Improved Representation Learning with Multitasking in Blind Sweep Obsetric Ultrasound Videos

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2026

Abstract

Blind Sweep Obstetric Ultrasound (BSOU), based on predefined abdominal trajectories, enable non-experts to capture ultrasound videos and are increasingly used for AI-based estimation of obstetric measures like Gestational Age and Fetal Presentation. However, existing work focuses on single-task models, overlooking the potential of joint learning. We propose the first multi-task framework for BSOU-based AI models, leveraging spatio-temporal constraints inherent in sweep protocols and fetal anatomy. Our approach includes multi-head cross-entropy (MTCE) and a novel approach to Multi-head Supervised Contrastive Loss (MTSCon) for BSOU datasets, treating videos with matching labels across patients and sweep types as augmented versions of the same input in a contrastive setting. We introduce new applications, Deepest Vertical Pocket estimation and sweep type prediction, and show that carefully selected task combinations improve both in-domain performance and generalization to out-of-domain settings. The implementation code will be made publicly available upon publication.

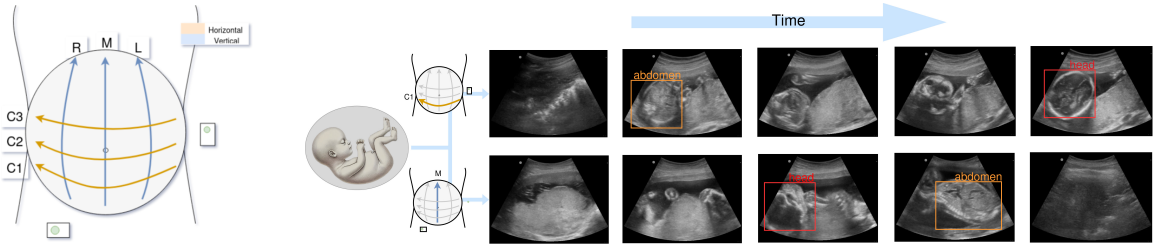
Keywords: Blind Sweep, Fetal, Multi-task, Multi-head, Contrastive

1. Introduction

Pregnancy complications and related mortality remain high in Low Middle IncomeCountries (LMICs) due to limited access to prenatal imaging, primarily from a shortage of ultrasound devices and trained operators. While ultrasound has become more affordable and portable, the lack of skilled personnel remains a key barrier (DeStigter et al., 2011; Abuhamad et al., 2016). Imaging the World (ITW) had developed the Blind Obstetric Ultrasound Sweep (BSOU) protocol, enabling minimally trained workers to acquire ultrasound images using predefined abdominal sweep paths (DeStigter et al., 2011)(Figure 1a). Experts could then access these videos remotely to evaluate key obstetric indicators such as multiple pregnancy, presentation, and placental location. However, these clips, which contain useful information spread across frames and sweeps, make it difficult for experts accustomed to acquiring standard views in real time that contain the desired information.

Recently, there has been renewed interest in BSOU videos, with multiple initiatives worldwide collecting them to develop deep learning models that automatically assess various obstetric parameters (van den Heuvel et al., 2019; Self et al., 2022; Pokaprakarn et al., 2022; AIMIX Project, 2025). These deep learning models have shown promise in estimating important obstetrics markers from BSOU videos, such as Gestational Age (GA) estimation (Pokaprakarn et al., 2022; Gomes et al., 2022; Stringer et al., 2024; Patel et al., 2024; Akumu et al., 2025), fetal presentation classification (Gomes et al., 2022; Gleed et al., 2023a, 2024; Wiśniewski et al., 2025), placenta localization (Gleed et al., 2023b; Schilpzand

et al., 2022), abdominal circumference estimation for fetal growth restriction (Sappia et al., 2025) and multiple pregnancy detection (Kalantari et al., 2024). Some of these works have demonstrated the potential to match or even exceed the performance of trained sonographers in GA estimation (Pokaprakarn et al., 2022; Gomes et al., 2022).

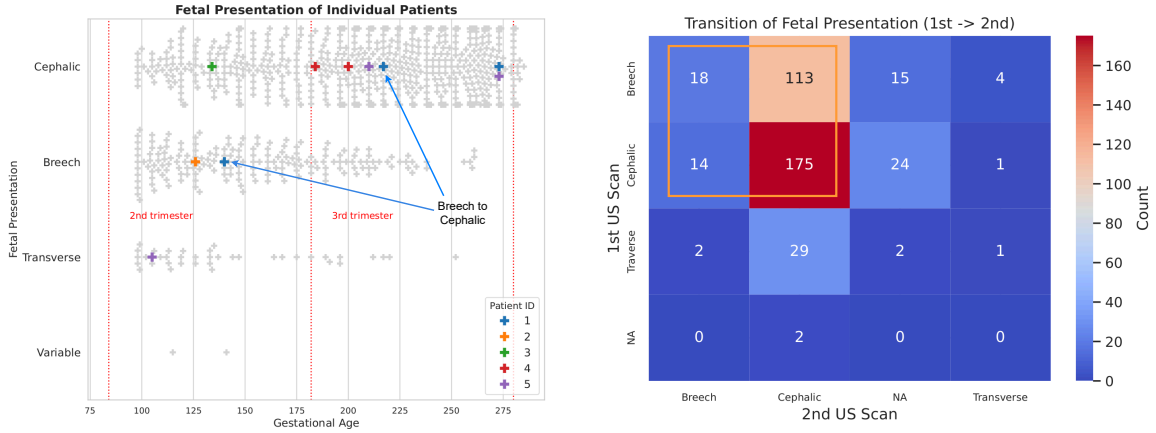


(a) (b)

Figure 1: (a) 6-sweep Blind Sweep Obstetrics Ultrasound (BSOU) protocol, (b) Different sweep directions: horizontal (C1) and vertical (M) produce ultrasound videos where same fetal structures, such as the fetal head (red) and abdomen (brown), appear at varying locations, orientations, and scales. The variability of fetal appearance across different sweeps highlights the value of jointly modeling sweep trajectories and key clinical parameters, such as fetal presentation, to better capture and interpret spatio-temporal patterns in BSOU data. The leftmost graphic in (b) was adapted from Wiśniewski et al. (2025).

Although BSOU datasets collected during routine care provide multiple clinical labels such as fetal biometry, presentation, and placental characteristics etc., most deep learning studies address these tasks independently. For instance, GA estimation (Gomes et al., 2022; Pokaprakarn et al., 2022) and fetal presentation (Gomes et al., 2022) tasks have separate models trained independently, missing out on the opportunity to learn jointly by leveraging interrelations among the tasks.

BSOU data inherently have spatio-temporal constraints being confined in a specific anatomical region of a pregnant women where the fetus is growing. Since the fetus moves within the womb and grows over time, there are interesting dependencies of various tasks such as GA estimation, fetal presentation classification with respect to which sweep is taken during what GA. For instance, identical fetal presentations can appear differently across sweep types (see Figure 1b), and anatomical features vary with gestational age even under a fixed acquisition protocol. Further, key fetal attributes such as presentation and placenta location are dynamic and may change considerably with gestational age. For instance, longitudinal tracking of fetal presentation (Figure 2) shows that it shifts markedly with gestational age, typically trending toward a cephalic position. This shows clear correlation can exist among labels such as Gestational Age, Fetal Presentation, Placenta Location, and Sweep type and exploiting such interrelationships may enhance model performance. We posit that the predefined sweep trajectories and the anatomically constrained uterine environment create spatio-temporal regularities that can be effectively leveraged, especially when large-scale patient data are available.



(a) (b)

Figure 2: Temporal dynamics of fetal presentation during pregnancy. Panel (a) illustrates individual longitudinal changes in presentation across gestational age, while panel (b) summarizes the overall pattern of presentation transitions between consecutive scans, demonstrating that fetal presentation is dynamic and commonly changes towards cephalic as pregnancy advances.

In this paper, we propose the first multi-task framework for obstetrics AI models using BSOU vidoes and study nuances of combining subsets of different tasks for more robust representation learning. In addition to the standard multi-head crossentropy loss (MTCE), we propose novel adaptation of Multi-head Supervised Contrastive Loss (MTSCon), where a sweep video from one patient is considered as 'in-class' with another patient's same trajectory sweep video when specific class label for them match, such as when both have the same fetal presentation. Our key contributions are:

- The first two-stage multi-task framework for BSOU using Multiheaded Supervised Contrastive Learning and Multitask Cross-Entropy, benchmarked against single-task and unsupervised methods.
- First to perform Deepest Vertical Pocket (DVP) estimation and sweep type prediction.
- Improved generalization across domains, including varied clinical settings, devices, and patient characteristics.

2. Related Works

2.1. BSOU data and AI

Recent advances in AI have renewed interest in blind sweep obstetric ultrasound (BSOU) as a means to expand access to obstetric imaging in low-resource settings. This has motivated the development and evaluation of a variety of sweep acquisition protocols. Large-scale datasets such as FAMLI2 have collected up to 15 distinct blind sweeps (Gomes et al., 2022),

whereas other studies have adopted more structured multi-step protocols, most notably the five-step approach inspired by Abuhamad et al. (2016) and explored in Self et al. (2022), in which each step comprises targeted sweeps designed for specific clinical objectives. Adaptive strategies have also been proposed, where the number of sweeps varies according to the symphysio-fundal height (SFH) (AIMIX Project, 2025; Akumu et al., 2025). Despite this, many studies have focused on developing deep learning methods for the standard six-sweep protocol (Gomes et al., 2022; Sappia et al., 2025).

The predominant focus of BSOU-based AI research has been gestational age (GA) estimation. Approaches range from attention-based feature aggregation using large-scale datasets (Pokaprakarn et al., 2022), to demonstrations of lightweight models capable of robust GA prediction (Gomes et al., 2022). More interpretable pipelines have also emerged; for example, Akumu et al. (2025) identifies clinically meaningful and diverse standard frames prior to GA estimation, enabling strong performance with limited data. Beyond GA, recent work has begun to explore broader clinical tasks. He et al. (2025) investigated the detection of frames suitable for fetal biometry, while fetal presentation has been addressed using both optimized deep learning models on large datasets (Gomes et al., 2022) and interpretable feature-based methods (Gleed et al., 2023a). Placental assessment has likewise been studied, with methods for classifying normal versus low-lying placenta or placenta previa (Schilpzand et al., 2022), and for placenta segmentation (Gleed et al., 2023b).

However, several clinically important aspects of BSOU remain underexplored. These include the ability of models to estimate the deepest vertical pocket (DVP) i.e. critical for assessing amniotic fluid volume as well as operator-dependent quality factors, such as accurate sweep classification to confirm whether a complete and clinically adequate sweep set has been acquired. Further, the potential benefits of jointly learning sweep type alongside important clinical labels are yet to be explored.

2.2. Multitask Representation Learning

Supervised representation learning has been explored in fetal ultrasound to improve feature quality. For instance, Fu et al. (2022) used supervised contrastive learning (Khosla et al., 2020) to pull anatomically similar image pairs closer in feature space. However, in BSOU, each video can belong to multiple classes simultaneously with respect to sweep type, gestational age, fetal presentation, and placental findings etc. This motivates using multi-task learning (MTL) to enable models to learn multiple related tasks concurrently by sharing a common representation. MTL often improves performance relative to single-task models by enabling inductive transfer, providing implicit regularization, and reducing sample complexity (Caruana, 1997; Zhang and Yang, 2022). In fetal ultrasound MTL has been applied to tasks such as fetal biometric measurement (Płotka et al., 2021), standard plane detection (Guo et al., 2023), and image quality assessment (Lin et al., 2019; Zhang et al., 2021), showing that jointly learning complementary clinical tasks can improve accuracy and robustness.

Despite these successes, MTL has not been explored in BSOU. Although these datasets include diverse, clinically related labels such as sweep type, gestational age, fetal presentation, and placental findings, current models treat each task independently, missing the opportunity to leverage their interdependencies for more robust representations.

3. Multi-task Learning from BSOU Videos

3.1. Learning from BSOU Videos: Problem Formulation

BSOU examinations follow standardized protocols, capturing video sweeps through predefined trajectories over the abdomen. Let Ω represent the space of all ultrasound video sweeps, with each sweep $x \in \Omega$ forming a 4D tensor in $\mathbb{R}^{F\times H\times W\times C}$, where F, H, W, and C denote frames, height, width, and channels, respectively. For a set of T tasks, each task $t \in \{1,\dots,T\}$ is associated with a label space $\mathcal{Y}^{(t)}$, either discrete $\{1,\dots,K_t\}$ for classification or continuous $[a_t, b_t] \subset \mathbb{R}$ for regression. Labels exist at both Patient-Level ($y_p^{(t)}$) and Sweep-Level ($y_{ps}^{(t)}$). Each patient $p \in \{1,\dots,N\}$ undergoes a predefined number of standard S sweeps indexed by $s \in \{1,\dots,S\}$, with the complete set of sweeps denoted as $X_p = \{x_{ps}\}_{s=1}^{S}$ (Figure 1a, $S = 6$).

3.2. Multi-Task Learning Framework

Our approach consists of two sequential stages designed to leverage multi-task learning benefits while enabling task-specific optimization (Figure 3).

First stage: Multi-Task Representation Learning : The initial stage focuses on representation learning at the **sweep level**. A deep encoder, $f_\theta : \Omega \to \mathbb{R}^d$, produces latent representations $z_{ps} = f_\theta(x_{ps}) \in \mathbb{R}^d$ that capture spatio-temporal anatomical patterns of the video. For training, patient-level annotations are propagated to all sweeps ($y_{ps}^{(t)} = y_p^{(t)}, \forall s \in \{1,\dots,6\}$). The dataset is defined as $\mathcal{D}_{sweep} = \{(x_{ps}, y_{ps})\}_{p=1,s=1}^{N,6}$ (Figure. 3).

We use a subset of tasks $\mathcal{T}_{pre} \subseteq \mathcal{T}$ selected for their ability to capture underlying commonalities in ultrasound structures: GA Category Classification (4 categories spanning 2nd and 3rd trimesters), Fetal Presentation (Cephalic and non-Cephalic), Placenta Location (Anterior and Posterior), and Sweep Type ($S = 6$ categories). The goal is to explore if the joint learning of these tasks can leverage the unique spatio-temporal constraints in BSOU videos and learn a more robust representations generalizable across clinical tasks.

We present experiments with two loss functions: **Cross Entropy for multitask classification, and the unweighted multi-head supervised contrastive loss** (Mu et al., 2023), which generalizes the supervised contrastive loss (Khosla et al., 2020) in multi-label setting.

Specifically, the multitask classification loss function is defined as:

$$\mathcal{L}_{cls}(\theta) = \frac{1}{NT}\sum_{p=1}^{N}\sum_{t\in\mathcal{T}_{pre}} \mathcal{L}_{CE}(f_t(z_{ps}), y_s^{(t)}) \tag{1}$$

where $z_{ps} = f_\theta(x_{ps})$ is the latent representation of sweep s for patient p, and $y_s^{(t)}$ is the corresponding label for task t.

And, the multi-head supervised contrastive learning:

$$\mathcal{L}_{m-supcont}(\theta) = \frac{1}{|\mathcal{B}|}\sum_{(p,s)\in\mathcal{B}} \frac{-1}{|\mathcal{P}(p,s)|}\sum_{p'\in\mathcal{P}(p,s)}\sum_{t\in\mathcal{T}_{pre}} \log \frac{\exp(z_{ps}\cdot \tilde{z}^t_{p'}/\tau)}{\sum_{a\in\mathcal{A}(p,s)}\exp(z_{ps}\cdot z^t_a/\tau)} \tag{2}$$

where $\tilde{z}_{p'}^t$ is the latent embedding of the positive sample for task t, and $\mathcal{P}(p,s)$ are positives with the same label for task t.

For comparison as an expected lower bound, we also use unsupervised contrastive learning, using single-head unsupervised learning loss(Chen et al., 2020), defined as:

$$\mathcal{L}_{unsupcont}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(p,s)\in\mathcal{B}} -\log \frac{\exp(z_{ps} \cdot \tilde{z}_{ps}/\tau)}{\sum_{a\in\mathcal{A}(p,s)} \exp(z_{ps} \cdot z_a/\tau)} \tag{3}$$

where $\tilde{z}_{ps}$ is the augmented positive sample, $\mathcal{A}(p,s)$ are negative samples, and τ is the temperature parameter.

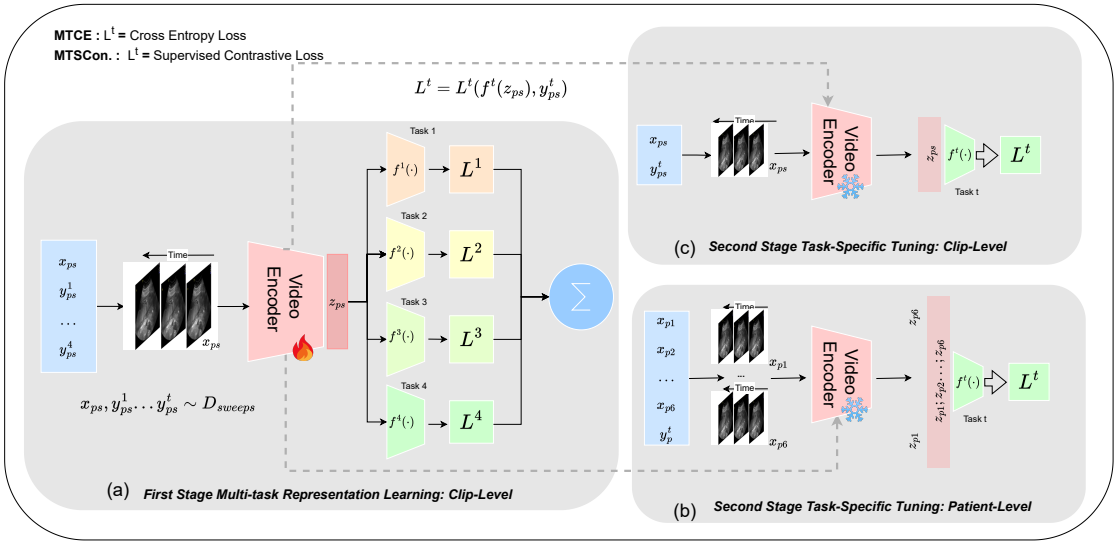


Figure 3: First stage pretrains an encoder using each sweep video as a sample, with either MTCE or MTSCon loss (a). Second stage finetunes the pretrained encoder with a single task-specific head for downstream prediction tasks, either at patient level using concatenated patient-level representations (b), or at sweep video level(c).

Second Stage: Task-Specific Tuning : These were performed either on **patient-level** or **sweep-level** depending upon tasks. The dataset is defined as $\mathcal{D}_{patient} = \{(X_p, y_p)\}_{p=1}^N$, where $X_p = \{x_{ps}\}_{s=1}^6$ and $y_p \in \mathcal{Y}^{(t)}$. After pretraining, we fine-tune the model either at the patient level or the sweep level depending on the downstream task. For patient-level fine-tuning, each patient p is associated with a collection of sweeps $X_p = \{x_{ps}\}_{s=1}^6$, and the representations of all sweeps are aggregated using a function $g(\cdot)$, specifically concatenation: $Z_p = [z_{p1}; z_{p2}; \ldots; z_{p6}]$, where $z_{ps} = f_\theta(x_{ps})$ is the representation of the s-th sweep for patient p. The aggregated representation Z_p is then used for downstream prediction tasks such as classification or regression via $\hat{y}_p = h_\phi(Z_p)$, where h_ϕ is the task-specific head trained during fine-tuning. For sweep-level fine-tuning, each sweep is treated independently, using $z_{ps} = f_\theta(x_{ps})$ directly with the task-specific head: $\hat{y}_{ps} = h_\phi(z_{ps})$. After sweep-level

fine-tuning, predictions can either be aggregated across all sweeps of a patient for patient-level tasks, or used directly for sweep-level tasks, depending on the evaluation setting. Aggregating multiple views at the patient level could enable the model to leverage richer contextual information, making patient-level predictions more robust compared to sweep-level predictions (Figure 3).

4. Experiments

4.1. Datasets and Task Description

Table 1: Training stages, tasks, and data distribution.

Stage	Purpose	Tasks	Data Split (study, 6 sweep each)	Notes
1	Pretraining *(Clip-level)*	**GAcls_bin** (4 age groups, $2^{nd}/3^{rd}$ Trim.); **FPcls**; **PLcls**; **STcls**	2,000/400 (Train/Valid)	Novel task **STcls** added to enhance multi-task learning via anatomical variation.
2-a	In-domain task tuning	**GAreg** (days); **FPcls**; **PLcls**; **STcls**	463/152/152 (Train/Valid/Test)	GA treated as regression for clinical realism.
2-b	Out-domain tasks Generalization	**DVPreg** (cm); **MPcls**	**DVP**: Same split as 2a; **MP***: 85/28/29 (Train/Valid/Test)	MP subset excludes most singleton pregnancies.
2-c	Robustness (Dist. Shift)	*Same as in-domain tasks*	**Obese**:40/19/23; **Butterfly**:66/22/22 (Train/Valid/Test)	Held out from Stg 1 for BMI & device-based shift testing.

Abbreviations: **GAcls_bin**: Gestational Age Classification (4 binned age groups in 2^{nd} and 3^{rd} Trimester); **FPcls**: Fetal Presentation Classification(**Cephalic vs Non-Cephalic**); **PLcls**: Placenta Location Classification(**Anterior vs Posterior**); **STcls**: Sweep Type Classification(**6 sweeps** (Figure 1a)); **GAreg**: Gestational Age Regression; **DVPreg**: Deepest Vertical Pocket Regression; **MPcls**: Multiple Pregnancy Classification. Splits for **MP** are at clip level rather than studies.

We used subsets of the FAMLI2 dataset (Pokaprakarn et al., 2022) in a two-stage framework (see Table 1 and Figure 3). The access to this dataset was obtained after an external request to the funding agency of this dataset, who is funding us a separate project related to obstetrics ultrasound. Stage 1 jointly trained selected classification tasks on large-scale clip-level data (Table 1, 1^{st} row), while Stage 2 fine-tuned and evaluated on data splits for in-domain tasks i.e. the ones used for Stage-1 training (Table 1, 2^{nd} row), out-domain classification and regression tasks (Table 1, 3^{rd} row) and including robustness scenarios with distribution shift (Table 1, 4^{th} row). Tuning for classification tasks i.e. Sweep Tag (ST_{cls}) and Multiple Pregnancy detection (MP_{cls}) were done at sweep level (Figure 3c), Fetal Presentation and Placenta Location at patient level (Figure 3b) whereas, regression

tasks i.e. Gestational Age Estimation(GA_{reg} - days) and Deepest Vertical Pocket Estimation(DVP_{reg} - cm) were done at sweep level with test prediction done at patient level using inverse variance weighting, which was outputed alongside regressed value and trained with heteroscedastic Gaussian loss. See abbreviations in Table 1 for categories used for classification tasks.

4.2. Implementation Details

We use a 3D ResNet (Tran et al., 2018) encoder initialized with Kinetics-400 pretrained weights, producing 512-dimensional sweep-level representations that are projected to 256 dimensions per contrastive head. First-stage training employs the Adam optimizer with a learning rate of 6×10^{-5}, batch size of 32, and up to 30 epochs, with learning rate scheduling via `ReduceLROnPlateau` (patience = 6, factor = 0.8). In the second stage, task-specific classification heads consist of two hidden layers of 256 dimensions with ReLU activations and a dropout rate of 0.15, trained with the same optimizer and scheduler but a higher learning rate of 3×10^{-4} for maximum of 15 epochs. Weighted cross-entropy and heteroscedastic Gaussian losses are used for classification and regression tasks, respectively. To improve robustness, domain-specific augmentations including speckle noise, random black patches, motion blur, brightness artifacts and rotations are applied to the blind ultrasound sweeps. Same augmentations were applied for our training our unsupervised model.

5. Results

Table 2: Performance of the proposed joint representation learning paradigms: MTCE and MTSCon, compared to the STSL baseline and UCon. Here, UCon is the expected lower bound. **Metrics**: Accuracy / macro-averaged F1 score (F1-macro) for classification tasks; Mean Absolute Error (MAE) / std deviation (std) for regression tasks.

	In-domain tasks				Out-of-domain tasks	
Loss	**GA**$_{\text{reg}}(\downarrow)$	**FP**$_{\text{cls}}(\uparrow)$	**PL**$_{\text{cls}}(\uparrow)$	**ST**$_{\text{cls}}(\uparrow)$	**DVP**$_{\text{reg}}(\downarrow)$	**MP**$_{\text{cls}}(\uparrow)$
STSL	16.2/13.2	85.7/77.2	93.5/93.5	**82.3/82.3**	1.13/0.92	46.4/46.4
UCon.	24.8/17.4	70.8/58.8	61.7/61.6	37.4/37.1	**1.11/0.89**	53.6/48.2
MTCE	15.7/12.2	**88.3/82.2**	**96.1/96.1**	82.0/82.1	1.13/0.87	53.4/44.9
MTSCon.	**14.3/11.2**	79.9/72.7	90.9/90.8	80.6/80.6	**1.11/0.91**	**57.1/53.3**

We conduct three primary analyses:

- Compare our multi-task representation learning with two baselines—Single Task Supervised Learning (STSL), using the same two-stage setup for in-domain tasks and single-stage for out-of-domain tasks, and Unsupervised Contrastive Learning (SIMCLR (Chen et al., 2020)) as a lower bound.
- Perform ablations by removing individual pretext tasks (Table 3).

- Assess robustness against distribution shifts (Table 4).

Multi-head Representation Learning outperforms single-task models: Supervised multi-task and multi-head contrastive models consistently outperform single-task, with the exception of Sweep Tag Classification where MTCE nearly matches the best-performing STSL method (Table 2). MTCE shows particularly strong performance on Fetal Presentation and Placenta Location classification, achieving accuracies of **88.3%** and **96.1%**, respectively. MTSCon further improves upon MTCE, especially on GA regression and out-of-domain tasks, highlighting its ability to more effectively integrate multi-task features for robust generalization to unseen scenarios. STSL also attains competitive accuracy on FP and ST, likely benefiting from the large-scale first-phase supervised pretraining. Overall, these results support our hypothesis that joint representation learning guided by BSOU-specific constraints yields richer and more transferable representations.

Role of Task Interrelationships :

Table 3 illustrates how different task combinations influence performance under the MTCE framework. While multi-task learning generally outperforms single-task models (Table 2), its effectiveness depends strongly on the specific set of jointly learned tasks. All configurations except ST+PL+GA yield notably high accuracy for fetal presentation (88.3% and 87.7%). The relatively lower performance of ST+PL+GA is likely due to presentation labels being omitted during its first-phase training. Nevertheless, it still achieves a 5.1% higher accuracy (75.9%) than the unsupervised model (70.8%, Table 2), indicating that the additional tasks positively contributed to its performance. ST+PL+GA and PL+FP+GA provide the best results for gestational age estimation. Moreover, ST+PL+GA also proves to be a strong configuration for multiple pregnancy classification. However with some task combinations the results are poorer than even the single task models. These findings highlight further need to carefully consider tasks when performing joint learning.

Table 3: Impact of different task combinations during 1^{st} stage multi-task training on individual tasks' performance after second stage task-specific training. **Metrics**: Accuracy / macro-averaged F1 score (F1-macro) for classification tasks; Mean Absolute Error (MAE) / std deviation (std) for regression tasks.

	In-domain tasks		Out-of-domain tasks	
Multi-Tasks	**GA**$_{reg}(\downarrow)$	**FP**$_{cls}(\uparrow)$	**DVP**$_{reg}(\downarrow)$	**MP**$_{cls}(\uparrow)$
ST+PL+FP	21.0/15.0	**88.3/82.6**	1.09/0.94	60.7/44.9
ST+PL+GA	**14.0/11.8**	75.9/69.9	1.10/0.92	**71.4/67.2**
ST+FP+GA	17.7/12.6	87.7/83.6	1.09/0.89	50.0/33.3
PL+FP+GA	14.15/12.35	87.7/83.6	1.09/0.98	64.2/59.1

Robustness against distribution shift :

As shown in Table 4, the multi-head approach maintains strong performance under distribution shifts for fetal presentation classification, achieving 81.8% accuracy on obese patients with MTCE and 66.7% accuracy on handheld devices. In contrast, for placenta

Table 4: Robustness against distribution shift: a) **Obese** Patients($BMI > 30$) and b) Blind sweeps taken by **novices** using **low-cost handheld device(Butterfly)**. Binary classification was performed for Fetal Presentation (Cephalic vs. Non-cephalic) and Placenta Location (Anterior vs. Posterior). **Metrics**: Accuracy / macro-averaged F1 score (F1-macro) for classification tasks; Mean Absolute Error (MAE) / std deviation (std) for regression tasks.

Domain	**Loss**	**In-domain tasks**			**Out-of-domain tasks**	
		$\mathbf{GA}_{reg}(\downarrow)$	$\mathbf{FP}_{cls}(\uparrow)$	$\mathbf{PL}_{cls}(\uparrow)$	$\mathbf{ST}_{cls}(\uparrow)$	$\mathbf{DVP}_{reg}(\downarrow)$
Obese	STSL	38.4/18.8	68.2/51.1	**81.8/77.1**	56.6/56.4	**1.18/0.81**
	UCon.	36.6/22.6	63.6/38.9	63.6/38.9	21.7/19.9	1.23/0.74
	MTCE	**33.7/20.8**	72.7/68.6	68.2/51.1	**59.4/59.2**	1.22/0.67
	MTSCon.	36.2/18.3	**81.8/74.1**	63.6/38.9	54.5/53.1	1.24/0.70
Handheld	STSL	**36.5/22.4**	61.9/57.1	**71.4/69.7**	42.7/42.5	0.98/0.74
Device	UCon.	41.6/23.7	57.1/53.3	61.9/38.2	22.1/21.1	**0.96/0.73**
(Butterfly)	MTCE	38.8/35.0	**66.7/61.0**	66.7/66.3	**42.7/42.5**	1.01/0.77
	MTSCon.	40.9/24.1	66.6/40.0	76.2/72.1	35.1./31.2	1.04/0.79

location, the multitask models fail to match the robustness of the single-task baseline under these shifts. Similarly, no multitask method surpasses the single-task models on $\mathrm{DVP}_{\mathrm{reg}}$, highlighting a persistent gap in performance under distribution shift. These observations underscore the need for further research on device-level generalization and task-specific robustness.

6. Discussion and Conclusion

We presented a novel multi-head representation learning framework for blind ultrasound sweeps demonstrating its capacity to improve performance on both in-domain and out-of-domain tasks. We found that joint representation learning approach often offers superior generalization and robustness compared to single-task approach. Our findings also show that careful task combination and feature selection are important for BSOU dataset in multi-task setup.

While the framework exhibits strong performance across various patient characteristics, we also identified a critical challenge: maintaining robustness under distribution shifts caused by different ultrasound device types. This area represents a significant frontier for future research, with promising directions including the development of models capable of effectively adapting to diverse device characteristics and the exploration of foundation models specifically designed for ultrasound sweep videos to further enhance robustness and scalability in this domain. Similarly, during the first stage of training, we propagated patient-level labels that come with the dataset to each of the sweeps, similar to all current literature. However, these labels are not always correct for each sweep videos, potentially reducing the performance of multitasking models we proposed. We will look into building robustness against this label noise in future.

References

Alfred Abuhamad, Yili Zhao, Sharon Abuhamad, Elena Sinkovskaya, Rashmi Rao, Camille Kanaan, and Lawrence Platt. Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination. *American journal of perinatology*, 2(01): 090–098, 2016.

AIMIX Project. Inclusive artificial intelligence for accessible medical imaging across resource-limited settings. https://aimix-erc.eu/, 2025. Accessed: 2025-07-03.

Tanya Akumu, Marawan Elbatel, Victor M. Campello, Richard Osuala, Carlos Martin-Isla, Ignacio Valenzuela, Xiaomeng Li, Bishesh Khanal, and Karim Lekadir. Adaptive Frame Selection for Gestational Age Estimation from Blind Sweep Fetal Ultrasound Videos . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15973. Springer Nature Switzerland, September 2025.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. URL https://api.semanticscholar.org/CorpusID:45998148.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Kristen K. DeStigter, G. Eli Morey, Brian S. Garra, Matthew R. Rielly, Martin E. Anderson, Michael G. Kawooya, Alphonsus Matovu, and Frank R. Miele. Low-cost teleradiology for rural ultrasound. In *2011 IEEE Global Humanitarian Technology Conference*, pages 290–295, 2011. doi: 10.1109/GHTC.2011.39.

Zeyu Fu, Jianbo Jiao, Robail Yasrab, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Anatomy-aware contrastive representation learning for fetal ultrasound. In *European Conference on Computer Vision*, pages 422–436. Springer, 2022.

A. D. Gleed, D. Mishra, V. Chandramohan, Z. Fu, A. Self, S. Bhatnagar, A. T. Papageorghiou, and J. A. Noble. Towards multi-sweep ultrasound video understanding: Application in detection of breech position using statistical priors. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023a. doi: 10.1109/ISBI53787.2023.10230662.

Alexander D. Gleed, Qingchao Chen, James Jackman, Divyanshu Mishra, Varun Chandramohan, Alice Self, Shinjini Bhatnagar, Aris T. Papageorghiou, and J. Alison Noble. Automatic image guidance for assessment of placenta location in ultrasound video sweeps. *Ultrasound in Medicine Biology*, 49(1):106–121, 2023b. ISSN 0301-5629. doi: https://doi.org/10.1016/j.ultrasmedbio.2022.08.006. URL https://www.sciencedirect.com/science/article/pii/S0301562922005294.

Alexander D. Gleed, Divyanshu Mishra, Alice Self, Ramachandran Thiruvengadam, Bapu Koundinya Desiraju, Shinjini Bhatnagar, Aris T. Papageorghiou, and J. Alison Noble. Statistical characterisation of fetal anatomy in simple obstetric ultrasound

video sweeps. *Ultrasound in Medicine Biology*, 50(7):985–993, 2024. ISSN 0301-5629. doi: https://doi.org/10.1016/j.ultrasmedbio.2024.03.006. URL https://www.sciencedirect.com/science/article/pii/S0301562924001352.

Ryan G Gomes, Bellington Vwalika, Chace Lee, Angelica Willis, Marcin Sieniek, Joan T Price, Christina Chen, Margaret P Kasaro, James A Taylor, Elizabeth M Stringer, et al. A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment. *Communications Medicine*, 2(1):128, 2022.

Juncheng Guo, Guanghua Tan, Fan Wu, Huaxuan Wen, and Kenli Li. Fetal ultrasound standard plane detection with coarse-to-fine multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 27(10):5023–5031, 2023. doi: 10.1109/JBHI.2022.3209589.

Dongli He, Hu Wang, and Mohammad Yaqub. Advancing fetal ultrasound image quality assessment in low-resource settings. *arXiv preprint arXiv:2507.22802*, 2025.

Leila Kalantari, Subhendu Seth, Manikanda Krishnan, Jonathan Sutton, Melanie Jutras, and Anuradha Rao. Multiple pregnancy detection from ultrasound blind sweeps. In *2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS)*, pages 1–4, 2024. doi: 10.1109/UFFC-JS60046.2024.10794008.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Zehui Lin, Shengli Li, Dong Ni, Yimei Liao, Huaxuan Wen, Jie Du, Siping Chen, Tianfu Wang, and Baiying Lei. Multi-task learning for quality assessment of fetal head ultrasound images. *Medical Image Analysis*, 58:101548, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101548. URL https://www.sciencedirect.com/science/article/pii/S1361841519300830.

Emily Mu, John Guttag, and Maggie Makar. Multi-similarity contrastive learning. *arXiv preprint arXiv:2307.02712*, 2023.

Priyam Patel, Leila Kalantari, Shyam Bharat, Soheil Borhani, Stephen Schmidt, Melanie Jutras, and Jonathan Sutton. Deep learning based gestational age estimation with outlier elimination in blind sweep fetal ultrasound. In *2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS)*, pages 1–5. IEEE, 2024.

Szymon Płotka, Tomasz Włodarczyk, Adam Klasa, Michał Lipa, Arkadiusz Sitek, and Tomasz Trzciński. Fetalnet: multi-task deep learning framework for fetal ultrasound biometric measurements. In *International Conference on Neural Information Processing*, pages 257–265. Springer, 2021.

Teeranan Pokaprakarn, Juan C Prieto, Joan T Price, Margaret P Kasaro, Ntazana Sindano, Hina R Shah, Marc Peterson, Mutinta M Akapelwa, Filson M Kapilya, Yuri V Sebastião, et al. Ai estimation of gestational age from blind ultrasound sweeps in low-resource settings. *NEJM evidence*, 1(5):EVIDoa2100058, 2022.

M. Sofia Sappia, Chris L. de Korte, Bram van Ginneken, Dean Ninalga, Satoshi Kondo, Satoshi Kasai, Kousuke Hirasawa, Tanya Akumu, Carlos Martín-Isla, Karim Lekadir, Victor M. Campello, Jorge Fabila, Anette Beverdam, Jeroen van Dillen, Chase Neff, and Keelin Murphy. Acouslic-ai challenge report: Fetal abdominal circumference measurement on blind-sweep ultrasound data from low-income countries. *Medical Image Analysis*, 105:103640, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103640. URL https://www.sciencedirect.com/science/article/pii/S1361841525001872.

Martijn Schilpzand, Chase Neff, Jeroen van Dillen, Bram van Ginneken, Tom Heskes, Chris de Korte, and Thomas van den Heuvel. Automatic placenta localization from ultrasound imaging in a resource-limited setting using a predefined ultrasound acquisition protocol and deep learning. *Ultrasound in Medicine Biology*, 48(4):663–674, 2022. ISSN 0301-5629. doi: https://doi.org/10.1016/j.ultrasmedbio.2021.12.006. URL https://www.sciencedirect.com/science/article/pii/S0301562921005202.

Alice Self, Qingchao Chen, Bapu Koundinya Desiraju, Sumeet Dhariwal, Alexander D Gleed, Divyanshu Mishra, Ramachandran Thiruvengadam, Varun Chandramohan, Rachel Craik, Elizabeth Wilden, 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 Research Protocols*, 11 (9):e37374, 2022.

Jeffrey SA Stringer, Teeranan Pokaprakarn, Juan C Prieto, Bellington Vwalika, Srihari V Chari, Ntazana Sindano, Bethany L Freeman, Bridget Sikapande, Nicole M Davis, Yuri V Sebastião, et al. Diagnostic accuracy of an integrated ai tool to estimate gestational age from blind ultrasound sweeps. *Jama*, 332(8):649–657, 2024.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. doi: 10.1109/CVPR.2018.00675.

Thomas L.A. van den Heuvel, Hezkiel Petros, Stefano Santini, Chris L. de Korte, and Bram van Ginneken. Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries. *Ultrasound in Medicine and Biology*, 45(3):773–785, 2019. doi: 10.1016/j.ultrasmedbio.2018.09.015. URL https://doi.org/10.1016/j.ultrasmedbio.2018.09.015.

Jakub Maciej Wiśniewski, Anders Nymark Christensen, Mary Le Ngo, Martin Grønnebæk Tolsgaard, and Chun Kit Wong. Determining fetal orientations from blind sweep ultrasound video. In Jens Petersen and Vedrana Andersen Dahl, editors, *Image Analysis*, pages 254–263, Cham, 2025. Springer Nature Switzerland.

Bo Zhang, Han Liu, Hong Luo, and Kejun Li. Automatic quality assessment for 2d fetal sonographic standard plane based on multitask learning. *Medicine (Baltimore)*, 100(4): e24427, Jan 2021. doi: 10.1097/MD.0000000000024427. URL https://doi.org/10.1097/MD.0000000000024427. PMCID: PMC7850658.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022. doi: 10.1109/TKDE.2021.3070203.