This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

FOCUS: Towards Fetal Obstetric Corrective UltraSound Guidance

Hala Lamdouar[1], Angela Feixue Wang[1], Xiaoqing Guo[2], Qianhui Men[3], Jayne Lander[4], Aris T. Papageorghiou[4], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, University of Oxford, UK
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[3] School of Engineering Mathematics and Technology, University of Bristol, UK
[4] Nuffield Department of Women's & Reproductive Health, University of Oxford, UK
hala.lamdouar@eng.ox.ac.uk

Abstract. Guidance is essential in teaching the required skills for performing highly technical obstetric ultrasound examination. These skills include selecting the appropriate control panel functions, making correct image adjustments, and positioning the probe effectively to acquire high quality anatomical views. Existing work mainly focused on assessing image quality, without identifying where errors occur during ultrasound scanning or how a trainee should correct them. We introduce FOCUS, the Fetal Obstetric Corrective UltraSound instructor, a new multimodal framework designed to learn and deliver corrective guidance for trainee sonographers using both video and text. FOCUS analyses ultrasound video sequences and provides corrective feedback indicating specific actions needed to improve technical performance. To develop this framework, we collect obstetric ultrasound data captured by trainees, annotated with corrective instruction labels, anatomical labels, and accompanying expert textual feedback. Our experiments demonstrate that FOCUS effectively learns to recognise suboptimal image patterns and generate targeted corrective feedback. We provide the code at https://github.com/hlamdouar/focus-ultrasound-guidance.

Keywords: Fetal Ultrasound · Generated feedback · Corrective instruction · Human AI collaboration

1 Introduction

Obstetric ultrasound examination is a highly complex task that involves the detection and recognition of key structures, the capture of specific anatomical views, and the measurement of biometry to assess the progress of the pregnancy and its associated care plan. Typically, during the examination, the operator aims to maintain the target fetal or maternal anatomical area in the centre of the screen, and to maximise image quality. To this end, the operator proceeds with a sequence of actions: *(i)* manipulating the transducer, *i.e.* through movements of rotation, translation, and inclination; and *(ii)* selecting optimal parameters

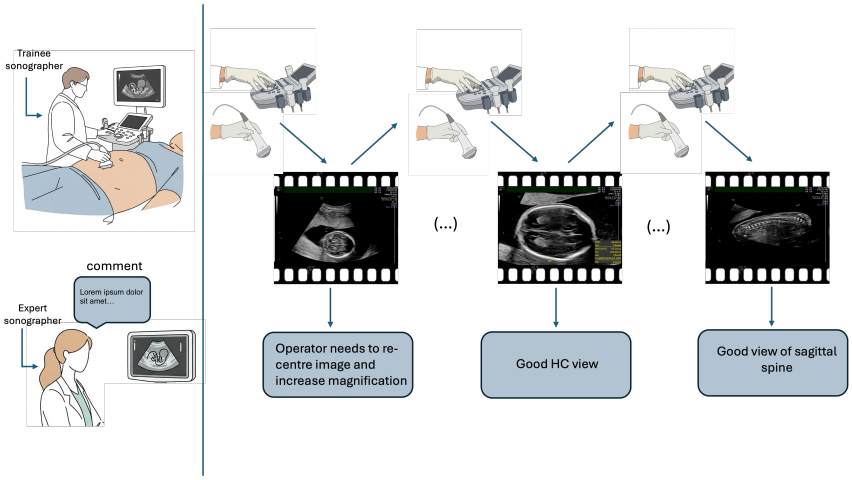


Fig. 1: Obstetric Ultrasound Trainee dataset: In our setting, a trainee ultrasound operator examines different fetal and maternal anatomies by performing a sequence of probe motions and keyboard actions. Retrospectively, an expert sonographer observes the ultrasound video and comments on the subsequences.

on the control panel, *i.e.* depth, sector width, focal zones, magnification *etc.* In addition to the complexity of actions, sonographers have to react to other factors that may occur during the scan that increase the difficulty *e.g.*, fetal movements, acoustic shadow. Learning to scan well can take years of practice to achieve the very high levels of proficiency needed for tasks such as fetal echocardiography. Even for basic skills such as fetal biometry, there is a worldwide lack of qualified sonographers, and a scarcity of training resources, particularly in resource-constrained environments where demand is high [20].

Current advances in Artificial Intelligence (AI) offer a great foundation for developing tools to support trainee sonographers and to help with their learning process. This capability is also relevant to non-ultrasound specialists who want to develop basic ultrasound skills for simple tasks on an occasional basis (such as a General Practitioner in a community or rural settings). Previous works have tackled ultrasound image quality problems, such as acoustic shadow estimation [10, 12, 13, 21], and more recently Point-of-Care Ultrasound (POCUS), as acquired by non-experts, it suffers from noisy suboptimal views. Recent work has explored language-based approaches for ultrasound captioning [1], and multimodal frameworks, including large language models, for report generation, visual grounding, and summarisation [18, 14, 7, 9]. More recently, research has shifted towards human–AI collaboration [8], with interactive systems providing real-time feedback and guidance to sonographers during acquisition [4, 15].

In this work, we reframe ultrasound guidance around the concept of sonographer learners. Rather than treating guidance as static-image interpretation, we model it as training an AI instructor that behaves like an expert standing

over a trainee's shoulder. The system evaluates live scanning behavior and issues corrective feedback when needed—e.g., "structures are unclear; increase magnification and re-centre the view" or "adjust the probe to eliminate shadowing from ...". To support this vision, we construct a dedicated dataset of trainee scanning sessions paired with retrospective expert commentary, and propose a multimodal framework FOCUS that generates temporally grounded, actionable feedback, shifting ultrasound AI guidance from static image analysis towards interaction-based, context-aware instruction that mirrors real-world teaching.

We make the following contributions: *(i)* We introduce CorLM, a lightweight transformer-based language module explicitly conditioned on anatomy, corrective instruction, and visual embeddings. Unlike generic captioning systems, CorLM generates clinically meaningful, anatomy-aware corrective feedback tailored to obstetric ultrasound training. *(ii)* We propose a dual-head framework that disentangles anatomical context from prescriptive guidance via dedicated decoders, producing semantic embeddings that consistently ground the language generator. This design enables fine-grained visual understanding and improves the reliability of downstream feedback. *(iii)* Finally, we describe a dataset framework of trainee sonographer videos paired with retrospective expert commentary and introduce an embedding-based semantic detection metric to assess whether generated comments capture the intended anatomical or corrective concept. While the dataset is private, we provide a transparent construction protocol.

2 Method

Motivation Consider a dataset $\{\mathcal{V}_i, t_i\}$, comprising video sequences acquired during obstetric examinations, here performed by trainee sonographers, paired with corresponding retrospective expert commentary providing corrective feedback. We argue that each corrective instruction (e.g., increasing gain, adjusting magnification) is intrinsically tied to the specific anatomy, fetal or maternal, being examined. As anatomical structures differ in appearance, echotexture, depth, and acoustic context, the visual manifestation of a required correction varies accordingly across views. To capture these dependencies, we adopt a conditioned language module whose generative output is grounded in learned semantic representations associated with both the corrective instruction and the underlying anatomy. This conditioning ensures that generated feedback is context-aware, anatomically appropriate, and aligned with expert guidance.

Visual stream The goal of this component is to provide the downstream language module with stable anatomical and contextual cues necessary for grounded corrective feedback. To extract semantically meaningful representations from ultrasound video sequences, we adopt a pretrained vision transformer (ViT)–style [5] encoder with frozen backbone weights. Formally, given an ultrasound frame $\mathcal{V}_i^{(t)}$, the encoder produces a d-dimensional representation

$$f_{\mathcal{V}}^{(t)} = \mathrm{ViT}(\mathcal{V}_i^{(t)}) \in \mathbb{R}^d, \tag{1}$$

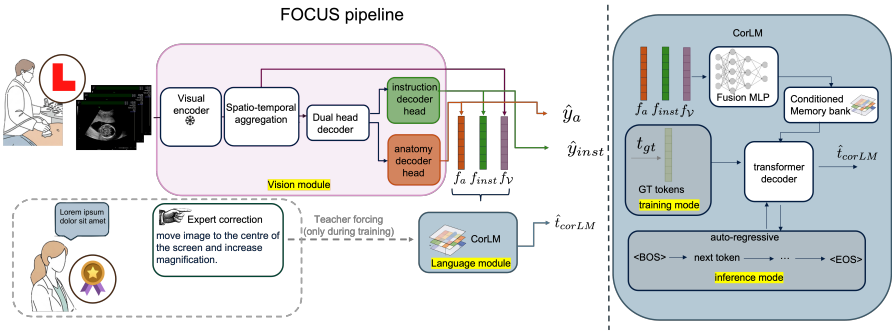


Fig. 2: **FOCUS pipeline.** We leverage retrospective expert commentary to train CorLM, a conditioned language module that generates corrective feedback for trainee sonographers. CorLM distills rich semantic information by attending to anatomy embeddings f_a, instruction embeddings f_{inst}, and vision embeddings $f_{\mathcal{V}}$. Conditioned on these complementary signals, the model produces detailed, context-aware instructive feedback to guide skill acquisition.

These embeddings are aggregated in a temporal fusion using a lightweight pooling operator:

$$f_{\mathcal{V}} = \mathrm{Pool}\Big(f_{\mathcal{V}}^{(1)}, \ldots, f_{\mathcal{V}}^{(T)}\Big)\,, \tag{2}$$

which captures anatomical appearance and overall image appearance.

Dual decoder head On top of the shared visual encoder, we employ a dual–decoder design consisting of (i) an *anatomy decoder* and (ii) an *instruction decoder*. Each decoder receives the clip-level visual representation $f_{\mathcal{V}} \in \mathbb{R}^D$ and produces a semantically meaningful embedding aligned to its respective prediction task. The anatomy decoder maps the visual features to an embedding $f_a = \mathrm{MLP}_a(f_{\mathcal{V}})$, which is subsequently classified into one of the annotated fetal or maternal anatomical structures via a softmax operator. Similarly, the instruction decoder forms an embedding $f_{\mathrm{inst}} = \mathrm{MLP}_{\mathrm{inst}}(f_{\mathcal{V}})$, which is used to classify the corrective action appropriate for the given view (e.g., adjust gain, magnify, or re-centre the structure). This dual-headed design encourages the encoder to learn disentangled representations of anatomy and corrective instruction, providing richer supervision and improving the grounding of generated feedback.

Conditioned language module To generate anatomically grounded and contextually appropriate corrective feedback, we incorporate a conditioned language module that leverages the semantic signals extracted by the dual decoders and the visual encoder. Let f_a and f_{inst} denote the anatomy and instruction embeddings, respectively, and let $f_{\mathcal{V}}$ denote the visual embedding produced by the backbone. These components are fused by a lightweight multilayer perceptron:

$$f_{\mathrm{cond}} = \mathrm{MLP}_{\mathrm{fusion}}([\, f_{\mathcal{V}},\, f_a,\, f_{\mathrm{inst}}\,])\,, \tag{3}$$

which yields a compact conditioning vector that summarises the anatomical context, the intended corrective action, and the visual characteristics of the ultrasound clip.

This fused vector is inserted into the transformer language decoder as a *conditioning memory token*. Formally, we treat $M = f_{\text{cond}} \in \mathbb{R}^{1\times d}$ as the key–value memory for all cross-attention layers within the decoder. During training, the decoder operates in a teacher-forcing regime; it receives the ground-truth token sequence, attends to the conditioning memory M, and predicts next-token logits which are optimised with a cross-entropy loss. At inference time, the decoder generates text autoregressively, producing one token at a time while continually attending to the same conditioning vector M.

By integrating both explicit semantic embeddings and trainable memory slots, the conditioned language module produces text that is structurally coherent and semantically grounded in the underlying anatomy, the corrective instruction, and the visual appearance of the ultrasound clip.

Training objective: We adopt the cross entropy loss to train the visual module. Given logits $\hat{\mathbf{y}}_{inst}$ and $\hat{\mathbf{y}}_a$ for corrective instruction and anatomy respectively:

$$\mathcal{L}_{\text{inst}} = \mathcal{L}_{\text{CE}}(\mathbf{y}^{\text{gt}}_{inst}, \hat{\mathbf{y}}_{inst}), \qquad \mathcal{L}_{\text{a}} = \mathcal{L}_{\text{CE}}(\mathbf{y}^{\text{gt}}_{a}, \hat{\mathbf{y}}_{a}), \tag{4}$$

For training the corrective feedback module, we adopt a masked language loss where the ground-truth tokens are shifted by one position to define next-token targets, and padding tokens are masked:

$$\mathcal{L}_{\text{corLM}} = -\frac{1}{|\mathcal{M}|} \sum_{(i,t)\in\mathcal{M}} \sum_{v} q_{i,t}(v)\, \log p_{i,t}(v), \tag{5}$$

where $\mathcal{M}$ is the set of non-<pad> positions, $q_{i,t}$ is the smoothed target distribution, and $p_{i,t}$ the probability distribution over the vocabulary at timestep t.

The total objective is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{inst}}\mathcal{L}_{\text{inst}} + \lambda_{\text{a}}\mathcal{L}_{\text{a}} + \lambda_{\text{corLM}}\mathcal{L}_{\text{corLM}}. \tag{6}$$

3 Experiments

3.1 Obstetric Ultrasound Skill Dataset

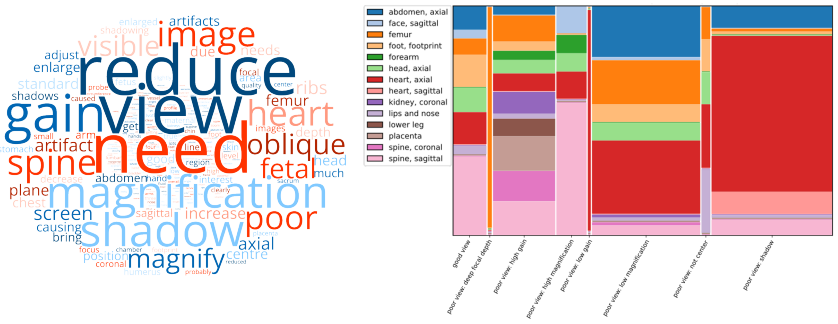
We collected and annotated 7 second trimester obstetrics ultrasound examination videos with expert audio commentary. Video durations range from 20 minutes to 31 minutes, totalling 2 h 53 min 32 s, with more than 260K frames at an imaging frame rate of 25fps. These videos were extracted from the PULSE project [6] with ethic approval and written consent from the participants.

Expert audio commentary was provided by an experienced sonographer trainer for each video, addressing their suggested corrections as well as the fetal anatomies observed, as presented in Figure 1. The audio was processed in two stages. First,

WhisperX [2] was used to transcribe the audio into text with sentence-level timestamps. Next, video clips corresponding to correction and anatomy commentary were extracted by a trained researcher, with manual validation of transcripts and clip boundaries. This produced a dataset of 324 clips, ranging from 0.042s to 60.665s, with a variety of examined anatomies and expert suggested corrections, as visualised in a mosaic plot in Figure 3. The dataset is split into 70% for training, 10% for validation, and 20% for testing, ensuring that all correction categories are present in the validation and test sets for unbiased model evaluation and avoiding data leakage.

The eight expert-suggested correction categories identified from the extracted clips are described below, with explanations detailing actions that the trainee could take to improve the quality of the image:

- "good view": when a standard plane was obtained according to guidelines.
- "poor view: deep focal depth": indicates that the trainee should decrease focal depth.
- "poor view: high gain": indicates that the trainee should decrease the gain.
- "poor view: high magnification": indicates that the trainee should decrease magnification.
- "poor view: low gain": indicates that the trainee should increase the gain.
- "poor view: low magnification": indicates that the trainee should increase the magnification.
- "poor view: not center": indicates that the trainee should adjust the probe so that the anatomy of interest moves to the center of the screen.
- "poor view: shadow": indicates presence of acoustic shadowing over the anatomy of interest, either from fetal or maternal anatomy. The trainee should avoid these.



(a) Word Cloud (b) Mosaic plot

Fig. 3: Distribution of dataset across instructions and fetal anatomies

Model	Instruction				Anatomy				Commentary		
	Acc	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**B**	**R-L**	**M**
Ours-w/o dual	65.4	64.9	65.5	64.7	63.7	64.1	63.8	61.1	8.4	27.6	21.8
Ours-ResNet50	**68.4**	**68.9**	**68.4**	**68.2**	64.8	67.2	64.8	61.7	**13.4**	**31.9**	**28.4**
Ours-SwinT	64.7	67.6	61.8	63.3	**74.5**	**74.1**	**71.4**	**70.3**	10.2	28.3	24.4

Table 1: Comparison on US instruction and fetal anatomy classification, and commentary generation metrics ($Rouge_L$ [16], Bleu [19] and METEOR [3]).

3.2 Implementation details

Models were trained with a batch size of 8 on ultrasound clips of 8 frames, sampled at one frame every five time steps. Each frame was resized to 448×448 and normalised using the pretrained backbone statistics and encoded into embeddings of dimension $d = 512$. We train all components end-to-end with the Adam optimizer and an initial learning rate of 10^{-4}. We use $\lambda_{\text{inst}} = \lambda_{\text{corLM}} = 1.0$ and $\lambda_{\text{a}} = 0.5$. All experiments are conducted on a single NVIDIA Quadro RTX 5000 GPU. Text is tokenized with a custom 230-word vocabulary extracted from the expert commentary. We use <bos>, <eos>, and <pad> special tokens. During training, teacher forcing was applied for sequence generation.

3.3 Quantitative Results

Table 1 summarizes performance across three dimensions: (i) anatomy recognition, (ii) corrective instruction classification, and (iii) the semantic quality of generated feedback. We evaluate our architecture with two visual backbones: ResNet50 [11] and Swin-T [17]. Interestingly, the model that achieves the strongest performance on corrective instruction prediction (the ResNet50-based variant) and delivers superior results on the corrective comment generation task, outperforming its Swin-T counterpart. This suggests that accurate instruction grounding directly benefits downstream language generation quality. In addition, we train an ablated model in which the language decoder is deprived of instruction–anatomy conditioning: "ours-w/o dual". This results in a significant decrease in performance showing the effectiveness of our conditioning design. A per-class analysis of the results in Table 1 reveals performance discrepancies, likely due to class imbalance. For instance, Ours w/o dual achieves 62.5% on Good View, 33.3% on High Gain, 22.2% on High Magnification, 81.2% on Low Magnification, and 66.3% on Shadow.

3.4 Qualitative Results

We present qualitative test examples from our dataset in Figure 4. The model consistently produces rich, context-aware feedback, in some cases offering observations that go beyond the original expert commentary. For instance, while

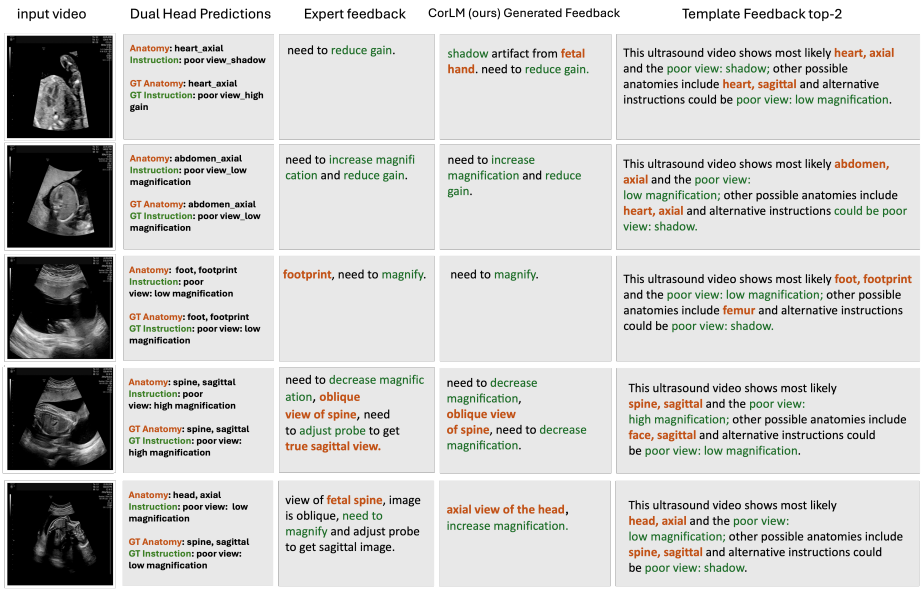


Fig. 4: **Qualitative Results.** From left to right: image from ultrasound video sequence, Dual head predictions and ground truth, Ground truth expert feedback, our generated corrective feedback and a template generated feedback.

the ground-truth instruction in the first example was simply to reduce the gain, the model also identifies the presence of acoustic shadowing and, given that the scanned structure is a fetal heart, correctly attributes it to a plausible cause ("shadow artifact from fetal hand"). In some instances, the generated feedback omits explicit reference to the underlying anatomy, as illustrated in the second and third examples. This reflects the nature of the training annotations, where the expert did not always include anatomical descriptions when they were not considered clinically relevant for the clip. We also observe failure modes for severely degraded inputs. In the final example, the model incorrectly predicts an axial head view, whereas the correct anatomy is the sagittal spine. Notably, the correct label appears among the model's top-2 predictions (cf. template).

3.5 Limitations

Our dataset is annotated by an expert sonographer instructor, following standard guidelines for obstetric ultrasound examinations. However, the subjective nature of human judgment means that the expert commentary may reflect a degree of subjectivity. For example, one expert might consider an anatomical view good, while another expert would continue to change acquisition setting, *e.g.* adjusting the gain and magnification in search of a more optimal view. While we carefully curated our dataset to capture a variety of challenging subtasks encountered

by sonographer trainees, its current scale and imbalance reflect the practical challenges of obtaining expert annotations in obstetric ultrasound. Larger-scale validation remains necessary for clinical deployment, and future work will extend this dataset with broader expert annotations.

4 Conclusion

In this paper, we present a framework for the automated generation of corrective feedback for ultrasound trainee performance during obstetric examinations. Our approach leverages a conditioned multimodal language model to produce anatomically grounded feedback from visual anatomical and instructional context. Experimental results demonstrate that the proposed method generates feedback closely aligned with expert commentary while remaining lightweight and efficient. These findings highlight the potential of compact multimodal language models to provide efficient, data-driven guidance for trainee sonographers, supporting more effective and scalable ultrasound training.

Acknowledgement. This research is funded by the EPSRC Turing AI Fellowship: Ultra Sound Multi-modal video-based human-machine collaboration EP/X040186/1.

Disclosure of Interests. The authors have no financial conflicts of interest to disclose related to this work.

References

1. Alsharid, M., Sharma, H., Drukker, L., Chatelain, P., Papageorghiou, A.T., Noble, J.A.: Captioning ultrasound images automatically. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 338–346. Springer (2019)
2. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio. INTERSPEECH 2023 (2023)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Bashir, Z., Lin, M., Feragen, A., Mikolaj, K., Taksøe-Vester, C., Christensen, A.N., Svendsen, M.B., Fabricius, M.H., Andreasen, L., Nielsen, M., et al.: Clinical validation of explainable ai for fetal growth scans through multi-level, cross-institutional prospective end-user evaluation. Scientific Reports **15**(1), 2074 (2025)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
6. Drukker, L., Sharma, H., Droste, R., Alsharid, M., Chatelain, P., Noble, J.A., Papageorghiou, A.T.: Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Scientific Reports **11**(1), 14109 (2021)

7. 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 pp. 1–17 (2026)
8. Guo, X., Jin, Y., Lamdouar, H., Men, Q., Ouyang, C., Sahu, M., Vedula, S.S.: Human-AI Collaboration: First International Workshop, HAIC 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 27, 2025, Proceedings. Springer (2026)
9. Guo, X., Men, Q., Noble, J.A.: Mmsummary: multimodal summary generation for fetal ultrasound video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 678–688. Springer (2024)
10. Hacihaliloglu, I.: Enhancement of bone shadow region using local phase-based ultrasound transmission maps. International journal of computer assisted radiology and surgery **12**(6), 951–960 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hellier, P., Coupé, P., Morandi, X., Collins, D.L.: An automatic geometrical and statistical method to detect acoustic shadows in intraoperative ultrasound brain images. Medical Image Analysis **14**(2), 195–204 (2010)
13. Hu, R., Singla, R., Deeba, F., Rohling, R.N.: Acoustic shadow detection: study and statistics of b-mode and radiofrequency data. Ultrasound in medicine & biology **45**(8), 2248–2257 (2019)
14. Li, J., Su, T., Zhao, B., Lv, F., Wang, Q., Navab, N., Hu, Y., Jiang, Z.: Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. IEEE Transactions on Medical Imaging **44**(1), 19–30 (2024)
15. Li, X., Huang, D., Zhang, Y., Navab, N., Jiang, Z.: Semantic scene graph for ultrasound image explanation and scanning guidance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 500–510. Springer (2025)
16. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. Maani, F., Saeed, N., Saleem, T.J., Farooq, Z., Alasmawi, H., Diehl, W., Mohammad, A., Waring, G., Valappil, S., Bricker, L., et al.: Fetalclip: A visual-language foundation model for fetal ultrasound image analysis. npj Digital Medicine (2026)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
20. Stewart, K.A., Navarro, S.M., Kambala, S., Tan, G., Poondla, R., Lederman, S., Barbour, K., Lavy, C.: Trends in ultrasound use in low and middle income countries: a systematic review. International Journal of Maternal and Child Health and AIDS **9**(1), 103 (2020)
21. Yasutomi, S., Arakaki, T., Matsuoka, R., Sakai, A., Komatsu, R., Shozu, K., Dozen, A., Machino, H., Asada, K., Kaneko, S., et al.: Shadow estimation for ultrasound images using auto-encoding structures and synthetic shadows. Applied Sciences **11**(3), 1127 (2021)