ADAPTIVE NESTED-PATH CONFORMAL PREDICTION FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty quantification is essential for medical image segmentation, where missed object regions can affect downstream clinical decisions. We present Conformal Prediction Adaptive Segmentation (CPAS), a conformal prediction framework for image segmentation that constructs compact prediction sets with finite-sample coverage guarantees. Unlike confidence-thresholding methods that treat pixels independently, and morphological dilation methods, which expand masks uniformly, CPAS defines a nested expansion path that jointly exploits model confidence and spatial structure. Starting from the predicted mask, CPAS iteratively adds boundary pixels through a confidence-guided exploitation step combined with a spatial exploration mechanism, and calibrates a threshold on the accumulated uncertainty along this path. The resulting sets adapt to both the model's soft predictions and the object geometry while preserving conformal validity. More generally, we formalize conformal calibration over input-dependent nested prediction paths. Experiments on multiple medical imaging benchmarks show that CPAS achieves the target object coverage while producing substantially more compact prediction sets than conformal segmentation baselines. On average, CPAS reduces stretch by approximately 20% compared to previous methods. These results demonstrate that confidence-guided spatial expansion is an effective approach for producing compact, statistically valid segmentation uncertainty sets.

1 INTRODUCTION

Accurate and trustworthy image segmentation is essential in many medical imaging pipelines, supporting tasks such as organ delineation, tumor characterization, and treatment planning. Despite major advances in deep neural networks, current segmentation models typically operate as deterministic predictors, providing little insight into where they may be unreliable or how large the errors may be. This lack of calibrated uncertainty is particularly problematic in safety-critical clinical settings, where segmentation errors can propagate into downstream diagnostic or interventional decisions. Sources of variability such as heterogeneous acquisition protocols, subtle pathologies, annotation ambiguity, and distribution shift between training and deployment environments further exacerbate this issue. Consequently, there is a critical need for segmentation frameworks that not only produce accurate masks but also quantify their uncertainty with statistical guarantees, enabling clinicians to understand when model outputs can be trusted and when caution is warranted.

Conformal Prediction (CP) (Vovk et al., 2005; Angelopoulos et al., 2024a) offers a principled, model-agnostic framework for constructing prediction sets (i.e. segmentations) that contain the ground–truth segmentation with finite-sample guarantees. Prior work on CP for segmentation can be broadly divided into two main approaches. (1) *Confidence-aware calibration methods*, such as Conformal Risk Control (CRC) (Angelopoulos et al., 2024b; Blot et al., 2025), rely exclusively on per-pixel confidence information. They define the prediction set as the set of pixels whose object class probability is above a calibrated threshold. (2) *Spatially-aware calibration methods*, such as Conformal Prediction for Segmentation (CPS) (Mossina & Friedrich, 2025) which defines a nonconformity score for each sample as the minimum number of morphological dilations required for the model's predicted mask to cover a prescribed fraction of the ground-truth mask. However, existing methods do not formulate conformal segmentation as an explicit image-dependent nested expansion path in which model confidence determines where the set grows and spatial structure constrains the expansion.

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.

Our key idea is to formulate segmentation uncertainty as conformal calibration over an image-dependent nested expansion path. Rather than calibrating a fixed pixel-wise threshold or a uniform geometric dilation, CPAS uses the model predictions and spatial structure to determine an ordered sequence of candidate segmentation sets, while conformal calibration determines how far along this sequence to expand. This separation allows the expansion policy to adapt to each image while preserving finite-sample validity through a nested expansion path and a nondecreasing path score. We instantiate this principle using a confidence-guided exploitation–exploration rule and an uncertainty-weighted path score. The resulting prediction sets concentrate expansion in regions that are plausible according to both the model confidence and the object geometry, yielding substantially more compact sets while retaining the desired coverage guarantee. More generally, we show that any input-dependent nested prediction path constructed without access to the ground-truth label can be conformally calibrated through a nondecreasing path score, and we derive the binary and joint multi-class CPAS constructions as instances of this principle.

2 Related Work

Standard segmentation networks usually return a single deterministic mask, or a pixel-wise probability map, without providing a statistically valid statement about the reliability of the predicted region. This shortcoming is particularly important in medical imaging, where missed object regions can affect diagnosis, treatment planning, or downstream quantitative measurements. A large body of work has studied uncertainty quantification for segmentation, including Bayesian approximations and Monte Carlo dropout (Gal & Ghahramani, 2016; Kendall & Gal, 2017), deep ensembles (Lakshminarayanan et al., 2017; Mehrtash et al., 2020), test-time augmentation (Wang et al., 2019), probabilistic segmentation models (Kohl et al., 2018), and post-hoc calibration methods (Guo et al., 2017; Mehrtash et al., 2020). While these methods can provide informative uncertainty estimates, they generally do not translate uncertainty into prediction sets with distribution-free finite-sample coverage guarantees, motivating conformal prediction as a more principled framework for reliable medical image segmentation.

CP provides a distribution-free framework for constructing prediction sets with rigorous coverage guarantees. This makes conformal methods especially attractive for medical imaging, where reliability requirements are high and where model confidence scores may be miscalibrated due to limited data or annotation variability. CP has been applied to several medical imaging classification and regression tasks, including skin-lesion classification, intracranial hemorrhage detection, and anatomical landmark localization (see e.g. (Fayyad et al., 2024; Gamble et al., 2025; Jonkers et al., 2026)). These studies and many others illustrate the growing use of CP across classification and regression problems in medical imaging, while highlighting the need for methods tailored to structured outputs such as segmentation.

More recently, CP has been extended to structured prediction problems, where the output space is high-dimensional and the goal is to control a user-defined risk functional rather than a simple misclassification event. Conformal Risk Control (CRC) extends the conformal framework to settings in which the quality of a prediction set is measured by a general loss or risk function (Angelopoulos et al., 2024b). In image segmentation, CRC can be used to calibrate a threshold on the model's pixel-wise confidence scores. The resulting prediction set contains all pixels whose object probability exceeds the calibrated threshold. This approach is simple, model-agnostic, and has finite-sample guarantees. However, since the calibrated set is determined independently at each pixel, the resulting segmentation may be spatially fragmented. Moreover, CRC relies solely on the confidence values and does not explicitly exploit the geometric structure of the predicted object.

A complementary line of work constructs CP sets for segmentation using spatial operations. CP for Segmentation (CPS) calibrates the number of morphological dilation steps required for the predicted mask to cover a prescribed fraction of the ground-truth object (Mossina & Friedrich, 2025). This produces nested prediction sets and thus fits naturally within the CP framework. However, purely spatial expansion also has an important limitation. Since standard morphological dilation expands the mask uniformly in all directions, it does not use the model's confidence map to decide where the object is more likely to extend. As a result, CPS can add many low-confidence pixels that are geometrically close to the predicted mask but unlikely to belong to the target object.

Several recent studies have extended conformal prediction for segmentation by modifying the prediction-set construction or incorporating spatial information. Kandinsky conformal prediction groups pixels according to similarities between their nonconformity curves and calibrates the resulting groups separately (Brunekreef et al., 2024). SACP introduces spatially weighted, class-conditional conformity scores, with particular emphasis on clinically important interfaces (Bereska et al., 2025). Davenport constructs inner and outer segmentation sets that provide a two-sided characterization of mask uncertainty (Davenport, 2025). CONSIGN accounts for correlations between pixels by decomposing stochastic segmentation outputs into spatial components and calibrating prediction sets in the resulting representation (Viti et al., 2026). RW-CP uses a random walk over a graph derived from pretrained foundation-model features to diffuse the model's probability scores, producing conformal sets with improved spatial coherence (Gaillochet et al., 2026).

These approaches demonstrate several ways of incorporating spatial dependence into conformal segmentation, including spatial weighting, score transformation, decomposition, and graph-based diffusion. CPAS instead constructs an explicit image-dependent nested expansion path: the confidence map and local boundary structure determine where the prediction set grows, while conformal calibration determines how far along this path the expansion should proceed.

3 Conformal Prediction for Adaptive Segmentation

Problem setup. Consider a binary image segmentation problem with a target object. Let Ω denote the set of pixels in image I. Let f denote a segmentation model trained on a labeled image dataset to identify the object of interest in the image I. The segmentation network outputs a soft object probability map $f(x) = p(x \in \text{object}|I) \in [0, 1]$ for each pixel $x \in \Omega$. The predicted segmentation $\hat{Y}$ is identified using a fixed threshold τ, i.e., $\hat{Y}(I) = \{x \in \Omega | f(x) > \tau\}$.

Although segmentation methods aim to reduce false negatives in the predictions, they do not provide any statistical guarantees regarding the coverage of the segmentation results. To do so, one can apply the CP procedure by augmenting the predicted segmentation $\hat{Y}$. For a new test sample (I, Y), where the ground-truth mask Y is unknown at inference time, our goal is to construct a prediction set $C(I) \subset \Omega$ that is as small as possible while satisfying

$$\mathbb{P}\left[\frac{|Y \cap C(I)|}{|Y|} \geq \rho\right] \geq 1 - \alpha, \tag{1}$$

where α represents the allowable failure probability and ρ specifies the desired coverage fraction. This criterion ensures that the prediction set covers at least a fraction ρ of the true object Y, thereby controlling false negatives in clinically relevant regions.

Validity guarantee. The key observation is that conformal validity does not depend on the particular mechanism used to construct the prediction path. It is sufficient that the candidate sets are nested and equipped with a nondecreasing path score, as formalized below.

Theorem 1 (Conformal calibration of discrete adaptive nested paths)**.** *Let $\{(I_i, Y_i)\}_{i=1}^n$ and a test sample (I, Y) be exchangeable. For each input I, consider an input-dependent sequence of prediction sets $C_0(I) \subseteq C_1(I) \subseteq \cdots \subseteq C_T(I)$, constructed without using the corresponding ground-truth label, together with a nondecreasing sequence of path scores*

$$\lambda_0(I) \leq \lambda_1(I) \leq \cdots \leq \lambda_T(I). \tag{2}$$

Define $G(Y, C) = |Y \cap C|/|Y|$. For a prescribed target level ρ, define the oracle stopping index

$$T_\rho(I, Y) = \min\{t : G(Y, C_t(I)) \geq \rho\}, \tag{3}$$

and the corresponding nonconformity score $S(I, Y) = \lambda_{T_\rho(I,Y)}(I)$. Let $q = S_{(\lceil (n+1)(1-\alpha) \rceil)}$ be the standard split-conformal quantile of the calibration scores $S(I_1, Y_1), \ldots, S(I_n, Y_n)$. For a test input I, define $T_q(I) = \min\{t : \lambda_t(I) > q\}$, with $T_q(I) = T$ if no such index exists, and return $C^q(I) = C_{T_q(I)}(I)$. Then $\Pr(G(Y, C^q(I)) \geq \rho) \geq 1 - \alpha$.

Proof. Suppose that the test-sample score satisfies $S(I, Y) \leq q$. By definition, $\lambda_{T_\rho(I,Y)}(I) \leq q$. Since the sequence $\{\lambda_t(I)\}_{t=0}^T$ is nondecreasing, the first index at which $\lambda_t(I) > q$ cannot occur before $T_\rho(I, Y)$. Hence, $T_q(I) \geq T_\rho(I, Y)$. Because the prediction sets are nested,

$$C_{T_\rho(I,Y)}(I) \subseteq C_{T_q(I)}(I) = C^q(I).$$

Using the monotonicity of G with respect to set inclusion, we obtain

$$G(Y, C^q(I)) \geq G\big(Y, C_{T_\rho(I,Y)}(I)\big) \geq \rho.$$

Therefore,

$$\{S(I,Y) \leq q\} \subseteq \{G(Y, C^q(I)) \geq \rho\}.$$

Finally, by exchangeability of the calibration scores $S(I_1, Y_1), \ldots, S(I_n, Y_n)$ and the test score $S(I, Y)$, the standard split-conformal quantile argument gives $\Pr(S(I,Y) \leq q) \geq 1 - \alpha$. Consequently, $\Pr(G(Y, C^q(I)) \geq \rho) \geq \Pr(S(I,Y) \leq q) \geq 1 - \alpha$, which proves the result. □

Theorem 1 does not introduce a new conformal proof technique; rather, it makes explicit sufficient structural conditions—an input-dependent nested path and a nondecreasing path score—under which an adaptive expansion policy can be calibrated. Existing constructions can be viewed as special cases. As an example, CPS is the special case $C_t(I) = \text{dilate}(\hat{Y}, t)$ with path score $\lambda_t = t$. Theorem 1 also accommodates paths that are not parameterized solely by an iteration count or a global confidence threshold. We next instantiate this general construction using an expansion policy that combines model confidence and spatial structure.

The CPAS algorithm. For each annotated image (I, Y) in the calibration set, we compute a CP nonconformity score using the following iterative process. We initialize the path with the predicted segmentation $C_0(I) = \hat{Y}$. At each iteration t, we select a subset of the outer boundary of $C_{t-1}(I)$ denoted by $\Delta_t(I)$, to form the next prediction set $C_t(I) = C_{t-1}(I) \cup \Delta_t(I)$. This process monotonically increases the size of the prediction set and continues until the prediction set achieves the desired coverage level ρ.

The outer boundary pixels at each step t, denoted by $B_t(I)$, are computed using morphological dilation, a fundamental operator in mathematical morphology (Gonzalez & Woods, 2017). Binary dilation on a discrete set (e.g., a binary image) is performed using a *structuring element* (SE), which defines pixel connectivity. Common choices include a 3×3 cross (4-connectivity) and a 3×3 square (8-connectivity). A pixel is added to the boundary set $B_t(I)$ if it has at least one neighboring pixel in $C_{t-1}(I)$.

After constructing the boundary set $B_t(I)$, we select a subset $\Delta_t(I)$. Our selection strategy balances exploitation, which expands into regions where the model is confident, with exploration, which allows expansion into low-confidence regions that may be necessary to achieve the desired coverage. In the confidence-based exploitation step, we choose the pixels $x \in B_t(I)$ whose object confidence $f(x)$ is the highest and thus the most likely to be part of the object. The geometric exploration score encourages selecting pixels that are spatially adjacent to regions incorporated during early iterations of the segmentation expansion procedure. For each $x \in B_t(I)$ we define

$$s(x) = \frac{1}{|\mathcal{N}(x)|} \sum_{h \in \mathcal{N}(x) \cap C_{t-1}(I)} \frac{1}{1 + d(h)}, \tag{4}$$

where $\mathcal{N}(x)$ denotes the neighbors of pixel x, and $d(h)$ is the iteration number at which h was added. Intuitively, $s(x)$ assigns higher values to pixels x that lie near regions added during the early stages of the iterative procedure.

At each iteration, we first select the top boundary pixels based on the confidence score $f(x)$ and then select additional boundary pixels with the highest values of $s(x)$. We add a fixed fraction $k \in (0, 1)$ of boundary pixels. A fraction $r \in (0, 1)$ of these are selected based on confidence (exploitation), whereas the remaining pixels are selected based on spatial adjacency (exploration). In other words, k controls the expansion rate and r controls the balance between confidence-based and spatially guided expansion.

The union of these two subsets, $\Delta_t(I)$, is then added to $C_{t-1}(I)$ to form $C_t(I)$. At the end of the t-th iteration, we set the iteration score λ_t as

$$\lambda_t = \lambda_{t-1} + \frac{\frac{1}{|\Delta_t(I)|} \sum_{x \in \Delta_t(I)} \big(1 - f(x)\big)}{|C_0(I)|}, \tag{5}$$

where $\lambda_0 = 0$. Intuitively, the score measures the average uncertainty of the pixels added during each iteration and normalizes it by the size of the predicted segmentation. Averaging over $|\Delta_t(I)|$ removes

Algorithm 1 Conformal Prediction Adaptive Segmentation (CPAS)

Input: Calibration set $\{(I_i, Y_i)\}_{i=1}^n$, test image I_{test}, model f, significance level α and coverage ratio ρ.
for each image (I_i, Y_i) in calibration set **do**
Start with $C_0(I_i) = \hat{Y}_i$ and iteratively expand $C(I_i)$ until the coverage rate of $|Y_i \cap C(I_i)|/|Y_i|$ is at least ρ.
Set $S(I_i)$ to be the score (5) at the last iteration.
end for
Sort scores: $S_{(1)} \leq S_{(2)} \leq \ldots \leq S_{(n)}$ and set $q \leftarrow S_{(\lceil (1-\alpha)(n+1) \rceil)}$.
Test image: Apply the same iterative expansion to I_{test} until $\lambda_t(I_{\text{test}}) > q$ and return the expansion result $C_t(I_{\text{test}})$.

dependence on the expansion rate, and normalization by $|C_0(I)|$ reduces sensitivity to the size of the initial predicted object and makes scores more comparable across images with different object sizes. The iteration process stops when the desired coverage ρ is reached. Let T denote the final iteration index. The CP score for image I is then defined as $S(I) = \lambda_T$. Finally, given the scores of n calibration images $S(I_1), \ldots, S(I_n)$, we apply the CP procedure by selecting the threshold q to be the $1 - \alpha$ quantile among these n scores.

For a new test image I, we run the iterative algorithm described above, continuing until the first index T such that $\lambda_T(I) > q$. Denote the prediction set at the end of the algorithm by $C^q(I)$. We call this algorithm Conformal Prediction Adaptive Segmentation (CPAS). A complete description of the method is provided in Algorithm 1.

Properties of the CPAS Expansion Path. The validity of CPAS relies on the fact that the expansion procedure defines a nested family of candidate prediction sets. For a given image I, the method constructs a sequence

$$C_0(I) \subseteq C_1(I) \subseteq \cdots \subseteq C_t(I) \subseteq \cdots,$$

where $C_0(I) = \hat{Y}$ is the original segmentation mask and each subsequent set is obtained by adding a subset of the current outer boundary. This monotonicity is essential: once a pixel is included in the prediction set, it is never removed. Therefore, increasing the conformal threshold can only enlarge the returned set and cannot reduce the achieved object coverage.

This nested structure distinguishes CPAS from general heuristic post-processing methods. A heuristic refinement of the predicted mask may improve visual quality, but it does not necessarily define an ordered sequence of prediction sets suitable for conformal calibration. In CPAS, the ordered sequence is constructed explicitly by the expansion process. The conformal calibration step then determines how far along this sequence one must proceed in order to achieve the required coverage level. As a result, the expansion policy can be adaptive and image-dependent, while the final prediction set still satisfies the standard finite-sample conformal guarantee.

This expansion rule also has a natural interpretation in terms of exploitation and exploration. The confidence-based component exploits the model's soft prediction map by selecting boundary pixels with high object probability. This encourages the conformal set to grow in directions that are supported by the model. The spatial component, on the other hand, prevents the method from relying only on the probability map, which may be poorly calibrated or locally overconfident. By favoring pixels adjacent to regions that were included early in the expansion process, the spatial score encourages coherent growth around the predicted object. The parameter r controls the relative importance of these two mechanisms, whereas k controls the speed of expansion.

This score differs from the iteration-count score used in purely morphological methods. An iteration-count score treats all expansion steps as equally costly, regardless of whether the added pixels are likely or unlikely according to the model. In contrast, CPAS assigns a lower cost to expansion through confident regions and a higher cost to expansion through uncertain regions. This allows the calibrated threshold to reflect not only how far the mask must expand, but also how uncertain the required expansion is. The ablation results support this design choice: removing either the confidence-weighted score or its normalization terms leads to larger prediction sets.

Extension to multi-class segmentation. While the core CPAS algorithm is formulated for binary targets, many clinical scenarios, such as multi-organ segmentation or cardiac chamber delineation, demand a multi-class framework. We next generalize CPAS into a *synchronized joint multi-class expansion process*. Let K denote the total number of semantic target classes. For a given multi-channel image I, the ground-truth annotations are represented as a collection of binary masks $\{Y_1, \ldots, Y_K\}$, and the model outputs a multi-channel soft probability map where $f_c(x) = \mathbb{P}(y(x) = c \mid I) \in [0, 1]$ for each pixel $x \in \Omega$ and class $c \in \{1, \ldots, K\}$.

For each class c, we initialize $C_{0,c}(I) = \{x \in \Omega : f_c(x) > \tau\}$. At each iteration, the candidate boundary pixel–class pairs (x, c) are pooled across all classes and selected using the same confidence-guided exploitation and spatial exploration criteria as in the binary case. The selected pairs define $\Delta_{t,c}(I)$, and we set

$$C_{t,c}(I) = C_{t-1,c}(I) \cup \Delta_{t,c}(I).$$

Define the augmented pixel–class space: $\widetilde{\Omega} = \Omega \times \{1, \ldots, K\}$, and represent the multi-class ground truth and the prediction set at iteration t as

$$\widetilde{Y} = \{(x, c) \in \widetilde{\Omega} : x \in Y_c\}, \qquad \widetilde{C}_t(I) = \{(x, c) \in \widetilde{\Omega} : x \in C_{t,c}(I)\}.$$

Since $C_{t-1,c}(I) \subseteq C_{t,c}(I)$ for every class c, the induced sets form a nested sequence

$$\widetilde{C}_0(I) \subseteq \widetilde{C}_1(I) \subseteq \cdots .$$

Moreover, the coverage functional of Theorem 1 becomes

$$G(\widetilde{Y}, \widetilde{C}_t(I)) = \frac{\left|\widetilde{Y} \cap \widetilde{C}_t(I)\right|}{|\widetilde{Y}|} = \frac{\sum_{c=1}^{K} |Y_c \cap C_{t,c}(I)|}{\sum_{c=1}^{K} |Y_c|}.$$

Under this augmented representation, the oracle stopping index in Eq. (3) becomes:

$$T_\rho(I, Y) = \min\{t : G(\widetilde{Y}, \widetilde{C}_t(I)) \geq \rho\},$$

which is precisely the desired aggregate multi-class coverage criterion and is therefore an instance of the oracle stopping rule in Theorem 1.

To mirror global risk calibration, the multi-class score accumulates the normalized uncertainty across all active channels simultaneously:

$$\lambda_t = \lambda_{t-1} + \frac{\frac{1}{|\Delta_t|} \sum_{c=1}^{K} \sum_{x \in \Delta_{t,c}(I)} (1 - f_c(x))}{\sum_{c=1}^{K} |C_{0,c}(I)|} \tag{6}$$

where $\lambda_0 = 0$ and $|\Delta_t| = \sum_{c=1}^{K} |\Delta_{t,c}(I)|$ represents the total number of pixel-class updates made during step t. The multi-class path score is nondecreasing because every score increment is nonnegative. Therefore, all the conditions of Theorem 1 are satisfied.

Let q denote the conformal quantile obtained from the calibration scores, and let $\widetilde{C}^{\,q}(I)$ denote the calibrated set in the augmented space. Equivalently, writing

$$C_c^q(I) = \{x \in \Omega : (x, c) \in \widetilde{C}^{\,q}(I)\},$$

the multi-class prediction sets satisfy

$$\Pr\left(\frac{\sum_{c=1}^{K} |Y_c \cap C_c^q(I)|}{\sum_{c=1}^{K} |Y_c|} \geq \rho\right) \geq 1 - \alpha.$$

Thus, the joint multi-class CPAS construction inherits the distribution-free finite-sample coverage guarantee of Theorem 1 directly. Note that this guarantee controls aggregate coverage across all semantic classes; it does not imply a separate $1 - \alpha$ guarantee for each class.

4 EXPERIMENTS

In this section, we describe the evaluation process used to assess the performance of the CPAS method.

Datasets. We evaluated the binary CPAS method on three public datasets: OASIS (neuroimaging dataset) (Marcus et al., 2007; Hoopes et al., 2022)[1], WBC (White Blood Cells) (Zheng et al., 2018)[2], and Kvasir (gastrointestinal polyps)(Jha et al., 2020). As in (Angelopoulos et al., 2024b; Mossina & Friedrich, 2025), we additionally assembled a Polyps dataset by combining five existing datasets: ETIS (Silva et al., 2013), CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Tajbakhsh et al., 2015), EndoScene (Vázquez et al., 2017), and Kvasir (Pogorelov et al., 2017). To validate the multi-class capabilities of the extended CPAS algorithm, we used the multi-class Cardiac Image Segmentation (M&Ms) Challenge dataset (Campello et al., 2021). This benchmark comprises multi-vendor cardiac MRI volumes spanning four distinct clinical domains: Siemens (Domain A, 192 volumes), Philips (Domain B, 252 volumes), General Electric (Domain C, 150 volumes), and Canon (Domain D, 100 volumes). The segmentation objective targets three coupled cardiac structures simultaneously: the Left Ventricle (LV), the Right Ventricle (RV), and the Myocardium (MYO).

Implementation details. Our binary image experiments implemented four different segmentation models, each applied to a different dataset. For the OASIS dataset, we used the pretrained UniverSeg model (Butoi et al., 2023)[3]. For the WBC dataset, we employed a U-Net architecture trained specifically for white blood cell segmentation. For the Kvasir dataset, we similarly employed a U-Net architecture for gastrointestinal polyp segmentation. Finally, for polyp segmentation, we used the PraNet model (Fan et al., 2020), which was trained on the Polyps dataset[4]. For the M&Ms dataset, a standard U-Net architecture was trained independently for each of the four distinct clinical domains. To perform conformalization, we randomly reshuffled the test split of each dataset and repeated the procedure 36 times, each time dividing the data evenly into calibration and test sets. All reported results are the mean (standard deviation) over the 36 splits. This resulted in calibration/test set sizes of 250, 35, 62, and 106 samples for the Polyps, WBC, OASIS, and Kvasir datasets, respectively. For the multi-domain M&Ms dataset, the calibration and test split sizes varied between 65 and 190 samples, depending on the specific scanner vendor domain. The hyperparameters $k = 0.15$ and $r = 0.6$ were selected once using the OASIS dataset via a small grid search and were then fixed across all other datasets and coverage settings. No per-dataset tuning was performed.

Computational Efficiency. We measured the average per-image runtime for computing the nonconformity score on 125 OASIS images using an Apple M4 CPU. CRC was fastest at approximately 0.11 ms per image, followed by CPS at 18.90 ms. CPAS required 124.85 ms due to its iterative confidence- and spatially-guided expansion. Despite this additional computation, CPAS remains practical, requiring only about 0.125 s per image while producing substantially more compact prediction sets.

Compared methods. We evaluated the performance of CPAS against two established conformal prediction baselines that provide rigorous finite-sample coverage guarantees, adapting them for both binary and multi-class configurations. The first is CPS (Mossina & Friedrich, 2025), a spatially driven framework that constructs prediction sets via calibrated morphological dilations; while it expands the predicted mask uniformly in the binary setting, we extend it to the multi-class task by decomposing the multi-class masks into independent binary channels and executing the morphological dilation steps channel-wise. The second is CRC (Angelopoulos et al., 2024b; Blot et al., 2025). While it filters a single probability map in the binary case, its multi-class extension applies a single global threshold across all class channels simultaneously. Other recent methods are not directly comparable because they target different guarantees or tasks, or require additional model components such as foundation-model features, MC dropout, or ensembles. To our knowledge, CPS is the only previous conformal segmentation method that progressively expands the boundary of the initial predicted mask.

[1] Source: https://sites.wustl.edu/oasisbrains, obtained via https://github.com/JJGO/UniverSeg.

[2] Distributed as open source at: https://github.com/zxaoyou/segmentation_WBC

[3] https://link.springer.com/article/10.1007/s43681-023-00400-7

[4] We reused the precomputed predictions and the training/test splits released by the authors of (Angelopoulos et al., 2024b). See https://github.com/aangelopoulos/conformal-prediction.

Table 1: Calibration results (Cov. and ϕ) on Polyps, WBC, OASIS, and Kvasir using cross SE.

Dataset			**CRC**		**CPS**		**CPAS**	
	ρ	$1-\alpha$	Cov.	ϕ	Cov.	ϕ	Cov.	ϕ
Polyps	0.9	0.9	0.90 (0.03)	1.20 (0.10)	0.90 (0.02)	1.21 (0.07)	0.90 (0.02)	**1.16 (0.06)**
	0.99	0.9	0.91 (0.02)	10.78 (3.52)	0.91 (0.02)	1.89 (0.28)	0.91 (0.02)	**1.71 (0.11)**
	0.9	0.95	0.96 (0.02)	3.68 (2.80)	0.96 (0.02)	3.26 (0.78)	0.95 (0.02)	**1.87 (0.25)**
	0.99	0.95	0.96 (0.02)	15.41 (0.76)	0.96 (0.02)	4.23 (0.88)	0.96 (0.02)	**2.91 (0.53)**
WBC	0.9	0.9	0.91 (0.06)	1.06 (0.03)	0.93 (0.04)	1.11 (0.04)	0.91 (0.06)	**1.05 (0.02)**
	0.99	0.9	0.91 (0.07)	1.37 (0.04)	0.93 (0.06)	1.60 (0.10)	0.91 (0.06)	**1.33 (0.03)**
	0.9	0.95	0.96 (0.05)	**1.11 (0.03)**	0.96 (0.04)	1.26 (0.09)	0.96 (0.04)	1.13 (0.03)
	0.99	0.95	0.98 (0.03)	1.67 (0.22)	0.97 (0.04)	2.08 (0.35)	0.97 (0.04)	**1.60 (0.22)**
OASIS	0.9	0.9	0.92 (0.05)	1.55 (0.01)	0.96 (0.02)	1.72 (0.02)	0.91 (0.05)	**1.53 (0.01)**
	0.99	0.9	1.00 (0.00)	2.06 (0.01)	0.95 (0.02)	2.18 (0.01)	0.93 (0.04)	**1.98 (0.01)**
	0.9	0.95	0.98 (0.02)	**1.57** (0.01)	0.98 (0.03)	1.79 (0.05)	0.97 (0.02)	**1.57 (0.02)**
	0.99	0.95	1.00 (0.00)	2.06 (0.01)	0.97 (0.02)	2.21 (0.01)	0.95 (0.03)	**2.00 (0.01)**
Kvasir	0.9	0.9	0.93 (0.04)	3.76 (0.62)	0.91 (0.04)	4.87 (0.49)	0.92 (0.04)	**3.38 (0.67)**
	0.99	0.9	0.98 (0.05)	8.44 (1.28)	0.92 (0.04)	7.00 (0.45)	0.92 (0.04)	**4.61 (0.68)**
	0.9	0.95	0.98 (0.02)	**5.29 (1.11)**	0.95 (0.03)	6.51 (0.87)	0.95 (0.03)	6.01 (1.27)
	0.99	0.95	1.00 (0.00)	8.76 (0.52)	0.96 (0.02)	8.44 (0.84)	0.96 (0.03)	**6.89 (1.03)**

Evaluation Metrics. To compare the methods, we used the following metrics computed on the test dataset:

$$\text{Empirical coverage: } \text{Cov}(\mathcal{C}, \rho) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\frac{|Y_i \cap \mathcal{C}(I_i)|}{|Y_i|} \geq \rho\right\}, \tag{7}$$

$$\text{Stretch: } \phi(\mathcal{C}) = \frac{1}{n}\sum_{i=1}^{n} \frac{|\mathcal{C}(I_i)|}{|\hat{Y}_i|}, \tag{8}$$

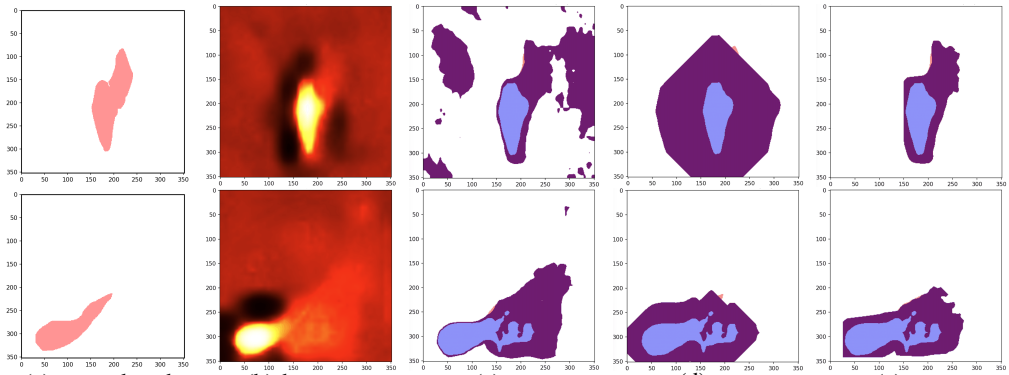
where n is the number of test images. Here, the empirical coverage serves as an estimator of the algorithm's coverage on a random test sample. The *stretch* metric serves as our primary efficiency benchmark and competitive focus. This metric quantifies, on average, how much larger the prediction sets are compared to the initial predicted labels $\hat{Y}$, where smaller values indicate superior, more compact prediction volume. In the appendix we also report calibration results using the Hausdorff boundary-distance primarily as a geometric sanity check to verify the geometric boundary behavior of the prediction boundaries.

For the multi-class setting, the coverage and stretch metrics are generalized globally by pooling the pixel counts across all K semantic channels for each test image. Specifically, the multi-class coverage condition checks if the aggregate intersection across all classes satisfies the fraction ρ, while the multi-class stretch quantifies the total volume of all generated prediction sets relative to the combined volume of the initial predicted masks.

Results. Table 1 compares coverage and stretch of the proposed CPAS method with CRC and CPS across the Polyps, WBC, OASIS, and Kvasir datasets. Consistent with their finite-sample conformal guarantees, all methods achieve empirical coverage close to the target levels. However, CPAS consistently produces more compact prediction sets, reflected by lower stretch values across nearly all datasets and confidence levels. For instance, on Polyps ($\rho = 0.9$, $1 - \alpha = 0.95$), CPAS attains coverage comparable to CRC and CPS while reducing stretch from 3.68 and 3.26 to 1.87. Similar behavior is observed on WBC, OASIS, and Kvasir, where CPAS maintains nominal coverage with smaller prediction volumes. Averaged over all datasets, CPAS achieves a mean stretch of 2.54, compared to 3.21 (CPS) and 4.36 (CRC), corresponding to average reductions of approximately 20% and 40%, respectively. These results demonstrate that the adaptive pixel selection and iterative refinement in CPAS yield tighter, well-calibrated segmentation sets without compromising statistical validity.

Table 2: Calibration results on multi-class images (Cov. and ϕ) on the M&Ms datasets using cross SE, averaged across all four domains.

Dataset			**CRC**		**CPS**		**CPAS**	
	ρ	$1-\alpha$	Cov.	ϕ	Cov.	ϕ	Cov.	ϕ
M&Ms	0.9	0.9	0.91 (0.04)	5.27 (2.61)	0.92 (0.04)	2.05 (0.48)	0.91 (0.04)	**1.74 (0.20)**
	0.99	0.9	0.92 (0.03)	67.19 (40.76)	0.93 (0.03)	3.90 (0.98)	0.92 (0.03)	**3.42 (0.40)**
	0.9	0.95	0.96 (0.03)	40.95 (34.34)	0.97 (0.02)	4.90 (1.49)	0.96 (0.03)	**2.73 (0.49)**
	0.99	0.95	0.96 (0.03)	168.19 (69.47)	0.97 (0.02)	8.44 (2.64)	0.97 (0.02)	**5.84 (1.28)**



(a) ground truth (b) heatmap (c) CRC (d) CPS (e) CPAS

Figure 1: Visual comparison from the Polyps dataset. (a) Ground truth. (b) Softmax output of the model f. (c-e) calibration results, blue—segmentation, red- ground truth, purple—pixels added during calibration (coverage = 0.99).

Fig. 1 provides a visual illustration of why the proposed CPAS method outperforms CPS and CRC. CRC (Fig. 1(c)) expands the prediction set solely based on confidence thresholds, resulting in spatially fragmented additions that do not respect object geometry. In contrast, CPS (Fig. 1(d)) applies isotropic morphological dilation, resulting in uniform expansion into low-confidence regions, regardless of the model's belief. CPAS (Fig. 1(e)) selectively expands along high-confidence boundaries while maintaining spatial coherence through adjacency-based exploration. As a result, CPAS achieves a substantially more compact prediction set for the same coverage level, directly explaining the reduced stretch observed in Table 1.

We next show multi-class results. Table 2 reports results on the M&Ms cardiac segmentation dataset, averaged across four scanner domains. All methods attain the target coverage, but CPAS consistently produces the most compact prediction sets. The advantage is particularly pronounced at $\rho = 0.99$ and $1 - \alpha = 0.95$, where CPAS reduces stretch from 8.44 for CPS to 5.84. CRC becomes highly inefficient because its global threshold introduces extensive overlap between class-wise prediction sets. These results demonstrate that CPAS remains effective beyond binary segmentation.

5 Conclusions

We introduced a general view of segmentation uncertainty as conformal calibration over an input-dependent nested prediction path, and instantiated this principle in CPAS using confidence-guided spatial expansion. By combining model confidence with spatially coherent expansion, CPAS bridges the gap between confidence-based conformal segmentation and purely morphological conformal dilation. Experiments on multiple medical imaging datasets show that CPAS preserves the desired coverage while substantially reducing prediction-set size. These results suggest that confidence-guided spatial expansion is a promising direction for reliable and interpretable segmentation uncertainty quantification. Future work will extend CPAS to distribution-shift settings and learned shape-aware expansion operators.

AI USE STATEMENT

Generative AI tools were used in the preparation of this manuscript solely for language editing: identifying typographical errors and improving grammatical and stylistic clarity of the text. AI tools were not used to generate research ideas, formulate hypotheses, design the methodology or experiments, perform data analysis, interpret results, or produce any part of the code or numerical results reported in this paper. All scientific content, claims, and conclusions are the authors' own.

REFERENCES

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024a.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024b.

Jacqueline Isabel Bereska, Hamed Karimi, and Reza Samavi. SACP: Spatially-adaptive conformal prediction in uncertainty quantification of medical image segmentation. In *Medical Imaging with Deep Learning (MIDL)*, 2025.

Jorge Bernal, F. Javier Sanchez, Genis Fernandez-Esparrach, Debora Gil, Cristina Rodriguez, and Francesc Vilarino. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

Vincent Blot, Anastasios Nikolas Angelopoulos, Michael Jordan, and Nicolas J-B. Brunel. Automatically adaptive conformal risk control. In *The International Conference on Artificial Intelligence and Statistics*, 2025.

Joren Brunekreef, Eric Marcus, Ray Sheombarsing, Jan-Jakob Sonke, and Jonas Teuwen. Kandinsky conformal prediction: Efficient calibration of image segmentation algorithms. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. UniverSeg: Universal medical image segmentation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

V. M. Campello et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. *IEEE Trans. on Medical Imaging*, 40(12):3543–3554, 2021.

Samuel Davenport. Conformal confidence sets for biomedical image segmentation. In *International Conference on Learning Representations (ICLR)*, 2025.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

Jamil Fayyad, Shadi Alijani, and Homayoun Najjaran. Empirical validation of conformal prediction for trustworthy skin lesions classification. *Computer Methods and Programs in Biomedicine*, 253: 108231, 2024. doi: 10.1016/j.cmpb.2024.108231.

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

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Cooper Gamble, Shahriar Faghani, and Bradley J. Erickson. Applying conformal prediction to a deep learning model for intracranial hemorrhage detection to improve trustworthiness. *Radiology: Artificial Intelligence*, 7(2):e240032, 2025.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4 edition, March 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Alexander Hoopes, Michael Hoffmann, Douglas N. Greve, Bruce Fischl, John Guttag, and Adrian V. Dalca. Learning the effect of registration hyperparameters with hypermorph. *Journal of Machine Learning for Biomedical Imaging*, 1:1–30, 2022.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pp. 451–462. Springer, 2020.

Jef Jonkers, Frank Coopman, Luc Duchateau, Glenn Van Wallendael, and Sofie Van Hoecke. Reliable uncertainty quantification for 2d/3d anatomical landmark localization using multi-output conformal prediction. *Medical Image Analysis*, 110:103953, 2026.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Daniel S. Marcus, Tracy H. Wang, Joel Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9): 1498–1507, 2007.

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020.

Luca Mossina and Corentin Friedrich. Conformal prediction for image segmentation using morphological prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun L. Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter T. Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer-aided gastrointestinal disease detection. In *Proceedings of the ACM on Multimedia Systems Conference (MMSys)*, 2017.

Jaime Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2013.

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

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

Bruno Viti, Elias Karabelas, and Martin Holler. CONSIGN: Conformal segmentation informed by spatial groupings via decomposition. In *International Conference on Learning Representations (ICLR)*, 2026.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*, volume 29. Springer, 2005.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Xue Zheng, Yanan Wang, Guohua Wang, and Jie Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018.

A Appendix

A.1 Ablation Study

CPAS is governed by two main hyperparameters: k, which controls the proportion of candidate pixels added to the prediction set at each iteration, and r, which defines the fraction of pixels selected based on f confidence scores. We evaluated k using values of 0.05, 0.15, and 0.25, and r using values of 0.3, 0.6, and 0.9. Our results indicated that setting $k = 0.15$ and $r = 0.6$ provided a good trade-off between exploration and exploitation. Notably, performance remained relatively stable across the examined ranges of k and r, suggesting that CPAS was not highly sensitive to these hyperparameters. This robustness suggests that CPAS can be deployed without extensive hyperparameter tuning, which is important in clinical workflows. The results on the OASIS dataset are summarized in Table 4.

CPAS incorporates model confidence both in pixel selection and in the nonconformity score. To assess the contribution of each component, we considered several variants: (i) using only the iteration count as the score, $S(I_i) = t_i$, as in CPS (CPAS$_{\text{iter}}$); (ii) removing normalization by $|\Delta_t|$ (CPAS$_{-\Delta}$); and (iii) removing normalization by $|C_0(I)|$ (CPAS$_{-C_0}$). We found that the full score formulation consistently yielded more compact prediction sets. Table 3 reports results on the Kvasir dataset.

Table 3: Comparison between the standard CPAS method and its variants using alternative nonconformity score formulations.

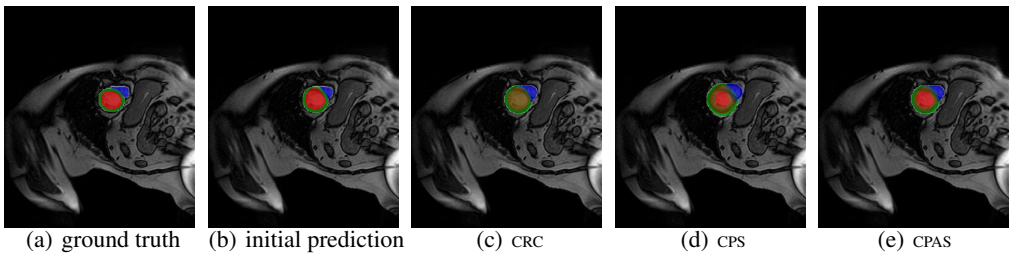
ρ	$1-\alpha$	CPAS$_{\text{iter}}$	CPAS$_{-\Delta}$	CPAS$_{-C_0}$	CPAS
0.9	0.9	3.46 (0.44)	4.41 (0.55)	3.38 (0.44)	3.38 (0.67)
0.99	0.9	5.20 (0.79)	5.50 (0.71)	5.13 (0.75)	4.61 (0.68)
0.9	0.95	5.62 (0.92)	6.70 (1.04)	5.56 (0.90)	6.01 (1.27)
0.99	0.95	7.32 (0.72)	7.46 (1.16)	7.17 (0.64)	6.89 (1.03)
Avg		5.40	6.02	5.31	**5.22**

Table 4: Calibration results on OASIS dataset for different k and r values.

		k Values			r Values		
ρ	$1-\alpha$	0.05	0.25	0.15 (CPAS)	0.3	0.9	0.6 (CPAS)
0.9	0.9	**1.52 (0.01)**	1.55 (0.01)	1.53 (0.01)	1.59 (0.01)	1.53 (0.01)	**1.53 (0.01)**
0.99	0.9	1.99 (0.01)	2.00 (0.01)	**1.98 (0.01)**	**1.95 (0.01)**	2.15 (0.01)	1.98 (0.01)
0.9	0.95	**1.55 (0.01)**	1.59 (0.01)	1.57 (0.02)	1.62 (0.01)	1.57 (0.01)	**1.57 (0.02)**
0.99	0.95	2.02 (0.01)	2.02 (0.01)	**2.00 (0.01)**	**1.98 (0.02)**	2.18 (0.01)	2.00 (0.01)
Avg		**1.77**	1.79	**1.77**	1.79	1.86	**1.77**

Table 5: Calibration results (d_H) on WBC and OASIS using cross SE.

Dataset	ρ	$1-\alpha$	CRC	CPS	CPAS
WBC	0.9	0.9	9.33 (1.35)	9.26 (1.37)	**8.43 (1.10)**
	0.99	0.9	21.72 (4.35)	18.85 (3.15)	**13.32 (3.34)**
	0.9	0.95	12.63 (2.16)	16.44 (4.81)	**11.60 (1.65)**
	0.99	0.95	31.21 (4.89)	29.22 (6.01)	**26.43 (8.21)**
OASIS	0.9	0.9	**13.53 (0.14)**	16.70 (0.10)	14.29 (0.38)
	0.99	0.9	17.77 (0.32)	26.74 (0.26)	**16.97 (0.36)**
	0.9	0.95	13.54 (0.14)	16.57 (0.13)	**13.30 (0.50)**
	0.99	0.95	17.83 (0.27)	29.07 (1.61)	**17.63 (0.72)**



(a) ground truth (b) initial prediction (c) CRC (d) CPS (e) CPAS

Figure 2: Visual comparison of multi-class calibration results on the M&Ms dataset at a target coverage level of $\rho = 0.99$. (a) Ground truth annotations. (b) Initial prediction ($\hat{Y}$). (c–e) Conformal prediction sets generated by CRC, CPS, and our proposed CPAS method, where the distinct colors represent the different semantic labels (LV, RV, and MYO) included in the finalized prediction region.

A.2 Calibration results measured by the shape.

Table 5 presents the boundary evaluation based on Hausdorff distance:

$$d_H(A, B) = \max(\max_{a \in A} \min_{b \in B} d(a, b), \max_{b \in B} \min_{a \in A} d(b, a))$$

between the true segmentation and the CP prediction set across the WBC and OASIS datasets. The empirical results demonstrate that while CPAS is heavily optimized for volume efficiency (stretch), it consistently achieves highly comparable geometric performance to both CPS and CRC. Crucially, CPAS demonstrates structural stability and never produces the worst-case performance under any calibration configuration.

A.3 Multi-class Segmentation results

Table 2 shows the multi-class evaluation results on the M&Ms dataset, with all performance metrics averaged across the four distinct vendor domains. While all three methods achieve the required coverage, the structural weaknesses of CRC and CPS become highly visible in this setting. For CRC, if the model confuses two classes on even a small group of pixels, the global threshold must change drastically to cover them. This forces the method to lower its bar everywhere, sweeping in many unnecessary pixels across the image and blowing up the stretch to 168.19. For CPS, because the class channels expand completely independently, classes that do not need much expansion are heavily over-expanded and penalized. This happens because they are forced to use a large expansion size dictated by a completely different class in the same image. In contrast, CPAS avoids these pitfalls by selectively growing boundaries using both model confidence and local structure, maintaining a highly compact stretch of only 5.84.

These quantitative trends are visually evident in Fig. 2. For CRC, the independent thresholding results in massive label overlaps, where many pixels are assigned more than one semantic label in the final prediction set (Fig. 2c). For CPS, because every class channel is expanded by the exact same size, structures that do not need much growth are forced to become much larger than necessary (Fig. 2d). In contrast, CPAS avoids both pitfalls, keeping the prediction sets highly compact and anatomically realistic (Fig. 2e).