

FetalAgents: A Multi-Agent System for Fetal Ultrasound Image and Video Analysis

Xiaotian Hu^{1,*}, Junwei Huang^{1,*}, Mingxuan Liu^{1,*}, Kasidit Anmahapong¹,
Yifei Chen¹, Yitong Luo¹, Yiming Huang², Xuguang Bai¹, Zihan Li¹,
Yi Liao³, Haibo Qu³, and Qiyuan Tian^{1,✉}

¹ Tsinghua University, Beijing, China

² University of California San Diego, La Jolla, CA, USA

³ West China Second University Hospital, Sichuan University, Chengdu, China

qiyuantian@tsinghua.edu.cn

Abstract. Fetal ultrasound (US) is the primary imaging modality for prenatal screening, yet its interpretation relies heavily on the expertise of the clinician. Despite advances in deep learning and foundation models, existing automated tools for fetal US analysis struggle to balance task-specific accuracy with the whole-process versatility required to support end-to-end clinical workflows. To address these limitations, we propose **FetalAgents**, the first multi-agent system for comprehensive fetal US analysis. Through a lightweight, agentic coordination framework, **FetalAgents** dynamically orchestrates specialized vision experts to maximize performance across diagnosis, measurement, and segmentation. Furthermore, **FetalAgents** advances beyond static image analysis by supporting end-to-end video stream summarization, where keyframes are automatically identified across multiple anatomical planes, analyzed by coordinated experts, and synthesized with patient metadata into a structured clinical report. Extensive multi-center external evaluations across eight clinical tasks demonstrate that **FetalAgents** consistently delivers the most robust and accurate performance when compared against specialized models and multimodal large language models (MLLMs), ultimately providing an auditable, workflow-aligned solution for fetal ultrasound analysis and reporting. Source code is available at: <https://github.com/birthlab/FetalAgents>.

Keywords: Fetal Ultrasound · Medical Agents · Video Summarization.

1 Introduction

Fetal ultrasound (US) is the cornerstone imaging modality for routine prenatal screening and structural anomaly assessment [32]. However, its interpretation relies heavily on examiner expertise, and the scarcity of experienced sonographers

* These authors contributed equally.

✉ Corresponding author.

further amplifies disparities in care quality [36]. To overcome these limitations, deep learning has been extensively explored to automate and standardize fetal US analysis [10]. Specifically, task-specific architectures, ranging from residual networks [14] and vision transformers [8] for plane classification and gestational age (GA) regression, to U-Net [31] variants [18] and SAM-based adaptations [21,24,45] for anatomical segmentation and biometry, have established highly accurate automated baselines across individual fetal US subtasks. More recently, foundation models such as FetalCLIP [28] and USFM [19] leverage large-scale pre-training to extract versatile representations generalizable across diverse downstream clinical tasks. However, most automated models remain narrowly specialized and workflow-fragmented, typically optimized for a single subtask and deployed as standalone predictors that leave clinicians to manually coordinate tool selection and translate predictions into structured documentation [10,42].

Recent advances in agentic systems offer a promising approach to overcoming these limitations. By augmenting large language models (LLMs) with planning, memory, and tool-use capabilities [17], agentic systems can orchestrate specialized models to execute multi-step workflows, as demonstrated by multi-disciplinary consultation agents [20,5], longitudinal disease-tracking systems [23,44], and multimodal medical imaging frameworks [9,27,37]. However, no existing agentic system addresses the unique demands of fetal US, where clinical reasoning spans multiple anatomical planes, tightly couples classification, segmentation, and biometry, and must handle temporal video streams beyond static images.

To bridge this gap, we present **FetalAgents**, the first multi-agent system for fetal US analysis. **FetalAgents** orchestrates a diverse suite of domain-specific models through a lightweight, language-driven coordination framework that dynamically interprets clinical intent, identifies anatomical context, and dispatches tasks to dedicated expert modules, thereby emulating the holistic reasoning process of human sonographers. Furthermore, recognizing that real-world fetal US examination is a temporal and continuous process, **FetalAgents** extends beyond static image analysis to support end-to-end video stream summarization. Specifically, an automated keyframe extraction module distills continuous scans into diagnostically relevant frames, which are analyzed by coordinated experts and synthesized into structured clinical reports. Extensive experiments on multi-center external datasets confirm that **FetalAgents** consistently achieves the best performance across eight clinical tasks when compared with specialized vision models and MLLMs, validating its potential to standardize and automate real-world sonographic workflows.

2 Methods

2.1 Overview of FetalAgents

FetalAgents is a collaborative, multi-agent system designed for comprehensive fetal ultrasound (US) analysis. Built upon the AutoGen [40] framework, **FetalAgents** integrates specialized vision tools with reasoning LLM agents, bridging the gap between isolated point-prediction models and comprehensive,

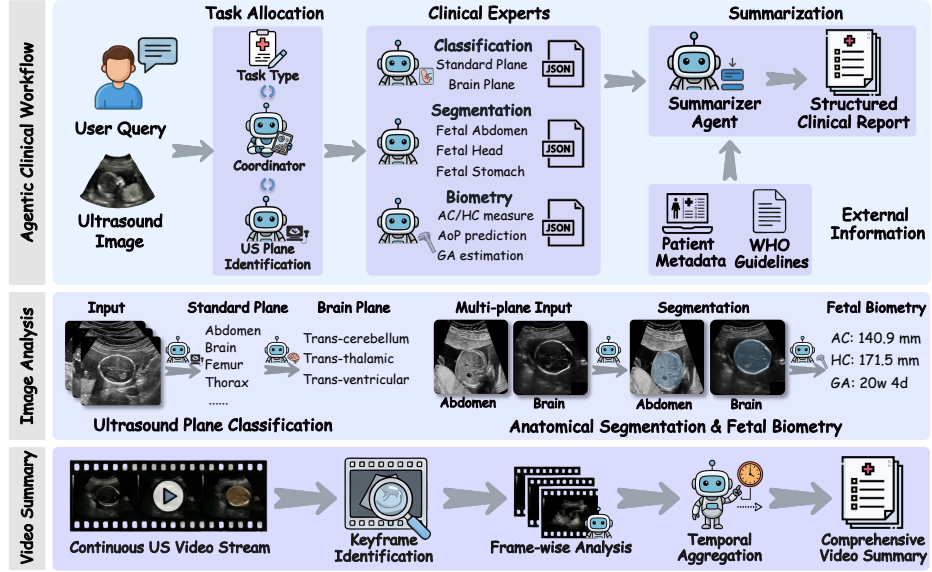


Fig. 1. System diagram of FetalAgents. The framework comprises three distinct modules: (1) Agentic Clinical Workflow, where the Coordinator interprets queries and dispatches tasks; (2) Image Analysis, where specialized experts execute specific vision tasks (classification, segmentation, biometry); and (3) Video Summary, which aggregates multi-plane keyframes and temporal findings into a comprehensive report.

real-life clinical workflows. To provide stable and reliable fetal US analysis while minimizing hallucination risks, **FetalAgents** distributes distinct responsibilities across three specialized agent types: Coordinator, Experts, and Summarizer:

(1) **Coordinator Agent.** Based on the GPT-5-mini [30], the Coordinator acts as an intelligent central orchestrator of the entire system. Given an input US image I (or a video stream V) and a user query Q , the Coordinator first classifies the query into a task type τ (specific clinical tasks or report generation), and identifies the standard ultrasound plane c . Based on the combined information (Q, τ, c) , it rephrases the query into a structured prompt P , and dynamically dispatches (I, P) to a subset of relevant specialized experts.

(2) **Expert Agents.** Modular Expert Agents are dedicated to addressing specific clinical tasks. Each expert $\mathcal{E}_k$ wraps a set of task-specific vision models $\mathcal{M}_k = \{m_{k,1}, m_{k,2}, \dots, m_{k,N}\}$ as callable tools and applies deterministic fusion rules $\mathcal{F}_k$ to generate robust results $\hat{o}_k = \mathcal{F}(m_{k,1}(I), \dots, m_{k,N}(I))$ across multiple outputs. Crucially, experts communicate upstream exclusively via structured JSON formats, eliminating subjective hallucinations from intermediate steps.

(3) **Summarizer Agent.** The Summarizer functions as a comprehensive report generator. It aggregates the structured outputs and tool-decision notes from the Experts to formulate a cohesive, human-readable report. By cross-referencing

biometric findings from related tasks, the Summarizer acts as a final clinical safeguard before formatting the final response.

2.2 Fetal Ultrasound Analysis

To support diverse clinical scenarios, **FetalAgents** is equipped with a robust toolset that addresses three primary categories of fetal US tasks:

Plane Classification. Accurate identification of standard fetal planes is a critical prerequisite for morphological diagnosis. To achieve this, **FetalAgents** integrates an ensemble of four complementary models $\mathcal{M}_{plane}$ to recognize standard fetal views and three brain sub-planes: (1) FetalCLIP [28], a vision-language foundation model capable of generating universal representations of fetal US images; (2) FU-LoRA [38], a Latent Diffusion Model fine-tuned via Low-Rank Adaptation (LoRA) [16]; (3) a ResNet-50 backbone initialized with RadImageNet [29] weights; and (4) a Vision Transformer (ViT-B/16) [8] initialized with ImageNet-1K [7] weights.

Anatomical Segmentation. Fine-grained segmentation of anatomical structures, such as the fetal head, abdomen, and stomach, provides crucial geometric priors for downstream biometry. For fetal head segmentation, the system ensembles nnU-Net [18], the Ultrasound Foundation Model (USFM) [19], and a lightweight CNN [43]. For abdominal segmentation, we utilize FetalCLIP to generate a coarse initial mask, subsequently extracting its geometric center and bounding box to serve as input prompts for SAMUS [24], an interactive, SAM-based US segmentation model. Finally, for the fetal stomach, we deploy a robust ensemble comprising nnU-Net, FetalCLIP, and SAMUS to ensure maximal segmentation fidelity.

Fetal Biometry. Fetal biometry is essential for tracking fetal development and guiding intrapartum care. To automatically measure the Angle of Progression (AoP), the system employs an ensemble of AoP-SAM [45] (a tailored SAM for pubic symphysis and fetal head segmentation), USFM, and an adapted UperNet [41] model. For precise GA estimation, we integrate FetalCLIP with dedicated regression heads fine-tuned using ResNet-50 and ConvNeXt-Tiny [25] backbones. Routine biometric indices, including Head Circumference (HC) and Abdominal Circumference (AC), are deterministically derived from the high-fidelity masks generated by the anatomical segmentation experts using ellipse fitting.

2.3 Comprehensive Clinical Workflow

Ultrasound scans are inherently multi-plane and temporal. Traditional models typically operate on a single static frame and a single task, leaving major workflow steps manual and operator-dependent [13]. **FetalAgents** mirrors real-world clinical workflows by extending beyond isolated tasks to support comprehensive image captioning and video stream summarization.

Image Caption Generation. When presented with a broad request, **FetalAgents** leverages its inherent agentic intelligence to synthesize findings. The Coordinator first identifies the anatomical plane and invokes the full suite of plane-appropriate

experts. The Summarizer then integrates these multi-faceted outputs $\mathcal{J}$ into a structured, comprehensive report template, ensuring all available findings are consolidated consistently. Additionally, during image caption generation, **FetalAgents** evaluates the biometric measurement (AC, HC) against the estimated GA by calculating the normative growth percentile p using the WHO fetal growth charts [22]. Importantly, the normative percentile doubles as an automated consistency safeguard. Measurements that deviate beyond plausible growth bounds prompt the Summarizer to invoke a reflection step over the expert outputs, dynamically replacing the outlier with a more trustworthy ensemble prediction to uphold reporting accuracy.

Video Summarization. Manually extracting keyframes and documenting frame-wise findings from US video streams $V = \{I_1, I_2, \dots, I_T\}$ is highly labor-intensive and operator-dependent. To automate this process, we fine-tuned a 6-class keyframe identification model utilizing the FetalCLIP encoder backbone. **FetalAgents** then deploys this model to automatically extract a subset of diagnostic keyframes $\mathcal{K} = \{I_{k_1}, \dots, I_{k_M}\} \subset V$, assigns appropriate Experts for frame-wise analysis, and finally, aggregates the temporal findings into a video-level clinical report. Through this end-to-end video summarization pipeline, **FetalAgents** effectively mirrors real-world sonographer workflows, offering significant adaptability over existing rigid video-summarization pipelines and seamlessly accommodating diverse clinical scenarios without extensive workflow-specific fine-tuning.

3 Experiments

3.1 Datasets and Experimental Setup

Training and external validation datasets. **FetalAgents** was trained using multiple public fetal US datasets for different task-specific purposes, including Fetal_PLANES_DB [3] (plane classification), ACOUSLIC-AI [33] and the dataset from [6] (abdomen and stomach segmentation), PSFHS [4] (AoP prediction), HC18 [15] (head circumference measurement), and PBF-US1 [11] (keyframe identification). **FetalAgents** and comparison methods were evaluated on several independent external datasets to assess out-of-distribution performance, including multi-center African data [35,39] (standard plane classification, HC measurement), datasets from [2] (brain plane classification, stomach segmentation), [1] (abdomen segmentation), [26] (AoP prediction), and a 511-case subset of HC18 [15] (GA estimation).

Baselines and Metrics. **FetalAgents** was compared against specialized vision models [38,14,8,18,45,41,43,25], US-specific foundation models [28,19], general MLLMs[30,12], and medical MLLMs[34]. For plane classification, we report Accuracy, Precision, Recall, F1-score, Cohen’s κ , and AUROC. For anatomical segmentation, we adopt the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), 95th-percentile Hausdorff Distance (HD95), Average Symmetric Surface Distance (ASSD), Positive Predictive Value (PPV), and Sensitivity (Sens). For fetal biometry, segmentation quality is again assessed with DSC and HD95, while measurement accuracy is quantified by Mean Absolute Error

Table 1. Quantitative evaluation on external datasets across eight clinical tasks grouped into three categories. **Bold:** best. Underline: runner-up.

I. Plane Classification						
Method	Accuracy	Precision	Recall	F1-score	Cohen’s κ	AUROC
<i>Standard Plane Classif. (Ext. Val.: 233 cases)</i>						
FetalCLIP[28]	0.914	0.942	0.915	0.927	0.887	99.7%
FU-LoRA[38]	0.764	0.977	0.759	0.847	0.717	98.0%
GPT-5-mini[30]	0.605	0.533	0.548	0.523	0.464	0.816
Gemini-3-flash[12]	0.880	0.894	0.843	0.856	0.857	0.961
MedGemma[34]	0.708	0.831	0.649	0.617	0.597	N/A
FetalAgents	0.927	0.937	0.927	0.931	0.903	99.7%
<i>Brain Plane Classif. (Ext. Val.: 569 cases)</i>						
FetalCLIP[28]	0.893	0.880	0.911	0.883	0.830	98.6%
ResNet[14,29]	0.824	0.798	0.853	0.810	0.727	94.7%
ViT-16[8,7]	0.881	0.864	0.896	0.874	0.810	96.5%
GPT-5-mini[30]	0.689	0.644	0.632	0.638	0.518	0.789
Gemini-3-flash[12]	0.636	0.657	0.661	0.606	0.473	0.755
MedGemma[34]	0.691	0.723	0.693	0.684	0.509	N/A
FetalAgents	0.909	0.884	0.925	0.896	0.856	96.3%
II. Anatomical Segmentation						
Method	DSC	IoU	HD95	ASSD	PPV	Sens
<i>Abdomen Seg. (Ext. Val.: 187 cases)</i>						
FetalCLIP[28]	0.902	0.834	4.20	0.447	0.891	0.940
nnUNet[18]	0.848	0.748	129.2	15.20	0.809	0.912
FetalAgents	0.937	0.937	2.32	0.202	0.922	0.955
<i>Stomach Seg. (Ext. Val.: 253 cases)</i>						
FetalCLIP[28]	0.771	0.653	52.56	8.62	0.829	0.764
SAMUS[24]	0.826	0.730	14.07	5.68	0.797	0.893
nnUNet[18]	0.847	0.754	11.30	1.95	0.841	0.891
FetalAgents	0.859	0.767	10.64	3.62	0.844	0.896
III. Fetal Biometry						
Method	DSC	HD95	MAE	MdAE	MdRAE	Acc@5%
<i>AoP Estimation (Ext. Val.: 64 cases)</i>						
AoP-SAM[45]	0.927	38.0	5.59°	4.86°	3.54%	65.6%
USFM[19]	0.870	68.3	10.03°	8.16°	6.46%	43.8%
UperNet[41]	0.913	42.9	6.74°	6.31°	4.77%	53.1%
FetalAgents	0.928	37.2	5.44°	4.68°	3.50%	67.2%
<i>AC Measurement (Ext. Val.: 187 cases)</i>						
FetalCLIP[28]	0.902	4.20	9.94 mm	6.59 mm	3.77%	65.6%
nnUNet[18]	0.848	129.2	11.7 mm	8.80 mm	5.28%	47.1%
FetalAgents	0.937	2.32	5.53 mm	4.29 mm	2.34%	84.0%
Method	DSC	HD95	MRAE	MdRAE	P95 RAE	Acc@5%
<i>HC Measurement (Ext. Val.: 75 cases)</i>						
CSM[43]	0.956	1.92	2.05%	1.48%	5.75%	89.3%
USFM[19]	0.924	11.54	3.21%	2.06%	13.0%	86.7%
nnUNet[18]	0.964	1.72	1.57%	1.21%	3.96%	98.7%
FetalAgents	0.961	1.93	1.40%	1.12%	3.75%	96.0%
Method	Valid.	MAE	MdAE	MRAE	P95 AE	Acc@5%
<i>GA Prediction (Ext. Val.: 511 cases)</i>						
ResNet[14,29]	71.62%	1.73w	1.08w	6.74%	5.41w	52.1%
FetalCLIP[28]	82.97%	1.32w	0.72w	5.13%	4.69w	68.9%
ConvNext[25]	74.17%	1.53w	0.96w	5.95%	5.19w	56.75%
FetalAgents	84.93%	1.23w	0.66w	4.70%	4.51w	68.7%

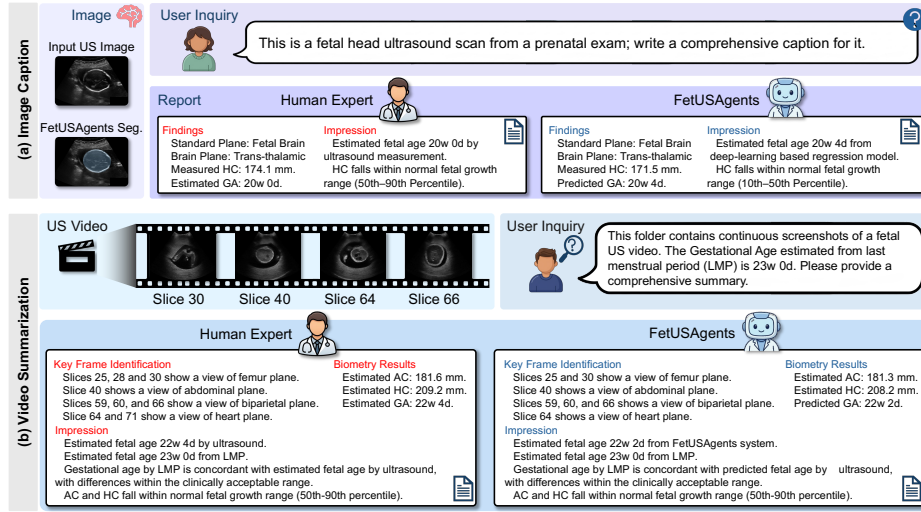


Fig. 2. Qualitative evaluation of clinical reporting. (a) Image Caption Generation: FetalAgents correctly identifies the standard plane and performs biometry to generate a report concordant with human experts. (b) Video Summarization: The system autonomously extracts multi-plane keyframes from a continuous video stream and cross-references ultrasound findings with patient metadata (LMP) for clinical consistency.

(MAE), Median Absolute Error (MdAE), Mean/Median Relative Absolute Error (MRAE/MdRAE), 95th-percentile Absolute/Relative Error (P95 AE/P95 RAE), and the percentage of predictions within 5% of the ground truth (Acc@5%). For gestational age (GA) prediction in particular, following [28], we introduce an additional Validity Rate metric, which tests whether the ground-truth HC falls within the 2.5th–97.5th percentile range of the predicted GA on the WHO fetal growth charts [22], thereby providing a clinically grounded sanity check on the plausibility of the estimated GA.

3.2 Quantitative Results

FetalAgents consistently achieves the most accurate and stable performance across 8 clinical tasks on external validation (Table 1), confirming the efficacy of agentic orchestration and dynamic model ensembling. Specifically, (1) Plane Classification: Generic and medical MLLMs struggled significantly with the acoustically noisy US modality, with GPT-5-mini achieving only 0.605 accuracy on standard planes. By dynamically routing images to specialized visual backbones, FetalAgents outperforms the strongest standalone model FetalCLIP, improving accuracy from 0.914 to 0.927 and Cohen’s κ from 0.887 to 0.903 on standard plane classification. On the more challenging brain sub-plane task, FetalAgents achieved F1-score 0.896 and κ 0.856, improving over FetalCLIP

by 1.3% and 2.6%, respectively. (2) Anatomical Segmentation: Our fusion rules effectively correct boundary errors typical of monolithic models. On abdomen segmentation, **FetalAgents** attained DSC 0.937 and HD95 2.32, improving over the runner-up FetalCLIP by 3.5% in DSC and 44.8% in HD95. For stomach segmentation, **FetalAgents** reached DSC 0.859 and HD95 10.64, surpassing nnUNet by 1.2% and 5.8%, respectively. (3) Fetal Biometry: **FetalAgents** achieves the lowest errors across all biometric tasks. For AC measurement, it reduced MAE from 9.94 mm (FetalCLIP) to 5.53 mm and raised Acc@5% from 65.6% to 84.0%. For HC measurement, it achieved MRAE 1.40% and P95 RAE 3.75%, outperforming nnUNet by 10.8% and 5.3% in relative reduction. On GA prediction, **FetalAgents** attained the highest validity rate of 84.93% and the lowest MAE of 1.23 w, improving over FetalCLIP by 2.0 percentage points and 6.8%, respectively. For AoP estimation, it achieved MAE 5.44° and Acc@5% 67.2%, consistently outperforming all standalone models.

3.3 Image Captioning and Video Summarization Results

To demonstrate the end-to-end clinical utility of **FetalAgents** beyond isolated metrics, we qualitatively evaluated its report generation capabilities. (1) Image Caption Generation: Given a broad inquiry and a single US image, **FetalAgents** correctly identified the trans-thalamic plane, coordinated relevant experts for biometry and GA estimation, and synthesized all outputs into a structured clinical report (Fig. 2 (a)). (2) Video Summarization: Given a continuous multi-plane video stream with patient metadata, **FetalAgents** autonomously extracted keyframes across diverse anatomies, aggregated frame-wise biometrics, and cross-referenced the derived GA against the LMP-derived GA to confirm clinical consistency (Fig. 2 (b)), closely mirroring real-world sonographer workflows.

4 Conclusion

This paper introduced **FetalAgents**, the first multi-agent system that bridges the gap between isolated point-prediction models and the complex realities of fetal ultrasound workflows. By leveraging inherent intelligence and tool-use capabilities of LLMs within an agentic framework, our system dynamically orchestrates diverse specialized vision backbones to conduct accurate fetal US analysis. **FetalAgents** successfully advances beyond static image interpretation by automating the labor-intensive process of continuous US video summarization, generating structured, expert-level clinical reports automatically. Extensive external evaluations confirm that **FetalAgents** consistently outperforms state-of-the-art standalone vision models and general-purpose MLLMs across a wide array of clinical tasks. Future work will focus on integrating broader longitudinal patient histories and evaluating the system in prospective clinical settings to further validate its impact.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 82572198), the Scientific Research Project of Sichuan Medical Association (Grant No. 2024HR130), Science and Technology Department of Sichuan Province (Grant No. 25SYSX0255), Tsinghua University Startup Fund, and Tsinghua University Dushi Program (Grant No. 20241080026, 20251080056).

Disclosure of Interests

The authors confirm there are no conflicts of interest associated with this study.

References

1. Ashkani Chenarlogh, V., Ghelich Oghli, M., Shabanzadeh, A., et al.: Fast and accurate u-net model for fetal ultrasound image segmentation. *Ultrason. Imaging* **44**(1), 25–38 (2022)
2. Belciug, S.: Fetal planes and organs (2024). <https://doi.org/10.5281/zenodo.14093338>
3. Burgos-Artizzu, X.P., Coronado-Gutierrez, D., Valenzuela-Alcaraz, B., et al.: Fetal_planes_db: Common maternal-fetal ultrasound images (2020). <https://doi.org/10.5281/zenodo.3904280>
4. Chen, G., Bai, J., Ou, Z., et al.: PSFHS: Intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head. *Sci. Data* **11**, 436 (2024)
5. Chen, K., Li, X., Yang, T., et al.: Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856* (2025)
6. Da Correggio, K.S., Noya Galluzzo, R., Santos, L.O., et al.: Fetal abdominal structures segmentation dataset using ultrasonic images (2023). <https://doi.org/10.17632/4gcpm9dsc3.1>
7. Deng, J., Dong, W., Socher, R., et al.: Imagenet: A large-scale hierarchical image database. In: *CVPR*. pp. 248–255 (2009)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR* (2021)
9. Fallahpour, A., Ma, J., Munim, A., et al.: Medrax: medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673* (2025)
10. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., et al.: A review on deep-learning algorithms for fetal ultrasound-image analysis. *Med. Image Anal.* **83**, 102629 (2023)
11. González, D., et al.: Natalia: Pbf-us1 (phantom blind-sweeps for fetal ultrasound scanning) (2024). <https://doi.org/10.5281/zenodo.14193949>
12. Google: Gemini 3 Flash. Large language model (2026)
13. Guo, X., Men, Q., Noble, J.A.: Mmsummary: Multimodal summary generation for fetal ultrasound video. In: *MICCAI* (2024)
14. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778 (2016)
15. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., van Ginneken, B.: Automated measurement of fetal head circumference using 2d ultrasound images (Jul 2018). <https://doi.org/10.5281/zenodo.1327317>

16. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
17. Hu, X., Qian, Y., Yu, J., et al.: The landscape of medical agents: A survey. TechRxiv (2026)
18. Isensee, F., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**, 203–211 (2021)
19. Jiao, J., Zhou, J., Li, X., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Med. Image Anal. **96**, 103202 (2024)
20. Kim, Y., Park, C., Jeong, H., et al.: Mdagents: an adaptive collaboration of llms for medical decision-making. arXiv preprint arXiv:2404.15155 (2024)
21. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
22. Kiserud, T., Piaggio, G., Carroli, G., et al.: The world health organization fetal growth charts: A multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. PLoS Med. **14**(1), e1002220 (2017)
23. Li, R., Wang, X., Berlowitz, D., et al.: Care-ad: a multi-agent large language model framework for alzheimer’s disease prediction using longitudinal clinical notes. npj Digit. Med. **8**(1), 541 (2025)
24. Lin, X., et al.: Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. arXiv preprint arXiv:2309.06824 (2023)
25. Liu, Z., Mao, H., Wu, C.Y., et al.: A convnet for the 2020s. CVPR (2022)
26. Lu, Y., Zhou, M., Zhi, D., et al.: The JNU-IFM dataset for segmenting pubic symphysis-fetal head. Data Brief **41**, 107904 (2022)
27. Lyu, X., Liang, Y., Chen, W., et al.: Wsi-agents: a collaborative multi-agent system for multi-modal whole slide image analysis. In: MICCAI. pp. 680–690 (2025)
28. Maani, F., Saeed, N., Saleem, T., et al.: Fetalclip: a vision-language foundation model for fetal ultrasound image analysis. arXiv preprint arXiv:2502.14807 (2025)
29. Mei, X., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiol. Artif. Intell. **4**(5), e210315 (2022)
30. OpenAI: Gpt-5 mini. Large language model (2026)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
32. Salomon, L.J., et al.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstet. Gynecol. **37**(1), 116–126 (2011)
33. Sappia, M., de Korte, C., van Ginneken, B., et al.: ACOUSLIC-AI challenge report: Fetal abdominal circumference measurement on blind-sweep ultrasound data from low-income countries. Med. Image Anal. **105**, 103640 (2025)
34. Selligren, A., Kazemzadeh, S., Jaroensri, T., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
35. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., et al.: Maternal fetal ultrasound planes from low-resource imaging settings in five african countries (2023). <https://doi.org/10.5281/zenodo.7540448>
36. Sippel, S., Muruganandan, K., Levine, A., et al.: Review article: Use of ultrasound in the developing world. Int. J. Emerg. Med. **4**, 72 (2011)
37. Vaidya, S., et al.: MedPAO: A protocol-driven agent for structuring medical reports. In: AI for Clinical Applications. pp. 33–45 (2025)
38. Wang, F., Whelan, K., Silvestre, G., et al.: Generative diffusion model bootstraps zero-shot classification of fetal ultrasound images in underrepresented african populations. arXiv preprint arXiv:2407.20072 (2024)
39. Wang, F., Silvestre, G., Curran, K.M.: Segmenting fetal head with efficient fine-tuning strategies in low-resource settings: an empirical study with u-net. arXiv preprint arXiv:2407.20086 (2024)

40. Wu, Q., Bansal, G., Zhang, J., et al.: Autogen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155 (2023)
41. Xiao, T., Liu, Y., Zhou, B., et al.: Unified perceptual parsing for scene understanding. arXiv preprint arXiv:1807.10221 (2018)
42. Yan, L., Li, Q., Fu, K., et al.: Progress in the application of artificial intelligence in ultrasound-assisted medical diagnosis. *Bioengineering* **12**(3), 288 (2025)
43. Zeng, W., Luo, J., Cheng, J., et al.: Efficient fetal ultrasound image segmentation for automatic head circumference measurement using a lightweight deep convolutional neural network. *Med. Phys.* **49**(8), 5081–5092 (2022)
44. Zhao, W., Wu, C., Fan, Y., et al.: An agentic system for rare disease diagnosis with traceable reasoning. *Nature* (2026)
45. Zhou, Z., Lu, Y., Bai, J., et al.: Segment anything model for fetal head-pubic symphysis segmentation in intrapartum ultrasound image analysis. *Expert Syst. Appl.* **263**, 125699 (2025)