Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Comparing label-free quality-control scores for automated angle-of-progression measurement in intrapartum ultra- sound: a retrospective benchmark study

Liangji Zhou
The Sixth Affiliated Hospital of Sun Yat-sen University
Yuanqiu Chen
The Sixth Affiliated Hospital of Sun Yat-sen University
Tongtong Chen
The Sixth Affiliated Hospital of Sun Yat-sen University
Yijuan Huang
The Sixth Affiliated Hospital of Sun Yat-sen University
Qihao Pan
The Sixth Affiliated Hospital of Sun Yat-sen University
Si Li
The Sixth Affiliated Hospital of Sun Yat-sen University
Chenyu Gou
The Sixth Affiliated Hospital of Sun Yat-sen University
Meiqiong Fan
The Sixth Affiliated Hospital of Sun Yat-sen University
Yu Gao
gaoy57@mail.sysu.edu.cn

The Sixth Affiliated Hospital of Sun Yat-sen University

Research Article

Keywords: intrapartum ultrasound, angle of progression, quality control, selective prediction, ensemble disagreement, uncertainty estimation, landmark detection, cephalometry

Posted Date: October 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11202847/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Comparing label-free quality-control scores for automated angle-of-progression measurement in intrapartum ultrasound: a retrospective benchmark study

Liangji Zhou[1,2], Yuanqiu Chen[1,2], Tongtong Chen[1,2], Yijuan Huang[1,2], Qihao Pan[1,2], Si Li[1,2], Chenyu Gou[1,2], Meiqiong Fan[1,2] and Yu Gao[1,2,*]

[1]Department of Obstetrics, The Sixth Affiliated Hospital of Sun Yat-sen University, No. 26 Yuancun Second Cross Road, Guangzhou 510655, Guangdong, China
[2]Biomedical Innovation Center, The Sixth Affiliated Hospital, Sun Yat-sen University, No. 26 Yuancun Second Cross Road, Guangzhou 510655, Guangdong, China
[*]Corresponding author: Yu Gao, gaoy57@mail.sysu.edu.cn

Abstract

Background: Automated angle-of-progression (AoP) measurement from intrapartum ultrasound is accurate on average but occasionally fails by tens of degrees without warning. We compared label-free reliability scores for identifying measurements to refer for manual review, at what workload, and where they fail.

Methods: Retrospective secondary analysis of the public IUGC2025 benchmark (501 test files, 493 distinct images, two file-name prefix groups without a patient mapping). Six landmark detectors (three HRNet-W48, three ConvNeXt-V2) gave a coordinate-wise median prediction. On that prediction we compared AoP-space and coordinate disagreement between models, single-model input consistency, the ensemble heatmap peak (E-MHA) and heatmap confidence, with the new comparisons pre-specified: retained error at 90% coverage (primary), residual failure rates above 10, 15 and 20° at review budgets of 2–20%, validation-to-test threshold transfer, and an audit of every missed failure. A three-model ensemble on the held-out ISBI 2015 cephalometric Test2 split served as replication.

Results: Mean AoP error was 4.62°. Referring the 10% most-disagreed frames lowered the retained error to 3.97° (coordinate disagreement 3.99°, difference 0.02°, 95% CI −0.10 to 0.10; input consistency 3.96°; confidence 4.21°; E-MHA 4.25°). Referring 5% of frames caught five

of six errors above 20° with either disagreement score (noisy-student input consistency three, confidence four), lowering the residual rate of such errors from 1.2% to 0.2%, but for errors above 10° only from 9.4% to 7.8% with AoP-space and 7.4% with coordinate disagreement. Validation thresholds flagged 11% of test frames instead of the intended 5% for AoP disagreement (6% for coordinate disagreement). The missed error above 20° was an underestimation shared by all six detectors with little spread between them. On cephalometric Test2, disagreement lowered the retained error by 0.28 mm (95% CI 0.13 to 0.46) at 90% coverage using the averaged reference, with similar reductions using either annotator separately.

Conclusions: In this frame-level benchmark, ensemble disagreement referred most gross AoP failures at a 5% review workload but could miss errors shared by all detectors; no advantage of angle-space over coordinate-space disagreement was detected, and thresholds needed local calibration. These exploratory results do not establish patient-level safety or clinical benefit.

Keywords: intrapartum ultrasound; angle of progression; quality control; selective prediction; ensemble disagreement; uncertainty estimation; landmark detection; cephalometry

1 Background

The angle of progression (AoP) is measured on transperineal ultrasound between the long axis of the pubic symphysis and a line from its inferior edge tangential to the fetal skull, and it increases as the fetal head descends; its acquisition and use in labour are described in the practice guidelines of the International Society of Ultrasound in Obstetrics and Gynecology [1]. Manual measurement requires three landmarks to be placed on a moving image and depends on the operator, which has motivated automated landmark-based measurement. The Intrapartum Ultrasound Grand Challenge (IUGC) benchmarks provide public labelled data for this task [2–6]. On these benchmarks the mean error of automated AoP measurement is a few degrees [5,7], but the error distribution has a long tail: a small fraction of measurements is wrong by tens of degrees, and nothing in the output marks them.

For a measurement that informs a clinical decision, the practical question is therefore not only how accurate a system is on average but which of its outputs should not be trusted.

Selective prediction formalises this as refusing the least reliable fraction of inputs and reporting the error on the rest [8]; in a clinical workflow the refused measurements would be referred for manual review. The refusal needs a reliability score that is available without the reference label. Candidate scores differ in what they measure and what they cost: the peak value of a detector's heatmap; the consistency of a single detector under input perturbations; and the disagreement among an ensemble of detectors, which is the strongest simple baseline in benchmarks of medical failure detection [9–11] and has served as label-free quality control for segmentation [12]. For heatmap-based landmark localisation, Schobs et al. ranked predictions by the spread of the ensemble's coordinate predictions (E-CPV) and by the peak of its averaged heatmap (E-MHA), and showed that discarding the most uncertain predictions removes gross errors [13]; in ultrasound landmark video, single-model uncertainty has been used to select reliable frames [14].

Three questions remain open for a derived measurement such as the AoP. First, whether disagreement should be computed on the measurement itself or on the landmark coordinates from which it is derived: the AoP is a nonlinear function of three landmarks, so coordinate spread and angle spread need not rank frames in the same order. Second, what a given review budget buys: how many large errors are caught, how many remain among the accepted measurements, and how the answer depends on what counts as large. Third, whether a threshold chosen on development data keeps its intended workload on new data, and which failures the scores miss. An alternative to referral, choosing the best of several candidate predictions for each image without labels, fails on the detector pool studied here: no label-free selection or aggregation rule we tested recovered more than 3.1% of what a label-aware per-frame oracle could, or 5.8% when a diffusion refiner's prediction was added as a seventh candidate (Additional file 1, Sections S19, S20 and S8), which is why we study abstention rather than selection.

We therefore report a retrospective comparative evaluation of label-free reliability scores for automated AoP measurement on the public IUGC2025 benchmark, with the new comparisons fixed in a dated analysis plan and a replication on a held-out split of a second landmark dataset. The aim is not a new detector but an account of what quality control built on these scores can and cannot do: the retained error at fixed coverage, the residual failure rate at a fixed review

workload, the transfer of thresholds from validation to test data, and an audit of the failures that are missed.

2 Methods

2.1 Study design and data

This is a retrospective secondary analysis of two public, de-identified benchmark datasets. It follows earlier, unpublished analyses of the same detector pools, which had already reported, on the IUGC2025 test split, the retained error and the recall of gross failures for AoP-space disagreement, heatmap confidence and noisy-student input consistency, and the disagreement result on the cephalometric Test2 split. A dated analysis plan written after those analyses fixed, before any of the new reliability scores was computed, the primary endpoint and its decision rule, the new scores (coordinate disagreement, E-MHA and input consistency on the validation frames), the review-workload grid, the fixed-threshold protocol, the failure audit, the reference sensitivity analysis and the cost comparison. The plan was not deposited in a public registry; the plan, its hash, the results known when it was written and the deviations from it are reported in Additional file 1, Section S1. The study design is summarised in Fig. 1.

The IUGC2025 benchmark [5, 6] provides transperineal ultrasound frames labelled with three landmarks: the two endpoints of the pubic symphysis and the tangent point on the fetal head, from which the AoP is computed. The release contains 300 labelled training frames, 33,466 unlabelled training frames, 100 validation frames and 501 test files. File names carry a numeric prefix that is renumbered within each split and is not documented as a patient, clip or centre. The validation frames carry prefixes 2 (45 frames) and 4 (55), and the test files prefixes 1 (84) and 4 (417); the numbering of prefix 4 interleaves the validation and test frames. The test split holds 493 distinct images: eight files repeat other files byte for byte (all with prefix 1) but carry different annotations, whose AoP differs by up to 5.3°. The challenge design states that several raters independently annotated a subset of the test cases [6], but the release does not identify them. The split that the release calls validation served for tuning and checkpoint selection (tuning data in the terminology of CLAIM [15]); every reported evaluation uses the

test split. All IUGC2025 statistics count files unless stated otherwise; analyses over distinct images are sensitivity analyses.

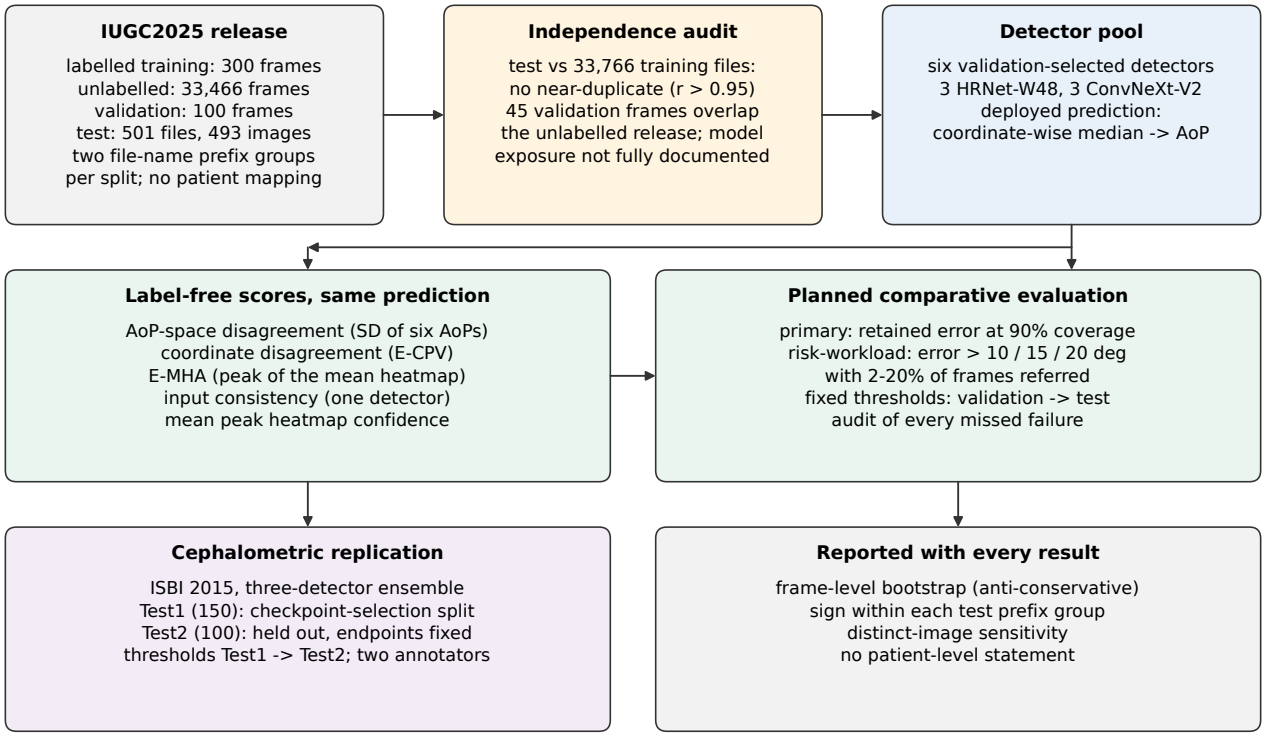

Figure 1: **Study design.** Data, independence audit, detector pool and deployed prediction, the label-free scores compared on that prediction, the evaluation fixed in the analysis plan, and the cephalometric replication. All statistics are frame-level; the IUGC2025 release provides no patient mapping.

The ISBI 2015 cephalometric dataset [16] provides 400 lateral cephalograms, each with 19 landmarks annotated by a senior and a junior orthodontist; the reference is their average. The cephalometric detectors described below were selected on the Test1 split (150 images), so Test1 is development data; the Test2 split (100 images) was excluded from model selection, and its initial evaluation used endpoints specified in an earlier dated script. Subsequent analyses reused the stored predictions (Additional file 1, Section S22).

2.2 Independence audit

We screened every validation and test frame against all 33,766 training files, labelled and unlabelled, by the Pearson correlation of 48×48 grayscale thumbnails, flagging pairs above 0.95, the construction used for the screen of the earlier IUGC2024 release (Additional file 1,

Section S16). The byte-identical duplicate pairs inside the test split served as positive control, and the correlations between test frames and 2,000 random training files as the null band. Because one detector was trained on pseudo-labels of unlabelled frames, we also counted, in every pseudo-label pool derived from the release that we could inspect, the frames drawn from unlabelled clips that share their identifier with a validation frame. Every flagged pair between 0.90 and 0.95 was inspected visually.

2.3 Detector pool and deployed prediction

The IUGC2025 pool comprises six heatmap-based landmark detectors trained on the labelled training frames and selected on the validation split. Three use an HRNet-W48 backbone [17]: a supervised detector; a detector fine-tuned with a circular AoP loss and exponential weight averaging; and a noisy-student detector trained on pseudo-labels of unlabelled frames produced by the supervised one. Three use a ConvNeXt-V2-Base backbone [18]: two random seeds of one supervised recipe, and one retrained with 1,000 additional synthetic images from a conditional flow-matching generator fitted to the labelled training frames. An audit of every checkpoint's saved selection metric found no selection on a test metric. The deployed prediction for each frame is the coordinate-wise median of the six landmark configurations, from which the AoP is computed; its absolute difference from the reference AoP is the error on which every score is judged. Inference used bfloat16 autocast in batches of four frames. In a first re-run on the analysis server, all six detectors reproduced the stored coordinates exactly (maximum difference 0.000 px over 601 frames), and a batch of two frames instead moved one coordinate by 1.9 px, so the batch size is part of the inference protocol; after the server had been re-rented, the same protocol no longer reproduced the coordinates bit for bit (Results) (Additional file 1, Section S2).

The cephalometric pool comprises three detectors trained on the 150 training images from ImageNet-initialised backbones: HRNet-W48, ConvNeXt-V2-Base and a lightweight `hrnet_w18_small_v2` network (Additional file 1, Sections S16 and S22). Its deployed prediction is the mean of the three configurations and its error the mean radial error (MRE) over the 19 landmarks, in millimetres.

2.4 Label-free reliability scores

All scores are computed from the predictions alone; a higher score means a less reliable frame. *AoP-space disagreement* is the standard deviation of the six per-detector AoPs. *Coordinate disagreement* follows the E-CPV construction of Schobs et al. [13]: for each landmark, the mean Euclidean distance of the six detector coordinates from their mean, averaged over the three landmarks (the maximum over landmarks is a secondary form). *E-MHA* [13] averages the detectors' heatmaps and takes the peak: because the six detectors produce heatmaps at different resolutions, each sigmoid heatmap was resampled to a common 128×128 grid and divided by its spatial sum before averaging, and the score is minus the mean over landmarks of the peak of the averaged map; averaging the raw sigmoid maps is a secondary form. *Input consistency* is a single-detector score: the mean absolute change in the AoP of the noisy-student detector over eight Gaussian perturbations of its input (standard deviation 0.05 on the normalised tensor), which costs one clean and eight perturbed forward passes; this comparison was pre-specified in an earlier analysis, and the consistency scores of the other five detectors are reported as secondary (Additional file 1, Table S35). *Heatmap confidence* is minus the mean over detectors of the average peak sigmoid value. *Random refusal* is reported as its analytical expectation. On the cephalometric pool, disagreement is the mean over the 19 landmarks of the Euclidean standard deviation of the three detectors' coordinates, as in the earlier analysis of this pool; the E-CPV form is secondary.

2.5 Evaluation

Coverage. For each score, the most suspicious frames are refused until a fraction c of frames remains, and the mean error of the retained frames is reported (ties are refused in frame order). The primary endpoint fixed in the plan is the retained error at 90% coverage of the deployed prediction ranked by coordinate disagreement minus the same ranked by AoP-space disagreement. The plan fixed three readings: an interval above zero would favour the angle, an interval below zero the coordinates, and an interval including zero would be reported as no detectable difference. The comparison of input consistency against AoP-space disagreement at 90% coverage is reported as pre-specified earlier.

Review workload. A failure is an absolute AoP error above 10, 15 or 20°. These are research thresholds for measurements that would change a clinical reading, not outcome-validated limits. For review budgets of 2, 5, 10 and 20% of frames (the k most suspicious frames, $k = 10$, 25, 50 and 100), we report the failures caught, the precision of referral and the residual failure rate, the fraction of the retained frames that are failures. For the 5% and 10% budgets we cross-tabulated the failures caught by AoP-space disagreement and by input consistency.

Fixed thresholds. A deployed system needs a score threshold rather than a rank. For each score, the $(1 - q)$ quantile of its values on the 100 validation frames was applied unchanged to the test frames, for $q = 5$, 10 and 20%, a frame being flagged when its score exceeded the threshold, and the realised fraction of flagged test frames was reported overall and for each test prefix group. Because all 45 prefix-2 validation frames overlapped the unlabelled release (Results), the calibration was repeated using the 55 prefix-4 validation frames, for which this overlap was not detected. Uncertainty in the threshold was estimated by bootstrapping the calibration frames. On the cephalometric pool, Test1 quantiles were applied to Test2.

Failure audit. Every failure above 20°, every frame flagged at the 5% budget and every failure missed at that budget were inspected: per-landmark radial error, the landmark dominating the error, the six per-detector AoPs, and an overlay of the image with the reference, the deployed prediction and the six detectors.

Reference sensitivity. The cephalometric Test2 endpoints were recomputed with the senior, the junior and the averaged annotation as reference, after confirming that the average of the raw annotations reproduces the reference used throughout (maximum difference 0.000 px).

System cost. We compared four systems: one noisy-student detector alone; the same detector with its input-consistency score; the three-detector subset chosen on the validation frames by the lowest retained error at 90% coverage of its own median prediction ranked by its own AoP-space disagreement; and the six-detector pool. Latency per frame was measured on one NVIDIA RTX 4090 with bfloat16 autocast, batch size 1, 20 warm-up and 200 timed iterations

with CUDA synchronisation, together with peak allocated GPU memory with only that system's detectors loaded. The accuracy side of each system was computed from the stored test predictions.

Analyses added after review. Two analyses were added after a pre-submission review and were not in the analysis plan. First, the evaluation was repeated for the three detectors whose configurations specify supervised training on the labelled frames alone from ImageNet-pretrained backbones, without pseudo-labels, synthetic images or a documented warm start (the supervised HRNet-W48 detector and the two ConvNeXt-V2 seeds; the label-only subset), and, separately, with the fine-tuned HRNet-W48 detector added, which according to its configuration was fine-tuned from the supervised HRNet-W48 checkpoint, a warm start that its run directory does not record. Each subset formed its own median prediction and was judged on its own failures. Second, the six detectors were re-run under four inference settings (bfloat16 or float32, batch size 4 or 1) to measure how the prediction, its failures and the rankings depend on the inference protocol, and heatmap confidence was also computed from the peak logit, which does not saturate.

2.6 Statistical analysis

Intervals are 95% percentile intervals over 2,000 frame-bootstrap resamples; comparisons between two scores are paired, with both scores re-ranked inside each resample. Frames within a prefix group are correlated, so these intervals are anti-conservative, the more so because 417 of the 501 test files carry one prefix. With two groups no cluster interval is possible (a bootstrap over two clusters admits three distinct resamples); instead, the sign of the primary contrast is reported within each prefix group. For the cephalometric replication, intervals are image-bootstrap intervals, and a one-sided permutation test (1,000 permutations of the score) was added for the 90% coverage gap. For the cephalometric recall and precision intervals, worst-decile membership was fixed from the full sample, whereas the disagreement ranking and the referral set were recomputed within each resample. Secondary analyses are reported without correction for multiplicity. Positive and negative controls preceded every analysis: each script

had to reproduce the values of the earlier analyses of this pool (for example the mean error of 4.615° and the retained error of 3.969° at 90% coverage) before any new number was written, and a permuted score had to rank at chance. The analysis scripts are named in Additional file 1. The analysis, plotting and verification scripts were drafted with the assistance of Claude Code (Anthropic), a large-language-model coding assistant, and OpenAI Codex was used for internal audits; the authors reviewed and ran every script and checked each reported number against the stored result files.

3 Results

3.1 Data audit and detector pool

No test frame had a near-duplicate among the 33,766 training files: the highest thumbnail correlation of any test frame was 0.93, against 1.000 for the known duplicate pairs and a 99.99th percentile of 0.824 for random pairs. Five test frames scored between 0.90 and 0.93. Visual inspection showed different acquisitions of similar appearance from the same scanner geometry; for the two most similar pairs the mean absolute pixel difference at full resolution was 7.8 and 9.3 grey levels, against 0.74 between adjacent frames of one clip (Additional file 1, Section S2). Without a patient mapping, the same patient cannot be excluded.

The validation split was not independent of the unlabelled release used for self-training. All 45 prefix-2 validation frames correspond, up to re-encoding, to frames of the unlabelled clip with the same identifier (thumbnail correlation 1.000), and 13 of the 14 pseudo-label pools derived from the release that we could inspect contain frames from these clips. The pseudo-label file used by the noisy-student detector was not retained, so its exposure to these clips cannot be fully reconstructed. Consistent with exposure, the noisy-student detector improved on its supervised teacher by 1.43° on the prefix-2 validation frames and by 0.38° on the prefix-4 frames, so its validation-based selection was probably optimistic. For the test split, on which every result below is computed, the screen found no near-duplicate image, but without a patient mapping its independence from the training data at the patient level cannot be established.

3.2 Accuracy and failures

The deployed median prediction had a mean AoP error of 4.615° over the 501 test files (4.655° over the 493 distinct images). The errors exceeded 10, 15 and 20° on 47 (9.4%), 13 (2.6%) and 6 (1.2%) files. A per-frame oracle that picks the best detector with knowledge of the reference reached 2.28°, a bound that no label-free selection rule approached (Additional file 1, Sections S19, S20 and S7).

3.3 Comparison of reliability scores

Ranking by AoP-space disagreement and refusing 10% of frames lowered the retained error from 4.615 to 3.969° (Table 1, Fig. 2a). Coordinate disagreement retained 3.986°. The primary contrast, coordinate minus AoP-space disagreement, was +0.017° (95% CI −0.101 to 0.099), with the same sign in both prefix groups (+0.102 and +0.066°) but the opposite sign over the 493 distinct images (−0.015°); no difference between the two scores was detected at this endpoint. The secondary measures were mixed: coordinate disagreement had the larger area under the curve for detecting the worst decile of frames (0.775 against 0.741) and the lower retained error at 50% coverage (3.308 against 3.548°), whereas AoP-space disagreement retained slightly less error at 80 and 70% coverage (3.751 against 3.761° and 3.621 against 3.636°). Input consistency retained 3.957° at 90% coverage (difference from AoP-space disagreement −0.013°, 95% CI −0.175 to 0.343) but levelled off below that coverage (3.889° at 50%). Heatmap confidence retained 4.210°. E-MHA did not rank the errors: it retained 4.247° at 90% coverage, and its area under the curve for the worst decile was 0.457; its raw-sigmoid form retained 4.152°.

3.4 Review workload and residual failures

At a 5% review budget (25 frames), either disagreement score caught five of the six failures above 20°, heatmap confidence and E-MHA four and input consistency three (Table 2, Fig. 3). With AoP-space disagreement, one failure above 20° remained among the 476 retained frames, a residual rate of 0.2% (95% CI 0.0 to 0.6%) against 1.2% under random referral; 20 of the 25 referred frames were not failures by this definition. A 10% budget caught all six with either

Table 1: **Retained AoP error and ranking quality of each label-free score on the same prediction**

	Retained error (°) at coverage					
Score	90%	80%	70%	50%	AUC[c]	ρ
AoP-space disagreement	3.969	3.751	3.621	3.548	0.741	0.242
Coordinate disagreement (E-CPV)	3.986	3.761	3.636	3.308	0.775	0.298
maximum over landmarks[a]	4.026	3.820	3.612	3.306	0.770	0.313
Input consistency (one detector)	3.957	3.873	3.857	3.889	0.718	0.097
E-MHA, normalised maps	4.247	4.167	4.206	4.366	0.457	−0.035
raw sigmoid maps[a]	4.152	4.220	4.138	3.923	0.642	0.118
Mean peak confidence	4.210	4.207	4.277	4.176	0.591	0.097
Random (expected)	4.615	4.615	4.615	4.615	0.500	0
Oracle rank (uses labels)[b]	3.405	2.808	2.339	1.587	–	–

IUGC2025 test split, 501 files; the deployed median prediction is the same in every row, only the order of refusal changes. Retained error: mean absolute AoP error of the frames kept after refusing the most suspicious ones. ρ: Spearman rank correlation of score and error. [a]Secondary form. [b]Refuses by the true error; a bound that no label-free score can reach, not a comparator. [c]Area under the receiver operating characteristic curve for detecting the worst decile (the 50 highest-error frames). Primary contrast (fixed in the analysis plan): coordinate minus AoP-space disagreement at 90% coverage, +0.017° (95% CI −0.101 to 0.099).

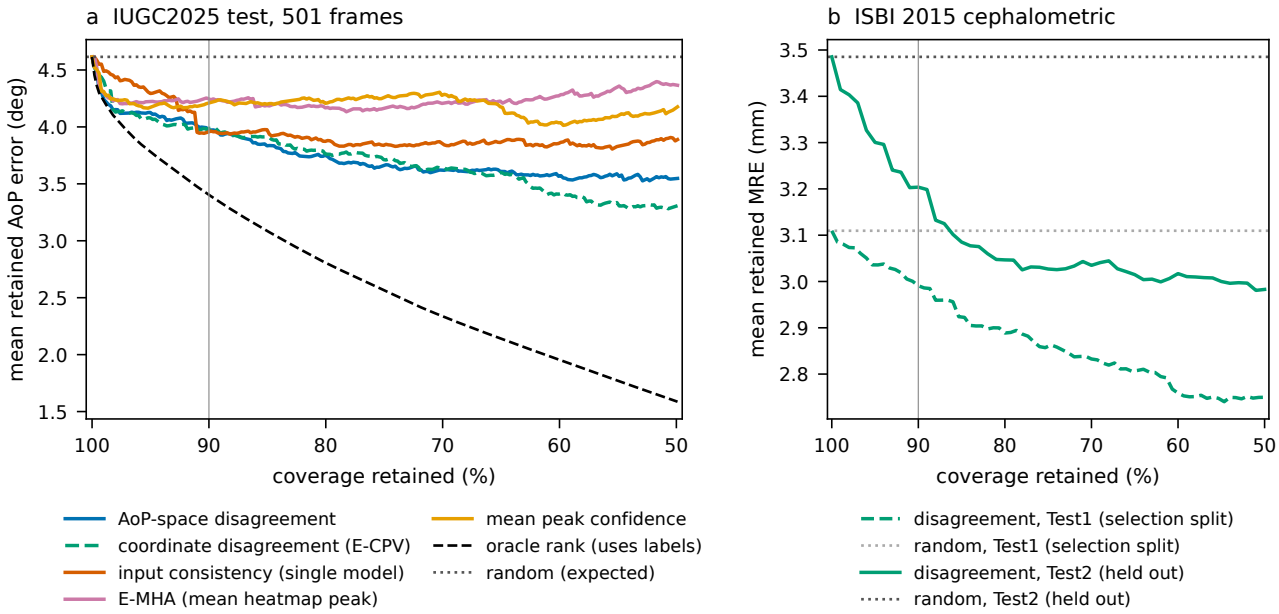

Figure 2: **Retained error as frames are refused in the order of each label-free score. a** IUGC2025 test split (501 frames): mean AoP error of the retained frames of the deployed median prediction against the fraction of frames retained. The dotted line is the expected value under random refusal; the black dashed curve refuses frames by their true error and is a bound, not a method. The vertical line marks the primary 90% coverage. **b** ISBI 2015 cephalometric replication with a three-detector ensemble: retained mean radial error when images are refused by cross-model disagreement, on Test1 (dashed, the split on which the detectors were selected) and on the held-out Test2 (solid), each with its random-refusal expectation (dotted).

disagreement score and five with input consistency. The gain shrank as the failure threshold was lowered. For errors above 15°, AoP-space and coordinate disagreement caught 7 and 8 of 13 at the 5% budget and left residual rates of 1.3 and 1.1%, against 2.6% under random referral. For errors above 10° they caught 10 and 12 of 47 and left 7.8 and 7.4%, against 9.4%. At a 10% budget input consistency caught the most failures above 10° (20 of 47, against 18 for either disagreement score).

The two kinds of score overlapped but were not interchangeable. At the 5% budget, AoP-space disagreement and input consistency both caught three of the six failures above 20°, disagreement alone caught two, consistency alone none, and one was missed by both. For failures above 10°, each caught failures the other missed (8 by both, 2 by disagreement only, 4 by consistency only), and 33 were missed by both. Counted over distinct images, the results were unchanged: five of six failures above 20° at 5% and all six at 10% with AoP-space disagreement.

Table 2: **Failures caught and residual failure rate at 5% and 10% review budgets**

	5% referred (25 frames)			10% referred (50 frames)		
Score	>20°	>15°	>10°	>20°	>15°	>10°
AoP-space disagreement	5/6 (0.2)	7/13 (1.3)	10/47 (7.8)	6/6 (0.0)	9/13 (0.9)	18/47 (6.4)
95% CI, residual	[0.0, 0.6]	[0.4, 2.3]	[5.3, 10.1]	[0.0, 0.2]	[0.0, 1.8]	[4.4, 9.1]
Coordinate disagreement	5/6 (0.2)	8/13 (1.1)	12/47 (7.4)	6/6 (0.0)	10/13 (0.7)	18/47 (6.4)
Input consistency	3/6 (0.6)	7/13 (1.3)	12/47 (7.4)	5/6 (0.2)	10/13 (0.7)	20/47 (6.0)
E-MHA	4/6 (0.4)	6/13 (1.5)	8/47 (8.2)	4/6 (0.4)	7/13 (1.3)	9/47 (8.4)
Mean peak confidence	4/6 (0.4)	6/13 (1.5)	9/47 (8.0)	4/6 (0.4)	6/13 (1.6)	11/47 (8.0)
Random (expected)	0.3/6 (1.2)	0.6/13 (2.6)	2.3/47 (9.4)	0.6/6 (1.2)	1.3/13 (2.6)	4.7/47 (9.4)

IUGC2025 test split, 501 files. Each cell: failures among the referred frames / all failures at that threshold (residual failure rate, the percentage of the retained frames whose error exceeds the threshold). Intervals: frame bootstrap, shown for the first row; all rows in Additional file 1, Section S3. Random: analytical expectation. The 2% and 20% budgets are in Additional file 1.

3.5 Fixed thresholds

Thresholds estimated on the validation frames did not reproduce the intended workload for every score (Table 3). The threshold of AoP-space disagreement intended to flag 5% of frames flagged 11.2% of the test frames (15.5% in prefix group 1, 10.3% in group 4) and still caught all six failures above 20°, at more than twice the planned workload. The intended 5% lies within the range of realised rates produced by resampling the 100 calibration frames alone (95% interval 3.8 to 14.6%), because a 5% threshold rests on the few highest validation scores;

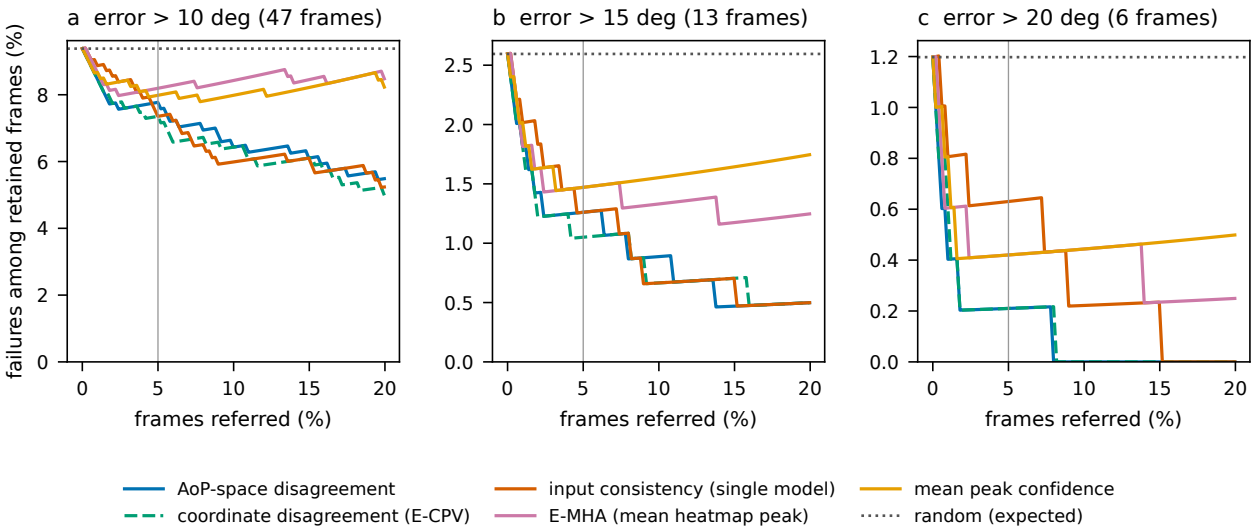

Figure 3: **Residual failure rate against the fraction of frames referred for review.** Each panel defines a failure as an absolute AoP error above 10° (**a**), 15° (**b**) or 20° (**c**); the number of failures among the 501 test frames is given in the panel title. Curves show the percentage of the retained frames that are failures when the given percentage of frames, the most suspicious under each score, is referred. The dotted line is the expectation under random referral; the vertical line marks the 5% budget. Coordinate disagreement is dashed so that it remains visible where it coincides with AoP-space disagreement.

the data therefore do not separate a difference between the splits from the imprecision of the calibration. Coordinate disagreement and input consistency came closer to their targets (6.4 and 7.8% at a nominal 5%). The E-MHA threshold for 20% flagged 42.1%. Heatmap confidence flagged 3.6, 7.4 and 17.8% at nominal 5, 10 and 20%. Calibrating on the 55 prefix-4 validation frames only, for which overlap with the unlabelled release was not detected, AoP-space disagreement flagged 12.6% (95% interval 2.8 to 23.2%), 21.4% and 28.1% at nominal 5, 10 and 20% (Additional file 1, Section S4). On the cephalometric pool, Test1 thresholds flagged 4.0, 8.0 and 20.0% of the Test2 images at nominal 5, 10 and 20%.

3.6 Detector subsets trained on labelled frames only

The label-only subset of three detectors, pooled as a system of its own, had a mean error of 4.773° and eight failures above 20°, including the six of the full pool. At a 5% budget, AoP-space and coordinate disagreement referred five and six of these eight (0.4 expected at random), and at 10% six and seven; refusing 10% of frames lowered the retained error from 4.773 to 4.295 and 4.187°, and the primary contrast was −0.107° (95% CI −0.531 to 0.190).

Table 3: **Realised referral rate of thresholds fixed on validation data**

	Nominal 5%		Nominal 10%		Nominal 20%	
Score	flagged %	>20°	flagged %	>20°	flagged %	>20°
AoP-space disagreement	11.2 (15.5/10.3)	6/6	13.8 (17.9/12.9)	6/6	21.8 (32.1/19.7)	6/6
Coordinate disagreement	6.4 (8.3/6.0)	5/6	9.8 (10.7/9.6)	6/6	18.0 (17.9/18.0)	6/6
Input consistency	7.8 (4.8/8.4)	4/6	12.6 (11.9/12.7)	5/6	21.4 (21.4/21.3)	6/6
E-MHA, normalised maps	9.8 (1.2/11.5)	4/6	22.6 (8.3/25.4)	5/6	42.1 (32.1/44.1)	5/6
raw sigmoid maps[a]	4.8 (3.6/5.0)	4/6	10.8 (9.5/11.0)	4/6	19.0 (14.3/19.9)	4/6
Mean peak confidence	3.6 (0.0/4.3)	4/6	7.4 (8.3/7.2)	4/6	17.8 (16.7/18.0)	4/6
Cephalometric, Test1 → Test2	4.0	4/10	8.0	6/10	20.0	9/10

IUGC2025: thresholds are quantiles of each score over the 100 validation frames, applied unchanged to the 501 test files; flagged %: realised percentage of test files flagged, overall (prefix group 1 / prefix group 4); >20°: failures above 20° among the flagged files. Cephalometric: quantiles over Test1 applied to Test2; the second column counts the worst-decile images (10) among the flagged ones. A file is flagged when its score exceeds the threshold. Threshold values, calibration on prefix-4 validation frames only and bootstrap intervals: Additional file 1, Section S4. [a]Secondary form.

With the fine-tuned HRNet-W48 detector added, the subset had four failures above 20°; either score referred two of them at 5% and all four at 10%, and the primary contrast was −0.002° (−0.116 to 0.108). The AoP-space threshold for a nominal 5% flagged 6.0% of the test frames for the three-detector subset and 10.6% for the four-detector subset (Additional file 1, Section S6).

3.7 Inference protocol and forms of confidence

Re-running the six detectors with the stored protocol on another machine did not reproduce the stored coordinates bit for bit: the median difference was 0.010 px, but 352 of the 9,018 landmark positions moved by more than 0.1 px (at most 5.6 px), and the deployed AoP of a frame changed by up to 1.3°. Batch size 1 and float32 inference moved the prediction further (float32: nearly every landmark position, by up to 32.8 px; the deployed AoP by up to 3.1°). Numerical identity was not achieved, but the endpoints we checked did not change materially: the mean error ranged from 4.558 to 4.618°, every setting had six failures above 20°, AoP-space disagreement referred five of them at 5% and all six at 10%, the frames referred at 5% were identical in three of the four settings, the retained error at 90% coverage ranged from 3.933 to 3.981° (stored predictions 3.969°), and the primary contrast ranged from −0.029 to +0.008° with every 95% CI including zero (stored predictions +0.017°). Computing heatmap confidence from the peak logit, which does not saturate, ranked the worst decile better than

the peak sigmoid (area under the curve 0.73 against 0.59) but referred one rather than four of the six gross failures at 5%, and its retained error at 90% coverage did not differ (4.269 against 4.216° in float32; difference 0.053°, 95% CI −0.245 to 0.525). E-MHA was unchanged in float32 (4.244 against 4.247°; Additional file 1, Section S6).

3.8 Failure audit

In each of the six failures above 20° the error was dominated by the fetal-head tangent point (Fig. 4). In five of them the detectors disagreed widely about that point, and AoP-space disagreement ranked the frames 1st, 2nd, 3rd, 5th and 9th of 501. A shared direction of error did not by itself hide a failure: in the frame with the largest error (84.2°), all six detectors underestimated the angle, but their spread ranked it 3rd. The sixth failure, missed at the 5% budget, combined a shared error with little spread: the six detectors agreed within 15° (123.7 to 138.6°) while the reference was 154.2°, so that frame ranked 40th and was referred only at the 10% budget.

The same pattern appeared across the test split in an analysis that was not pre-specified (Fig. 5). The deployed prediction overestimated small reference angles and underestimated large ones (correlation between signed error and reference AoP −0.63). The twelve frames with a reference above 140° were underestimated by 14.8° on average between 140 and 150° ($n = 7$) and by 29.0° above 150° ($n = 5$), and only three of the twelve were referred at the 5% budget. Part of this correlation is expected from noise in the reference itself, which regresses the error towards the reference values. Among the 37 failures above 10° missed at the 5% budget, 65% were underestimates, and their median AoP-space disagreement was 2.9°, against 17.6° for the ten that were caught.

3.9 Cephalometric replication

On the held-out Test2 split, refusing the 10% most-disagreed images lowered the retained MRE by 0.281 mm (95% CI 0.135 to 0.460) against random refusal, and by 0.439 mm at 80% coverage (Fig. 2b); none of 1,000 permutations of the score reached the observed gap (one-sided $P = 0.001$). At a 5% budget all five referred images belonged to the worst decile, which is the

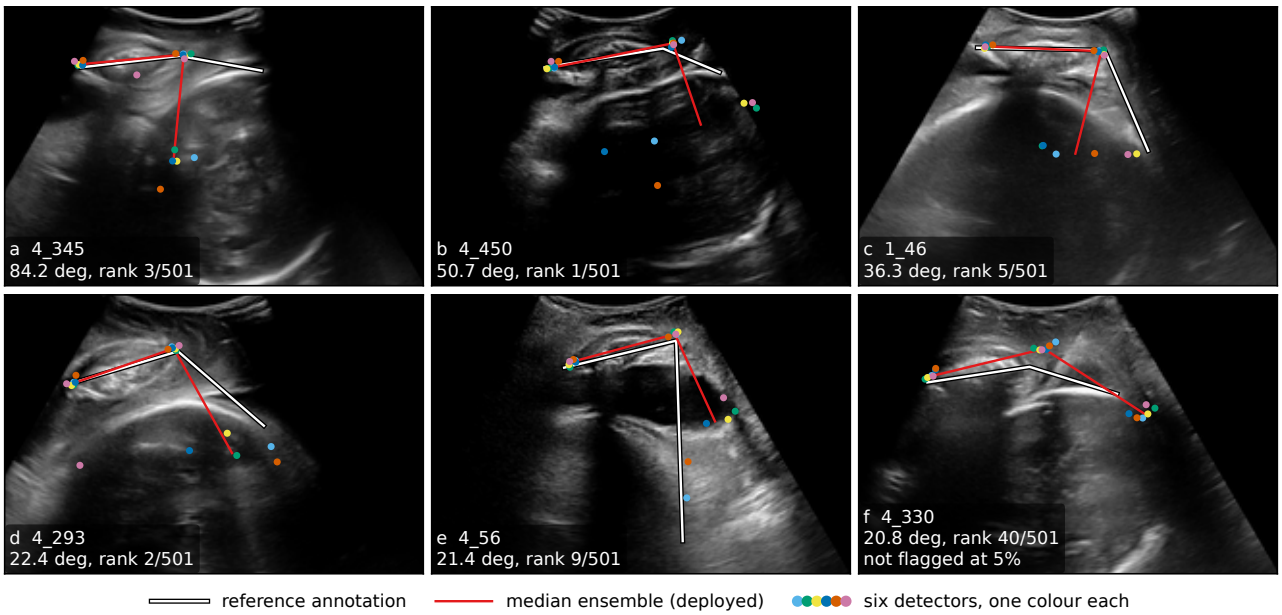

Figure 4: **The six test frames with an AoP error above 20°.** Each panel shows the image (cropped to the landmark region) with the reference landmarks (white), the deployed median prediction (red) and the predictions of the six detectors (dots, one colour per detector); the lines join the two pubic-symphysis endpoints and the fetal-head tangent point. The label gives the frame, its AoP error and its rank under AoP-space disagreement (1 = most suspicious). Panels **a**–**e** were referred at a 5% budget; in **a** all six detectors underestimate the angle, but their spread still ranks the frame 3rd. In **f** the six detectors agree with each other but not with the reference; the frame ranks 40th and is referred only at a 10% budget.

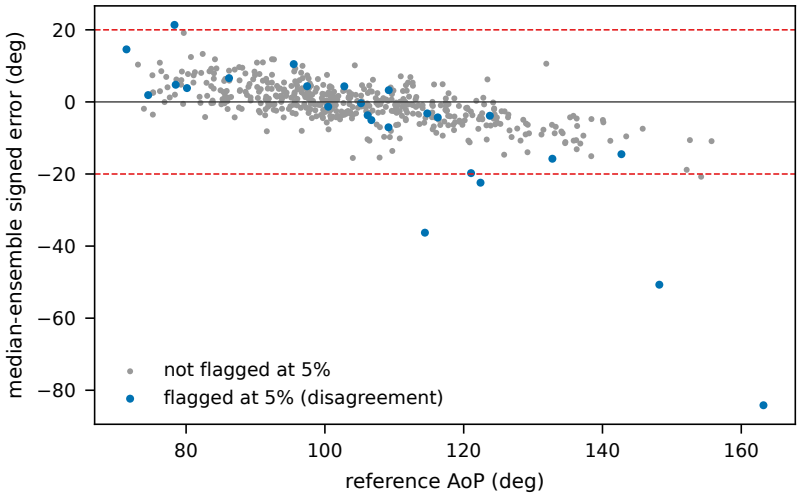

Figure 5: **Signed AoP error of the deployed prediction against the reference AoP (post hoc).** Each point is a test frame (501); frames referred at a 5% budget by AoP-space disagreement are highlighted. Dashed lines mark errors of ±20°. In this sample, larger reference angles were more often underestimated, and three of the twelve frames with a reference above 140° were referred at the 5% budget. Part of this trend is expected from noise in the reference itself. This analysis was added after the failure audit and was not pre-specified.

most a five-image budget can recover of a ten-image decile (recall 0.50; precision 1.00, 95% CI 0.40 to 1.00); at 10% the recall was 0.70 (95% CI 0.46 to 1.00). The E-CPV form gave a 90% coverage gap of 0.288 mm. The result did not depend on the reference: with the senior annotator the gap was 0.289 mm (95% CI 0.149 to 0.456), and with the junior annotator 0.272 mm (0.131 to 0.445). The two annotators differed from each other by an MRE of 1.51 mm on this split. On Test1, the split on which the detectors had been selected, the gap was 0.119 mm (0.067 to 0.175).

3.10 Cost of the systems

A single detector took 35.8 ms per frame (Table 4). Adding its input-consistency score raised this to 290.2 ms with the nine passes run one after another, or 67.9 ms when they were batched. The six-detector pool took 159.9 ms and 2.20 GiB of GPU memory. The three-detector subset chosen on the validation frames (two HRNet-W48 detectors and the ConvNeXt-V2 detector trained with synthetic images) took 89.9 ms. Its retained error at 90% coverage, 3.972°, matched the full pool, but its mean error was higher (4.877 against 4.615°) and it had eight failures above 20° instead of six. The single detector had a mean error of 5.237° and ten failures above 20°, of which its own consistency score referred five at a 5% budget.

Table 4: **Cost and accuracy of four systems on one NVIDIA RTX 4090**

	Cost per frame			AoP error (°)		Failures
System	Passes	Latency (ms)	GiB	Mean	At 90%	caught[b]
One detector, no score	1	35.8 (37.8)	0.56	–	–	–
One detector + input consistency	1 + 8	290.2 (297.9)	0.57	5.237	4.282	5/10
nine passes as one batch[a]	1 + 8	67.9 (69.8)	0.74			
Three detectors (validation-selected)	3	89.9 (92.2)	1.25	4.877	3.972	7/8
Six detectors (deployed pool)	6	159.9 (162.9)	2.20	4.615	3.969	5/6

Passes: forward passes per frame. Latency: median (95th percentile) over 200 timed iterations at batch size 1 with bfloat16 autocast. GiB: peak allocated GPU memory with only that system's detectors loaded. At 90%: retained error at 90% coverage. Accuracy on the 501 IUGC2025 test files, each system ranked by its own score (input consistency for the single detector, AoP-space disagreement otherwise); the single detector is the noisy-student detector. The accuracy columns use the stored predictions (batches of four); at batch size 1, the timing setting, the six-detector pool had a mean error of 4.603° and six failures above 20°, five of them referred at 5% (Additional file 1, Section S6). [a]Secondary timing protocol. [b]The system's own failures above 20° that are referred at a 5% review budget.

4 Discussion

On a public intrapartum ultrasound benchmark, disagreement among six detectors flagged most gross AoP failures: with 5% of frames referred for review, five of the six errors above 20° were caught and the residual rate of such errors fell from 1.2% to 0.2%. No difference was detected between disagreement computed on the angle and on the landmark coordinates at the primary endpoint. The scores did much less for moderate errors, thresholds set on 100 validation frames did not keep their intended workload for every score, and an error shared by all detectors with little spread between them escaped the 5% budget. The replication on a held-out cephalometric split, with either annotator as reference, suggests that the ranking effect is not specific to one dataset or one reference standard, although the cephalometric task has no comparable tail of gross failures.

The null result of the primary contrast is informative. The AoP is a nonlinear function of three landmarks, so we expected the spread of the angle to be a better guide to angle errors than the spread of the coordinates. We found no evidence for this: no difference was detected at 90% coverage, the sign of the contrast reversed over the distinct images and in float32 inference, the secondary measures were mixed (coordinate disagreement better on the worst-decile measures and at 50% coverage, AoP-space disagreement slightly better at 80 and 70%), and the thresholds of coordinate disagreement came closer to their target (6.4 against 11.2% flagged at a nominal 5%). One plausible reason is that the sensitivity of the angle to a given landmark displacement depends on the geometry of each image, which adds a component to the angle spread that is unrelated to error and differs between splits; we did not test this. For practice, coordinate disagreement has the advantage of not depending on the particular measurement derived from the landmarks.

Our adaptation of E-MHA performed poorly as a ranking score for the fixed median prediction of this pool, although the score was proposed for heatmap-based landmark localisation [13]. It was developed for ensembles of identical architectures, whose heatmaps are comparable. A plausible reason, which we did not test, is that here the detectors differ in architecture, heatmap resolution and sharpness, so that the peak of their averaged map reflects which detectors produce sharp maps as much as whether they agree. Scores that compare predictions rather than

internal outputs, such as coordinate or angle disagreement, are not affected by such differences in calibration. Neither weakness is an artefact of numerical precision: E-MHA was unchanged in float32, and a form of confidence that cannot saturate referred fewer gross failures, not more. The single-detector input-consistency score was a closer competitor. It matched disagreement at 90% coverage and caught the most failures above 10° at a 10% budget, but the pre-specified noisy-student score caught three rather than five of the six gross failures at 5%, its ranking levelled off below 90% coverage, and the detector behind it was less accurate than the pool (5.24 against 4.62°). How much consistency caught depended on the detector that supplied it: among the secondary consistency scores, that of the supervised HRNet-W48 detector caught five of the six at 5% (retained error 4.041° at 90% coverage), whereas the AoP-form consistency scores of the three ConvNeXt-V2 detectors caught at most two (Additional file 1, Table S35). With its perturbations batched it was cheaper than the pool (68 against 160 ms), so the choice between the two is a trade between accuracy, tail detection and cost, and a consistency score would need to be validated for the detector that supplies it.

The failure audit marks a limit of the disagreement scores. Disagreement flags frames on which the detectors fail differently. An error that all detectors share does not escape it by itself: in the frame with the largest error all six underestimated the angle, yet their spread ranked it third. A shared error with little spread between the detectors, however, can go unflagged at a small budget, as the one missed failure did. In this sample, underestimation was most frequent at large reference angles, and only three of the twelve frames with a reference above 140° were referred at the 5% budget. Large angles are the values reached in advanced descent [1], so this weakness falls in a clinically relevant range. A quality-control rule for AoP measurement may therefore need a second component that does not rely on disagreement, for example a check of the input distribution or a detector trained with more frames at advanced descent. Referral of predicted angles above a set value would have to allow for the underestimation itself, because an underestimated large angle need not cross such a cut-off. None of these was evaluated here. The same observation explains why the gain shrank for moderate errors: most errors between 10 and 20° were small shared biases rather than disagreements.

The threshold results bear directly on deployment. A rank-based analysis, which refuses a

fixed fraction of a completed test set, cannot be used at the bedside, where each measurement must be accepted or referred when it is made. When thresholds were carried from validation to test frames, the referral rate of AoP-space disagreement more than doubled (11.2% for a nominal 5%). This increased the review workload, while in this sample the flagged frames still included all six failures above 20°; the realised rates also differed between prefix groups and between scores. With 100 calibration frames, a 5% threshold rests on a handful of extreme scores, and resampling the calibration frames alone produced referral rates from 3.8 to 14.6%, an interval that contains the intended 5%; a shift between the splits cannot be separated from this imprecision. On the cephalometric pool the thresholds transferred almost exactly. A site that adopts such a rule would need a calibration set large enough to estimate the relevant tail of the score, and would need to monitor its referral rate over time.

These findings are consistent with previous work in which ensemble disagreement was the most reliable simple signal for failure detection [10,11] and ensemble coordinate spread identified gross landmark errors [13]. They add a measurement-level evaluation with a fixed review workload, a threshold-transfer test and a failure audit. They also suggest that the choice between the disagreement scores matters less than the structure of the errors: the two disagreement scores were close, and the gross failure they missed was one the detectors shared with little spread between them.

4.1 Limitations

All IUGC2025 results are frame-level. The test files come from two prefix groups without a patient mapping, so no patient-level statement is possible and the intervals are anti-conservative; the per-group signs agree for the primary contrast, but two groups cannot support a cluster interval. The validation split overlaps the unlabelled release used for self-training, which probably made the validation-based selection of one detector optimistic and may have affected the threshold calibration; calibrating on the prefix-4 validation frames alone gave similar referral rates, and 55 or 100 calibration frames estimate the tail of a score imprecisely. The study uses one public benchmark retrospectively; the earlier IUGC2024 release is not independent of IUGC2025 (Additional file 1, Section S16), and no external patient cohort was available. The

failure thresholds of 10, 15 and 20° are research definitions that have not been validated against labour outcomes. Each IUGC2025 frame has one reference annotation, and the byte-identical duplicates show that annotations of the same image can differ by up to 5.3°; three of the six failures above 20° exceeded that threshold by less than 2.5° (20.75, 21.37 and 22.41°), so which frames count as gross failures depends on the reference; no clinician re-annotation was available, and the reference sensitivity analysis was possible only for the cephalometric data. The pool contains a detector trained with synthetic images whose image-quality gate was not met (Fréchet inception distance 232.7 against a planned limit of 50), and the pseudo-label file of the noisy-student detector was not retained (Additional file 1, Section S2); in the label-only subset of three detectors the ranking effect held, but how many gross failures a 5% budget caught depended on which detectors were pooled (Additional file 1, Section S6). Our adaptation of E-MHA for a heterogeneous pool, with a common grid and normalised maps, is one of several possible, and it was evaluated only as a ranking score for the median prediction, not as the complete method, which also localises the landmarks from the averaged heatmap. The stored predictions were made with bfloat16 inference, whose numerics depend on the batch and the machine; re-running them moved individual coordinates by up to several pixels but left the six failures above 20° and their referral at 5 and 10% unchanged (Additional file 1, Section S6). The mean error of the pool (4.62°) is above the best published result on this split (3.81° [7]), and a more accurate system may leave a different tail. Finally, the tail rests on six failures above 20°, so the counts at that threshold are fragile (an exact binomial 95% CI for five of six is 36 to 100% even if frames were independent), and the post hoc analysis of large reference angles needs confirmation in independent data.

5 Conclusions

In this retrospective frame-level benchmark, disagreement among six landmark detectors enriched gross AoP errors among the measurements referred for review: five of the six errors above 20° were referred at a 5% budget, and the one missed was an underestimation shared by all detectors with little spread between them. No advantage of disagreement computed on the

angle over disagreement computed on the landmark coordinates was detected. Disagreement did much less for moderate errors, and thresholds derived from 100 validation frames did not keep their intended referral rates on the test frames. Ensemble disagreement is therefore a candidate quality-control signal for this measurement that needs a complementary check for shared errors and local calibration; these findings do not establish patient-level safety or clinical benefit, which require prospective, patient-level studies.

Abbreviations

AoP: angle of progression; AUC: area under the receiver operating characteristic curve; CI: confidence interval; CLAIM: Checklist for Artificial Intelligence in Medical Imaging; E-CPV: ensemble coordinate prediction variance; E-MHA: ensemble maximum heatmap activation; GiB: gibibyte; GPU: graphics processing unit; HRNet: high-resolution network; ISBI: International Symposium on Biomedical Imaging; IUGC: Intrapartum Ultrasound Grand Challenge; MRE: mean radial error.

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified datasets released for open challenges (IUGC2025 [5,6] and ISBI 2015 [16]); no patient-identifiable information was accessed. Ethical approval and consent to participate were therefore not required, in accordance with Article 32 of China's Measures for Ethical Review of Life Sciences and Medical Research Involving Humans (2023).

Consent for publication

Not applicable; Figure 4 shows de-identified images from a public dataset.

Availability of data and materials

The datasets analysed in this study are publicly available at the following URLs. The IUGC2025 dataset [5,6] is available at `https://zenodo.org/records/17355570` (DOI: `10.5281/zenodo.17355570`) and from the challenge organisers on request under their data usage agreement, under which the copy analysed here was obtained. The ISBI 2015 cephalometric dataset [16, 19], including the images and the annotations of both orthodontists, is available from its challenge organisers at `https://figshare.com/s/37ec464af8e81ae6ebbf`. The earlier IUGC2024 release, used only for the overlap screen reported in Additional file 1, Section S16, and not as an external validation cohort, is available at `https://zenodo.org/records/17655183` (DOI: `10.5281/zenodo.17655183`). Access to and reuse of these datasets are subject to the terms of their providers. The analysis code, the figure and table scripts, and the dated analysis plan with its documented deviations will be released in a public GitHub repository and archived on Zenodo upon acceptance; until then they are available to the editors and reviewers from the corresponding author on request. Per-frame outputs of the detectors on the IUGC2025 validation and test images are derived from the benchmark data and are available from the corresponding author to researchers who hold the IUGC2025 data usage agreement.

Competing interests

The authors declare that they have no competing interests. None of the authors is affiliated with the organising committees of the IUGC or ISBI 2015 challenges.

Funding

This work was supported by the Basic and Applied Basic Research Foundation of Guangdong Province (Grant 2021A1515110418, to SL). The funder had no role in study design, data analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

LZ conceived and designed the study, implemented the software, performed the analyses and wrote the first draft. YC validated the evaluation and curated the data, including the detector pool and its audit. TC and YH trained the detectors and generative models whose outputs form the pools. QP and SL carried out the abstention analyses and their clinical interpretation, and SL acquired the funding. CG and MF carried out the selection analyses and the statistical analysis. YG supervised and administered the project, provided resources and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the organisers of the Intrapartum Ultrasound Grand Challenge and of the ISBI 2015 cephalometric challenge for making their data available. Liangji Zhou thanks family members for their continuous support and encouragement.

Clinical trial number

Not applicable.

References

[1] Ghi T, Eggebø T, Lees C, Kalache K, Rozenberg P, Youssef A, et al. ISUOG Practice Guidelines: intrapartum ultrasound. Ultrasound in Obstetrics & Gynecology. 2018;52(1):128-39.

[2] Lu Y, Zhou M, Zhi D, Zhou M, Jiang X, Qiu R, et al. The JNU-IFM dataset for segmenting pubic symphysis-fetal head. Data in Brief. 2022;41:107904.

[3] Chen G, Bai J, Ou Z, Lu Y, Wang H. PSFHS: Intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head. Scientific Data. 2024;11:436.

[4] Bai J, Zhou Z, Ou Z, Koehler G, Stock R, Maier-Hein K, et al. PSFHS challenge report: Pubic symphysis and fetal head segmentation from intrapartum ultrasound images. Medical Image Analysis. 2025;99:103353.

[5] Bai J, Tang Y, Liu X, Hu J, Li Y, Chen X, et al. IUGC: A benchmark of landmark detection in end-to-end intrapartum ultrasound biometry. Medical Image Analysis. 2026;110:103960.

[6] Bai J, Khobo I, Lu Y, Ni D, Yaqub M, Lekadir K, et al.. Landmark Detection Challenge for Intrapartum Ultrasound Measurement Meeting the Actual Clinical Assessment of Labor Progress; 2025. Zenodo, challenge design document, version 2.

[7] Ma C, Li Y, Guo B, Jiao J, Huang Y, Wang Y, et al. Unlabeled Data-Driven Fetal Landmark Detection in Intrapartum Ultrasound. In: Intrapartum Ultrasound. Lecture Notes in Computer Science. Springer; 2026. p. 14-23.

[8] Geifman Y, El-Yaniv R. Selective Classification for Deep Neural Networks. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017.

[9] Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: Advances in Neural Information Processing Systems (NeurIPS); 2017. ArXiv:1612.01474.

[10] Jaeger PF, Lüth CT, Klein L, Bungert TJ. A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification. In: International Conference on Learning Representations (ICLR); 2023.

[11] Zenk M, Zimmerer D, Isensee F, Traub J, Norajitra T, Jäger PF, et al. Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. Medical Image Analysis. 2025;101:103392.

[12] Roy AG, Conjeti S, Navab N, Wachinger C. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage. 2019;195:11-22.

[13] Schobs LA, Swift AJ, Lu H. Uncertainty Estimation for Heatmap-Based Landmark Localization. IEEE Transactions on Medical Imaging. 2023;42(4):1021-34.

[14] Jafari MH, Luong C, Tsang M, Gu AN, Van Woudenberg N, Rohling R, et al. U-LanD: Uncertainty-Driven Video Landmark Detection. IEEE Transactions on Medical Imaging. 2022;41(4):793-804.

[15] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiology: Artificial Intelligence. 2024;6(4):e240300.

[16] Wang CW, Huang CT, Lee JH, Li CH, Chang SW, Siao MJ, et al. A benchmark for comparison of dental radiography analysis algorithms. Medical Image Analysis. 2016;31:63-76.

[17] Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, et al. Deep High-Resolution Representation Learning for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;43(10):3349-64.

[18] Woo S, Debnath S, Hu R, Chen X, Liu Z, Kweon IS, et al. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 16133-42.

[19] Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms. Scientific Reports. 2016;6:33581.

Supplementary Information

Additional file 1 (PDF). Supplementary methods, results and tables. Part A documents the present evaluation: S1, the analysis plan, the results known when it was written and the deviations from it; S2, the independence audit, the composition and provenance of the detector

pool, and the inference protocol; S3, the full risk-workload tables with intervals, the complementarity of the scores and the distinct-image analysis; S4, the fixed-threshold analyses with both calibrations, the threshold values and the implementation of E-MHA; S5, the reference sensitivity, the failure audit table and the timing protocol; S6, the analyses added after review, on detector subsets trained on labelled frames only and on the inference protocol; S7, the oracle headroom of the pool, its dependence on pool size and three dispersion-matched nulls; S8, why disagreement flags the failures, a single-policy control and the robustness to pool composition. Part B (S9 to S22) reproduces the earlier exploratory work on the same pools, whose aims differ from those of this study: the apparatus that generated the candidate pools, the full sweep of label-free selection rules and the cephalometric scaling analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1B.pdf