SonoCLIP: Mask-Guided Region-Aware Vision–Language Pretraining for Fetal Ultrasound Analysis

Hang Su[1⋆], Chao Sun[1,2⋆], Zhaofan Li[1], Wei Hu[3], Juhua Liu[1,2,4,5(✉)], and Bo Du[1,2,4,5(✉)]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Institute of Artificial Intelligence, Wuhan University, Wuhan, China
[3] Department of Ultrasound, Renmin Hospital of Wuhan University, Wuhan, China
[4] National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
[5] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China
{liujuhua@whu.edu.cn, dubo@whu.edu.cn}

Abstract. Vision–language foundation models have shown strong potential in medical image analysis. Although foundation models for ultrasound imaging have recently emerged, the domain is particularly challenging due to severe speckle noise, acquisition variability, and subtle anatomical boundaries, leading to high inter-observer variability. Existing CLIP-based models rely primarily on global image–text alignment, limiting their sensitivity to clinically decisive local structures. We propose SonoCLIP, the first million-scale region-controllable fetal ultrasound vision-language foundation model that integrates segmentation masks as mask-channel visual prompts within an encoder, enabling joint global–local contrastive representation learning. To support scalable region–text alignment, we introduce a sigmoid-based pairwise contrastive loss that improves stability under large-scale supervision. We further curate a 1.44M-image multimodal fetal ultrasound dataset spanning 24 standard planes for large-scale pretraining. Extensive cross-center evaluations demonstrate that SonoCLIP achieves superior zero-shot transfer performance under both global and mask-guided inference, establishing a controllable and clinically oriented foundation model for fetal ultrasound analysis. Our code and data are available at: https://github.com/Harrison-one/SonoCLIP.

Keywords: Fetal Ultrasound · Vision–Language Foundation Model · Region-Controllable Learning · Contrastive Alignment.

1 Introduction

Prenatal care has been significantly advanced by ultrasound technology. Due to its safety, accessibility, and cost-effectiveness, fetal ultrasound is routinely used

⋆ Equal contribution. (✉) Corresponding authors.

for real-time monitoring of fetal development and early detection of congenital abnormalities [2,8,16]. However, ultrasound interpretation remains subjective and operator-dependent, leading to substantial inter-observer variability. Image quality is often degraded by speckle noise and view-dependent artifacts. As a result, less-experienced clinicians may fail to identify subtle anatomical structures [17,5]. This limitation causes inconsistent fetal biometry measurements and hinders standardized assessment in clinical practice [3]. The problem is more severe in resource-limited settings, where experienced sonographers are scarce. These challenges highlight the urgent need for robust and trustworthy artificial intelligence methods to enhance diagnostic consistency and reliability in fetal ultrasound imaging [11,6,20].

In recent years, foundation models (FMs) trained with self-supervised learning have achieved notable success and advanced medical artificial intelligence [12,13,19]. By learning generic representations from large-scale unlabeled data, FMs can be adapted to diverse downstream tasks with minimal annotation. This paradigm outperforms traditional task-specific models that require extensive retraining. Vision–language models such as CLIP align visual and textual features through contrastive learning and demonstrate strong transferability [15,9]. Recent extensions to biomedicine further validate this potential [22,10,14]. However, a widely accepted foundation model for fetal ultrasound is still lacking. Existing approaches face several limitations. First, most methods adopt generic visual backbones [7] and rely on relatively small ultrasound datasets, which constrain robust fine-tuning. Models such as CLIP, pretrained on natural images, lack domain-specific anatomical knowledge. Although UniMed-CLIP [10] leverages large-scale multimodal medical data, its modality-specific performance remains inconsistent and underexplored for fetal ultrasound. Second, fetal view recognition and cross-center generalization require sensitivity to subtle local structures, such as small chambers and fine boundaries. These cues are often degraded by speckle noise and acquisition variability. Global image–text alignment captures coarse semantics but may overlook clinically critical details. Moreover, FetalCLIP [14] does not explicitly address the localization–context trade-off in fetal imaging, nor does it incorporate architectural designs tailored to the spatial hierarchy and morphological complexity of fetal anatomy.

We propose **SonoCLIP**, a foundation model for fetal ultrasound that introduces explicit region controllability during large-scale contrastive fine-tuning. Segmentation annotations are treated as visual prompts. Specifically, masks are appended as an additional channel to the image encoder, enabling region-guided representation learning while preserving global context. To effectively leverage both global- and region-level image–text pairs, we replace the conventional softmax-based contrastive objective with a sigmoid pairwise alignment loss. This formulation optimizes each pair independently and reduces sensitivity to batch composition. SonoCLIP is trained on a million-scale dataset covering 24 standard fetal ultrasound planes. We evaluate cross-center zero-shot plane classification under both global and mask-guided inference.

Our main contributions are summarized as follows:

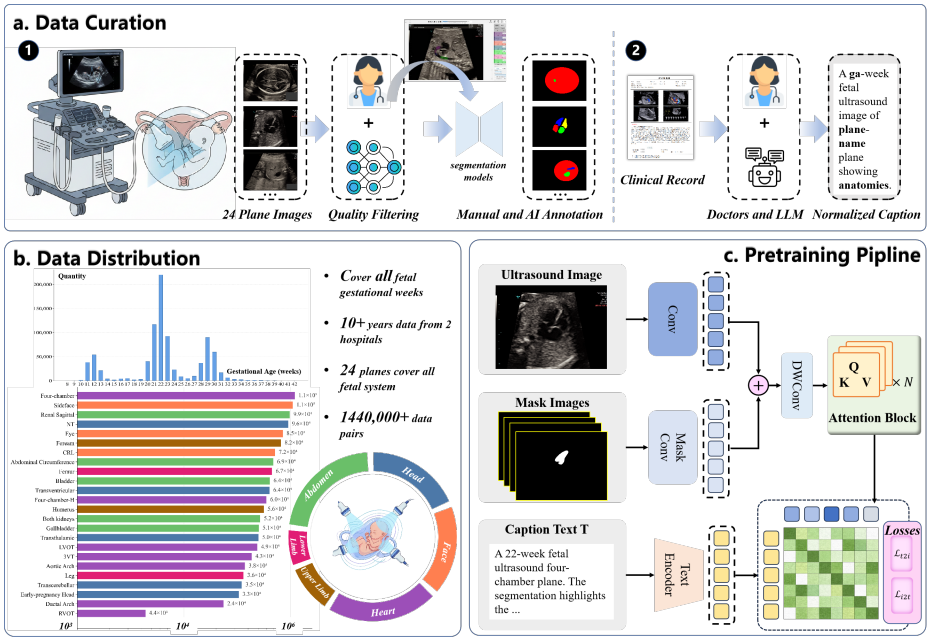


Fig. 1. Overview of SonoCLIP. (a) Data curation pipeline, including quality control, manual and AI-assisted mask annotation, and standardized caption generation from clinical records. (b) Dataset distribution across gestational weeks, hospitals, and anatomical planes. (c) Pre-training framework: ultrasound images and masks are jointly encoded via a mask-channel visual pathway, and aligned with paired text using a sigmoid-based objective.

- We propose a region-controllable fetal ultrasound foundation model, SonoCLIP, that incorporates mask-channel prompts to enable region-aware contrastive representation learning.
- We introduce a sigmoid-based pairwise contrastive loss to enhance stability and scalability for large-scale region–text fine-tuning.
- We curate and utilize a **1.44M**-image multimodal fetal ultrasound dataset with plane labels, segmentation annotations, and structured captions, and demonstrate strong cross-center zero-shot transfer on **24** plane classes with both global and mask-guided inference.

2 Method

2.1 Data Curation and Region–Text Pair Generation

AI-assisted Mask Extraction and Curation As illustrated in Fig. 1(a), we first apply a pretrained segmentation model to each ultrasound image to generate candidate anatomical masks with predicted labels and confidence scores.

To address fragmented predictions or duplicate instances under the same label, we perform label-wise consolidation. For multiple disconnected regions with identical labels, only the largest connected component is retained. We further filter masks using confidence thresholds and basic geometric constraints, such as non-empty foreground and plausible spatial extent, to remove unreliable predictions. This procedure produces a curated set of structural masks for each image, reducing noise in region-level supervision.

Region–Text Pair Generation for Fetal Ultrasound Following mask filtering, we construct region-text pairs for training. For each filtered image-mask pair, we employ a large language model and templates to generate structure-specific descriptions, exemplified by:

> *A **ga**-week fetal ultrasound **plane-name** plane. The segmentation highlights the **anatomy**.*

To retain plane-level semantics, we generate a global description for each image:

> *A **ga**-week fetal ultrasound image of the **plane-name** plane showing **anatomies**.*

During training, global image–text pairs are combined with mask-based region–text pairs. This strategy enables joint learning of plane-level recognition and anatomically grounded alignment at scale.

2.2 Mask-channel Visual Pathway

Inspired by Alpha-CLIP [18], we introduce a mask-channel visual pathway to jointly encode each ultrasound image and its corresponding anatomical mask. This design injects region-specific cues into the visual encoder while retaining the global contextual representations of the original CLIP image branch. Given a pseudo-color ultrasound image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ and a binary anatomical mask $\mathbf{A} = \mathbf{M} \in \{0,1\}^{H\times W}$, where 1 denotes the foreground and 0 the background, we process them through two parallel convolutional stems and fuse their features at the embedding level:

$$\mathbf{E} = \mathrm{DWConv}(\mathrm{Conv}(\mathbf{X}) + \mathrm{Conv}_m(\mathbf{A})), \tag{1}$$

where DWConv denotes depthwise convolution, Conv is the original image convolution, and Conv_m is the mask-channel convolution with the same kernel size and stride. As a result, both branches produce embeddings with identical spatial resolution and feature dimension.

To preserve the pretrained CLIP behavior at initialization, we initialize the Mask Conv kernel weights to zero, which ensures $\mathrm{Conv}_m(\mathbf{A}) = 0$ initially and thus $\mathbf{E} = \mathrm{Conv}(\mathbf{X})$. During fine-tuning, Conv_m learns to inject spatially localised cues into the early representations based on mask prompts, thereby achieving region-specific encoding while preserving the prior information embedded in the pre-trained original path.

2.3 Sigmoid Pairwise Contrastive Loss

Our goal is to learn a joint embedding space where prompted ultrasound representations align with their matched texts at both global and structural levels, while remaining distinct from unmatched texts within the same mini-batch. The supervision is heterogeneous, combining global captions and mask-based region descriptions. In addition, the number of implicit negatives increases with batch size. Under this setting, softmax-normalized objectives such as InfoNCE couple all samples through batch-level normalization and are sensitive to batch composition. Inspired by SigLIP [21], we adopt a pairwise sigmoid alignment loss. Each image–text pair is optimized independently as a binary matching objective, while still benefiting from large numbers of implicit negatives.

Let v_i and p_i be L2-normalized image and text embeddings:

$$v_i = \frac{f_\theta(\hat{x}_i)}{\|f_\theta(\hat{x}_i)\|}, \quad p_j = \frac{g(\hat{l}_j)}{\|g(\hat{l}_j)\|}. \tag{2}$$

We compute pairwise logits:

$$s_{ij} = t \cdot \langle v_i, t_j \rangle + b, \tag{3}$$

where t is a learnable logit scale and b is a learnable bias. For a matched pair we set $y_{ij} = +1$, otherwise $y_{ij} = -1$. The image-to-text loss is defined as:

$$\mathcal{L}_{i2t} = \frac{1}{n} \sum_i \sum_j \log(1 + \exp(-y_{ij} s_{ij})) \,, \tag{4}$$

and we analogously compute a text-to-image loss $\mathcal{L}_{t2i}$ by swapping the roles of image and text features. Final alignment loss is the symmetric average:

$$\mathcal{L}_{sig} = \frac{1}{2} \left(\mathcal{L}_{i2t} + \mathcal{L}_{t2i} \right). \tag{5}$$

3 Experiments

3.1 Datasets and Experimental Details

FetalP24 Dataset: FetalP24 dataset(Center A) contains 1.44 million fetal ultrasound images, covering 24 standard anatomical planes. The dataset currently includes 36,482 patients, mainly acquired from GE Voluson E8/E10 systems. Each image is annotated with anatomical masks and gestational age (GA) labels. Textual descriptions of both global and structure-specific features were generated using templates. A FetalP24(Center B) test dataset comprises the same 24 categories, with approximately 100 images per category, for zero-shot evaluation under both global and mask-guided inference. Images from Centre B were never utilised during training.

Table 1. Zero-shot classification performance on the FetalP24 dataset (Center B). Best results are in **bold**; second-best are underlined.

Method	Top-1	Top-5
CLIP	10.52	29.59
UniMed-CLIP	16.69	50.34
FetalCLIP	39.78	83.25
SonoCLIP (w/o mask)	58.38	94.47
SonoCLIP (w/ mask)	**85.01**	**99.01**

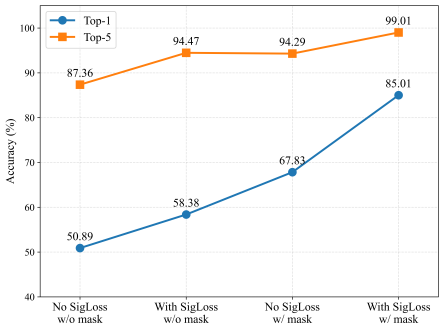


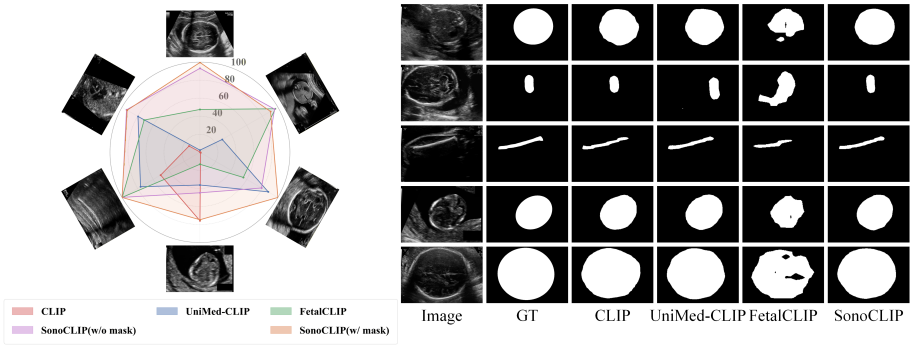
Fig. 2. Ablation study of SigLoss and mask guidance in zero-shot classification on the FetalP24 dataset.

FetalP6 Dataset: We further constructed a FetalP6 dataset comprising 5034 images by integrating two publicly available benchmark datasets: MFP[1] and FETAL-PLANES-DB[4]. Additionally, we refined the HC18 subset within MFP into early pregnancy head and corpus callosum plane categories. The final dataset comprises six standard planes: Abdominal Circumference(AC), Cerebellum(Cereb), Early-pregnancy Head(EarlyHead), femur, Four-chamber, and Transthalamic. Concurrently, we curated a segmentation FetalP5 dataset comprising five planes: Abdominal Circumference(AC), Cerebellum(Cereb), Early-pregnancy Head, Femur, and Transthalamic.

Experimental Setup. For SonoCLIP pre-training, we employed CLIP (ViT-L/14@336px) as the foundational architecture, training on four NVIDIA RTX 3090 GPUs with an effective batch size of 32. We employed the AdamW algorithm with a base learning rate of 1×10^{-4} and a weight decay rate of 0.02. The masking branch was optimised using the full learning rate, while the remaining visual encoder parameters, learnable log probability scaling factors, and biases were trained with a learning rate multiplier of $0.01\times$. The learning rate was scheduled via 5000 warm-up steps followed by cosine decay. To maintain global semantic alignment, 10% of training pairs employed all-ones masks and full descriptions. Here, w/ mask and w/o mask denote inference with anatomical masks and all-ones masks, respectively. For downstream evaluation, we adopt a frozen-encoder protocol for both classification and segmentation. For classification, we train only a lightweight linear head. For segmentation, we optimize a lightweight decoder on top of dense patch-level features. To prevent label leakage, segmentation uses an all-ones mask-channel input rather than the ground-truth mask.

3.2 Zero-shot Classification on the FetalP24 Dataset

As shown in Table 1, CLIP and UniMed-CLIP achieve relatively low performance, indicating limited transferability to fetal ultrasound. FetalCLIP performs



(a) (b)

Fig. 3. (a) Zero-shot classification results of each model on the FetalP6 dataset. (b) Qualitative comparison of segmentation results on the FetalP5 dataset.

substantially better, highlighting the importance of domain adaptation. SonoCLIP (w/o mask) further improves over FetalCLIP by 18.60 and 11.22 percentage points in Top-1 and Top-5 accuracy, respectively, showing that the proposed training framework learns more transferable representations under mixed global and region-level supervision. With mask-guided inference, SonoCLIP (w/ mask) gains a further 26.63% and 4.54% over SonoCLIP (w/o mask), indicating that mask guidance provides additional benefits by emphasizing clinically relevant anatomical regions during zero-shot matching.

3.3 Ablation Studies

Fig. 2 presents ablation results for SigLoss and mask-guided learning in zero-shot classification. When mask-guided learning is disabled, incorporating SigLoss alone yields a modest improvement in model performance. However, enabling both mask-guided learning and SigLoss simultaneously results in a significant increase in both Top-1 and Top-5 performance. This indicates that SigLoss's advantages become more pronounced when combined with mask guidance, and also demonstrates a strong synergistic effect between paired alignment objectives and mask-assisted inference.

3.4 Zero-shot Classification on the FetalP6 Dataset

Fig. 3(a) reports zero-shot classification on the processed FetalP6 dataset, with no training or fine-tuning on this dataset. SonoCLIP without masks (w/o mask) performs strongly across representative planes and outperforms baseline models in overall balance, indicating improved cross-dataset generalization. Mask-assisted (w/ mask) inference further boosts accuracy on several planes, suggesting additional gains from region guidance.

Table 2. Linear-probe classification performance on the FetalP6 dataset."w/" and "w/o" denote with and without the mask-channel pathway, respectively. Best results are highlighted in **bold**, and second-best results are underlined.

Model	AC		Cereb		Earlyhead		Femur		Four-chamber		Transthalamic		Avg	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CLIP	89.3	87.4	50.7	58.0	91.4	95.5	97.6	96.2	90.7	92.6	89.7	84.9	87.1	85.8
UniMed-CLIP	91.6	87.7	69.1	57.7	**100.0**	87.5	98.6	97.9	88.1	91.1	66.9	74.6	83.5	82.7
FetalCLIP	97.3	97.8	80.1	80.7	77.1	85.7	**100.0**	99.3	99.1	99.1	92.3	91.1	94.6	92.3
SonoCLIP(w/o)	**98.2**	**98.9**	83.1	85.0	94.3	**95.7**	**100.0**	99.5	**100.0**	99.9	94.2	93.2	96.3	95.3
SonoCLIP(w/)	**98.2**	98.7	**100.0**	**100.0**	91.4	95.5	**100.0**	**99.8**	**100.0**	**100.0**	**99.4**	**98.7**	**99.3**	**98.8**

3.5 Linear Probe Classification on the FetalP6 Dataset

Table 2 reports linear-probe classification results. SonoCLIP achieves the best overall performance among all methods, improving mean accuracy and F1 by 3.0% and 3.5% over FetalCLIP, respectively. Notably, this method achieves top results across most planes, demonstrating consistent robust classification capabilities across multi-class fetal ultrasound planes. This also indicates that the proposed pre-training strategy can learn more discriminative and robust fetal ultrasound representations, while mask-guided techniques further enhance downstream classification performance.

Table 3. Segmentation performance on the FetalP5 dataset. "w/o" denotes without the mask-channel pathway. Best results are highlighted in **bold**, and second-best results are underlined.

Model	AC		Cereb		Earlyhead		Femur		Transthalamic		Avg	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
CLIP	93.7	88.8	69.0	58.2	91.3	86.2	74.6	61.3	92.3	87.7	85.1	77.5
UniMed-CLIP	**94.3**	89.7	62.9	53.1	**94.2**	**89.6**	77.9	65.1	**96.0**	**93.0**	86.5	79.8
FetalCLIP	83.4	73.7	21.1	11.9	77.9	67.0	54.3	40.2	90.8	84.3	69.8	60.2
SonoCLIP(w/o)	**94.3**	**90.0**	**74.1**	**64.8**	91.5	86.5	**78.1**	**65.6**	93.2	89.6	**87.2**	**80.5**

3.6 Segmentation on the FetalP5 Dataset

Table 3 demonstrates that SonoCLIP achieves the best overall segmentation performance, with a Dice score of 87.2% and an mIoU of 80.5%. All segmentation results were obtained using the same frozen-encoder and lightweight-decoder setting. The qualitative results in Fig. 3(b) further demonstrate that the generated mask predictions are clearer and more complete. These findings indicate that the proposed pre-training strategy generates stronger dense visual representations, which effectively transfer to downstream tasks in fetal ultrasound segmentation.

4 Conclusions

We present SonoCLIP, a vision–language foundation model for fetal ultrasound. By introducing a mask-channel visual pathway and a sigmoid pairwise contrastive loss, SonoCLIP learns region-aware representations robust to heterogeneous supervision and large-scale implicit negatives. Extensive experiments show consistent gains in cross-center zero-shot transfer and cross-dataset generalization, as well as strong performance under frozen-encoder classification and segmentation. These results indicate that SonoCLIP can serve as a transferable backbone for fetal ultrasound applications, including standardized plane recognition and anatomy-aware analysis, with potential to support reliable assessment in routine and resource-limited clinical settings.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC2705705, in part by the National Natural Science Foundation of China under Grants 62225113 and U23B2048, in part by the Innovative Research Group Project of Hubei Province under Grants 2024AFA017, in part by the Science and Technology Major Project of Hubei Province under Grants 2024BAB046 and 2025BCB026, and in part by the New Cornerstone Science Foundation through the XPLORER PRIZE. This work was also supported by WHU-Kingsoft Joint Lab. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

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

References

1. Ashkani Chenarlogh, V., Ghelich Oghli, M., Shabanzadeh, A., Sirjani, N., Akhavan, A., Shiri, I., Arabi, H., Sanei Taheri, M., Tarzamni, M.K.: Fast and accurate u-net model for fetal ultrasound image segmentation. Ultrasonic imaging **44**(1), 25–38 (2022)
2. Avola, D., Cinque, L., Fagioli, A., Foresti, G., Mecca, A.: Ultrasound medical imaging techniques: A survey. ACM Comput. Surv. **54**(3) (Apr 2021)
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**, 10200 (2020)
4. 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)
5. Guo, X., Alsharid, M., Zhao, H., Wang, Y., Lander, J., Papageorghiou, A.T., Noble, J.A.: A visually grounded language model for fetal ultrasound understanding. Nature Biomedical Engineering (2026), published online 15 Jan 2026

6. He, J., Yang, L., Liang, B., Li, S., Xu, C.: Fetal cardiac ultrasound standard section detection model based on multitask learning and mixed attention mechanism. Neurocomputing **579**, 127443 (2024)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
8. Khalil, A., Sotiriadis, A., D'Antonio, F., Da, Silva Costa, F., Odibo, A., Prefumo, F., Papageorghiou, A.T., Salomon, L.J.: Isuog practice guidelines: performance of third-trimester obstetric ultrasound scan. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology p. 63 (2024)
9. Khan, W., Leem, S., See, K.B., Wong, J.K., Zhang, S., Fang, R.: A comprehensive survey of foundation models in medicine. IEEE Reviews in Biomedical Engineering **19**, 283–304 (2026)
10. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities (2024)
11. Krishna, T.B., Kokil, P.: Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. Expert Systems with Applications **238**, 122153 (2024)
12. Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open AI foundation model applied to chest radiography. Nature **643**, 488–498 (2025)
13. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., Yan, F., Cai, Y., Zhu, Z., Jin, C., Lin, Y., Jiang, X., Zhao, C., Li, D., Han, A., Li, Z., Chan, R.C.K., Wang, J., Fei, P., Cheng, K.T., Zhang, S., Li, L.L., Chen, H.: A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nature Biomedical Engineering (2025), published online 02 Sep 2025
14. Maani, F., Saeed, N., Saleem, T., Farooq, Z., Alasmawi, H., Diehl, W., Mohammad, A., Waring, G., Valappi, S., Bricker, L., Yaqub, M.: Fetalclip: A visual-language foundation model for fetal ultrasound image analysis (2025)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
16. Salomon, L.J., Alfirevic, Z., Bilardo, C.M., Chalouhi, G.E., In, O.I.S.O.U.: Isuog practice guidelines: performance of first-trimester fetal ultrasound scan (vol 41, pg 102, 2013). Ultrasound in Obstetrics and Gynecology **41**(2), 102–113 (2013)
17. Singh, R., Gupta, S., Mohamed, H.G., Bharany, S., Rehman, A.U., Ghadi, Y.Y., Hussen, S.: Advancing prenatal healthcare by explainable AI enhanced fetal ultrasound image segmentation using U-Net++ with attention mechanisms. Scientific Reports **15**, 19612 (2025)
18. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13019–13029 (June 2024)
19. Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., Tang, V., Ng, A.B., Powell, D., Bonnington, P., See, S., Magnaterra, E., Ferguson, P., Nguyen, J., Guitera, P., Banuls, J., Janda, M., Mar, V., Kittler, H., Soyer, H.P., Ge, Z.: A multimodal vision foundation model for clinical dermatology. Nature Medicine **31**, 2691–2702 (2025)

20. Yeung, P.H., Hesse, L.S., Aliasi, M., Haak, M.C., 21st Consortium, I., Xie, W., Namburete, A.I.L.: Sensorless volumetric reconstruction of fetal brain freehand ultrasound scans with deep implicit representation. Medical Image Analysis **94**, 103147 (2024)
21. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11941–11952 (2023)
22. Zhang, S., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2023)