Accurate but Not Equally Trustworthy: Benchmarking Foundation Models for Fetal Ultrasound in Africa

Alice Bagyiereyele Lakyiere(✉), Richard Adusei, Kwabena Owusu-Agyemang, Rose-Mary Owusuaa Mensah Gyening

Department of Computer Science, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana
lakyierealice@gmail.com

Abstract. Fetal ultrasound AI benchmarks for Africa remain scarce, and the few that exist report accuracy alone, leaving calibration and explanation faithfulness – the properties clinicians need to trust a model – almost entirely unmeasured. We present a fully reproducible, CPU-only benchmark of four frozen, publicly-available image encoders (ImageNet ResNet-50, BiomedCLIP, EchoCLIP, and DINOv2) on the open five-country African fetal-plane dataset, using leave-one-country-out linear probing, with significance assessed via bootstrap confidence intervals rather than raw averages alone. We extend this three ways, an ablation over the classifier head (logistic regression, linear SVM, k-nearest-neighbours) to rule out classifier-specific artifacts; a transfer experiment that trains on one resource setting and tests on the other, in both directions, using the original high-resource European fetal-plane dataset; and an explainability protocol combining deletion, insertion, and sanity-check tests. We find that modern self-supervised and contrastive encoders (BiomedCLIP, DINOv2) significantly outperform an anatomy-mismatched ultrasound-specific model (EchoCLIP, pretrained on cardiac scans) and an older supervised CNN (ResNet-50) on accuracy, holding up across countries, continents, and classifier heads – yet BiomedCLIP and DINOv2 are not statistically distinguishable from each other, reframing "domain-specific pretraining" as less important than the underlying representation-learning paradigm. Critically, this accuracy advantage does *not* extend to explanation quality. BiomedCLIP's predictions are significantly *less* faithful, by the insertion metric specifically, than ResNet-50's and DINOv2's ($p < 0.01$; deletion-AUC and a parameter-randomization sanity check point the same direction but do not reach significance at this sample size) – a real but metric-specific dissociation between how well a model performs and how trustworthy its explanations are. This argues for routine, statistically-tested, multi-axis trust evaluation, not single-metric leaderboards, as the standard for African medical-imaging AI research.

Keywords: Fetal ultrasound · Foundation models · Domain generalization · Calibration · Explainability · Statistical significance · Low-resource settings · Africa

1 Introduction

Prenatal ultrasound screening is a cornerstone of maternal-fetal care, yet access to trained sonographers is severely constrained across much of sub-Saharan and North Africa, and health workers already using AI-assisted point-of-care ultrasound tools have raised concrete concerns about when such tools can be trusted [6]. A recent scoping review of AI for obstetric ultrasound found that accuracy is reported almost everywhere in this literature, while calibration and explainability are reported almost nowhere [9], and a separate audit of publicly available fetal ultrasound datasets, including the African dataset used here, raised concerns about demographic and technological representativeness [7]. Collectively, these findings highlight a measurable trust gap that existing benchmarks have not yet addressed with statistical rigor.

Recent work in this area largely focuses on reporting a single performance measure, usually accuracy, while practitioners also need to know whether that performance remains reliable across different sites and whether the model's explanations for its decisions are trustworthy. Rather than proposing a new model, this paper re-examines a set of already-popular pretrained encoders through that lens, we test whether cross-country generalization, calibration, and explanation faithfulness rankings (i) hold up under bootstrap significance testing rather than raw averages, (ii) are robust to which classifier is trained on top of the frozen features, and (iii) replicate when distribution shift is reversed against a fully independent, high-resource European dataset.

Contributions. (1) The first cross-country benchmark of four frozen foundation models on the five-country African fetal-plane dataset that reports accuracy, macro-F1, and calibration (ECE) together, backed by bootstrap confidence intervals and pairwise significance tests rather than point estimates alone, and confirms that the resulting ranking survives a change of classifier head. (2) A transfer experiment between resource settings, training on Africa and testing on the original high-resource European fetal-plane dataset [3], and vice versa, showing the same ranking holds in both directions. (3) An explainability protocol combining deletion, insertion, and sanity-check tests that uncovers a statistically significant *dissociation*, the most accurate encoder is not the most faithfully explainable one.

2 Related Work

African fetal ultrasound AI. Sendra-Balcells et al. [13] trained a plane classifier on 1,792 Spanish and 1,008 Danish patients and showed transfer learning with 25 patients per African center could raise recall to 0.92 ± 0.04; the underlying task follows Burgos-Artizzu et al. [3], whose original two-hospital dataset we also use here. Follow-up work has used the African subset for super-resolution [2] and diffusion augmentation [14], and a parallel effort, ACOUSLIC-AI [12], benchmarks fetal abdominal-circumference estimation in Sierra Leone and Tanzania; none of these offer a statistically-tested comparison of competing pretrained encoders, and all report accuracy only. Commercial deployment is also underway,

with DeepEcho recently receiving FDA clearance for use by African ministries of health [5], making rigorous trust evaluation increasingly important beyond the academic setting.

Foundation models. EchoCLIP [4] is a ConvNeXt contrastive model trained on cardiac echocardiogram video-text pairs; BiomedCLIP [15] is a ViT-B/16 trained contrastively on 15M biomedical figure-caption pairs; DINOv2 [10] is a ViT-S/14 trained purely self-supervised on natural images, with no text or medical supervision at all.

Calibration, Faithfulness, and Sanity Checks. We adopt ECE [8] per country to probe calibration under African cross-site shift, addressing the calibration-reporting gap identified in [9]. For faithfulness we use the deletion and insertion metrics of Petsiuk et al. [11], and we additionally apply the model-parameter randomization sanity check of Adebayo et al. [1], which flags saliency methods whose output barely changes when the classifier's weights are replaced with noise. Neither insertion-metric nor sanity-check evaluation has previously been applied to fetal ultrasound foundation models.

3 Data

African dataset. The open *Maternal Fetal Ultrasound Planes from African Countries* dataset comprises 450 frames from five centers, Algeria, Egypt, Ghana, Malawi, and Uganda, labeled into four planes: abdomen, brain, femur, and thorax. Auditing the label distribution revealed that Ghana and Uganda contain zero thorax-plane images (25 patients $\times$ 3 planes vs. 25 $\times$ 4 for the other three centers).

European dataset. To test whether our findings generalize beyond within-Africa shift, we additionally use the original two-hospital FETAL_PLANES_DB [3] (Spain, 12,400 images, 1,792 patients). Its public metadata distinguishes operator and ultrasound machine but not hospital of origin, so we treat it as one pooled high-resource site ("Europe"). We draw a stratified random sample of 60 images per target plane (240 total, fixed seed) to keep compute comparable to a single African center.

4 Methods

Feature extractors. We compare four frozen, publicly-available image encoders that span a range from generic to domain-specific pretraining: ImageNet ResNet-50, a conventional supervised convolutional network trained on natural photographs; BiomedCLIP ViT-B/16, a Vision Transformer trained contrastively on captioned biomedical figures spanning many modalities; EchoCLIP, a contrastive model trained specifically on ultrasound video, but of a different organ (the heart, not the fetus); and DINOv2 ViT-S/14, a Vision Transformer trained with no text or medical supervision at all, purely self-supervised on natural images. In every case the encoder is used only as a fixed feature extractor. Each image is passed through once to obtain an embedding vector, the encoder's own

weights are never updated, and only a lightweight classifier is trained on top (described next). This "frozen probing" design is what keeps the entire pipeline runnable on CPU-only hardware, since no backpropagation through the large encoder is ever required.

Leave-one-country-out (LOCO) linear probing. To measure how well a classifier trained in some African countries generalizes to one it has never seen, we repeatedly hold out one of the five countries as a test set, standardize the frozen features using only the statistics of the remaining four, and fit an L_2-regularized multinomial logistic regression with balanced class weights on those four countries. The held-out country is used purely for evaluation. This simulates the realistic scenario of deploying a model at a new clinical site with no local training data, and is a stricter test than the Europe-to-Africa transfer studied in [13], since it measures generalization *within* Africa itself, across sites that differ in equipment, operators, and patient population.

Statistical testing. Reporting only the mean accuracy over five folds risks over-interpreting differences that could plausibly be noise, given how small each fold is. We therefore report, for every metric, both the fold-wise mean $\pm$ standard deviation (accuracy, macro-F1 computed only over classes present in each country, and Expected Calibration Error (ECE, 10 bins) [8]) and a separate, pooled analysis. We combine every model's out-of-country predictions across all five folds and repeatedly resample them with replacement (bootstrapping, N=2000 resamples) to obtain a 95% confidence interval for each metric. Two models are only called significantly different on a given metric if the bootstrap confidence interval for their difference excludes zero; throughout this paper "significant" refers strictly to this bootstrap criterion, not to subjective importance.

Ablation on classifier choice. Because the classifier trained on top of the frozen features, not the encoder itself, is what makes the final decision, an apparent ranking between encoders could in principle just reflect which encoder happens to suit logistic regression best. To check this, we repeat the entire LOCO evaluation with two alternative classifiers on the same frozen features – a linear support vector machine and a k-nearest-neighbour classifier (k=5) – and verify whether the encoder ranking is preserved regardless of which classifier sits on top.

Transfer between resource settings. The evaluation above only tests generalization *within* the African dataset. To check whether the same ranking holds under a different, independent kind of distribution shift, we fit a classifier on all 450 pooled African images and evaluate it, with no further training, on the 240-image European sample described in Section 3; we then repeat this in reverse, fitting on the European sample and evaluating on the pooled African images. This extends [13], which studied only the Europe-to-Africa direction and only accuracy, to both directions and to calibration.

Explanation faithfulness and sanity checks. For test images that a given encoder-classifier pair gets right (up to 6 per country per model), we divide the image into a 4×4 grid of 16 patches and measure how much blanking out each individual patch changes the predicted probability of the correct class, giving a

coarse map of which patches the model's decision relies on. We then evaluate that map two complementary ways, following Petsiuk et al. [11], a *deletion* test, which removes patches in order of importance and checks how quickly the predicted probability collapses (a fast collapse means the map correctly identified *necessary* regions), and an *insertion* test, which starts from a blank image and reveals the same patches in the same order, checking how quickly the probability recovers (a fast recovery means the map identified regions that are *sufficient* on their own). Finally, we apply the model-parameter randomization sanity check of Adebayo et al. [1], we recompute the same importance map after replacing the trained classifier's weights with random noise of matching scale, then correlate the real and randomized maps. A low correlation is the desired outcome, since it shows the map genuinely depends on what the classifier learned, rather than on incidental structure in the image or encoder alone. Because the set of correctly-classified images differs in size across models (n=19–28), we use a two-sample bootstrap (N=5000) to test whether deletion-AUC, insertion-AUC, or sanity-check correlation differ significantly between encoders.

5 Results

Fig. 1 previews the paper's central finding before the step-by-step evidence below, accuracy and insertion-based explanation faithfulness do not align across the four encoders. BiomedCLIP is the most accurate but the least faithful *by the insertion metric*; on deletion-AUC the ranking reverses and BiomedCLIP is the most faithful of the four (Table 2). Faithfulness is therefore not one-dimensional, the rest of this section establishes, one statistical test at a time, which of these differences are real and which are noise.

Cross-country generalization. Table 1 summarizes LOCO performance with bootstrap 95% CIs. BiomedCLIP (pooled 84.4%, CI [81.1, 87.6]; fold-wise mean 84.5%, Table 1) and DINOv2 (pooled 80.2%, CI [76.4, 83.8]; fold-wise mean 80.5%) significantly outperform EchoCLIP (pooled 66.9%, CI [62.7, 71.3], $p < 0.001$) and ResNet-50 (pooled 75.6%, CI [71.6, 79.3], $p < 0.001$ vs. BiomedCLIP). Critically, **BiomedCLIP vs. DINOv2 is not statistically significant** on accuracy (p=0.057), macro-F1 (p=0.051), or ECE (p=0.520) – nor is BiomedCLIP vs. ResNet-50 on ECE (p=0.293). The robust, significant conclusions are, (a) EchoCLIP, the only encoder pretrained on ultrasound of the wrong organ, is significantly worse than every other model on both accuracy and calibration – domain-matched pretraining without anatomy-matched pretraining is not enough; (b) BiomedCLIP and DINOv2, the two models built on more modern contrastive or self-supervised objectives, both significantly beat the older supervised ResNet-50 on accuracy; and (c) BiomedCLIP is *not* significantly better calibrated than the plain ImageNet CNN once uncertainty is quantified properly, even though its point estimate looks better in Table 1. Comparing the two ECE columns suggests that BiomedCLIP is the clear calibration winner, but the intervals overlap substantially, and the bootstrap test indicates that the observed gap could plausibly be noise given only five country-level folds.

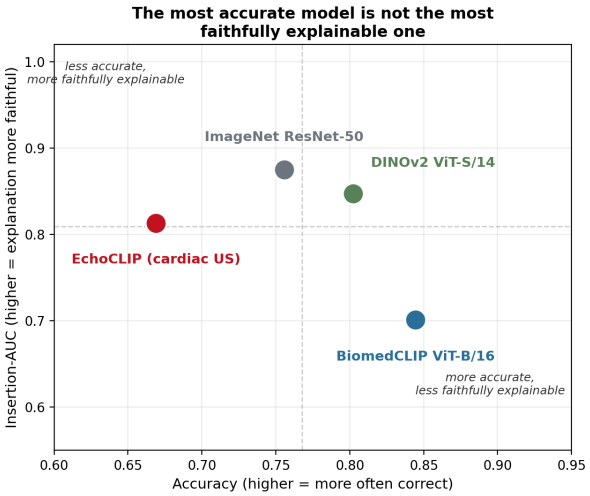

Fig. 1. The central finding at a glance: accuracy (x-axis) and insertion-AUC (y-axis) rank the four encoders differently. BiomedCLIP is the most accurate but the least faithful *by the insertion metric*; it in fact has the best deletion-AUC of the four (Table 2), so this panel deliberately shows one axis of a multi-axis result, not overall faithfulness.

Table 1. LOCO results. "Accuracy" reports the fold-wise mean±std alongside the pooled bootstrap estimate and 95% CI used for all significance testing; the two differ slightly since held-out centers have unequal patient counts (Ghana, Uganda n=75; others n=100).

Model	Accuracy: fold-mean±std / pooled [95% CI]	Macro-F1 (fold-mean±std)	ECE ↓ (pooled [95% CI])
ImageNet ResNet-50	0.759 ± 0.053 / 0.756 [0.716, 0.793]	0.765 ± 0.086	0.111 [0.085, 0.151]
BiomedCLIP ViT-B/16	$\mathbf{0.845 \pm 0.039}$ / **0.844** [0.811, 0.876]	$\mathbf{0.852 \pm 0.050}$	0.091 [0.067, 0.125]
EchoCLIP (cardiac US)	0.663 ± 0.122 / 0.669 [0.627, 0.713]	0.653 ± 0.106	0.232 [0.192, 0.274]
DINOv2 ViT-S/14	0.805 ± 0.066 / 0.802 [0.764, 0.838]	0.798 ± 0.089	0.105 [0.079, 0.143]

Ablation on classifier choice. Repeating the LOCO evaluation under two additional classifier heads (linear SVM; k-nearest-neighbours, k=5) confirms BiomedCLIP remains top-ranked and EchoCLIP at or near the bottom under every classifier. The one exception is k-nearest-neighbours, which degrades every encoder considerably, an expected consequence of a distance-based classifier on high-dimensional embeddings with only a few hundred training images per fold, not evidence against any particular encoder. The *relative ranking* survives the two better-behaved classifiers (logistic regression, linear SVM), so the cross-country conclusions above are not an artifact of choosing logistic regression.

Transfer between resource settings. Fitting on all pooled African data and testing on the independent European sample, and vice versa (full per-direction table released with the code), shows the ranking replicates in *both* directions. BiomedCLIP leads training-on-Africa-testing-on-Europe (80.8%) and training-on-Europe-testing-on-Africa (78.4%); EchoCLIP trails in both (65.0%,

Table 2. Faithfulness (deletion/insertion-AUC) and sanity-check correlation, mean±std, over correctly-classified test images (n=19–28 per model).

Model	Deletion-AUC ↓	Insertion-AUC ↑	Sanity-check corr. ↓ |0|
ImageNet ResNet-50	0.436 ± 0.334	$\mathbf{0.875 \pm 0.157}$	0.025 ± 0.295
BiomedCLIP ViT-B/16	$\mathbf{0.295 \pm 0.230}$	0.701 ± 0.215	-0.066 ± 0.416
EchoCLIP (cardiac US)	0.342 ± 0.289	0.813 ± 0.226	0.142 ± 0.419
DINOv2 ViT-S/14	0.427 ± 0.332	0.847 ± 0.126	-0.013 ± 0.318

60.2%). The test population here is an entirely different continent, hospital system, and patient cohort, so this is stronger evidence for a genuine encoder-quality difference than the within-Africa result alone. Calibration does not always track accuracy, however, DINOv2 has the best training-on-Europe-testing-on-Africa ECE (0.137), edging out BiomedCLIP (0.151).

Confusion patterns. Across all four encoders the fetal thorax plane is the most frequently misclassified class, most often confused with abdomen. Thorax and abdomen are the two most inferior transverse sections of the fetal trunk, so a small probe-angle change can visually shift one into the other, an effect compounded by two of the five centers never having acquired a thorax image at all (Section 3).

Explanation faithfulness and sanity checks. Table 2 and Fig. 2 reveal a genuine dissociation, not a uniform ranking. On **deletion-AUC, BiomedCLIP is the most faithful of the four** (lowest, 0.295), though the gap to each other encoder individually is *not* statistically significant given the small per-model sample (n=19–28: vs. DINOv2 p=0.090, vs. ResNet-50 p=0.092, vs. EchoCLIP p=0.536). On **insertion-AUC, the ranking reverses: BiomedCLIP is significantly *worse*** than DINOv2 (0.701 vs. 0.847, p=0.001) and ResNet-50 (0.701 vs. 0.875, p=0.001). Taken together, BiomedCLIP's correct predictions depend on a small, sharply concentrated set of patches – removing them collapses confidence fastest (best deletion-AUC) – but that same small set, revealed in isolation, is *insufficient* to reconstruct full confidence (worst insertion-AUC), while ResNet-50 and DINOv2's decisions are more evenly supported by multiple regions. In practical terms, if a clinician were shown only the single most-highlighted region behind a BiomedCLIP prediction, that region would correctly be necessary but would still be a poor stand-alone summary of the decision, whereas the same kind of summary for a ResNet-50 or DINOv2 prediction would be more informative on its own – even though ResNet-50 and DINOv2 are wrong more often in the first place. Deletion-AUC and insertion-AUC are therefore both needed, since a model can win on one and lose on the other. Sanity-check correlations are low (near zero) for all models, with EchoCLIP nominally highest (0.142, i.e. closest to failing the check) though not significantly different from the others ($p > 0.09$).

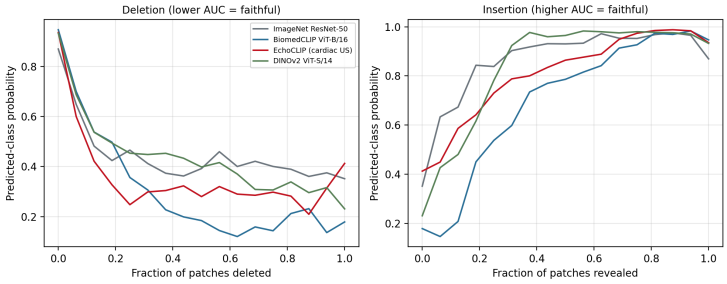

Fig. 2. Mean deletion and insertion curves. BiomedCLIP collapses fastest under deletion but rises slowest under insertion – the trade-off underlying Table 2, not a single-axis ranking.

6 Discussion

Three findings carry practical implications for groups selecting a pretrained backbone under compute constraints. First, domain-specificity alone does not predict cross-site generalization, the two best performers differ mainly in pretraining *paradigm*, not medical-domain relevance, so practitioners should validate candidate encoders empirically rather than assume an "ultrasound-labeled" model is automatically safer. Second, the ranking replicates across two independent datasets and three classifier heads, reassuring for external validity, but the finer-grained calibration and insertion-faithfulness rankings do not always track accuracy and must be checked on their own terms. Third, and most concretely, accuracy leaderboards can actively mislead trust-sensitive deployment decisions. A team selecting BiomedCLIP purely for its accuracy and calibration advantage would deploy the model whose explanations are, by a statistically confirmed margin, the least sufficient of the four by the insertion metric, illustrating why joint accuracy/calibration/explainability reporting [9] can change a deployment recommendation. The entire pipeline runs in under an hour on CPU-only hardware, supporting its practicality in resource-constrained settings where access to GPU infrastructure may be limited.

7 Limitations

Both datasets are small (25 patients/country for Africa, a 240-image stratified sample for Europe), so per-site and per-model estimates carry real uncertainty, which is why we report bootstrap CIs throughout rather than point estimates alone; the thorax class is entirely absent in two of five African countries. We evaluate frozen linear probing only, not fine-tuning, as a deliberate CPU-only scope choice, and the encoder ranking established here may not hold once these models are fine-tuned end-to-end. The European dataset's public metadata does not disaggregate by hospital, so our "Europe" site is pooled rather than multi-site internally. Our faithfulness analysis uses a coarse 4×4 occlusion grid and is

restricted to correctly-classified images (n=19–28 per model), which limits the power of the faithfulness significance tests, only the insertion-AUC gap reached significance, and the seemingly larger deletion-AUC gap favoring BiomedCLIP did not ($p > 0.09$), a result that should be treated as preliminary until replicated on a larger sample. We also apply no multiple-comparisons correction across the many pairwise tests reported; the two significant insertion-AUC gaps ($p = 0.001$) survive a conservative Bonferroni threshold for the six faithfulness comparisons ($\alpha = 0.0083$), but the boundary-case accuracy/macro-F1 gaps ($p = 0.051$–0.057) would not. Finally, our trust evaluation is purely computational; whether sonographers would actually experience BiomedCLIP's explanations as less useful is left to human-factors follow-up work.

8 Conclusion

We presented a statistically-tested benchmark of frozen foundation models for African fetal ultrasound plane classification, spanning country- and continent-level distribution shift. Our central finding is not simply which model wins, but that *accuracy, calibration, and explanation faithfulness rank models differently, and only some of the differences are statistically robust* – this argues for routine, statistically-tested, multi-axis trust evaluation, not single-metric leaderboards, as the standard for African medical-imaging AI research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
2. Boumeridja, H., Ammar, M., Alzubaidi, M., Mahmoudi, S., Benamer, L.N., Agus, M., Househ, M., Lekadir, K., El Habib Daho, M.: Enhancing fetal ultrasound image quality and anatomical plane recognition in low-resource settings using super-resolution models. Scientific Reports **15**(1), 8376 (2025)
3. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
4. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine **30**, 1481–1488 (2024)
5. DeepEcho: Deepecho reaches new milestone with fda 510(k) clearance for ai-powered prenatal imaging. Press release, June 20, 2025 (2025)
6. Della Ripa, S., Santos, N., Walker, D.: Ai-enabled obstetric point-of-care ultrasound as an emerging technology in low- and middle-income countries: provider and health system perspectives. BMC Pregnancy and Childbirth **25**(1), 729 (2025)

7. Fiorentino, M.C., Moccia, S., Di Cosmo, M., Frontoni, E., Giovanola, B., Tiribelli, S.: Uncovering ethical biases in publicly available fetal ultrasound datasets. npj Digital Medicine **8**, 355 (2025)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330 (2017)
9. Gupta, V., Santos, N., Della Ripa, S., Walker, D.: Emerging applications of artificial intelligence for obstetric ultrasound: A scoping review. International Journal of Gynecology & Obstetrics **174**, 34–43 (2026)
10. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
11. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (BMVC) (2018)
12. Sappia, M.S., de Korte, C.L., van Ginneken, B., Ninalga, D., Kondo, S., Kasai, S., Hirasawa, K., Akumu, T., Martín-Isla, C., Lekadir, K., Campello, V.M., Fabila, J., Beverdam, A., van Dillen, J., Neff, C., Murphy, K.: Acouslic-ai challenge report: Fetal abdominal circumference measurement on blind-sweep ultrasound data from low-income countries. Medical Image Analysis **105**, 103640 (2025)
13. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ahmed, Y.A., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L.N., Kisembo, H.N., Sereke, S.G., Wanyonyi, S.Z., Temmerman, M., Gratacós, E., Bonet, E., Eixarch, E., Mikolaj, K., Tolsgaard, M.G., Lekadir, K.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five african countries. Scientific Reports **13**(1), 2728 (2023)
14. Wang, F., Whelan, K., Silvestre, G., Curran, K.M.: Generative diffusion model bootstraps zero-shot classification of fetal ultrasound images in underrepresented african populations. In: MICCAI Workshop on Perinatal, Preterm and Paediatric Image Analysis (PIPPI) (2024), arXiv:2407.20072
15. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2024), arXiv:2303.00915