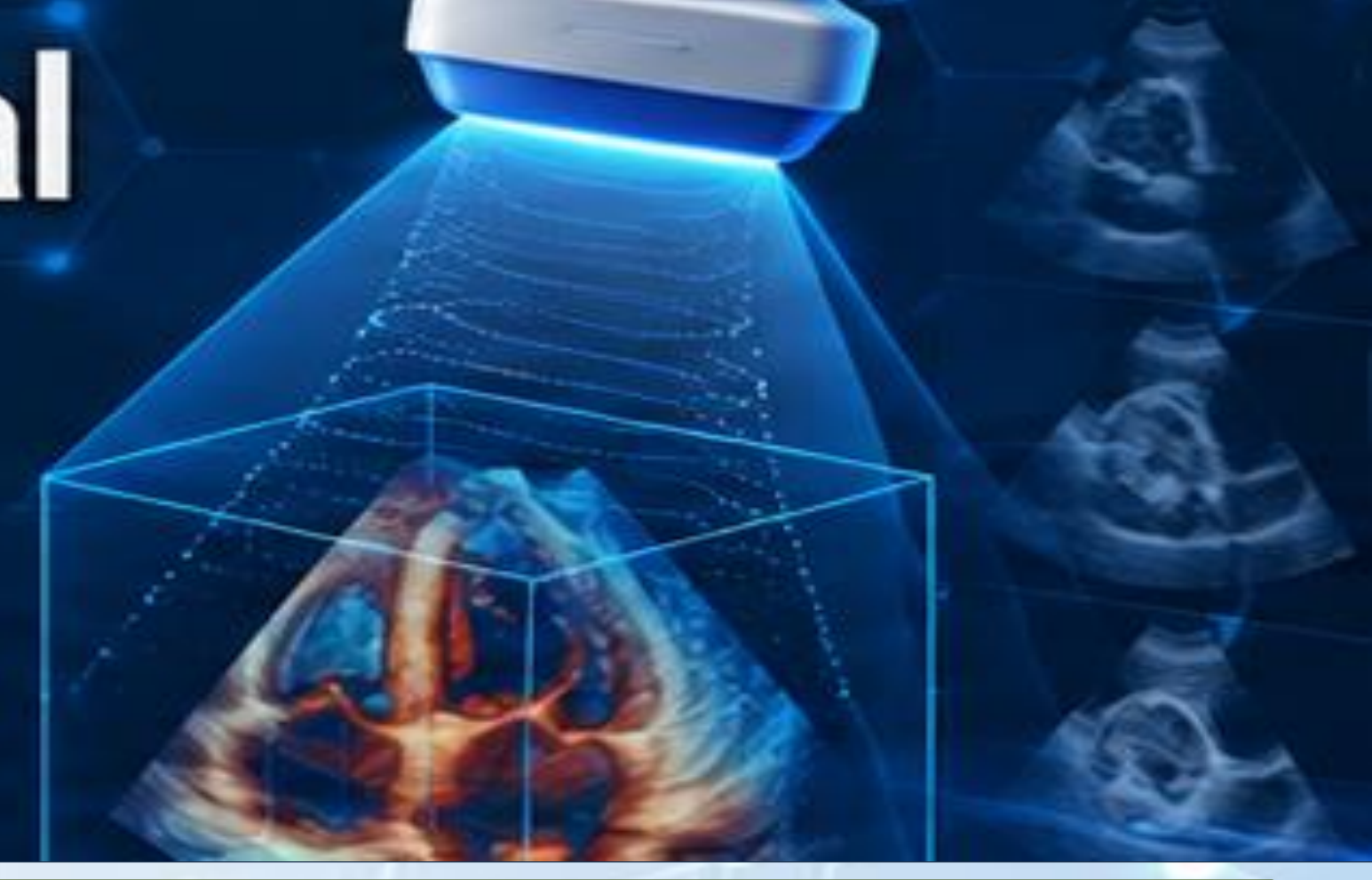




Large-scale foundation model for automatic interpretation of routine clinical 3D ultrasound

Arkadiusz Sitek^{1,2}, Sarah C. Lassey^{3,2*}, Pawel Renc^{1,2}, Kira Bromwich^{3,2}, Szymon Plotka⁴, Thomas D. Shipp^{3,2}

¹Massachusetts General Hospital, ²Harvard Medical School, ³Brigham and Women's Hospital, ⁴Jagiellonian University, *Presenting author (slassey@bwh.harvard.edu)



1. PURPOSE / BACKGROUND

- 3D ultrasound provides rich volumetric information, but most clinical interpretation still relies on task-specific measurements and manual review.
- Self-supervised foundation models can learn reusable representations directly from large collections of unlabeled clinical images.
- Routine 3DUS is particularly attractive for foundation-model development because large numbers of volumes are already acquired during clinical care and each volume contains a vast amount of imaging data.

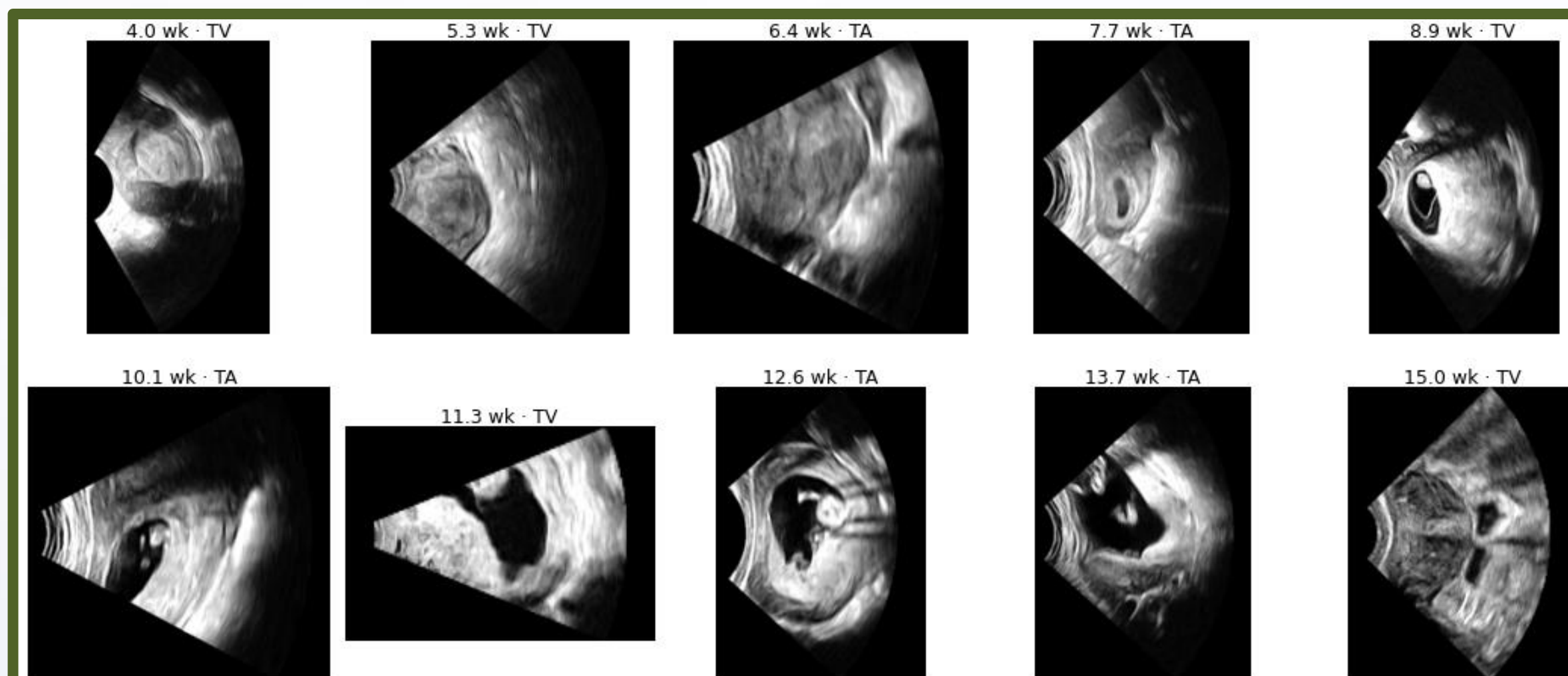
Goal: Develop a large-scale foundation model for routine clinical 3D ultrasound and determine whether its learned representations support gestational age estimation on temporally held-out data.

2. DATA & METHODS

- 300,092 clinical 3DUS volumes from 50,682 patients
- Acquired during routine examinations on Voluson systems (GE)
- 82,026 transabdominal volumes (27%)
- 73% transvaginal acquisitions
- All available volumes included without task-specific filtering
- Trained self-supervised 3D masked autoencoder
- 82% development: acquired before September 1, 2024
- 18% independent test: acquired after September 1, 2024
- Gestational age estimation evaluated in pregnant women from the temporally held-out test cohort

3. RESULTS

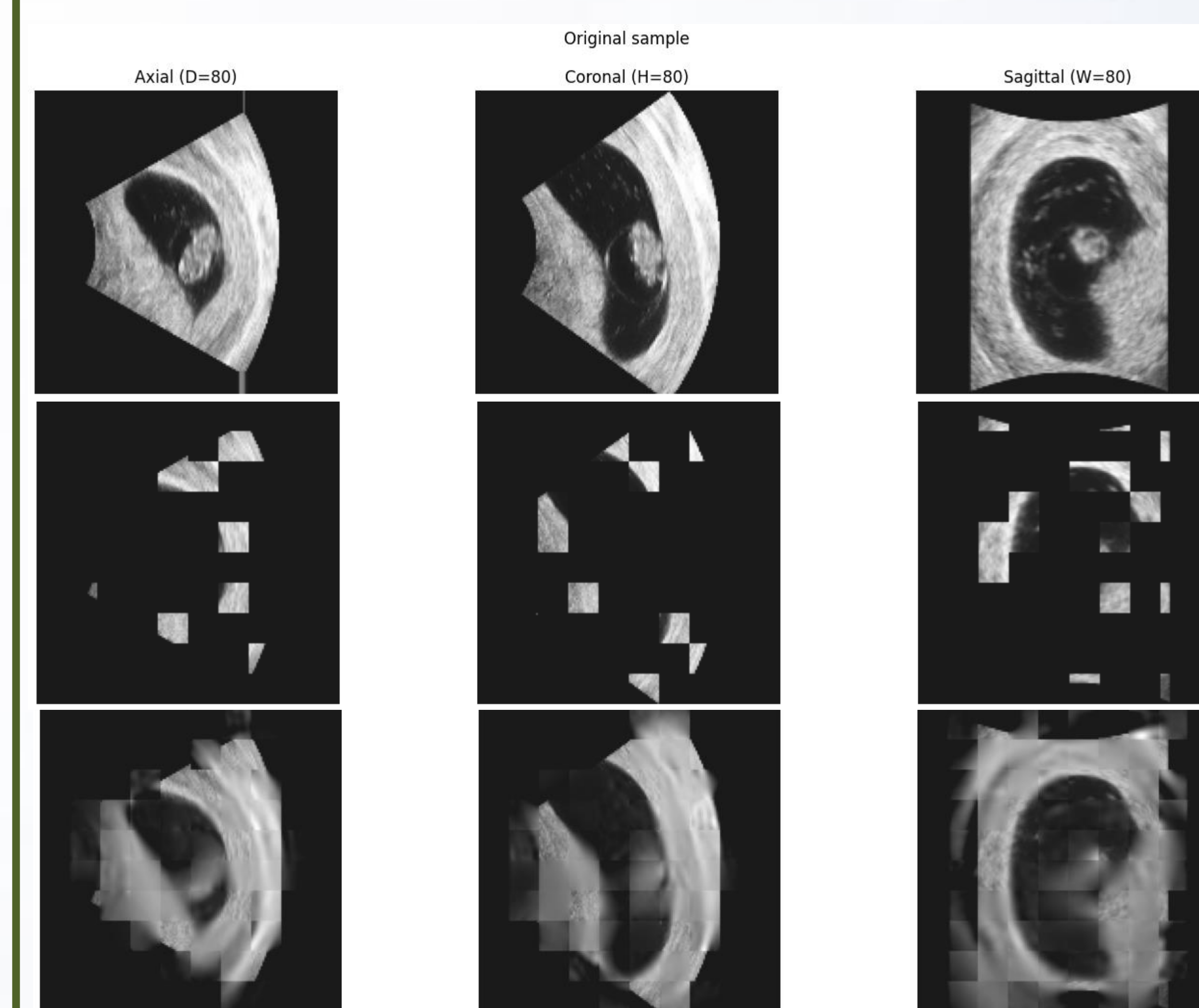
- The foundation model learned representations containing clinically meaningful information about pregnancy development and gestational age.
- Predicted gestational age showed strong agreement with reference gestational age in the temporally held-out cohort.
- Mean absolute error: 0.71 weeks 95% CI: 0.68–0.76 weeks
- Performance on examinations acquired after the training period demonstrates temporal generalization of the learned representation.



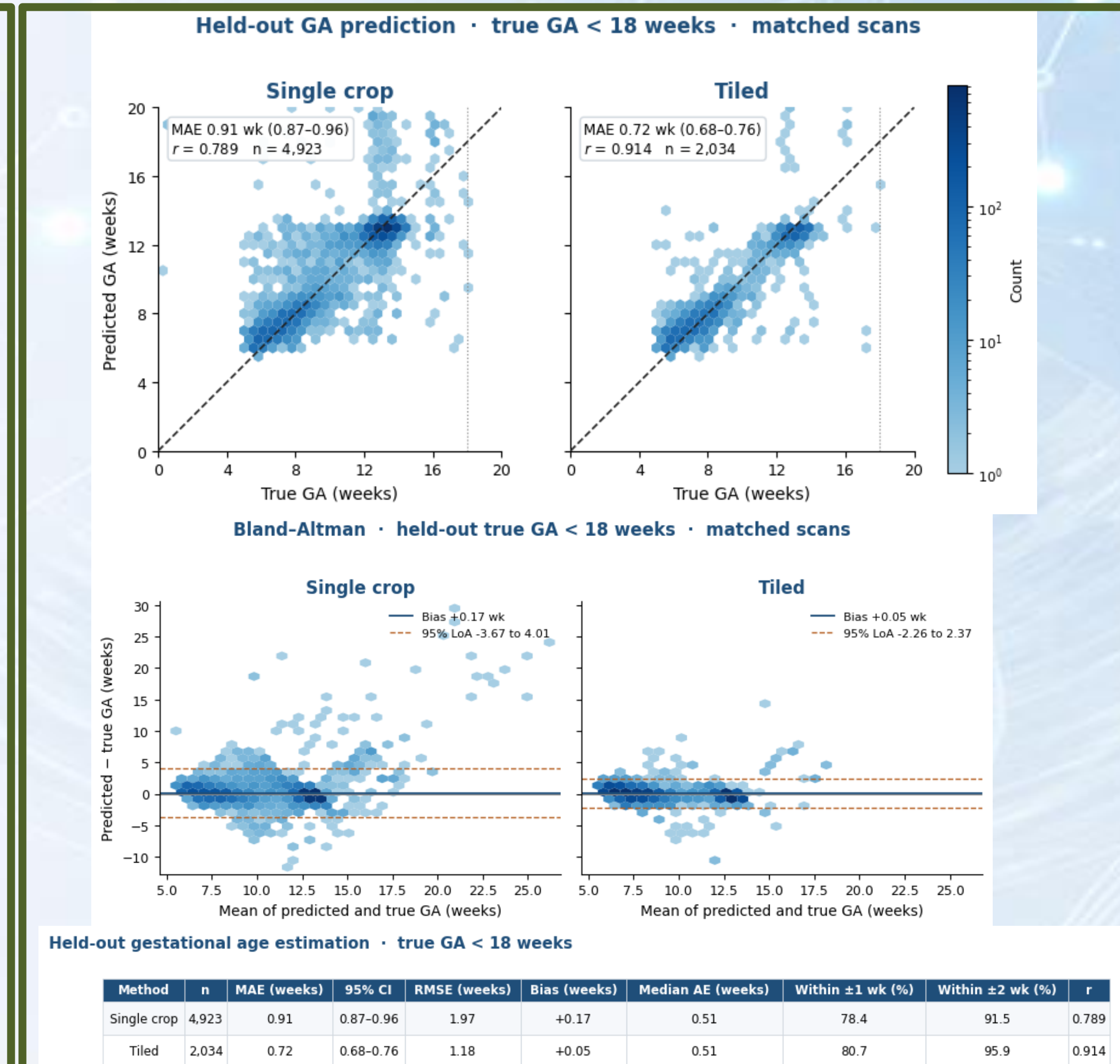
Representative routine clinical 3D ultrasound volumes across early gestation. Example images spanning 4.0–15.0 weeks of gestation demonstrate the substantial variation in fetal development, anatomy, field of view, and acquisition geometry present in the training data. Both transvaginal (TV) and transabdominal (TA) examinations were included, reflecting the heterogeneity of routine clinical 3D ultrasound.

4. CONCLUSIONS

Large-scale self-supervised learning can extract generalizable volumetric representations from routine clinical 3D ultrasound. These representations support clinically meaningful downstream analysis, demonstrated here using gestational age estimation. The results provide a foundation for extending the model beyond a single task toward automated interpretation of routine 3DUS. Future applications include automated biometry, anatomical assessment, detection of abnormal fetal or placental development, and automatic report generation.



Self-supervised masked autoencoder training example. Representative axial, coronal, and sagittal views from a clinical 3D ultrasound volume are shown before masking (top), after random masking of image patches (middle), and following reconstruction by the 3D masked autoencoder (bottom). The model is trained to recover the missing volumetric information from the visible image context, encouraging learning of generalizable 3D anatomical representations.



Predicted versus reference gestational age for scans with true GA <18 weeks using a single-crop approach (left) and tiled inference across the 3D volume (right). Tiled inference showed stronger agreement with reference GA and lower error (MAE 0.72 weeks, $r = 0.914$) than single-crop inference (MAE 0.91 weeks, $r = 0.789$). Bland–Altman analysis (bottom) demonstrated reduced bias and narrower limits of agreement with tiled inference.

KEY FINDINGS

- ★ A foundation model can be trained directly on hundreds of thousands of routine clinical 3D ultrasound volumes.
- ★ No task-specific filtering of the training dataset was required.
- ★ The learned representation encoded clinically meaningful information sufficient for accurate gestational age estimation.
- ★ Performance generalized to examinations acquired in a later, temporally independent cohort.