RISK-CONTROLLED SAFE STOPPING FOR SEQUENTIAL DATA ACQUISITION

Anonymous authors
Paper under double-blind review

ABSTRACT

In clinical diagnosis and tomographic imaging, data arrive one measurement at a time and each measurement costs time, money, or patient burden. Deciding *when* enough has been acquired is as consequential as deciding what to acquire, yet existing acquisition policies and heuristic stopping rules offer no guarantee on the risk of the decision made when acquisition halts. We introduce risk-controlled stopping (RCS), a post-hoc, distribution-free procedure that selects a stopping rule so that a user-specified risk, such as the false negative or false positive rate, is controlled among the samples on which the system commits to a decision. RCS treats each candidate rule as a trajectory-level selective predictor and certifies it with the Learn-Then-Test framework, giving a finite-sample guarantee on the risk *conditional on stopping*. Importantly, RCS can be applied over any acquisition pipeline without retraining, whether it is fixed, random, or learned. Because the controlled risk is absolute rather than relative to full acquisition, no rule is safe by construction; when none meets the target, RCS certifies that fact and defers to full acquisition instead of returning an unsafe rule. On a synthetic CIFAR-10 benchmark, on FASTMRI knee k-space acquisition, and on an internal prostate MRI cohort, RCS holds the target risk at every acquisition rate while stopping a substantial fraction of samples early, and it certifies a learned reinforcement-learning acquisition policy as a post-hoc wrapper.

1 INTRODUCTION

Many high-stakes decisions rest on data that arrive one measurement at a time. A physician orders tests in sequence and uses signals from all the test results thus far to decide whether the next test is needed; an MRI scanner fills in the k-space matrix line by line; a diffusion-weighted sequence in MR acquisition is repeated to raise its signal-to-noise ratio. In each case, every additional measurement costs time, money, or patient burden (Lustig et al., 2008; Zbontar et al., 2018), so a sequential decision system faces two coupled questions: 1) *what* to acquire next? and 2) *when* enough has been acquired to stop safely?

The first question has received most of the attention. Dynamic feature acquisition methods learn policies that select informative measurements under a cost budget (Bakker et al., 2020; Singhal et al., 2023; Covert et al., 2023; Yen et al., 2024). The second question is usually settled by a fixed budget, a confidence heuristic, or a learned policy tuned for average accuracy (Shim et al., 2018; Bakker et al., 2020). None of these approaches offer a guarantee on the risk of the decision made when acquisition halts. This is consequential in safety-critical settings like healthcare, where stopping too early could produce unsafe decisions, and stopping too late wastes costs and increases patient burden.

We propose risk-controlled stopping (RCS), a post-hoc, distribution-free procedure for deciding when to stop acquisition so that a user-specified risk, such as misclassification, the false negative rate (FNR), or the false positive rate (FPR), stays below a chosen level. Two properties of the controlled risk set RCS apart. First, it is an *absolute* risk, measured against the ground-truth label rather than against the prediction the model would have made with all the data. Second, it is evaluated *conditional on stopping*: only the samples on which the policy commits to a decision count. The second property is important, because a policy that rarely stops but is wrong on almost every sample

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.

where it stops can still satisfy a marginal risk bound over the full population. Whereas a conditional risk bound rules out this behavior.

The statistical engine behind RCS is the Learn-Then-Test (LTT) framework (Angelopoulos et al., 2025), which certifies a family of parameterized predictors on a held-out calibration set with finite-sample, distribution-free guarantees. Our key observation is that a stopping rule is a form of selective predictor: it decides both *whether* and *when* to make a prediction, and thus on which partial observation the prediction is made. Treating each rule as a trajectory-level policy lets LTT certify the risk conditional on stopping directly. This guarantee comes from LTT and we do not claim a new bound or proof technique. Our contribution is the problem formulation, the definition of the risk being controlled, and the procedure needed to control it. An absolute risk admits no configuration that is safe by construction, so the set of certifiable rules can be non-monotone or empty. RCS therefore calibrates with a step-down procedure rather than fixed-sequence testing, and when no rule meets the target it certifies that fact and defers to full acquisition instead of returning an unsafe rule.

We evaluate RCS on one synthetic and two clinical tasks. The synthetic task, sequential patch acquisition on CIFAR-10 (Krizhevsky et al., 2009), validates the procedure under a controlled acquisition model. We draw no clinical conclusions from it. The clinical tasks use raw scanner measurements: slice-level abnormality detection from FASTMRI knee k-space (Zbontar et al., 2018), and detection of clinically significant prostate cancer from diffusion-weighted averages in an internal prostate MRI cohort. Prior work has shown that a well-chosen subset of measurements can support reliable detection at a fixed undersampling factor (Singhal et al., 2023; Yen et al., 2024); it does not say when, for a given patient, acquisition can stop. RCS answers that question with a certificate. At every acquisition rate we test, the class-conditional FNR among stopped scans stays below the prescribed level, and the same procedure wraps a learned acquisition policy (Shim et al., 2018) without retraining.

Contributions

- We propose *risk-controlled stopping*: choosing when to halt sequential acquisition so that an absolute, user-specified risk against the ground-truth label (for example, FNR) is controlled on the samples the policy commits to. To our knowledge, no prior work casts acquisition stopping as a distribution-free, finite-sample risk-control problem.
- We show that treating each stopping rule as a trajectory-level selective policy makes this object certifiable by LTT (Corollary 1 in Section 2.1), and that the absolute-risk setting forces two design choices absent from relative-risk calibration: Holm's step-down in place of fixed-sequence testing, and certified infeasibility in place of a guaranteed fallback.
- We instantiate the framework with class-conditional FNR/FPR and demonstrate certified control on scanner-ordered MRI pipelines, on a synthetic CIFAR-10 benchmark, and as a post-hoc wrapper around a learned reinforcement-learning acquisition policy.

2 Background and Related Work

Dynamic feature acquisition. Acquiring features sequentially at test time has been studied extensively. Early methods select greedily by a utility score (Melville et al., 2005), these methods fall into three families distinguished by how they decide to stop. *Fixed budgets and information gain.* EDDI (Ma et al., 2018) selects the feature with the largest expected information gain and reports curves over budgets. Adapted to dynamic stopping, such methods halt when the largest expected gain falls below a cost threshold τ. The threshold trades information against cost, it bears no direct relation to downstream risk such as an error rate below 5%. *Policy-based stopping.* Shim et al. (2018) and Ghosh & Lan (2023) cast acquisition as a Markov decision process with "stop" as an action and maximize a reward $R = \mathbb{I}(y_{\text{pred}} = y_{\text{true}}) - \lambda\,\mathbf{c}$. Optimizing an average reward encourages the agent to stop early on hard, expensive instances whenever the cost of more information exceeds the expected loss from an error, which is the opposite of a per-decision safety guarantee.

Early classification and adaptive halting. A parallel literature on time series learns when to halt, and sometimes what to measure next, by optimizing an average objective. Chang et al. (2019) schedule measurements with deep reinforcement learning; EARLIEST (Hartvigsen et al., 2019) and the Recurrent Halting Chain (Hartvigsen et al., 2020) train halting networks that trade earliness against accuracy. These are strong halting backbones, but the halting point is chosen to be good on average, not certified to control a specified risk on the samples where the model actually stops. RCS

complements these approaches: as a post-hoc wrapper, it can use the stopping scores of any such policy to certify risk conditional on stopping, as we show for JAFA in Section 5.4.

Conformal prediction and risk control. Conformal prediction constructs prediction sets with finite-sample coverage under exchangeability (Shafer & Vovk, 2008; Angelopoulos & Bates, 2022), and conformal risk control extends the guarantee from coverage to any bounded risk (Angelopoulos et al., 2022). Both certify a *fixed* rule on a *fixed* input. Sequential acquisition violates both premises: the prediction is formed from a partial observation, and the time at which it is formed is chosen from the data. Gauthier et al. (2025) bring conformal ideas to sequential settings, and Jazbec et al. (2024) calibrate early exits in neural networks to save inference compute. The latter bounds a marginal risk over all inputs and economizes internal computation, we reduce external physical acquisition and bound the risk on the stopped subset.

Closest work. Ringel et al. (2024) also calibrate a stopping rule with LTT, for early time-series classification. The main difference is the risk being controlled. They bound the *accuracy gap*, the rate at which stopping early changes the prediction relative to using the full series. This measures agreement with the full-data model, rather than correctness. For example, a cancer case that the model misses at every step has an accuracy gap of zero and is still a false negative. RCS instead bounds the error itself, which changes the calibration problem. The accuracy gap is zero at full acquisition, giving their family a configuration that is safe by construction. An absolute risk has no such guaranteed safe configuration.

2.1 The Learn-Then-Test framework

Learn-Then-Test (LTT) (Angelopoulos et al., 2025) calibrates a pre-trained model post hoc. Given a family of predictors $f(\mathbf{x}; \lambda)$ indexed by a finite grid $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$, a bounded loss $\ell(y, \hat{y}) \in [0, B]$, a target level $\alpha \in (0, 1)$, and a confidence level $\delta \in (0, 1)$, LTT selects $\hat{\lambda} \in \Lambda$ such that

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left[\mathop{\mathbb{E}}_{\mathbf{x},y\sim\mathcal{P}^{\text{data}}}[\ell(y, f(\mathbf{x}; \hat{\lambda}))] \leq \alpha\right] \geq 1 - \delta. \tag{1}$$

It does so by treating calibration as multiple hypothesis testing on a held-out set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}^i, y^i)\}_{i=1}^n$. For each λ, the null hypothesis $H_{0,\lambda} : R(\lambda) > \alpha$ states that λ is unsafe. A concentration inequality for bounded variables (Hoeffding or empirical Bernstein) converts the empirical risk $\widehat{R}(\lambda) = \frac{1}{n}\sum_i \ell(y^i, f(\mathbf{x}^i; \lambda))$ into a valid p-value $p(\lambda)$. A family-wise error rate procedure at level δ, such as Holm's step-down (Holm, 1979), then rejects a subset of the nulls; the corresponding parameters form the safe set $\widehat{\Lambda}_{\text{safe}}$, and every member of it satisfies Equation (1). The user is free to deploy the member that maximizes an application-specific utility $U(\lambda)$.

3 Problem Setup

An input $\mathbf{x} = [x_1, \ldots, x_d]$ consists of d measurements that are acquired over time. The partial observation after t steps is $\mathbf{x}_t = \mathbf{x} \circ \mathbf{s}_t$, where $\mathbf{s}_t \in \{0, 1\}^d$ is a mask marking the components acquired so far and $\circ$ is element-wise multiplication. The acquisition policy that produces $\mathbf{s}_t$ may be fixed in advance, random, or adaptive (a function of $\mathbf{x}_{t-1}$); in every case the observations are nested. Full trajectories $(\mathbf{x}^i, y^i)$ are drawn i.i.d. from an unknown distribution $\mathcal{P}^{\text{data}}$, and we hold out a calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}^i, y^i)\}_{i=1}^n$. A pre-trained predictor p_θ returns a prediction $\hat{y}(\mathbf{x}_t)$ from any partial observation. We seek a data-dependent *stopping time* $\tau(\mathbf{x}) \in \{1, \ldots, d\} \cup \{\infty\}$ at which acquisition halts and the system commits to $\hat{y}(\mathbf{x}_\tau)$; $\tau(\mathbf{x}) = \infty$ means the policy defers to full acquisition.

Why bounding the marginal risk is not sufficient. It is tempting to bound the marginal risk of the stopped output, $\mathbb{P}(E_\tau)$, where E_τ is the event that the policy stopped and made an error. Let $S = \{\tau(\mathbf{x}) < \infty\}$ be the event that the policy stops and commits to a prediction. Since, deferred samples do not count as errors from stopping, we have $E_\tau \subseteq S$ and

$$\mathbb{P}(E_\tau) = \mathbb{P}(E_\tau \mid S)\,\mathbb{P}(S). \tag{2}$$

A marginal bound $\mathbb{P}(E_\tau) \leq \alpha$ therefore says nothing about the error among committed decisions, $\mathbb{P}(E_\tau \mid S)$, unless $\mathbb{P}(S)$ is close to one. A policy that commits on 1% of samples and is wrong on every one of them satisfies $\mathbb{P}(E_\tau) = 0.01 \leq \alpha$ while being useless. What a clinician needs is the reliability of the decisions that are actually made. RCS therefore controls the risk *conditional on stopping*.

4 RISK-CONTROLLED SAFE STOPPING VIA POLICY-LEVEL CERTIFICATION

We first define the risk conditional on stopping and show how to test each candidate policy using the calibration samples on which it stops. We then combine these tests to certify a set of policies and select one based on a user-chosen utility, defaulting to full acquisition if none can be certified. Finally, we extend the procedure to class-conditional false negative and false positive rates and describe the stopping rules used in our experiments.

4.1 CONDITIONAL RISK OF A STOPPING POLICY

Consider a family of stopping rules $\mathcal{D}_{\text{STOP}}(\mathbf{x}_t; \lambda) \in \{0, 1\}$ indexed by a finite grid $\lambda \in \Lambda = \{\lambda_1, \ldots, \lambda_m\}$. Each rule induces a stopping time, the first step at which it fires,

$$\tau_\lambda(\mathbf{x}) = \inf\{t \in \{1, \ldots, d\} \mid \mathcal{D}_{\text{STOP}}(\mathbf{x}_t; \lambda) = 1\}, \tag{3}$$

with $\tau_\lambda(\mathbf{x}) = \infty$ if the rule never fires, in which case the policy defers to full acquisition. When it fires at $t \leq d$, the system commits to $\hat{y}(\mathbf{x}_t)$. For a bounded loss $\ell(y, \hat{y}) \in [0, B]$, the quantity we control is the expected loss given that the policy stops,

$$R_{\text{cond}}(\lambda) = \mathop{\mathbb{E}}_{\mathbf{x}, y \sim \mathcal{P}^{\text{data}}} \left[\ell\left(y, \hat{y}(\mathbf{x}_{\tau_\lambda(\mathbf{x})})\right) \;\middle|\; \tau_\lambda(\mathbf{x}) < \infty \right]. \tag{4}$$

Our goal is to select a $\hat{\lambda} \in \Lambda$ with $R_{\text{cond}}(\hat{\lambda}) \leq \alpha$ with probability at least $1 - \delta$ over the draw of $\mathcal{D}_{\text{cal}}$. Changing λ changes not only *which* samples the policy stops on, but also changes *when* it stops. It therefore also changes the partial observation each prediction is made. The map $\lambda \mapsto R_{\text{cond}}(\lambda)$ can therefore jump and need not be monotone.

Testing each policy on the samples it stops. For each λ, let $\mathcal{S}(\lambda) = \{i \leq n \mid \tau_\lambda(\mathbf{x}_i) < \infty\}$ be the indices of the calibration samples where the policy stops, and let $n(\lambda) = |\mathcal{S}(\lambda)|$. The empirical conditional risk is the mean loss over this subset,

$$\widehat{R}_{\text{cond}}(\lambda) = \frac{1}{n(\lambda)} \sum_{i \in \mathcal{S}(\lambda)} \ell\left(y^i, \hat{y}(\mathbf{x}^i_{\tau_\lambda(\mathbf{x}^i)})\right), \tag{5}$$

and we test $H_{0,\lambda} : R_{\text{cond}}(\lambda) > \alpha$. The set $\mathcal{S}(\lambda)$ is selected by the very rule under test. Because the calibration trajectories are i.i.d. and membership in $\mathcal{S}(\lambda)$ depends only on the trajectory itself, the stopped samples are conditional on $n(\lambda)$ draws from the distribution conditioned on stopping. Hoeffding's inequality then yields a p-value that is valid conditional on $n(\lambda)$,

$$p(\lambda) = \begin{cases} 1, & \text{if } n(\lambda) = 0 \text{ or } \widehat{R}_{\text{cond}}(\lambda) \geq \alpha, \\ \exp\left(-2n(\lambda)\left(\alpha - \widehat{R}_{\text{cond}}(\lambda)\right)^2 / B^2\right), & \text{otherwise.} \end{cases} \tag{6}$$

This conditional validity also makes the p-value valid over the random draw of the calibration set.

Multiplicity correction and selection. Testing m hypotheses at once calls for family-wise error control. We apply Holm's step-down procedure (Holm, 1979): sort the p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, reject $H_{(j)}$ as long as $p_{(j)} \leq \delta/(m - j + 1)$, and stop at the first failure. The rejected hypotheses form the safe set $\widehat{\Lambda}_{\text{safe}}$, and every λ in it controls the conditional risk. We deploy the safe rule that maximizes a user-chosen utility $U(\lambda)$, typically the fraction of samples stopped or the negative mean stopping time. If $\widehat{\Lambda}_{\text{safe}}$ is empty, RCS reports that no rule in the family meets the target and every sample is fully acquired.

Corollary 1 (Conditional risk control; corollary of LTT)**.** *Let $\mathcal{D}_{cal}$ consist of n i.i.d. sample trajectories and let $\hat{\lambda}$ be selected by the procedure above. Then, provided $\widehat{\Lambda}_{safe} \neq \emptyset$,*

$$\mathbb{P}_{\mathcal{D}_{cal}}\left[\mathop{\mathbb{E}}_{\mathbf{x}, y}\left[\ell\left(y, \hat{y}(\mathbf{x}_{\tau_{\hat{\lambda}}(\mathbf{x})})\right) \;\middle|\; \tau_{\hat{\lambda}}(\mathbf{x}) < \infty \right] \leq \alpha \right] \geq 1 - \delta. \tag{7}$$

The result is a direct application of LTT (Angelopoulos et al., 2025): each λ is a selective predictor on trajectories, and LTT already certifies a selective risk over a data-dependent selected subset. The argument requires only that the selection event be fixed before calibration, so it also applies to any other pre-specified event. Pseudocode is given in Algorithm 1 in the appendix.

4.2 CLASS-CONDITIONAL FNR AND FPR

Equation (4) covers mis-classification directly. Clinical practice asks instead for class-conditional rates: the false negative rate among truly positive cases and the false positive rate among truly negative ones, rather than a pooled or prevalence-weighted error. For a task with labels $y \in \{0,1\}^K$ (binary when $K = 1$) we certify, for each label k,

$$\mathrm{FNR}_k(\lambda) = \mathbb{P}(\hat{y}_k(\mathbf{x}_{\tau_\lambda}) = 0 \mid y_k = 1,\ \tau_\lambda(\mathbf{x}) < \infty) \le \alpha_k, \tag{8}$$

and FPR_k symmetrically. This is the conditional risk of Equation (4) under the refined selection event $\{y_k = 1,\ \tau_\lambda < \infty\}$: the effective calibration set is the stopped positives $\mathcal{S}_k^+(\lambda) = \{i : \tau_\lambda(\mathbf{x}^i) < \infty,\ y_k^i = 1\}$ and the loss is $\mathbf{1}\{\hat{y}_k^i = 0\}$. Because the loss is binary we replace Hoeffding with the exact binomial tail, which is tighter, and we combine labels by $p(\lambda) = \max_k p_k(\lambda)$, which is valid under arbitrary dependence between labels and accommodates label-specific targets α_k. The policy family, the stopping rule, and the Holm correction are unchanged.

Policy families. RCS places no restriction on the form of the stopping rule beyond the family being fixed before calibration. We use three. A *confidence threshold* on the softmax score (CIFAR-10); the JAFA *Q-margin*, the gap between the Q-value of "stop" and that of the best acquisition action (Section 5.4); and *ambiguity bands* for the MRI tasks, where $\lambda \in [0,1]$ scales a per-label uncertainty band from never stopping ($\lambda = 0$) to an aggressive data-driven band ($\lambda = 1$), and the rule fires once every label's score has left its band (Section A.3).

5 EXPERIMENTS

We evaluate the proposed stopping-calibration procedure in two complementary settings. In Section 5.3, we consider fixed acquisition trajectories across vision and medical-imaging benchmarks to assess whether the method controls the target risk while reducing the number of measurements acquired. In Section 5.4, we extend the evaluation to learned adaptive acquisition policies by applying our method as a post-hoc wrapper around JAFA (Shim et al., 2018), a reinforcement-learning-based acquisition model. We assess whether the wrapped system controls the target risk while retaining the efficiency of the underlying adaptive policy.

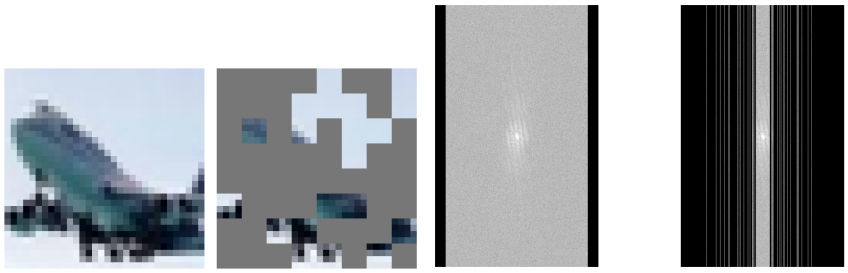

Figure 1: **Examples of sequential data acquisition in synthetic and MR settings. Left:** in the synthetic regime, the fully observed image $\mathbf{x}$ is partitioned into non-overlapping patches, each treated as a feature to be acquired sequentially; the right panel within that image shows the partially observed image $\mathbf{x}_t$ after acquisition of t patches. **Right:** in the MR setting, the left panel shows a fully sampled *k*-space, and the right panel shows the same *k*-space at an intermediate acquisition step.

5.1 DATASETS

We use the following datasets that span both synthetic and real-world acquisition settings:

1. **CIFAR-10** is a widely used benchmark dataset for image classification, consisting of $60,000$ images of size $32 \times 32 \times 3$ pixels, divided into 10 object classes, with each class equally represented (Krizhevsky et al., 2009).
2. **FastMRI Knee** dataset (Zbontar et al., 2018) provides open-source fully-sampled *k-space* data associated with the knee MR scans. It consists of a total of 41679 2D slices, refer to Table 1 in appendices, for the full data statistics. The diagnostic task involves detecting two musculoskeletal abnormalities: ACL tear and meniscus tear, a standard set of pathologies diagnosed in clinical tasks (Singhal et al., 2023; Yen et al., 2024). Since these labels are not mutually exclusive, we treat this as a multi-label classification problem.



Figure 2: **Control of misclassification rates on CIFAR-10. Left:** The cumulative fraction of samples halted at each sampling stage. **Right:** Predictive accuracy of stopped samples as a function of the number of acquired patches. We observe that RCS successfully maintains the misclassification rate below the specified risk level ($\alpha = 0.1$), whereas heuristic baselines consistently violate the safety threshold.

3. **Clinical Prostate MRI (Internal dataset)** is a large de-identified prostate MRI cohort from an urban health system. The task is binary PI-RADS-positive classification, defined as radiologist-assigned PI-RADS ≥ 3, an MRI-based marker of suspicion for clinically significant prostate cancer (Turkbey et al., 2019). Each scan contains an axial T2-weighted volume and diffusion-weighted imaging (DWI). Split statistics and class prevalence are reported in appendices Table 2

5.2 Score Functions

Our calibration procedures operate on score functions evaluated on a held-out calibration set. We train deep neural network classifiers to obtain these score functions. For CIFAR-10, we train a ResNet-18 (He et al., 2016) on randomly masked images, to maximize the log likelihood of the associated labels. For the FastMRI Knee and Prostate datasets, we follow the recipe of Singhal et al. (2023) to train classifiers on undersampled k-space data.

5.3 Evaluation on Sequential Acquisition Tasks

We first evaluate the proposed methods on three datasets: CIFAR-10 for object classification with sequential acquisition simulated through patch masking; FastMRI Knee for slice-level musculoskeletal abnormality detection from sequentially acquired k-space lines; and Clinical Prostate MRI for PI-RADS-positive classification from sequentially acquired DWI averages, with axial T2 held fixed. Examples of the CIFAR-10 and FastMRI Knee acquisition processes appear in Figure 1. For each dataset, we plot the cumulative fraction of examples stopped and the performance on those examples as functions of the acquisition rate. We report stopped-sample accuracy for CIFAR-10. For FastMRI Knee, we report the class-conditional false negative rate (FNR) among stopped positives for each controlled pathology (ACL tear and meniscus tear) alongside balanced accuracy. For Clinical Prostate MRI, we report FNR among stopped PI-RADS-positive exams (PI-RADS ≥ 3), also alongside balanced accuracy.

Acquisition setup. For CIFAR-10, we simulate sequential acquisition by partitioning each image into 4×4 non-overlapping patches and acquiring these patches progressively over time. For the knee MRI dataset, we simulate sequential acquisition directly in the raw k-space domain, where each acquisition step reveals additional frequency-domain lines. This matches the underlying measurement process and allows us to evaluate stopping decisions in a clinically realistic acquisition regime. For Clinical Prostate MRI, sequential acquisition follows the scanner order of the DWI sequence. The axial T2-weighted volume is acquired before the DWI sequence and is fixed across all states. The full DWI protocol contains 48 acquired DWI measurements: 12 at $b = 50\,\mathrm{s/mm^2}$, with 4 repeats per diffusion direction, and 36 at $b = 1000\,\mathrm{s/mm^2}$, with 12 repeats per direction. Stopping policies are evaluated at 15 scanner-ordered reconstruction checkpoints. At each checkpoint, ADC and computed high-b ($b = 1500\,\mathrm{s/mm^2}$) DWI volumes are reconstructed only from the measurements acquired so far, consistent with PI-RADS v2.1 guidance (Turkbey et al., 2019). Later checkpoints therefore contain progressively higher-SNR diffusion information. A fixed 3D

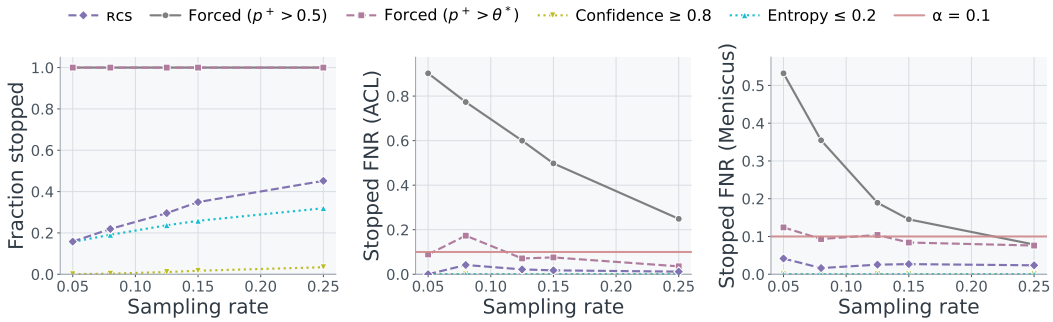

Figure 3: **Class-conditional false negative control on FastMRI knee (ACL tear and meniscus tear).** We decide whether to halt slice-level k-space collection at sampling rates $\mathcal{T} = \{5\%, 8\%, 12.5\%, 15\%, 25\%\}$. The panel reports stopped fraction together with class-conditional FNR among stopped positives for ACL tears and meniscus tears We observe that RCS controls each pathology-specific FNR at the user-specified level ($\alpha = 0.1$) across acquisition rates, including at low undersampling such as 5%, while heuristic baselines violate the threshold.

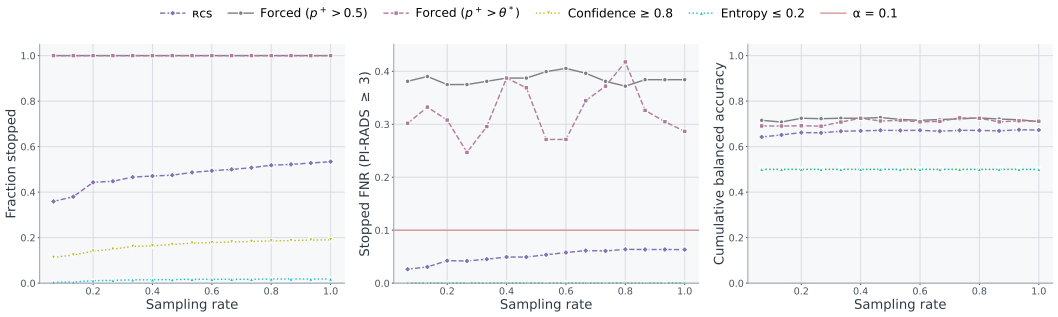

Figure 4: **Controlling false negative rate (FNR) among stopped exams on Prostate PI-RADS classification.** We decide whether to halt acquiring diffusion averages for DWI sequences along the scanner-ordered reconstruction checkpoints. **Left:** fraction of exams stopped by RCS and heuristic criteria. **Center/Right:** FNR and balanced accuracy on the stopped cohort, where the controlled risk is the false-negative rate among PI-RADS-positive exams (PI-RADS ≥ 3) that have already been halted. We observe that RCS maintains this stopped-sample FNR at the user-specified level ($\alpha = 0.1$) across acquisition stages, whereas baseline strategies violate the safety threshold.

T2+DWI fusion classifier produces a PI-RADS-positive score at each checkpoint (Johnson et al., 2025). Appendix Figure 6 visualizes DWI-derived reconstructions across the acquisition trajectory.

Baselines. As baselines, we compare RCS to three potentially "unsafe" strategies. First, **Forced with Static Decision Thresholds** where the model is forced to make a prediction for every sample across all sampling rates. The score threshold used to generate the prediction is a static scalar (we use $p^+ > 0.5$). This is referred to as "Forced ($p^+ > 0.5$)" in Figures 3 and 4. Second, **Forced with Calibrated Decision Thresholds** where the model is forced to make a prediction for every sample across all sampling rates. The score threshold used to make a prediction at each sampling rate is calibrated to maximize some metric of interest (we use balanced accuracy). This is referred to as "Forced ($p^+ > \theta^*$)" in Figures 3 and 4. Third, **Heuristic Stopping Thresholds** determine dynamic stopping times based on some heuristic threshold on the scores. We use Confidence and Entropy of the predictions.

Results. For our synthetic CIFAR-10 experiment, we use the misclassification rate as the risk ($R(y, C_\alpha(\mathbf{x}_t, t)) = \mathbf{I}(y \notin C_\alpha(\mathbf{x}_t, t))$), and control it at the rate of 10% ($\alpha = 0.1$), implying a target coverage/accuracy of 90%. For both datasets, we report performance as a function of the percentage of features acquired (a.k.a., the acquisition rate) and at every acquisition rate, we track

the cumulative percentage of stopped samples. This is a direct indicator of our proposed guarantee: for all stopped samples, we maintain an accuracy level corresponding to the input miss-coverage rate (α). We compare the performance of RCS against the Fixed Sampling Rate Stopping or "Forced", Confidence-based Thresholding, and Entropy-based Thresholding. See Figure 2 for results on CIFAR-10 dataset. Thresholds for the Entropy and Confidence baselines are set to 0.5.

In our clinical diagnosis tasks, the risk of interest is false-negative error among sample that have been stopped. For FastMRI knee, this is instantiated as a *class-conditional* multi-label guarantee: we control pathology-specific FNR for ACL tear and meniscus tear separately, and we report the corresponding per-pathology curves in Figure 3. For Prostate PI-RADS, the task is binary, so the same stopped-sample FNR reduces to control among PI-RADS-positive exams in Figure 4.

We instantiate $\mathcal{D}_{\text{STOP}}(\mathbf{x}_t; \lambda)$ using *ambiguity-band stopping rules*. For each acquisition time t and task k, we define an uncertainty interval $[\ell_{t,k}(\lambda), u_{t,k}(\lambda)]$. The parameter $\lambda \in [0, 1]$ scales the band's width, interpolating from maximal ambiguity (highly conservative) to a narrow, data-driven baseline (aggressive). The policy halts acquisition at the first step where predictive scores for *all* controlled tasks exit their respective bands, assigning terminal predictions based on the side of the exit. While we detail this specific band construction in Appendix A.3, our RCS guarantees remain valid for any arbitrary stopping rule parameterized by λ.

In Figures 3 and 4 we observe that RCS shows efficient stopping behavior while always adhering to the risk constraint. Furthermore, across both sets of experiments, we observe that even though the baseline method "Forced ($p^+ > \theta^*$)" is calibrated to maximize the balanced accuracy, RCS shows comparable performance while never violating the risk threshold.

Sensitivity to the risk level α. Beyond fixed operating points, we study how stopping behavior changes as the user varies the allowed risk α. For each α on a grid, we re-select the certified threshold $\widehat{\lambda}(\alpha)$ on the same calibration scores and evaluate the resulting policy on the held-out test set. Figure 5 summarizes these sweeps for FastMRI knee and Prostate PI-RADS. The top panels show the cumulative stopped fraction versus acquisition rate, with one curve per α: larger α (more permissive risk) yields earlier and more frequent stopping. The bottom panel reports the expected acquisition rate at stopping, $\mathbb{E}[r_{\text{stop}}]$, as a function of α, quantifying the efficiency–risk tradeoff induced by calibration. As expected, $\mathbb{E}[r_{\text{stop}}]$ decreases monotonically as α increases, while the fixed-α panels above confirm that the corresponding stopped cohorts remain consistent with the requested risk level.

5.4 Compatibility with Adaptive Acquisition Policies

Our second set of experiments evaluate whether the proposed stopping-calibration procedures can serve as post hoc wrappers on top of learned adaptive acquisition policies. Since our method is designed to be agnostic to the acquisition mechanism itself, it should be applicable not only to fixed or random acquisition trajectories, but also to learned adaptive policies, provided that the calibration and test trajectories are sampled from the same induced distribution. To demonstrate this, we use **JAFA** (Shim et al., 2018), a reinforcement-learning-based adaptive feature acquisition method that jointly learns which feature to acquire next and when to stop. JAFA provides a strong representative of the adaptive acquisition literature, in which acquisition decisions are optimized for average utility or reward but are not, by themselves, equipped with post hoc statistical guarantees on the risk of the final stopped decision.

Wrapper construction We treat JAFA as an acquisition backbone and construct a parameterized stopping wrapper on top of it. Concretely, JAFA generates a sequential acquisition trajectory and exposes a stopping score at each acquisition step. Specifically, JAFA learns a Deep-Q network, that produces values for a given $(\text{state}, \text{action})$ pair. The policy selects the next action by picking $a_{t+1} = \arg\max_a Q(a, s_t)$, where Q is the Q-network and s is the state at time t. By taking the difference between the Q-value for the "STOP" action and the argmax action (excluding $a_{t+1} =$ "STOP"), we can define a score function. We then instantiate a family of wrapped stopping policies by thresholding this score with a tunable parameter λ, and apply RCS to select a risk-certified member of this wrapped family. This preserves JAFA's adaptive acquisition behavior while calibrating the final decision to satisfy a user-specified risk target.

Results. To isolate the effect of the wrapper cleanly, we perform this experiment on **CIFAR-10** only. As in 5.3, we report: 1) the cumulative fraction of stopped samples versus acquisition rate, and

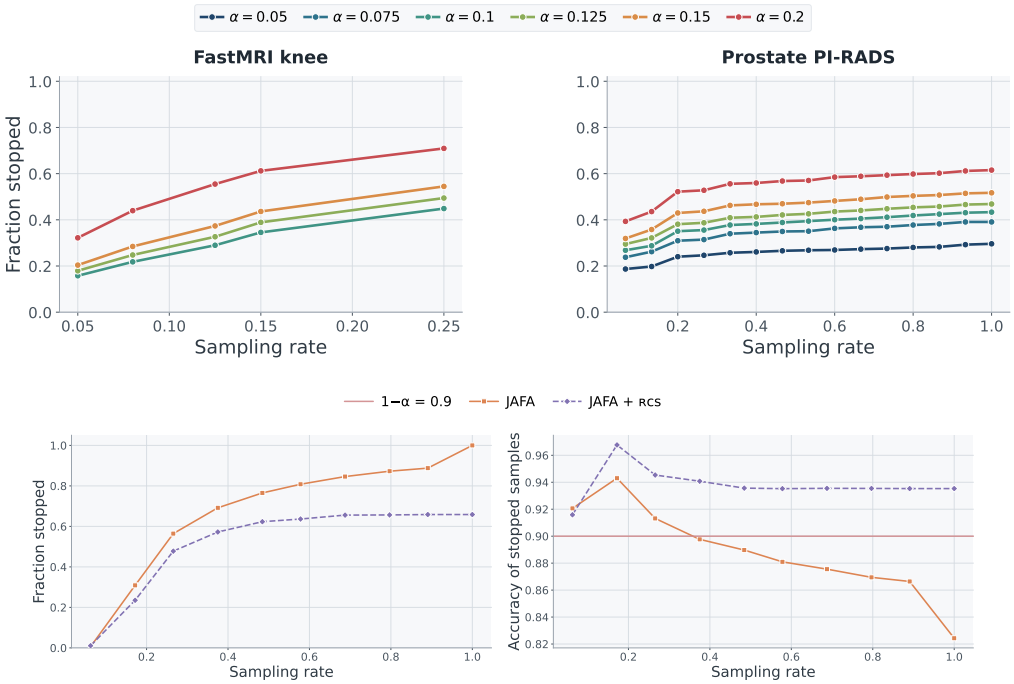

Figure 5: **Upper Panel: Effect of the risk level α on calibrated stopping (FastMRI knee and Prostate PI-RADS).** We show the cumulative fraction of stopped examples versus sampling rate for several values of α (one curve per α). **Lower Panel: Complementing a learned acquisition policy with RCS (CIFAR-10).** In this experiment we wrap JAFA (Shim et al., 2018) with RCS, and compare the efficiency and safety of the resulting policy with JAFA alone. In the left plot, we plot the fraction of samples stopped by each procedure across acquisition stages. In the right plot, we monitor the risk (miscoverage) of both across acquisition stages. We observe that at the cost of slightly lower efficiency, JAFA+RCS adheres to the safety constraint while JAFA alone does not.

2) he stopped-sample performance versus acquisition rate. From Figure 5 (lower panel), we observe that calibrating on JAFA-generated acquisition trajectories enhances the efficiency of the stopping procedure. At the same time, our wrapper ensures that the accuracy of the stopped samples remains above 90%. Because both methods use the same underlying predictor, RCS achieves this result by abstaining on the few samples that JAFA would otherwise misclassify.

6 Discussion

In this work, we introduce risk-controlled stopping (RCS), a method to safely stop data acquisition in sequential classification tasks. RCS provides post-hoc, distribution-free risk control at user desired thresholds and notions of risk. These risk functions and thresholds can be changed without requiring any further training. Our experiments on the FASTMRI knee and the prostate dataset show that RCS can significantly accelerate clinical decision making pipelines, while rigorously controlling class-conditional (or stopped-sample) false-negative risk despite accelerations. Furthermore, we demonstrate that RCS is complementary to learned acquisition policies, and provides rigorous statistical guarantees on their behavior.

6.1 Limitations

Producing the RCS guarantee requires access to fully-sampled data or data sampled from the same policy. However, in future work, we plan on using masked data that could be collected during a deployment of early stopping methods. Moreover, the class of policies parameterized by λ could be enriched by allowing λ to be multi-dimensional, thereby capturing the strong correlation structure among errors across nearby acquisition steps.

ETHICS STATEMENT

This work concerns risk-controlled early stopping of medical-image acquisition. The prostate MRI data were analyzed under an institutional IRB-approved research protocol as a retrospective study of pre-existing, de-identified clinical examinations from patients evaluated for suspected prostate cancer; no patients were enrolled and no imaging was performed for research. Splits were made at the patient level so that no patient appears in more than one of the training, validation, calibration, and test sets. RCS is meant to support clinical judgment, not replace it, and any deployment would require prospective validation; in particular, faster screening can enlarge the screened population and so raise the absolute number of false positives. By design, any sample RCS does not confidently stop on proceeds to the full standard-of-care protocol, so the end-to-end workflow is never less safe than current practice.

REPRODUCIBILITY STATEMENT

The calibration procedure is specified in Section 4, with pseudocode in **??** and the proof of Corollary 1 in **??**. The three policy families (confidence threshold, JAFA Q-margin, ambiguity bands) are defined in Section 4 and Section A.3. Dataset construction, acquisition ordering, and split statistics are given in Section 5.3 and the appendix. CIFAR-10 and FASTMRI are public; code for the calibration procedure and experiments will be released as anonymized supplementary material.

AI USE STATEMENT

The authors used generative AI tools to edit prose for clarity and concision and to help draft the schematic in **??**. All technical claims, proofs, experimental designs, and results were produced and verified by the authors, who reviewed and edited every AI-assisted passage and take full responsibility for the content of this work.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Tim Bakker, Herke van Hoof, and Max Welling. Experimental design for mri by greedy policy search. *Advances in Neural Information Processing Systems*, 33, 2020.

Chun-Hao Chang, Mingjie Mai, and Anna Goldenberg. Dynamic measurement scheduling for event forecasting using deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pp. 6424–6447. PMLR, 2023.

Etienne Gauthier, Francis Bach, and Michael I Jordan. E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050*, 2025.

Aritra Ghosh and Andrew Lan. Difa: Differentiable feature acquisition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7705–7713, Jun. 2023. doi: 10.1609/aaai.v37i6.25934. URL https://ojs.aaai.org/index.php/AAAI/article/view/25934.

Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. Adaptive-halting policy network for early classification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019.

Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. Recurrent halting chain for early multi-label classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Metod Jazbec, Alexander Timans, Tin H Veljković, Kaspar Sakmann, Dan Zhang, Christian A Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. *Advances in Neural Information Processing Systems*, 37:129825–129854, 2024.

Patricia M. Johnson, Tarun Dutt, Luke A. Ginocchio, Amanpreet Singh Saimbhi, Lavanya Umapathy, Kai Tobias Block, Daniel K. Sodickson, Sumit Chopra, Angela Tong, and Hersh Chandarana. Prostate cancer risk stratification and scan tailoring using deep learning on abbreviated prostate mri. *Journal of Magnetic Resonance Imaging*, 62(3):858–866, 2025. doi: 10.1002/jmri.29798.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: efficient dynamic discovery of high-value information with partial VAE. *CoRR*, abs/1809.11142, 2018. URL http://arxiv.org/abs/1809.11142.

Prem Melville, Foster Provost, Maytal Saar-Tsechansky, and Raymond Mooney. Economical active feature-value acquisition through expected utility estimation. In *Proceedings of the 1st international workshop on Utility-based data mining*, pp. 10–16, 2005.

Liran Ringel, Regev Cohen, Daniel Freedman, Michael Elad, and Yaniv Romano. Early time classification with accumulated accuracy gap control. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e5841df2166dd424a57127423d276bbe-Paper.pdf.

Raghav Singhal, Mukund Sudarshan, Anish Mahishi, Sri Kaushik, Luke Ginocchio, Angela Tong, Hersh Chandarana, Daniel K Sodickson, Rajesh Ranganath, and Sumit Chopra. On the feasibility of machine learning augmented magnetic resonance for point-of-care identification of disease. *arXiv preprint arXiv:2301.11962*, 2023.

Baris Turkbey, Andrew B. Rosenkrantz, Masoom A. Haider, Anwar R. Padhani, Geert Villeirs, Katarzyna J. Macura, Clare M. Tempany, Peter L. Choyke, Francois Cornud, Daniel J. Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European Urology*, 76(3):340–351, 2019. doi: 10.1016/j.eururo.2019.02.033.

Chen-Yu Yen, Raghav Singhal, Umang Sharma, Rajesh Ranganath, Sumit Chopra, and Lerrel Pinto. Adaptive sampling of k-space in magnetic resonance for rapid pathology prediction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 PROOF OF RISK CONTROL

We now briefly describe the statistical properties that lead to the guarantees underlying RCS.

Exchangeability of sequence-level losses. For any fixed policy parameter, the stopping rule and terminal decision are deterministic functions of the full trajectory X_i. Since the calibration trajectories (X_i, Y_i) are i.i.d., the induced sequence-level losses are exchangeable across samples. This allows Hoeffding's inequality to be applied to the empirical risk for each fixed candidate policy.

Validity of the conditional p-values. For the LTT procedure, the pairs $(S_i^{(j)}, L_i^{(j)})$ are i.i.d. across calibration samples because each depends only on the corresponding trajectory-label pair (X_i, Y_i). Conditional on the stopped index set A_j, the selected losses $\{L_i^{(j)} : i \in A_j\}$ remain independent. Therefore, conditional on A_j, the empirical conditional risk $\widehat{R}_{\mathrm{cond}}(\lambda_j)$ is the mean of n_j independent $[0,1]$-valued random variables with expectation $R_{\mathrm{cond}}(\lambda_j)$ and Hoeffding's inequality yields the valid one-sided p-value (6). Marginalizing over A_j proves unconditional p-value validity.

Algorithm 1 RCS: Risk Control on Stopped Samples

Input: Calibration set $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^{n_{cal}}$, candidate grid $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$, target conditional risk α, confidence level δ, utility $U(\lambda)$.

for $j = 1, \ldots, m$ **do**
  for $i = 1, \ldots, n_{cal}$ **do**
    Evaluate policy D_{λ_j} on trajectory X_i and compute stopping time

$$\tau_{\lambda_j}(X_i) = \inf\{t \in \{1, \ldots, T\} : D_{\lambda_j}(X_i^t) \neq \text{CONTINUE}\}$$

    Set

$$S_i^{(j)} = \mathbf{1}\{\tau_{\lambda_j}(X_i) < \infty\}$$

    if $S_i^{(j)} = 1$ **then**
      Compute terminal decision

$$\hat{Y}_{\lambda_j}(X_i) = D_{\lambda_j}(X_i^{\tau_{\lambda_j}(X_i)})$$

      Compute loss

$$L_i^{(j)} = \mathcal{L}\big(\hat{Y}_{\lambda_j}(X_i), Y_i\big)$$

    end if
  end for
  Let

$$A_j = \{i : S_i^{(j)} = 1\}, \qquad n_j = |A_j|$$

  if $n_j = 0$ **then**
    Set $p_j = 1$.
  else
    Compute empirical conditional risk

$$\widehat{R}_{\text{cond}}(\lambda_j) = \frac{1}{n_j} \sum_{i \in A_j} L_i^{(j)}$$

    Compute p-value

$$p_j = \exp\Big(-2n_j\big(\alpha - \widehat{R}_{\text{cond}}(\lambda_j)\big)_+^2\Big)$$

  end if
end for
Apply Holm's step-down procedure to $\{p_j\}_{j=1}^m$ at level δ.
Let

$$\widehat{\mathcal{S}} = \{\lambda_j \in \Lambda : H_{0,j} \text{ is rejected}\}$$

if $\widehat{\mathcal{S}} = \varnothing$ **then**
  return default conservative policy or abstain.
else
  Select

$$\hat{\lambda} \in \arg\max_{\lambda_j \in \widehat{\mathcal{S}}} U(\lambda_j).$$

  return $\hat{\lambda}$.
end if

A.2 STATISTICS FOR FASTMRI KNEE AND PROSTATE DATASETS

Table 1: Dataset statistics for FASTMRI Knee. $N = 41,679$ slices with $80\%, 10\%, 10\%$ split for the train, validation and test sets.

Class Label	Fraction of Positive labels
ACL Sprain	0.03
Meniscal Tear	0.11
Abnormal	0.16

Table 2: Clinical Prostate MRI split statistics. PI-RADS positive is defined as radiologist-assigned PI-RADS ≥ 3.

Split	Total	Fraction of Positive Labels
Train	6413	0.48
Validation	952	0.49
Calibration	945	0.45
Test	824	0.40

A.3 AMBIGUITY-BAND STOPPING POLICIES

The RCS procedure is agnostic to the form of the stopping policy D_λ, provided that the policy family is fixed prior to calibration and yields a bounded trajectory-level loss. In practice, however, the efficiency of safe stopping depends strongly on how this policy family is parameterized. We therefore instantiate D_λ using a family of *ambiguity-band stopping rules* defined over time-dependent predictive scores. This construction is used primarily for the MRI experiments, where clinically relevant false negative and false positive tradeoffs make thresholded score bands a natural decision primitive.

Predictive scores. Suppose that at acquisition time $t \in \{0, \dots, T-1\}$, the model outputs a positive-class score $S_{i,t,k} \in [0,1]$ for sample i and label k, based on the cumulative partial observation X_i^t. In the multi-task setting, $k \in \{1, \dots, K\}$. Larger values of $S_{i,t,k}$ indicate stronger evidence for the positive class.

Baseline ambiguity anchors. For each acquisition step t and task k, we define a pair of nominal ambiguity anchors $b_{t,k}^-, b_{t,k}^+$, with $0 \leq b_{t,k}^- \leq b_{t,k}^+ \leq 1$, that identify a region of the score space where the model is considered ambiguous. Intuitively, scores below $b_{t,k}^-$ provide evidence for a negative decision, while scores above $b_{t,k}^+$ provide evidence for a positive decision. The interval $[b_{t,k}^-, b_{t,k}^+]$ therefore serves as a time-specific uncertainty band. In our experiments, these anchors are estimated from a small held-out estimation set by comparing the empirical score distributions of positive and negative examples at each acquisition step. Concretely, for a fixed (t,k), let

$$S_{t,k}^+ = \{S_{i,t,k} : y_{i,k} = 1\}, \qquad S_{t,k}^- = \{S_{i,t,k} : y_{i,k} = 0\},$$

denote the positive- and negative-class score collections. We then define the nominal ambiguity anchors as the narrowest non-inverted quantile bracket within a pre-specified quantile family, yielding a data-driven aggressive endpoint for the band family. Importantly, these anchors are not themselves claimed to be risk-calibrated; rather, they serve to define an ordered family of stopping policies, while the actual statistical guarantee is supplied downstream by the RCS calibration procedure.

Band modulation by λ Given the nominal anchors $(b_{t,k}^-, b_{t,k}^+)$, we define a one-parameter family of effective ambiguity bands indexed by $\lambda \in [0,1]$:

$$\ell_{t,k}(\lambda) = \lambda\, b_{t,k}^-, \qquad u_{t,k}(\lambda) = \lambda\, b_{t,k}^+ + (1-\lambda). \tag{9}$$

This construction interpolates between:

- **maximal ambiguity** at $\lambda = 0$, where $[\ell_{t,k}(\lambda), u_{t,k}(\lambda)] = [0,1]$, and
- the **nominal ambiguity band** at $\lambda = 1$, where $[\ell_{t,k}(\lambda), u_{t,k}(\lambda)] = [b_{t,k}^-, b_{t,k}^+]$

Larger values of λ therefore correspond to narrower ambiguity bands and more aggressive stopping rules.

Stopping rule. For a fixed λ, the stopping rule halts acquisition at the first time all tasks simultaneously exit their ambiguity bands. Formally, define the stopping time

$$\tau_\lambda(X_i) = \inf\{t \in \{0, \dots, T-1\} : \forall k,\ S_{i,t,k} \notin [\ell_{t,k}(\lambda), u_{t,k}(\lambda)]\}$$

If no such time exists, we set $\tau_\lambda(X_i) = \infty$ and allow the policy to abstain or defer to the terminal horizon, depending on the risk formulation under consideration.

Terminal prediction. The ambiguity band jointly determines both stopping and prediction. Acquisition continues while the score remains inside the effective ambiguity interval $[\ell_{tk}(\lambda), u_{tk}(\lambda)]$. Once the score exits this interval, the side of exit determines the terminal class label: scores below the lower boundary are assigned to the negative class, while scores above the upper boundary are assigned to the positive class. Thus, for a single task, the sequential policy is

$$D_\lambda(X^t) = \begin{cases} 0, & S_t(X^t) < \ell_t(\lambda), \\ 1, & S_t(X^t) > u_t(\lambda), \\ \text{CONTINUE}, & S_t(X^t) \in [\ell_t(\lambda), u_t(\lambda)]. \end{cases}$$

In the multi-task setting, acquisition halts once every task has exited its corresponding ambiguity band, and the terminal label for each task is determined by the side on which its score exits. This yields a complete stopping-and-prediction policy D_λ, which can then be certified by the procedure in Section 4.

Interpretation. The ambiguity-band construction should be viewed as a practical policy-family parameterization rather than as the source of the final guarantee. The role of the nominal anchors is to define an aggressive score-space ambiguity template at each acquisition time, and the role of λ is to control how aggressively that ambiguity region is shrunk. The final risk guarantee is not inherited from the bands themselves, but from the downstream calibration procedure applied to the resulting trajectory-level stopping policies.

A.4 CLINICAL PROSTATE MRI ACQUISITION STATES

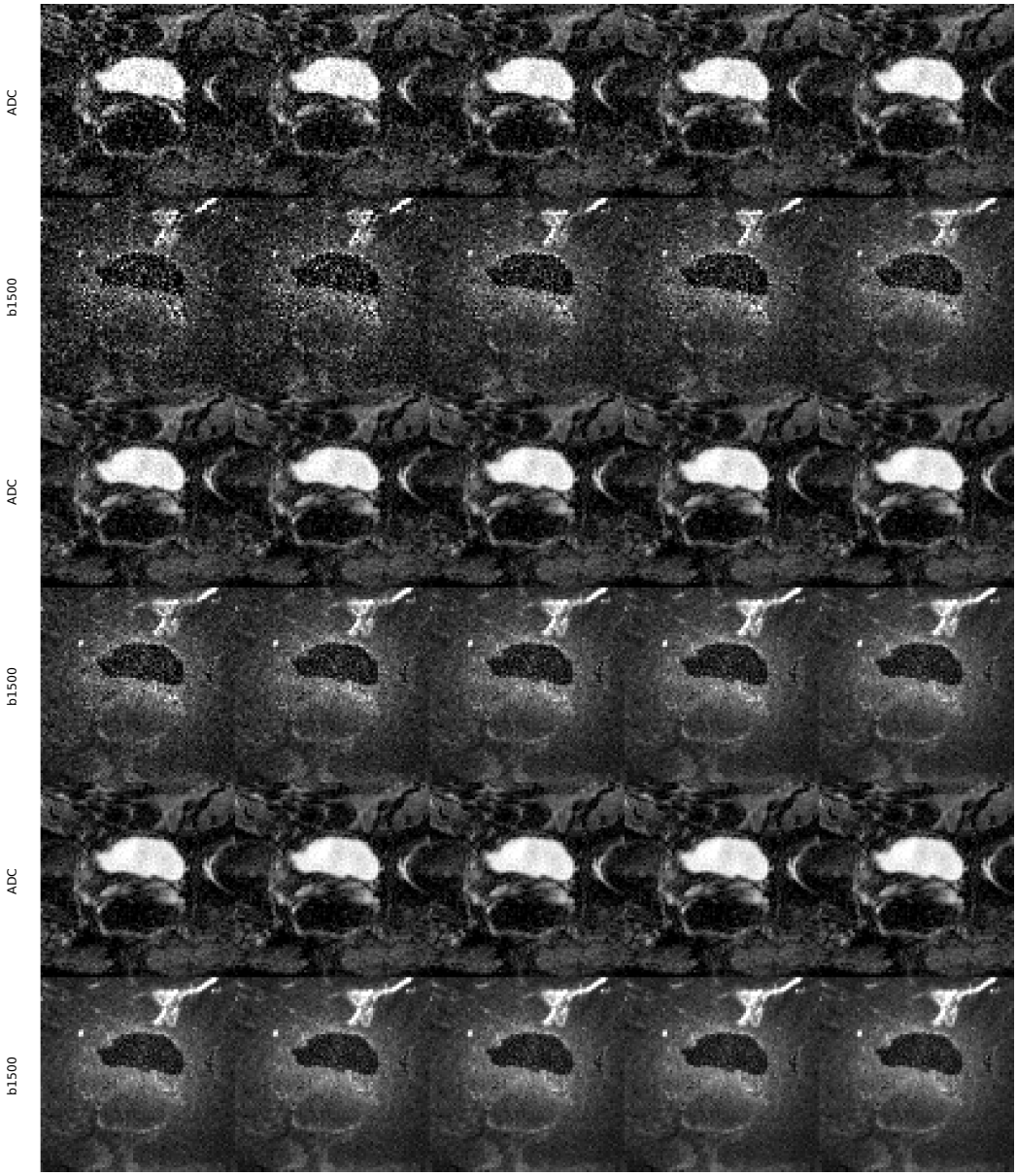

Figure 6: Prostate DWI-derived images across sequential acquisition states. Each tile shows ADC above computed high-b-value DWI ($b = 1500$) for the same scan and slice. Tiles are ordered left-to-right and top-to-bottom according to scanner acquisition order, from $1/1$ to $4/12$ acquired $b = 50/b = 1000$ averages.