This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

From PACS to Morphometrics: A Hospital-Integrated Workflow for Hip Digital Twin Research

Sections follow the Phase 1 pipeline in the order it runs, then the Phase-2 detail. Cohort names follow the main manuscript, and every experiment states the hips it ran on.

S1 Bone and labrum inference policies

MedSAM2-Anatomy takes a localization prior M_a^{exp} from a frozen expert, prompts a frozen MedSAM2 with it, and keeps the refinement only if fixed plausibility rules accept it. No weight is updated and there are no human prompts. Bone and labrum share this composition and differ only in the execution policy (Table S1): for each plane $v \in \mathcal{V}$ the volume is re-oriented so that v becomes the frame axis, a seed set $\mathcal{S}_v$ of frames is prompted, each prompt is propagated with half-width k, and the passes are unioned into one 3D candidate; the candidates are combined by a vote with quorum κ and, for the labrum, intersected with a band of width δ around the prior. The frame axis is resolved per subject from the direction cosines, the acquisition axis being anatomically axial in 5 of the 12 expert-annotated hips and sagittal in the other 7.

The osseous harness is the published configuration unchanged, and its gain comes from memory propagation along a thick, high-contrast cross-section. The labral harness inverts nearly every setting, and Section S3 shows that each inversion earns its place. Two choices carry most of it. The prior is cropped to its bounding box dilated by 10 mm and re-windowed inside that crop, because the labrum occupies under 1 % of the field of view and global windowing maps fibrocartilage and contrast-filled joint space to the same grey levels. Propagation is switched off and every prior-bearing frame becomes its own seed, so MedSAM2 acts as a 2D boundary refiner rather than a tracker; the positive points sit at maxima of the within-component distance transform, placing them on the ridge of a structure three to five voxels thick rather than on its edge.

S2 The nnU-Net HipLabrum localizer

HipLabrum supplies the labral prior: nnU-Net v2 in the `3d_fullres` configuration, a six-stage PlainConvUNet trained from scratch for 100 epochs on $80 \times 160 \times 192$ patches, batch size two, Dice plus cross-entropy loss, polynomial learning-rate decay, target spacing $0.60 \times 0.46 \times 0.44$ mm. It localises; it does not delineate.

The model is trained by four-fold cross-validation and deployed as the four-fold ensemble. The 12 expert-annotated volumes were partitioned into four folds of three (fold 0 KLI_11/36/41, fold 1 KLI_19/40/8, fold 2 KLI_14/38/44, fold 3 KLI_2/18/4); each fold trained one checkpoint on the other nine volumes, and

Table S1: Bone and labrum inference policies. Both instantiate the same MedSAM2-Anatomy composition with frozen weights and no human prompting and differ only in the execution policy. $k{=}\infty$ is bidirectional propagation over the whole stack, $k{=}0$ confines each prompt to its own frame; P counts foundation-model passes per anatomy.

	Osseous harness (Ac, Fe)	Labral harness (La)
Expert localizer E_a	TotalSegmentator, frozen, trained outside the cohort	HipLabrum nnU-Net, four-fold ensemble on this cohort
Preprocessing	Native grid, global window	10 mm ROI crop, local window
Prompt encoding	Dense mask (bypasses decoder)	Box + 6 positive points
Negative prompts	—	≤ 3 inside partner bone
Prompt granularity	One object per slice	One object per component
Seed set $\mathcal{S}_v$	3 interior frames	Every prior-bearing frame
Propagation k	∞ (bidirectional)	0 (none)
Planes $\mathcal{V}$	Acquisition plane (1)	Axial, sagittal, coronal (3)
Within-plane fusion	Union over seeds	Concatenation over frames
Cross-plane quorum κ	1 (vacuous)	2 of 3
Prior constraint	Laterality gate, $\rho{=}0.15$	Band $\delta{=}3$ mm
Passes P	3	Tens to low hundreds
Source of gain	Memory propagation	Per-frame prompt decoding

Table S2: Distributional statistics behind the labrum rows of main-text Table 2, on the Expert-Annotated Evaluation Cohort ($n{=}12$): every nnU-Net prior comes from the fold that did not train on that hip, and both MedSAM arms were prompted from that same prior. Subject-level DSC; 95 % CI from 10,000 bootstrap resamples. Mean HD95 is 5.00, 4.99 and 4.87 mm, no pair differing significantly.

Method	Mean ± SD	Median [IQR]	95 % CI
nnU-Net (held-out fold)	**0.635 ± 0.115**	0.641 [0.590, 0.723]	[0.571, 0.695]
MedSAM2, single plane	0.572 ± 0.079	0.575 [0.538, 0.599]	[0.530, 0.615]
MedSAM2-Anatomy	0.620 ± 0.101	0.621 [0.578, 0.688]	[0.565, 0.673]

production inference averages the four checkpoints. Phase 1 accuracy (main-text Table 2) is scored from the fold that held each hip out, so no hip is scored by a model that trained on it; Phase 2 consumes the production ensemble masks and is in-sample. On the production masks the ensemble prior reaches mean DSC 0.839 against 0.776 after MedSAM2-Anatomy refinement, the same ordering as the held-out comparison.

Held-out per-hip Dice spans 0.431–0.803 (Table S2); validation pseudo-Dice plateaus at 0.60–0.68 by epoch 40 while training loss keeps falling, and no fold is anomalous (Fig. S1). The spread tracks acquisition rather than pathology: KLI_02 is near-isotropic at 0.56×0.56×0.60 mm over 144 slices, whereas KLI_08 is 0.44×0.44×1.00 mm over 96 slices with a prior that over-segments at 1.24 times the reference volume. Both refinements inherit that error. Paired Wilcoxon tests (Holm-adjusted) give MedSAM2 below the prior by 0.063 ($p{=}0.029$), MedSAM2-Anatomy below it by 0.015 ($p{=}0.31$), and MedSAM2-Anatomy above MedSAM2 by 0.048 ($p{=}0.046$).

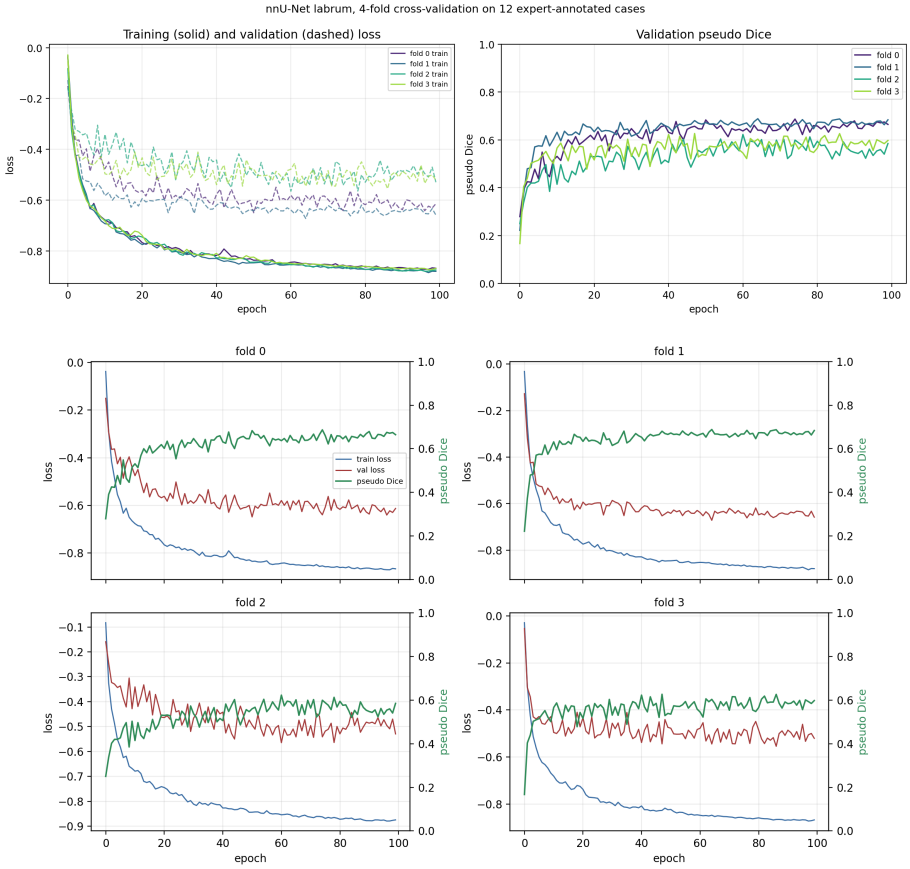


Fig. S1: HipLabrum four-fold training curves used to build the production ensemble. Top: training (solid) and validation (dashed) loss for all folds (nnU-Net convention: negative Dice plus cross-entropy) and validation pseudo-Dice. Bottom: fold-by-fold runs. No fold diverges within 100 epochs.

S3 How the labral policy was derived

The labral harness was screened, not designed, starting from the osseous one. Applied unchanged to the labrum, the osseous policy does not degrade, it fails: DSC 0.048 ± 0.064, median HD95 28.42 mm on all 12 expert-annotated hips, and not one output accepted by quality assurance. Two mechanisms explain it. The released checkpoint re-emits a dense mask prompt verbatim, so the three seeded frames reproduce the prior and everything between them comes from propagation; and an axial labral cross-section is a 2–3 mm arc whose bounding box encloses the femoral head, so memory attention locks onto the head–acetabulum interface and never returns. The failure belongs to the policy, not to the prior or the foundation model.

Table S3 is the screen that followed. It is sequential, not fully crossed: each stage fixes the winner of the previous one, and in stage 1 several factors move

Table S3: Derivation of the labral inference policy; each stage fixes the winner of the previous stage (bold). Stage 1 is a single-hip screen (KLI_11, prior DSC 0.894) slicing along the acquisition axis; stage 4 uses n=4 hips (mean prior DSC 0.904) and resolves the anatomical plane from the direction cosines. Stages 2 and 3, the threshold and point-density sweeps on the same four hips, are reported in the text. k, propagation half-width in slices; Bone, fraction of predicted voxels inside bone.

Configuration	DSC ↑	Prec.	Rec.	HD95 ↓	Bone ↓
Stage 1 – prompt encoding and propagation (n=1)					
Mask prompts, 3 slices (osseous policy)	0.152	0.241	0.111	7.54	0.511
Single ROI box, whole volume	0.315	0.380	0.268	5.49	0.144
Per-component box, 3 anchors, k=2	0.199	0.472	0.126	16.27	0.008
Per-comp. box + points, 5 anchors, k=2	0.345	0.620	0.239	7.00	0.020
Per-comp. box + points, dense (k=0)	0.770	0.698	0.858	1.66	0.031
+ negative points (anatomy layer)	**0.771**	0.714	0.838	1.60	0.026
Stage 4 – slicing plane and fusion (n=4; τ=0, six points, 1 voxel padding)					
Axial, dense	0.757	0.724	0.808	1.63	0.026
Sagittal, dense	0.726	0.753	0.715	1.44	0.030
Coronal, dense	0.779	0.758	0.808	1.26	0.018
Sagittal, 5 anchors, k=3	0.411	0.680	0.297	4.37	0.037
Multi-planar, union	0.741	0.622	0.935	1.93	0.032
Multi-planar, 2-of-3 majority	**0.818**	0.828	0.814	1.07	0.019

together, so the rows rank configurations rather than isolate marginal effects. Stage 1 ran on one hip (KLI_11, prior DSC 0.894), stages 2–4 on four (mean prior DSC 0.904), and the two closing experiments on all 12. A box that reaches the mask decoder doubled DSC over the dense mask, per-component boxes with ridge points raised it further, and switching propagation off was the largest single gain, 0.345 to 0.770.

Stages 2 and 3 characterise the thin-structure regime, which is recall-limited. Raising the logit threshold from τ=0 to τ=3 raised precision monotonically (0.801 to 0.952) while recall fell (0.711 to 0.368) and DSC dropped (0.752 to 0.525), so τ=0 is used throughout; six positive points per component with one voxel of box padding added +0.039 DSC over two. For a structure a few voxels across, almost every voxel is a boundary voxel, so trading recall for precision removes the structure instead of sharpening it. Stage 4 settled plane and fusion. Single planes order coronal > axial > sagittal, and the sagittal plane, where the labrum appears as the longest continuous arc and a reader finds it most informative, was weakest for the model, that arc also yielding the least specific box. Two-of-three majority fusion beat the best single plane and the union rule used for bone, whose recall rose to 0.935 as precision fell to 0.622.

The anatomy-derived negative points, the only input separating MedSAM2-Anatomy from MedSAM2 in this branch, were measured alone on all 12 expert-annotated hips with production ensemble priors. They barely move overlap (0.776 vs. 0.773 DSC) but do move the operating point, precision rising from 0.757 to 0.794 as recall falls from 0.798 to 0.767. Telling the decoder what the target is not trims predictions that crossed into bone without recovering labral voxels the prompt never reached, so here the anatomy layer is a plausibility rather than an accuracy mechanism.

S4 What does not transfer back to bone

The MedSAM2 column of main-text Table 2 is a single anatomical-axial, box-seeded run without negative points, delivering 0.788 ± 0.219 (acetabulum) and 0.871 ± 0.137 (femur) after quality assurance. Frozen localizers introduce localization failures under acquisition shift, which the QA layer records per mask rather than hiding (Section S5), so scaling to further modalities and acquisition protocols needs acquisition-aware preprocessing and out-of-distribution routing ahead of the frozen expert, not further prompt tuning alone.

Multi-planar majority fusion, the largest single effect on the labrum, does not transfer back to bone. Both arms ran through the same harness, so the comparison isolates plane fusion. Being an internal ablation and not an accuracy report, it uses the ten of the twelve hips whose osseous prior is fully automatic; main-text Table 2 keeps all twelve. On those ten hips the single-plane policy reaches DSC 0.883 ± 0.072 (acetabulum) and 0.940 ± 0.025 (femur), multi-planar two-of-three fusion 0.874 ± 0.051 and 0.937 ± 0.033. No pooled metric moves (DSC -0.006, two-sided Wilcoxon p=0.52 over n=20 paired structure instances; HD95 -0.98 mm, p=0.70; ASSD -0.08 mm, p=0.87); only dispersion improves, the acetabular SD falling to 0.051 and the worst case rising from 0.704 to 0.787 DSC. At roughly three times the inference cost that does not justify adoption, so the osseous branch keeps union fusion over seed slices.

Since the labral refinement does not beat its prior on average, five rules for combining the two were tested on the same 12 hips with production ensemble priors, from prior minus predicted bone (mean DSC 0.837) down to the intersection (0.800). None improved on that prior at 0.839, and the ordering is monotone in how much of the prior each rule retains. A selective policy is no better, failing priors not being identifiable without GT (ρ=+0.27, p=0.40).

S5 Delivery record: quality assurance, provenance and prior failures

Phase 1 feeds Phase 2 with no human in the loop, so every delivered mask carries a machine-checkable acceptance decision and a provenance label. A bilateral prior is admitted only if $|M_a^{\mathrm{exp}}| \geq 0.15\,|M_{\bar{a}}^{\mathrm{exp}}|$, tested before prompting. A refinement is rejected if it is empty, if its volume exceeds the prior by more than 8× (bone) or 2× (labrum), if it recovers less than 30% of the prior, or if more than 10% (bone) or 15% (labrum) of it falls inside neighbouring bone; the osseous ceiling is looser because a correct osseous refinement is routinely 3–5× its under-segmenting prior. A rejected refinement is replaced by the prior and labelled a fallback, so a fallback is never reported as model output. Across the 26 Phase-2 hips the gate accepted all osseous refinements and 25 of 26 labral ones; the single rejection (KLI_38, raw DSC 0.838) shows that GT-free rules also discard good masks. On the 14 hips without expert GT the indicators sit close to the expert-annotated arm (volume ratio 0.997, prior recall 0.873, bone overlap 0.033), which evidences configuration consistency, not accuracy.

No inference policy recovers a failed prior, since boxes, points, crop and band all derive from it. Across the candidate automatic-inference pool, expert localization returned absent or unusable osseous priors on three hips and unusable labral priors on two; three of those five were excluded from Phase 2, and two expert-annotated hips carry a reference-seeded osseous localization, recorded as such in the provenance log, their labral outputs remaining automatic. They stay inside every Phase-1 denominator, so main-text Table 2 scores all three methods on the same 12 hips; osseous localisation is image-derived on 10 of those 12 and on 24 of the 26 Phase-2 hips. Failure is observable without GT through the provenance indicators (degenerate priors yield far fewer anatomy-derived negative points than a usable prior). Frozen models therefore require acquisition-aware preprocessing as the number of modalities and protocols grows; the QA layer records the failure rather than inventing a mask.

S6 Phase 2: morphology–torsion detail and shape-signature agreement

Table S5 gives the strongest morphometric correlation per anatomy and pathology in both arms, the expert-GT cohort (n=12) and the Combined Phase-2 delivered-mask cohort (n=26; 24 hips with fully automatic osseous localisation, two reference-seeded, labral output automatic throughout). Both ran the same pipeline at K=256, so the columns differ only in mask source and cohort size, and three patterns separate replication from small-sample noise. The labral rows keep their sign and gain significance in both pathologies as n grows, which is how a reproducible correlation behaves. The femoral rows do not: the HD mid-axis extent coefficient ($r = -0.96$, the only result to survive correction anywhere) disappears and the FAI major-extent one reverses sign. Cohort extension appears to be the main contributor, and the two arms can be compared at all because the delivered-mask cohort contains the expert-annotated hips: on those seven HD hips the delivered masks reproduce the coefficient at $r = -0.81$ (p=0.029) against $r = -0.96$ (p=0.0006), a loss consistent with segmentation noise, while adding the seven remaining HD hips takes it to $r = +0.09$ (p=0.75); the neighbouring major-extent descriptor behaves identically (-0.77 expert, -0.66 delivered on seven hips, $+0.22$ on fourteen). Segmentation effects in the unlabelled hips cannot be excluded. The acetabular rows weaken but keep their direction, and the combined components stay weaker than the best labral descriptor in both arms.

Two per-arm quantities support this. Cumulative PCA variance over the retained modes is 0.64–0.87 at n=26 against 0.77–0.99 at n=12; the compression at the smaller size is a symptom of too few subjects, not a simpler shape space, the clearest case being the HD femur, whose first mode holds 91% of the variance over seven subjects and 53% over fourteen. The acetabular and femoral HD models groomed 14 of 15 hips, one timeout each on KLI_35 and KLI_23, the labral model all 15; the combined signature uses all 26.

S6.1 Independent bases, and the exported source data

No principal basis is shared between cohorts. Pipeline B optimises one ShapeWorks correspondence system per cohort, anatomy and diagnosis, each with its own medoid reference, so mode k of one cell is a different axis from mode k of another; pipeline A fits one joint PCA per cohort, and its spectra differ likewise. Eigenvector signs are unidentified throughout. Mode and component coefficients are therefore read only inside their own cell, which is why main-text Fig. 3b draws one panel per cell and Table 3 separates the two bases into blocks. The classical descriptors are exempt, being named measurements computed identically in both arms, and Fig. 3a puts the cohorts on shared rows for that reason. A cross-cohort comparison of components would need one n=26 correspondence and PCA basis with the 12 expert-GT hips projected onto it; that variant was not run and is the recommended route if such a comparison is wanted.

Every point of that figure is released with its cohort, diagnosis, anatomy, feature, n, Pearson r, 95% Fisher interval, raw p and BH q: 156 descriptor coefficients, flagging the 12 drawn per panel; 57 mode coefficients with their variance shares, of which the two leading ones per anatomy are drawn; and the 20 joint components of Table 3. Figure and table are built from the delivered correlation tables by the released plotting code and exported as editable PDF and SVG next to PNG. The intervals and the component-level q values are added there, the delivery reporting r and p only.

S6.2 Particle-count sensitivity, $K \in \{256, 2048\}$

Morphometrics are invariant to particle count. On the expert-GT cohort (n=12) every mesh morphometric of acetabulum, femur and labrum is bit-identical at K=256 and K=2048 (maximum absolute difference 0 over the 30 key features). The combined multi-anatomy signature is the joint PCA of the 23 z-normalised morphometric descriptors and takes no particle input; we reproduced the delivered joint spectra and torsion correlations from the morphometrics alone to 10^{-15}. The combined signature, the tertile statistics and the single-versus-combined comparison of Table S5 are therefore invariant to K by construction, and a second optimisation of the n=26 arm is not needed for that claim. That arm ran at K=256 only; the delivered K=2048 directory labelled MedSAM2-Anatomy n=26 is a byte-identical copy of the expert-GT run (twelve subjects from `gt_segs`).

The particle shape modes do move (Table S4). Modes are compared by the absolute correlation between the two score vectors, the eigenvector sign being unidentified, and by $|r|$ against torsion. The leading mode is reproduced (identity 0.996, mean $|\Delta|r||$ 0.020 over the six anatomy-by-pathology models) and identity falls to 0.426 by PC5. Over the 15 of 27 modes preserved at identity ≥ 0.95 the mean $|\Delta|r||$ is 0.062, against 0.290 for the 12 that are not. Two nominal shape-mode correlations survive (acetabular HD PC2, femoral HD PC2), two do not (labral FAI PC3, labral HD PC2, the latter at identity 0.535). Shape-mode results should be read at PC1–PC2 only.

Table S4: Particle-count sensitivity of the shape modes on the expert-GT cohort, aggregated over the six anatomy-by-pathology models. Mode identity is the absolute correlation between the K=256 and K=2048 score vectors, the eigenvector sign being unidentified; $|\Delta|r||$ is the change in absolute correlation with torsion. The morphometric descriptors are bit-identical at the two counts, so only the particle modes move.

Mode	Identity (mean)	Identity (min)	Mean $\|\Delta\|r\|\|$	Max $\|\Delta\|r\|\|$
PC1	0.996	0.985	0.020	0.041
PC2	0.890	0.535	0.132	0.314
PC3	0.796	0.439	0.248	0.442
PC4	0.726	0.154	0.245	0.532
PC5	0.426	0.150	0.180	0.388

S6.3 Torsion correlation within torsion tertiles

Tertiles are reported here rather than in the main text because they are a proxy for diagnosis. Cut points are arm-specific: ≤ 17, ≤ 30 and $> 30°$ at n=12 (4/4/4 hips), ≤ 13, ≤ 25 and $> 25°$ at n=26 (9/8/9). The composition is the problem. The lowest tertile is FAI-dominated (3 of 4 hips at n=12, 7 of 9 at n=26) and the highest holds no FAI hip in either arm, so a tertile contrast is a diagnosis contrast with a restricted torsion range attached; the main text stratifies by diagnosis for that reason. The coefficients follow suit. At n=12 no interval excludes zero, four hips per cell being too few. At n=26 five do, and the two leading descriptors reverse across the range: labral mean width runs -0.72 in the low tertile and $+0.72$ in the high one against $+0.42$ overall, acetabular socket asphericity is negative in both outer tertiles (-0.80, -0.85) against $+0.30$. Reversals are expected once a cell spans one diagnosis and a fraction of the range, and do not indicate an effect that changes direction with torsion. The across-tertile Kruskal–Wallis test is the safer statistic, nominal at n=26 for acetabular rim range and socket asphericity only and reaching nothing below p=0.15 at n=12. Both particle counts share these numbers exactly. Particle modes cannot be tertile-stratified at all, their basis being pathology-specific.

S6.4 Paired shape-signature agreement in a common basis

Main-text Fig. 3 asks whether torsion correlations persist as the cohort grows; this section answers the different question of whether delivered and expert-GT descriptors of the *same* hip agree. Agreement is scored over the whole morphometric signature: both vectors are z-normalised in the expert-GT space and compared by cosine similarity (1 = same position in shape space, 0 = no better than an unrelated hip). Particle shape modes have no shared basis across models and are excluded (last paragraph).

Both refinements clearly beat the frozen expert prior on the medians (acetabulum 0.75 and 0.63 against 0.45, femur 0.94 and 0.95 against 0.23), but none of the pairwise differences is significant on 12 hips. MedSAM2-Anatomy over MedSAM2 is $+0.05$ mean cosine on the acetabulum (two-sided p=1.00) and -0.03 on the femur (p=0.38), and the margin over the prior ($+0.17$ and $+0.34$) is likewise

Table S5: Strongest morphometric correlation with report-documented femoral torsion per anatomy and diagnosis, expert-GT cohort against the Combined Phase-2 delivered-mask cohort, both at K=256: the numeric summary of main-text Fig. 3a, whose 95% intervals are omitted here for space and released in full in the source data. Pearson r, raw p and BH q within each anatomy-by-diagnosis-by-cohort family. Bold marks the only coefficient surviving correction anywhere in Phase 2. The combined signature is main-text Table 3.

			Expert GT (n=12)			Delivered masks (n=26)		
Anatomy	Group	Descriptor	r	p	q	r	p	q
Acetabulum	FAI	Compactness	−0.93	0.021	0.38	−0.64	0.034	0.38
	HD	Volume / elongation*	−0.68	0.093	0.81	−0.50	0.066	0.54
Femur	FAI	Major extent	+0.86	0.063	0.87	−0.47	0.145	0.61
	HD	Mid / minor extent*	**−0.96**	**0.001**	**0.012**	−0.22	0.452	0.92
Labrum	FAI	Compactness	+0.77	0.129	0.80	+0.69	0.020	0.33
	HD	Mean width	+0.83	0.022	0.30	+0.66	0.008	0.13

*Expert-GT / delivered-mask descriptor where the leading one differs between arms. Femoral HD at n=26 also has a nominal shape-mode correlation (PC5, side-controlled partial r=0.59, p=0.025), which no morphometric descriptor supports.

Table S6: Paired agreement, not main-text Table 3: cosine similarity of delivered vs. expert-GT morphometric descriptors on the same 12 hips in one common z-normalised basis. p: one-sided Wilcoxon vs. zero. Frozen expert prior = TotalSegmentator (bone) / HipLabrum (labrum); its osseous rows rest on 11 hips and carry no reference-seeded value, the two hips of Section S5 contributing none (one has no frozen osseous prior at all, the other is scored from that prior and not from the reference), so this column is not flattered. Labrum uses the production four-fold ensemble (in-sample); signature = volume, mean width, mean thickness, compactness.

			Whole cohort		Median by group		
Structure	Model	n	Mean ± SD	Median [IQR]	FAI	HD	p
Acetabulum	Frozen expert prior	11	0.35 ± 0.43	0.45 [0.09, 0.65]	0.54	0.34	0.027
	MedSAM2	12	0.48 ± 0.47	0.63 [0.09, 0.85]	0.37	0.66	0.002
	MedSAM2-Anatomy	12	0.53 ± 0.51	**0.75** [0.44, 0.86]	0.82	0.74	0.005
Femur	Frozen expert prior	11	0.18 ± 0.54	0.23 [−0.16, 0.60]	−0.05	0.41	0.183
	MedSAM2	12	0.59 ± 0.64	**0.95** [0.51, 0.99]	0.99	0.92	0.005
	MedSAM2-Anatomy	12	0.56 ± 0.70	0.94 [0.48, 0.99]	0.88	0.97	0.005
Labrum	MedSAM2-Anatomy	12	0.87 ± 0.18	**0.96** [0.82, 0.99]	0.94	0.99	< 0.001

within noise (p=0.32, p=0.21). The distribution carries what the mean does not: on the acetabulum MedSAM2-Anatomy exceeds 0.44 on 9 of 12 hips and falls below −0.7 on one, so the single failure stays individually identifiable.

The particle shape modes cannot be compared this way: each model, structure and pathology has its own correspondence optimisation and ICP reference, so a GT shape and its delivered counterpart live in different bases, and over all 105 pairs their cosine similarity is 0.03 ± 0.52, the value expected of unrelated vectors. Table S6 therefore uses the morphometric space, a common basis by construction.