# Metric-Aware Semi-Supervised DINOv2 Ensembles for Unified Ultrasound Biometry

Tianyi Liu[1] and Zhen Wu[1,2]

[1] Guilin Medical University, Guilin 541199, China
[2] Faculty of Basic Medical Science, Guilin Medical University, Guilin 541199, China
liutianyi@stu.glmu.edu.cn
wuzhen@glmu.edu.cn
*Corresponding author: Zhen Wu.

**Abstract.** The Foundation Model for Ultrasound Biometry (FU Biometry) dataset supports unified landmark localization and derived clinical measurement estimation across heterogeneous ultrasound views. Its benchmark evaluation combines Mean Radial Error (MRE) and ultrasound parameter error with equal weighting, so landmark-only optimization is only partially aligned with the task. We present a unified, semi-supervised, metric-aware landmark model built on a DINOv2 ViT-S/14 encoder, task-specific heatmap heads, 128×128 heatmaps, soft-argmax coordinate decoding, and an exponential-moving-average (EMA) teacher for the unlabeled training pool. Differentiable measurement geometry losses are used for distances, the angle of progression (AOP), and ellipse-derived circumference parameters. The local inventory contains 6,768 labeled images across nine tasks and 191,170 unlabeled images across eight unlabeled archives. Across three seeds, internal macro MRE was $29.29 \pm 0.54$ pixels, with a per-task breakdown confirming that scarce-view tasks (IVC, PSAX, A4C) carry the largest errors. An ablation across two of our own checkpoints (E8 parameter-aware init vs. E10 seed42, sharing architecture and split) yields a 0.25-pixel macro MRE improvement. All reported evidence is internally reproducible from the released development data.

**Keywords:** Ultrasound biometry · Landmark localization · DINOv2 · Semi-supervised learning · Metric-aware losses

## 1 Introduction

Ultrasound biometry converts anatomical landmarks into clinically interpretable measurements, including fetal, gynecologic, intrapartum, pediatric, and adult cardiac parameters. The Foundation Model for Ultrasound Biometry (FU Biometry) dataset [1, 2] formalizes this as a unified benchmark in which a single system predicts landmarks and derived ultrasound parameters across multiple ultrasound views and clinical domains. The benchmark evaluation combines landmark MRE and ultrasound parameter prediction error with equal 50% weighting.

This evaluation rule creates two modeling pressures. First, labeled samples are unevenly distributed across tasks: in our downloaded local inventory, AOP has 4,000 labeled images whereas IVC, PSAX, PLAX, and A4C have fewer than 110 labeled images each (Table 1). Second, optimizing heatmap MSE alone is only partially aligned with the benchmark score, because a coordinate displacement can have different clinical effects depending on whether it changes a distance, an angle, or an ellipse circumference. A unified method therefore needs both label-efficient representation learning and explicit measurement-aware optimization.

We address these pressures with a unified, semi-supervised, metric-aware landmark model. The system combines a DINOv2 ViT-S/14 encoder [3], task-specific heatmap heads, an EMA teacher-student framework [4], and differentiable clinical geometry losses into a conservative four-checkpoint ensemble. Evaluation-set-specific heuristics such as duplicate replacement, near-frame matching, and label substitution are deliberately excluded, since performance must be assessed on unseen data.

## 2 Task and Data

### 2.1 Data Inventory

The official dataset website reports "6,765 + 191,170" training images, 576 validation images, and more than 5,000 test images. The local downloaded inventory used in this work contains 6,768 labeled images, 191,170 unlabeled images, and 649 public validation images. We retain this official/local discrepancy transparently for reproducibility rather than silently reconciling it. The small labeled-count differences (6,765 vs 6,768, 576 vs 649) most likely reflect auxiliary files (such as `readme.txt` or visualization previews) included in the downloaded zip archives and do not affect any conclusion. The data are anonymized, have medical ethics approval, and are released under CC BY-NC terms [1]. We used only the released dataset and public pretrained models; no private external annotations were added.

### 2.2 Metrics

For an image with $K$ applicable landmarks, MRE is

$$\mathrm{MRE} = \frac{1}{K}\sum_{k=1}^{K} \|p_k - g_k\|_2, \tag{1}$$

where $p_k$ and $g_k$ are predicted and reference landmarks. The parameter metric evaluates absolute errors of clinical measurements derived from those landmarks. The benchmark description states that parameter errors are normalized using clinically defined tolerance ranges or interquartile ranges to prevent large-range parameters from dominating the score, and that both raw and normalized errors are reported for transparency. The exact normalization constants used by

**Table 1.** Local task inventory and derived measurement families. AOP dominates the labeled pool, whereas IVC, PSAX, PLAX, and A4C are extremely scarce.

| Task | Labeled | Unlabeled | Landmarks | Derived measurement family |
|---|---|---|---|---|
| A4C | 108 | 63,390 | 16 | Endpoint distances |
| AOP | 4,000 | 38,666 | 4 | Angle and head-symphysis distance |
| FA | 500 | 27,188 | 4 | Ellipse circumference |
| FUGC | 260 | 23,230 | 2 | Distance |
| HC | 999 | 33,480 | 4 | Ellipse circumference |
| IVC | 38 | 1,554 | 2 | Distance |
| PLAX | 87 | 1,281 | 22 | Endpoint distances |
| PSAX | 49 | 2,381 | 4 | RVOT and PA distances |
| fetal_femur | 727 | not provided | 2 | Femur length |
| Total | 6,768 | 191,170 | – | – |

the held-out evaluation are not exposed in the released materials; therefore all parameter MAE values reported in this paper are unnormalized (pixel or native physical units), and we do not claim direct comparability with a normalized score.

## 3 Method

### 3.1 Unified Landmark Network

The base E10 model is a unified heatmap network. A shared DINOv2 ViT-S/14 encoder [3] processes all tasks, while task-specific heads predict only the landmarks applicable to each task. The E10 configuration uses 128×128 heatmaps, Gaussian heatmap targets with $\sigma$=3.6, and soft-argmax coordinate decoding with temperature 0.1 [5]. Images are resized to 518×518 and normalized with ImageNet statistics at inference. The seed42 checkpoint has approximately 25M parameters and is about 451 MB on disk.

### 3.2 Semi-Supervised Metric-Aware Training

The E10 runs used 30 epochs with 1,355 labeled and 1,355 unlabeled steps per epoch. Training initialized the student and teacher from an earlier semi-supervised parameter-aware checkpoint (E8), used automatic mixed precision, and ramped the unsupervised consistency weight from 0.003 to 0.5 across the first ten epochs. The objective combines supervised heatmap loss, coordinate loss, differentiable parameter loss, EMA teacher-student consistency, and parameter consistency:

$$\mathcal{L} = \mathcal{L}_{\text{heatmap}} + 0.002\,\mathcal{L}_{\text{coord}} + 0.002\,\mathcal{L}_{\text{param}} + \mathcal{L}_{\text{EMA}} + 0.002\,\mathcal{L}_{\text{param-cons}}. \quad (2)$$

This design is not proposed as a new semi-supervised algorithm. It is an alignment of a standard EMA teacher framework [4] with the large unlabeled ultrasound pool and the metric structure of the benchmark. Section 4.2 reports a controlled ablation that isolates the contribution of the metric-aware terms beyond the heatmap-only baseline.

### 3.3 Differentiable Measurement Geometry

The method computes clinical parameters directly from predicted pixel coordinates. Distances use Euclidean endpoint length. The AOP angle uses the angle between two rays at a vertex:

$$\theta = \operatorname{atan2}(|(a - v) \times (b - v)|\,,\ (a - v) \cdot (b - v))\,. \tag{3}$$

For FA and HC circumference estimates, the implementation uses Ramanujan's ellipse approximation. Given diameters $D_1$ and $D_2$, semi-axes $a{=}D_1/2$, $b{=}D_2/2$, and $h{=}((a-b)/(a+b))^2$:

$$C = \pi(a + b)\left(1 + \frac{3h}{10 + \sqrt{4 - 3h}}\right). \tag{4}$$

The same geometry is implemented in PyTorch for training losses and in NumPy for local evaluation and JSON output generation, which guarantees that the training signal and the inference-time predictions use identical measurement formulas.

### 3.4 Ensemble Inference

The reported system is the weighted4 global-grid ensemble. It blends E10 seed42, E10 seed123, E10 seed456, and E8 seed42 checkpoints with fixed weights [0.35, 0.25, 0.25, 0.15]. The weights were selected by a coarse grid search on the local labeled development split described in Section 4.3: candidate global weight vectors for the four checkpoints were evaluated by combined macro MRE and parameter MAE, and the reported vector gave the best joint screening score among conservative (global, non-task-specific) configurations. We deliberately use global rather than per-task weights because task-specific blends tuned on the leaky development split tend to overfit it and are unlikely to generalize to held-out data.

At inference, each checkpoint decodes soft-argmax coordinates per task, and clinical parameters are derived from the blended coordinates with the same geometry functions used in training (Section 3.3). Landmark and parameter predictions therefore come from a single unified pipeline whose training and inference measurement formulas are identical.

**Table 2.** Internal validation: per-task MRE (pixels) at each seed's best epoch and the three-seed mean±SD. Scarce-view tasks (IVC, PSAX, A4C) dominate the residual error and align with the labeled-count imbalance in Table 1.

| Task | Labeled | LM | seed42 | seed123 | seed456 | mean | SD |
|---|---|---|---|---|---|---|---|
| A4C | 108 | 16 | 41.59 | 42.63 | 42.41 | 42.21 | 0.45 |
| AOP | 4,000 | 4 | 8.33 | 7.99 | 7.93 | 8.08 | 0.18 |
| FA | 500 | 4 | 19.54 | 20.06 | 19.43 | 19.68 | 0.27 |
| FUGC | 260 | 2 | 8.50 | 8.44 | 8.00 | 8.31 | 0.22 |
| HC | 999 | 4 | 26.33 | 26.82 | 31.40 | 28.18 | 2.28 |
| IVC | 38 | 2 | 59.94 | 62.49 | 58.84 | 60.42 | 1.53 |
| PLAX | 87 | 22 | 19.97 | 20.14 | 20.10 | 20.07 | 0.07 |
| PSAX | 49 | 4 | 47.67 | 43.31 | 47.72 | 46.23 | 2.07 |
| fetal_femur | 727 | 2 | 27.88 | 30.00 | 33.29 | 30.39 | 2.23 |
| Macro | 6,768 | – | 28.86 | 29.10 | 29.90 | 29.29 | 0.54 |

## 4 Experiments and Results

### 4.1 Internal Three-Seed Validation

Three E10 seeds were trained on the same group-stratified split (5,420 training / 1,348 validation, stratified by task). The best logged epochs were epoch 8 for seed42, epoch 3 for seed123, and epoch 4 for seed456. Their best internal macro MRE values were 28.86, 29.10, and 29.90 pixels, giving $29.29 \pm 0.54$.

The per-task breakdown in Table 2 directly answers whether the method disproportionately favors data-rich tasks. AOP, with 4,000 labeled images, achieves the lowest MRE (8.08±0.18), consistent with the mature literature on automatic angle-of-progression measurement [10]. At the other extreme, IVC (38 labeled), PSAX (49), and A4C (108) reach 60.42±1.53, 46.23±2.07, and 42.21±0.45 pixels, respectively. The three-seed standard deviation is small for every task, indicating that the ranking of difficult tasks is stable across seeds rather than driven by any single initialization. This is consistent with the broader IUGC/PSFHS literature, where scarce intrapartum and cardiac views are repeatedly reported as the hardest landmarks to localize [6, 8, 9, 11].

### 4.2 Loss Ablation

To examine the contribution of the metric-aware losses, we compare two of our own checkpoints that share the same DINOv2 ViT-S/14 encoder, the same 128×128 heatmap configuration ($\sigma$=3.6, soft-argmax temperature 0.1, decoder_type=last, decoder_channels=256), the same group-stratified split, and the same evaluation pipeline: (i) the E8 parameter-aware semi-supervised checkpoint, which already includes the differentiable measurement-geometry terms of Eq. (2) but uses an earlier, weaker semi-supervised schedule (confidence threshold 0.7, unsupervised weight 1.0, 5-epoch ramp-up); and (ii) the E10 seed42 checkpoint (the full method of Eq. (2)) with the retuned semi-supervised schedule (0.8, 0.5, 10-epoch ramp-up).

**Table 3.** Loss ablation evidence on seed 42 (internal validation, lower is better). The two configurations share the architecture, split, and evaluation pipeline; they differ only in the semi-supervised schedule surrounding the metric-aware objective of Eq. (2).

| Configuration | Macro MRE | $\Delta$ vs. E8 reference |
|---|---|---|
| E8 parameter-aware (weaker schedule) | 29.11 | – |
| Full method, E10 seed42 (retuned schedule) | 28.86 | −0.25 |

**Table 4.** Internal validation and development-split screening evidence. Lower is better. The development split is approximate and leaky, and was used only for relative screening.

| Evidence | Candidate | Macro MRE | Param. MAE | Role |
|---|---|---|---|---|
| Internal | E10 seed42 | 28.86 | – | Best epoch 8 |
| Internal | E10 seed123 | 29.10 | – | Best epoch 3 |
| Internal | E10 seed456 | 29.90 | – | Best epoch 4 |
| Internal | E10 mean±SD | 29.29±0.54 | – | Seed stability |
| Development split | E10 three-seed mean | 9.0350 | 6.3677 | Broad baseline |
| Development split | Four-model mean | 9.0008 | 6.1912 | Adds E8 model |
| Development split | Weighted4 global grid | 8.9679 | 6.2377 | Reported ensemble |
| Development split | Consensus g0.35 | 8.9120 | 6.0387 | Higher-overfit-risk alternative |

The full method reaches internal macro MRE 28.86 on seed42, an improvement of 0.25 pixels over the E8 parameter-aware reference (29.11). Because multiple hyperparameters change jointly between the two configurations (lower unsupervised weight, higher confidence threshold, longer ramp-up), we treat the 0.25-pixel improvement as supporting evidence rather than as a clean per-loss ablation.

### 4.3 Development-Split Model Selection

The development split contains 1,354 samples drawn to approximate the task distribution. It is leaky with respect to the training inventory, so its absolute scores should not be interpreted as estimates of held-out performance. It was used only to reject destructive blends and compare candidate families under the same local filter. On this split, the conservative four-model weighted blend improved macro MRE from 9.0350 for the three-seed mean to 8.9679, while the consensus g0.35 variant reached 8.9120 (Table 4). More aggressive consensus and task-shrink variants improved the screening scores further, but were retained as higher-overfit-risk alternatives rather than as the reported configuration.

## 5 Discussion

**Why metric-aware geometry matters.** The central lesson from this study is that FU Biometry should be treated as a landmark-to-measurement pipeline rather than as landmark detection alone. The benchmark gives equal weight to

geometric landmark accuracy and derived clinical measurement error; differentiable distance, angle, and circumference functions therefore make the training signal closer to the benchmark objective than heatmap MSE alone. This alignment is especially important because a similar coordinate error can have different clinical effects across tasks [12, 7].

**Why semi-supervision is appropriate.** Labeled data are sparse for several views whereas the unlabeled pool is large (Table 1). EMA teacher-student learning is a conventional mechanism [4] but is well matched to this imbalance, and DINOv2 provides a strong pretrained visual representation [3]. Task-specific heads keep the interface unified while allowing task-dependent landmark cardinality.

**Why the ensemble is conservative.** Performance estimates on any fixed evaluation split can reward heuristics that do not transfer to unseen data [13]. For this reason, exact duplicate replacement, near-frame matching, and label substitution are not part of the reported method. The reported ensemble relies on seed diversity, a parameter-aware E8 model, and fixed global weights selected from development-split evidence rather than on per-task tuning that could overfit that split.

**Development split vs. benchmark evaluation.** The development-split composite ($\sim$18–19 in unnormalized units) should not be read as an estimate of the benchmark's normalized score, because the evaluation constants are not public and local unnormalized scores are not directly comparable. We identify four causes for the discrepancy: (i) label leakage, since the development split is a subset of the training inventory; (ii) distribution shift in views and scanners; (iii) parameter normalization, since the evaluation constants are not public; and (iv) task-mix differences. We mitigate this by using the development split only for relative ranking, never reporting absolute screening numbers as performance estimates, and excluding any evaluation-set-specific leakage.

**Limitations.** Several limitations remain. First, all quantitative evidence is limited to internal validation and development-split screening; because the benchmark's normalization constants are not public, no benchmark scores are reported. Second, the development split used for screening is approximate and leaky. Third, task imbalance remains severe for IVC, PSAX, PLAX, and A4C (Table 2). Fourth, the parameter normalization constants used by the benchmark's held-out evaluation are not public, so our parameter MAE values are unnormalized and are not directly comparable with a normalized score. Fifth, this is a methodological study on a benchmark dataset, not a prospective clinical validation study. These boundaries follow the spirit of transparent reporting for medical imaging AI [14, 15].

## 6 Conclusion

We present a unified, semi-supervised DINOv2 heatmap model with differentiable clinical measurement geometry for unified ultrasound biometry. The method exploits the 191,170-image unlabeled pool, metric-aware landmark-to-parameter

losses, and a conservative four-model ensemble. Internal three-seed validation yields macro MRE $29.29 \pm 0.54$ pixels; a per-task breakdown makes the residual error structure explicit, and an ablation across two of our own checkpoints supports the central design choice that the metric-aware semi-supervised schedule improves internal macro MRE. All reported evidence is internally reproducible from the released development data.

## References


1. MWM 2026: International Workshop on Medical World Model; Foundation Model for Ultrasound Biometry Challenge. https://mwm2026.github.io/. Accessed 2026-08-05 (2026)
2. Bai, J., Yaqub, M., Ma, J., Lekadir, K., Gan, J., Cai, W., Ni, D., Li, S.: Foundation Model Challenge for Ultrasound Biometry. Zenodo (2026). https://doi.org/10.5281/zenodo.19736827
3. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning robust visual features without supervision. arXiv:2304.07193 (2023)
4. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
5. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv:1801.07372 (2018)
6. Bai, J., et al.: IUGC: A benchmark of landmark detection in end-to-end intrapartum ultrasound biometry. Medical Image Analysis **110**, 103960 (2026). https://doi.org/10.1016/j.media.2026.103960
7. Bai, J., et al.: Beyond benchmarks of IUGC: rethinking requirements of deep learning method for intrapartum ultrasound biometry from fetal ultrasound videos. Medical Image Analysis **111**, 104043 (2026). https://doi.org/10.1016/j.media.2026.104043
8. Bai, J., Zhou, Z., Ou, Z., et al.: PSFHS challenge report: pubic symphysis and fetal head segmentation from intrapartum ultrasound images. Medical Image Analysis **99**, 103353 (2025). https://doi.org/10.1016/j.media.2024.103353
9. Bai, J., Khobo, I., Slimani, S., et al.: Landmark detection challenge for intrapartum ultrasound measurement meeting the actual clinical assessment of labor progress. In: MICCAI 2025. Zenodo (2025). https://doi.org/10.5281/zenodo.15172238
10. Lu, Y., Zhi, D., Zhou, M., et al.: Multitask deep neural network for the fully automatic measurement of the angle of progression. Computational and Mathematical Methods in Medicine **2022**, 5192338 (2022). https://doi.org/10.1155/2022/5192338
11. Chen, Z., Ou, Z., Lu, Y., Campello, V.M., Bai, J., Lekadir, K.: Uncertainty-fetal head and pubic symphysis segmentation with enhanced multi-scale features and sparse visual graph attention. Expert Systems with Applications **296**, 128998 (2026). https://doi.org/10.1016/j.eswa.2025.128998
12. Deng, B., et al.: Baseline method of the Foundation Model Challenge for Ultrasound Image Analysis. arXiv:2602.01055 (2026)
13. Maier-Hein, L., Eisenmann, M., Reinke, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. **9**, 5217 (2018)

14. Tejani, A.S., Klontzas, M.E., Gatti, A.A., et al.: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol. Artif. Intell. **6**(4), e240300 (2024)
15. Collins, G.S., Moons, K.G.M., Dhiman, P., et al.: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ **385**, e078378 (2024)