Understanding Task Aggregation for Generalizable Ultrasound Foundation Models

Fangyijie Wang[1,3†⋆], Tanya Akumu[2†], Vien Ngoc Dang[2], Amelia Jimńez-Sánchez[2], Jieyun Bai[6,7], Guénolé Silvestre[1,4] Karim Lekadir[2,5], and Kathleen M. Curran[1,3⋆]

[1] Research Ireland Centre for Research Training in Machine Learning
[2] Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
[3] School of Medicine, University College Dublin, Dublin, Ireland
[4] School of Computer Science, University College Dublin, Dublin, Ireland
[5] Institució Catalana de Recerca i Estudis Avançats (ICREA)
[6] Department of Cardiovascular Surgery, The First Affiliated Hospital of Jinan University, Jinan University, Guangzhou, China
[7] Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand
[†] Equal contribution

Abstract. Foundation models promise to unify multiple clinical tasks within a single framework, but recent ultrasound studies report that unified models can underperform task-specific baselines. We hypothesize that this degradation arises not from model capacity limitations, but from task aggregation strategies that ignore interactions between task heterogeneity and available training data scale. In this work, we systematically analyze when heterogeneous ultrasound tasks can be jointly learned without performance loss, establishing practical criteria for task aggregation in unified clinical imaging models. We introduce M2DINO, a multi-organ, multi-task framework built on DINOv3 with task-conditioned Mixture-of-Experts blocks for adaptive capacity allocation. We systematically evaluate 27 ultrasound tasks spanning segmentation, classification, detection, and regression under three paradigms: task-specific, clinically-grouped, and all-task unified training. Our results show that aggregation effectiveness depends strongly on training data scale. While clinically-grouped training can improve performance in data-rich settings, it may induce substantial negative transfer in low-data settings. In contrast, all-task unified training exhibits more consistent performance across clinical groups. We further observe that task sensitivity varies by task type in our experiments: segmentation shows the largest performance drops compared with regression and classification. These findings provide practical guidance for ultrasound foundation models, emphasizing that aggregation strategies should jointly consider training data availability and task characteristics rather than relying on clinical taxonomy alone.

Keywords: Foundation models · Ultrasound imaging · Multi-Task learning.

[⋆] Corresponding authors: fangyijie.wang@ucdconnect.ie, kathleen.curran@ucd.ie

1 Introduction

Ultrasound imaging is a cornerstone of clinical care, including obstetrics [15], cardiology [25], oncology [14], and point-of-care (POC) settings [21]. It enables rapid, non-invasive, and cost-effective assessment of diverse anatomical structures at the bedside. However, ultrasound image appearance varies substantially across operators, devices, and acquisition protocols, complicating robust generalization [19,5,24]. Despite recent advances in Deep Learning (DL) for ultrasound, most models focus on isolated task instances (e.g., single-organ segmentation [11], multi-organ classification [10], or multi-organ segmentation [2]) or limited combinations of tasks (e.g., joint classification and segmentation [9]), rather than enabling unified multi-organ, multi-task analysis. Such specialization limits clinical applicability, as real-world workflows require simultaneous multi-organ and multi-task assessment. Foundation models therefore aim to streamline deployment, promote cross-task knowledge sharing, and enable comprehensive ultrasound analysis [1]. However, developing a single unified model that reliably performs segmentation, detection, classification, and regression across heterogeneous ultrasound tasks remains an open challenge.

Recent multi-task and foundation-style approaches aim to unify clinical tasks within a single model, thereby simplifying deployment and enabling cross-task knowledge sharing [9,2,10,13]. While these methods report promising results on selected task combinations, a systematic study of how task aggregation strategies influence performance across organs and task types is still lacking. In particular, it remains unclear which tasks can be effectively unified without inducing negative transfer, and how training data scale modulates such interactions.

To address these questions, we introduce Multi-organ and Multi-task DINO framework (M2DINO), a DINOv3-based encoder augmented with task-conditioned Mixture-of-Experts (MoE) blocks for large-scale multi-task learning across 27 ultrasound tasks spanning segmentation, classification, detection, and regression. We investigate three training paradigms: (1) task-specific (TS) training, where each task is optimized independently; (2) clinically-grouped (CG) training, where clinically related tasks are trained jointly; and (3) all-task unified (AU) training, where all tasks are learned simultaneously within a single model.

Our contributions are: (1) We introduce M2DINO, a unified multi-organ, multi-task ultrasound framework built on DINOv3 with task-conditioned MoE for adaptive capacity allocation across heterogeneous task objectives. (2) We introduce a structured framework for evaluating clinical task aggregation and compatibility across organ systems and prediction types. (3) Through experiments on 27 tasks, we show scale-dependent aggregation effects, identify conditions under which CG training induces negative transfer, and provide practical design guidelines for developing unified ultrasound foundation models.

2 Methodology

This section first formalizes the problem setting and training paradigms (Section 2.1). We then detail the proposed M2DINO architecture, including the backbone [22], the task-conditioned MoE, and the heads with the multi-task objective (Sections 2.2–2.4). Fig. 1 illustrates an overview of the M2DINO framework.

2.1 Problem Setting and Training Paradigms

Let $\mathcal{D} = \{\mathcal{D}_t\}_{t=1}^{T}$ denote a collection of T ultrasound tasks spanning segmentation, classification, regression, and detection, covering diverse anatomical regions. Each task $\mathcal{D}_t = \{(\mathbf{X}_i^t, \mathbf{Y}_i^t)\}$ consists of ultrasound images $\mathbf{X}_i^t$ and TS labels $\mathbf{Y}_i^t$. Under the unified training paradigms, our objective is to learn a shared encoder f_θ that maps an input image $\mathbf{X}$ to a latent representation, which is subsequently optimized by heterogeneous TS prediction heads.

We study how different task aggregation strategies affect model performance and transfer behavior within a common DINOv3-based foundation model. Specifically, we evaluate three training paradigms in a controlled comparison setting:

- **task-specific (TS)**: A separate DINOv3 model is trained independently for each task t, without any parameter sharing or cross-task interaction.
- **clinically-grouped (CG)**: Tasks are jointly trained within predefined clinical groups based on shared organ systems and examination context (e.g., obstetric tasks (OB), breast imaging tasks (Breast), and lung ultrasound tasks (Lung)). Each group shares a DINOv3 Vision Transformer (ViT) encoder and task-conditioned MoE routing while optimizing heterogeneous prediction objectives (segmentation, classification, detection, or regression).
- **all-task unified (AU)**: All T tasks are trained simultaneously within a single shared DINOv3 ViT encoder.

For unified settings (CG and AU), the multi-task loss function is defined as: $\mathcal{L} = \sum_{t=1}^{T} \lambda_t \mathcal{L}_t$, where T denotes the number of tasks trained jointly in the current paradigm (e.g., $[3-27]$), and $\mathcal{L}_t$ denotes the TS loss and λ_t denotes balancing coefficients. Unless otherwise specified, losses are equally weighted.

In our controlled comparison, all training paradigms share the same pre-trained DINOv3 backbone, MoE configuration (when enabled), input resolution, data pre-processing, and optimization settings. As the effective training data size varies across paradigms (e.g., TS vs. AU), we perform a limited learning rate search within a fixed range for each setting to ensure stable optimization, using a consistent validation-based selection protocol. Table 1 summarizes the experimental settings for each training paradigm.

2.2 DINO Backbone

We use the pre-trained DINOv3 [22] model as our encoder backbone. DINOv3 provides ViT-S/B/L variants; we adopt the ViT-B/16 backbone to balance model capacity with dataset scale and computational efficiency. Given an ultrasound image, we convert it in RGB format to obtain the input $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$. The DINOv3-based ViT encoder f_θ produces token embeddings $\mathbf{Z}$ and corresponding spatial feature maps $\mathbf{F}$: $(\mathbf{Z}, \mathbf{F}) = f_\theta(\mathbf{X})$. Unlike prior multi-task formulations [23], we use spatial feature maps as the unified interface across all tasks.

Table 1. Experimental settings for evaluating different training paradigms. Seg: segmentation; Cls: classification; Reg: regression; Det: detection. MO: Multi-organ.

Training Paradigm	**MoE**	**# Tasks**	**# Images**	**Reg**	**Cls**	**Seg**	**Det**	**Target Anatomy**	**Clinical Group**
TS	✗	1	1,144		✓			Breast	Breast
TS	✗	1	1,776			✓		Breast	Breast
TS	✗	1	208	✓				Cervical	OB
TS	✗	1	2,818	✓				PS/Fetal head	OB
TS	✗	1	762			✓		Fetal abdomen	OB
TS	✗	1	624	✓				Fetal femur	OB
TS	✗	1	1,849			✓		Fetal head	OB
TS	✗	1	5,952		✓			Fetal organs	OB
TS	✗	1	483		✓			Fetal breech	OB
TS	✗	1	1,482			✓		Lung	Lung
TS	✗	1	772		✓			Lung	Lung
CG	✓	7	11,910	✓	✓	✓		Fetal anatomy	OB
CG	✓	3	2,254		✓	✓		Lung	Lung
CG	✓	3	2,920		✓	✓		Breast	Breast
AU	✓	27	32,311	✓	✓	✓	✓	MO	All

Downstream task-specific heads, including a dense prediction transformer (DPT) decoder [18] for segmentation, take the feature maps $\mathbf{F}$ as input.

Although the f_θ produces token embeddings, we use only the spatial feature maps $\mathbf{F}$ for downstream heads. This design provides a consistent dense feature representation across segmentation, detection, classification, and regression tasks. Using global token pooling (e.g., the classification token) could favor global prediction tasks over dense prediction tasks such as segmentation and detection. By adopting feature maps $\mathbf{F}$ as the unified interface, we maintain architectural consistency and isolate the effect of task grouping in our compatibility analysis.

2.3 Mixture of Experts with Task-Conditioned Routing

To mitigate task interference in unified training paradigms (CG/AU), we integrate task-conditioned MoE blocks into the DINOv3 encoder, inspired by [8,12]. Each task is assigned a unique identifier t, which is mapped to a learnable embedding vector: $\mathbf{e}_t = \text{Embedding}(t)$. The gating network (shown in Fig. 1) conditions expert selection on both token embeddings $\mathbf{h}$ and the task embedding $\mathbf{e}_t$: $g\left(\mathbf{h}, \mathbf{e}_t\right) = \text{Softmax}\left(W_g\left[\mathbf{h}; \mathbf{e}_t\right]\right)$. The output of the MoE block is computed as a weighted combination of expert outputs: $\mathbf{h}' = \sum_{i=1}^{K} g_i\left(\mathbf{h}, \mathbf{e}_t\right) E_i(\mathbf{h})$, where E_i denotes the i-th expert and K is the total number of experts. This design enables task-adaptive capacity while maintaining a shared backbone.

Instead of integrating MoE into all ViT layers, we integrate the MoE blocks into the later layers (layers $7-12$, i.e., the last six layers). Early transformer layers tend to encode generic low-level image representations, while later layers encode task-specific representations [4]. Restricting MoE blocks to later layers

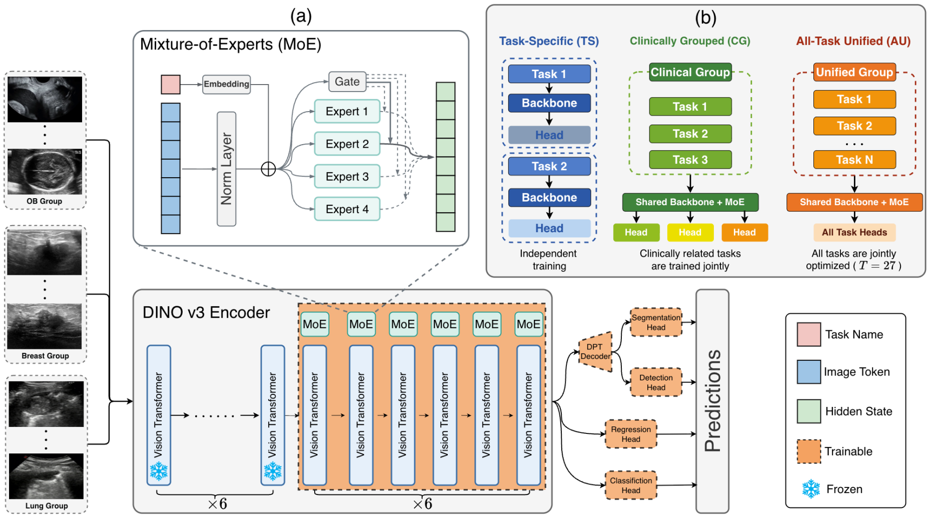

Fig. 1. Overview of our M2DINO framework. **(a)** Ultrasound images are processed by a shared DINOv3 encoder augmented with task-conditioned MoE blocks. The unified representation is optimized for segmentation, detection, regression, and classification via task-specific prediction heads. Frozen and trainable components are indicated. **(b)** A conceptual comparison of the three training paradigms. Although the architecture remains the same, task-specific (TS), clinically-grouped (CG), and all-task unified (AU) differ in how tasks are aggregated during training and in whether the MoE is enabled.

enables efficient conditional capacity allocation. Given the scale of our dataset (32,311 training and 8,077 validation samples), we adopt a partial-MoE design to balance task-adaptive capacity with computational efficiency.

2.4 Task-Specific Heads and Multi-Task Learning

For four different task types, including segmentation, classification, regression, and detection, we develop four lightweight heads to improve computational efficiency. Let $\mathbf{F}$ denote the shared feature maps produced by f_θ. Each task employs a lightweight prediction head h_t to output $\hat{y}_t = h_t(\mathbf{F})$, where the head parameters are task-specific. For segmentation tasks, we adopt a DPT-style [18] decoder to generate dense pixel-wise predictions. Classification and regression tasks utilize global pooling followed by fully connected layers, while detection tasks adopt a task-specific detection head.

Each task is optimized using an appropriate loss function $\mathcal{L}_t$. Specifically, we use Dice loss for segmentation, cross-entropy loss for classification, and $L1$ loss for regression. For detection, we use a single-stage detection loss combining focal loss for pixel-wise supervision and Smooth $L1$ loss for normalized bounding box regression at the corresponding ground-truth center cell. For unified settings (CG/AU), the overall objective is defined in Sec 2.1.

3 Experiments

Dataset. The dataset is designed to evaluate the model's ability to generalize across four fundamental task categories:

- **Segmentation** (12 tasks): Pixel-level annotations for fetal organs (e.g., the head, heart, and abdomen), maternal structures, and lesions. The training set contains 16,615 samples and the test set includes 2,674 samples.
- **Classification** (9 tasks): Includes fetal standard-plane and fetal position classification, lung disease recognition, and tumor malignancy assessment. The training set has 16,361 samples, and test set has 2,727 samples.
- **Detection** (3 tasks): Localization of thyroid nodules, uterine fibroids, and spinal cord injuries (4,333 training / 725 test samples).
- **Regression** (3 tasks): Biometric measurements including angle of progression, cervical length, and fetal femur length. The training set includes 3,078 samples, and the test set contains 617 samples.

During training, 20% of the training data are selected for validation.

Implementation Details. All methods were trained for 200 epochs with a batch size of 16 using AdamW (initial learning rate $1e-5$, weight decay $1e-4$). The backbone learning rate was set to $2e-5$, the DPT head to $1e-5$, MoE to $2e-4$, and task-specific heads to $1e-3$ to accelerate convergence. Implementation was based on PyTorch (2.1.2) and Segmentation Models PyTorch [7] with CUDA (12.2), and experiments were conducted on a NVIDIA 4090 GPU. Models were evaluated on the validation set after each epoch, and the best-performing model's weights were saved. Data augmentation and preprocessing followed standard protocols. Full implementation details and code are available at: GitHub.

Evaluation Metrics. We define standardized evaluation metrics for each of the task types: **Segmentation**: We report the Dice Similarity Coefficient (DSC) [3] for region overlap and Hausdorff Distance (HD) [6] for boundary accuracy. **Classification**: We use the Area Under the Curve (AUC) [17], F1-score [20], and Matthews Correlation Coefficient (MCC) [16]. **Detection**: We use the Intersection over Union (IoU) [26] to measure the localization accuracy of predicted bounding boxes. **Regression**: The Mean Radial Error (MRE), reported in pixels, reflects the real-world clinical measurement precision, as it is computed at the original image resolution (i.e., predictions are mapped back from resized inputs).

4 Results

Fig. 2 presents absolute performance comparisons across training paradigms. In the data-rich obstetrics (OB) group (11,910 training samples), both CG and AU training paradigms generally improve over TS on most tasks. Specifically, AU reduces cervial regression error (MRE: 30.4 → 15.6) and increases fetal abdomen segmentation overlap (DSC: 0.217 → 0.481). For fetal head segmentation and

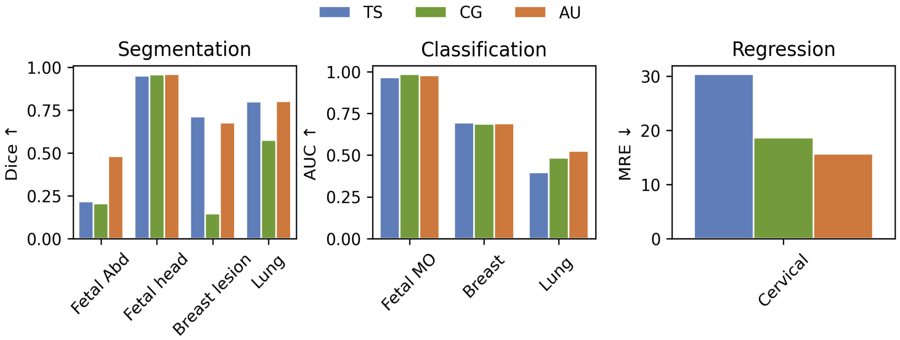

Fig. 2. Absolute performance of TS, CG, and AU training paradigms across representative tasks: segmentation (DSC ↑), classification (AUC ↑), and regression (MRE ↓). Abd: Abdomen; MO: Multi-organ.

multi-organ classification, CG and AU yield small, modest gains. However, the Breast and Lung groups exhibit different trends. AU improves lung classification (AUC: 0.396 → 0.525). In contrast, CG shows large performance drops in breast lesion segmentation (DSC: 0.713 → 0.145) and lung segmentation (DSC: 0.801 → 0.576). These results suggest that task aggregation strategies (CG/AU) benefit from data-rich settings, whereas CG is less reliable in low-data settings.

To quantify task aggregation effects, Fig. 3 reports relative performance changes with respect to TS. It shows that the impact of CG and AU depends strongly on data scale. In the OB group (11,910 training samples), both AU and CG outperform TS across most tasks, with the largest improvements in regression and segmentation. CG shows a 5.1% performance drop in fetal abdomen segmentation. In contrast, CG shows significant performance drops in smaller groups (Breast and Lung), especially in breast lesion segmentation (-79.7%). By comparison, AU exhibits comparatively less performance changes (-4.9%). These results suggest that task aggregation interacts strongly with data availability, and that CG is more prone to negative transfer in low-data settings.

Table 2 summarizes the group-wise average performance change relative to TS training. CG yields positive improvements in the OB ($\Delta = +2.93$; 11,910 samples), but shows slight average decreases in Breast ($\Delta = -0.29$) and Lung ($\Delta = -0.07$). In contrast, AU exhibits more stable performance across datasets (+3.76 in OB, −0.02 in Breast, and +0.07 in Lung) and generally outperforms CG in smaller-scale settings. These results suggest that the effectiveness of CG training paradigm depends strongly on data scale.

5 Discussion

Our study shows that the effectiveness of task aggregation strategies (clinically-grouped (CG)/all-task unified (AU)) in ultrasound imaging is strongly dependent on training data scale. In the data-rich obstetrics (OB) group, both CG and

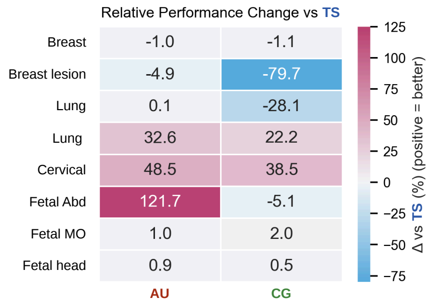

Fig. 3. Relative performance change (Δ, %) with respect to TS.

Table 2. Group-wise average performance change (Δ) relative to TS. Positive values indicate improvement; for regression (MRE), the sign is adjusted accordingly.

Group	**# Images**	**Δ(CG)**	**Δ(AU)**
OB	11,910	+**2.93**	+**3.76**
Breast	2,920	−0.29	−0.02
Lung	2,254	−0.07	+**0.07**

AU improve performance over task-specific (TS) (Table 2). However, in smaller groups (Breast and Lung), CG induces significant negative transfer, indicating that clinical grouping alone does not guarantee positive transfer.

Importantly, AU shows more stable performance across groups and fewer large performance drops than CG (Fig. 3). This suggests that broader task aggregation may provide a regularizing effect that reducing overfitting when data are limited. Our findings highlight that partial grouping (i.e., CG) can be more prone to negative transfer in small datasets, whereas all-task aggregation yields more reliable transfer behavior.

Furthermore, we observe task-type-dependent effects in our experiments. Segmentation shows the largest performance drops and negative transfer, while regression and classification remain comparatively stable (Fig. 2). These results suggest that the design of aggregation strategies for foundation models should consider clinical taxonomy together with data scale and task characteristics.

This study has several limitations. First, we focus on a single backbone (DINOv3) and predefined clinical grouping strategies. Alternative architectures, such as ultrasound-specific foundation models (e.g., USFM [9] and TinyUSFM [13]), or data-driven grouping schemes, may lead to different outcomes. Second, our analysis is limited to ultrasound imaging. Future work should examine whether similar transfer patterns generalize to other 2D modalities (e.g., radiography or digital pathology) as well as 3D domains such as CT and MRI. Despite these limitations, our findings provide empirical evidence that aggregation strategy and data scale are important factors influencing the performance and stability of unified medical foundation models.

6 Conclusion

We present a large-scale empirical analysis of task aggregation strategies for multi-task ultrasound foundation models across 27 heterogeneous clinical tasks. Our findings show that aggregation effectiveness is governed not only by clinical taxonomy but also by data scale and task characteristics. Clinically-grouped aggregation improves performance in data-rich settings but can induce negative transfer in low-data settings. In contrast, anatomy-agnostic aggregation provides more stable cross-task transfer. Segmentation tasks are particularly sensitive to aggregation design, underscoring the need for principled task selection. These results demonstrate that naive task scaling does not guarantee improved foundation models and provide practical guidelines for constructing reliable and scalable ultrasound foundation models, with implications for broader medical imaging applications.

Acknowledgments. This work was funded by Taighde Éireann – Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

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

References

1. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundation models defining a new era in vision: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2245–2264 (2025). https://doi.org/10.1109/TPAMI.2024.3506283
2. Chen, H., Cai, Y., Wang, C., Chen, L., Zhang, B., Han, H., Guo, Y., Ding, H., Zhang, Q.: Multi-organ foundation model for universal ultrasound image segmentation with task prompt and anatomical prior. IEEE Transactions on Medical Imaging **44**(2), 1005–1018 (2025). https://doi.org/10.1109/TMI.2024.3472672
3. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
4. Dorszewski, T., Tětková, L., Jenssen, R., Hansen, L.K., Wickstrøm, K.K.: From colors to classes: Emergence of concepts in vision transformers. In: Explainable Artificial Intelligence. pp. 28–47. Springer Nature Switzerland (2026)
5. Huang, L., Zhou, J., Jiao, J., Zhou, S., Chang, C., Wang, Y., Guo, Y.: Standardization of ultrasound images across various centers: M2o-diffgan bridging the gaps among unpaired multi-domain ultrasound images. Medical Image Analysis **95**, 103187 (2024). https://doi.org/10.1016/j.media.2024.103187
6. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9), 850–863 (1993). https://doi.org/10.1109/34.232073
7. Iakubovskii, P.: Segmentation models pytorch (2019), https://github.com/qubvel/segmentation_models.pytorch
8. Jain, Y., Behl, H., Kira, Z., Vineet, V.: Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. Advances in Neural Information Processing Systems **36**, 69625–69637 (2023)
9. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., Guo, Y.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024). https://doi.org/10.1016/j.media.2024.103202

10. Kang, Q., Lao, Q., Gao, J., Bao, W., He, Z., Du, C., Lu, Q., Li, K.: Urfm: a general ultrasound representation foundation model for advancing ultrasound image diagnosis. IScience **28**(8) (2025)
11. Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., Li, Q.: Echofm: Foundation model for generalizable echocardiogram analysis. IEEE Transactions on Medical Imaging **44**(10), 4049–4062 (2025). https://doi.org/10.1109/TMI.2025.3580713
12. Lu, Y., Weng, M., Xiao, Z., Jiang, R., Su, W., Zheng, G., Lu, P., Li, X.: Dynamic-dino: Fine-grained mixture of experts tuning for real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20847–20856 (October 2025)
13. Ma, C., Jiao, J., Liang, S., Fu, J., Wang, Q., Li, Z., Wang, Y., Guo, Y.: Tinyusfm: Towards compact and efficient ultrasound foundation models. arXiv preprint arXiv:2510.19239 (2025)
14. Madsen, H.H.T., Rasmussen, F.: Contrast-enhanced ultrasound in oncology. Cancer Imaging **11**(1A), S167 (2011)
15. Maraci, M.A., Yaqub, M., Craik, R., Beriwal, S., Self, A., von Dadelszen, P., Papageorghiou, A., Noble, J.A.: Toward point-of-care ultrasound estimation of fetal gestational age from the trans-cerebellar diameter using CNN-based ultrasound image analysis. J Med Imaging (Bellingham) **7**(1), 014501 (Jan 2020)
16. Matthews, B.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure **405**(2), 442–451 (1975). https://doi.org/10.1016/0005-2795(75)90109-9
17. Peterson, W., Birdsall, T., Fox, W.: The theory of signal detectability. Transactions of the IRE Professional Group on Information Theory **4**(4), 171–212 (1954). https://doi.org/10.1109/TIT.1954.1057460
18. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
19. Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D.G., Papageorghiou, A.T., International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st): Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet. Gynecol. **39**(3), 266–273 (2012)
20. Sasaki, K., Sakamoto, S., Uchida, H., Shigeta, T., Matsunami, M., Kanazawa, H., Fukuda, A., Nakazawa, A., Sato, M., Ito, S., et al.: Two-step transplantation for primary hyperoxaluria: A winning strategy to prevent progression of systemic oxalosis in early onset renal insufficiency cases. Pediatric Transplantation **19**(1), E1–E6 (2015)
21. Self, A., Chen, Q., Desiraju, B.K., Dhariwal, S., Gleed, A.D., Mishra, D., et al.: Developing clinical artificial intelligence for obstetric ultrasound to improve access in underserved regions: Protocol for a computer-assisted low-cost point-of-care ultrasound (calopus) study. JMIR Res Protoc **11**(9), e37374 (Sep 2022)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
23. Song, X., Xu, X., Zhang, J., Machado Reyes, D., Yan, P.: Dino-reg: Efficient multimodal image registration with distilled features. IEEE Transactions on Medical Imaging **44**(9), 3809–3819 (2025). https://doi.org/10.1109/TMI.2025.3567247

24. Vega, R., Dehghan, M., Nagdev, A., Buchanan, B., Kapur, J., Jaremko, J.L., Zonoobi, D.: Overcoming barriers in the use of artificial intelligence in point of care ultrasound. npj Digital Medicine **8**(1), 213 (Apr 2025)
25. Villemain, O., Baranger, J., Friedberg, M.K., Papadacci, C., Dizeux, A., Messas, E., Tanter, M., Pernot, M., Mertens, L.: Ultrafast ultrasound imaging in pediatric and adult cardiology: Techniques, applications, and perspectives. JACC: Cardiovascular Imaging **13**(8), 1771–1791 (2020). https://doi.org/10.1016/j.jcmg.2019.09.019
26. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: The AAAI Conference on Artificial Intelligence (AAAI). pp. 12993–13000 (2020)