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
TABS: Transfer-Aware Backbone Selection for Cross-Centre Ultrasound Biometry

Daniil Sukhorukov, Ekaterina Antipushina, Yury Koush, and Ilya Makarov

This supplement provides the per-task evidence behind the claims of the main paper. Each item states how it was measured, since the numbers come from three different settings: the public validation leaderboard, a held-out split of the labelled source data, and a framing probe that re-runs the same checkpoints under a different preprocessing. No additional training is involved anywhere in this document.

Protocol corrections (Table S1). Both rows report the same unchanged checkpoint before and after a change to the evaluation specification, so the deltas measure the specification and not the model. "Unscored" means the submitted predictions were written for a different image subset than the one the scorer reads, so no sample of that task was evaluated. The AOP row is the correction of a train/evaluation point ordering that placed every team near 73 px regardless of method.

Table S1. Protocol corrections, separated from modeling. Both rows use an unchanged checkpoint; only the evaluation specification changed.

Task	Specification error	Before	After
Cervical length	predictions written for the wrong image subset	unscored	3.59
AOP	evaluation point ordering vs. released convention	73.4	15.82

Per-task effect of sub-pixel decoding (Table S2). The same DINOv2 checkpoint is decoded twice on the same held-out split, once with the integer argmax of the 128^2 heatmap and once with DARK, so the difference isolates grid quantization. Reading it by column shows that the relative gain tracks how small the error already is: tasks whose decoded error is close to one heatmap cell (IVC, PLAX, cervical length) gain most, while tasks limited by something other than rounding (AOP, HC) gain least.

Blind routing probe (Table S3). Each encoder is evaluated on the same held-out *source* images under two preprocessings: the square resize used during training, and an aspect-preserving letterbox that changes framing while leaving appearance untouched. No target-domain labels and no leaderboard feedback enter this table. Applying Eq. (1) of the main paper, that is routing to the encoder

Table S2. Per-task contribution of DARK sub-pixel decoding (held-out split, same DINOv2 checkpoint, 128^2 heatmaps; MRE in px). The relative gain tracks how small the error already is, not the anatomy.

	FUGC	AOP	PLAX	FL	FA	A4C	IVC	PSAX	HC	Macro
Argmax	3.81	6.80	6.59	8.93	8.55	9.95	4.71	12.90	13.43	8.41
+ DARK	**1.81**	6.15	**2.80**	6.26	5.74	7.10	**1.23**	8.26	11.73	**5.68**
Gain (%)	−53	−9	−57	−30	−33	−29	−74	−36	−13	−32

with the lower worst case over the two framings, yields the row $r(t)$; the last row records whether that blind choice matches the route we actually deployed. The one disagreement, cervical length, is also the task with the largest framing penalty for DINOv2 ($1.80 \rightarrow 9.14$ px), which is what the target-shaped canvas randomization later removed.

Table S3. Blind routing probe on held-out source images (MRE, px). R is the worst case over the two framings, and $r(t)$ the route selected by Eq. (1) of the main paper; the last row marks agreement with the deployed system.

	FUGC	AOP	PLAX	FL	FA	A4C	IVC	PSAX	HC
DINOv2 square	1.80	6.17	2.88	6.25	5.75	7.22	1.24	8.36	11.79
DINOv2 letterb.	9.14	6.17	10.49	16.57	16.03	12.61	5.26	11.56	21.62
DINOv3 square	5.54	5.93	16.93	22.31	14.97	36.05	21.43	73.02	23.19
DINOv3 letterb.	5.69	5.93	16.61	21.60	15.08	36.33	16.28	67.19	20.07
$r(t)$ by Eq. (1)	v3	v3	v2	v2	v3	v2	v2	v2	v2
Matches deployed	no	yes	yes	yes	yes	yes	yes	yes	yes

Routing versus unified single-backbone systems (Table S4). Both baselines are complete multi-task systems that share the head, the 128^2 heatmap, the decoding code and the inverse transform with TABS; they differ from it only in that one encoder is used for every task. The four tasks listed are those for which both encoders were measured on the leaderboard, and the values precede the FA geometry projection so that the comparison isolates the encoder choice. TABS never evaluates two encoders on the same image, so the difference is a selection effect rather than added capacity.

Deployed routing recipe (Table S5). This is the operational table for the submitted validation run: the route trusted for each task and the reason it was trusted. It is a record of the system as submitted, not a claim that one backbone is universally stronger.

Decoding components (Table S6). Components are added cumulatively to the DINOv2 path on the ablation split, starting from the challenge-baseline con-

Table S4. Routing versus unified single-backbone systems (validation MRE, px) on the four tasks where both encoders were measured, before the FA geometry projection so that the comparison isolates the encoder choice. Both baselines use the identical head, 128^2 heatmaps and DARK decoding; only the encoder and resize policy differ.

System	AOP	FA	HC	IVC	Mean
All DINOv2-S/14 (square)	15.82	23.03	**42.02**	**24.36**	26.31
All DINOv3-B/16 (letterbox)	**14.44**	**20.95**	48.0	39.0	30.60
TABS routing	**14.44**	**20.95**	**42.02**	**24.36**	**25.44**

Table S5. Task-specific routing used in the final validation submission.

Task	Route	Main reason
FUGC	DINOv2-S/14, square	Short-distance endpoint task; DARK removes most heatmap quantization.
HC	DINOv2-S/14, square	Dense token grid is still preferable for perimeter landmarks despite high evaluation error.
FL	DINOv2-S/14, square	Elongated bone endpoints are localization-limited rather than transfer-limited.
FA	DINOv3-B/16, letterbox	Aspect and field-of-view shift made the precision path overfit the source crop.
AOP	DINOv3-B/16, letterbox	Sector geometry mismatch dominated in-domain precision.
A4C	DINOv2-S/14, square	Cardiac points stayed more stable with the denser feature grid.
PLAX	DINOv2-S/14, square	Local landmark appearance was more important than global geometry robustness.
PSAX	DINOv2-S/14, square	High absolute error, but alternative routes did not improve validation feedback.
IVC	DINOv2-S/14, square	Too few labels for blind tuning; the lowest-risk model was kept.

figuration (argmax decoding, 64^2 heatmaps, MSE loss). DARK contributes the largest single step and requires no retraining; the finer heatmap and the Adaptive Wing loss add smaller amounts. This split differs from the one used in Table S2, so the macro values are not identical.

Table S6. Decoding-component ablation on the DINOv2-S/14 path (held-out macro-MRE, px). The per-encoder transfer probe is reported per task in Table S3.

Configuration	Baseline (argmax, 64^2, MSE)	+ DARK	+ 128^2	+ AWing
Macro-MRE ↓	9.34	6.43	6.10	**5.88**

Qualitative predictions (Fig. S1). Predictions after DARK decoding on representative prenatal, intrapartum and cardiac frames, with ground truth in green and predictions in red. The panels are chosen to show both the typical case and the failure modes that an aggregate MRE hides, in particular endpoint drift along elongated structures and displaced points on views whose framing differs from training.

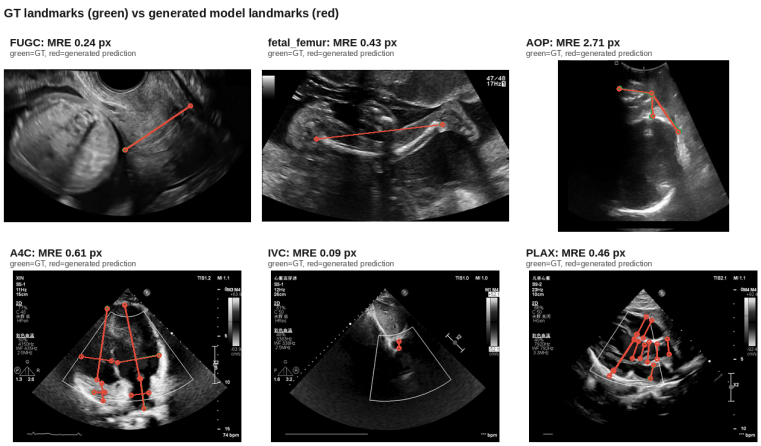


Fig. S1. Representative local predictions after DARK decoding. Green marks ground truth and red marks predicted landmarks.

System overview (Fig. S2). The routing decision is the only task-dependent part of the pipeline. Once an encoder is selected, every task shares the same convolutional heatmap head, heatmap resolution, DARK decoding, inverse coordinate transform and clinical-parameter computation, and this is what makes the single-backbone baselines in Table S4 directly comparable.

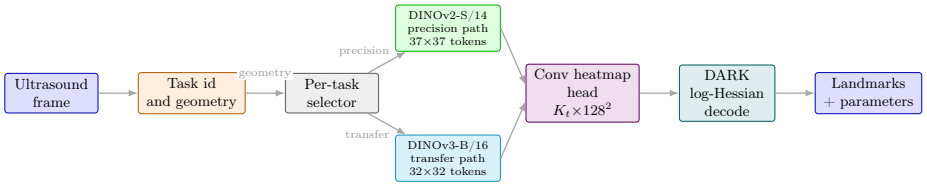


Fig. S2. TABS overview. Each task is routed to the encoder family that transfers best, then decoded through the same heatmap and DARK coordinate interface.

S1 Negative results in detail

The main paper summarises the approaches that did not work. This section reports what was run and what each failure implies about the nature of the shift. TENT-style test-time adaptation and the feature-pyramid decoder are cited in the main paper's bibliography.

Adaptation was fragile as well, and each failure narrows down the kind of shift this benchmark contains. *Test-time adaptation:* entropy minimization sharpened heatmaps that were already centred on the wrong location, raising confidence without moving the coordinate, and recovered a negative fraction on a measurable synthetic-shift proxy.

Heatmap ensembling: averaging maps from checkpoints that disagree yields a bimodal map whose argmax lies between the modes, a location neither model predicted; coordinate-level per-task selection was consistently better.

Photometric (Fourier) adaptation: replacing the low-frequency amplitude of each evaluation image with a source donor, a training-free style transfer that works well for MRI scanner shift, degraded every task it touched (IVC 24.4 → 76.3, A4C 23.7 → 32.0, PSAX 37.8 → 41.7 px). If the shift were photometric this should have helped, so the residual is geometric.

Capacity: doubling the input and adding a feature-pyramid decoder left the DINOv3 path near 26 px in-domain, because keypoint identity lives in the deep, coarse tokens.

Two further failures concern per-task retraining. On IVC an eight-image in-domain split rewarded variants that worsened evaluation from 24 to 38 and 70 px. Retraining a femur specialist after removing 25 mirrored training images (a correct data fix) also scored worse than the multi-task checkpoint trained with them (20.44 → 25.52 px): the specialist overfits a small task faster than the cleaning helps. Small-data tasks are therefore left on the shared checkpoint by design.

S2 Training and inference details

Backbones are fine-tuned end-to-end with task-specific heads (DINOv2: square 518×518, batch 4; DINOv3: letterbox 512×512, batch 10; both 40 epochs, encoder lr 2×10^{-5}, head lr 10^{-3}). The offline harness reports per-task MRE, pa-

rameter error, bootstrap intervals, and both square-resize and letterbox scores. Unless noted, decoding is DARK at 128^2 heatmap resolution, and leaderboard numbers are validation-phase MRE.

Heatmap targets are Gaussians with $\sigma = 3.6$ cells. Training uses task-balanced batches, AdamW, base learning rate 10^{-4}, cosine decay to 10^{-6}, encoder lr $0.2\times$, head lr $10\times$, 40 epochs, and EMA 0.999 for domain-generalized runs. the Adaptive Wing loss is used where reported in the ablation.

At inference the pipeline is deterministic: images are resized or letterboxed with recorded scale and pad, the heatmap is sigmoided and smoothed with an 11×11 Gaussian kernel before DARK refinement, the refined offset is accepted only when the Hessian is invertible and the displacement is below one heatmap cell, and coordinates are mapped back through the inverse transform before the clinical parameters are recomputed.