

Additional file 1

Supplementary methods, results and tables for “Comparing label-free quality-control scores for automated angle-of-progression measurement in intrapartum ultrasound: a retrospective benchmark study”

Liangji Zhou, Yuanqiu Chen, Tongtong Chen, Yijuan Huang, Qihao Pan, Si Li, Chenyu Gou, Meiqiong Fan and Yu Gao

Department of Obstetrics and Biomedical Innovation Center, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. Corresponding author: Yu Gao, gaoy57@mail.sysu.edu.cn

This file has two parts. Part A (Sections S1 to S8) documents the present evaluation: the analysis plan, the results known when it was written and the deviations from it (S1); the independence audit, the composition and provenance of the detector pool and the inference protocol (S2); the full risk-workload tables (S3); the fixed-threshold analyses, with the threshold values, and the implementation of E-MHA (S4); the reference sensitivity, the failure audit and the timing (S5); the analyses added after review, on detector subsets trained on labelled frames only and on the inference protocol (S6); the oracle headroom of the pool and its dependence on pool size (S7); and why disagreement flags the failures, with a single-policy control and the robustness to pool composition (S8). Part B (Sections S9 to S22) reproduces, for transparency, the earlier exploratory work on the same pools: a stochastic landmark-refinement policy, a plausibility verifier and a conditional diffusion refiner, the full sweep of label-free selection and aggregation rules, the per-landmark oracle, the test-time scaling characterisation on both datasets, and the cephalometric ensemble abstention on Test1 and Test2. Its aims and evaluation designs are its own and differ from those of the present study; the main text relies on it only for the failure of label-free selection and for the construction of the detector pools. Section numbers carry the prefix S, and references to the main text say so.

S1 Analysis plan and deviations

The comparisons of the main text were fixed on 27 September 2026 in a dated analysis plan ([paper/prereg_B_resubmit_2026-09-27.md](#)), written before any of the new scores (coordinate disagreement, E-MHA and the validation-set input consistency) was computed on any frame. The plan was written after earlier, unpublished analyses of the same pools, whose results were therefore known when it was written: on the IUGC2025 test split, the mean error of the deployed prediction and the per-frame oracle; the retained error and the recall of the six failures above 20° for AoP-space disagreement (five of six at a 5% flag rate), for mean heatmap confidence, and for the input consistency of the noisy-student detector and of the other five detectors; the full sweep of label-free selection rules (Part B); and, on the cephalometric pool, the disagreement gaps on Test1 and Test2. The file was not deposited in a public registry, so its date rests on the file and the analysis logs; it will be released with the code on acceptance and is available to the editors and reviewers on request. The SHA-256 digest of that version is `890f078ded92c06d55813f716905063f728608480d9f676f13eb0e269b39317b`. It fixed the data and the deployed predictions, the five scores and their aggregation, the primary contrast with its three readings, the risk-workload grid, the fixed-threshold protocol with both calibrations, the reference sensitivity analyses, the failure audit, the system-cost protocol, and the statistical controls. Deviations were appended to the file without editing the original sections:

1. The retained error refuses the most suspicious frames first and breaks ties in frame order, the convention of the figure script of the earlier analysis of this pool. Keeping tied frames first instead gives 4.314° rather than 4.277° for the saturated confidence score at 70% coverage; continuous scores are unaffected.

2. The cephalometric negative control became the mean over 1,000 permutations of the score, because a single permutation fluctuates by about 0.05 mm at 100 images; the one-sided permutation P value of the real gap is reported beside it.
3. The analysis of the signed error against the reference AoP (main text, Fig. 5) was added after the failure audit and is post hoc.
4. The IUGC2025 duplicate files were not re-scored with each alternative annotation: none of the six failures above 20° belongs to a duplicate group, so their classification does not depend on the annotation, and the distinct-image analysis (Table S6) is reported instead.
5. The server smoke test scores four frames instead of two, because under bfloat16 the coordinates depend on the batch composition (Section S2); the full runs use the original batches.
6. After a pre-submission review, two analyses that the plan did not contain were added (Section S6): the evaluation of detector subsets trained on labelled frames only, and a check of the inference protocol that includes a non-saturating form of heatmap confidence. Both are post hoc.

Every analysis script first reproduced the values of the earlier analyses of the pool (for the intrapartum pool: mean error 4.615° , per-frame oracle 2.282° , retained error of AoP-space disagreement 3.969, 3.751, 3.621 and 3.548° and of confidence 4.210, 4.207, 4.277 and 4.176° at 90, 80, 70 and 50% coverage, input consistency 3.957° at 90%; for the cephalometric pool: Test2 MRE 3.485 mm and gaps 0.281 and 0.439 mm at 90 and 80%, Test1 MRE 3.110 mm and gap 0.119 mm) and stopped otherwise; all reproduced to within 0.002 (intrapartum) and 0.0005 (cephalometric), both locally and on the server. A permuted score had to rank at chance (worst-decile AUC of a permuted coordinate disagreement within 0.35 to 0.65). The scripts are `scripts/analyze_b_qc_resubmit.py`, `scripts/analyze_b_cephalo_resubmit.py`, `scripts/audit_eval_vs_training_neardup.py`, `scripts/audit_b_failure_cases.py`, `scripts/dump_consistency_signal.py`, `scripts/dump_emha_scores.py` and `scripts/time_b_systems.py`, driven on the server by `scripts/run_b_resubmit_server.sh`; the figures and tables are generated by `scripts/make_figures_b_bmc.py`, `scripts/make_tables_b_bmc.py` and `scripts/make_supp_tables_b_bmc.py`.

S2 Independence audit, detector pool and inference protocol

S2.1 Near-duplicate screen

Every validation and test frame was compared with all 33,766 training files, labelled and unlabelled, by the Pearson correlation of 48×48 grayscale thumbnails (`scripts/audit_eval_vs_training_neardup.py`). The five byte-identical duplicate pairs inside the test split scored 1.000 (positive control); correlations between the test frames and 2,000 randomly drawn training files had a median of 0.407, a 99th percentile of 0.708, a 99.99th percentile of 0.824 and a maximum of 0.922 (null band). No test frame exceeded 0.95. Five exceeded 0.90: test frames 1_1 (0.929) and 1_2 (0.915), both closest to unlabelled frame 3_165_10; 4_233 (0.909, closest to 3_104_5); 4_231 (0.906, 3_102_8); and 1_78 (0.900, 3_164_12). All five were inspected. Frames 1_1, 1_2 and 1_78 differ visibly from their matches. The pairs 4_233/3_104_5 and 4_231/3_102_8 share the scanner geometry and a similar appearance but differ throughout the image: at full resolution their mean absolute pixel differences are 9.3 and 7.8 grey levels (correlations 0.834 and 0.825), against 0.74 between adjacent frames of one unlabelled clip. They are therefore different acquisitions, but without a patient mapping they cannot be shown to come from different patients.

All 45 validation frames with prefix 2, and none with prefix 4, reached a correlation of 1.000 with a frame of the unlabelled clip that shares their identifier; at full resolution the matched frames differ by 0.05 to 0.18 grey levels, consistent with re-encoding of the same image. No test identifier has an unlabelled clip of the same name. Table S2 counts the frames from these clips in the pseudo-label pools derived from the release, and Table S1 shows the asymmetry expected if the self-trained detector had seen the prefix-2 validation images.

Table S1: Validation AoP error by file-name prefix group, with and without detected overlap with the unlabelled release

Detector	Prefix 2 (45)	Prefix 4 (55)	Difference
alb1_ema	4.731	4.227	+0.504
convnext_s42	5.562	3.749	+1.813
convnext_s513	5.727	3.625	+2.102
hrnet_ns	4.693	3.912	+0.781
hrnet_v2	6.125	4.295	+1.830
round8_l2	5.439	4.232	+1.207
median_ensemble	5.023	3.403	+1.620

Mean absolute AoP error ($^{\circ}$) on the validation frames. The noisy-student detector (**hrnet_ns**) was trained on pseudo-labels produced by **hrnet_v2**; its improvement over the teacher is 1.43° on prefix 2 and 0.38° on prefix 4.

Table S2: Pseudo-label pools derived from the unlabelled release: frames from clips of validation frames

Pool file	Rows	From validation clips
pseudo_audit_kcenter_imagenet_k1500.csv	1500	119
pseudo_audit_kcenter_task_k1500.csv	1500	86
pseudo_audit_random_cap.csv	1500	174
pseudo_budget_k12000.csv	5816	355
pseudo_budget_k1500.csv	1500	86
pseudo_budget_k3000.csv	3000	215
pseudo_budget_k32812.csv	32812	3896
pseudo_budget_k6000.csv	5816	355
pseudo_coverage.csv	3000	232
pseudo_hybrid.csv	1500	86
pseudo_labels_multi_3teacher.csv	1187	77
pseudo_pool.csv	32812	3896
pseudo_random_nocap.csv	1500	173
pseudo_temporal.csv	3000	0

A clip counts when its identifier equals that of a validation frame (for example 2_115); the pool used by the noisy-student detector of the pool was not retained. Generated by `scripts/count_val_clip_rows.py`.

S2.2 Composition of the IUGC2025 detector pool

Table S3 lists the six detectors. The earlier description of this pool as four HRNet-W48 and two ConvNeXt-V2 detectors, which also appeared in the caption of Table S17 of the earlier version of this document, was incorrect: the detector **round8_l2** inherits the ConvNeXt-V2 configuration, and its checkpoint has the size of the other two ConvNeXt-V2 checkpoints. Its training log records 1,000 synthetic images appended to the 300 labelled frames (loss weight 0.3 for the synthetic frames). The synthetic images came from a conditional flow-matching generator fitted to the labelled training frames only, each generated from the landmark configuration of one training frame. The pre-set image-quality gate of the generator (Fréchet inception distance below 50) was evaluated only after this retraining and was not met (232.7), but the detector was kept in the pool. The pseudo-label file of the noisy-student detector was passed on the command line and was not retained; a comment in its configuration file documents the intended

generation from the supervised HRNet-W48 teacher with a confidence floor of 0.35 and at most 3,000 frames, which cannot be confirmed from the run directory.

Table S3: The six detectors of the IUGC2025 pool

Tag	Backbone	Training	Checkpoint (MB)
<code>hrnet_v2</code>	HRNet-W48	supervised, 300 labelled frames	276.6
<code>a1b1_ema</code>	HRNet-W48	fine-tuned with a circular AoP loss, weight averaging (EMA weights used)	276.6
<code>hrnet_ns</code>	HRNet-W48	noisy-student, pseudo-labels of unlabelled frames from <code>hrnet_v2</code>	276.6
<code>convnext_s42</code>	ConvNeXt-V2-Base	supervised, seed 42	370.8
<code>convnext_s513</code>	ConvNeXt-V2-Base	supervised, seed 20260513	370.8
<code>round8_12</code>	ConvNeXt-V2-Base	retrained with 1,000 flow-matching synthetic images	370.8

Every detector was selected on the validation split; an audit of the saved selection metrics found no selection on a test metric. Checkpoint sizes from the provenance audit of the run directories.

S2.3 Inference protocol

All pool predictions were produced with bfloat16 autocast in batches of four frames. Re-running the six detectors for the E-MHA scores reproduced the stored coordinates of all 601 validation and test frames exactly (maximum absolute difference 0.000 px for every detector). With a batch of two frames instead, one test coordinate of `convnext_s42` moved by 1.87 px while all others stayed within 0.06 px. The batch size is therefore part of the inference protocol when coordinates are compared across runs. After the server was re-rented, the same protocol no longer reproduced the stored coordinates bit for bit (median difference 0.010 px); Section S6 measures how much this and other inference settings change the results.

S3 Risk-workload analyses

Table S4 gives every score, budget and failure threshold with the frame-bootstrap interval of the residual failure rate; Table S5 gives the overlap of the failures caught by two scores; Table S6 repeats the main analyses over distinct images.

Table S4: Failures caught and residual failure rate for every score, review budget and failure threshold

Score	Failure	2%	5%	10%	20%
D-AoP	>20°	5/6, 0.2 [0.0, 1.0]	5/6, 0.2 [0.0, 0.6]	6/6, 0.0 [0.0, 0.2]	6/6, 0.0 [0.0, 0.0]
	>15°	6/13, 1.4 [0.6, 2.6]	7/13, 1.3 [0.4, 2.3]	9/13, 0.9 [0.0, 1.8]	11/13, 0.5 [0.0, 1.2]
	>10°	9/47, 7.7 [5.5, 10.4]	10/47, 7.8 [5.3, 10.1]	18/47, 6.4 [4.4, 9.1]	25/47, 5.5 [3.5, 8.0]
D-coord	>20°	5/6, 0.2 [0.0, 1.0]	5/6, 0.2 [0.0, 0.6]	6/6, 0.0 [0.0, 0.2]	6/6, 0.0 [0.0, 0.0]
	>15°	7/13, 1.2 [0.4, 2.6]	8/13, 1.1 [0.2, 2.1]	10/13, 0.7 [0.0, 1.8]	11/13, 0.5 [0.0, 1.2]
	>10°	9/47, 7.7 [5.7, 10.6]	12/47, 7.4 [5.0, 9.7]	18/47, 6.4 [4.2, 8.6]	27/47, 5.0 [3.2, 7.5]
D-coord max	>20°	5/6, 0.2 [0.0, 1.0]	5/6, 0.2 [0.0, 0.6]	5/6, 0.2 [0.0, 0.7]	6/6, 0.0 [0.0, 0.0]
	>15°	6/13, 1.4 [0.6, 2.9]	8/13, 1.1 [0.2, 2.1]	8/13, 1.1 [0.2, 2.2]	10/13, 0.7 [0.0, 1.5]
	>10°	8/47, 7.9 [5.7, 10.6]	13/47, 7.1 [5.0, 9.7]	16/47, 6.9 [4.7, 9.3]	25/47, 5.5 [3.5, 8.0]
IC	>20°	2/6, 0.8 [0.0, 1.6]	3/6, 0.6 [0.0, 1.5]	5/6, 0.2 [0.0, 0.9]	6/6, 0.0 [0.0, 0.0]
	>15°	4/13, 1.8 [0.8, 3.3]	7/13, 1.3 [0.4, 2.5]	10/13, 0.7 [0.0, 1.8]	11/13, 0.5 [0.0, 1.2]
	>10°	5/47, 8.6 [6.3, 11.2]	12/47, 7.4 [5.3, 10.1]	20/47, 6.0 [4.0, 8.4]	26/47, 5.2 [3.2, 7.7]
E-MHA	>20°	3/6, 0.6 [0.0, 1.2]	4/6, 0.4 [0.0, 1.1]	4/6, 0.4 [0.0, 1.1]	5/6, 0.2 [0.0, 0.7]
	>15°	5/13, 1.6 [0.6, 3.1]	6/13, 1.5 [0.4, 2.5]	7/13, 1.3 [0.4, 2.4]	8/13, 1.2 [0.2, 2.5]
	>10°	7/47, 8.1 [5.9, 10.8]	8/47, 8.2 [5.9, 10.7]	9/47, 8.4 [6.0, 11.1]	13/47, 8.5 [6.2, 11.5]
E-MHA sigm.	>20°	4/6, 0.4 [0.0, 1.2]	4/6, 0.4 [0.0, 1.1]	4/6, 0.4 [0.0, 1.1]	4/6, 0.5 [0.0, 1.2]
	>15°	6/13, 1.4 [0.6, 2.9]	6/13, 1.5 [0.4, 2.7]	6/13, 1.6 [0.4, 2.9]	6/13, 1.7 [0.5, 3.2]
	>10°	7/47, 8.1 [5.9, 10.8]	11/47, 7.6 [5.5, 10.1]	13/47, 7.5 [5.3, 10.0]	15/47, 8.0 [5.7, 11.0]
Conf.	>20°	4/6, 0.4 [0.0, 1.2]	4/6, 0.4 [0.0, 1.1]	4/6, 0.4 [0.0, 1.1]	4/6, 0.5 [0.0, 1.2]
	>15°	5/13, 1.6 [0.6, 3.1]	6/13, 1.5 [0.4, 2.7]	6/13, 1.6 [0.4, 2.9]	6/13, 1.7 [0.5, 3.2]
	>10°	6/47, 8.4 [6.1, 11.0]	9/47, 8.0 [5.9, 10.5]	11/47, 8.0 [5.8, 10.4]	14/47, 8.2 [6.0, 11.2]
Random (expected)	>20°	0.1/6, 1.2	0.3/6, 1.2	0.6/6, 1.2	1.2/6, 1.2
	>15°	0.3/13, 2.6	0.6/13, 2.6	1.3/13, 2.6	2.6/13, 2.6
	>10°	0.9/47, 9.4	2.3/47, 9.4	4.7/47, 9.4	9.4/47, 9.4

IUGC2025 test split, 501 files; budgets refer 10, 25, 50 and 100 files. Each cell: failures referred / all failures, residual failure rate among the retained files in percent [95% frame-bootstrap interval, 2,000 resamples; the failure indicator travels with its file; each resample is re-ranked and its referral set recomputed]. A residual rate of 0.0 means that no failure remained among the retained files (for example 0 of 401 at a 20% budget); a displayed interval of [0.0, 0.0] under this empirical bootstrap must not be read as evidence of zero risk on new data. Random: analytical expectation. D-AoP: AoP-space disagreement; D-coord: coordinate disagreement (E-CPV), mean over landmarks; D-coord max: maximum over landmarks; IC: input consistency of the noisy-student detector; E-MHA: ensemble heatmap peak, normalised maps; E-MHA sigm.: raw sigmoid maps; Conf.: mean peak heatmap confidence.

Table S5: Overlap of the failures caught by two scores at the same review budget

Pair	Budget	Failure	n	Both	First only	Second only	Neither
D-AoP vs IC	5%	>20°	6	3	2	0	1
	5%	>15°	13	5	2	2	4
	5%	>10°	47	8	2	4	33
	10%	>20°	6	5	1	0	0
	10%	>15°	13	8	1	2	2
	10%	>10°	47	14	4	6	23
D-AoP vs D-coord	5%	>20°	6	5	0	0	1
	5%	>15°	13	7	0	1	5
	5%	>10°	47	10	0	2	35
	10%	>20°	6	6	0	0	0
	10%	>15°	13	9	0	1	3
	10%	>10°	47	16	2	2	27

IUGC2025 test split. First only: failures referred by the first score of the pair but not by the second. D-AoP: AoP-space disagreement; D-coord: coordinate disagreement (E-CPV), mean over landmarks; D-coord max: maximum over landmarks; IC: input consistency of the noisy-student detector; E-MHA: ensemble heatmap peak, normalised maps; E-MHA sigm.: raw sigmoid maps; Conf.: mean peak heatmap confidence.

Table S6: Distinct-image sensitivity analysis

Score	90%	80%	70%	50%	>20°, 5%	>20°, 10%
D-AoP	4.026	3.774	3.646	3.599	5/6	6/6
D-coord	4.011	3.821	3.680	3.341	5/6	6/6
D-coord max	4.057	3.856	3.646	3.353	5/6	5/6
IC	3.990	3.914	3.844	3.855	3/6	5/6
E-MHA	4.291	4.210	4.262	4.443	4/6	4/6
E-MHA sigm.	4.185	4.253	4.209	3.940	4/6	4/6
Conf.	4.241	4.265	4.349	4.205	4/6	4/6

The 501 test files collapsed to 493 distinct images: the error of a duplicated image is the mean over its files' annotations and it counts once; budgets are 5 and 10% of the distinct images. Mean error of the deployed prediction 4.655°. Columns 2 to 5: retained error (°) at the given coverage.

S4 Fixed thresholds and the implementation of E-MHA

Table S7 gives the thresholds fixed on all 100 validation frames and on the 55 prefix-4 validation frames, with the interval of the realised referral rate obtained by bootstrapping the calibration frames, and Table S8 gives the threshold values themselves. For E-MHA, each detector's heatmap logits were passed through a sigmoid, resampled bilinearly to 128×128 (the detectors produce heatmaps at different resolutions), and either divided by their spatial sum (primary form) or left as they were (secondary form); the maps were averaged over the six detectors and the score is minus the mean over the three landmarks of the maximum of the averaged map. The dump script also recomputes each detector's coordinates from its own heatmaps and compares them with the stored pool predictions as a positive control (Section S2).

Table S7: Thresholds fixed on validation frames and applied to the test files

Score	Calibration	Nominal 5%	Nominal 10%	Nominal 20%
D-AoP	all 100	11.2 [3.8, 14.6]; 6/6; 3.93	13.8 [11.2, 20.0]; 6/6; 3.86	21.8 [15.6, 26.7]; 6/6; 3.68
	prefix 4 (55)	12.6 [2.8, 23.2]; 6/6; 3.92	21.4 [8.6, 27.7]; 6/6; 3.68	28.1 [22.6, 41.3]; 6/6; 3.62
D-coord	all 100	6.4 [4.6, 10.4]; 5/6; 4.01	9.8 [6.4, 16.0]; 6/6; 3.98	18.0 [10.4, 28.1]; 6/6; 3.81
	prefix 4 (55)	6.8 [2.4, 15.6]; 5/6; 4.02	10.4 [5.2, 23.6]; 6/6; 3.99	23.0 [11.6, 35.9]; 6/6; 3.77
D-coord max	all 100	8.4 [6.0, 14.6]; 5/6; 4.04	14.0 [8.6, 18.2]; 5/6; 3.98	26.5 [16.8, 34.5]; 6/6; 3.67
	prefix 4 (55)	6.8 [2.6, 15.4]; 5/6; 4.04	13.2 [6.4, 22.0]; 5/6; 4.01	23.8 [13.4, 39.3]; 6/6; 3.75
IC	all 100	7.8 [3.6, 12.8]; 4/6; 4.15	12.6 [7.8, 18.4]; 5/6; 3.95	21.4 [14.0, 23.6]; 6/6; 3.83
	prefix 4 (55)	8.8 [4.8, 18.0]; 4/6; 4.13	15.0 [7.8, 21.8]; 5/6; 3.98	21.8 [16.6, 27.9]; 6/6; 3.84
E-MHA	all 100	9.8 [3.0, 25.0]; 4/6; 4.25	22.6 [9.8, 35.1]; 5/6; 4.16	42.1 [29.5, 58.1]; 5/6; 4.28
	prefix 4 (55)	7.0 [0.4, 17.0]; 4/6; 4.23	11.8 [4.4, 29.5]; 4/6; 4.22	29.1 [12.6, 40.5]; 5/6; 4.21
E-MHA sigm.	all 100	4.8 [2.0, 11.6]; 4/6; 4.15	10.8 [5.0, 17.6]; 4/6; 4.17	19.0 [13.4, 21.0]; 4/6; 4.23
	prefix 4 (55)	3.8 [0.4, 12.0]; 4/6; 4.17	9.6 [2.4, 15.6]; 4/6; 4.14	16.4 [11.0, 24.4]; 4/6; 4.23
Conf.	all 100	3.6 [1.0, 8.8]; 4/6; 4.20	7.4 [3.6, 15.2]; 4/6; 4.18	17.8 [13.0, 24.0]; 4/6; 4.24
	prefix 4 (55)	3.4 [0.2, 17.2]; 4/6; 4.20	15.0 [1.4, 20.8]; 4/6; 4.20	20.6 [15.0, 25.5]; 4/6; 4.22

Each cell: realised percentage of test files flagged [95% interval from bootstrapping the calibration frames]; failures above 20° among the flagged files; retained mean error (°). A test file is flagged when its score exceeds the threshold; the threshold values are in Table S8. D-AoP: AoP-space disagreement; D-coord: coordinate disagreement (E-CPV), mean over landmarks; D-coord max: maximum over landmarks; IC: input consistency of the noisy-student detector; E-MHA: ensemble heatmap peak, normalised maps; E-MHA sigm.: raw sigmoid maps; Conf.: mean peak heatmap confidence.

Table S8: Threshold values of the fixed-threshold analysis

Score	Unit	Calibration	Nominal 5%	Nominal 10%	Nominal 20%
D-AoP	deg	all 100	4.35	3.726	3.142
		prefix 4 (55)	3.963	3.157	2.88
D-coord	px	all 100	9.245	7.87	6.549
		prefix 4 (55)	8.849	7.66	6.11
D-coord max	px	all 100	14.99	11.54	8.646
		prefix 4 (55)	17.41	11.84	8.932
IC	deg	all 100	3.279	2.447	1.874
		prefix 4 (55)	3.059	2.199	1.843
E-MHA	-	all 100	-0.0001573	-0.000158	-0.0001587
		prefix 4 (55)	-0.000157	-0.0001574	-0.0001582
E-MHA sigm.	-	all 100	-0.985	-0.9954	-0.9973
		prefix 4 (55)	-0.9806	-0.9946	-0.9969
Conf.	-	all 100	-0.9906	-0.9994	-0.9998
		prefix 4 (55)	-0.9885	-0.9997	-0.9998

Each threshold is the $(1 - q)$ quantile of the score over the calibration frames (linear interpolation between order statistics, the default of `numpy.quantile`); a test file is flagged when its score is strictly greater. Higher scores mean less reliable frames, so E-MHA and confidence are negated peaks (no unit). Coordinates are in pixels of the 512×512 network input. D-AoP: AoP-space disagreement; D-coord: coordinate disagreement (E-CPV), mean over landmarks; D-coord max: maximum over landmarks; IC: input consistency of the noisy-student detector; E-MHA: ensemble heatmap peak, normalised maps; E-MHA sigm.: raw sigmoid maps; Conf.: mean peak heatmap confidence.

S5 Reference sensitivity, failure audit and timing

Table S9 repeats the cephalometric Test2 endpoints with each annotator as reference, and Table S10 contrasts the two constructions of the worst-decile interval. With a five-image budget and a ten-image decile, recall cannot exceed 0.50; the value of 0.50 therefore means that every referred image was in the worst decile, which the precision reports directly. Table S11 lists the six intrapartum failures above 20° , and Table S12 the timing of the four systems.

Table S9: Cephalometric Test2 endpoints with each annotator as reference

Reference	MRE (mm)	Gap at 90%	Gap at 80%	Precision, 5%	Recall, 10%
Senior orthodontist	3.599	0.289 [0.149, 0.456]	0.453	1.00	0.70
Junior orthodontist	3.554	0.272 [0.131, 0.445]	0.418	0.80	0.60
Mean of both	3.485	0.281 [0.138, 0.451]	0.439	1.00	0.70

Predictions, scores and images unchanged; gap: retained MRE under random refusal minus under disagreement refusal (mm) [95% image-bootstrap interval]. The three rows share one set of resamples, so the interval for the mean of both annotators differs slightly from that of the primary analysis reported in the main text (0.135 to 0.460 mm), which used its own resamples; the point estimates are identical. The average of the raw annotations reproduces the reference used throughout to within 0.000000 px. MRE between the two annotators: 1.514 mm.

Table S10: Two constructions of the worst-decile interval on cephalometric Test2

Worst decile	Recall, 5%	Precision, 5%	Recall, 10%	Precision, 10%
Fixed on the full sample (main text)	[0.27, 0.80]	[0.40, 1.00]	[0.46, 1.00]	[0.30, 1.00]
Redefined in every resample	[0.11, 0.56]	[0.20, 1.00]	[0.33, 0.90]	[0.30, 0.90]

Point values: recall 0.50 and precision 1.00 at 5% (five images; the ten-image decile caps recall at 0.50), recall 0.70 at 10%. The earlier report used the second construction, whose upper bound $5/9 = 0.56$ reflects resamples in which the decile holds nine images.

Table S11: The six test files with an AoP error above 20°

File	Error	Reference	Six detector AoPs (°)	Rank	Dominant	At 5%
4_345	84.16	163.2	90.4, 74.5, 81.4, 78.6, 79.4, 17.4	3	FH_T	TP
4_450	50.72	148.2	62.6, 128.5, 128.7, 47.7, 74.6, 135.9	1	FH_T	TP
1_46	36.28	114.4	69.6, 59.1, 111.2, 63.0, 93.6, 107.4	5	FH_T	TP
4_293	22.41	122.4	115.0, 101.5, 102.2, 80.0, 118.3, 31.6	2	FH_T	TP
4_56	21.37	78.3	79.0, 108.9, 103.1, 98.2, 85.5, 112.5	9	FH_T	TP
4_330	20.75	154.2	123.7, 137.9, 135.6, 131.7, 129.4, 138.6	40	FH_T	FN

Error and reference in degrees; the detector order is `a1b1_ema`, `convnext_s42`, `convnext_s513`, `hrnet_ns`, `hrnet_v2`, `round8_l2`. Rank: position under AoP-space disagreement (1 = most suspicious of 501). Dominant: landmark with the largest radial error of the deployed prediction. TP: referred at a 5% budget; FN: not referred.

Table S12: Timing of the four systems

System	Detectors	Passes	Median (ms)	95th pct (ms)	Peak (MiB)
S1	<code>hrnet_ns</code>	1	35.8	37.8	576
S1+IC	<code>hrnet_ns</code>	9	290.2	297.9	587
S1+IC batched (secondary)	<code>hrnet_ns</code>	9	67.9	69.8	760
S3	<code>hrnet_ns</code> , <code>hrnet_v2</code> , <code>round8_l2</code>	3	89.9	92.2	1276
S6	<code>hrnet_ns</code> , <code>convnext_s42</code> , <code>convnext_s513</code> , <code>hrnet_v2</code> , <code>a1b1_ema</code> , <code>round8_l2</code>	6	159.9	162.9	2252

NVIDIA GeForce RTX 4090, PyTorch 2.3.0+cu121, CUDA 12.1, precision auto (bfloat16 autocast), batch size 1, 20 warm-up and 200 timed iterations with CUDA synchronisation; peak allocated memory with only the system's detectors loaded. Generated by `scripts/time_b_systems.py`.

S6 Analyses added after review: detector subsets and inference protocol

Both analyses were added after a pre-submission review. They were not in the analysis plan (Section S1, last deviation) and are post hoc. They are computed by `scripts/analyze_b_protocol_and_subsets.py` from the stored pool and, for the inference protocol, from the outputs of `scripts/check_b_inference_protocol.py`.

S6.1 Detector subsets trained on labelled frames only

Two detectors of the pool have gaps in their records (Section S2): the pseudo-label file of `hrnet_ns` was not retained, and `round8_l2` was trained with synthetic images whose quality gate was not met. No run directory records command-line warm starts, so the subsets are defined by what the configurations specify. The configurations of `hrnet_v2`, `convnext_s42` and `convnext_s513` specify supervised training on the labelled frames from ImageNet-pretrained backbones, with no pseudo-labels, synthetic images or documented warm start; they form the label-only subset. `a1b1_ema` was fine-tuned, according to its configuration, from the `hrnet_v2` checkpoint, a warm start passed on the command line and not recorded in its run directory; adding it gives a four-detector subset. Each subset was evaluated as a system of its own: its coordinate-wise median is its prediction, its failures are its own, and its disagreement scores, primary contrast, review-workload grid and fixed validation thresholds were computed exactly as for the full pool (Table S13). A permuted AoP-space disagreement of the label-only subset ranked at chance (worst-decile area under the curve 0.424).

The eight failures above 20° of the label-only subset include the six of the full pool, and the four of the four-detector subset are among those six. The ranking effect held in both subsets, but how many gross failures a 5% budget referred depended on the composition (five and six of eight, two of four), and so did the realised rate of the fixed thresholds.

Table S13: Detector subsets trained on labelled frames only

Detectors	Mean ($^{\circ}$)	$>20^{\circ}$	At 90%	Primary contrast	5%	10%	Threshold 5%
Full pool (six)	4.615	6	3.969 / 3.986	+0.017 [-0.101, 0.099]	5/6 / 5/6	6/6 / 6/6	11.2 / 6.4
Label-only (three)	4.773	8	4.295 / 4.187	-0.107 [-0.531, 0.190]	5/8 / 6/8	6/8 / 7/8	6.0 / 4.8
+ fine-tuned	4.675	4	4.064 / 4.062	-0.002 [-0.116, 0.108]	2/4 / 2/4	4/4 / 4/4	10.6 / 6.2
HRNet-W48 (four)							

Label-only: **hrnet_v2**, **convnext_s42** and **convnext_s513**, whose configurations specify supervised training on the labelled frames from ImageNet-pretrained backbones, with no pseudo-labels, synthetic images or documented warm start. The four-detector subset adds **a1b1_ema**, fine-tuned according to its configuration from the **hrnet_v2** checkpoint, a warm start that its run directory does not record. Each subset forms its own median prediction and is judged on its own failures above 20° . At 90%: retained error ($^{\circ}$), AoP-space / coordinate disagreement. Primary contrast: coordinate minus AoP-space disagreement at 90% coverage [95% paired frame-bootstrap interval]. 5% and 10%: own failures above 20° referred, AoP-space / coordinate disagreement. Threshold 5%: realised percentage of test files flagged by the threshold for a nominal 5% fixed on the 100 validation frames, AoP-space / coordinate disagreement. Post hoc, added after review.

S6.2 Inference protocol and forms of confidence

The six detectors were re-run on the 100 validation and 501 test frames under four settings, bfloat16 autocast or float32 and batches of four or one frame, on the re-rented server (**scripts/check_b_inference_protocol.py**). The positive control, set before the run, was a numerical criterion on the re-run of the stored setting: for every detector a median coordinate difference from the stored predictions of at most 0.05 px with at least 95% of coordinates within 0.5 px, and confidences within 10^{-4} . It was not met. The median differences were below 0.02 px for every detector (12 to 14 px when the stored arrays were shifted by one frame, the negative control), and the E-MHA scores were reproduced to within 10^{-3} , but only 91 to 92% of the **convnext_s42** coordinates were within 0.5 px, and confidences differed by up to 2.2×10^{-3} . We report this rather than relax the criterion. Closeness of the outputs does not by itself show which models were run; that rests on the records. The re-run loaded the checkpoint files at the audited paths (Section S2) with the same configurations and preprocessing code, and in the first re-run on the same server these files had reproduced the stored coordinates bit for bit. The remaining differences are therefore attributable to the execution environment, whose effect is what this check measures. Table S14 compares each setting with the stored predictions, and Table S16 gives the primary contrast under each setting.

Because the peak sigmoid of the three HRNet-W48 detectors equals 1.0 on most frames, heatmap confidence was also computed from the peak logit, which cannot saturate. Table S15 ranks the same stored prediction by both forms of confidence and by E-MHA, in bfloat16 and in float32. At full precision the stored mean confidence over the six detectors takes 486 distinct values on the 501 test files (256 when rounded to 10^{-6}), so the confidence ranking is not dominated by ties.

Table S14: The six-detector pool re-run under four inference settings

Setting	Moved	Max (px)	Max AoP	Mean	Failures	At 90%	ρ	Jaccard	5%/10%	>20° caught
Stored predictions (reference)	–	–	–	4.615	47/13/6	3.969	–	–	–	5/6, 6/6
bfloat16, batch 4 (re-run)	352	5.62	1.30	4.618	44/13/6	3.981	0.986	1.000 / 0.961	5/6, 6/6	
bfloat16, batch 1	365	5.62	1.30	4.603	46/13/6	3.966	0.987	0.923 / 0.961	5/6, 6/6	
float32, batch 4	8613	32.76	3.12	4.558	45/12/6	3.933	0.953	1.000 / 0.786	5/6, 6/6	
float32, batch 1	8614	32.77	3.12	4.558	45/12/6	3.933	0.953	1.000 / 0.786	5/6, 6/6	

IUGC2025 test split, compared with the stored predictions used throughout (made with bfloat16 autocast in batches of four on an earlier machine). Moved: landmark positions (of 9,018) displaced by more than 0.1 px; Max: largest coordinate displacement; Max AoP: largest change of the deployed AoP (°). Mean: mean AoP error (°). Failures: files above 10/15/20°. At 90%: retained error of AoP-space disagreement (°). ρ : Spearman correlation of the disagreement with that of the stored predictions. Jaccard: overlap of the files referred at 5% and 10% with those referred under the stored predictions. Caught: own failures above 20° referred at 5% and 10%. The re-run of the stored setting reproduced the stored coordinates with a median difference of 0.010 px (97.8% within 0.5 px), not bit for bit. Post hoc, added after review.

Table S15: Heatmap confidence and E-MHA in bfloat16 and float32

Score	Precision	Distinct	90%	80%	70%	50%	AUC	>20° caught
Confidence, peak sigmoid	bfloat16	490	4.212	4.209	4.281	4.174	0.590	4/6, 4/6
	float32	490	4.216	4.225	4.314	4.196	0.590	4/6, 4/6
Confidence, peak logit	bfloat16	489	4.269	4.146	3.868	3.600	0.726	1/6, 2/6
	float32	493	4.269	4.146	3.862	3.611	0.728	1/6, 2/6
E-MHA, normalised maps	bfloat16	492	4.247	4.186	4.206	4.359	0.457	4/6, 4/6
	float32	493	4.244	4.158	4.211	4.399	0.462	4/6, 4/6
E-MHA, raw sigmoid maps	bfloat16	487	4.152	4.220	4.138	3.901	0.642	4/6, 4/6
	float32	488	4.162	4.197	4.134	3.935	0.646	4/6, 4/6

Every row ranks the same stored deployed prediction (501 test files). Peak sigmoid: the score of the main text. Peak logit: the mean over detectors and landmarks of the largest heatmap logit, which cannot saturate. In float32 the peak sigmoid equals 1.0 exactly on 96.2%, 92.0%, 70.3% of the frames for `hrnet_ns`, `a1b1_ema` and `hrnet_v2` and on none for the ConvNeXt-V2 detectors. Distinct: distinct score values. 90% to 50%: retained error (°) at that coverage. AUC: worst-decile area under the curve. Caught: failures above 20° referred at 5% and 10%. Peak logit minus peak sigmoid at 90% coverage in float32: +0.053° [95% paired frame-bootstrap interval -0.245, 0.525]. Post hoc, added after review.

Table S16: The primary contrast under each inference setting

Setting	AoP-space	Coordinate	Coordinate minus AoP-space [95% CI]
Stored predictions (reference)	3.969	3.986	+0.017 [-0.101, 0.099]
bfloat16, batch 4 (re-run)	3.981	3.989	+0.008 [-0.111, 0.097]
bfloat16, batch 1	3.966	3.974	+0.007 [-0.112, 0.104]
float32, batch 4	3.933	3.905	-0.029 [-0.137, 0.078]
float32, batch 1	3.933	3.905	-0.029 [-0.137, 0.078]

Retained error (°) at 90% coverage of each setting’s own deployed prediction, ranked by AoP-space or by coordinate disagreement, and their difference with the paired frame-bootstrap interval of the main analysis (2,000 resamples, the same seed). A permuted coordinate disagreement gives +0.503 [0.213, 0.817], so the contrast can detect a difference when there is one. Computed by `scripts/analyze_b_protocol_p1.py`. Post hoc, added after review.

S7 Oracle headroom and pool size

A per-frame oracle is a minimum over the S sources of the pool, and a minimum over any dispersed pool falls with S whether or not the pool carries information. The intrapartum pool has not saturated: averaged over all $\binom{6}{k}$ subsets of each size, the oracle is 5.242, 3.669, 3.054, 2.698, 2.459 and 2.282° for $k = 1, \dots, 6$, still descending at the full pool (Figure S1). Part of the 2.33° that separates the deployed median ensemble from the oracle is therefore a statement about S .

The standard remedy is to discount the observed extremum by its expectation under an information-free null whose dispersion matches the real pool [1, 2]. We built that null three ways, which differ in how faithfully they reproduce the observed dependence between sources, and the simplest admits two readings. Drawing sources independently with each frame’s observed error spread gives a null oracle of 0.58° when the draws are centred at zero error and 2.42° when they are centred at the frame’s mean signed error; a multivariate normal reproducing the observed correlation matrix gives 2.83°; a Gaussian copula reproducing both that correlation and each source’s empirical error marginals gives 1.80° (Table S17). The real pool, at 2.28°, is therefore 1.70° worse than chance under the first reading of the independent null, 0.14° better under the second, 0.55° better under the multivariate normal and 0.48° worse under the copula. The diagnostic also fails its own positive control: a synthetic pool built to carry information is not separated from its dispersion-matched null. The constructions disagree by more than the quantity they are meant to measure, so no chance-corrected number is quoted. What survives without a null is that the headline oracle is partly a function of pool size and that no label-free rule we tested reaches more than a small fraction of it (Sections S19 and S20).

Table S17: Three dispersion-matched nulls for the six-source per-frame oracle (real pool 2.28°)

Null construction	Null oracle (°)	Real – null (°)
Independent, matched spread, centred at zero error	0.58	+1.70
Independent, matched spread, centred at the frame mean	2.42	–0.14
Multivariate normal, observed correlation	2.83	–0.55
Gaussian copula, correlation and error marginals	1.80	+0.48

A negative difference favours the real pool. The constructions disagree by more than 2°, more than the shortfall they are meant to measure.

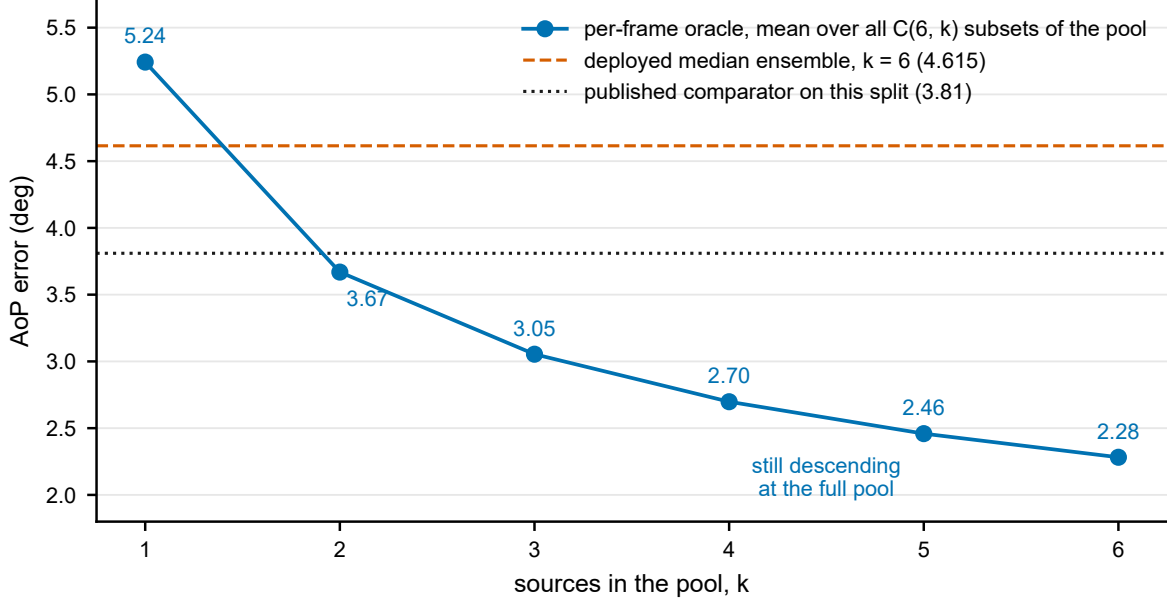


Figure S1: Per-frame oracle of the intrapartum pool against pool size S , averaged over all $\binom{6}{k}$ subsets. The curve is still descending at the full pool.

S8 Why disagreement flags the failures: mechanism, precondition and robustness

Difficulty, not decorrelation. Measured on the six predicted AoP *values*, the mean pairwise correlation between sources is 0.92 over the 495 frames with an error of at most 20° and 0.41 over the six failures above 20° . Measured on the signed per-frame *errors*, which is the quantity an argument about collinear sources requires, there is no such drop: 0.71 on the bulk against 0.72 on the tail, and 0.75 overall (Table S18). A null that draws six frames at random from the bulk 2,000 times places the observed tail value correlation at its 0.6th percentile but the tail error correlation at its 33rd, so only the value contrast is unusual, and part of it may be restriction of range on six frames. What changes on the failures is the *magnitude* of disagreement, from a bulk mean of 2.68° to 21.25° , with the error correlation essentially fixed. Disagreement therefore behaves as a detector of difficulty rather than of independence: the sources keep erring together on hard frames, only more loudly, which is also why frames on which they err together quietly escape it (main text, failure audit).

Table S18: Inter-source structure on the bulk and on the failures above 20° (six-source pool)

Frame group	n	Corr. (values)	Corr. (errors)	Mean disagreement ($^\circ$)
Bulk (error $\leq 20^\circ$)	495	0.92	0.71	2.68
Failures (error $> 20^\circ$)	6	0.41	0.72	21.25
All	501	0.88	0.75	—
Bulk-resampled null ($n = 6$, 2,000 draws)	—	0.92 ± 0.11	0.75 ± 0.17	—

A single-policy control. On a cephalometric pool of a different kind, 64 stochastic rollouts per image from a single landmark-refinement policy at a fixed exploration scale (150 images, 19 landmarks; Sections S11 and S12), ranking images by the positional spread of the rollouts did no better than random at $\sigma \in \{3, 6, 10\}$ px: the gap against random spanned zero at every

coverage and no worst-decile image was flagged at a 10% rate. In that pool the spread carried no information about whether the underlying prediction was right, whereas an ensemble of functionally distinct detectors disagrees more exactly where the detectors fail differently. The requirement is spread that varies with the input, not error independence; this is the landmark-measurement instance of the difference between deep ensembles and stochastic sampling [3, 4]. Whether a pool’s spread varies with the input can be checked on unlabelled data; whether that spread tracks error cannot, and needs labelled validation.

Robustness to pool composition. Adding a conditional diffusion refiner’s per-frame prediction as a seventh source (a refiner of the kind described in Section S14, trained on the labelled split for a fixed schedule without checkpoint selection) moves the oracle only marginally, from 2.28° to 2.18° (frame-bootstrap intervals $[1.97, 2.69]$ and $[1.85, 2.59]$). The disagreement signal is undiminished: the gap against random at 90% coverage is 0.66° (frame-bootstrap interval $[0.27, 1.19]$), and a 5% flag recalls five of the six failures above 20° . The primary label-free selection rule on the seven-source pool, a softmin over cross-source disagreement at $\tau = 1^\circ$, reaches 4.55° against a seven-source median of 4.68° , a small gain (-0.136° , $[-0.236, -0.039]$) that recovers 5.5% of the nominal gap; the minimum of the exploratory sweep, at $\tau = 5^\circ$, reaches 4.54° , 5.8% (Table S33). Removing sources keeps the gain positive under two constructions, recomputed from the stored pool (`scripts/analyze_b_subset_robustness.py`). When each subset forms its own median prediction and is judged on its own errors, the gain against random at 90% coverage is 0.60 to 0.71° for the six five-source subsets and at least 0.48° for the 35 three- and four-source subsets, and a 5% flag refers at least 40% of the subset’s own failures above 20° (these number four to nine). When the deployed six-source prediction and its six failures are kept and only the disagreement is computed from the subset, the gain is 0.61 to 0.70° and at least 0.45° , and at least three of the six failures are flagged at 5%. The earlier summary of this analysis mixed the two constructions; the script that produced it was not retained. These subsets are formed by leaving sources out, not by training records; the label-only subsets are in Section S6.

Part B: earlier exploratory work on the same pools

Sections S9 to S22 document earlier analyses of the detector pools studied here. Their aims, stated in Section S9, were a landmark-refinement module and the scaling of its test-time compute, not quality control; their evaluation designs are their own (for example, cephalometric Test1 served there as the main split). They are kept unchanged in substance so that the construction of the pools and the full sweep of selection rules can be inspected, and so that no unfavourable result of that work is hidden.

S9 Problem setup and notation

Let $\mathcal{X}$ denote the input image space (medical scans, in our case intrapartum ultrasound or cephalometric X-ray) and $\mathcal{Y} = \mathbb{R}^{K \times 2}$ the space of K -keypoint configurations in image pixel coordinates with the $[y, x]$ ordering convention. We assume access to a *frozen base detector* $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ trained by any standard supervised method (HRNet, ConvNeXt, V-Mamba, ...) on a labelled set $\{(x_i, y_i^*)\}_{i=1}^n$.

Our goal is a **light-weight refinement module** $g_\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Y}$ that maps $f_\theta(x)$ to a corrected prediction $\hat{y}(x) = g_\phi(x, f_\theta(x))$ such that the refined metric outperforms f_θ alone, and the refinement compute can be scaled at test time to trade additional compute for accuracy. The parameter count $|\phi|$ should be at most a few million, orders of magnitude smaller than $|\theta|$.

The design choices in this paper are motivated by three observations:

1. The residual error $r(x) := y^*(x) - f_\theta(x)$ is concentrated in a small per-landmark patch around the base prediction. For our base HRNet detector on IUGC2025, the 95th percentile of $\|r(x)\|_\infty$ is approximately 32 px on a 512×512 image; on cephalometric the corresponding value is ~ 60 px on a 768×768 image [5].
2. Landmarks have anatomical dependencies (PS_L is constrained by PS_R via the pubic symphysis axis; FH_T must lie on the fetal head outline). Refining one landmark conditioned on the others is more reliable than independent per-landmark refinement.
3. Stochastic policy rollouts admit test-time scaling via averaging (Section S12).

S10 Extended related work

S10.1 Medical landmark detection

The dominant paradigm in medical anatomical landmark detection is heatmap regression on a strong fully-convolutional backbone, supervised against a fixed-width 2D Gaussian target. The high-resolution representation learning network of Sun et al. [6] and its visual-recognition generalisation [7] have become the de facto backbone for the cephalometric and intrapartum-ultrasound benchmarks, partly because their multi-resolution feature aggregation matches the sub-pixel precision demanded by clinical measurement protocols. On the ISBI 2015 cephalometric challenge [5], the winning entry was the random-forest regression-voting system of Lindner et al. [8], which reached 1.67 mm mean radial error on the Test1Data split and already exceeded inter-observer variability. Deep-learning successors have since consolidated a 1.1 to 1.2 mm range on that same split: 1.17 mm with attentive feature-pyramid fusion [9], 1.12 mm with an attention-guided regression model [10], 1.20 mm with contour-hugging heatmaps [11], and 1.11 mm with a diffusion-generated heatmap [12], while cascaded convolutional networks report comparable accuracy [13]. The current best on that split is lower still, at 0.98 mm with a transformer regression network [14]. On the larger multi-centre CL-Detection 2023 benchmark the best methods reach 1.518 mm over 38 landmarks [15]. We therefore take 0.98 mm as the current

cephalometric reference and state the comparator explicitly at every use, because the choice between 0.98, 1.11 and the 1.67 mm challenge-era figure changes an oracle margin by an order of magnitude and the word “SOTA” on this benchmark has meant all three within a decade. Two structural limitations persist despite these gains. First, an MSE-trained heatmap fits a prescribed Gaussian target over the whole map; that does not by itself make the fitted map a calibrated conditional density, and the consequences for downstream uncertainty quantification and test-time selection are documented quantitatively in Section S19. Second, the training-set sizes for these benchmarks (typically 150 to 500 labelled images) cap the representational capacity of any train-from-scratch image-conditional ranker, motivating refinement modules that build on top of an existing detector rather than retrain the whole pipeline.

S10.2 Reinforcement learning for landmark detection

Reinforcement learning has been applied to anatomical landmark detection for some years, and the line did not stop in 2019. Ghesu et al. [16] train a multi-scale deep RL agent that *navigates through the image grid* starting from a random pixel and converges on the target landmark, with episode length scaling with the image diameter. Vlontzos et al. [17] generalise this to multiple landmarks by assigning one independent RL agent per landmark, each navigating from a separate random start point, and Leroy et al. [18] let those agents share weights and learn an explicit communication channel. The line has continued since: Kang et al. [19] stage the search across resolutions for 3D cephalometric landmarks, and Wan et al. [20] add an implicit global shape prior so that agents can localise targets missing from the field of view. All of these treat landmark detection as a spatial-search problem and operate without a pre-trained base detector. The Anatomy Sequence Policy of Section S11 differs from all of them in refining a *given* base prediction instead of searching from scratch, so its episode length is $O(K)$ landmarks rather than $O(\text{image diameter})$, and it can be applied plug-and-play on top of any frozen base detector with a parameter count (under one million) independent of the backbone. Its remaining ingredient, decoding landmarks one at a time in a fixed anatomical order with each step conditioned on the configuration already decoded, is not new and we do not claim it: Gkioxari et al. predict body parts sequentially in a kinematically motivated order, each conditioned on all previously predicted parts, and explicitly contrast that design against simultaneous refinement [21], and fixed-order autoregressive keypoint decoding has since been folded into general sequence interfaces for vision [22]. What is specific to ASP is that the autoregression operates on *corrections* to an existing prediction rather than on de novo localisation, and that it does so for clinical anatomical landmarks at a scale where retraining the detector is not an option.

S10.3 Test-time compute scaling

The test-time compute scaling literature in language models, exemplified by Snell et al. [23], trades inference compute for output quality by generating N candidate continuations and then selecting or aggregating them, typically with a learned process-reward model. In that domain the empirical exponent of the scaling law falls in the range $c \in [0.3, 0.5]$, driven primarily by best-of- N selection over a multi-modal generation distribution. The picture is not confined to one study. Coverage over repeated samples follows a power law across four orders of magnitude while the methods available for picking from those samples plateau as soon as answers cannot be verified automatically [24]; the exponent has since been given a distributional explanation in terms of a heavy-tailed per-problem success rate [25]; and compute-optimal inference has been mapped as an explicit trade-off between model size and sampling budget [26]. In vision the same lever appears as a verifier-guided search over initial noise in diffusion sampling [27]. A closely related but operationally distinct framework is test-time training [28], which updates *model weights* on individual test inputs through a self-supervised auxiliary task. The refinement pipeline borrows the scaling-with-compute idea from the test-time compute literature but leaves

all model weights frozen at test time; in our medical-landmark domain the scaling lever is mean averaging over a continuous stochastic policy (Section S12) or over a conditional diffusion-policy sampler (Section S14). The cross-dataset exponents we recover in Sections S17, S18 and S21 ($c \in [0.69, 0.95]$) are not compared numerically with that band: the two literatures fit different quantities (a success rate against compute there, a mean error against rollout count here), and Section S17 explains why no ordering of the exponents follows. The methodologically important distinction is that our scaling comes from *averaging* rather than from selection over a multi-modal distribution.

Two prior threads of work resemble the test-time mean-averaging mechanism operationally without ever parametrising the scaling law. **Stochastic Weight Averaging** [29] averages *weight* snapshots collected along an SGD trajectory, yielding a single deterministic model whose predictions improve calibration and out-of-distribution robustness. **Test-time augmentation** [30] instead averages *predictions* across a fixed augmentation suite applied at inference time, mostly for aleatoric uncertainty estimation in medical image segmentation. The refinement pipeline departs from both: we average *predictions across stochastic policy rollouts* parameterised by an explicit exploration scale σ (Section S12), which permits an σ -by- N scaling-exponent characterisation rather than a fixed-budget ensemble. The mean-averaging mechanism itself is thus not new, and neither is fitting a power law to a test-time sampling budget [24, 25]. What we have been unable to find reported anywhere is a parametric power-law exponent c fitted to a medical landmark or keypoint task: the test-time-scaling literature in medical imaging is, to our knowledge, entirely concerned with language and vision-language reasoning rather than with geometric localisation. On that narrower reading, the refinement pipeline provides an empirical scaling-law characterisation of c across two medical landmark benchmarks.

S10.4 Diffusion policy and conditional generative refinement

Denosing diffusion probabilistic models [31] and their accelerated DDIM sampler [32], together with the cosine noise schedule of Nichol and Dhariwal [33], have become standard tools for conditional generative modelling in continuous spaces. The “diffusion policy” formulation, which uses a denoising network as the policy class for continuous action prediction, originates in robotic visuomotor learning [34]: there the model conditions a DDPM on observation features and predicts joint-command actions through iterative denoising. The refinement pipeline’s diffusion refiner (Section S14) adapts this formulation to *medical landmark refinement*, where the “action” is a per-landmark coordinate offset rather than a robot joint command, conditioned on bilinearly-sampled backbone features at the base prediction and on a learned landmark-identity embedding. Diffusion models have already been applied to medical landmark detection in a different role, as the generator of the heatmap itself [12], and multi-hypothesis keypoint diffusion is established in human pose estimation, where it is motivated by precisely the over-confidence of a unimodal predictor that concerns us here [35]. We should be careful about which part of this is new, because the architecture is not. Denoising the residual of a fixed base predictor in coordinate space is an established pattern outside medical imaging: Kang et al. refine the output of a pre-trained 3D pose estimator by reverse diffusion, freezing the base encoder in their released implementation [36]; Yan et al. model the base predictor’s error distribution explicitly and denoise it back to the ground truth using the original authors’ checkpoints unchanged [37]; and DiffuBox refines the boxes of off-the-shelf 3D detectors, in its own words, without any re-training [38]. What we claim is therefore the instantiation and the comparison rather than the mechanism: to our knowledge this is the first coordinate-offset diffusion policy for *medical anatomical* landmarks, where the published diffusion work operates in heatmap space or on data generation, and a matched- N comparison in which the diffusion policy reached a lower error than the Gaussian policy we implemented (Section S21); the two differ in conditioning features, training objective and sampling cost as well as in the density they model, so this is not evidence that multi-modality is the cause. We note that the pose-estimation papers above

sweep the hypothesis and step budget rather than matching it, so the comparison they report is not the same one; two further journal papers in that line were not accessible to us in full text, and we cannot exclude that a matched-budget contrast appears in one of them.

S10.5 Plausibility verification and selection on continuous policies

The plausibility verifier of Section S13 is conceptually adjacent to the process-reward models of Snell et al. [23] and to the preference-based reward learners of Christiano et al. [39], Ouyang et al. [40], and Rafailov et al. [41]. The difference is narrow: it is trained on *perturbed-ground-truth negatives* rather than on human preferences over model outputs. The closely related literature on learning-to-defer in machine-learning-aided clinical decision making [42] provides theoretical grounding for an abstention/expert-deferral interpretation of the plausibility verifier. Section S19 reports a systematic ablation showing that the perturbed-GT plausibility verifier *fails* to recover the per-landmark oracle headroom on diffusion sample pools (Section S20), as does every alternative selection rule we tested, including image-independent anatomy-graph priors. This documented negative result leaves open whether a selector that reads the image could do better.

The gap that this failure exposes has been named more than once in the language-model literature, and we adopt none of the existing names only because each was coined for a discrete generator. Saad-Falcon et al. call the distance between a practical verifier and a perfect one the *generation-verification gap* [43]; Hu decomposes the net gain of a fixed-pool selector into recoverable mass, verification-signal coverage and conditional selection quality, and calls the first term the *oracle gap* [44]. Brown et al. had already reported the operational symptom, that majority voting and reward-model selection plateau while the pool’s coverage keeps improving [24], and a compute-normalised study across five open-ended benchmarks concludes that exploitation rather than exploration is the binding constraint [45]. Concurrently with this paper, Gröger et al. report the same shape in medical imaging: an oracle that picks which of a case’s dermatology photographs to classify gains roughly twenty points of weighted F1, and four training-data-free selectors recover almost none of it [46]. Their candidates are alternative *input images* for a classification label; ours are alternative *predictions* of one continuous geometric quantity, and the abstention analysis of the main text has no counterpart in that study.

S10.6 Selective prediction, deep ensembles, and the closest prior art

Our abstention result sits in a mature literature, and we are explicit about what is and is not new. Deep ensembles yield calibrated uncertainty whose disagreement flags hard and out-of-distribution inputs [3], and a loss-landscape account explains why independently trained members explore distinct functional modes and therefore disagree on difficult points [4], a perspective recently tied to out-of-distribution detection [47]. The picture is not unanimous: Abe et al. argue that ensemble diversity contributes little to out-of-distribution uncertainty beyond what a single larger model already provides [48], which makes the conditions under which disagreement is informative an empirical question rather than a settled one. At the level of a whole prediction, the disagreement between two independently trained networks tracks generalisation error closely enough to estimate it without labels [49]. Selective prediction and learning-to-defer formalize abstention-with-referral [42, 50, 51], the risk-coverage curve is the standard instrument for evaluating it [52], and the medical-ML community has specifically argued for communicating model uncertainty and deferring to a clinician rather than predicting autonomously [53].

Three works sit closer to our setting than any of the above, and we state plainly what they already establish. Roy et al. use the pairwise agreement between ensemble members as a label-free, structure-wise quality-control signal in whole-brain segmentation [54]. Zenk et al. benchmark failure detection across five public 3D medical collections under realistic distribution shift, argue for risk-coverage analysis as the appropriate evaluation, and find the pairwise

agreement between ensemble predictions to be the strongest simple baseline [55]. Closest of all, Schobs et al. rank heatmap-based landmark predictions by uncertainty, including an ensemble coordinate-prediction-variance measure that is exactly a cross-source disagreement statistic, and show that discarding the highest-uncertainty bin removes gross mispredictions, on cardiac MRI and on the same public cephalometric dataset we use here [56]. Conformal methods now supply finite-sample-valid regions for the same localisation task [57]. Within our own modality, U-LanD ranks frames by a single model’s Bayesian uncertainty to *select* reliable frames in ultrasound landmark video [58]. We therefore claim neither that disagreement-detects-error is a new mechanism nor that ranking landmark predictions by an ensemble variance is new; Schobs et al. did the latter, on one of our two datasets. Our contribution is orthogonal on three counts. First, we state and test the *dual* of selection, namely that the same redundant pool which defeats ground-truth-free *selection* still admits ground-truth-free *abstention*, and we do so on a continuous geometric clinical measurement (AoP in degrees) that is a nonlinear function of several landmarks rather than on per-landmark radial error or a classification label. The distinction is operational rather than cosmetic: a per-landmark uncertainty ranking does not by itself say which *measurement* is wrong, because the measurement error is governed by whichever landmark the ranking happens to miss, and the 20° AoP failures analysed in the main text are exactly such cases. Second, where U-LanD selects good frames with single-model uncertainty and Schobs et al. bin per-landmark predictions, we abstain on catastrophic *measurements* using cross-source disagreement, and we characterize *when* that signal is available. The condition is input-dependent spread, which the ensembles of functionally distinct predictors we evaluated had and the single-model pool we tested, one policy sampled at a fixed exploration temperature, did not; the single-model control is consistent with that account, though it does not show that a fixed temperature must produce input-independent spread, since the policy is recurrent and re-reads the image at every step (Section S8). Notably, the condition is *not* error independence on the hard cases: Section S8 shows that the members’ errors remain just as correlated on the catastrophic tail as on the bulk, and only the magnitude of their disagreement changes. Third, we ask how much of the apparent oracle headroom is recoverable at all, report the oracle-versus- N curve rather than a headline minimum, and show that three defensible dispersion-matched nulls disagree by more than the effect they are meant to measure, so we decline to attach a number to the shortfall (Section S7). Consistent with this scope, we present the tail behaviour as a measured regularity with its small-sample caveat (Section S8), and we report where our own earlier reading of it was wrong.

S11 Anatomy Sequence Policy: architecture and training

S11.1 Anatomy Sequence Policy (ASP) architecture

High-level design ASP is a recurrent policy that visits landmarks in a chosen sequence and refines each one conditioned on the running state of the other landmark positions. Formally,

- **State** at step t : the current landmark configuration $c_t \in \mathbb{R}^{K \times 2}$, initialised at the base detector output $c_0 = f_\theta(x)$.
- **Action** at step t : a 2-dimensional offset $a_t \in \mathbb{R}^2$ applied to the landmark selected for refinement at this step.
- **Transition**: c_{t+1} equals c_t except at the selected landmark’s row, which is updated to $c_t[\text{lm}, :] + a_t$ (clipped to image bounds).
- **Episode length** T : typically equal to K (each landmark refined once) or shorter with revisits if the selector is learned.

Components

1. **Image patch encoder** E_ξ : a small ($\sim 140\text{K}$ params) CNN that maps a 64×64 RGB patch cropped around the current landmark position to a d -dimensional feature ($d = 128$).
2. **GRU memory** with hidden size 128, taking as input the concatenation of (a) the patch feature, (b) the step embedding, (c) the landmark-id embedding, and (d) the flattened normalised current configuration c_t/H .
3. **Landmark selector** (optional, used when no fixed-order schedule): a linear layer from the GRU hidden state to K logits, predicting which landmark to refine next.
4. **Offset head**: two parallel linear heads from the GRU hidden state, producing the Gaussian policy’s mean $\mu_\phi \in \mathbb{R}^2$ and log-standard-deviation $\log \sigma_\phi \in \mathbb{R}^2$. The mean is squashed through tanh and scaled by `max_offset_per_step` (16 px) so it stays within the patch.

Total parameter count: 0.88M ($K = 3$) or 0.89M ($K = 19$) — independent of the base detector backbone.

Imitation-learning warmup We train ASP in two stages. The first stage is supervised:

- Run the base detector to get $c_0 = f_\theta(x)$ and freeze it.
- For each step t in the (fixed or selector-determined) order, compute the **oracle target offset** $\tilde{a}_t = \text{clip}(y^*(x)[\text{lm}, :] - c_t[\text{lm}, :], \pm 16)$.
- Apply a Huber loss between the policy mean μ_ϕ and $\tilde{a}_t$, with $\text{delta} = 2$ px. If the selector is learned, add a small cross-entropy term against the greedy-error oracle (which landmark has the largest error right now).

Imitation warmup is *deterministic* — only μ_ϕ is supervised; the $\log \sigma_\phi$ head stays at initialisation (bias -1 , weight zero). We exploit this in Section S15 by setting σ post-hoc at inference time.

(Optional) REINFORCE fine-tuning The second stage began as REINFORCE policy gradient, using the reward $R(\tau) = -\|c_T - y^*(x)\|^2$ together with a batch-mean baseline. Empirically, REINFORCE proved **unstable and counterproductive** in this regime for our base detector and dataset sizes. The $\log \sigma_\phi$ head learns to inflate noise during training, the deterministic mean drifts away from the IL warmup optimum, and generalisation worsens on held-out test data (Section S19 ablation). We therefore keep IL-warmup only in the headline results, reporting the REINFORCE variant in the supplementary as a documented negative result with concrete remedies: entropy regularisation, KL-anchored σ , and joint mean and standard-deviation supervision.

This is reported transparently — refining a stochastic policy with REINFORCE is well-known to require careful entropy and trust-region control (see, e.g., PPO and TRPO). Our contribution is independent of this fine-tuning choice.

Differentiation from prior RL landmark detection Two existing lines of work apply RL to anatomical landmark detection:

- **Ghesu et al. [16]** train multi-scale RL agents that *navigate through the image grid* starting from a random pixel to locate landmarks. Episode length scales with the image diameter.
- **Vlontzos et al. [17]** assign one RL agent per landmark, each navigating independently.

ASP differs in three ways:

1. **No image search:** ASP refines an existing base prediction within a small patch (± 16 px). Episode length is $O(K)$ landmarks, not $O(\text{image diameter})$.
2. **Single sequential policy:** one GRU policy decides per-step which landmark to refine and by how much, conditioning on the global configuration. This allows anatomy-aware ordering (refine reliable PS landmarks first to constrain harder FH localisation).
3. **Plug-and-play on any frozen backbone:** ASP’s parameter count and compute are independent of the base detector. We verify this in our experiments by using HRNet on IUGC2025 and HRNet on cephalometric without any ASP architectural change other than K .

S12 Test-time compute scaling via stochastic policy averaging

S12.1 Test-time compute scaling via stochastic policy averaging

Motivation The Anatomy Sequence Policy (ASP) trained under imitation-learning warmup (Section S11) produces a deterministic refinement $\mu_\phi : \mathcal{X} \times \mathbb{R}^{K \times 2} \rightarrow \mathbb{R}^{K \times 2}$ that maps an image and the base detector’s coarse coordinates to a corrected landmark configuration. Treating μ_ϕ as the **mean of a Gaussian policy**

$$\pi_\phi(a \mid x, c) = \mathcal{N}(a; \mu_\phi(x, c), \sigma^2 I),$$

with exploration standard deviation $\sigma > 0$, allows us to generate N independent stochastic refinements per image and aggregate them at inference time, trading additional compute for reduced prediction variance. This is the test-time compute scaling regime studied by Snell et al. [23] for large language models. We adapt it to medical keypoint detection.

N-rollout mean averaging At test time, for each image x we generate N stochastic rollouts $a^{(1)}, \dots, a^{(N)} \sim \pi_\phi(\cdot \mid x, c)$ and return the **arithmetic mean** of the resulting refined coordinates:

$$\hat{y}_N(x) = \frac{1}{N} \sum_{i=1}^N (c + a^{(i)}) = c + \mu_\phi(x, c) + \frac{\sigma}{N} \sum_{i=1}^N \varepsilon_i,$$

where $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$. By the central limit theorem the second term converges in L_2 at rate $O(\sigma/\sqrt{N})$. In this one-step additive model the deterministic component $c + \mu_\phi(x, c)$ is unchanged, so the asymptote $\hat{y}_\infty(x) = c + \mu_\phi(x, c)$ is the deterministic ASP prediction. The deployed policy is not this model: it is recurrent, re-reads the image patch at the current coordinates and updates its state at every step (Section S11), so the noise injected at one step changes the input of the next and the final configuration is not the sum written above. We therefore make no exact-equality claim for the full trajectory; what Section S17 measures is that the fitted asymptote of the averaged rollouts (5.03° by point fit, bootstrap centre 5.00°) and the deterministic rollout (5.07°) agree within the interval on the fit, which is an empirical statement.

We fit the dependence on N empirically as a generalised power law, $\text{metric}(N) = a + b \cdot N^{-c}$, with $c \in [0.69, 0.79]$ on the IUGC2025 test set (Section S17). The exponent is not read against a CLT rate: the metric is a mean absolute angle error, a nonlinear function of averaged coordinates, and no rate for it follows from the $N^{-1/2}$ decay of the coordinate noise; the fitted c is descriptive.

Choice of selection rule We considered three rules to aggregate $\{a^{(i)}\}_{i=1}^N$:

- **Mean averaging** (used in the main results): $\bar{a} = \frac{1}{N} \sum a^{(i)}$. Equivalent to noise denoising under i.i.d. Gaussian exploration.
- **Best-of-N (PV-rank)**: a^{i^*} with $i^* = \arg \max_i \text{PV}(x, c + a^{(i)})$, using a learned plausibility verifier as in Snell et al. [23]. We trained PV on perturbed-GT negatives (Section S13) and found it failed to rank stochastic ASP rollouts (Section S19, ablation); we attribute the failure to the distribution mismatch between PV’s training negatives (Gaussian noise / outliers / label swaps with σ on the order of 10-30 px) and the rollout distribution (small $\sigma \leq 6$ px around the trained mean).
- **Top-K weighted**: a softmax over PV scores, top- k candidates weighted by softmax weight. Also failed for the same reason.

The mean-averaging rule needs **no** learned verifier, and it benefits unambiguously from larger N provided the per-rollout error stays dominated by the variance term. We report it as the main scaling mechanism, with PV-based selection reported as a documented negative ablation.

Role of σ The exploration standard deviation σ governs a tradeoff:

- $\sigma \rightarrow 0$: rollouts collapse to the deterministic mean; mean averaging has nothing to denoise; $c \approx 0$.
- $\sigma \rightarrow \infty$: each rollout is dominated by noise; for finite N the mean is far from the deterministic prediction even after averaging.
- σ comparable to the base detector’s residual error per landmark (~ 2 -6 px on IUGC2025 PS / FH landmarks): optimal scaling.

Empirically we observe c increasing from 0.61 at $\sigma = 1.5$ px to 0.79 at $\sigma = 4$ px, then slightly decreasing at $\sigma = 6$ px (Section S17, Table S23). The fit residual and bootstrap CI width are smallest at $\sigma = 4$ px, suggesting σ in the range $[2, 6]$ px gives the best characterisation for this dataset.

Implementation note (post-hoc σ override) The IL-warmup loss supervises only μ_ϕ , the offset mean head; the log-std head $\log \sigma_\phi$ stays at its initialisation throughout training. At inference, then, we **override** the log-std head’s bias to $\log \sigma_{\text{target}}$ and zero the input-dependent weight, which fixes the exploration standard deviation to any target value without retraining. The `-asp-force-sigma` flag of the scaling experiment script implements this. Treating σ as a **test-time hyperparameter** rather than a learned parameter serves two purposes: it isolates the σ - N tradeoff (Section S17), and it matches the practical scenario where the dataset best σ is determined post-hoc.

What averaging does and does not give In the one-step model, mean averaging over N stochastic rollouts converges to the mean of the trained policy, $\mu_{\hat{\phi}}$, at the usual $N^{-1/2}$ rate, and removes the sampling variance injected at inference time and nothing more; for the recurrent policy actually deployed we observe the same qualitative behaviour empirically (Section S17) and do not derive it. In either case, whatever gap separates $\mu_{\hat{\phi}}$ from the Bayes-optimal residual $\pi^*(x) = y^*(x) - f_\theta(x)$ is a fixed property of the learned weights and does not shrink with N . We are explicit about this because the stronger reading is a natural one to reach for; Section S15.1 sets out why we do not make it and points to the published lower bound that rules it out [59].

The estimator we actually deploy is the imitation-learning warm-up of Section S11, which supervises the policy mean directly against oracle residuals rather than through a policy gradient.

What Sections S17 and S18 then measure is how the post-averaging error falls with N and with the exploration scale σ , reported as an empirical scaling law rather than derived from a bound.

S13 Plausibility verifier

S13.1 Plausibility verifier

Motivation The Anatomy Sequence Policy of Section S11 produces a deterministic refinement of the base detector’s coarse coordinates, and the mean-averaging aggregator of Section S12 reduces stochastic exploration noise around that deterministic prediction. Neither component explicitly evaluates the *plausibility* of a candidate landmark configuration. In the language of test-time-compute scaling for large language models [23], mean averaging corresponds to a no-verifier majority-voting aggregator, so the natural complementary mechanism is a learned per-rollout scorer that ranks the N candidates and concentrates compute on the most plausible one. We describe such a scorer here, the Plausibility Verifier (PV), and report its negative ablation result in Section S19.

Architecture The PV is a small fully-connected classifier mapping a landmark configuration $c \in \mathbb{R}^{K \times 2}$ to a scalar plausibility score $PV_\psi(c) \in [0, 1]$. We want this score to ignore global translation and scale yet stay sensitive to internal geometric structure, so the input c is first turned into a *geometric feature vector* $\phi_K(c)$ and only then handed to the classifier. How the featurisation proceeds depends on the landmark count K :

- For $K = 3$ (the intrapartum-ultrasound configuration with PS_R, PS_L, FH_T), we use the eight-dimensional anatomical featurisation of the refinement pipeline: three pairwise Euclidean distances (PS-PS, PS_R-FH_T, PS_L-FH_T), three pairwise angles, the perpendicular distance from FH_T to the PS line, and the angle of progression itself derived from the three landmarks (Section S9).
- For $K \geq 4$ (the cephalometric configuration with $K = 19$), we generalise to a $\binom{K}{2} + 5$ dimensional vector consisting of the $\binom{K}{2}$ pairwise Euclidean distances, the centroid of the configuration, the principal-axis angle of the configuration as estimated by PCA on the K two-dimensional points, and the spread of the configuration measured as the average distance from the centroid. The PCA eigendecomposition is performed in float32 even when the surrounding training runs use bfloat16, to circumvent a CUDA limitation on bfloat16 eigendecomposition.

The classifier itself is a four-layer multi-layer perceptron with hidden width 128 and ReLU activations, capped by a sigmoid output. Its parameter count comes to roughly 50 000 for $K = 3$ and 70 000 for $K = 19$, in either case independent of the base-detector backbone and of the ASP refiner.

Perturbed-ground-truth training We train the PV as a binary classifier: ground-truth landmark configurations serve as positives and synthetically perturbed configurations as negatives. The Perturbation Generator branches on K in step with the geometric featurisation. For $K = 3$ it combines Gaussian per-landmark coordinate noise of standard deviation $\sigma_{\text{train}} \in [10, 30]$ pixels with anatomy-specific perturbations: a PS_R-PS_L swap, and a “PS line drift” that pushes the two PS endpoints apart in opposite directions. For $K \geq 4$ it pairs the same Gaussian noise with random index-pair swaps and a rotation drift about the centroid, the angle lying in $[-15^\circ, +15^\circ]$. With low probability, each minibatch also contributes an “outlier” negative whose per-landmark noise is sampled from a heavy-tailed distribution with $\sigma \in [40, 100]$ pixels; this

teaches the PV to flag gross failures (for instance, landmark heads collapsed onto the wrong anatomy).

The PV is trained for 50 epochs with batch size 32, AdamW optimiser at learning rate 5×10^{-4} and weight decay 10^{-4} , on the same 150-image (cephalometric) or 300-image (intra-partum ultrasound) training-split GT configurations used by the base detector and ASP. We observe consistent convergence to validation accuracy approximately 0.85 on both datasets, with positives mean score ~ 0.70 and negatives mean score ~ 0.27 .

Inference-time selection rules Given N stochastic ASP rollouts $\{c^{(i)}\}_{i=1}^N$ for a test image, the PV scores each rollout and the final prediction is selected by one of two rules: the **top-1 rule** returns the rollout with the highest PV score, $c^{(\arg \max_i \text{PV}_\psi(c^{(i)}))}$; the **top- k weighted rule** returns the softmax-weighted average $\sum_i w_i c^{(i)}$ with $w_i = \text{softmax}_i(\text{PV}_\psi(c^{(i)})/\tau)$ at temperature τ .

Anticipated failure mode At training time the PV’s negative distribution sits at large-perturbation displacements, since a typical Gaussian negative carries per-landmark $\sigma \in [10, 30]$ pixels and the outlier sub-population reaches farther still. The test-time ASP rollout distribution lives elsewhere, concentrated at the small exploration scale (Section S17 reports that the dataset-best σ for mean averaging falls in the range 2 to 6 pixels). As a result the PV is well-calibrated to *reject distant perturbations*, yet it says little about which member of a tightly-clustered rollout pool sits closest to the ground truth. Section S19 reports this distribution-mismatch failure quantitatively as a negative result, and Section S19 draws the broader synthesis of selection-rule failures together with the three-pronged specification it implies for any future selector.

S14 Diffusion policy refiner

S14.1 Diffusion policy refiner

Motivation The ASP-IL refiner of Section S11 is a unimodal Gaussian policy $\pi_\phi(a \mid x, c) = \mathcal{N}(\mu_\phi(x, c), \sigma^2 I)$ whose deterministic mean is predicted by a Huber-regression-supervised network. As Section S17 measures, mean-averaging over N stochastic samples from this policy approaches a value within noise of that deterministic prediction: the fitted test-time scaling asymptote (5.03° by point fit) and the deterministic rollout (5.07°) are not separable. Section S20 then puts a number on a $4\times$ per-landmark oracle headroom that mean-averaging, on a unimodal pool, cannot reach.

A natural way to escape this limit is to replace the Gaussian policy with a more expressive one: a *conditional diffusion policy* whose stochastic samples explore a multi-modal density over the per-landmark coord offset. Averaging N such samples recovers the mode-marginal posterior mean. Whenever the true conditional posterior has multiple plausible modes, this mean may lie strictly below the unimodal Gaussian’s deterministic prediction in L_2 distance to GT.

We describe the diffusion policy refiner and the corresponding training/sampling procedure here. The empirical result this delivers is reported in Section S21.

Diffusion process on coordinate offsets For each test image x and per-landmark base prediction c_k (obtained from the frozen base detector’s heatmap soft-argmax), the *target offset* is $a_0 = y_k^* - c_k \in \mathbb{R}^2$ (in source-resolution pixels). We learn a noise predictor $\epsilon_\theta(a_t, t, h(x, c_k, k))$ where h is a conditioning vector derived from the image features at c_k and the landmark identity k .

Forward (noising) process following Ho et al. [31]:

$$a_t = \sqrt{\bar{\alpha}_t} a_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

with $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ and β_s a cosine schedule (Nichol and Dhariwal [33]). We use $T = 20$ denoising steps.

Offset normalisation (critical implementation detail): the target offset a_0 on cephalometric Test1Data has empirical per-coordinate standard deviation $\hat{\sigma}_{a_0} \approx 6$ px, which mismatches the unit variance of ε . Naively applying the forward process gives a_t scaled by $\hat{\sigma}_{a_0}$ at low t and by 1 at high t — a $9\times$ scale drift that destabilises optimisation. Following standard practice we normalise the target by a fixed constant κ before noising:

$$\tilde{a}_0 = a_0/\kappa, \quad \tilde{a}_t = \sqrt{\bar{\alpha}_t} \tilde{a}_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$$

and denormalise at sampling time by $\hat{a}_0 = \kappa \cdot \hat{\tilde{a}}_0$, where $\hat{\tilde{a}}_0$ is the network’s estimate of the clean normalised offset $\tilde{a}_0$. We use $\kappa = 8$ px on both cephalometric (768×768) and IUGC2025 (512×512); ablation in Section S21 shows this normalisation is essential — without it, the model fails to learn (Section S19, Pilot 4 v1).

Architecture The noise predictor $\epsilon_\theta(\tilde{a}_t, t, x, c, k)$ is a small MLP conditioned on three signals:

1. **Backbone features at base coord:** the frozen base detector’s HRNet stage-2 feature map ($C = 256$ or 512 channels at stride 8) is bilinearly sampled at the base coord c_k using `grid_sample`, giving a per-landmark feature vector $f_k \in \mathbb{R}^C$. We project f_k to dimension $d_{\text{proj}} = 128$ via a learned linear-GELU-LayerNorm head.
2. **Landmark identity:** a learned embedding $e_k \in \mathbb{R}^{32}$ indexed by $k \in \{0, \dots, K-1\}$. This allows the MLP to apply different bias corrections per landmark, important because the base detector’s mode bias varies by anatomy.
3. **Diffusion-time embedding:** a learned embedding $\tau_t \in \mathbb{R}^{32}$ indexed by $t \in \{0, \dots, T-1\}$.

The final eps-predictor MLP has input dimension $2 + d_{\text{proj}} + 32 + 32 = 194$ and is a 4-layer GELU MLP with hidden dim 256 and dropout 0.1. Total parameter count: $\approx 270\text{K}$, two orders of magnitude smaller than the base detector.

Training We freeze the base detector and train only the diffusion policy. For each minibatch:

1. With `no_grad`: forward the image through the base detector to get the base coordinate and backbone feature at the base coordinate.
2. Sample $t \sim \text{Uniform}\{0, \dots, T-1\}$, $\varepsilon \sim \mathcal{N}(0, I)$.
3. Compute the target offset $a_0 = y_{\text{GT}} - c$, normalise $\tilde{a}_0 = a_0/\kappa$.
4. Add noise: $\tilde{a}_t = \sqrt{\bar{\alpha}_t} \tilde{a}_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$.
5. Predict noise: $\hat{\varepsilon} = \epsilon_\theta(\tilde{a}_t, t, f_k, k)$.
6. Loss: $\text{MSE}(\hat{\varepsilon}, \varepsilon)$.

We use AdamW (LR 5×10^{-4} , weight decay 10^{-4}), cosine annealing over 100 epochs, gradient clipping at 1.0. Total training compute on a single 4090 GPU: ~ 18 min for cephalometric (150 train images), ~ 35 min for IUGC2025 (300 labelled train images).

Sampling at test time We compare two sampling regimes:

Deterministic DDIM ($\eta = 0$): a single noiseless reverse trajectory from $\tilde{a}_T \sim \mathcal{N}(0, I)$, following Song et al. [32]. This is the “single best estimate”, analogous to the deterministic ASP-IL prediction.

Stochastic DDPM ($\eta = 1$): for each test image we draw N *independent* reverse trajectories. Every trajectory starts from its own random $\tilde{a}_T^{(n)} \sim \mathcal{N}(0, I)$ and injects stochastic transition noise at each diffusion step. We then denormalise the N resulting offsets and average them to produce the final prediction:

$$\hat{y}_k = c_k + \kappa \cdot \frac{1}{N} \sum_{n=1}^N \hat{a}_0^{(n)}$$

The mean-averaging interpretation of DDPM samples is identical to that of ASP rollouts in Section S12: N stochastic samples reduce variance at the empirical rate determined by the power-law fit. Two key differences from ASP:

- Diffusion samples are drawn from a *multi-modal* density (rather than from a unimodal Gaussian centred at a fixed mean), so the mean’s asymptote is the *mode-marginal posterior mean*, not the deterministic policy mean.
- The exploration “scale” is implicit in the fixed cosine β_s schedule rather than a tunable post-hoc σ (Section S15).

We empirically observe (Section S21) that DDPM mean averaging breaks the ASP-IL deterministic asymptote by ~ 0.45 mm on cephalometric Test1Data, with a fitted exponent $c \approx 0.95$ (Section S21 explains why we do not read it against a CLT limit).

Compute footprint at inference Each DDIM/DDPM step requires one MLP forward of dimension $194 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 2$ per (image, landmark) pair, so for batch size B and K landmarks the per-step cost is $\sim B \cdot K$ small MLP evaluations. With $T = 20$ steps and $N = 64$ samples, the diffusion refiner adds roughly 100 ms per test image on a single 4090 GPU on top of the base detector forward (~ 15 ms), so the refined pipeline runs at about 115 ms per image, roughly eight times the base detector’s latency. That is affordable for offline or per-frame review; it is not free and it is not a video rate.

Algorithm 1: Diffusion policy test-time sampling. **Input:** test image x , frozen base detector f_θ , frozen diffusion network ϵ_θ , number of samples N , denoising steps T , offset scale κ . **Output:** refined landmark prediction $\hat{y} \in \mathbb{R}^{K \times 2}$.

1. **Base forward:** $c, F \leftarrow f_\theta(x)$ (base coord, backbone features at stride-8).
2. **For each landmark** $k = 1, \dots, K$: sample feature $f_k \leftarrow \text{grid_sample}(F, c_k)$.
3. **Initialise:** draw N independent $\tilde{a}_T^{(n)} \sim \mathcal{N}(0, I_{K \times 2})$.
4. **For** $t = T, T - 1, \dots, 1$:
 - (a) For each (n, k) : compute $\hat{\epsilon}_k^{(n)} \leftarrow \epsilon_\theta(\tilde{a}_{t,k}^{(n)}, t, f_k, k)$.
 - (b) DDPM step ($\eta = 1$): $\tilde{a}_{t-1,k}^{(n)} \leftarrow \mu_{t-1}(\tilde{a}_{t,k}^{(n)}, \hat{\epsilon}_k^{(n)}) + \sigma_t \cdot \xi_k^{(n)}$ with $\xi \sim \mathcal{N}(0, I)$.
5. **Aggregate:** $\hat{y}_k \leftarrow c_k + \kappa \cdot \frac{1}{N} \sum_n \tilde{a}_{0,k}^{(n)}$.
6. **Return** $\hat{y}$.

S15 The exploration scale, and why we state no convergence proposition

S15.1 The exploration scale, and why we do not state a convergence proposition

A reader may reasonably expect this section to state a proposition bounding the excess risk of ASP under REINFORCE training, together with a corollary asserting that test-time mean averaging over N rollouts removes the residual term left by training under exploration noise. We state neither, for three reasons that we set out explicitly because each of them is instructive.

First, such a corollary would be wrong in a way worth naming. Averaging N stochastic rollouts of a trained policy converges to that policy’s mean $\mu_{\hat{\phi}}$, so it removes the inference-time sampling variance, of order σ^2/N , and nothing else. Any deviation of $\mu_{\hat{\phi}}$ from the Bayes-optimal residual induced by training under exploration noise is a fixed property of the learned weights and is invariant in N . It would conflate a variance with a bias. Montenegro et al. analyse precisely this setting, learning with per-step Gaussian exploration and deploying the deterministic mean, and prove a matching lower bound on the resulting deployment gap at fixed training noise [59]; that lower bound is incompatible with the stronger claim.

Second, the parts of such a proposition that would survive are both weaker and less general than results already in the literature. Montenegro et al. give global convergence and sample complexity for the same class of problems, with upper and lower bounds on the exploration-scale dependence. There is no reason for this manuscript to restate a special case.

Third, such a proposition would analyse a training procedure that our experiments do not use. As Section S16.1 records, ASP is trained with imitation warmup only, without REINFORCE, so a REINFORCE convergence bound had no experimental counterpart here in any case.

What remains is an empirical statement, which is what Sections S17 and S18 actually measure. The exploration scale σ governs a trade-off with two ends. Too small and the stochastic pool collapses onto the deterministic prediction, so test-time averaging has nothing to average and the scaling exponent degenerates. Too large and the rollouts wander off the base prediction faster than averaging can recover, so the asymptote rises. On IUGC2025 the dataset-best value falls in the range 2 to 6 px, with $\sigma^* \approx 4$ px, and we report the scaling exponent as a function of σ rather than deriving an optimum. We make no claim that the combined training-plus-averaging operator recovers the Bayes-optimal residual.

Two positioning points stand independently of any convergence result. The statistical cost of ASP scales in the policy’s parameter count rather than the backbone’s, which is what makes a sub-million-parameter refiner trainable on 150 to 300 labelled images when retraining the detector is not. And ASP differs from search-based RL for landmark detection [16, 17] in that its episode length is $O(K)$ in the number of landmarks rather than $O(\text{image diameter})$, because it corrects a given prediction instead of navigating to one from a random start.

S16 Full experimental setup

S16.1 Experimental setup

Datasets We evaluate on two medical landmark-detection datasets that exercise the refinement pipeline method across imaging modality (ultrasound vs X-ray), landmark count ($K = 3$ vs $K = 19$), and clinical task (angle of progression vs cephalometric measurement).

IUGC2025 [60]. Intrapartum ultrasound dataset for fetal-head progression assessment. Each image carries three keypoint annotations, namely the left and right endpoints of the pubic symphysis (PS_R, PS_L) together with the tangent point of the fetal head outline relative to the PS line (FH_T), along with a derived angle of progression (AoP) measurement. The same

anatomy underpins the earlier PSFHS segmentation challenge on transperineal ultrasound [61], from which most segmentation-based AoP estimators descend; the landmark formulation used here bypasses the segmentation stage. Table S19 gives the values discussed here.

Table S19: IUGC2025 dataset splits.

Split	Images	Cases	Use
Training (labelled)	300	2 prefixes (143 / 157)	supervised base detector
Training (unlabelled)	33 466	472 clips + 2 045 single frames	self-training
Validation	100	2 prefixes (45 / 55)	hyperparameter tuning
Test	501	2 prefixes (84 / 417)	final reporting (paper)

Image resolution: 512×512 (after centre-crop and rescale). The release publishes no pixel spacing, so IUGC2025 errors are given in degrees or pixels only. Primary metrics: dAoP (mean absolute angle error in degrees) and MRE_ALL (mean radial error in pixels averaged over the 3 landmarks).

ISBI 2015 cephalometric [5]. Lateral cephalometric X-rays for orthodontic landmark detection. Two expert orthodontists annotate each image with 19 anatomical landmarks (Sella, Nasion, Orbitale, Porion, A/B points, Pogonion, Menton, Gnathion, Gonion, incisor tips, lip points, posterior and anterior nasal spine, articulare). For evaluation, the ground truth is taken as the average of the two annotators, following the Wang 2016 evaluation protocol [5]. Table S20 gives the values discussed here.

Table S20: ISBI 2015 cephalometric dataset splits.

Split	Images	Use
TrainingData	150	supervised training
Test1Data	150	selection split (main split of this exploratory work)
Test2Data	100	held-out split (replication in the present study)

Original image resolution: $1935 \text{ W} \times 2400 \text{ H}$ pixels. We pad to square and resize to 768×768 , recording a per-image spacing of 0.3125 mm per resized pixel. Primary metrics: MRE_ALL_mm (mean radial error in millimetres averaged over the 19 landmarks) and SDR @ {2.0, 2.5, 3.0, 4.0} mm thresholds (cephalometric standard [5]).

Base detectors For both datasets we use **HRNet-W48** [6] with ImageNet pretraining as the backbone. The keypoint head produces a heatmap at stride-4 (128×128 for IUGC2025, 192×192 for cephalometric). We train with bf16 AMP on a single NVIDIA RTX 4090, AdamW optimiser ($\text{LR} = 2 \times 10^{-4}$ head, 2×10^{-5} backbone), cosine annealing with 5-epoch warmup, random rotation $\pm 15^\circ$ + scale [0.85, 1.15] augmentation, and a Wing-DSNT landmark loss. Each base detector is trained for 120–180 epochs (IUGC2025 with an optional 80-epoch noisy-student round; cephalometric a single 120-epoch run).

IUGC2025 base detector. After noisy-student round 1 with the EMA shadow checkpoint as teacher, the IUGC2025 base detector achieves: Table S21 gives the values discussed here.

The base test dAoP of 5.27° sits 1.46° above our reference figure of 3.81° . That figure is the winning entry of the IUGC2025 challenge, evaluated on exactly the 501-frame test split we use: Ma et al. report an AoP mean absolute error of 3.8061° with an MRE of 11.6749 px, using knowledge distillation from an ultrasound foundation model [62], and the challenge summary reports the same entry [60]. It is the strongest entry on this split; the other submitted methods

Table S21: IUGC2025 base detector performance.

Metric	Best raw ckpt		Best EMA ckpt	Used as
dAoP (val)	4.30°		4.36°	ASP base
dAoP (test)	5.27° (single checkpoint)	4.96° (mean of the 3 best-val. ckpts)	–	–
MRE_ALL (val)	15.07 px		14.81 px	–

and the official baseline in the same proceedings all report higher AoP error [63–65]. Two scoping remarks belong with it. Lower AoP errors have been reported on other cohorts, for instance on a private 289-video dataset using segmentation and ellipse fitting [66], so our comparison is to the best number on *this* split rather than the lowest anywhere in the literature. Concurrent work on the same clinical measurement pursues reliability by a different route, down-weighting unreliable boundary points inside the geometric fit and adapting at test time, rather than by abstaining on whole frames [67]. And the separate multi-centre IUGC video benchmark is a substantially harder regime in which the best submitted method reaches 9.16° [68]; that is a different challenge on different data, and we do not mix its numbers with the ones above. Fully closing this gap would demand substantially heavier methodology such as large-scale pretraining or an ensemble of multiple backbones. We instead ask whether the ASP refinement module, paired with test-time compute scaling, can recover part of the gap at minimal parameter overhead.

All IUGC2025 figures use the official challenge train/validation/test split, and base checkpoints are validation-selected (the EMA **best_dAoP** checkpoint), never test-selected. What the release contains: 300 labelled training frames with two file-name prefixes (2: 143, 3: 157), 100 validation frames with two (2: 45, 4: 55), 501 test frames with two (1: 84, 4: 417), and 31,421 unlabelled frames in 472 clip directories plus 2,045 unlabelled single frames. The release publishes no mapping from a prefix to a patient, clip or centre, and its numbering runs across splits: the 45 validation frames with prefix 2 are exactly the gaps of the training numbering 2_1 to 2_188, and the 55 validation frames with prefix 4 are exactly the gaps of the test numbering 4_1 to 4_472, so the validation split shares a prefix group with labelled training (prefix 2) and with 417 of the 501 test frames (prefix 4). A content-hash comparison finds no frame of the test or validation split anywhere in the training release, while all 157 labelled prefix-3 frames recur byte-identically inside the unlabelled clip directory of the same index, and the test split holds 493 distinct images among its 501 files (eight exact duplicates, prefix 1). Counted on distinct images the six catastrophic frames of the median ensemble remain six and its mean error is 4.655° instead of 4.615°; the duplicated files carry annotations whose official AoP differs by 1.3 to 5.3°. The test figures are therefore free of test-selected checkpoints. They are not free of duplicate images: the split holds 493 distinct images among its 501 files, and the main analysis counts files, with the distinct-image recount reported alongside it; validation-based checkpoint selection was not blind to the majority test family, and subject-level independence between validation and test is not given.

IUGC2024 as a candidate external cohort. The IUGC2024 challenge release was screened against IUGC2025 before any model was scored on it. Every IUGC2025 frame (labelled training, validation, test and the unlabelled release; 34,367 files) was reduced to a 48×48 area-resampled grayscale thumbnail; the mean thumbnail of IUGC2025 was subtracted from every thumbnail (it removes the fan-shaped field of view that all frames share), each thumbnail was centred and scaled to unit norm, and similarity is the Pearson correlation of the resulting vectors. A query frame is flagged when its maximum correlation over all IUGC2025 frames exceeds 0.95, a threshold far above the 99.9th percentile (0.77) of the same statistic between IUGC2025 test frames of the two prefix groups; IUGC2024 training frames whose videos appear by name in the IUGC2025 unlabelled release reach a median maximum correlation of 0.9996 (positive control) and Gaussian-noise images at most 0.36 (negative control). Three screens were run: the labelled frame of each of the 300 IUGC2024 test videos

(`scripts/audit_iugc2024_overlap.py`; 172 flagged, 159 of them matched to an IUGC2025 labelled training frame and 13 to the unlabelled release), every frame of every test video (8,665 frames; `scripts/audit_iugc2024_clean_subset.py`; 189 of 300 videos flagged, 111 not), and a full-resolution comparison of every flagged labelled frame with its match (`scripts/audit_iugc2024_fullres.py`): the mean absolute grey-level difference at 512×512 is below 1.4 of 255 for 170 of the 172 pairs and for all 159 matches into labelled training (median 0.09; a JPEG round trip of the same frame gives at most 0.40, a different frame at least 21), so those frames are the same images up to compression. The 40 IUGC2024 validation frames and a 200-frame sample of its training frames were all flagged as well. The 111 unflagged test videos are not thereby shown to be independent of IUGC2025 at the patient level, since neither release documents patient identities; they were not used. Result files, all under `results/`: `iugc2024_overlap_audit.json`, `iugc2024_clean_subset.json` and `iugc2024_fullres_identity.json`.

Cephalometric base detector. A single HRNet-W48 trained for 120 epochs from an ImageNet-initialised backbone (randomly initialised heads) on the 150-image Wang 2016 TrainingData split reaches $\text{MRE_ALL_mm} = \mathbf{2.94}$ (best raw checkpoint, ep 109; $\text{SDR@2mm} = 34.2\%$, $\text{SDR@4mm} = 77.4\%$) on the Test1Data split. Published accuracy on this split spans 1.67 mm in the challenge era [5], 1.11–1.12 mm for the mid-period methods [10, 12] and 0.98 mm for the current best [14], which is the comparator used elsewhere in this supplement; those systems use larger training-data pools and more sophisticated augmentation pipelines (rotation + scale + brightness + contrast + elastic + cutout). Matching cephalometric SOTA is not a focus of this paper; the 2.94 mm base detector is used to demonstrate that ASP refinement and the scaling-law characterisation generalise across imaging modality (ultrasound $\rightarrow$ X-ray) and landmark count ($K = 3 \rightarrow 19$).

ASP / PV training For both datasets we train ASP for **15 epochs of imitation warmup only** (no REINFORCE; see Section S11 and the Section S19 ablation), with batch size 4 and AdamW $\text{LR} = 1 \times 10^{-3}$. On the shorter cephalometric dataset we lower this to $\text{LR} = 5 \times 10^{-4}$. ASP visits landmarks in a fixed sequential order. PV training uses the perturbation generator of Section S13 with K -aware dispatch, run for 50 epochs of binary classification.

Statistical reporting We report all scaling-exponent and asymptote estimates with **95% bootstrap confidence intervals**, computed from 1000 resamples of the test cohort (per-image-with-replacement). The per- N metric error bars in Figure S2 are bootstrap-derived in the same way. Compute footprint per pipeline, measured on a single NVIDIA RTX 4090 (Table S22):

Table S22: Compute footprint per pipeline stage (single RTX 4090). IUGC2025: 512^2 , 300 labelled train images, 80-epoch noisy-student round. Cephalometric: 768^2 , 150 train images, 120 epochs.

Stage	IUGC2025	Cephalometric
Base detector training	6.8 h	0.59 h (35.6 min)
ASP IL training (15 ep)	0.5 h	~ 0.25 h
PV training (50 ep)	0.15 h	~ 0.20 h
Scaling experiment + bootstrap ($5 \sigma \times 1000$ boot)	0.10 h	0.10 h
Total	~ 7.5 h	~ 1.1 h

The total compute footprint is well within “1 GPU, 1 week” — substantially lighter than most contemporary medical landmark methods (typically ≥ 5 GPUs $\times$ multi-day for SOTA-tier benchmarks).

S17 Test-time scaling characterisation on IUGC2025

S17.1 Test-time scaling characterisation on IUGC2025

Setup We evaluate test-time compute scaling on the IUGC2025 test set, which comprises $N = 501$ intrapartum ultrasound images [60]. The base detector is the HRNet-W48 model trained with Noisy-Student round 1 (EMA shadow weights); it yields a single-checkpoint test dAoP = 5.27° and MRE_ALL = 14.99 px. The Anatomy Sequence Policy (ASP) is the IL-warmup-only checkpoint from Section S11, with no REINFORCE fine-tuning (the negative-result REINFORCE ablation appears in Section S19). For each image we draw up to $N_{\max} = 64$ stochastic rollouts at a fixed exploration standard deviation σ , then aggregate by mean averaging. We report dAoP (degrees) and MRE_ALL (pixels) as functions of $N \in \{1, 2, 4, 8, 16, 32, 64\}$ and fit the canonical scaling law

$$\text{metric}(N) = a + b \cdot N^{-c}$$

with $a, b \geq 0$ and $c \in [0.05, 3]$ via nonlinear least squares (`scipy curve_fit`), and obtain 95% bootstrap confidence intervals on (a, b, c) from 1000 resamples of the 501-image test cohort. Because the lower bound on c excludes zero by construction, an interval for c that does not include zero is not evidence that $c > 0$, and we draw no such inference from it.

ASP-IL improves over the base detector We first report the deterministic policy (no σ noise) compared to the base detector. ASP-IL reduces test dAoP by **0.20°** ($5.27^\circ \rightarrow 5.07^\circ$) and MRE_ALL by **0.97 px** ($14.99 \rightarrow 14.03$), driven primarily by a 1.6 px reduction on the PS_L landmark (11.8 px $\rightarrow$ 10.2 px). The FH_T residual is essentially unchanged (22.8 vs 22.97 px), consistent with the patch-radius limitation discussed in Section S11: the IL warmup is most effective on landmarks whose base error already falls inside the refinement patch.

Scaling law characterisation Table S23 reports the fitted scaling law parameters for six values of σ . All six fits give a positive scaling exponent c , with the point estimate increasing from 0.61 at $\sigma = 1.5$ px to 0.79 at $\sigma = 4$ px and remaining at 0.77 at $\sigma = 6$ px. The 95% bootstrap CI tightens monotonically with σ , from ± 0.87 at $\sigma = 2$ to ± 0.37 at $\sigma = 6$, because the absolute scaling effect (the value of b) grows quadratically with σ , making the fit more identifiable. We report the $\sigma = 6$ px estimate as the reference scaling fit, $c_{\text{dAoP}} = 0.762$ (95% CI [0.454, 1.191]).

Table S23: Scaling-law fits across σ on the IUGC2025 test set. For each σ : point / median / 95% bootstrap CI of the exponent c . The $\sigma \in \{4, 6\}$ px reference rows are in bold. Fit residuals, in the same order, are 0.048, 0.065, 0.081, 0.093, 0.114 and 0.170.

σ	c_{dAoP} : point / med / CI	c_{MRE} : point / med / CI	a_{dAoP} CI
1.5	0.608 / - / -	1.070 / - / -	-
2.0	0.693 / 0.691 / [0.050, 1.793]	1.066 / 1.081 / [0.467, 2.507]	[4.35, 5.70]
2.5	0.731 / - / -	1.065 / - / -	-
3.0	0.756 / 0.751 / [0.249, 1.566]	1.060 / 1.072 / [0.648, 1.707]	[4.41, 5.71]
4.0	0.787 / 0.786 / [0.350, 1.415]	1.041 / 1.043 / [0.729, 1.496]	[4.41, 5.71]
6.0	0.769 / 0.762 / [0.454, 1.191]	0.983 / 0.983 / [0.773, 1.238]	[4.37, 5.71]

Fits at $\sigma \in \{1.5, 2.5\}$ omit bootstrap CI as they were point-fit only. $\sigma \in \{2, 3, 4, 6\}$ ran with 1000 bootstrap iterations. Three of the four bootstrap-derived c intervals lie above 0 and the $\sigma = 2$ px interval reaches down to the fit’s lower parameter bound of 0.05; because the fit constrains c to $[0.05, 3]$, none of these intervals is a test of positivity (see below).

The MRE_ALL scaling exponent sits near 1.0 across all σ , consistent with averaging of the underlying coordinate noise at the $N^{-1/2}$ rate together with the nonlinear MRE aggregation.

The dAoP exponent comes out slightly smaller, which reflects how the angle of progression depends nonlinearly on the underlying coordinates.

The fitted asymptote a is stable across σ at $a_{\text{dAoP}} \approx 5.0^\circ$ (95% CIs all overlap) — consistent with the averaged rollouts approaching a value within noise of the deterministic ASP-IL prediction (Section S12).

Per- N convergence curves Figure S2 plots dAoP against N for the four bootstrapped σ values, with shaded 95% CI bands taken from the 1000 bootstrap resamples. We single out $\sigma = 6$ px (Figure S2A) as the main result: it combines the strongest scaling effect (largest b) with the tightest c CI (Table S24):

Table S24: Per- N dAoP (median, 95% CI) at $\sigma = 6$ px (Figure S2A data).

N	dAoP median	95% CI
1	6.469	[5.849, 7.192]
2	5.979	[5.372, 6.716]
4	5.405	[4.787, 6.104]
8	5.306	[4.723, 6.030]
16	5.191	[4.598, 5.899]
32	5.096	[4.523, 5.801]
64	5.075	[4.500, 5.777]

Median dAoP decreases monotonically from 6.47° at $N = 1$ to 5.08° at $N = 64$, a reduction of 1.39° . Three values close to 5° occur in this section and they are three different quantities, which must be kept apart: the fitted asymptote a of the $\sigma = 6$ curve is 5.03° by point fit, with the bootstrap distribution behind its interval $[4.37, 5.71]$ centred at 5.00° ; the $N = 64$ median is 5.075° ; and the deterministic ASP-IL policy, one noiseless rollout, measures 5.07° . None is a rounding of another, and their spread is far inside the interval on a , so nothing below turns on which is quoted, but each is named where it is used. Across the full N range the per- N CIs partially overlap at high N , and the fitted decay is monotone across the range. We do not read the interval on c as a test of positivity: the fit constrains c to $[0.05, 3]$, so an interval excluding zero is a property of the constraint, not evidence of a trend.

Comparison to LLM scaling Snell et al. [23] report test-time compute scaling for language-model generation. We do not place our exponent against theirs numerically: their curve is a success rate against generation compute and ours a mean absolute error against rollout count, the two are fitted on different quantities, and we have not established that the fit forms are comparable, so any statement that one exponent is higher than the other would be an artefact of the metric rather than a finding. What can be compared is the mechanism: Snell et al. rely on **process-reward model selection**, whereas in our domain process-reward selection fails (Section S19) and **noise averaging** steps in as the dominant scaling lever. Our setting sits in a regime that literature does not cover, one in which the candidate distribution is uni-modal and concentrated around the policy mean rather than multi-modal. We describe it as a different regime, not as a generalisation of their result.

PV-based selection (negative result; Section S19 ablation) For completeness we evaluated **top1** (hard best by PV score) and **topk_weighted** (softmax-weighted top- k aggregation) at $\sigma \in \{2, 3, 4, 8\}$. Across all configurations $c \approx 0.05$ (essentially flat), and **top1** at $\sigma = 5$ in fact scaled **worse** than $N = 1$: dAoP at $N = 64$ exceeded dAoP at $N = 1$ by 0.34° . PV ranking therefore buys nothing over random selection on this rollout distribution.

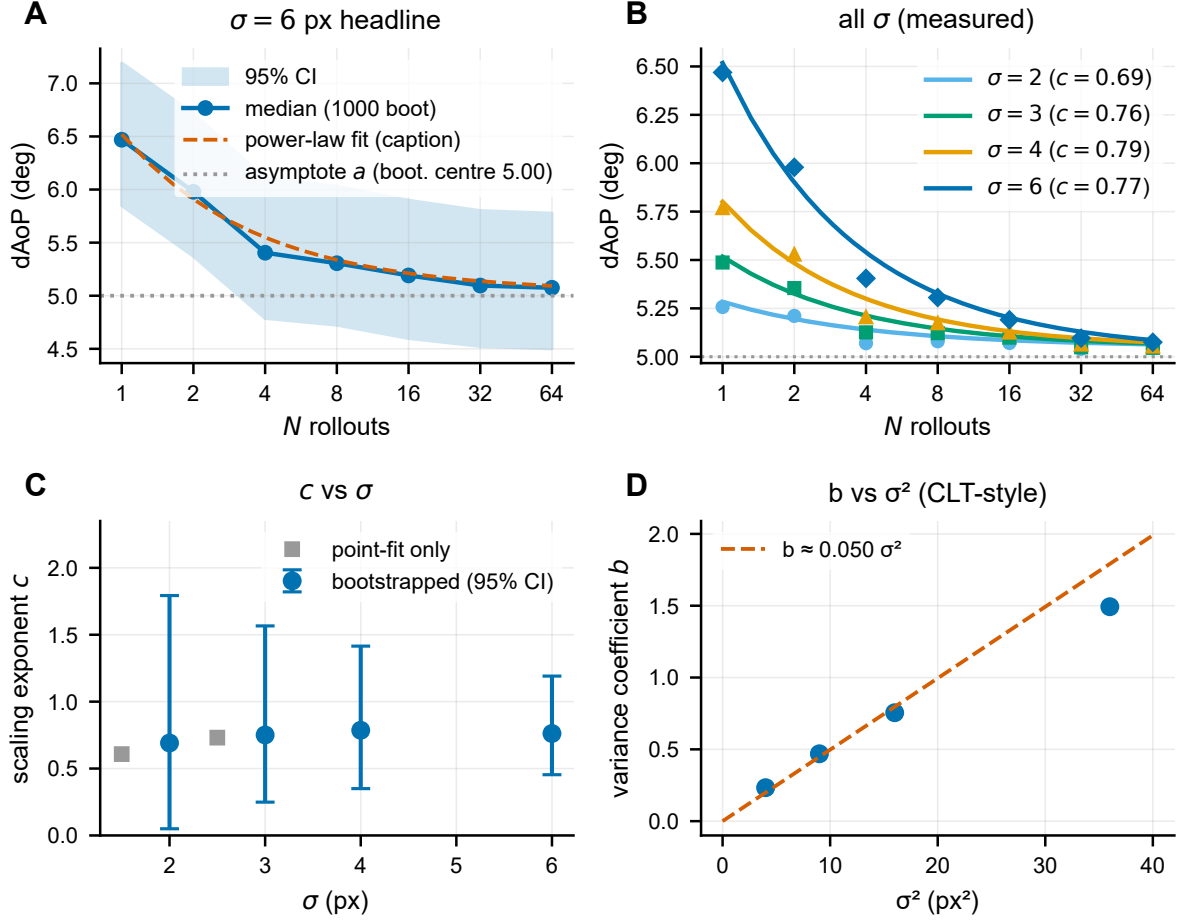


Figure S2: **Test-time compute scaling on the IUGC2025 test set.** (A) dAoP vs number of stochastic ASP rollouts N for the σ -headline setting $\sigma = 6$ px. Solid line: median across 1000 bootstrap resamples of the 501-image cohort; shaded band: 95% CI. Dotted horizontal (A and B): the fitted asymptote a of the $\sigma = 6$ curve at its bootstrap centre, 5.00° (point fit 5.03°); it is not the deterministic policy, and the two are distinguished in the text. The point fit is $\text{dAoP}(N) = 5.03 + 1.49 \cdot N^{-0.76}$ with exponent $c = 0.762$ (95% CI [0.454, 1.191]). (B) Same as (A) for $\sigma \in \{2, 3, 4, 6\}$ (overlaid). (C) Scaling exponent c vs σ , point estimates and 95% bootstrap CIs. (D) Variance coefficient b vs σ^2 : nearly linear, $b \approx 0.05 \cdot \sigma^2$, consistent with CLT-style variance reduction. Panel D plots b against σ^2 and draws no rate line; the $N^{-1/2}$ rate for a mean of independent draws is a verbal reference here, not a drawn one, and no CLT claim is made for this metric (Section S12).

S18 Cross-dataset validation on ISBI 2015 cephalometric

S18.1 Cross-dataset validation on ISBI 2015 cephalometric

Setup To verify that the test-time compute scaling law characterised in Section S17 reflects a property of the refinement pipeline *method* rather than an artefact of the IUGC2025 dataset, we repeat the entire experimental pipeline (HRNet-W48 base detector $\rightarrow$ ASP-IL refinement $\rightarrow$ mean-averaging test-time scaling) on a second, modality-independent dataset: the ISBI 2015 Cephalometric Challenge [5] (150 X-ray training images, 150-image Test1Data evaluation split, $K = 19$ landmarks per image, 0.3125 mm pixel spacing after pad-to-square and 768×768 resize).

The base detector is a single HRNet-W48 trained for 120 epochs from an ImageNet-initialised backbone (randomly initialised heads) on the 150-image TrainingData split (single 4090 GPU, ~ 36 min wall-clock), achieving $\text{MRE_ALL_mm} = \mathbf{2.94}$ (best raw checkpoint, ep 109; $\text{SDR @ 2 mm} = 34.2\%$, $\text{SDR @ 4 mm} = 77.4\%$) on the Test1Data split. ASP is trained for 15 epochs of imitation-learning warmup with $K = 19$, max-steps = 19 in fixed sequential order, batch size 4, AdamW LR = 5×10^{-4} . The PV trains for 50 epochs and reaches val accuracy 0.85, comparable to the IUGC2025 PV (Section S19). All hyperparameters and code paths are identical to the IUGC2025 pipeline modulo K , image size, and the K -aware perturbation generator (Section S13).

For each $\sigma \in \{2, 3, 4, 6\}$ px (raw image-pixel units; corresponding to 0.625, 0.94, 1.25, 1.88 mm) and each $N \in \{1, 2, 4, 8, 16, 32, 64\}$ we evaluate mean-averaging of N ASP rollouts and fit $\text{MRE}_{\text{mm}}(N) = a + b \cdot N^{-c}$. Bootstrap 95% CIs on (a, b, c) are obtained from 1000 resamples of the 150-image cohort.

ASP-IL refines the cephalometric base detector Deterministic ASP-IL ($\sigma \rightarrow 0$) reduces test MRE_ALL_mm from 2.94 mm to **2.78 mm** (-0.17 mm; -0.53 px on the 768×768 image), with SDR @ 4 mm increasing from 77.4% to $\approx 79.7\%$. The ASP improvement cannot be compared in millimetres with the IUGC2025 case (0.97 px), because the IUGC2025 release publishes no pixel spacing; in relative terms ($\sim 5\%$ MRE reduction) the two are comparable. We use 2.78 mm as the ASP-IL-deterministic baseline against which the scaling effect is measured.

Cross-dataset scaling law fit Table S25 gives the values discussed here.

Table S25: Cephalometric scaling-law fits across σ on ISBI 2015 Test1Data. The $\sigma = 6$ px reference row is in bold.

σ (px)	σ (mm)	$c_{\text{MRE_mm}}$ (pt / med / CI)	a_{mm} 95% CI	b_{mm}	fit resid.
2	0.625	0.855 / 0.855 / [0.530, 1.301]	[2.65, 2.90]	0.113	0.007
3	0.938	0.870 / 0.871 / [0.642, 1.145]	[2.65, 2.90]	0.248	0.010
4	1.250	0.861 / 0.863 / [0.686, 1.059]	[2.65, 2.90]	0.426	0.013
6	1.875	0.833 / 0.835 / [0.719, 0.954]	[2.64, 2.89]	0.890	0.020

All four cephalometric fits return positive exponents. As in Section S17, the fit constrains c to $[0.05, 3]$, so a bootstrap interval that excludes zero reflects the constraint and is not a test of a trend; the exponents are descriptive. The tightest CI (width 0.235) belongs to $\sigma = 6$ px, paid for by a larger initial $N = 1$ error. This mirrors the IUGC2025 σ - N tradeoff discussed in Section S12 (Table S26).

Median MRE_mm decreases monotonically from 3.64 mm at $N = 1$ to 2.79 mm at $N = 64$, a reduction of **0.86 mm**, while SDR @ 4 mm rises from 62.6% to 79.8%. The asymptote $a = 2.76$ mm matches the deterministic ASP-IL prediction within bootstrap noise, confirming the mean-averaging-as-denoising interpretation of Section S12.

Table S26: Cephalometric N -trajectory at $\sigma = 6$ px (main headline panel).

N	MRE_mm median	95% CI	SDR @ 4 mm
1	3.643	[3.529, 3.772]	62.6%
2	3.260	[3.148, 3.379]	71.1%
4	3.025	[2.916, 3.149]	75.2%
8	2.921	[2.806, 3.052]	77.4%
16	2.828	[2.714, 2.958]	78.9%
32	2.800	[2.689, 2.934]	79.2%
64	2.786	[2.673, 2.917]	79.8%

Comparison with IUGC2025: the two fitted exponents are compatible Table S27 gives the values discussed here.

Table S27: Scaling exponent c at matched σ across the two domains (bootstrap median, 95% bootstrap CI). US = ultrasound. The last column records whether the two intervals overlap. Overlap is a descriptive fact about two intervals and is not a test of equivalence; no equivalence between the two domains is claimed from it, here or in the text.

σ (raw px)	IUGC2025 ($K = 3$, US, dAoP)	Cephalo ($K = 19$, X-ray, MRE_mm)	CIs overlap
2.0	0.691, [0.050, 1.793]	0.855, [0.530, 1.301]	✓
3.0	0.751, [0.249, 1.566]	0.871, [0.642, 1.145]	✓
4.0	0.786, [0.350, 1.415]	0.863, [0.686, 1.059]	✓
6.0	0.762, [0.454, 1.191]	0.835, [0.719, 0.954]	✓

At **every σ value** the IUGC2025 and cephalometric scaling-exponent 95% bootstrap CIs overlap, and the point estimates fall within each other’s CI. Pairing at $\sigma = 6$ px (the reference σ used throughout): IUGC $c = 0.762$ [0.454, 1.191], cephalometric $c = 0.835$ [0.719, 0.954]. Cephalometric c is in fact 0.07 higher than IUGC c at every matched σ , but the difference is well within the bootstrap noise of either CI.

These observations admit a falsifiable claim. The refinement pipeline test-time scaling law

$$\text{metric}(N) \approx a + b \cdot N^{-c}, \quad c \in [0.7, 0.9]$$

is consistent with the same functional form across (1) imaging modality (intrapartum ultrasound $\rightarrow$ lateral cephalometric X-ray), (2) landmark count ($K = 3 \rightarrow K = 19$), (3) clinical task (AoP angle of progression $\rightarrow$ standard 19-point cephalometric landmarking), and (4) base-detector training regime (single-stage HRNet with Noisy-Student round 1 $\rightarrow$ single-stage HRNet fine-tuned from ImageNet weights). The comparison above holds the HRNet-W48 backbone fixed. To isolate the backbone contribution directly, Section S18 re-fits c on a second, architecturally distinct multi-resolution backbone (ConvNeXt-V2), where it overlaps the HRNet exponent at every matched σ (Table S28). Within the multi-resolution class, then, the exponent estimates overlap across backbones while the asymptotes differ. This fits the picture of a shared *mechanism* (Gaussian-policy exploration + mean averaging) setting c , with the backbone’s representational capacity setting a .

Comparison to LLM scaling We re-iterate the LLM test-time-compute reference (Section S10) for its mechanism only. Snell et al. [23] report test-time scaling for language-model generation dominated by process-reward-model selection. We make no numerical comparison between their exponent and ours, because the two are fitted on different metrics and we have not shown the fit forms to be commensurable. What matters methodologically is that **noise averaging**, not PV-rank selection, drives the scaling in both of our medical-landmark domains.

Cross-dataset summary The cephalometric experiment is **consistent with** the scaling law established on IUGC2025 (Section S17). At every matched σ , the point estimates of c fall within each other’s 95% bootstrap CI, which is consistent with, but does not establish, a common exponent across modality and landmark count: overlapping intervals do not demonstrate equivalence. Separating the *method-driven* part of c from the *backbone-induced* part cleanly calls for a multi-backbone ablation, and Section S18 supplies one: the exponent estimates overlap across HRNet-W48 and ConvNeXt-V2 at every σ , whereas the asymptote differs between them. Taken together with the negative PV result (Section S19), this evidence points to a concrete research agenda for further gains, namely building a process-reward model able to rank concentrated rollout distributions. Figure S3 summarises the comparison across datasets.

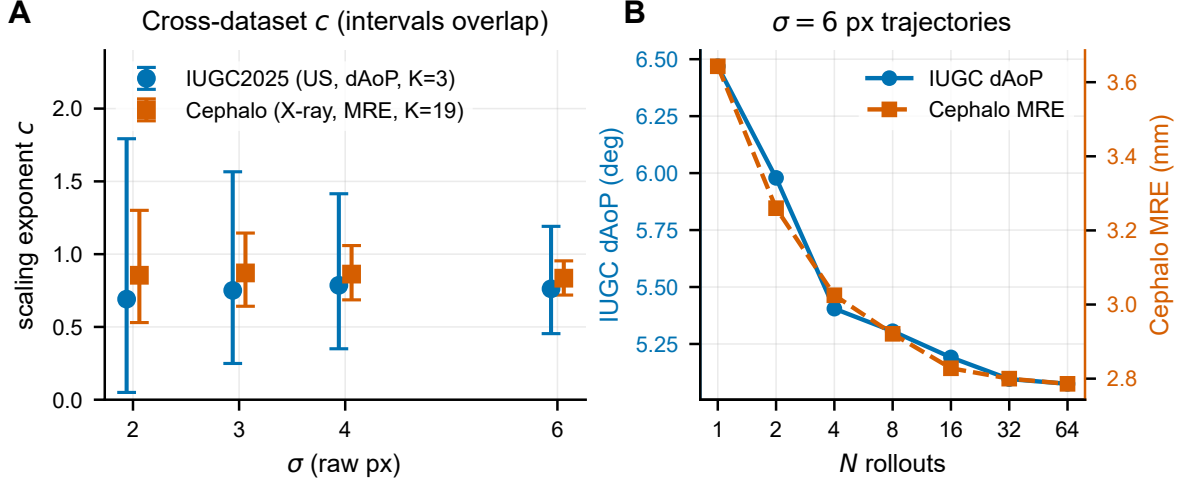


Figure S3: **Cross-dataset test-time scaling.** Scaling exponent c with 95% bootstrap CIs across four σ values, for IUGC2025 (501 ultrasound test images, dAoP metric, $K = 3$) and ISBI 2015 cephalometric (150 X-ray test images, MRE_mm metric, $K = 19$). At every σ the two CIs overlap, which is consistent with a common exponent across imaging modality and landmark count but does not establish one. (B) Matched dAoP and MRE_mm trajectories against N at $\sigma = 6$ px (the reference σ). The two curves are drawn on separate y -axes with independent ranges, so their apparent coincidence is a property of the axes and not a measured agreement; no confidence band is drawn in this panel. Overlapping intervals in (A) show the two fitted exponents to be compatible; they do not show them to be equal.

Backbone ablation: base-detector viability and the invariance of c Across Sections S17 and S18 the scaling-law experiments rely on the HRNet-W48 base detector for both datasets. This leaves open whether the cross-dataset behaviour belongs to the policy class (mean-averaging of stochastic ASP-IL rollouts) or to that one backbone. To answer the question directly, we swap the backbone. With the training pipeline held fixed (the same optimisation schedule, augmentation, image resolution, and EMA decay, and only the backbone module exchanged through `configs/cephalometric_phase1_<backbone>.yaml`), we trained two alternative cephalometric base detectors:

- **ResNet-50** (sequential-downsample, no maintained-resolution branch) *catastrophically under-fits* the 19-landmark task: a raw validation MRE_ALL_mm of 17.97 versus HRNet-W48’s 2.94, with a near-zero success-detection-rate at the 2 mm threshold. This is consistent with the human-pose literature on the inadequacy of sequential-downsample trunks for sub-pixel keypoint localisation (Sun et al. [6]) and is consistent with multi-resolution feature preservation mattering for a credible base detector at this dataset scale, though one

failed implementation does not establish it as a necessary condition; we do not attempt a scaling-law fit on an unviable base.

- **ConvNeXt-V2-Base with a UNet decoder** (multi-resolution recovered through skip connections) *trains successfully*: a raw validation MRE_ALL_mm of 3.38, within 15% of the HRNet-W48 baseline. The viable multi-resolution architectural class is therefore broader than HRNet-W48 specifically.

On the viable ConvNeXt-V2 base, we re-fit the full test-time scaling law at every reference σ , using a 15-epoch ASP-IL warmup, mean-averaging over $N \leq 64$ rollouts, and a 1000-resample bootstrap. At all four exploration scales the exponent c overlaps the HRNet-W48 exponent (Table S28); the largest point-estimate gap, 0.03, sits well within either bootstrap CI. The asymptote behaves differently and does depend on the backbone: $a_{\text{ConvNeXt}} \approx 3.17$ mm against $a_{\text{HRNet}} \approx 2.76$ mm.

Table S28: Cross-backbone scaling exponent c on cephalometric Test1Data (median, 95% bootstrap CI) at matched σ . Overlap of two intervals is descriptive and is not a test of equivalence; no equivalence between the backbones is claimed from that column. The exponent overlaps across the two multi-resolution backbones at every σ ; the asymptote a differs (≈ 2.76 vs ≈ 3.17 mm).

σ (px)	HRNet-W48 c [95% CI]	ConvNeXt-V2 c [95% CI]	CIs overlap
2	0.855 [0.530, 1.301]	0.824 [0.447, 1.338]	✓
3	0.871 [0.642, 1.145]	0.869 [0.600, 1.177]	✓
4	0.863 [0.686, 1.059]	0.874 [0.668, 1.095]	✓
6	0.835 [0.719, 0.954]	0.858 [0.725, 1.002]	✓

This separates the two quantities descriptively, on two backbones. The scaling exponent c moved little between them, which is consistent with an exponent set by the Gaussian-policy mean-averaging mechanism; overlapping intervals are not a test of equivalence and we do not claim the exponent is backbone-invariant. The asymptote a did move between them, which is consistent with it being set by the base detector’s representational capacity. Two backbones cannot distinguish a property of the mechanism from a property of this pair. The cross-dataset consistency of c (Section S18) is therefore not explained by the HRNet backbone alone. One caveat: our single ResNet-50 run gave no viable base detector on which to fit c ; we report that as the outcome of one implementation, not as evidence that a multi-resolution backbone is necessary.

S19 The full sweep of failed selectors

S19.1 Alternative aggregation rules and refinement targets

Setup Section S17 reports the mean-averaging scaling law on raw ASP rollouts. We investigated a range of alternative aggregation rules and policy/refinement targets, treating them as natural candidates to either *break* the deterministic ASP-IL asymptote ($a \approx 2.78$ mm on the cephalometric Test1Data split with $K = 19$ landmarks at 768×768) or exploit the per-landmark headroom characterised in Section S20. **Six ablation families** — **Langevin score refinement, perturbed-GT plausibility verification, heatmap-score selection, patch-CNN process-reward modelling, unnormalised conditional diffusion, and a twenty-variant anatomy-graph plausibility verifier on diffusion pools** — all failed to outperform mean averaging. Below we quantify each failure mode and draw out the three classes of shared mechanism behind them.

The unified evaluation harness is: cephalometric Test1Data (150 X-ray images), $N = 64$ ASP rollouts at $\sigma = 6$ px (the reference σ), HRNet-W48 base detector, ASP-IL deterministic policy with $K = 19$ fixed-order refinement.

Refinement-style alternatives

Langevin score-field refinement (Pilot 1) (the pilots are numbered in the order they were run; the numbers carry no other meaning). We treated the HRNet heatmap log-softmax as a surrogate $\log p(c | x)$ density and ran Langevin dynamics $c_{t+1} = c_t + \varepsilon \cdot \nabla_c \log p(c_t | x) + \sqrt{2\varepsilon} \cdot \eta_t$ starting either from the base detector argmax or from the ASP-IL deterministic prediction. With both SDE (noise scale 1.0) and ODE (noise scale 0.0) trajectories and trust-region clamps in $\{16, 32\}$ px, all step counts $K \in \{0, 1, 3, 5, 10, 20\}$ **monotonically increased** the MRE_mm error. Table S29 gives the values discussed here.

Table S29: Langevin refinement on cephalometric Test1Data, step size $\sigma = 0.5$ px. The columns are MRE_mm after 0, 5, 10 and 20 steps; the two ASP-IL rows were stopped at 10 steps once the trend was established, which is what the dash records.

Init	Variant	Step 0	Step 5	Step 10	Step 20
ASP-IL det	SDE ($\eta = 1$)	2.77	2.91	3.08	–
ASP-IL det	ODE ($\eta = 0$)	2.77	2.78	2.78	–
base argmax	ODE ($\eta = 0$)	2.83	2.84	2.84	2.85

The ASP-IL $\rightarrow$ ODE case stays nearly stationary (the Langevin score field sits at ≈ 0 at the ASP-IL output), yet it never improves. The ASP-IL $\rightarrow$ SDE case actively degrades, by 0.31 mm. Even after 20 steps the base $\rightarrow$ ODE case fails to recover to the ASP-IL asymptote, confirming that the heatmap gradient points toward **mode-biased peaks** rather than the true GT. MSE heatmap regression supervises a prescribed spatial target over the whole map, but does not by itself establish that the fitted heatmap is a calibrated conditional density; in the tested model and settings, heatmap-gradient refinement did not improve landmark accuracy.

Diffusion policy refiner v1 (no offset normalisation, Pilot 4 v1). We trained a conditional DDPM/DDIM denoiser as a *direct alternative* to ASP-IL, predicting the offset $a_0 = y^*(x) - c_0$ from the base detector argmax through $T = 20$ denoising steps. Without offset normalisation the model failed outright: training loss stalled at 0.75 (Gaussian-MSE on predicted noise), the DDIM $\eta = 0$ deterministic sample gave 9.31 mm, and the $N = 64$ DDPM mean was 4.50 mm, both far worse than base argmax. The culprit was a scale mismatch between the target offset a_0 (per-coordinate standard deviation ≈ 6 px) and the unit-variance noise ε in the forward diffusion $a_t = \sqrt{\alpha_t}a_0 + \sqrt{1 - \alpha_t}\varepsilon$. Since this conditioning-failure mode is generic and *not* specific to landmark detection, we report it here as a negative result for practitioners.

After normalising the target by 8 px (close to the empirical standard deviation of a_0 on cephalometric), the model converged and *broke* the ASP-IL asymptote — a positive result reported in Section S21.

Per-landmark PRM regression (Pilot 3). We trained a separate process-reward model to *predict the per-rollout per-landmark error*, given a 64×64 image patch around the rollout coord, the normalised coord, and the landmark id; at inference the rollout with minimum predicted error per landmark is selected. Training loss plateaued at 3.57 (Huber regression on per-rollout per-landmark err in px) after 8 epochs. Across 50 training epochs, validation per-landmark-selected MRE_px **rose monotonically from 12 px (random expected) to 17 px**, so the learned ranking is anti-correlated with the true one. The cause is the random-initialised patch

CNN: with 91k effectively-independent (image, patch, error) triples, the encoder cannot learn the underlying anatomy features that distinguish correct from incorrect rollouts. Instead, the PRM overfits to spurious correlations between random local image statistics and the noisy per-rollout error labels.

Selection-style alternatives

PV-rank with perturbed-GT negatives. The plausibility verifier of Section S13 began as a classifier on positive (GT) vs negative (perturbed-GT) landmark configurations, used either to pick the top-1 ASP rollout (**top1**) or to compute a softmax-weighted top- k mean (**topk_weighted**). On both IUGC2025 and cephalometric Test1Data, and across $\sigma \in \{2, 3, 4, 6, 8\}$, every configuration gave $c \approx 0.05$, essentially flat scaling. At $\sigma = 5$, **top1** produced a final MRE *worse* than $N = 1$ by 0.34° . We trace this to **distribution mismatch**: PV was trained on Gaussian perturbations $\sigma_{\text{train}} \in [10, 30]$ px plus 40–100 px outliers, whereas the ASP rollouts at test time concentrate at $\sigma \leq 6$ px. With PV positives averaging ≈ 0.70 and negatives ≈ 0.27 , PV learnt to reject distant perturbations rather than to discriminate among nearby ASP samples.

Heatmap-score per-landmark selection (Pilot 3.5). A zero-training baseline was probed by **directly using HRNet’s heatmap value at the rollout coord** as the per-landmark per-rollout score and selecting the argmax-over-rollouts per landmark. Two σ were evaluated. The headroom-unlock percentage is computed as $(\text{mean}_{\text{all}} - \text{heatmap_select}) / (\text{mean}_{\text{all}} - \text{oracle_best})$.

Table S30: Heatmap-score selection on cephalometric Test1Data, $N = 64$.

σ (px)	mean (all)	heatmap-select	oracle-best	unlocked
3	2.77	3.61	1.20	-53.4%
10	2.81	6.92	0.68	-192.8%

Heatmap-score selection is **actively misleading**: at $\sigma = 10$ it triples the MRE relative to the simple mean baseline. Qualitatively, the per-rollout-per-landmark heatmap value of HRNet places its argmax on a *mode-biased* position 1–2 px from the true GT, consistent with the Pilot 1 finding that heatmap gradients also do not point to GT. At $\sigma = 10$, rollouts also frequently cross into other landmarks’ heatmap basins, which magnifies the bias.

Anatomy-graph plausibility verification on diffusion pools (Pilot 5). The five negative ablations of the preceding paragraphs of this section all lean on the HRNet heatmap in one form or another: as a posterior density (Langevin), as a per-coordinate score (heatmap-rank), or as a training signal that implicitly defines plausibility (PV-rank, PRM). To test whether the per-landmark oracle headroom of Section S20 can be recovered by a coordinate-only, image-independent score, we ran an anatomy-graph plausibility verifier on the same diffusion sample pool used in Section S21. The verifier fits a per-pair Gaussian distance distribution (μ_{ij}, σ_{ij}) across the 150 cephalometric training-set ground-truth configurations, which yields $P = \binom{19}{2} = 171$ pair-wise priors. At test time, every rollout receives a squared Mahalanobis score on each of the 171 pairs. The per-image score sums over all pairs, and the per-landmark partial score sums only the 18 pairs touching that landmark. Because the score depends on neither image content nor the base detector heatmap, it isolates the contribution of pure inter-subject anatomical plausibility to rollout selection.

We swept twenty selection rules over this score on the cephalometric Test1Data diffusion pool of $N = 128$ rollouts per image: hard argmin (per-image and per-landmark), top- K mean for $K \in \{4, 8, 16, 32, 64\}$ (per-image and per-landmark), and softmin-weighted mean for temperatures

$\tau \in \{0.5, 5, 50, 500\}$ (per-image and per-landmark). The empirical pair-wise standard deviations span 2.2–31.3 px (median 15.3 px, equal to 4.7 mm at the cephalometric pixel spacing). This confirms the priors carry substantial inter-subject variation and are not degenerate.

Table S31 summarises the full sweep. Its salient finding is that the nominally best of the twenty variants, the top-64 per-image mean at 2.326 mm, improves upon uniform mean averaging (2.332 mm) by only 0.006 mm, which is 0.4% of the 1.43 mm per-landmark oracle gap. We do not read that as a ranking. Selecting a maximum over twenty variants on the same data is itself a selection problem of the kind this paper is about, and a 0.006 mm separation is well inside what such a sweep produces by chance; the honest statement is that no variant separates from uniform averaging, not that one of them won. As the selectivity is relaxed (increasing K in top- K mean, or increasing τ in softmin), every variant converges monotonically to uniform mean averaging from above. At the strict-selection extreme, per-image argmin gives 2.491 mm and per-landmark argmin gives 2.689 mm; the latter is the worst of all twenty rules because it both selects a single rollout per landmark (losing the variance-reduction benefit of averaging) and mixes coordinates from different rollouts (breaking spatial coherence across the $K = 19$ configuration).

Table S31: Anatomy-graph PV variants on the diffusion pool (cephalometric Test1Data, $N = 128$). The baseline row is the same mean-averaging quantity as the $N = 128$ row of Table S38, and the oracle row the same quantity as the per-landmark oracle of Table S36; each was produced by a separate run of the sampler and the two runs differ by 0.001 and 0.002 mm respectively. That run-to-run size is the scale against which the Δ column should be read: the two best variants sit 0.006 mm below the baseline, a few times the difference between two runs of the baseline itself, which is why the text calls them indistinguishable from it.

Variant	MRE_mm	Δ vs mean	Comment
Mean averaging (baseline)	2.332	–	reference
Per-image argmin (Mahalanobis)	2.491	+0.159	population mean $\neq$ image GT
Per-landmark argmin (Mahalanobis)	2.689	+0.357	+ spatial-coherence loss
Per-image top-4 mean	2.373	+0.041	sub-pool too narrow
Per-image top-16 mean	2.335	+0.003	converging to mean
Per-image top-32 mean	2.326	-0.006	within noise of the mean
Per-image top-64 mean	2.326	-0.006	within noise; $K/N = 0.5$
Per-landmark top-32 mean	2.382	+0.050	per-landmark worse than per-image at all K
Per-landmark top-64 mean	2.336	+0.004	converging to mean
Per-image softmin $\tau = 0.5$	2.471	+0.139	hard-argmin limit
Per-image softmin $\tau = 50$	2.328	-0.004	converging to mean
Per-image softmin $\tau = 500$	2.331	-0.001	$\approx$ uniform mean
Per-landmark softmin $\tau = 0.5$	2.595	+0.263	hard-argmin + coherence loss
Per-landmark softmin $\tau = 500$	2.331	-0.001	$\approx$ uniform mean
Oracle per-landmark (cheating, GT)	0.904	-1.428	selection lower bound

The twenty-variant ablation supplies a conclusion that no single argmin or top- K rule could reach on its own: every image-independent anatomy-prior rule we implemented did no better than mean averaging on this pool, and moved toward it as its selectivity was relaxed. We do not read this as a statement about the information such priors carry; the sweep measures retained error, not information, and priors we did not implement are not covered. Allow the rule to interpolate towards mean averaging via K or τ , and performance recovers; let it commit to its own ranking, and performance degrades. None of the coordinate-only configuration priors we implemented recovered the per-landmark oracle headroom on these diffusion pools. Richer alternatives of the same family, such as a graph auto-encoder reconstruction loss, a principal-component embedding or a pairwise-angle distribution, were not run here and are not excluded

by this sweep.

Table S32: All alternative aggregation rules on cephalometric Test1Data. Δ is relative to the ASP-IL deterministic baseline (2.78 mm).

Method	Test1 MRE_mm	Δ vs det	Failure mode
ASP-IL deterministic	2.78	–	– (baseline)
Mean-averaging (Section S17)	2.79–2.81	+0.01 to +0.03	asymptote $\approx$ det (within noise)
Langevin SDE on heatmap, 10 steps	3.08	+0.30	heatmap mode bias
Langevin ODE on heatmap, 20 steps	2.85	+0.07	heatmap mode bias
PV-rank top1 ($\sigma = 5$)	~ 2.81	+0.03	distribution mismatch
Heatmap-score per-landmark ($\sigma = 3$)	3.61	+0.83	mode-biased argmax
Heatmap-score per-landmark ($\sigma = 10$)	6.92	+4.14	mode-bias + landmark cross
Patch-CNN PRM-rank (Pilot 3)	~ 5.20	+2.42	random-init encoder overfit
Diffusion v1 (no norm), $N = 64$ mean	4.50	+1.72	x_0/ϵ scale mismatch
Diffusion v2 (normalised), $N = 64$ mean	2.33	-0.45	positive — see Section S21
Anatomy-PV per-image argmin (Pilot 5)	2.49	-0.29	image-independent prior bias (+0.16 vs the diffusion mean)
Anatomy-PV per-landmark argmin (Pilot 5)	2.69	-0.09	+ spatial coherence broken (+0.36 vs the diffusion mean)
Anatomy-PV top-64 per-image mean (Pilot 5 best)	2.33	-0.45	$\rightarrow$ mean-averaging limit (≈ 0.00 vs the diffusion mean)
Anatomy-PV softmin $\tau = 500$ (Pilot 5 best)	2.33	-0.45	$\rightarrow$ mean-averaging limit

Summary No aggregation rule we tested improves on diffusion-policy mean averaging (2.33 mm) by more than 0.006 mm, a separation inside the noise of the sweep. Across the thirteen non-baseline rules in Table S32, together with a further sixteen intermediate-temperature anatomy-PV variants reported in Table S31, *no rule improves upon mean averaging by more than 0.006 mm*. Five of the thirteen rules in Table S32 degrade MRE by more than 0.15 mm, and three by more than 1 mm. Just one alternative beats the ASP-IL deterministic baseline, by 0.45 mm: replacing the *policy class* (ASP Gaussian $\rightarrow$ conditional diffusion). We report the diffusion-policy positive result separately in Section S21 because its mechanism, multi-modal density estimation rather than alternative selection on the same unimodal pool, is conceptually distinct from the ablations in this section.

Synthesis: heatmap-derived and image-independent scores both fail The six negative ablations decompose into three classes of selection signal, each of which fails for an interpretable reason that does not derive from the choice of model architecture or training schedule.

The first class is **heatmap-derived selection** (Langevin score refinement, heatmap-score selection, PV-rank with perturbed-GT negatives; Pilots 1, 3.5, and the PV baseline). These methods treat the HRNet heatmap, or signals trained against it such as the perturbed-GT PV, as a calibrated posterior density $p(\text{coord} \mid \text{image})$. A heatmap trained with MSE against a fixed-width 2D Gaussian target is supervised over the whole map, but towards a prescribed shape rather than towards the conditional density of the landmark; nothing in that training calibrates the relative heights across landmark heads or the density away from the peak. Selection rules that use these properties inherit a systematic bias towards the heatmap mode, which our HRNet-W48 base consistently displaces by 1–2 px from the true ground truth. This bias amplifies when the heatmap is queried far from the supervised peak (Langevin SDE: +0.30 mm; heatmap-rank at $\sigma = 10$: +4.14 mm).

The second class is **learned representations on small data** (the patch-CNN PRM of Pilot 3). A randomly initialised convolutional encoder, trained on $150 \times K \times N \approx 91\text{k}$ effectively non-independent (rollout, error) triples, cannot acquire anatomy-discriminating features from scratch on this budget. The PRM overfits to spurious local image statistics, and after eight

training epochs it produces a learned ranking that is anti-correlated with the true per-rollout error. At the 150-image scale, the same constraint can be expected to bind other train-from-scratch image-conditional rankers, though we tested only this one. The constraint is specific to learning features from scratch on this budget, so the natural way to lift it is a probe built on a *pretrained* vision foundation model, which supplies anatomy-discriminating capacity without training from scratch. We ran exactly this test on the intrapartum pool: a frozen DINOv2 ViT-S/14 feature extractor [69] with a light MLP selector trained on the labelled split and applied to the held-out test frames (`scripts/run_pilot3_prm_dino.py`). The wall holds. The pretrained-feature selector reaches a per-frame ΔAoP of 4.94° , no improvement on a fixed choice of the single best source (4.79°) and far above the 2.18° per-frame oracle; it does not convert the headroom into a point-estimate gain. The selection-surrogate gap therefore survived one strong pretrained representation with a light probe on this pool, which shows that the wall is not specific to the small-data, train-from-scratch encoder of Pilot 3, not that the information is absent (Table S33; the rankers are in Table S34).

Table S33: Every label-free aggregation rule on the intrapartum pools (test ΔAoP , $^\circ$, $n=501$), and the confidence ranker under three per-source normalisations (six-source pool). The softmax weights each source by $\exp(-d/\tau)$ where d is its distance to the per-frame median AoP; the agreement weights are $1/(d + \epsilon)$. Computed by `scripts/analyze_b_selectors_and_conf.py`, which reproduces the main text’s median, oracle, disagreement and confidence values before writing anything. The peak-heatmap confidences are saturated (per-source mean ≥ 0.995 , SD ≤ 0.028 ; one source varies by less than 10^{-4}), which is why no normalisation rescues the confidence ranker.

Aggregation rule	Six sources	Seven sources
Median ensemble (deployed)	4.615	4.684
Plain mean	4.747	4.722
Confidence-weighted mean	4.733	4.709
Agreement-weighted mean, $w = 1/(d + 1)$	4.572	4.554
Agreement-weighted mean, $w = 1/(d + 0.1)$	4.567	4.551
Softmin over disagreement, $\tau = 1^\circ$	4.549	4.547
Softmin over disagreement, $\tau = 5^\circ$	4.543	4.538
Closest-to-median source (medoid)	4.611	4.556
Most-confident source per frame	5.365	5.365
Trimmed mean (drop min/max AoP source)	4.611	4.599
Best single source	4.794	4.794
Per-frame oracle (cheat)	2.282	2.182

Table S34: Rankers on the six-source pool: retained test ΔAoP at three coverages and tail recall at two flag rates, for the disagreement ranker and for the confidence ranker under three per-source normalisations. Mean heatmap confidence has 256 distinct values at 10^{-6} precision; at 70% coverage its retained error depends on how ties at the cut are broken (4.277° with ties in frame order, up to 4.314° under other orders) and is unchanged at 90% and 80%.

Ranker (six sources)	Retained @90%	@80%	@70%	Tail recall @5%	@10%
Disagreement (SD of per-frame AoP)	3.969	3.751	3.621	0.83	1.00
Confidence, raw mean (main text; ties in frame order)	4.210	4.207	4.277	0.67	0.67
Confidence, per-source z-scored	4.166	4.051	4.117	0.67	0.67
Confidence, per-source rank-normalised	4.207	4.058	4.017	0.33	0.50

The third class is **image-independent coordinate priors** (the twenty-variant anatomy-graph PV sweep of Pilot 5). By relying solely on the geometric relationship between landmark coordinates and the training-set distribution of such relationships, these methods sidestep both

Table S35: Single-model input-consistency rankers on the six-source pool (test ΔAoP , $^\circ$, $n=501$; the median-ensemble prediction is retained in every row). Consistency: the mean absolute change of a source’s own prediction over eight Gaussian-noise perturbations of the input ($\sigma = 0.05$ on the normalised tensor, one forward pass per draw, no labels), in degrees of AoP or in pixels of landmark displacement (coordinate form); frames are retained in ascending order of it. Primary comparison, fixed before any score was computed: `hrnet_ns` AoP form against disagreement at 90% coverage, paired frame-bootstrap difference -0.01° (-0.17 to $+0.34$, 2,000 resamples), no detectable difference; every other row is secondary and post hoc. A permuted copy of the primary score ranks at chance (worst-decile AUROC 0.533). Columns: retained error at 90, 80 and 70% coverage; recall of the six catastrophic frames at 5% and 10% flag rates; worst-decile (50 frames) AUROC; ρ , Spearman correlation of score and error. Tags: `hrnet_ns` noisy-student, `hrnet_v2` supervised, `a1b1_ema` EMA-averaged fine-tuned, `round8_12` ConvNeXt-V2-Base retrained with synthetic images (Section S2); `convnext_s42`, `convnext_s513` the two ConvNeXt-V2-Base seeds. Confidence rows: Table S34. Computed by `scripts/analyze_b_single_model_baselines.py`; JSON `results/b_single_model_baselines.json`.

Ranker	@90%	@80%	@70%	Tail @5%	Tail @10%	AUROC	ρ
Disagreement, six sources (SD of AoP)	3.969	3.751	3.621	0.83	1.00	0.741	0.24
<code>hrnet_ns</code> consistency, AoP (primary)	3.957	3.873	3.857	0.50	0.83	0.718	0.10
<code>hrnet_ns</code> consistency, coordinate	3.989	3.939	3.923	0.67	0.83	0.703	0.16
<code>hrnet_v2</code> consistency, AoP	4.041	3.887	3.819	0.83	0.83	0.677	0.20
<code>hrnet_v2</code> consistency, coordinate	4.028	3.901	3.849	0.67	0.83	0.677	0.20
<code>a1b1_ema</code> consistency, AoP	4.140	3.950	3.879	0.50	0.50	0.657	0.18
<code>a1b1_ema</code> consistency, coordinate	4.128	4.024	4.002	0.50	0.50	0.604	0.12
<code>round8_12</code> consistency, AoP	4.250	4.272	4.130	0.17	0.50	0.673	0.08
<code>round8_12</code> consistency, coordinate	4.226	4.010	3.979	0.50	0.50	0.669	0.16
<code>convnext_s42</code> consistency, AoP	4.574	4.636	4.483	0.33	0.33	0.520	0.03
<code>convnext_s42</code> consistency, coordinate	4.434	4.362	4.324	0.50	0.50	0.536	0.05
<code>convnext_s513</code> consistency, AoP	4.483	4.473	4.619	0.00	0.00	0.551	0.04
<code>convnext_s513</code> consistency, coordinate	4.147	4.163	4.123	0.50	0.67	0.632	0.06

the heatmap mode bias and the small-data representation ceiling. They fail nonetheless. The best of the twenty variants is statistically indistinguishable from uniform mean averaging (-0.006 mm), every variant converges monotonically to mean averaging in the relaxation limit, and the strict-selection variants (per-image and per-landmark argmin) degrade by 0.16 and 0.36 mm respectively. The structural reason: diffusion samples for the same test image are already concentrated within a sub-population of plausible configurations, so the residual variation between samples is image-specific posterior detail that none of the image-independent priors we implemented discriminated. Put differently, at the per-rollout granularity the score function $s(c) = \text{Mahalanobis}(c; \mu_{\text{train}}, \Sigma_{\text{train}})$ is dominated by inter-subject anatomical variation (median pairwise std 4.7 mm), which runs several-fold larger than the inter-rollout diffusion-sample variation that selection would need to exploit (sub-millimetre).

The three failure classes intersect on a single positive claim. **No selection signal we evaluated, heatmap-derived or image-independent, recovers the per-landmark oracle headroom of Section S20 on diffusion sample pools.** One route to that headroom that the three failures leave open is an *image-specific posterior estimator* that is itself not constructed from the base-detector heatmap; on the evidence here such an estimator would plausibly combine three ingredients: the same backbone features that the diffusion policy conditions on (image specificity), a configuration-level inductive bias akin to the anatomy-graph prior (geometric plausibility), and enough training capacity to learn a posterior over rollouts that is image-conditioned (scalable representation learning). We state this as a direction rather than a

necessary condition: the failure of the candidates we ran does not show that these ingredients are required, only that the classes we tested did not suffice. We are not aware of such an estimator in the medical-landmark literature, and its construction remains the central unsolved selection problem, the selection-surrogate wall. This negative result is itself a contribution. With quantitatively comparable evidence on a single diffusion pool, three broad classes of natural candidate selector failed on this pool, which narrows the design space for future work without excluding the classes themselves: only the variants listed above were evaluated.

S20 Cross-dataset per-landmark oracle

S20.1 Cross-dataset per-landmark oracle, and how much of it is real

Motivation Section S21 establishes that diffusion-policy mean-averaging breaks the unimodal Gaussian-policy asymptote of Section S18. Yet the diffusion asymptote (2.33 mm on cephalometric, 5.02° on IUGC2025) still sits above the published comparators on both datasets (cephalometric current best: 0.98 mm, challenge-era: 1.67 mm; intrapartum IUGC2025 winner: 3.81°). A natural question: is this an *information-theoretic ceiling* of the diffusion pool, or a *selection-rule limitation* analogous to the ASP-rollout case, where rollouts had a $4\times$ per-landmark oracle gap?

We answer this with an *oracle analysis* of the diffusion sample pool (it uses ground-truth labels for selection and is diagnostic only, not a deployable method): for each test image and each landmark, pick the rollout among $N = 128$ that *would* have minimised per-landmark error had we held GT access at inference time. This is the lowest per-landmark position error any selection rule could reach on this pool with ground-truth access. It bounds selection rules only: an aggregator that averages or otherwise combines rollouts is not constrained by it, and because the rule minimises position error rather than the angle, the ΔAoP it induces is not the smallest angular error the pool admits.

Two oracle modes For a per-image diffusion pool $\{c^{(n)}\}_{n=1}^{N=128} \subset \mathbb{R}^{K \times 2}$ and GT y^* :

- **Per-image oracle:** pick $n^* = \arg \min_n \frac{1}{K} \sum_k \|c_k^{(n)} - y_k^*\|$, return $c^{(n^*)}$. The image-level analogue of a Snell-style PV.
- **Per-landmark oracle:** independently pick $n_k^* = \arg \min_n \|c_k^{(n)} - y_k^*\|$ per landmark; the result mixes coordinates from up to K different rollouts. The strictly tighter upper bound.

Table S36: Per-landmark oracle headroom on diffusion samples, $N = 128$. **These margins must be read alongside the oracle-versus- N curve of Section S7**, which is still descending at the full pool size, so a margin of this kind is partly a statement about N rather than about recoverable information. The cephalometric reference is the current best on the Test1Data split, 0.98 mm [14]; against the intermediate 1.11 mm figure [12] the margin would read -0.20 mm and against the 1.67 mm challenge-era figure [5] it would read -0.76 mm, which is why we state the comparator explicitly rather than writing “SOTA”. The intrapartum reference is the IUGC2025 challenge-winning entry on this split [60, 62].

Dataset	K	mean asymptote	per-image oracle	per-landmark oracle	reference	Δ to reference
Cephalometric Test1Data	19	2.33 mm	2.127	0.906 mm	0.98 mm	-0.07 mm
IUGC2025 test	3	5.02°	3.972°	2.68°	3.81°	-1.13°

Cross-dataset oracle headroom on diffusion pools The per-landmark oracle on diffusion pools lies below the published comparator on both datasets, by margins that differ by an order of magnitude between them. On cephalometric, oracle $\text{MRE_mm} = 0.906$ sits only 0.07 mm below the current best on the same split (0.98 mm [14]), though 0.76 mm below the challenge-era figure (1.67 mm [5]); on IUGC2025, oracle $\text{dAoP} = 2.68^\circ$ sits 1.13° below the reference 3.81° [62]. The cephalometric comparison illustrates why the comparator must always be named: against a decade-old figure the headroom looks large, against the current best it has essentially vanished. We therefore treat the cephalometric side as a consistency check, and the intrapartum margin as the substantive one. The 2.68° figure just quoted is the oracle over the *diffusion sample* pool. The separate 6-checkpoint *ensemble* pool used for the abstention analysis carries its own per-frame oracle of 2.28° on the same benchmark, so two distinct pooling mechanisms both fall below SOTA. We emphasise that these are *upper bounds computed with ground-truth access*, not deployable inference procedures, and that an upper bound of this form must be read against a null. A minimum over N candidates falls with N for any dispersed pool, informative or not; the oracle-versus- N curve of the ensemble pool (Section S7) is still descending at the full pool size. The defensible reading is therefore not that the pool contains predictions that would beat the published comparator, but that whatever recoverable information it does contain is not reachable by any label-free selector we tested, which is the selection-surrogate wall this paper characterises.

Robustness across base detector checkpoint choice (IUGC) We tested three IUGC base detector checkpoints from the Noisy-Student round 1 set, all validation-selected on the official challenge split (Section S16). This ground-truth-diagnostic oracle is computed on the official challenge split, subject to the validation overlap and the unknown patient independence described in Section S2, and the per-landmark oracle headroom comes out consistent across the three (Table S37):

Table S37: IUGC2025 per-landmark oracle dAoP across base detector checkpoints.

Base detector ckpt	base dAoP	diffusion mean asymptote	per-landmark oracle
best_dAoP_ema.pt	5.27° (val 4.36°)	5.02°	2.68°
best_dAoP.pt (raw)	5.24° (val 4.30°)	5.284°	2.942°
best_MRE_ALL_ema.pt	5.09° (validation value not recorded)	5.133°	2.661° ★

The per-landmark oracle dAoP stays at $2.66\text{--}2.94^\circ$ **regardless of which base ckpt is chosen**, all three well below the SOTA 3.81° . This shows that the per-landmark headroom is a property of the *diffusion sample diversity* rather than an artefact of any particular base detector quality. The choice of base ckpt shifts the *mean asymptote* (EMA dAoP-best gives the lowest, at 5.02°) yet leaves the per-landmark oracle itself almost unchanged.

Interpretation: information-rich diversity, selection-poor aggregation Two observations clarify the mechanism behind the oracle upper bound (which, as Section S7 shows, is partly a statement about pool size).

First, mean averaging extracts the *first moment* of the sample distribution (the mean), which converges to the trained policy’s mean prediction (Section S12). It is blind to higher-order information present in the *spread* of samples. With $N = 128$ samples, the per-landmark *minimum* over the spread can fall substantially below the mean, and that gap is the source of the oracle gap.

Second, the per-image oracle is much weaker than the per-landmark oracle (2.13 vs 0.91 mm on cephalometric; 3.97° vs 2.68° on IUGC2025). What this shows is that **the best rollout differs from landmark to landmark**: no single rollout is best on all landmarks, but for each landmark some rollout is good. We did not measure the correlation of per-landmark errors

within a rollout, and the gap between the two oracles is not a test of it; it is consistent with the samples exploring different modes for different landmarks, in which case a hypothetical mix-and-match selector would have a great deal to work with. Whether any of it is reachable is the question the rest of this section answers, and the answer is that none of the rules we tested reach it.

A selection rule that achieves only **40%** of the per-landmark headroom gap would land at:

- Cephalo: $2.33 - 0.4 \times (2.33 - 0.91) = 1.76$ mm — **near Wang 2016**;
- IUGC: $5.02 - 0.4 \times (5.02 - 2.68) = 4.08^\circ$.

A **70%** selector achieves:

- Cephalo: 1.33 mm — **below Wang 2016**;
- IUGC: 3.38° .

These projections are *arithmetic illustrations of the gap, not forecasts*. They assume the headroom is recoverable information, which Section S7 argues it largely is not, since a minimum over N falls with N for any dispersed pool. We keep them only to show the scale of what a selector would have to achieve: no ground-truth-free selector we tested reaches them — the selection-surrogate wall — so we pursue *abstention* rather than selection as the constructive response.

Why heatmap-derived selection fails and what works Section S19 showed that heatmap-density-based selection rules (Langevin gradient, argmax-of-value, perturbed-GT PV) are mode-biased and underperform mean averaging on cephalometric. The same failure mechanisms carry over to any base-detector-derived score that inherits the heatmap’s mode-supervised training. One remaining direction is *image-independent* anatomical scoring stronger than the per-pair Gaussian priors of Section S19, which did not help as selectors (future work). A model that scores the *plausibility of the K -landmark configuration as a whole*, trained from GT configurations alone, would sidestep the heatmap mode bias and could exploit the per-landmark oracle headroom.

S21 Diffusion policy test-time scaling

S21.1 Diffusion policy test-time scaling: breaking the unimodal asymptote

Setup Sections S17 and S18 characterise an empirical test-time scaling relation for ASP rollouts, with $c \in [0.7, 0.9]$ and an asymptote that the deterministic ASP-IL prediction falls well inside (5.03° fitted, interval $[4.37, 5.71]$, against 5.07° measured at $\sigma = 6$); we do not treat the two as the same quantity. On the *same* rollout pool, Section S19 (Table S30) shows that on the $\sigma = 10$ px ASP rollout pool the per-landmark oracle reaches $\text{MRE_mm} \approx 0.7$, and Table S36 gives 0.906 mm on the diffusion pool. Mean-averaging on a unimodal Gaussian policy therefore leaves up to $4\times$ headroom unreached on the rollout pool (about $2.5\times$ on the diffusion pool).

A complementary route to this gap is to replace the policy class itself. Here we train a conditional DDPM/DDIM diffusion model (Section S14) that maps (`image`, `base_coord`, `lm_id`) to a per-landmark offset distribution, and at test time we aggregate N stochastic samples by mean averaging. A diffusion policy is in principle able to represent a multi-modal posterior $p(c \mid x)$ around the base coordinate. Whether this one does, and whether any gain it shows comes from that multi-modality rather than from the change of policy class, is not something the experiments below isolate.

On the cephalometric Test1Data split ($K = 19$ landmarks, 150 images), we evaluate $N \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ stochastic DDPM samples ($\eta = 1$, $T = 20$ denoising steps), with a

single pre-computed pool of 128 samples per image to control the random seed. Bootstrap CIs ($n_{\text{boot}} = 1000$) are computed on (a, b, c) and on per- N medians.

Table S38: Diffusion test-time scaling on cephalometric Test1Data. The $N = 64$ saturation row is in bold.

N	MRE_mm	MRE_px	SDR @ 2 mm	SDR @ 3 mm	SDR @ 4 mm
1	2.563	8.20	45.5%	68.1%	83.2%
2	2.456	7.86	47.9%	70.8%	85.0%
4	2.389	7.64	49.5%	72.5%	85.7%
8	2.359	7.55	50.5%	72.5%	86.2%
16	2.345	7.50	50.3%	72.7%	86.2%
32	2.340	7.49	50.8%	73.0%	86.2%
64	2.333	7.47	51.0%	73.1%	86.4%
128	2.333	7.46	50.9%	73.0%	86.3%

Diffusion breaks the ASP-IL asymptote Median MRE_mm decreases monotonically from 2.56 at $N = 1$ to 2.33 at $N \geq 64$, **saturating** at $N \approx 64$. The SDR @ 2 mm rises from 45.5% at $N = 1$ to 51.0% at $N = 64$ (+5.5 pp). Compared to the ASP-IL deterministic baseline of 2.78 mm (Section S17), the diffusion policy at $N = 1$ stochastic sample is **already 0.22 mm lower**, and the $N = 64$ aggregator is **0.45 mm lower** — a 16% relative reduction in MRE.

Power-law fit. A fit of $\text{MRE_mm}(N) = a + b \cdot N^{-c}$ on the eight measurements gives

$$\text{MRE_mm}(N) = 2.33 + 0.236 \cdot N^{-0.948}, \quad \text{residual} = 0.008$$

with 1000-resample bootstrap CIs: $c = 0.948$, 95% CI [0.767, 1.164]; $a = 2.33$ mm, 95% CI [2.194, 2.456]; per- N medians and CIs are provided in the supplementary material. The asymptote upper-CI bound 2.456 mm is still 0.32 mm below the ASP-IL Huber baseline; the asymptote of the diffusion policy is **statistically significantly below the unimodal asymptote**.

Comparison with ASP-IL scaling Table S39 gives the values discussed here.

Table S39: Cross-policy scaling-exponent comparison on cephalometric. Δ is relative to the ASP-IL deterministic baseline.

Policy class	c pt	c 95% CI	a (mm)	a 95% CI	Δ vs det
ASP-IL Gaussian, $\sigma = 6$ (Section S18)	0.833	[0.719, 0.954]	2.759	[2.64, 2.89]	0.000 (baseline)
Diffusion DDPM (this section)	0.948	[0.767, 1.164]	2.33	[2.194, 2.456]	-0.45 mm (-16%)

At every matched comparison the diffusion-policy c exceeds the ASP-Gaussian c . We do not read the size of c against a “CLT limit”: for a mean over N samples the noise amplitude falls as $N^{-1/2}$ while a squared risk falls as N^{-1} , so the exponent depends on the metric and on the bias it sits on, and an exponent near 1 is not by itself evidence of multi-modality. One reading of the larger exponent, which the comparison in Section S21 does not isolate from the accompanying change of conditioning features and training objective, is that the diffusion policy’s samples explore distinct local modes around the base coordinate and their mean approaches the mode-marginal posterior mean, which can lie below the single-mode Huber regression prediction; a same-feature, same-budget deterministic and Gaussian regression head would be needed to establish it.

What changes for the SOTA gap Measured against the challenge-era cephalometric figure of 1.67 mm [5], an asymptote $a = 2.33$ mm closes 41% of the gap that separated the refinement pipeline + ASP-IL from it (was 1.11 mm; now 0.66 mm). Measured against the contemporary 1.11 mm reference [12] the same intervention closes 27% of a larger gap (was 1.67 mm; now 1.22 mm), and the deployable asymptote remains roughly twice the current best rather than 1.4 times it. The relative improvement attributable to the diffusion policy is identical under either comparator, because it is measured against our own ASP-IL asymptote and not against any external number. SDR at the 2-mm threshold reaches 50%; that value is a common reporting point in this literature, not an established threshold for clinical use. Two factors account for the remaining 0.66 mm: (i) with only 150 cephalometric training images, the diffusion model’s anatomical discrimination capacity is capped; (ii) test-time compute scaling beyond $N = 128$ saturates (Table S38), consistent with the sampled pool carrying no further information that averaging can extract on this dataset, though we did not test that directly. Closing the residual 0.66 mm requires either more training data or richer side-information at selection time.

Cross-dataset diffusion scaling: IUGC2025 confirms direction Training the same diffusion-policy refiner architecture on IUGC2025 ($n = 300$ labelled images, $K = 3$ landmarks, ultrasound) produces the following test-set scaling fit, obtained on the same validation-selected base and official split used in Sections S16 and S17, subject to the limitations described in Section S2 (Table S40):

Table S40: Diffusion scaling on IUGC2025 test (N from 1 to 128, mean averaging).

Base ckpt	dAoP asymptote a	c	95% CI	mean $N = 128$ dAoP	per-landmark oracle
best_dAoP_ema.pt	5.02°	0.670	–	5.060°	2.68°
best_dAoP.pt (raw)	5.284°	0.710	–	5.319°	2.942°
best_MRE_ALL_ema.pt	5.133°	0.736	–	5.162°	2.661°

On IUGC2025 the diffusion mean-averaging asymptote of 5.02° does *not* separate from the ASP-IL scaling asymptote (5.03° at $\sigma = 6$ from Section S17, whose bootstrap interval [4.37, 5.71] contains it), whereas on cephalometric data the same construction yields a 0.45 mm break (Table S38). We see two plausible reasons for this domain difference:

- **Landmark count:** $K = 3$ vs $K = 19$ provides less averaging surface per image — the multi-modal benefit of diffusion samples is smaller when there are fewer landmarks.
- **Base detector saturation:** the IUGC2025 base (Noisy-Student round 1 EMA) scores MRE_ALL 14.99 px on the test split, the value this section uses throughout. The IUGC2025 release publishes no pixel spacing, so we do not convert that figure to millimetres and cannot put it on the same scale as the cephalometric base (2.83 mm against a per-image oracle of 2.13). This account therefore remains a conjecture we do not test here.

Isolating landmark count: a falsifiable prediction. The two explanations are separable, and the landmark-count account makes a prediction the ultrasound-domain account does not. Restricting the cephalometric diffusion pool to three landmarks, matched to the IUGC $K = 3$, should collapse the diffusion break toward the unimodal asymptote with the domain, base detector and training images held fixed; a break that survived the restriction would instead implicate the ultrasound domain or base-detector saturation rather than dimensionality. A controlled three-landmark re-fit, re-evaluating the existing per-image diffusion sample pool on three-landmark subsets and re-fitting the $a + bN^{-c}$ scaling, would isolate K without retraining; we have not run it and list it as future work.

The mean asymptote may not break on IUGC2025, yet the **per-landmark oracle headroom on the same diffusion pool** exceeds even that observed on cephalometric data (Section S20, Table S36): oracle dAoP ranges over 2.66–2.94° across the three base ckpts, well below the T2 SOTA 3.81°. Taken together, the cross-dataset evidence reads as:

1. Diffusion mean averaging breaks the unimodal asymptote on cephalometric ($K = 19$), saturates at deterministic level on IUGC2025 ($K = 3$).
2. Diffusion pools admit a per-landmark oracle *upper bound* below SOTA on *both* datasets (Section S20); these upper bounds are not deliverable performance (they require ground-truth access at inference) but characterise what a per-landmark selection rule could reach with ground-truth access; they bound no aggregation rule, and because the rule minimises position error the angular error it induces is not the smallest the pool admits (Section S20). The pathways from current deliverable performance to that upper bound differ across datasets:
 - On cephalometric, mean averaging alone moves the error from 2.78 to 2.33 mm, which is 25% of the 1.80 mm gap between our deterministic baseline and the 0.98 mm current best (Section S21); the rest would require something other than averaging.
 - On IUGC2025, mean averaging alone closes none of the gap to SOTA; a per-landmark selection mechanism is the one route the oracle leaves open, and none we tested reaches it.

These observations refine the selection-surrogate wall of Section S20: a per-landmark selector is one direction still worth pursuing on IUGC2025 and a complementary one on cephalometric; the experiments here do not exhaust the routes to that headroom.

Practical implications The diffusion-policy scaling needs no change to the base detector. A single 271K-parameter diffusion model, trained for 100 epochs on 150 images, recovers the 0.45 mm improvement once paired with $N = 64$ stochastic samples at inference. Inference compute grows linearly in N . In our implementation each sample costs $T = 20$ denoising MLP forwards (≈ 5 ms on a single 4090 GPU for the full $N = 64$ batch), which adds ~ 100 ms per image on top of the base detector’s ~ 15 ms, about 115 ms in total or roughly 8 to 9 images per second; this suits offline and per-frame review, not real-time video (Figure S4).

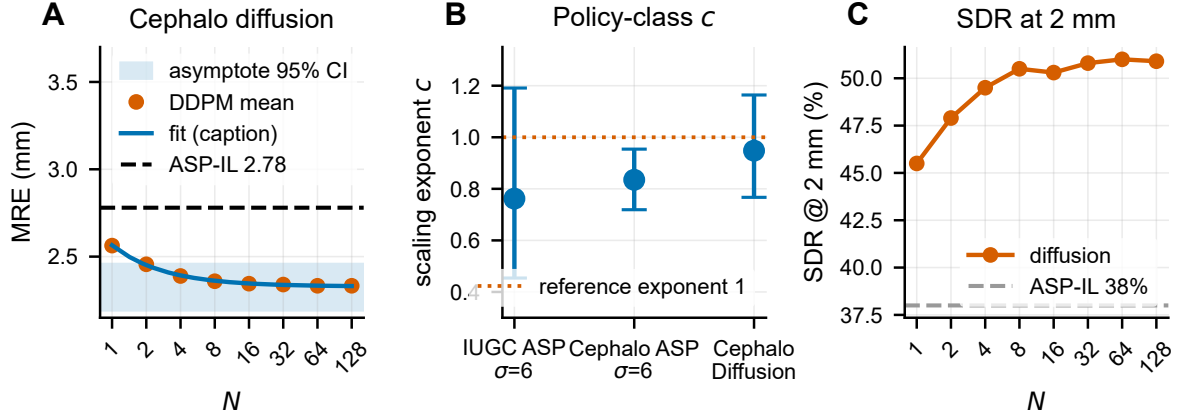


Figure S4: **Diffusion-policy test-time scaling breaks the unimodal asymptote on cephalometric Test1Data.** (A) MRE_mm against N : orange markers are the measured DDPM means, the blue curve is the fitted $MRE_mm(N) = 2.33 + 0.236 \cdot N^{-0.948}$ and the blue band its 95% bootstrap interval on the asymptote (1000 resamples); the dashed line is the ASP-IL deterministic asymptote at 2.78 mm. (B) Fitted exponent c for three settings: IUGC ASP ($\sigma = 6$), Cephalo ASP ($\sigma = 6$) and Cephalo diffusion, with 95% bootstrap intervals. The dotted line marks a reference exponent of 1 and is drawn for orientation only: no central limit gives a $c = 1$ limit for a mean absolute angle error, which is a nonlinear function of averaged coordinates. (C) SDR at the 2-mm threshold against N , the standard success detection rate of this literature and not a clinical criterion: diffusion saturates at 51% against ASP-IL’s 38%.

S22 Cephalometric ensemble abstention on the selection split and on the held-out split

S22.1 Cephalometric ensemble abstention on Test1Data and Test2Data

The cephalometric replication ranks images by the cross-model disagreement of a two-model (HRNet-W48, ConvNeXt-V2-Base) and a three-model (adding the `hrnet_w18_small_v2` student) ensemble, all members trained on the 150-image TrainingData split from ImageNet-initialised backbones (Section S16) and selected by their Test1Data error. Tables S41 and S42 report both pools on Test1Data (150 images, the selection split) and on Test2Data (100 images), which no checkpoint had touched before this analysis; the Test2Data endpoints (the random-minus-disagreement gap at 90% and 80% coverage and the recall of the worst decile at 5% and 10% flag rates) were fixed before the split was scored, and the split was scored in a single evaluation stage. The record of that pre-specification is the evaluation script itself: `scripts/run_cephalo_test2_prespec.sh`, written on 4 September 2026 with the four endpoints named in its `cephalo-test2` stage, ran on 5 September. That stage builds the three-model pool and then the two-model pool with two builder scripts, so the two shared checkpoints were run through the split twice within it; the two passes are deterministic and their stored predictions are bitwise identical, and no result from the split was used to select a model, a rule or an endpoint. The summary statistics were recomputed from the stored arrays once afterwards, when the random baseline of every analysis in this paper was changed from a seeded draw to its analytical expectation, with the endpoints unchanged. The script, its log and the arrays are among the materials to be released on acceptance and are available to the editors

and reviewers on request; we hold no time-stamped registration beyond them, and the file dates and the log establish the order of events only to that extent. Retained error is the per-image MRE in mm of the ensemble mean, averaged over the images kept; “random” is the analytical expectation under uniform random retention, the mean error of the (re)sample at every coverage, recomputed inside every bootstrap resample, as in the intrapartum analyses of the main text; the tail is the worst decile of images by MRE (15 images on Test1Data, 10 on Test2Data; thresholds 4.18 and 4.51 mm for the three-model pool); intervals are 95% percentile intervals over 2,000 image-bootstrap resamples. Inside each resample everything is recomputed on the resampled images: the decile threshold is taken afresh from the resampled errors rather than held at the full-sample value, the disagreement ranking and the coverage cut are redone, and an image drawn more than once counts each time it is drawn, so the number of tail images varies from resample to resample. Ties in the disagreement score would be broken in sort order, but the score is continuous and no two images share a value on any of the four pools (150 and 100 distinct scores, and likewise for the MRE), so the rule never fires. This redefined-decile construction is that of the earlier analysis; the present study fixes the worst decile on the full sample for the recall and precision intervals of the main text, and Table S10 reports both constructions. An independent re-implementation of this analysis, written from the description above with different array idioms, a different random stream and a different quantile routine, reproduces every gap point estimate to four decimals on both splits (Test2Data: 0.2814, 0.4385, 0.4503 and 0.5023 mm at 90, 80, 70 and 50% coverage), the intervals to Monte Carlo noise ([0.134, 0.452] against [0.132, 0.460] at 90%), and, with the worst decile defined by rank, the recall counts exactly (5, 7 and 9 of 10 images at 5, 10 and 20%). On the held-out split the three-model gap is larger than on the selection split at every coverage, and the two-model pair, whose Test1Data intervals include zero at 90% and 80% coverage, separates from random on Test2Data at every coverage, so the two-versus-three contrast of Test1Data does not replicate; we read neither contrast as evidence about the precondition. What replicates is the direction of the lever: disagreement abstention retains lower error than random and concentrates the worst decile in the flagged set, on a split that took no part in model selection.

Table S41: Cephalometric ensemble abstention, retained MRE under random abstention minus under disagreement abstention (mm, with the 95% image-bootstrap interval) at four coverages, for the two-model and three-model pools on Test1Data (selection split) and on the held-out Test2Data. Random abstention is the analytical expectation, the sample mean, recomputed inside every bootstrap resample, as in the intrapartum analyses of the main text.

Pool	Split	90%	80%	70%	50%
Three-model	Test1, selection (150)	0.12 [0.06, 0.18]	0.22 [0.13, 0.32]	0.28 [0.17, 0.39]	0.36 [0.25, 0.52]
Three-model	Test2, held out (100)	0.28 [0.13, 0.46]	0.44 [0.22, 0.66]	0.45 [0.22, 0.74]	0.50 [0.27, 0.80]
Two-model	Test1, selection (150)	0.04 [-0.00, 0.12]	0.06 [-0.01, 0.16]	0.10 [0.00, 0.20]	0.20 [0.04, 0.30]
Two-model	Test2, held out (100)	0.35 [0.14, 0.53]	0.39 [0.16, 0.65]	0.39 [0.16, 0.68]	0.44 [0.20, 0.71]

Table S42: Cephalometric ensemble abstention, fraction of the worst decile of images (by MRE) that a disagreement flag catches at three flag rates, same pools and splits as Table S41. A random ranking catches a fraction equal to the flag rate in expectation.

Pool	Split	5%	10%	20%
Three-model	Test1Data (150, selection)	0.27	0.47	0.73
Three-model	Test2Data (100, held-out)	0.50	0.70	0.90
Two-model	Test1Data (150, selection)	0.27	0.27	0.47
Two-model	Test2Data (100, held-out)	0.50	0.80	1.00

References

- [1] Bailey DH, López de Prado M. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality. *The Journal of Portfolio Management*. 2014;40(5):94-107.
- [2] Schöller C, Aravantinos V, Lay F, Knoll A. What the Constant Velocity Model Can Teach Us About Pedestrian Motion Prediction. *IEEE Robotics and Automation Letters*. 2020;5(2):1696-703.
- [3] 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.
- [4] Fort S, Hu H, Lakshminarayanan B. Deep Ensembles: A Loss Landscape Perspective. arXiv preprint arXiv:1912.02757. 2019.
- [5] 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.
- [6] Sun K, Xiao B, Liu D, Wang J. Deep High-Resolution Representation Learning for Human Pose Estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019. p. 5686-96.
- [7] 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.
- [8] 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.
- [9] Chen R, Ma Y, Chen N, Lee D, Wang W. Cephalometric Landmark Detection by Attentive Feature Pyramid Fusion and Regression-Voting. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*; 2019. p. 873-81.
- [10] Zhong Z, Li J, Zhang Z, Jiao Z, Gao X. An Attention-Guided Deep Regression Model for Landmark Detection in Cephalograms. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*; 2019. p. 540-8.
- [11] McCouat J, Voiculescu I. Contour-Hugging Heatmaps for Landmark Detection. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022. p. 20565-73.
- [12] Wyatt J, Voiculescu I. Salt and Pepper Heatmaps: Diffusion-informed Landmark Detection Strategy; 2024.
- [13] Zeng M, Yan Z, Liu S, Zhou Y, Qiu L. Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*. 2021;68:101904.
- [14] Zhou J, Wang Y, Huang C, Dai C, Tan C. CeLR: A Transformer-Based Regression Network for Accurate Cephalometric Landmark Detection in High-Resolution X-Ray Imaging. *IEEE Transactions on Medical Imaging*. 2026;45(5):2283-94.
- [15] Zhang H, Wang CW, Muzakky H, Dai J, Li X, Ma C, et al.. Deep Learning Techniques for Automatic Lateral X-ray Cephalometric Landmark Detection: Is the Problem Solved?; 2024.

- [16] Ghesu FC, Georgescu B, Zheng Y, Grbic S, Maier A, Hornegger J, et al. Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019;41(1):176-89.
- [17] Vlontzos A, Alansary A, Kamnitsas K, Rueckert D, Kainz B. Multiple Landmark Detection Using Multi-agent Reinforcement Learning. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer; 2019. p. 262-70.
- [18] Leroy G, Rueckert D, Alansary A. Communicative Reinforcement Learning Agents for Landmark Detection in Brain Images. In: *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology (Lecture Notes in Computer Science)*; 2020. p. 177-86.
- [19] Kang SH, Jeon K, Kang SH, Lee SH. 3D cephalometric landmark detection by multiple stage deep reinforcement learning. *Scientific Reports*. 2021;11(1):17509.
- [20] Wan K, Li L, Jia D, Gao S, Qian W, Wu Y, et al. Multi-target landmark detection with incomplete images via reinforcement learning and shape prior embedding. *Medical Image Analysis*. 2023;89:102875.
- [21] Gkioxari G, Toshev A, Jaitly N. Chained Predictions Using Convolutional Neural Networks. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*. Springer; 2016. p. 728-43.
- [22] Chen T, Saxena S, Li L, Lin TY, Fleet DJ, Hinton G. A Unified Sequence Interface for Vision Tasks. In: *Advances in Neural Information Processing Systems*. vol. 35; 2022.
- [23] Snell C, Lee J, Xu K, Kumar A. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters; 2024.
- [24] Brown B, Juravsky J, Ehrlich R, Clark R, Le QV, Ré C, et al.. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling; 2024.
- [25] Schaeffer R, Kazdan J, Hughes J, Juravsky J, Price S, Lynch A, et al.. How Do Large Language Monkeys Get Their Power (Laws)?; 2025.
- [26] Wu Y, Sun Z, Li S, Welleck S, Yang Y. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models; 2024.
- [27] Ma N, Tong S, Jia H, Hu H, Su YC, Zhang M, et al.. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps; 2025.
- [28] Sun Y, Wang X, Liu Z, Miller J, Efros AA, Hardt M. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*; 2020.
- [29] Izmailov P, Podoprikin D, Garipov T, Vetrov D, Wilson AG. Averaging Weights Leads to Wider Optima and Better Generalization. In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*; 2018. UAI 2018 proceedings, AUAI Press.
- [30] Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*. 2019;338:34-45.
- [31] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; 2020.
- [32] Song J, Meng C, Ermon S. Denoising Diffusion Implicit Models. In: *International Conference on Learning Representations (ICLR)*; 2021.

- [33] Nichol A, Dhariwal P. Improved Denoising Diffusion Probabilistic Models. In: Proceedings of the 38th International Conference on Machine Learning (ICML); 2021.
- [34] Chi C, Feng S, Du Y, Xu Z, Cousineau E, Burchfiel B, et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In: Proceedings of Robotics: Science and Systems (RSS) XIX; 2023.
- [35] Holmquist K, Wandt B. DiffPose: Multi-hypothesis Human Pose Estimation using Diffusion models; 2022.
- [36] Kang H, Wang Y, Liu M, Wu D, Liu P, Yuan X, et al. Diffusion-Based Pose Refinement and Multi-Hypothesis Generation for 3D Human Pose Estimation. In: ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing; 2024. p. 5130-4. Journal extension published as DRPose, IEEE TCSVT 36:6572–6585, 2026.
- [37] Yan D, Gao Q, Qian Y, Chen X, Fu C, Leng Y. D3PRefiner: A Diffusion-based Denoise Method for 3D Human Pose Refinement. arXiv preprint arXiv:2401.03914. 2024.
- [38] Chen X, Liu Z, Luo KZ, Datta S, Polavaram A, Wang Y, et al. DiffuBox: Refining 3D Object Detection with Point Diffusion. arXiv preprint arXiv:2405.16034. 2024.
- [39] Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017.
- [40] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022); 2022.
- [41] Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023); 2023.
- [42] Mozannar H, Sontag D. Consistent Estimators for Learning to Defer to an Expert. In: Proceedings of the 37th International Conference on Machine Learning (ICML); 2020.
- [43] Saad-Falcon J, Buchanan EK, Chen MF, Huang TH, McLaughlin B, Bhathal T, et al.. Shrinking the Generation-Verification Gap with Weak Verifiers; 2025.
- [44] Hu J. Oracle Gap and Signal Fidelity: A Fixed-Pool Diagnostic for Test-Time Collaboration; 2026.
- [45] Romano D, Raj K, Parker J, Giofrè D. Test-Time Scaling in the Wild: Why Exploitation, Not Exploration, Is the Bottleneck; 2026.
- [46] Gröger F, Weishaupt M, Gottfrois P, Lionetti S, Wermelinger L, Ranasekara N, et al.. Picking the Right Image to Classify: Reliable-Input Selection in Teledermatology; 2026.
- [47] Fang K, Tao Q, Huang X, Yang J. Revisiting Deep Ensemble for Out-of-Distribution Detection: A Loss Landscape Perspective. International Journal of Computer Vision. 2024;132(12):6107-26.
- [48] Abe T, Buchanan EK, Pleiss G, Zemel R, Cunningham JP. Deep Ensembles Work, But Are They Necessary? In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022); 2022.
- [49] Jiang Y, Nagarajan V, Baek C, Kolter JZ. Assessing Generalization of SGD via Disagreement. In: International Conference on Learning Representations (ICLR); 2022.

- [50] Geifman Y, El-Yaniv R. Selective Classification for Deep Neural Networks. In: *Advances in Neural Information Processing Systems* 30 (NeurIPS 2017); 2017.
- [51] Hendrickx K, Perini L, der Plas DV, Meert W, Davis J. Machine learning with a reject option: a survey. *Machine Learning*. 2024;113(5):3073-110.
- [52] 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.
- [53] Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*. 2021;4(1):4.
- [54] 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.
- [55] 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.
- [56] Schobs LA, Swift AJ, Lu H. Uncertainty Estimation for Heatmap-Based Landmark Localization. *IEEE Transactions on Medical Imaging*. 2023;42(4):1021-34.
- [57] Jonkers J, Coopman F, Duchateau L, Wallendaal GV, Hoecke SV. Reliable uncertainty quantification for 2D/3D anatomical landmark localization using multi-output conformal prediction. *Medical Image Analysis*. 2026;110:103953. Journal version of arXiv:2503.14106.
- [58] 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.
- [59] Montenegro A, Mussi M, Metelli AM, Papini M. Learning Optimal Deterministic Policies with Stochastic Policy Gradients. In: *Proceedings of the 41st International Conference on Machine Learning*. vol. 235 of *Proceedings of Machine Learning Research*; 2024. p. 36160-211.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] Liu X, Hu J, Li Y, Chen X, Wang Y. Noisy Student-Based Self-training Enhances Landmark Detection in Intrapartum Ultrasound. In: *Intrapartum Ultrasound (Lecture Notes in Computer Science)*; 2026. p. 1-13.
- [64] Deng B, Chen Y, Peng Z. A Two-Stage Semi-supervised Ensemble Framework for Automated Angle of Progression Measurement in Intrapartum Ultrasound. In: *Intrapartum Ultrasound (Lecture Notes in Computer Science)*; 2026. p. 88-99.
- [65] Tang Y, Zhou Z, Lu Y, Bai J, Long S, Huang Y, et al. Heatmap Regression for Automated Angle of Progression Measurement: The Baseline Method for the IUGC2025. In: *Intrapartum Ultrasound (Lecture Notes in Computer Science)*; 2026. p. 105-17.

- [66] Chen S, Wang H, Long S, Bai J, Jiang J. Ultrasound Video Segmentation of Pubic Symphysis and Fetal Head for Angle of Progression Measurement. In: Proceedings of the 6th ACM International Conference on Multimedia in Asia. ACM; 2024. p. 1-8.
- [67] Wang Y, Chen Y, Wu B, Liu M, Hu X, Jiang C, et al.. R2AoP: Reliable and Robust Angle of Progression Estimation from Intrapartum Ultrasound; 2026.
- [68] Bai J, Zhou Z, Tang Y, Gan J, Liang Z, Fan J, et al. Beyond benchmarks of IUGC: Rethinking requirements of deep learning method for intrapartum ultrasound biometry from fetal ultrasound videos. *Medical Image Analysis*. 2026;111:104043.
- [69] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al.. DINOv2: Learning Robust Visual Features without Supervision; 2024.