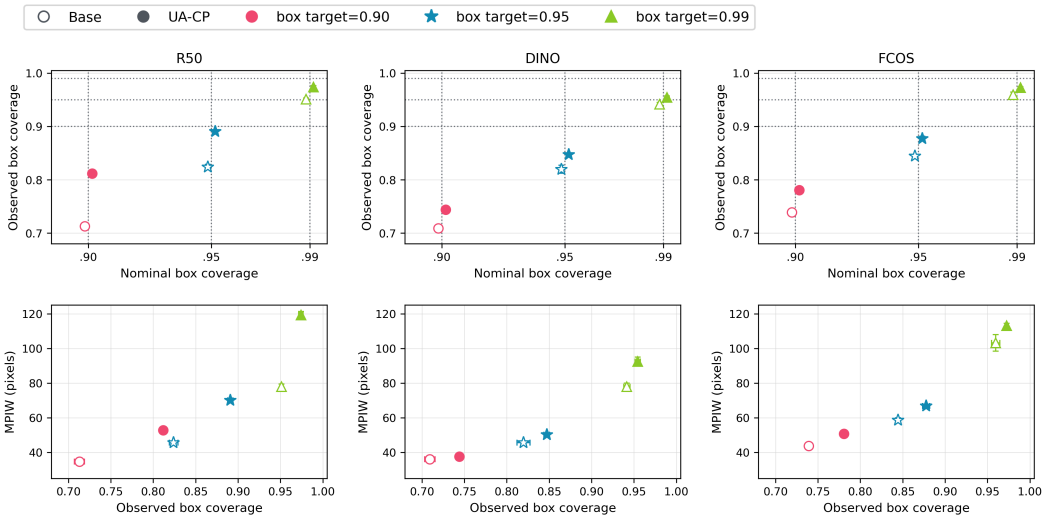

Figure 10: **Coverage–efficiency robustness on LDPolypVideo.** We compare Base and UA-CP for three detector configurations. We report observed joint coverage behavior (top) and uncertainty efficiency measured by mean label-set size and MPIW (bottom). Error bars indicate variability across repeated calibration subsets.

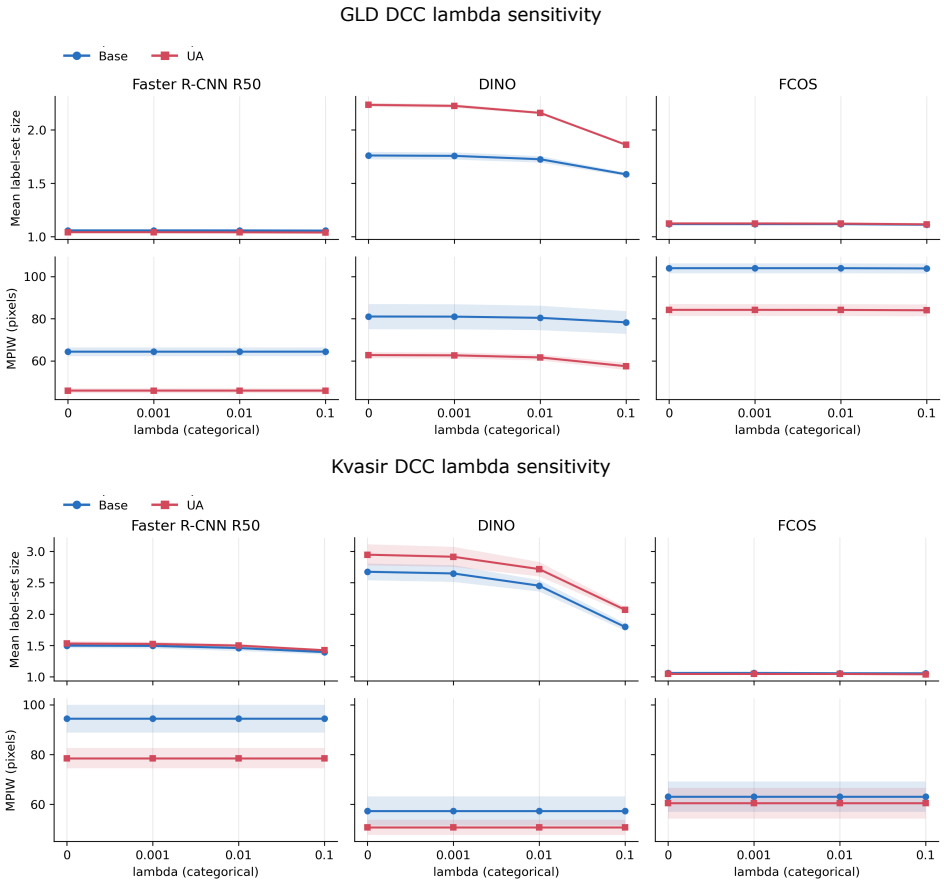

Figure 11: **Sensitivity to the RAPS regularization strength λ_R on LGLDD (top) and Kvasir-Capsule (bottom).** Mean label-set size and MPIW are reported under the Base and UA seetings across $\lambda_R \in \{0, 0.001, 0.01, 0.1\}$. Shaded regions indicate variability across calibration replicates, with lower label-set size and MPIW indicating greater category and localization efficiency, respectively.

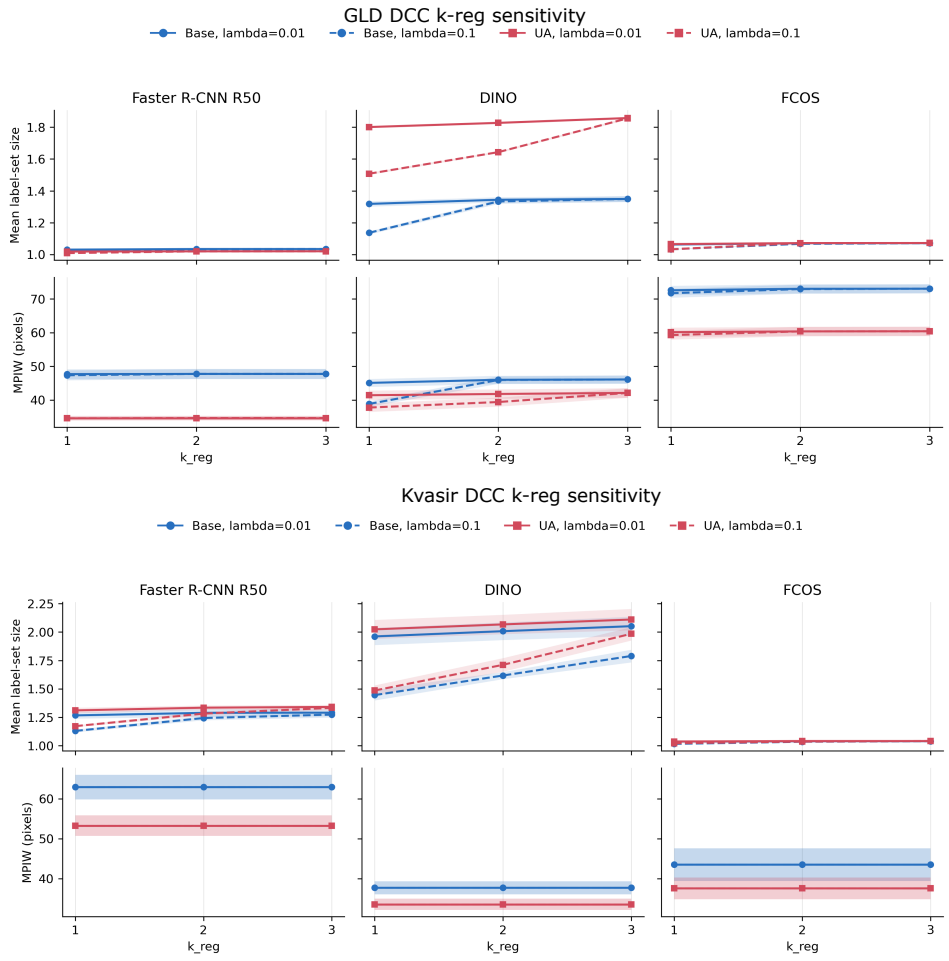

Figure 12: **Sensitivity to the RAPS onset rank k_{reg} on LGLDD (top) and Kvasir-Capsule (bottom).** Mean label-set size and MPIW are reported under the Base and UA settings across $k_{\text{reg}} \in \{1, 2, 3\}$. Solid and dashed curves correspond to $\lambda = 0.01$ and 0.1, respectively, and shaded regions indicate variability across calibration replicates. Lower label-set size and MPIW indicate greater uncertainty efficiency.

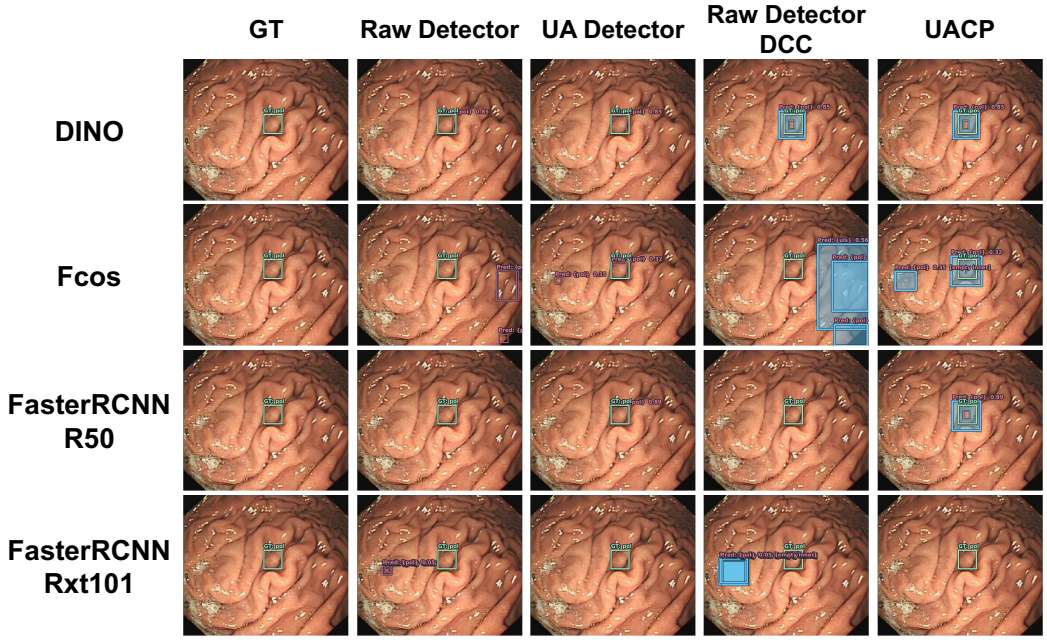

Figure 13: Small lesion on textured mucosa in Kvasir-Capsule. The textured background produces detector-dependent candidate behavior; UA-CP retains a lesion-aligned hypothesis and represents its residual localization uncertainty across all four architectures.

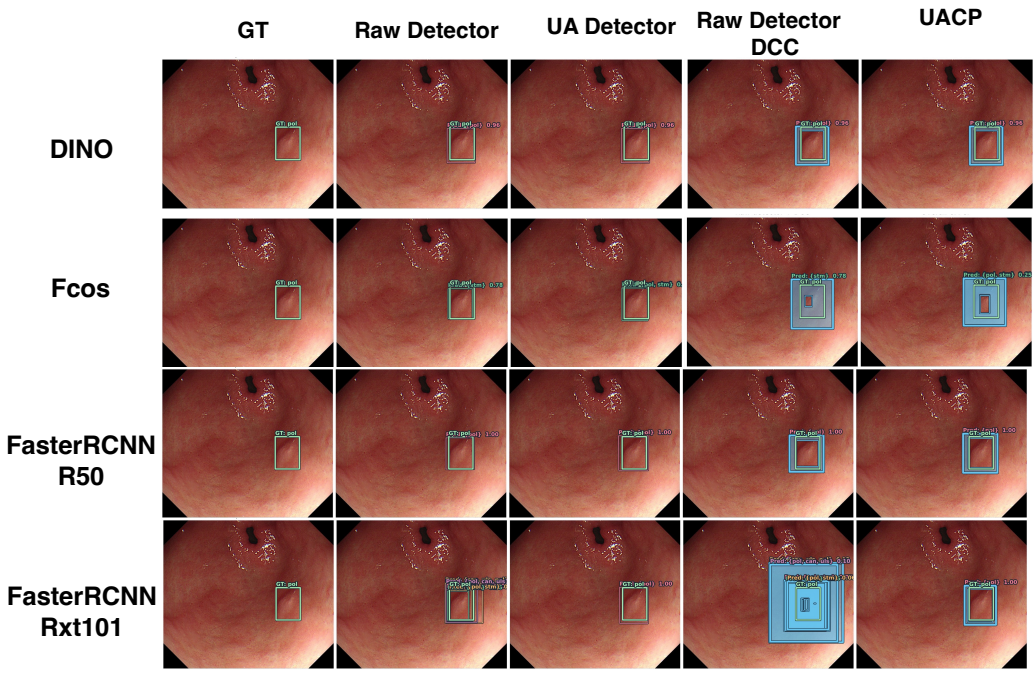

Figure 14: Small, low-contrast polyp in LGLDD. The lesion occupies a limited image area and resembles the adjacent mucosa, making candidate retention and precise boundary placement difficult.

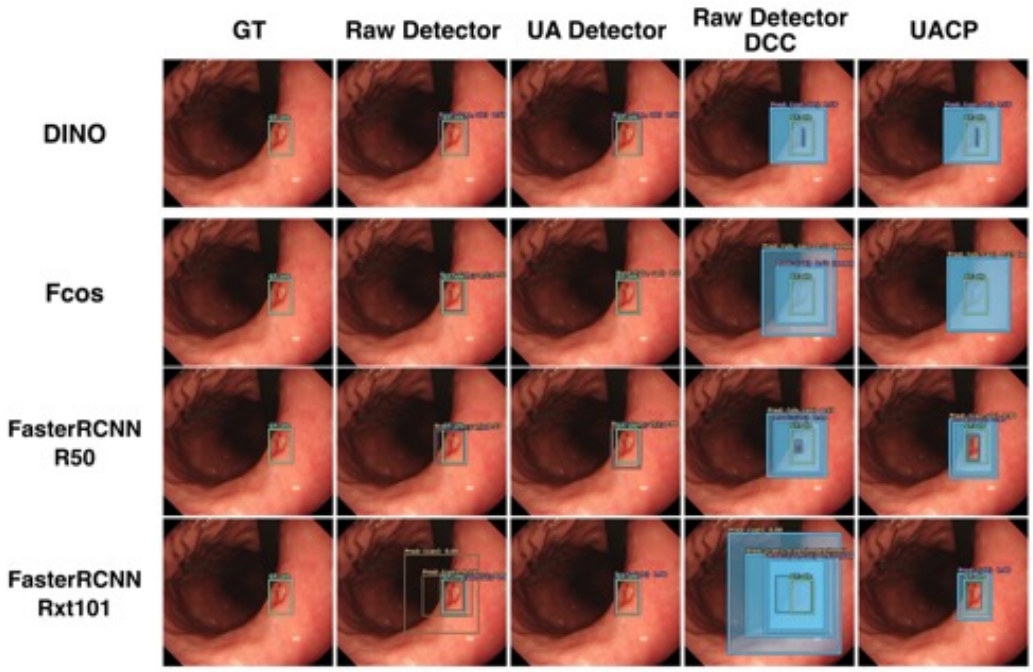

Figure 15: Small ulcer in LGLDD. The comparison separates candidate recovery from localization uncertainty: UA retains a lesion-aligned prediction, while the nested DCC frames show the detector-specific uncertainty in its spatial extent.

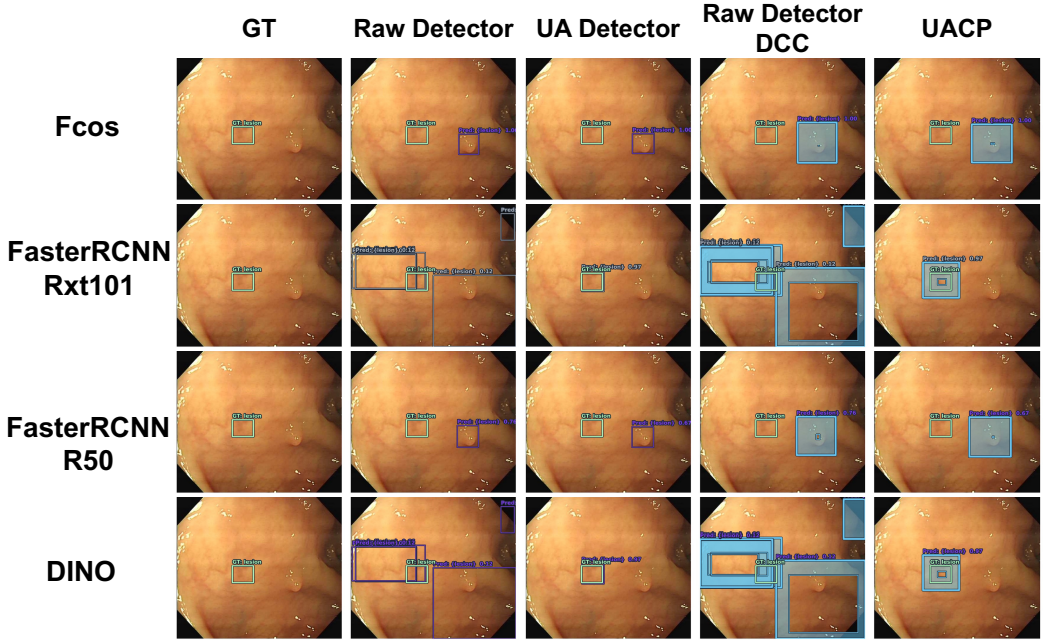

Figure 16: Small lesion in LDPolypVideo. Because this dataset is single-class, the comparison is restricted to candidate retention and localization uncertainty.

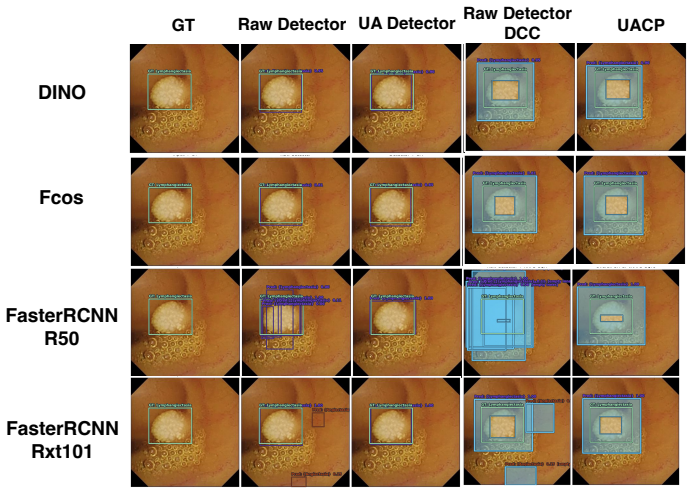

Figure 17: *Qualitative* comparison of *multi-object detection capability* on the LGLDD dataset.

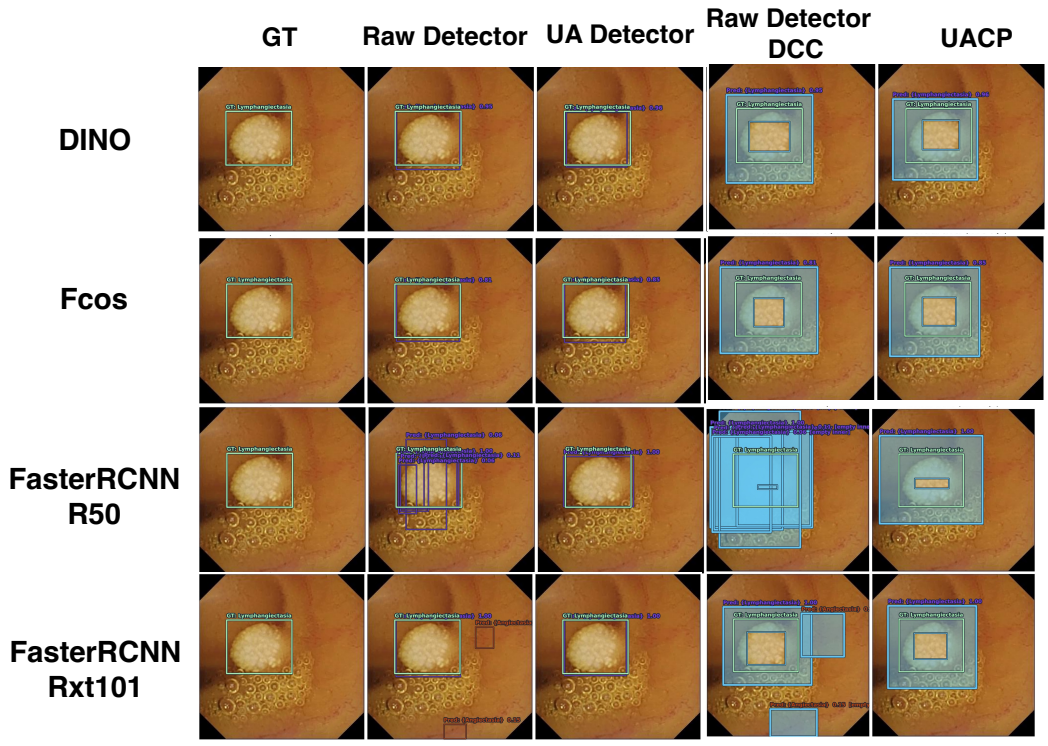

Figure 18: Cross-detector boundary rescue in Kvasir-Capsule. The same image and ground-truth box are shown for every architecture. Detector-specific point boxes differ, whereas the UA-CP regions make the remaining boundary uncertainty visible.

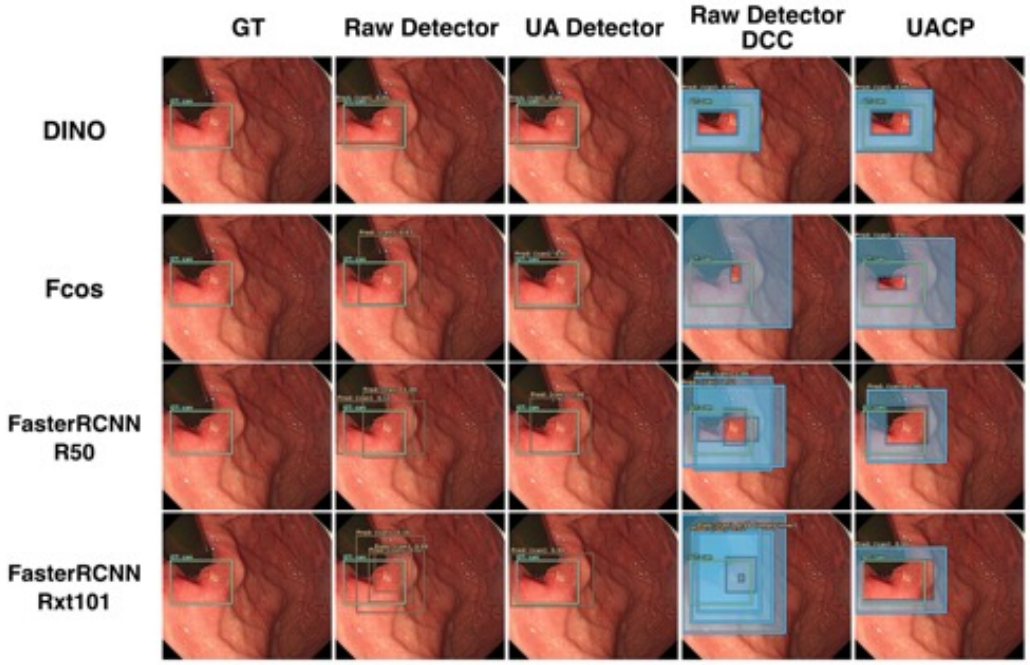

Figure 19: Cancer with an irregular boundary in LGLDD. The comparison illustrates why a single rigid point box is insufficient when the lesion-mucosa transition is spatially ambiguous.

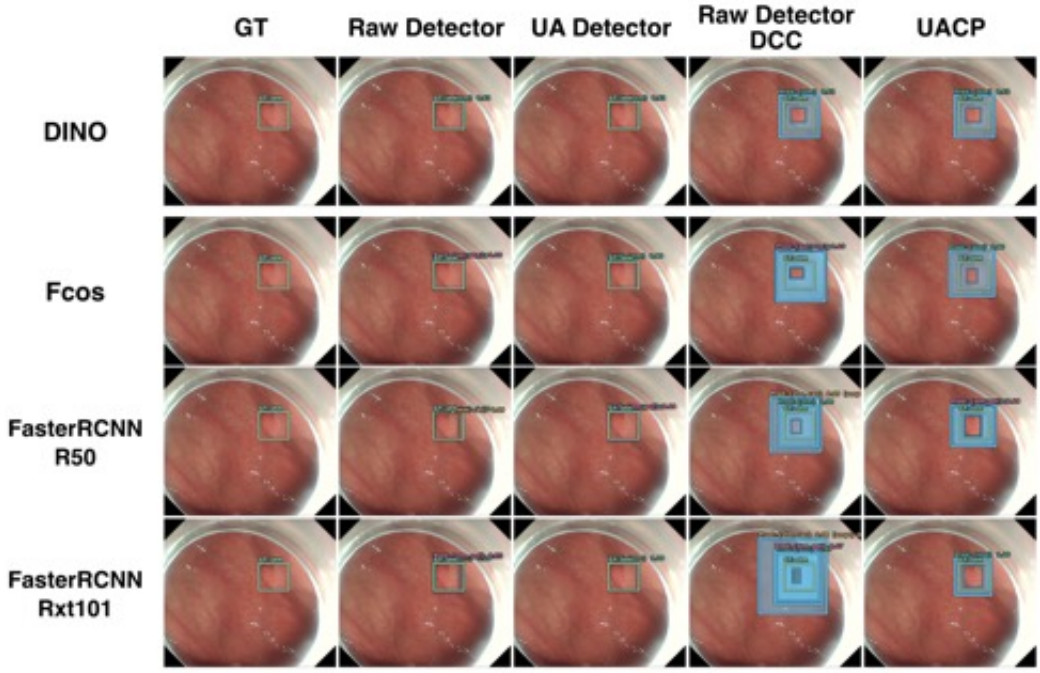

Figure 20: Diffuse submucosal-tumor boundary in LGLDD. The blue outer-minus-inner band represents the plausible localization range remaining after UA has selected the candidate.

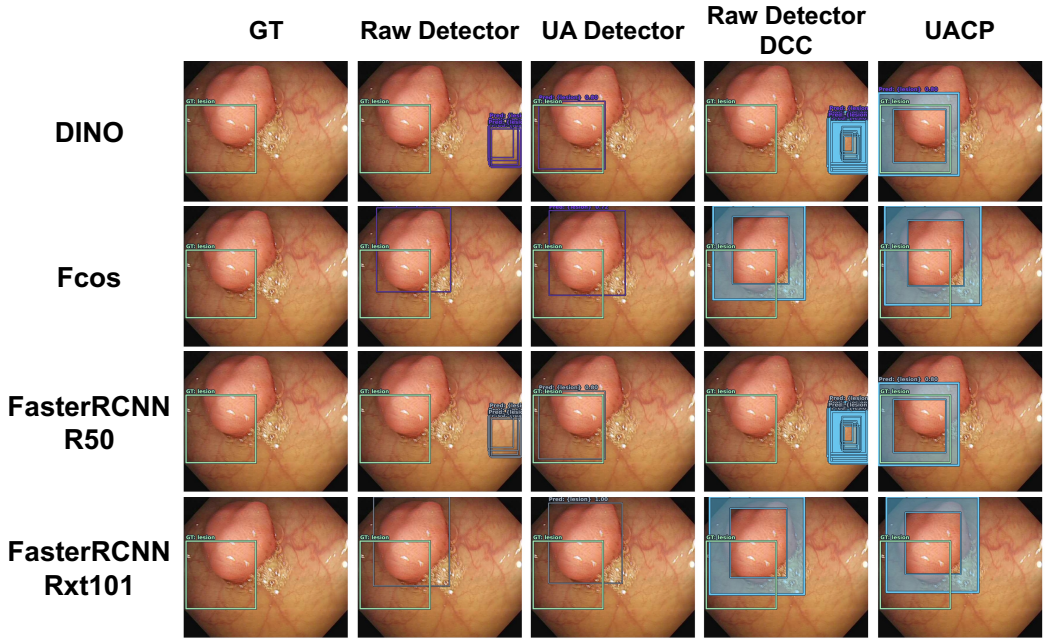

Figure 21: Diffuse lesion boundary in LDPolypVideo. The variation in conformal-region width across architectures reflects detector-dependent localization uncertainty on the same single-class target.

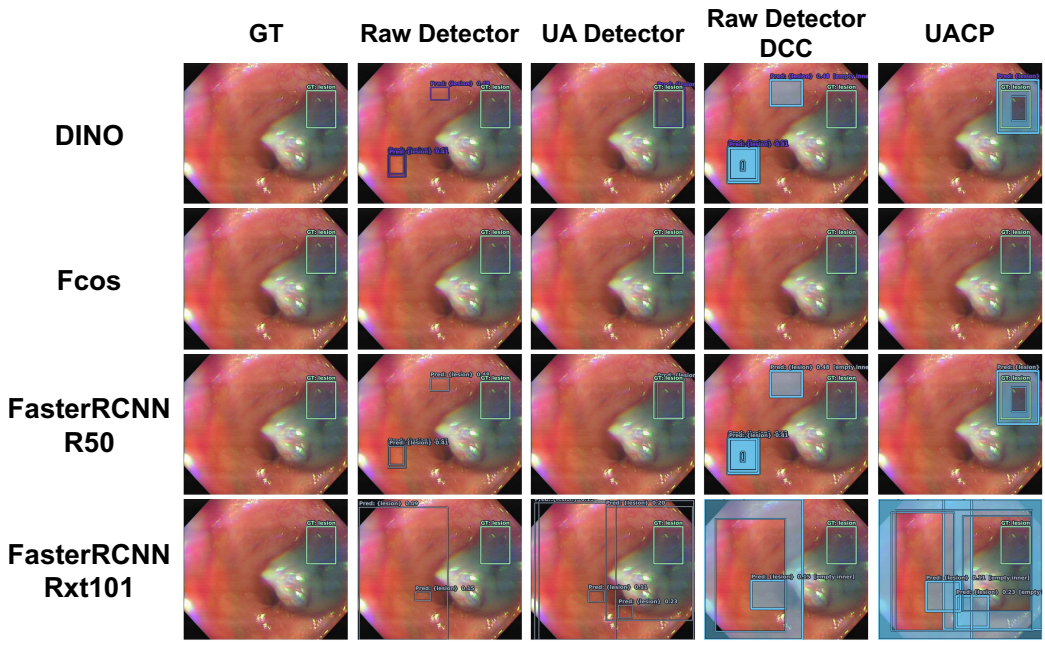

Figure 22: Peripheral, fine-scale localization challenge in LDPolypVideo. Limited surrounding context and a small spatial target lead to marked differences in the compactness of the detector-specific UA-CP regions.

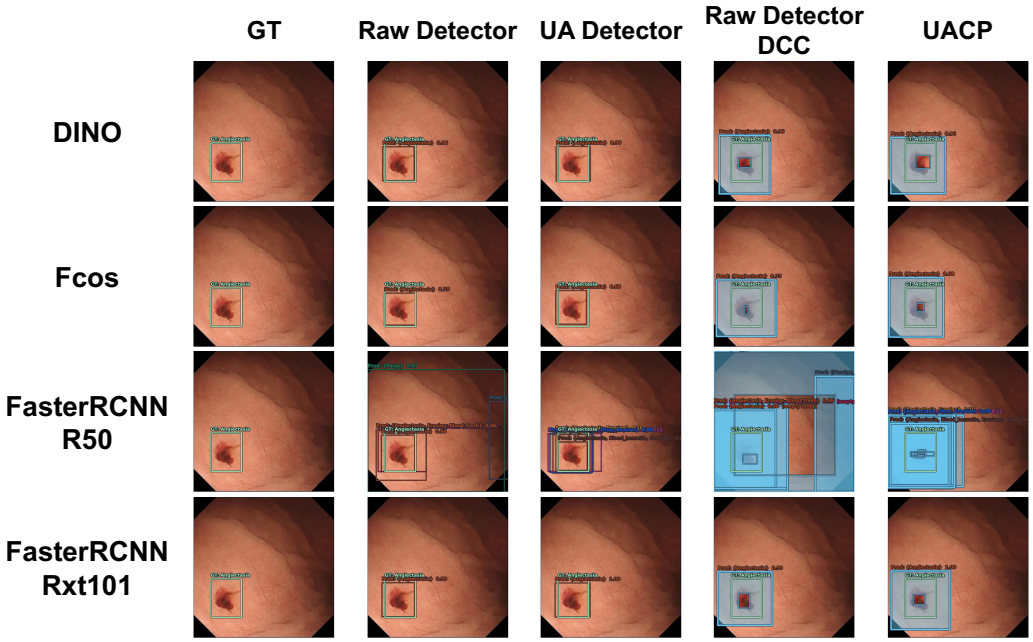

Figure 23: Semantic-set rescue in Kvasir-Capsule. The example shows how UA-CP represents residual category ambiguity after candidate selection, while simultaneously retaining a localized conformal region.

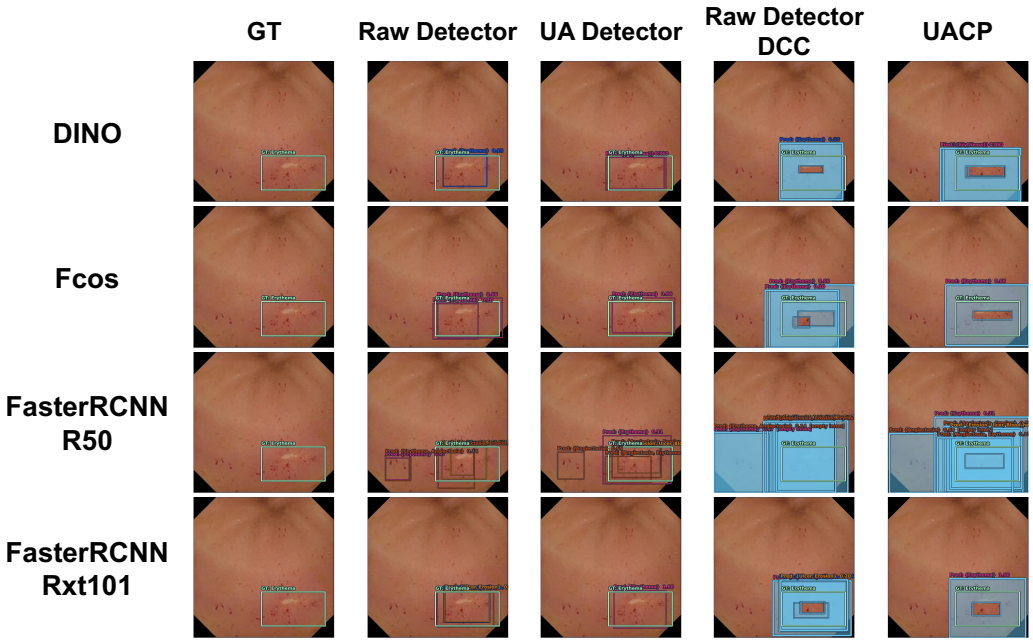

Figure 24: Coupled boundary rescue and category diffusion in Kvasir-Capsule. Cross-detector differences in both point localization and plausible categories motivate the joint semantic–spatial representation used by UA-CP.

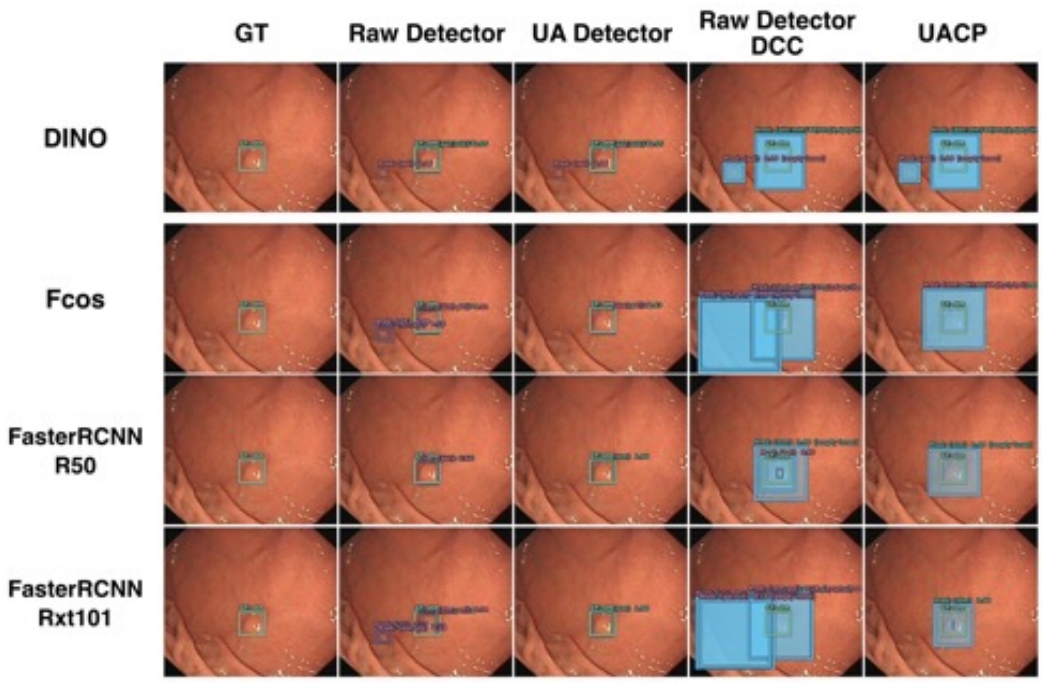

Figure 25: Submucosal-tumor example in LGLDD. UA reduces off-target candidate competition, and UA-CP records the remaining detector-dependent semantic and localization ambiguity.

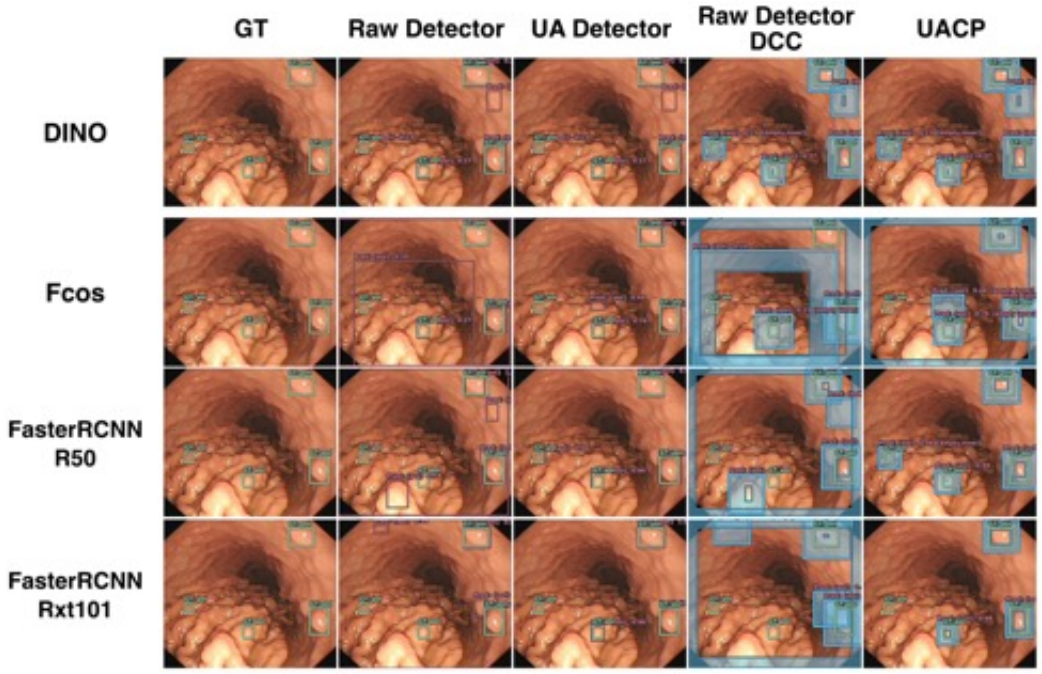

Figure 26: Multiple polyps in LGLDD. Raw conformalization propagates proposal clutter into multiple wide regions, whereas UA-CP applies uncertainty quantification to a cleaner lesion-aligned candidate set.

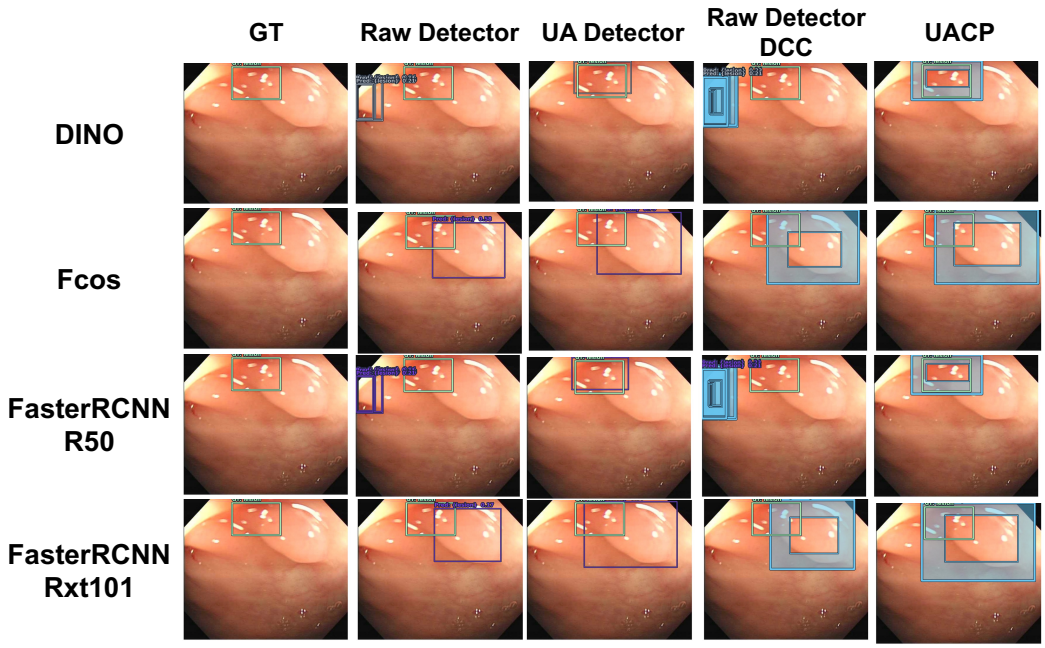

Figure 27: Candidate-selection and missed-detection challenge in LDPolypVideo. The single-class example isolates detection and localization behavior: UA changes which candidate is retained, and DCC represents the remaining spatial uncertainty.