Appendix for SonoDiNoV2: Geometry-Aware, Pseudo-Label Guided Multi-Task Ultrasound Model

Khurram Afroz, Aarush Pandey, and Monika Agrawal

Indian Institute of Technology Delhi

This appendix accompanies the main paper submitted to the MICCAI 2026 Foundation Model for Ultrasound Biometry challenge (FoundUS). Sections are lettered A–H to match citations in the main text (e.g., "Appendix A," "Appendix C," "Appendix G"); cross-references (e.g., "main §4.2," "main Table 1") use that document's numbering. The controlled experimental study is centered on **Stage 1**, the configuration submitted to the official test server, which achieves **24.42 px MRE** as the **Stage 1-family ensemble**. **Stage 2** was developed subsequently as a challenge extension and reached a later best score of **23.43 px MRE**; its broader evaluation lies outside the controlled Stage 1 study and is left for future work.
Both checkpoint families share the same 448×448-input, 32×32-token DINOv2 ViT-S/14 encoder; pseudo-label checkpoints widen the decoding heads to a 192×192 heatmap grid, giving 45.63M total stored parameters. All other configurations were development candidates, selected via leakage-free cleanval and, where reported, evaluated once on the test server for transparent reporting.

A Architecture and Parameter Accounting

A.1 Two checkpoint families

All checkpoints share DINOv2 ViT-S/14 with 448×448 input and a 32×32 grid of 384-dimensional tokens (main §2), differing only in decoding-head design:

- **Default family** (Stage 1 and its ablations; some Stage-2 candidates). Two 3×3 convolutional layers with bilinear upsampling per head, producing a $128\times128\times K_t$ heatmap (G=128). **29.4M** parameters total.
- **Pseudo-label family** (Stage 2, Attempt 2, Appendix F). Each head adds one 3×3 convolution and wider channels, producing a $192\times192\times K_t$ heatmap (G=192). **45.63M** parameters total, all nine widened heads included.

Bilinear upsampling is parameter-free, so the increase to 45.63M comes entirely from added convolutional capacity and wider channels supporting the finer grid; the shared encoder is identical across families, and the difference lies entirely in the heads.

Table 1. Parameter accounting for the two checkpoint families.

Component	Default family	Pseudo-label family
Shared encoder (DINOv2 ViT-S/14)	identical, 32×32 tokens	
Head design (per task)	2 conv layers	3 conv layers, wider
Heatmap grid G	128×128	192×192
Total stored checkpoint	**29.4M**	**45.63M**

B Gradient-Level Analysis of Sampling-Induced Interference

B.1 Motivation and methodology

Main §4.2 attributes the degradation under square-root sampling (τ=0.5) to gradient-magnitude dominance rather than directional conflict. We distinguish *directional conflict* (task gradients pointing in opposing directions) from *magnitude dominance* (one gradient substantially outweighing the rest regardless of direction): the two point to different fixes—gradient surgery for the former, gradient normalization or tighter sampling control for the latter—so separating them matters.

At three checkpoints (early: epoch 1; mid: epoch 15; late: best validation checkpoint) under uniform (τ=0) and square-root (τ=0.5) sampling, we draw one 8-image mini-batch per task and compute each task's gradient with respect to the shared encoder only (private heads are excluded), averaged over 8 draws per checkpoint/scheme. From the nine resulting encoder-gradient vectors we compute the full 9×9 pairwise cosine matrix and its mean off-diagonal value. For AOP—the task whose sampling share changes most (11.1% → 32.7%)—we additionally report its mean cosine with the other eight tasks, its fraction of negative pairs, and its gradient norm ratio $\|g_{\text{AOP}}\| / \text{mean}_{j \neq \text{AOP}}(\|g_j\|)$.

B.2 Global gradient alignment

Table 2 shows mean off-diagonal cosine similarity declining under both schemes (uniform 0.47 → 0.14; sqrt 0.43 → 0.14)—that is, gradients become more orthogonal over training, which is not by itself evidence of conflict. The scheme-to-scheme gap is small (at most 0.042) and reverses sign by the end of training, so global alignment alone cannot explain the 18.6 px degradation; the task-specific view below isolates the dominant effect.

Table 2. Global gradient alignment across the nine Stage-1 tasks.

Stage	Uniform (τ=0)	Square-root (τ=0.5)
Start (epoch 1)	0.471	0.429
Mid (epoch 15)	0.288	0.274
End (best checkpoint)	0.139	0.144

B.3 Task-specific alignment and magnitude

Table 3 isolates AOP, the task most affected by square-root sampling. Under uniform sampling, its gradient norm remains near or below the average of the other tasks, whereas square-root sampling amplifies it to 5.27×, 3.77×, and 2.89× at the three checkpoints. AOP remains positively aligned on average, indicating that the main effect is gradient-magnitude dominance rather than strong directional conflict.

Table 3. AOP-versus-rest gradient statistics on the shared Stage-1 encoder.

Stage	Scheme	Cosine (AOP vs. rest)	Neg. pairs (/8)	Norm ratio
Start	Uniform	0.523	0.8	1.42×
Start	Square-root	0.555	0.2	**5.27**×
Mid	Uniform	0.320	0.4	0.74×
Mid	Square-root	0.320	0.6	**3.77**×
End	Uniform	0.212	1.1	0.73×
End	Square-root	0.138	2.2	**2.89**×

Directional conflict is weak. AOP stays positively correlated with every other task at every checkpoint under both schemes; negative pairs remain a minority, peaking at 2.2/8 under sqrt. Strong angular opposition is therefore unlikely to be the primary mechanism.

Magnitude imbalance is persistent. Under uniform sampling, AOP's norm falls below parity by mid-training (0.74×, 0.73× at convergence). Under square-root sampling, its 32.7% share keeps the norm at 3–5× the mean of the other tasks throughout training. Oversampling AOP therefore sustains a disproportionately large encoder update that outweighs the other eight tasks' combined contribution.

B.4 Interpretation and caveats

We attribute the square-root degradation (main Table 4, §4.2) primarily to sustained magnitude imbalance, with directional conflict a smaller, late-training effect. Two caveats: (i) 8-draw averages leave early-checkpoint ratios subject to batch noise, though the trend is consistent across checkpoints; (ii) this analysis isolates AOP specifically, while smaller pairwise effects among the other eight tasks are captured only by the coarser global statistic in Table 2.

C Full Training and Optimization Settings

Main §2 states the core recipe; this section lists complete settings for reproducibility. Unless stated otherwise, all controlled experiments, ablations, and results in this paper use the accepted **Stage-1** configuration and protocol (24.42 px MRE).

- **Optimizer.** AdamW, base LR 10^{-4}, cosine-annealed to 10^{-6}; Stage 1 uses 30 epochs. Stage 2 uses longer, configuration-dependent schedules and is extension work, not part of the Stage-1 evidence. Encoder-LR multiplier λ_{enc}: default 0.2 (all Stage-1/default-family checkpoints); swept to 0.05 and 0.00 (frozen) in the main §4.3 ablation and retained as an ensemble candidate.
- **Batch composition.** Task sampled uniformly, p_t=1/9; 8 images per sampled task per step. Pseudo-label-family checkpoints additionally draw accepted pseudo-labeled frames for the sampled task at the same batch size; these belong to the Stage-2 extension and are not used to define the Stage-1 experiments.
- **Weight averaging.** EMA of model weights for checkpoint selection, decay 0.999.
- **Loss weights.** $\mathcal{L}_t = \mathcal{L}_{\text{hm}} + \lambda_{\text{graph}}\mathcal{L}_{\text{graph}} + \lambda_{\text{coord}}\mathcal{L}_{\text{coord}}$ (main §2). $\mathcal{L}_{\text{hm}}$: focal-weighted BCE, γ=2, positive weight 5. $\mathcal{L}_{\text{coord}}$: soft-argmax coordinate term (Smooth-L1-style); soft-argmax decoding was statistically indistinguishable from hard-argmax on cleanval, so hard-argmax is used for all reported inference. $\mathcal{L}_{\text{graph}}$: Huber loss ($\delta$=0.1) on scale-normalized pairwise landmark-graph distances. **Both auxiliary terms were evaluated and found unnecessary given the base heatmap objective; every submitted Stage-1 checkpoint uses $\mathcal{L}_{\text{hm}}$ alone** (λ_{graph}=λ_{coord}=0).
- **Augmentation.** Geometric: horizontal flip (p=0.5), translation up to 7%, isotropic scaling $[0.8, 1.2]$, rotation up to $\pm 30°$, applied jointly to image and landmarks before letterboxing. Photometric (deliberately aggressive): brightness/contrast jitter, gamma, CLAHE, global histogram equalization, unsharp-mask sharpening, Gaussian blur, additive noise. An additional explicit intensity-normalization step on top of this stack gave no further test benefit (main §5.1, §6), consistent with limited residual photometric shift after augmentation.
- **Checkpointing.** Full optimizer, scheduler, and EMA state saved every epoch for exact resumption under preemptible compute.
- **Letterbox.** Isotropic scale s=448/ $\max(H_0, W_0)$, symmetric zero-padding to 448×448; verified to round-trip a synthetic landmark to ≤ 7.35 px at 1920 px native resolution.
- **Stage-2 extension.** Confidence-filtered pseudo-labeling and its associated pseudo-label-family configurations, yielding a later 23.43 px MRE, were developed after the accepted Stage-1 submission and are therefore not used in the Stage-1 controlled experiments, ablations, or conclusions. We document the configuration here for completeness; a systematic study of its benefits and limitations is left for future work.

C.1 Backbone adaptation is U-shaped

Sweeping the encoder learning-rate multiplier (Table 4) yields a U-shaped relationship on cleanval: a fully frozen backbone is worse (29.98 px), the default multiplier adapts too aggressively (26.56 px), and a mostly frozen backbone is best (23.96 px). The backbone must adapt to ultrasound, but by roughly four

times less than the default. This provides partial, in-distribution support for constrained (parameter-efficient) adaptation. Importantly, this best-on-validation setting did not transfer to the test set: as a standalone model it reaches only 28.81 px on test, the validation-to-test dissociation discussed in main §5 and revisited in Appendix H.

Table 4. Encoder learning-rate multiplier (cleanval MRE, px).

Multiplier	Encoder LR	Cleanval MRE
0.00 (frozen)	0	29.98
0.05	5×10^{-6}	23.96
0.20 (default)	2×10^{-5}	26.56

C.2 Decoding and test-time augmentation

On cleanval, soft-argmax (26.24 px) and hard-argmax (26.40 px) are within noise, confirming that localization is insensitive to the decoder; we use hard argmax. Test-time augmentation, however, is orientation dependent (Table 5): horizontal-flip averaging benefits laterally ambiguous fetal planes (AOP by 1.4, A4C by 1.2, FA by 0.8, PLAX by 0.75, and FUGC by 0.3 px) but harms fixed-orientation cardiac views (IVC by 4.8 and PSAX by 2.7 px), and orientation search benefits only AOP. The appropriate policy is therefore per-task augmentation rather than a global setting. We note that the IVC and PSAX validation sets are small (n=8 and n=10), so these magnitudes should be read as indicative.

Table 5. Decode and TTA on cleanval (MRE, px), with the associated mechanism.

Decode	TTA	Cleanval	Effect
soft-argmax	none	26.24	insensitive to decoder
hard-argmax	none	26.40	baseline
hard-argmax	+hflip	26.83	benefits planes, harms cardiac
hard-argmax	+orient	26.78	benefits AOP only

D Characterizing the Validation-to-Test Shift

D.1 Methodology

For every task, on cleanval and the CodaBench test set, we compute four statistics from image content alone, using no test labels: mean brightness and contrast of content pixels (same standardized space used to construct cleanval, main §3); native resolution $\sqrt{HW}$; and an interface-clutter proxy, counting small, high-contrast, sharp-edged contours in the outer 12% border band of each image, where scanner overlays and calipers are conventionally rendered. The proxy is a coarse density measure, not a text detector, but it is computed identically across splits and tasks, which is what the correlation analysis requires.

D.2 Full per-task statistics

Table 6 reports all four statistics for all nine tasks under the Stage-1 protocol; main §5.1 discusses the extremes (HC, AOP, PSAX, A4C, femur) in the main text for space. Stage 2 is not included; its interaction with this shift is left for future work.

Table 6. Full cleanval/test comparison, all nine tasks, ordered by validation-to-test gap (most positive to most negative; main §5).

	Brightness		Contrast		Resolution		Overlay	
Task (gap)	val	test	val	test	val	test	val	test
HC (+24.8)	55.4	51.3	39.7	40.1	657	846	21.4	130.0
AOP (+13.9)	49.1	58.2	40.7	48.6	512	512	28.6	81.2
PSAX (+13.9)	84.0	82.6	61.4	61.8	776	729	143.9	162.5
FUGC (+5.0)	65.1	65.6	43.2	49.0	428	428	62.7	56.3
FA (+2.3)	62.9	59.9	45.5	47.5	887	887	131.9	129.9
PLAX (+2.2)	87.5	98.1	62.4	65.6	790	701	138.7	166.5
IVC (−2.6)	78.6	82.5	61.0	65.6	720	736	151.5	154.4
femur (−6.3)	58.5	50.7	43.8	39.5	771	712	111.9	89.0
A4C (−9.5)	84.9	90.7	65.7	72.4	732	693	135.9	161.4

D.3 Correlation with the validation-to-test gap

Spearman rank correlation (n=9) between each statistic's absolute cleanval-to-test change and the per-task gap gives ρ=−0.22 for brightness/contrast (flat—consistent with intensity normalization giving no test benefit, main §6) versus ρ=+0.50 each for resolution and overlay clutter. These correlations are suggestive, not conclusive, at n=9: A4C is an explicit counterexample, showing increased clutter alongside a *negative* gap. On balance, the Stage-1 evidence favors framing and on-screen presentation over photometric appearance as the dominant shift axis, without excluding task-specific factors (e.g. HC's acoustic-shadow difficulty, main §8) that this image-level analysis cannot capture.

D.4 Extended qualitative analysis

Main Figure 2 shows one representative cleanval/test pair per task for the five largest-gap tasks. Beyond that single example: for HC, the tight test-side cropping with border-rendered scanner text recurs across the additional pairs we inspected, not an isolated case, and the acoustic-shadowed poles are pushed toward or past the boundary in most higher-gap test images reviewed. For AOP

and PSAX, the positive gap coincides with visibly different probe-angling conventions between splits across the additional pairs inspected, consistent with a framing shift rather than a labeling or resolution artifact. For the negative-gap tasks (A4C, femur), cleanval's construction-time bias toward appearance outliers recurs: several cleanval A4C frames use color-Doppler overlays or non-standard imaging modes essentially absent from test. This is corroborating, illustrative evidence, not an independently powered analysis.

E Backbone Compute Accounting

Main Table 4 reports encoder GMACs alongside parameter counts to separate representational capacity from token-grid resolution as candidate explanations for the backbone finding (main §4.1); these comparisons are part of the accepted Stage-1 protocol.

GMACs are computed for a single encoder forward pass (excluding task heads, which are lightweight and near-identical in cost across the compared backbones) at each backbone's stated input resolution, using standard per-layer multiply–accumulate accounting for transformer blocks as a function of token count and embedding dimension. The 1.4× compute gap between patch-14 at 448 px (31.4 GMACs, 1024 tokens) and patch-16 at the same resolution (22.3 GMACs, 784 tokens) arises directly from the token-count difference at matched embedding width—an unavoidable consequence of patch size at fixed resolution, not an independent factor. A FLOP-matched comparison (patch-14 at a reduced resolution matching patch-16's token count) would isolate patch size from token count and is left to future work.

This accounting covers only the backbone-comparison checkpoints of main Table 4 (all G=128, default head design); it excludes the pseudo-label family's widened-head compute, which we did not separately profile since the heads are a small fraction of total compute relative to the shared 32×32-token encoder. A dedicated Stage-2 compute analysis is left for future work.

F Pseudo-Label Attempt History and Downstream Measurement Breakdown

F.1 Two attempts, and the subsequent per-task selection

Main §3.1 and Table 1 reference two pseudo-labeling attempts and a subsequent per-task selection step; we lay them out in full here, since the compressed main text states outcomes without the full configuration of each. The experiments below were developed *after* the accepted Stage-1 submission and are documented as extension work, not as evidence for the Stage-1 conclusions.

– **Attempt 1 (rejected).** Unfiltered pseudo-labels applied uniformly across all nine tasks, using the default-family architecture (29.4M, G=128) and default encoder-LR multiplier (0.2). This checkpoint obtained a test MRE

of 31.52 px, worse than the accepted Stage-1 result of 24.42 px (main Table 5). The configuration was rejected on cleanval evidence *before* this test number was obtained (Appendix H).

– **Attempt 2 (subsequent extension).** Confidence-thresholded pseudo-labels (τ=0.85, swept on cleanval) combined with the widened pseudo-label-family architecture (45.63M, G=192), default encoder-LR multiplier (0.2). Evaluated per task on cleanval against the Stage-1 family, this configuration improved only AOP, PLAX, and FUGC. As an internal development check, deploying it alone as a binary per-task-or-not ensemble reached 27.40 px aggregate test MRE (not submitted; main Table 1). A subsequent full per-task n-way search over all trained configurations selected this confidence-filtered candidate for AOP, PLAX, and FUGC specifically, yielding 16.11 px (AOP), 16.32 px (PLAX), and 10.16 px (FUGC).

These results represent the Stage-2 development path, not the accepted Stage-1 submission; the 27.40 px figure and the per-task results above are not used in the Stage-1 ablations or controlled analyses elsewhere in this paper. For HC, the same n-way search separately selected the best-validated *default-family* candidate from the expanded pool—the "Encoder LR 0.05 single model (29.4M)" row of main Table 1—reducing HC from 44.11 px to 40.85 px (main Table 3, §8); this improvement likewise belongs to the Stage-2 development. Every controlled experiment in this paper remains based on the Stage-1 protocol and its 24.42 px MRE.

F.2 Full downstream measurement breakdown

Main Table 3 reports landmark-level MRE for the later per-task ensemble. Table 7 additionally reports, per task, the mean absolute error (MAE) of the clinically relevant *downstream* measurement derived from those landmarks (e.g. head-circumference length from the predicted HC ellipse, or each cardiac chamber's extent from the predicted A4C landmarks), rather than the raw per-landmark radial error. These quantities are not directly comparable—MRE is a per-landmark Euclidean distance, while each downstream MAE is a task-specific derived quantity that can amplify or partially cancel individual landmark errors—but both matter to a reviewer or clinician assessing fitness for purpose: HC's derived circumference ($\mathrm{HC}_{\mathrm{mae}}$=113.36 px) is substantially noisier than its underlying BPD landmark distance ($\mathrm{BPD}_{\mathrm{mae}}$=20.93 px), consistent with the acoustic-shadow difficulty (main §8) affecting the ellipse fit more than the two pole landmarks individually.

These downstream measurements document the Stage-2 development; a systematic analysis of Stage-2 pseudo-labeling, per-task selection, and downstream measurement behavior is reserved for future work.

Table 7. Final ensemble: downstream measurement MAE (px, except AOP angle in degrees). Each entry pairs a measurement with its MAE; N is the test sample count for that task (of 619).

Task	N	Measurement (MAE)
HC	215	BPD (18.18), HC (133.64)
FA	188	FA (88.80)
fetal_femur	62	FL (18.71)
AOP	60	AOP (11.23°), HSD (11.82)
PLAX	26	AAO (12.94), AV (8.69), IVS (4.29), LA (6.25), LV (16.43), LVOT (7.14), LVPW (7.74), RV (8.53), RVOT (7.79), STJ (8.56), VAO (9.70)
A4C	20	LA-h (11.17), LA-v (14.14), LV-h (15.08), LV-v (13.41), RA-h (17.38), RA-v (23.87), RV-h (16.98), RV-v (21.21)
FUGC	20	CL (13.04)
PSAX	18	PA (27.00), RVOT (17.93)
IVC	10	IVC (8.33)

G Full Supporting Evidence for Practical Recommendations

Main §7 states six recommendations in prose for space. Table 8 gives the deployed setting and supporting evidence behind each.

Table 8. Recommendations, deployed setting, and supporting evidence.

Design choice	Recommended default	Deployed	Evidence
Backbone patch size	Patch-14 over patch-16 at matched resolution/budget	DINOv2 ViT-S/14, 32×32 grid	Main Table 4, §4.1: grid density outweighs pretraining domain and parameter count
Dataset sampling	Uniform across tasks when the metric is task-uniform; avoid τ=0.5	Uniform (τ=0)	Main §4.2, App. B: sqrt sampling causes sustained gradient-magnitude dominance from the most-oversampled task
Adaptation strength	Do not deploy the validation-optimal encoder LR standalone; keep it as an ensemble candidate	LR mult. 0.2; 0.05 retained as a per-task candidate	Main §4.3, §5: 0.05 wins the cleanval aggregate (23.96) but loses on test standalone (28.81)
Test-time augmentation	Apply per task by orientation semantics; require a CI excluding zero on low-n tasks	Flip on AOP, FA, PLAX only	Main §4.4: only these three have bootstrap CIs excluding zero
Pseudo-labeling & decoder capacity	Treat confidence filtering and widened heads as extension work; validate per task	Stage 1: not deployed. Stage 2: τ=0.85, 45.63M widened heads, AOP/PLAX/FUGC only	Main §3.1, §6, App. A and F: gains confined to selected tasks; systematic evaluation deferred
Ensembling	With no Pareto-optimal configuration, validate each task-level checkpoint choice individually	Final Stage-1 per-task selection	Main §4.5, §5: task-level selection gives the accepted 24.42 px MRE

H Methodological Note: Test-Server Evaluation vs. Official Deployment Decisions

Main §3.1 states that deployment decisions for the accepted Stage-1 submission were fixed on cleanval before test-server evaluation, and that test scores were

never used for model selection. Main Table 5 nevertheless reports test-server MRE for rejected or intermediate ideas, including Attempt-1 pseudo-labeling (31.52 px). These two facts are consistent: the test server was used for evaluation and reporting, never for selecting among candidates.

Deployment versus reporting. The accepted **Stage-1** configuration was selected entirely on cleanval, with task-level validation where applicable, then submitted to the official test server (**24.42 px**); candidates were never ranked by test score. For rejected or superseded approaches, we sometimes submitted a checkpoint once *after* its role had been settled on cleanval, so the paper could report actual test performance rather than only a qualitative rejection—confirmatory results, not inputs to model selection.

Worked example. Attempt 1 used unfiltered pseudo-labels across all nine tasks and was rejected on cleanval because it gave no aggregate benefit over Stage 1. The resulting checkpoint was then submitted once and obtained 31.52 px (main Table 5): the rejection predates that number, so it did not drive the decision. The same protocol applies to every other rejected, "ablation only," or "not submitted" row in the main paper's tables.

Why this matters. Searching over test performance across many candidates risks a multiple-comparisons problem: repeated evaluation against the same test set can favor a configuration by chance, inflating the apparent generalization estimate. Reserving the test server for confirmation avoids this. Main §5 gives a related example of why validation should not rest on a single aggregate score: the encoder-LR-0.05 configuration reaches 23.96 px on cleanval but only 28.81 px on test as a standalone model (Appendix C.1)—evidence for task-level validation and selection over choosing one configuration from an aggregate score.

Stage 1 and Stage 2. The Stage-1 family ensemble (main Table 2) reaches **24.42 px** and is the accepted submission; every controlled experiment here is centered on that protocol. **Stage 2** (confidence-filtered pseudo-labeling, widened heads, task-level selection) was built afterwards and is reported as an extension at **23.43 px**, not as evidence validating Stage 1.