Radiation-Preserving Selective Imaging for Pediatric Hip Dysplasia: A Cross-Modal Ultrasound-Xray Policy with Limited Labels

Duncan Stothers[1,3*], **Ben Stothers**[1,2], **Emily Schaeffer**[1,2], **Kishore Mulpuri**[1,2]

[1]HIPPY lab
[2]University of British Columbia
[3]Harvard University

Abstract

We study an ultrasound-first, radiation-preserving policy for developmental dysplasia of the hip (DDH) that requests an X-ray (XR) only when needed.

We (i) pretrain modality-specific encoders (ResNet-18) with SimSiam on a large unlabelled registry (37,186 ultrasound; 19,546 radiographs), (ii) freeze the backbones and fit small, measurement-faithful heads on DDH-relevant landmarks and measurements, (iii) calibrate a one-sided conformal deferral rule on ultrasound predictions that provides finite-sample *marginal* coverage guarantees under exchangeability, using a held-out calibration set. Ultrasound heads predict Graf α/β and femoral head coverage; X-ray heads predict acetabular index (AI), center-edge (CE) angle and IHDI grade. On our held-out labeled evaluation set, ultrasound measurement error is modest (e.g., α MAE $\approx$ 9.7°, coverage MAE $\approx$ 14.0 percentage points), while radiographic probes achieve AI and CE MAEs of $\approx$ 7.6° and $\approx$ 8.9°, respectively. The calibrated US-only policy is explored across rule families (alpha-only; alpha OR coverage; alpha AND coverage), conformal miscoverage levels ($\delta_\alpha, \delta_{\mathrm{cov}}$), and per-utility trade-offs using decision-curve analysis. Conservative settings yield high coverage (e.g., $\sim$0.90 for α) with near-zero US-only rates; permissive settings (e.g., alpha OR coverage at larger deltas) achieve non-zero US-only throughput with expected coverage trade-offs.

The result is a simple, reproducible pipeline that turns limited labels into interpretable measurements and tunable selective imaging curves suitable for clinical handoff and future external validation.

Background

Clinical context and age-aware measurements. Developmental dysplasia of the hip (DDH) spans acetabular undercoverage through dislocation across infancy and early childhood. In the radiographic domain, clinically accepted measurements include the acetabular index (AI), IHDI grading, and, when ossification permits, the lateral center-edge (Wiberg) angle and Sharp's acetabular angle (Tönnis 1987; Narayanan et al. 2015; Doski, Qadir et al. 2022; Sharp 1961; Wiberg 1939). In early infancy, standardized sonography under the Graf method (reporting α/β angles and femoral head coverage) is the dominant screening and early management tool (Graf 1980; Graf, Scott, and Lercher 2006; Zieger 1986; Omeroğlu 2014). As ossification progresses, anteroposterior (AP) pelvic radiographs become more informative for acetabular development and surgical planning; minimizing ionizing exposure in pediatrics motivates selective XR acquisition (Keller and Nijs 2009; Dezateux and Rosendahl 2007).

Ultrasound-only automation (what exists). A substantial body of work automates hip ultrasound tasks: standard-plane detection, quality gating, Graf typing, and direct prediction of α/β and femoral head coverage. Representative approaches include multitask CNNs that jointly detect plane adequacy and infer Graf measurements (Chen et al. 2022; Hu et al. 2022), lightweight real-time systems designed for point-of-care guidance (Hsu, Lin et al. 2025), and pipelines that integrate structure/landmark cues to improve robustness (Kinugasa, Kobayashi et al. 2023; Chen et al. 2024). Reviews of pediatric MSK ultrasound emphasize ultrasound's safety, accessibility, and the potential for AI-assisted standardization to reduce operator dependence (van Kouswijk, Yeung, and Jaremko 2025). These methods typically optimize unimodal accuracy/MAE and, when evaluated, report agreement with expert sonographers on Graf measures and coverage.

Radiograph-only automation (what exists). On AP pelvic radiographs, deep systems tackle triage, landmark detection, and explicit measurement prediction for AI, CE, and related angles. Recent efforts use local-global architectures and landmark-aware heads to increase interpretability and reduce measurement bias (Liu et al. 2020; Moon et al. 2024; Li et al. 2019). Additional work frames DDH classification and grading tasks with CNNs and detectors (Park et al. 2021; Zhang et al. 2020; Fraiwan et al. 2022; Den et al. 2023; Xu et al. 2022). Systematic reviews synthesize steady progress but highlight the need for stronger labeling protocols, external validation, and reporting of clinically meaningful measurement error rather than only image-level accuracy (Wu, Wang et al. 2023; Chen, Wang et al. 2024).

Cross-modal US-XR learning (what remains sparse). Despite mature unimodal pipelines, explicitly *paired* US-XR learning for DDH is scarce. Reviews note limited availability of temporally matched US-XR cohorts, inconsistent

[*]Corresponding author: duncanstothers@alumni.harvard.edu

annotation schemes across modalities, and few methods that directly model the decision of whether to acquire an XR given a US (Wu, Wang et al. 2023; Chen, Wang et al. 2024; van Kouswijk, Yeung, and Jaremko 2025). Emerging resources begin to address pairing and benchmarking (Qi, Zhang et al. 2025), and there are early bimodal/local-global ideas in related orthopedic settings (Shimizu, Tsukagoshi et al. 2024), but a practical, measurement-aware policy that trades radiation against risk with explicit guarantees remains underexplored.

Why the intersection matters: clinical and statistical rationale. US and XR encode overlapping-but not identical-morphology. Graf α and femoral head coverage on US reflect acetabular roof orientation and femoral containment in early infancy, while AI/CE on XR quantify acetabular slope and lateral coverage once ossification permits reliable landmarks (Graf 1980; Graf, Scott, and Lercher 2006; Tönnis 1987; Wiberg 1939). Clinically, a clearly normal US (e.g., high α, adequate coverage) often reduces the immediate value of XR in very young infants, whereas abnormal or borderline US increases the value of XR to confirm severity, assess symmetry, and plan management. Statistically, this suggests a selective acquisition problem: use US-derived measurements and calibrated uncertainty to predict whether XR would meaningfully shift the diagnostic or management decision boundary.

From ultrasound to radiographic surrogates (US→XR prediction). Several radiographic targets admit physiologic surrogates on ultrasound. Elevated α and adequate coverage reduce the likelihood of elevated AI or pathologic IHDI in age-appropriate cohorts; conversely, low α or poor coverage raise suspicion for acetabular underdevelopment that XR can quantify with AI/CE more precisely as ossification progresses. A measurement-aware proxy model can therefore (i) estimate XR-relevant risk from US measurements, (ii) calibrate one-sided lower bounds that certify "confidently normal" cases against clinical thresholds (e.g., $\alpha \geq 60°$; coverage $\geq 50\%$), and (iii) defer to XR when uncertainty or predicted abnormality exceeds a tunable margin. This framing preserves interpretability (decisions ride on named measurements) and keeps the bridge between modalities clinically legible.

Decision-theoretic selective imaging and risk control. Selective acquisition can be cast as a utility-optimization under uncertainty: balance a (small) radiation cost against the (potentially large) cost of missing a lesion that XR would reveal. In deployment, finite-sample guarantees matter. Conformal prediction provides distribution-free *marginal* coverage control for one-sided lower bounds on US measurements. We then use these bounds inside US-only rules while routing ambiguous cases to XR. Decision-curve analysis then summarizes net benefit across utility weights, exposing operating points along the safety-radiation frontier.

Positioning of this work. We contribute a pragmatic, label-efficient cross-modal policy: (1) self-supervised pretraining on large unlabeled US/XR registries with frozen backbones; (2) small, measurement-faithful heads trained on a curated set of paired studies; (3) a one-sided conformal layer that certifies "US-only" decisions against clinically standard thresholds; and (4) decision-curve analysis that makes explicit the trade between XR utilization and risk. The intersection we target, using calibrated US measurements to decide when XR adds value, addresses the practical gap highlighted by recent reviews (Wu, Wang et al. 2023; Chen, Wang et al. 2024; van Kouswijk, Yeung, and Jaremko 2025) and complements ongoing efforts in unimodal automation (Park et al. 2021; Zhang et al. 2020; Liu et al. 2020; Moon et al. 2024; Chen et al. 2022; Hu et al. 2022).

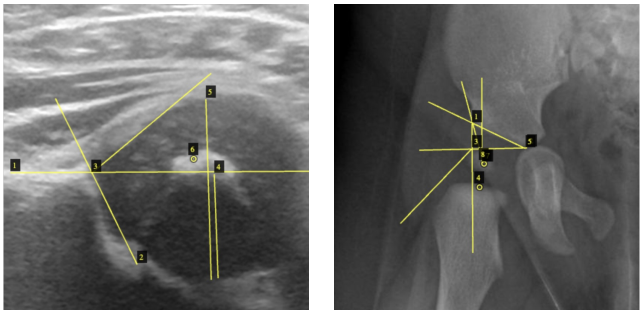

Figure 1: (Left) Pediatric hip ultrasound (US) and (Right) pediatric hip X-ray (XR). On US: annotations for the horizontal (1), Graf α (2), Graf β (3), percent femoral head coverage (4/5), and location of ossific nucleus (6). On XR: P-line (1), H-line (2), 45° line (3), IHDI quadrant and grade (4), Acetabular Index (AI) (5), Central Edge Vertical line (6), Central Edge Angle line (CE) (7), location of ossific nucleus (8).

Methods

Data, pairing, and splits

We assembled a large unlabeled corpus for self-supervised pretraining (US: 37,186; XR: $19{,}546$ grayscale images, resized to 512×512 and channel-repeated to 3). For supervised training and policy evaluation, we curated a small, paired subset with trainee line/point annotations and strict side/date matching. We removed frog-lateral and non-AP studies with a fixed set of view and QC rules, which eliminated eight subjects. The resulting labeled set contained 321 images from 75 subjects. To avoid leakage, we used subject-level splits: *post-train* (30 subjects; 136 images), *calibration* (7 subjects; 28 images; including 26 ultrasound images), and *evaluation* (38 subjects; 157 images). On *evaluation*, strict matching by (subject, date, side) yielded $N{=}77$ hip pairs with at least one XR ground truth label (AI and/or CE and/or IHDI).

For each US image, the trainee recorded a horizontal baseline and α/β lines, exposed/total femoral-head lengths, and an optional ossific-nucleus point. For each XR image, the trainee recorded Hilgenreiner's H-line, Perkin's P-line, a $45°$ reference, the H-point, an acetabular index line, and CE-angle rays, with laterality handled explicitly. Numeric targets were derived by the standard acute-angle and ratio formulas (Graf α/β; coverage = exposed/total; AI = angle(H-line, AI-line); CE = angle(vertical, CE-ray); IHDI from the quadrant of the H-point).

Self-supervised pretraining (SimSiam) and frozen encoders

We train separate encoders for US and XR with SimSiam on the unlabeled corpora and then freeze them. Let f_ϕ be a ResNet-18 encoder, h_ϕ a projection MLP, and q_ϕ a prediction MLP. For an image x, we draw two augmentations t_1, t_2 to obtain views $v_1{=}t_1(x)$ and $v_2{=}t_2(x)$. Define

$$\begin{aligned} z_1 &= h_\phi\big(f_\phi(v_1)\big), \quad z_2 = h_\phi\big(f_\phi(v_2)\big), \\ p_1 &= q_\phi(z_1), \quad p_2 = q_\phi(z_2). \end{aligned}$$

The SimSiam loss is the stop-gradient negative cosine similarity:

$$\mathcal{L}_{\mathrm{SSL}}(x;\phi) = -\tfrac{1}{2}\left[\frac{\langle p_1,\ \mathrm{sg}(z_2)\rangle}{\|p_1\|_2\,\|\mathrm{sg}(z_2)\|_2} + \frac{\langle p_2,\ \mathrm{sg}(z_1)\rangle}{\|p_2\|_2\,\|\mathrm{sg}(z_1)\|_2}\right].$$

We train for 10 epochs per modality with standard SimSiam augmentations (random crop/flip, color jitter for XR toned down for US), then discard h_ϕ, q_ϕ and freeze the encoder f_ϕ for downstream use.

Measurement heads on frozen encoders

Given a frozen modality-specific encoder f_ϕ, we extract a 512-dimensional feature $u = f_\phi(x)$ (global average pooled). We then fit small, measurement-faithful heads with task-appropriate losses.

Ultrasound head. We use a single MLP with one hidden layer (128 units) and three outputs predicting $\hat{\alpha}, \hat{\beta} \in \mathbb{R}$ (degrees) and $\widehat{\mathrm{cov}} \in \mathbb{R}$ (percentage points). Let $y^\alpha, y^\beta, y^{\mathrm{cov}}$ denote the trainee labels for a given image. The US loss is mean absolute error (MAE) with per-task scaling to balance magnitudes:

$$\mathcal{L}_{\mathrm{US}}(x) = \lambda_\alpha\left|\hat{\alpha} - y^\alpha\right| + \lambda_\beta\left|\hat{\beta} - y^\beta\right| + \lambda_{\mathrm{cov}}\left|\widehat{\mathrm{cov}} - y^{\mathrm{cov}}\right|.$$

In our implementation we set $\lambda_\alpha{=}\lambda_\beta{=}1$ and $\lambda_{\mathrm{cov}}{=}1$ (coverage reported in percentage points).

Radiograph heads. We fit angle regressors for AI and CE and, when present, an IHDI classifier. Let $\hat{a}, \hat{e} \in \mathbb{R}$ be AI and CE predictions and $\hat{\pi} \in \Delta^{K-1}$ be softmax probabilities over K IHDI grades. With angle labels $y^{\mathrm{AI}}, y^{\mathrm{CE}}$ and (optional) IHDI one-hot y^{IHDI},

$$\begin{aligned} \mathcal{L}_{\mathrm{XR}}(x) = \mu_{\mathrm{AI}}\left|\hat{a} - y^{\mathrm{AI}}\right| + \mu_{\mathrm{CE}}\left|\hat{e} - y^{\mathrm{CE}}\right| \\ + \mu_{\mathrm{IHDI}}\,\mathrm{cross\_entropy}\big(\hat{\pi}, y^{\mathrm{IHDI}}\big), \end{aligned}$$

where the crossentropy term is dropped if y^{IHDI} is absent. In all experiments we set $\mu_{\mathrm{AI}} = \mu_{\mathrm{CE}} = \mu_{\mathrm{IHDI}} = 1$, i.e., we use an unweighted sum of the available terms. When y^{IHDI} is missing, the classification term is omitted rather than reweighted. Given the limited dataset size and the relatively similar numeric scale of AI and CE errors, we did not attempt further loss reweighting. We train heads on the post-train split only using a fixed training schedule and freeze them for evaluation and policy analysis. The calibration split is reserved for affine bias correction and conformal calibration.

Calibration: affine bias correction and one-sided conformal bands

For selective imaging we require calibrated lower bounds on US measurements with finite-sample (marginal) coverage under exchangeability. These bounds are then used inside the US-only decision rules. We perform two steps on the US calibration set $\mathcal{C} = \{(x_i, y_i)\}$ for each target $t \in \{\alpha, \mathrm{cov}\}$.

(i) Affine bias correction. Let $\hat{y}_i^t$ be the head prediction on x_i. We fit a robust affine correction

$$\tilde{y}_i^t \;=\; a_t\,\hat{y}_i^t \;+\; b_t$$

$$\text{by} \quad (a_t, b_t) \;\in\; \arg\min_{a,b} \sum_{(x_i,y_i)\in\mathcal{C}} \left|a\,\hat{y}_i^t + b - y_i^t\right|.$$

This reduces small systematic bias without re-training the head.

(ii) One-sided residual quantiles. Define residuals $r_i^t = y_i^t - \tilde{y}_i^t$. For a desired miscoverage level $\delta_t \in (0,1)$ we compute a one-sided conformal radius

$$q_t^+(\delta_t) \;=\; \mathrm{Quantile}_{\lceil(|\mathcal{C}|+1)(1-\delta_t)\rceil/|\mathcal{C}|}\Big(-r_i^t \text{ over } (x_i, y_i) \in \mathcal{C}\Big),$$

so that, under exchangeability, with probability $\geq 1 - \delta_t$ a fresh sample will satisfy $y^t \geq \tilde{y}^t(x) - q_t^+(\delta_t)$. We then define the calibrated lower bound

$$\mathrm{LB}_t(x;\delta_t) \;=\; \tilde{y}^t(x) \;-\; q_t^+(\delta_t).$$

We sweep δ_α and δ_{cov} on a discrete grid (e.g., 0.10:0.40) to produce coverage–utilization curves; smaller δ corresponds to higher target coverage $(1 - \delta)$ and more conservative bounds.

Selective imaging rules (US-only vs. defer to XR)

Let T_α and T_{cov} denote clinical normality thresholds (we use $T_\alpha{=}60^\circ$ and $T_{\mathrm{cov}}{=}50\%$). For an evaluation US image x, we declare "US-only" if the calibrated lower bounds exceed thresholds. We study three rule families:

Alpha-only

$$d_\alpha(x) \;=\; \mathbb{I}[\mathrm{LB}_\alpha(x;\delta_\alpha) \;\geq\; T_\alpha]\,. \tag{1}$$

Alpha OR Coverage

$$\begin{aligned} d_{\alpha\vee\mathrm{cov}}(x) \;=\; \mathbb{I}[\mathrm{LB}_\alpha(x;\delta_\alpha) \;\geq\; T_\alpha \;\vee \\ \mathrm{LB}_{\mathrm{cov}}(x;\delta_{\mathrm{cov}}) \;\geq\; T_{\mathrm{cov}}]\,. \end{aligned} \tag{2}$$

Alpha AND Coverage

$$\begin{aligned} d_{\alpha\wedge\mathrm{cov}}(x) \;=\; \mathbb{I}[\mathrm{LB}_\alpha(x;\delta_\alpha) \;\geq\; T_\alpha \;\wedge \\ \mathrm{LB}_{\mathrm{cov}}(x;\delta_{\mathrm{cov}}) \;\geq\; T_{\mathrm{cov}}]\,. \end{aligned} \tag{3}$$

Here $d(\cdot) \in \{0,1\}$ indicates "US-only" (1) vs. "defer to XR" (0). Sweeping $(\delta_\alpha, \delta_{\mathrm{cov}})$ yields a grid of policies.

Ossification-proxy variant. As a pragmatic sensitivity analysis, we also evaluate a variant that conditions thresholds on an ossific-nucleus flag $o \in \{0,1\}$ (from the US point annotation), e.g.,

$$T_\alpha(o) = \begin{cases} 60^\circ, & o = 0 \\ 60^\circ, & o = 1 \end{cases} T_{\mathrm{cov}}(o) = \begin{cases} 50\%, & o = 0 \\ 50\%, & o = 1 \end{cases}$$

(kept equal in this study, but the machinery supports o-specific thresholds). This lets future work explore age/ossification-aware tuning without altering the conformal recipe.

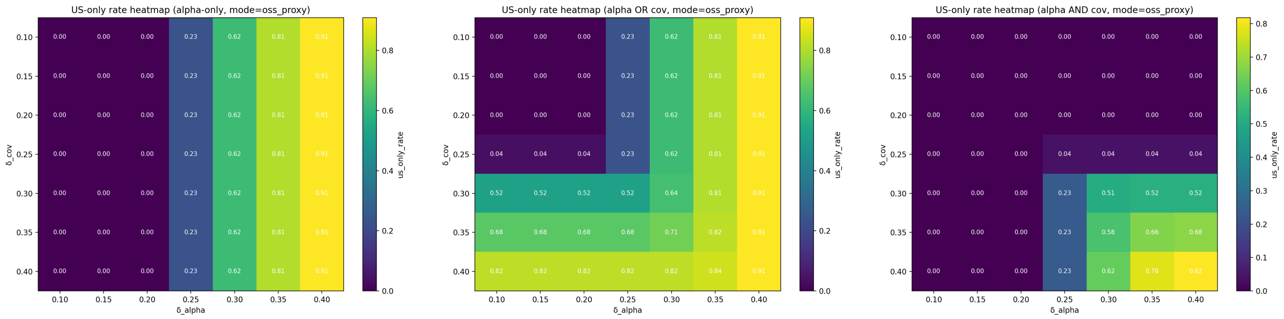

Figure 2: US-only rate across miscoverage levels $(\delta_\alpha, \delta_{\text{cov}})$ (target coverages $1 - \delta$) for the three policy families. Smaller δ is more conservative.

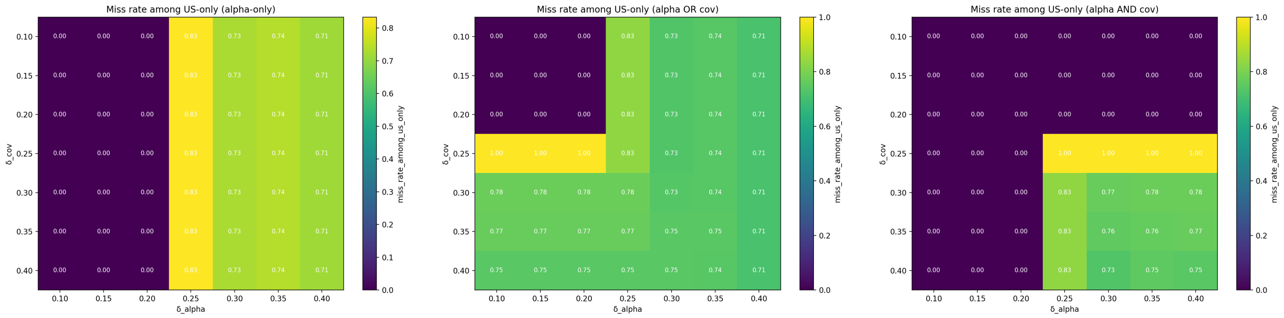

Figure 3: Miss rate among US-only decisions across the same miscoverage grid $(\delta_\alpha, \delta_{\text{cov}})$.

Pair construction and XR ground-truth events

To align US decisions with XR outcomes, we form strict pairs on the evaluation split using $(\text{subject}, \text{date}, \text{side})$. For each pair j, define a binary XR abnormality indicator $z_j \in \{0, 1\}$ that aggregates available radiographic ground truth (e.g., thresholding AI or CE when present, or using IHDI grades). We report all computations both per-rule and per-$(\delta_\alpha, \delta_{\text{cov}})$ on the subset with at least one XR label.

Empirical coverage and safety diagnostics

On the evaluation US set $\mathcal{E}$ we compute empirical one-sided coverage for each target,

$$\widehat{\text{cvg}}_t = \frac{1}{|\mathcal{E}|} \sum_{(x,y)\in\mathcal{E}} \mathbb{I}\big[y^t \geq \text{LB}_t(x; \delta_t)\big], \quad t \in \{\alpha, \text{cov}\}.$$

We also report miss rate among US-only decisions, $\text{MR} = \frac{\sum_j d_j z_j}{\sum_j d_j}$, and the US-only rate, $\text{UOR} = \frac{1}{N} \sum_j d_j$, where N is the number of strict pairs.

Decision-curve analysis (utility over cost-penalty grids)

To summarize policy desirability across clinical preferences, we define a simple per-pair utility with radiation cost $\lambda \geq 0$ and miss penalty $\mu \geq 0$:

$$u_j(d_j; \lambda, \mu, z_j) \;=\; -\lambda\,(1 - d_j) \;-\; \mu\, z_j\, d_j.$$

Acquiring XR (d_j=0) incurs cost λ; skipping XR (d_j=1) risks a penalty μ if the case is XR-abnormal (z_j=1). The average utility for a policy family Π over $(\delta_\alpha, \delta_{\text{cov}})$ and strict pairs $\{1, \ldots, N\}$ is

$$U_\Pi(\lambda, \mu) \;=\; \max_{(\delta_\alpha, \delta_{\text{cov}})\in\mathcal{G}} \frac{1}{N} \sum_{j=1}^{N} u_j\big(d_j(\delta_\alpha, \delta_{\text{cov}}); \lambda, \mu, z_j\big),$$

where $\mathcal{G}$ is the swept grid. We plot $U_\Pi(\lambda, \mu)$ versus λ for fixed μ and compare against "acquire-all" and "acquire-none" baselines; this is analogous to net-benefit curves and exposes the radiation-safety trade-off transparently.

Implementation details

All images were resized to 512×512 with aspect-preserving padded crops; grayscale channels were replicated to three to match the ResNet stem. We trained SimSiam for 10 epochs per modality. For supervised heads, we used MAE for angle/coverage targets and cross-entropy for IHDI when labels were present; heads were trained on *post-train* only using a fixed schedule and evaluated on *evaluation*. The *calibration* split is reserved for affine bias correction and conformal calibration. Encoders remained frozen throughout, aligning with the limited-label regime. Conformal calibration swept miscoverage levels $\delta_\alpha, \delta_{\text{cov}} \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$ (target coverages $1 - \delta_\alpha$ and $1 - \delta_{\text{cov}}$). We report example radii for $\delta_\alpha = \delta_{\text{cov}} = 0.10$ in Results. Pair construction used strict (subject, date, side) matching; all reported pairwise metrics are computed on the subset with XR ground truth present.

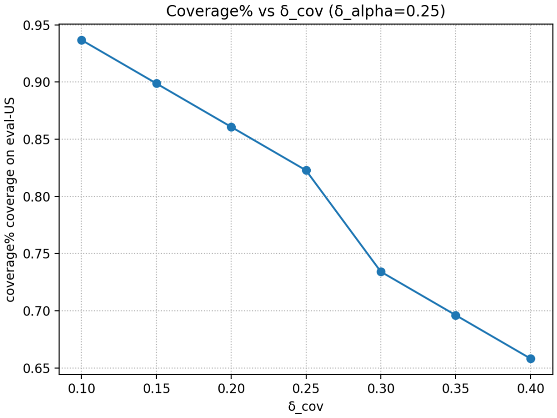

Figure 4: Empirical one-sided coverage for femoral head coverage versus miscoverage δ_{cov} (target coverage $1-\delta_{\text{cov}}$).

Table 1: Policy snapshots (strict eval pairs, N=77).

Rule	miscoverage $\delta_\alpha/\delta_{\text{cov}}$	US-only	XR use
AND	0.10 / 0.10	0.00	1.00
AND	0.20 / 0.20	0.00	1.00
OR	0.35 / 0.35	0.43	0.57
OR	0.40 / 0.40	0.55	0.45

Results

Probe Accuracy on Held-Out Evaluation

Frozen encoders with small measurement heads achieved low double-digit mean absolute error (MAE) on ultrasound and single-digit MAE on radiographs. On the strict evaluation split, ultrasound probes yielded α MAE $=$ 9.69°, β MAE $=$ 11.25°, and head-coverage MAE $=$ 13.97 percentage points. Radiographic probes achieved acetabular index (AI) MAE $=$ 7.60° and center-edge (CE) MAE $=$ 8.93°. These accuracies are competitive with unimodal reports that use larger bespoke networks, and they are achieved with frozen backbones and limited labels.

Conformal Calibration on Ultrasound

On the calibration split (7 subjects; 26 ultrasound images), we fitted per-target affine bias corrections then computed one-sided residual quantiles at miscoverage $\delta_\alpha = \delta_{\text{cov}} = 0.10$ (target coverage 0.90). For α we obtained q^+_α=10.75°; for coverage, q^+_{cov}=28.74 percentage points. Calibration MAEs were $\approx$ 6.4° (α) and $\approx$ 14.9 percentage points (coverage). These radii define conservative lower bounds used by the policies in the next sections.

Selective Imaging Policies: Coverage and Throughput

We evaluated three ultrasound-first deferral rules on the strict paired set (N=77 hips with any XR ground truth), sweeping conformal miscoverage levels $\delta_\alpha, \delta_{\text{cov}} \in \{0.10{:}0.40\}$ (target coverage $1-\delta$):

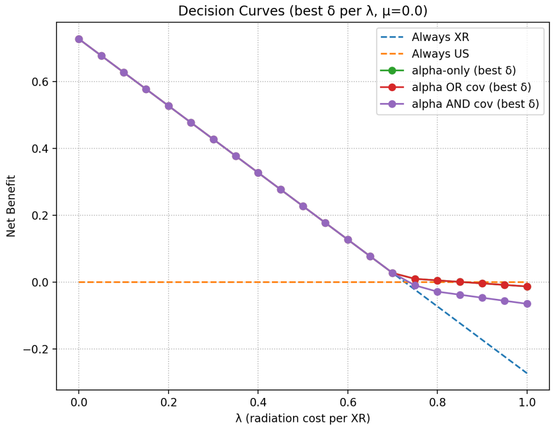

Figure 5: Decision-curve envelopes over radiation cost λ for miss penalty μ=0.0. When μ is low, ultrasound-only becomes higher utility when radiation cost λ is high. On the strict paired evaluation set, the OR policy (red) has higher utility than the AND policy (purple) at high radiation cost levels.

Alpha-only requests an X-ray unless the calibrated lower bound for α exceeds 60°. **Alpha OR coverage** requests an X-ray unless either the α lower bound exceeds 60° or the coverage lower bound exceeds 50%. **Alpha AND coverage** requests an X-ray unless both lower bounds exceed their thresholds.

Heatmaps of US-only rate (Fig. 2) and miss rate among US-only decisions (Fig. 3) expose the expected trade-offs. At conservative settings (e.g., δ_α=δ_{cov}=0.10), the AND rule keeps US-only rate near zero while reaching empirical coverage $\approx$ 0.90 for α and $\approx$ 0.94 for coverage on the evaluation US set. Permissive settings under the OR rule increase US-only throughput substantially (e.g., $\sim$ 0.43 at 0.35/0.35 and $\sim$ 0.55 at 0.40/0.40) with corresponding increases in miss-rate heatmap intensity.

Coverage as a Function of Conservatism

Figure 4 shows empirical one-sided coverage for femoral head coverage as a function of δ_{cov} (with δ_α fixed to the grid median), confirming monotone behavior with respect to miscoverage: empirical coverage decreases as δ_{cov} increases (equivalently, decreasing δ yields more conservative bounds and higher coverage). The same pattern holds for α (not shown to avoid redundancy).

Decision-Curve Analysis

We summarize the radiation-safety trade using decision-curve style utilities $U(\lambda, \mu)$ that penalize each radiograph with cost λ and each missed XR abnormality among US-only cases with penalty μ. Figure 5 shows decision-curve envelopes for miss penalty μ=0.0: for each radiation cost λ, we select the best $(\delta_\alpha, \delta_{\text{cov}})$ on our grid *within each rule family*. When radiation is costly (higher λ), more permissive policies that skip XR more often (higher US-only) tend to yield higher utility; as λ falls (XR becomes cheaper), more conservative operating points that defer to XR more frequently become preferable. (Larger μ values shift the optimum toward more conservative operating points; not shown here.)

Policy Snapshots for Reviewer Orientation

Table 1 lists a few interpretable operating points. Under AND at $0.10/0.10$, US-only rate is zero, XR utilization is unity, and empirical coverage is high (~ 0.90 for α; ~ 0.94 for coverage). Under OR at $0.35/0.35$ and $0.40/0.40$, US-only throughput increases to $\sim$ 0.43 and $\sim$ 0.55 while XR utilization drops accordingly. These snapshots match the contour gradients in the heatmaps.

Discussion

From measurements to decisions. Our goal is not to replace radiographs with ultrasound, but to formalize when ultrasound is enough and when an X-ray has positive value of information. The policy operates on named measurements that clinicians already use in practice, then converts them into calibrated lower bounds relative to clinical thresholds. This keeps the model on familiar ground and makes every decision traceable to a small number of interpretable quantities.

Collaborative decision support, not automation. The framework is designed to collaborate with sonographers, radiologists, and pediatric orthopedists. Clinicians retain control of thresholds and operating points through two dials. First, the rule family encodes clinical preference about how to combine evidence from α and femoral head coverage. Second, the miscoverage levels $(\delta_\alpha, \delta_{\text{cov}})$ translate appetite for risk into stricter or looser lower bounds: smaller δ targets higher coverage and yields more conservative bounds; larger δ is more permissive. Heatmaps in Figure 2 and Figure 3 are meant to be used as interactive guides during service setup. Teams can choose conservative points that preserve coverage and defer frequently to XR, or permissive points that reduce XR utilization while accepting a controlled increase in miss risk among US-only cases.

Radiation minimization as an explicit objective. Decision curves in Figure 5 make the radiation trade-off explicit. The utility weights (λ, μ) are not tuned by the model. They are handles for local policy. Sites that place a high cost on radiation (higher λ) will favor more permissive regions where US-only rates are higher; sites that place a lower cost on radiation (lower λ) can choose more conservative operating points with higher XR utilization. Sites that can tolerate a small increase in miss risk will move along the frontier to OR-rule settings that reduce X-ray utilization. This separates modeling from policy and puts radiation stewardship in the hands of the clinical team.

Safety properties and fail-safe behavior. The policy fails safe in two ways. First, one-sided conformal bounds give finite-sample (marginal) coverage control for the one-sided lower bounds. Empirical coverage trends in Figure 4 reflect the intended monotonicity as miscoverage level increases. Second, cases near the thresholds route to XR. The pairwise analysis shows that many borderline hips sit within a few degrees of the 60° alpha cutoff. Rather than stretching ultrasound to decide these, the system seeks a radiograph. This mirrors how experienced clinicians act under uncertainty.

Interpretability and auditability by design. The pipeline keeps the entire reasoning chain visible: line and point annotations, derived measurements, calibrated lower bounds, rule evaluation, and policy outcome. There are no hidden logits or opaque multi-class outputs gating safety-critical actions. This supports prospective quality assurance and after-action review. If a site wants to adjust a threshold, reweight the utility, or adopt ossification-aware criteria, the change reads like a guideline, not a re-training recipe.

How clinicians can use the figures. The three figure families provide complementary views for deployment. Heatmaps of US-only rate map operational throughput across $(\delta_\alpha, \delta_{\text{cov}})$. Miss-rate heatmaps bound downside risk on the subset of US-only decisions. Decision curves summarize net benefit across radiation costs and highlight policy dominance regions. Together they provide a conservative floor and a permissive ceiling, giving a spectrum of defensible operating points that align with local practice and regulatory constraints.

Strengths relative to prior art. Most prior work treats ultrasound and radiography in isolation, often with black-box classifiers. Our approach uses measurement-faithful heads on both modalities, aligns decisions with long-standing clinical thresholds, and introduces coverage-controlled selective prediction to turn a unimodal estimate into a cross-modal action. This retains clinical vocabulary, supports mixed cohorts with variable ossification, and exposes an explicit safety knob.

Generalization and subgroup reliability. Real deployments will face device heterogeneity, operator variability, and demographic shifts. The conformal layer can be recalibrated per site, per device family, or per age band. Subgroup coverage audits can detect drift and trigger automatic tightening (decreasing) of the miscoverage levels δ. Because the policy is measurement-based, these recalibrations are lightweight and do not require end-to-end re-training.

Clinical workflow integration. A practical path is ultrasound-first triage at the point of care. The model computes $\hat{\alpha}$ and $\widehat{\text{cov}}$, applies bias correction, and renders calibrated lower bounds with a simple traffic-light presentation: green for US-only above threshold, gray for defer to XR, and an explanation panel showing the margin to threshold and the chosen $(\delta_\alpha, \delta_{\text{cov}})$. Radiographs, when acquired, feed back into the registry and enlarge future calibration pools. This supports continuous learning while keeping the clinician in the loop.

Future work. Three extensions are natural. First, add modality-specific quality gates and age or ossification-aware thresholds so the rule respects validity domains automatically. Second, explore US$\rightarrow$XR surrogates that predict radiographic AI or IHDI risk from ultrasound measurements, then fold those calibrated risks into the decision curves. Third, run a prospective study that measures radiation saved per 100 infants, change in return-visit rates, and agreement with multidisciplinary adjudication. Each of these fits within the current policy scaffold without sacrificing interpretability.

Conclusion. This study shows that a small set of measurement-faithful probes, a calibration step that controls coverage, and a transparent policy grid are sufficient to build a collaborative assistant for selective imaging in DDH. The assistant is tunable to local radiation preferences, deferential near uncertainty, and legible to the clinicians who must own the decision. The figures and tables are intended as handrails for clinical setup and as a reproducible template for future multi-site validation.

References

Chen, T.; Zhang, Y.; Wang, B.; et al. 2022. Development of a fully automated Graf standard plane selection and measurement method for infant hip ultrasound. *Diagnostics*, 12(2): 488.

Chen, Y.-P.; Fan, T.-Y.; Chu, C.-J.; Lin, J.-J.; Ji, C.-Y.; Kuo, C.-F.; and Kao, H.-K. 2024. Automatic and human-level Graf's type identification for detecting developmental dysplasia of the hip. *Biomedical Journal*, 47(2): 100614.

Chen, Z.; Wang, X.; et al. 2024. Application of artificial intelligence in the diagnosis of developmental dysplasia of the hip based on imaging features: a systematic review. *Journal of Orthopaedic Surgery and Research*.

Den, H.; Koga, H.; Oda, T.; et al. 2023. Diagnostic accuracy of a deep learning model using YOLOv5 for developmental dysplasia of the hip. *Scientific Reports*, 13: 7383.

Dezateux, C.; and Rosendahl, K. 2007. Developmental dysplasia of the hip. *The Lancet*, 369(9572): 1541–1552.

Doski, J.; Qadir, R.; et al. 2022. An Upgrade of the International Hip Dysplasia Institute Classification for Developmental Dysplasia of the Hip. *Clinics in Orthopedic Surgery*, 14(1): 1–9.

Fraiwan, M.; Al-Kofahi, N.; Ibnian, A.; and Hanatleh, O. 2022. Detection of developmental dysplasia of the hip in X-ray images using deep transfer learning. *BMC Medical Informatics and Decision Making*, 22(216).

Graf, R. 1980. The diagnosis of congenital hip-joint dislocation by the ultrasonic compound treatment. *Archives of Orthopaedic and Traumatic Surgery*, 97(2): 117–133.

Graf, R.; Scott, S.; and Lercher, K. 2006. *Hip Sonography: Diagnosis and Management of Infant Hip Dysplasia*. Berlin: Springer, 2nd edition.

Hsu, W.-S.; Lin, Y.-C.; et al. 2025. Real-Time Ultrasound Diagnosis of Developmental Dysplasia of the Hip Using Deep Learning. *International Journal of Medical Sciences*, 22: 4236–4250.

Hu, X.; Wang, L.; Yang, X.; Zhou, X.; Xue, W.; Cao, Y.; Liu, S.; Huang, Y.; Guo, S.; Shang, N.; Ni, D.; and Gu, N. 2022. Joint Landmark and Structure Learning for Automatic Evaluation of Developmental Dysplasia of the Hip in Ultrasound. *IEEE Journal of Biomedical and Health Informatics*.

Keller, M.; and Nijs, E. 2009. The role of radiographs and ultrasound in developmental dysplasia of the hip: how good are they? *Pediatric Radiology*, 39(2): 211–220.

Kinugasa, K.; Kobayashi, A.; et al. 2023. Diagnosis of developmental dysplasia of the hip by ultrasound imaging using deep learning. *Journal of Pediatric Orthopaedics*.

Li, Q.; Zhong, L.; Huang, H.; et al. 2019. Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of Sharp's angle on standardized anteroposterior pelvic radiographs. *Medicine*, 98(51): e18500.

Liu, C.; Xie, H.; Zhang, S.; Mao, Z.; Sun, J.; and Zhang, Y. 2020. Misshapen Pelvis Landmark Detection With Local–Global Feature Learning for Diagnosing Developmental Dysplasia of the Hip. *IEEE Transactions on Medical Imaging*, 39(12): 3944–3954.

Moon, K.; Kim, Y.; Lee, H.; et al. 2024. Automated assessment of pelvic radiographs using deep learning: validation for multiple radiographic parameters. *Heliyon*. Article e-publication.

Narayanan, U.; Mulpuri, K.; Sankar, W.; Clarke, N.; Hosalkar, H.; and Price, C. 2015. Reliability of a New Radiographic Classification for Developmental Dysplasia of the Hip (IHDI classification). *Journal of Pediatric Orthopaedics*, 35(5): 478–484.

Omeroğlu, H. 2014. Use of ultrasonography in developmental dysplasia of the hip. *Journal of Children's Orthopaedics*, 8(2): 105–113.

Park, H.; Jeon, K.; Cho, K.; Kim, S.; and Hwang, H. 2021. Diagnostic performance of a convolutional neural network for detecting developmental dysplasia of the hip on anteroposterior pelvic radiographs. *Korean Journal of Radiology*.

Qi, G.; Zhang, H.; et al. 2025. A dataset for quality evaluation of pelvic X-ray and pelvic landmark detection (MTDDH). https://www.nature.com/articles/s41597-025-05146-x. Scientific Data article; pediatric pelvic radiograph dataset for DDH tasks.

Sharp, I. 1961. Acetabular Dysplasia: The Acetabular Angle. *Journal of Bone and Joint Surgery. British Volume*, 43-B(2): 268–272.

Shimizu, H.; Tsukagoshi, R.; et al. 2024. Bimodal machine learning model for unstable hips in infants: integration of radiographic images with automatically generated clinical measurements. *Scientific Reports*.

Tönnis, D. 1987. *Congenital Dysplasia and Dislocation of the Hip in Children and Adults*. Berlin: Springer.

van Kouswijk, H.; Yeung, M.; and Jaremko, J. 2025. Current and Emerging Applications of Artificial Intelligence in Pediatric Musculoskeletal Imaging. *Children*, 12(5): 645.

Wiberg, G. 1939. Studies on Dysplastic Acetabula and Congenital Subluxation of the Hip Joint: With Special Reference to the Complication of Osteoarthritis. *Acta Chirurgica Scandinavica (Supplement)*, 83(Suppl 58): 7–38.

Wu, X.; Wang, Y.; et al. 2023. Artificial intelligence for acetabular index measurement and DDH assessment: a systematic review and meta-analysis. *Frontiers in Pediatrics*, 11: 1094892.

Xu, L.; Zhao, Z.; Wang, L.; et al. 2022. Deep-Learning Aided Diagnostic System in Assessing Developmental Dysplasia of the Hip on Pediatric Pelvic Radiographs. *Frontiers in Pediatrics*, 10: 785480.

Zhang, Y.; Sun, J.; Zhang, S.; Li, C.; et al. 2020. Artificial intelligence-assisted diagnosis of paediatric developmental dysplasia of the hip using anteroposterior pelvic radiographs. *Bone & Joint Journal*, 102-B(11): 1650–1657.

Zieger, M. 1986. Ultrasound of the infant hip. Part 2. Validity of the method. *Pediatric Radiology*, 16: 488–492.