Automated Ultrasound Substructure Localization

Anastasiia Statcenko, Andrew Hinh, Gigi Patmore, Pranav Chainani

ENGR 110

Winter 2025

Dr. Jessica Kuczenski

March 18, 2025

Maternal Health Foundation

Introduction

The Maternal Health Foundation (MHF) has an overall vision of a world without childbirth injuries. Their fund supports renowned doctors and researchers who are enabling efficient and accurate treatment of complications during childbirth. It is an issue facing countries worldwide, with about 287,00 women dying during and following pregnancy in 2020 [1]. This issue is even more prevalent in low-income countries, where healthcare is not as advanced and harder to come by, and where most of these deaths occur. The current focus is in Sub-Saharan Africa where, "the nearest healthcare facility is often a two-day walk, resulting in a 1in 16 risk of dying from pregnancy or childbirth," [2]. The MHF hopes to make accurate maternal health diagnosis and treatment accessible in places where the problem is the most prominent. Working in these areas provides unique challenges due to the lack of healthcare infrastructure and availability for those living there, especially in remote areas. Therefore, MHF and other foundations must explore unique solutions that mix techniques for on-the-ground care and continuing education to prevent future injuries [2].

One proposed solution is the use of artificial intelligence (AI) as a way to increase the efficiency and accuracy of ultrasound labeling. By finding irregularities early and accurately, personalized care can be provided immediately increasing the safety of both mother and child. Projects using datasets such as the FPUS23 phantom dataset have proved that high levels of accuracy in labeling the fetus can be achieved using neural networks and other machine learning technology [3]. The growth of this idea has led to the creation of multiple datasets and challenges, which aim at encouraging teams to create accurate coding solutions for ultrasound and brain labeling across various datasets and label types. An example is the FeTA challenge which focuses on fetal MRI brain labeling of different tissue and matter, which could then be applied to detect deformities if percentages don't match the expected amount [4]. The implementation of this technology in remote areas solves a variety of problems. As ultrasound technology advances and prenatal screenings become more accessible, making these procedures quicker allows volunteer doctors to see more patients, and therefore potentially prevent more injuries and deaths. These programs could also label issues that doctors may have missed, offering an on-the-spot second opinion. Additionally, AI labeling, if developed enough, would be able to be used by a midwife or equivalent in combination with portable ultrasound equipment, allowing for continued care in areas that lack doctors. The wide applications of this technology provide evidence for why the development of an accurate large language model (LLM) for assisted labeling is in the interests of the MHF.

From discussions with MHF and research on maternal healthcare challenges, several key needs and constraints have been identified. Healthcare providers in underprivileged or low-resource areas require a reliable and efficient method for labeling fetal brain ultrasound images, particularly in environments where specialists are scarce. The system must offer high accuracy while being fast, intuitive, and easy to use, allowing midwives and clinicians with limited training to integrate it into their existing workflows without adding a significant burden.

One of the main challenges is the lack of access to abnormal fetal ultrasound datasets, which makes training and validating a robust labeling system more difficult. As it was mentioned above, our target users are located in very low-resource environments. Given these constraints, the project must balance efficiency, computational feasibility, and diagnostic accuracy to ensure practical implementation in real-world settings.

The primary objective of this project is to develop a method for improving ultrasound image labeling that is both effective and practical for use in low-resource settings. To achieve this, various models were proposed as a solution to assess how accuracy, efficiency, and computational complexity interact. The client has emphasized the need to minimize latency while also reducing the number of parameters without sacrificing accuracy. Since computational power is becoming more affordable, the goal is to design a system that can eventually operate independently of external computing resources. Successful implementation will be measured through the accuracy of labeled images, the processing time per scan, minimal latency, and ideally the system's ability to run on the given hardware without internet connectivity.

The primary users of this system include doctors, sonographers, and midwives working in areas where maternal healthcare services are limited or overburdened. Doctors and sonographers in small clinics and hospitals will benefit from a more structured and efficient way to assess fetal brain scans, while midwives, who often serve as the first point of contact for pregnant women in remote areas, can use the system to support their clinical judgment and identify potential concerns early. By enhancing the speed and reliability of ultrasound interpretation, this project contributes to MHF's broader mission of reducing maternal and infant mortality. With better tools to detect complications early, healthcare providers can offer more timely interventions, reduce preventable deaths, and improve overall maternal health outcomes. The long-term goal is to create a solution that is not only effective but also sustainable, ensuring that healthcare workers in underprivileged regions have access to the resources they need to provide quality care to every mother and child.

In the rest of this paper, we will begin with a discussion about how we designed this system to assist in labeling ultrasound images. More specifically, we will:

- Describe the constraints of the project, as specified by MHF's Chief AI Officer Mahni Shayganfar, and its practical impact and implications.
- Review existing solutions, their pitfalls, and the key differentiators for our solution,
- Discuss the dataset we chose to train and evaluate the system alongside alternatives we did not use, and
- Explain the system design we implemented and alternative designs we did not proceed with.
- Explore the tests we chose to evaluate our system and the corresponding technical and business success metrics.

Then, we will conclude with our opinion on the project's success and reasoning for its downsides. We will also discuss recommended next steps for anybody who chooses to expand on the ideas put forth in this project. Finally, we will attach appendices containing calculations, test results, a bill of materials, and team and project management materials.

Table of Contents

Abstract

This project aimed to develop a Large Language Model (LLM)-based approach for automated ultrasound labeling to assist healthcare providers and improve maternal care in low-resource settings. By using a small vision language model (Qwen2.5-VL-3B-Instruct) [6], we created a program to automate ultrasound labeling, allowing healthcare professionals such as midwives and doctors to more quickly and accurately identify potential fetal abnormalities. This model was trained and evaluated on a dataset containing seven fetal brain substructures and focused on metrics such as Hausdorff distance, Euclidean distance, precision, recall, and F1-score to assess performance. Initial evaluations showed significant challenges in terms of accuracy, including overfitting and high false positive rates. However, we observed improvements in our Hausdorff and Euclidean distance metrics through supervised fine turning and task reformulation, switching to predicting substructures individually rather than simultaneously. This came with trade-offs such as the occasional increase in false positives and issues in precision indicating areas for future optimization. Future work should focus on dataset expansion model optimization and real-world clinical testing to ensure the program's reliability and scalability. Overall, our findings serve as a proof of concept demonstrating the feasibility of using AI for ultrasound image analysis, though further refinements are necessary.

Discussion

The development of the LLM-based program for fetal brain anomaly detection posed a significant challenge due to the lack of a labeled pathological dataset. Without abnormal fetal brain scans, traditional supervised learning methods were not an option. To address this, three main approaches were explored.

1. Generating synthetic pathological data using AI.
2. Training models to detect deviations from normal patterns.
3. Using heuristic-based classification methods.

The final approach selected was training models to identify deviations from normal fetal brain structures, as it provided the highest accuracy while remaining feasible given the dataset constraints. By fine-tuning Qwen2.5-VL-3B using supervised fine-tuning (SFT) on the majority of the dataset (subset held out for validation and testing), the model was optimized for efficient and precise anomaly detection without requiring explicit abnormal case labels.

Existing research in fetal brain ultrasound analysis has primarily focused on deep-learning approaches that classify images as normal or abnormal. A study by Xie et al. (2020) developed a CNN-based supervised learning model, which was trained on a large dataset of 15,372 normal and 14,047 abnormal fetal brain images [1]. The model performed image segmentation, binary classification, and lesion localization using heat maps, achieving a 96.3% classification accuracy. However, it relied on a large dataset with verified abnormal cases, which was unavailable for our project. Unlike their fully supervised approach, our model was trained only on normal images and used self-supervised learning to identify abnormalities by detecting deviations from learned patterns. This made our approach more practical for real-world applications where labeled pathological datasets are limited.

A review by Weichert and Scharf (2024) examined AI applications in fetal neurosonography, highlighting image analysis, automated measurement, prediction models, and visualization techniques [2]. Their discussion of deep learning models for detecting neurodevelopmental deviations closely aligns with our approach, as both methods identify abnormalities based on deviations from expected patterns rather than direct labels. However, most models they reviewed used 3D ultrasound imaging and multiplanar reconstructions, whereas our approach focused on 2D segmented images. Additionally, their studies often relied on large-scale, multi-center datasets, while we worked within the constraints of a limited dataset containing only normal cases. Some models in their review used predefined neurodevelopmental atlases as reference points, whereas our method dynamically inferred abnormalities using contrastive learning.

Despite these differences, both studies support the validity of deviation-based anomaly detection as a viable alternative when labeled pathological data is unavailable. While fully supervised models like Xie et al. perform well when large datasets are accessible, our self-supervised learning approach offers greater adaptability in data-limited scenarios, making it a valuable contribution to LLM-based prenatal screening.

To assess the model's accuracy in detecting fetal brain substructures, evaluation metrics were established during the initial analysis phase. The original method involved substructure matching, where predicted structures were paired with ground truth structures, and false positives and false negatives were identified when mismatches occurred. Precision, recall, and F1-score were used at both the label level (substructure detection) and the point level (individual point matching). Euclidean distance averaging was initially chosen to measure spatial accuracy, while Hausdorff Distance was added to better capture the largest spatial discrepancies. Additionally, specificity and AUC-ROC were incorporated, and a thresholding method was introduced to refine precision and recall calculations at both levels.

Public organizations play a crucial role in shaping maternal healthcare policies and influencing the implementation of LLM-based solutions like our project. The Maternal Health Organization (MHF), our community partner, focuses on improving prenatal and postnatal care in Sub-Saharan Africa. Additionally, international organizations such as the World Health Organization (WHO) and the United Nations Population Fund (UNFPA) contribute significantly to maternal health initiatives. WHO promotes early detection and better diagnostic tools, aligning with our project's goal of LLM ultrasound data labeling analysis. UNFPA works to increase access to ultrasound technology in underserved regions, reinforcing the need for affordable AI-assisted maternal healthcare solutions.

Beyond international health organizations, regulatory agencies influence the deployment of medical AI technology. In the United States, the Food and Drug Administration (FDA), through its Center for Devices and Radiological Health (CDRH), regulates AI-powered medical software under the Software as a Medical Device (SaMD) framework. If our system were to be implemented in the U.S., it would need to undergo extensive clinical trials and regulatory evaluation to gain FDA approval. Internationally, agencies such as the European Medicines Agency (EMA) and the African Union Development Agency (AUDA-NEPAD) establish regional healthcare standards that would directly impact the deployment of our technology in different countries. These regulatory bodies ensure that AI-based diagnostics meet safety, efficacy, and ethical standards before clinical integration.

Policies on data privacy and security are particularly relevant since our LLM processes sensitive medical data. In the U.S., HIPAA (Health Insurance Portability and Accountability Act) and HITECH (Health Information Technology for Economic and Clinical Health Act) establish guidelines for protecting individually identifiable health information. Our project uses open-source ultrasound datasets that comply with existing privacy regulations, however, as the system expands, compliance with international data protection laws, such as the General Data Protection Regulation (GDPR) in Europe or African-specific healthcare privacy standards, would be necessary to ensure ethical and legal deployment. Additionally, any future implementation of this technology would need to account for ethical concerns related to LLM-based medical diagnostics, ensuring the system remains an assistive tool rather than replacing human medical judgment.

One of the central civic issues our project addresses is the lack of access to maternal healthcare in underserved and rural areas. In many regions, particularly in Sub-Saharan Africa, trained medical professionals are scarce, leaving midwives and nurses as the primary caregivers. By integrating AI-assisted ultrasound analysis into maternal healthcare, midwives can identify potential complications earlier, allowing for timely medical intervention. This approach has the potential to significantly reduce preventable maternal and infant mortality by ensuring that more individuals receive proper prenatal care. Furthermore, increasing the number of trained professionals capable of utilizing this technology would expand healthcare access in regions where hospitals and clinics are difficult to reach.

Public organizations, both governmental and non-governmental, play a critical role in addressing these healthcare challenges. WHO and UNFPA drive maternal health initiatives, influencing policies that prioritize early diagnostics and expand healthcare accessibility. Regulatory agencies such as the FDA and EMA enforce safety and ethical standards for medical AI deployment. Additionally, organizations focused on healthcare equity advocate for reducing racial and socioeconomic disparities, ensuring solutions are accessible, unbiased, and effective.

The role of public organizations extends beyond policy-setting, as their involvement determines the real-world feasibility of healthcare innovations. Collaboration with health organizations and regulatory bodies is essential for our project’s success. Partnering with institutions that support maternal healthcare advancement would help with regulatory compliance and infrastructure development. Through these collaborations, our AI-assisted ultrasound system could help reduce maternal mortality, improve healthcare accessibility, and ensure equitable health outcomes for underserved populations.

Results and Analysis

First, we conduct exploratory data analysis (EDA) on our data to ensure there are no missing or unclean data points. As shown in Figure 1, each ultrasound has seven substructures labeled with point-based outlines. This specifies the task we try to solve: given an ultrasound, can we return localized point-based outlines of each of the seven substructures? An important constraint given was the requirement to only use a small vision language model (VLM), in this case, Qwen2.5-VL-3B-Instruct, as the "predictor" for this task.

As shown in Figure 2, labels for all substructures (including an extra that was removed during ETL) are present for all ultrasounds. As shown in Figure 3, the number of points used to outline a substructure varies, ranging anywhere from two to eight points. The sample with the zero-point outline is removed during ETL. As shown in Figure 4, the distribution of xy-coordinates for each substructure varies and reveals interesting insight as to where each is located on average.

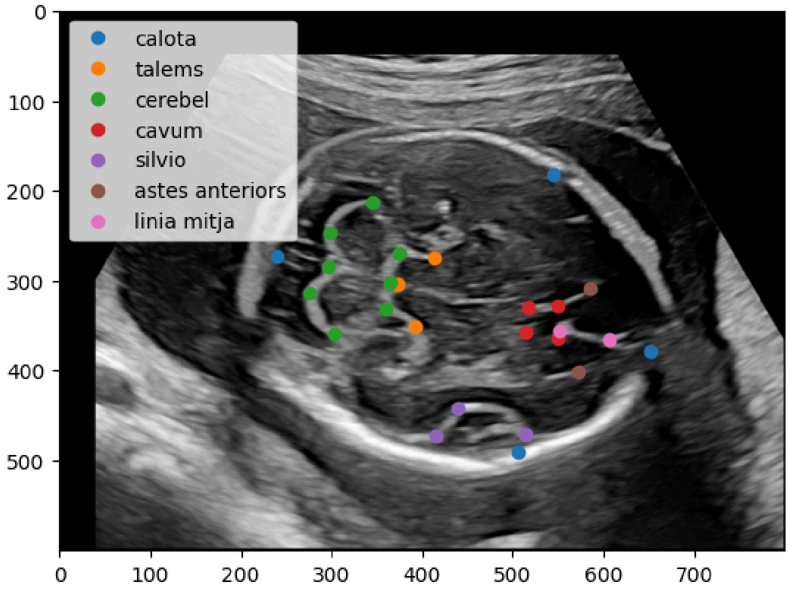

Figure 1: Example ultrasound with substructure locations displayed.

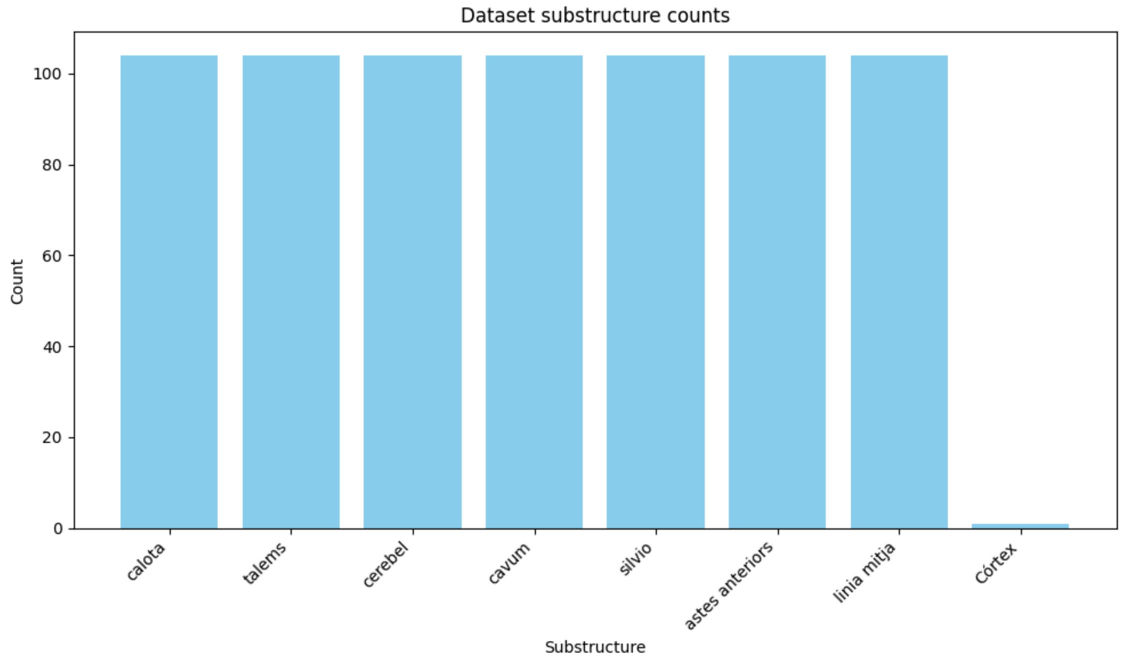

Figure 2: Counts of substructures in the dataset.

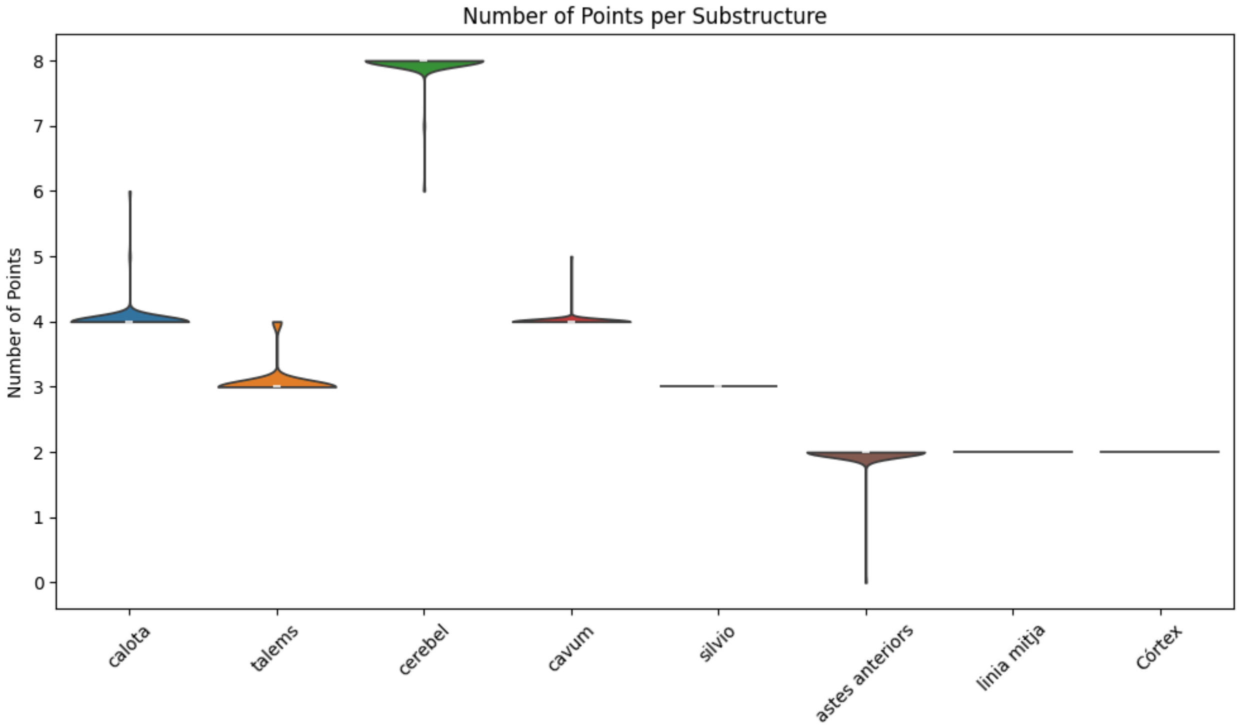

Figure 3: Number of points per substructure in the dataset.

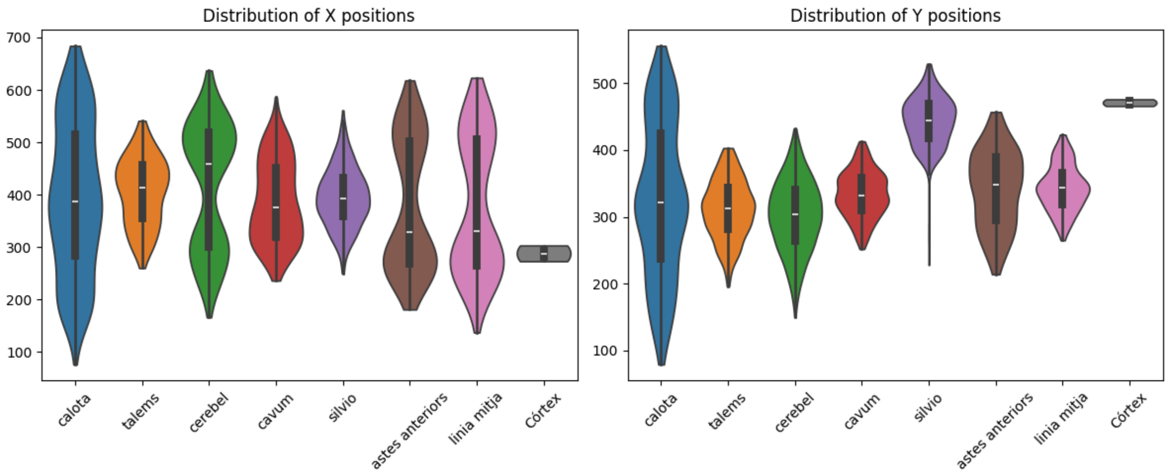

Figure 4: Distribution of x- and y- coordinates for each substructure in the dataset.

Then, we conduct a baseline evaluation where the model must predict the locations of all substructures in a single pass to gauge how well Qwen2.5-VL-3B-Instruct performs off the shelf. For all evaluations, a held-out test set of 11 samples is utilized. The metric we use is defined as follows:

- First, find the best one‑to‑one pairing between ground truth and predicted points to minimize total distance using Hungarian Matching, only matching pairs if they're within 20 pixels of each other.
- A true positive (TP) is a predicted point within the threshold of a real point.
- A false positive (FP) is an unmatched or too far-predicted point.
- A false negative (FN) is an unmatched ground-truth point (no close-enough prediction).
- Euclidean distance measures the average closeness of matched points.
- Hausdorff distance measures the worst-case distance among matched points
- Precision is the fraction of correct predictions.
- Recall is the fraction of real points successfully found.
- F1 is the harmonic mean of Precision & Recall.
- AUC-ROC defines how well TP is separated from FP across different thresholds.
- AUC-PR defines how well precision and recall are balanced across different thresholds.

Hausdorff and Euclidean distances are summed across all samples in the test set. The results of the baseline evaluation can be found in Table 1 below.

Table 1: Baseline evaluation, all substructures at once

substructure	hausdorff_distance	euclidean_distance	tp	fp	fn	precision	recall	f1	auc_roc	auc_pr
astes anteriors	3244.97	3708.49	0	62	22	0	0	0	0	0
calota	3351.69	7946.43	1	62	43	0.016	0.023	0.019	1	1
cavum	2798.88	6982.81	0	64	44	0	0	0	0	0
cerebel	2915.65	9029.97	1	69	87	0.014	0.011	0.013	1	1
linia mitja	3166.11	3712.88	0	61	22	0	0	0	0	0
silvio	3201.25	7328.32	0	61	33	0	0	0	0	0
talems	2358.98	4207.55	2	60	31	0.032	0.061	0.042	1	1

Then, we conduct an initial run of SFT on the aforementioned task and a corresponding evaluation in an attempt to improve and gauge model performance. As shown in Figure 5, four attempts to train the model were made. The initial run, represented by the blue curves, demonstrated extreme overfitting. The purple, red, and gray curves respectively show multiple attempts to correct for this by adjusting the following hyperparameters:

- *Gradient accumulation steps*: ranging from 1-8
  - Roughly the number of training examples to learn from at a time.
- *Learning rate*: ranging from 1e-6 to 3e-5
  - Roughly the speed of learning.
- *Warmup ratio*: ranging from 0.1 to 0.3
  - Roughly how fast to learn in the first few training steps.
- *Weight decay*: ranging from 1e-4 to 1e0
  - Roughly how much to dampen the model for better generalization.

To meet the constraint of low disk space, we quantize the model to a size of 3GB which can be found here: https://huggingface.co/andrewhinh/mhf-qwen2.5-vl-3b-instruct-full-sft

The evaluation results can be found in Table 2 below.

Table 2: SFT evaluation, all substructures at once

substructure	hausdorff_distance	euclidean_distance	tp	fp	fn	precision	recall	f1	auc_roc	auc_pr
astes anteriors	1713.24	3081.66	0	22	20	0	0	0	0	0
calota	2003.69	5521.22	2	42	44	0.045	0.043	0.044	1	1
cavum	1329.19	4994.83	0	40	40	0	0	0	0	0
cerebel	2499.23	16871.43	1	87	87	0.011	0.011	0.011	1	1
linia mitja	1757.64	3401.69	1	19	19	0.05	0.05	0.05	1	1
silvio	1164.87	3046.16	0	30	30	0	0	0	0	0
talems	1386.22	4089.02	0	33	33	0	0	0	0	0

Comparing the results between Tables 1 and 2, we see a 43.4% reduction in Hausdorff distance, 8.4% reduction in Euclidean distance, no change in number of true positives, 39% reduction in number of false positives, 4.6% reduction in number of false negatives, 19.96% increase in precision, 6.5% reduction in recall, a 5.4% increase in f1 score, a 33.3% reduction in AUC ROC, and a 33.3% reduction in AUC PR. Although the reduction in distance and false prediction errors are evident, they are not substantial.

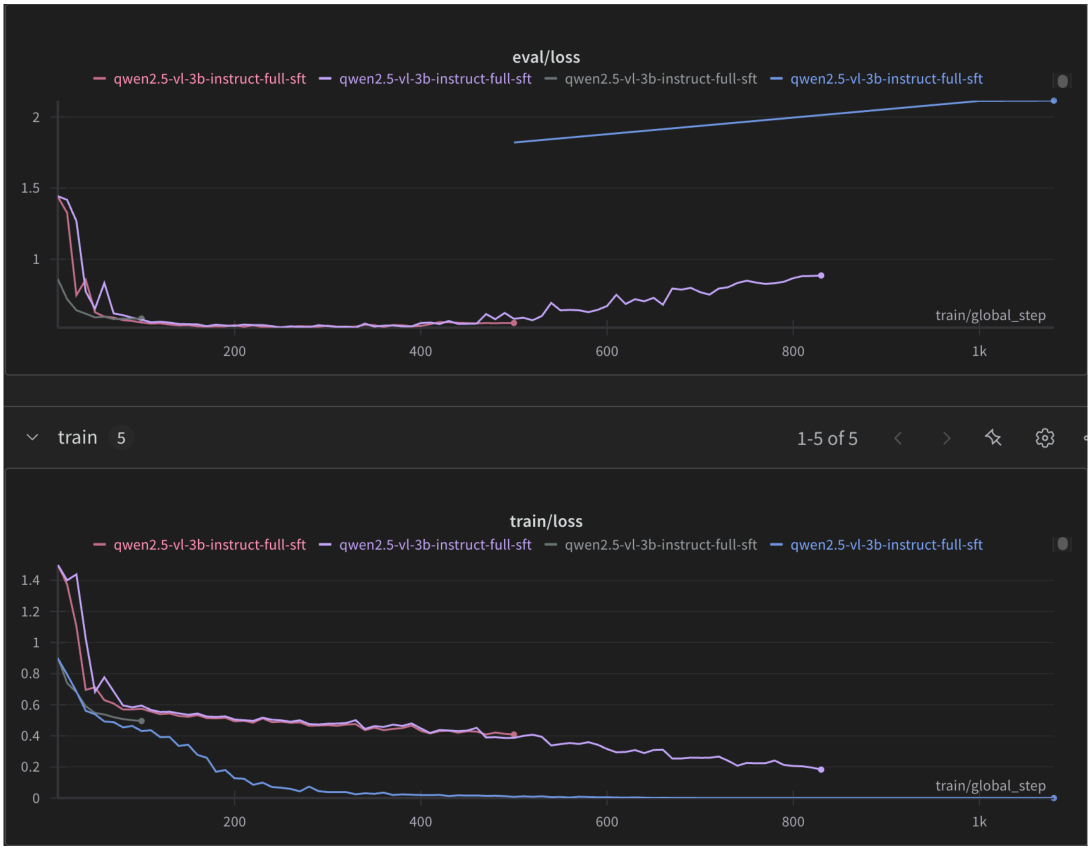

Figure 5: Training and validation loss curves for the initial SFT.

Then, we conduct a baseline evaluation where the model must predict the locations of only one substructure at a time, resulting in seven passes. This is done to answer the following question: can a simpler task for the model lead to less overfitting and improved performance? The results can be found in Table 3 below.

Table 3: Baseline evaluation, one substructure at a time

substructure	hausdorff_distance	euclidean_distance	tp	fp	fn	precision	recall	f1	auc_roc	auc_pr
astes anteriors	1927.33	2922.35	0	19	22	0	0	0	0	0
calota	2804.17	6890.46	0	43	46	0	0	0	0	0
cavum	1964.23	5789.37	0	44	44	0	0	0	0	0
cerebel	2080.79	8519.39	3	59	85	0.048	0.034	0.04	1	1
linia mitja	1808.33	1480.5	1	11	21	0.083	0.045	0.059	1	1
silvio	2248.34	3192.05	0	18	33	0	0	0	0	0
talems	1956.31	4671.32	0	33	33	0	0	0	0	0

Comparing the results between Tables 1 and 3, we see a 35.1% reduction in Hausdorff distance, a 30.9% reduction in Euclidean distance, a 14.3% reduction in the number of true positives, a 71.6% reduction in the number of false positives, a 0.9% increase in the number of false negatives, a 12.9% reduction in precision, a 14.0% reduction in recall, a 14.0% reduction in f1 score, a 28.6% reduction in AUC ROC, and a 28.6% reduction in AUC PR. The reductions in the distance and false prediction errors show similar if not better performance than SFT on the original task, clearly indicating this is a better formulated task that the model can perform better on.

Then, we conduct a second run of SFT on the aforementioned task and a corresponding evaluation for the same reason the initial SFT was done. As shown in Figure 6, two attempts to train the model were made. The two runs differ only in the number of steps. The following hyperparameters are used:

- *Freeze vision tower:* True
  - Whether to train the “vision” part of the VLM; more freezing enables less overfitting.
- *Freeze multimodal projector:* True
  - Whether to train the “vision to text” part of the VLM; more freezing enables less overfitting.
- *Train multimodal projector only:* True

- ○ Whether to only train the “vision to text” part of the VLM; more freezing enables less overfitting.

- *Gradient accumulation steps*: 1; allows for less overfitting.
- *Learning rate*: 3e-5; allows for the fastest training without performance degradation.
- *Learning rate scheduler type:* cosine
  - ○ How to vary learning rate, cosine generally performs well for Qwen models.
- *Warmup ratio*: 0.2; allows for fast training without overfitting.
- *Weight decay* = 1e-2; allows for maximal generalization.

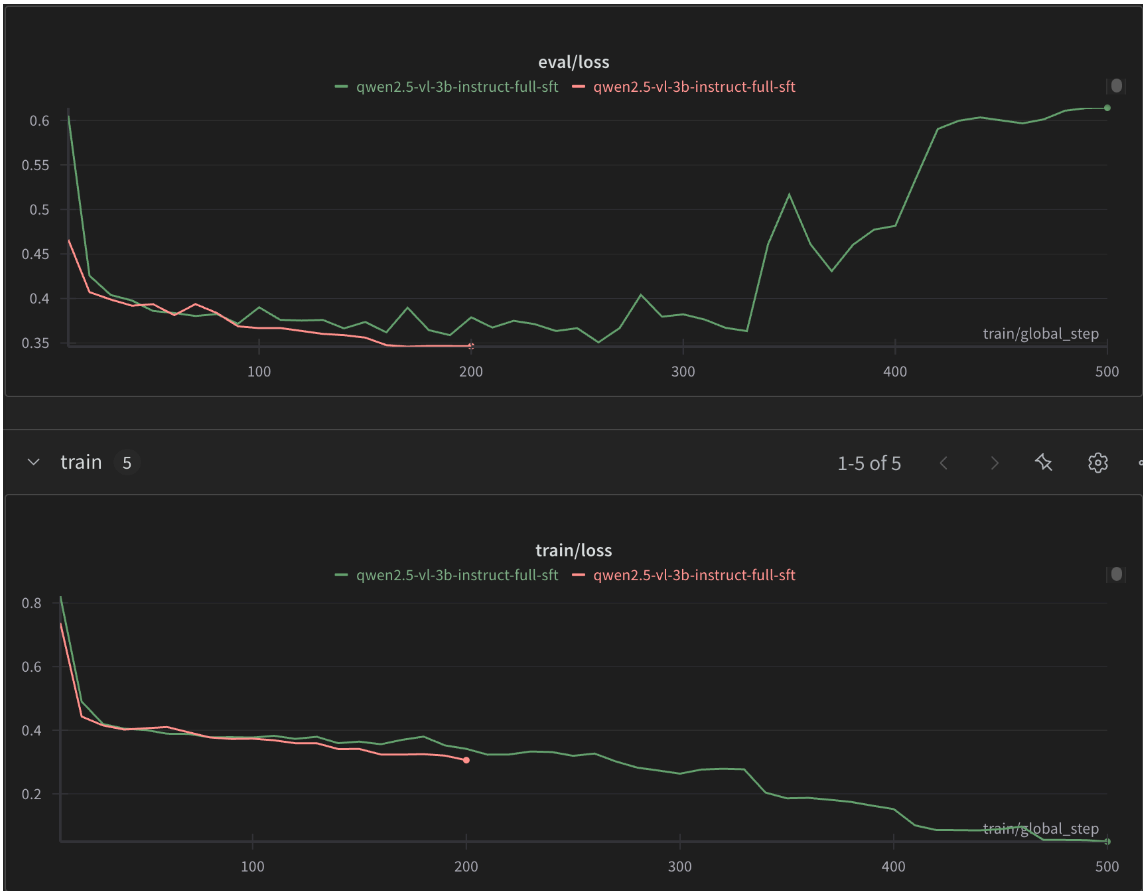

Figure 6: Training and validation loss curves for the second SFT.

Again, we quantize the model which can be found here:
https://huggingface.co/andrewhinh/mhf-qwen2.5-vl-3b-instruct-full-sft-awq

The evaluation results can be found in Table 4 below.

Table 4: SFT evaluation, one substructure at a time

substructure	hausdorff_distance	euclidean_distance	tp	fp	fn	precision	recall	f1	auc_roc	auc_pr
astes anteriors	1813.68	3359.78	0	20	20	0	0	0	0	0
calota	1155.2	3299.2	1	39	41	0.025	0.024	0.024	1	1
cavum	1023.78	3826.73	5	39	39	0.114	0.114	0.114	1	1
cerebel	1596.36	11241.99	4	84	84	0.045	0.045	0.045	1	1
linia mitja	1344.63	2573.79	0	22	22	0	0	0	0	0
silvio	883.69	2348	0	33	33	0	0	0	0	0
talems	1054.29	2742.11	0	33	34	0	0	0	0	0

Comparing the results between Tables 3 and 4, we see a 50.75% reduction in Hausdorff distance, a 14.06% reduction in Euclidean distance, a 32.65% increase in the number of true positives, a 20.54% increase in the number of false positives, a 1.07% reduction in the number of false negatives, a 27.65% increase in precision, a 32.55% increase in recall, a 30.25% increase in f1 score, and no change in AUC ROC or AUC PR. Although the improvements in distance metrics and performance measures such as precision, recall, and f1 are promising, the increase in false positives warrants further investigation. Regardless, as shown in Figure 7, the performance meets the success metric of accurately returning localized point-based outlines of each of the seven substructures utilizing only a small VLM.

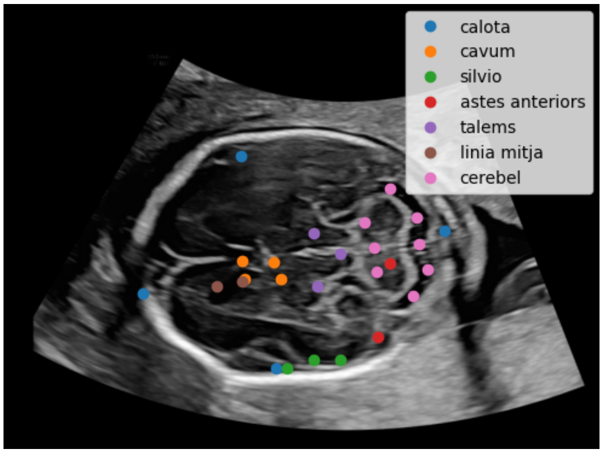

Figure 7: Per-substructure model predictions for a test set example.

A detailed bill of materials can be found in the Appendix. To summarize, everything from ETL to model quantization/evaluation/training costs under $2, with training the model and testing the API incurring most of the cost.

Conclusions and Recommendations

Overall, this project successfully developed an AI-driven approach to ultrasound substructure localization using Qwen2.5-VL-3B-Instruct as a basis for our LLM. This serves as a proof of concept for future exploration of such a system for aiding healthcare professionals, particularly midwives, and clinicians, with assisting in identifying fetal abnormalities early, thus improving maternal and infant health outcomes. Through data analysis and multiple rounds of fine-tuning, we observed significant reductions in error distances and false predictions while remaining within our original boundaries of correctness. While the system met key success metrics in terms of accurate labeling, challenges such as overfitting and false positives show that further improvements are needed for broader deployment. This is especially true when taking into account the memory size of portable ultrasound machines, which are the intended devices.

To enhance the model's reliability and usability, future work should focus on expanding the data used by testing with more diverse ultrasound scans. Addressing computational constraints is also key, with optimization needed to ensure the model continues to operate effectively on resource-limited hardware. Additionally, testing in real-world clinical environments would provide valuable insights into how to optimize for deployment in remote areas. Finally, long-term collaboration with healthcare organizations and continued innovation in AI-driven medical diagnostics such as the ones demonstrated by this project will be crucial in making maternal healthcare more accessible and effective worldwide.

Works Cited

[1] World Health Organization, “Maternal mortality,” *World Health Organization*, Apr. 26, 2024. https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[2] “Maternal Health Fund - Treatment and Prevention of Childbirth Injuries,” *Maternal Health Fund*, May 13, 2024. https://maternalhealthfoundation.org/

[3] B. S. Prabakaran, P. Hamelmann, E. Ostrowski, and M. Shafique, “FPUS23: An Ultrasound Fetus Phantom Dataset with Deep Neural Network Evaluations for Fetus Orientations, Fetal Planes, and Anatomical Features,” *IEEE Access*, vol. 11, pp. 58308–58317, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3284315.

[4] A. Jakab, “FeTA Dataset | Fetal and neonatal developmental imaging research,” *Neuroimaging.ch*, 2021. http://neuroimaging.ch/feta (accessed Feb. 07, 2025).

[5] https://qwenlm.github.io/blog/qwen2.5-vl/

[6] https://modal.com/pricing

[7] H. N. Xie et al., “Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal,” Ultrasound Obstet Gynecol., vol. 56, no. 4, pp. 579–587, Oct. 2020, doi: 10.1002/uog.21967.

[8] J. Weichert and J. L. Scharf, “Advancements in Artificial Intelligence for Fetal Neurosonography: A Comprehensive Review,” J. Clin. Med., vol. 13, no. 18, p. 5626, Sep. 2024, doi: 10.3390/jcm13185626.

Appendix

Bill of Materials

Since this project heavily utilized Modal, and because it was not done in an isolated account, we lack a clear cost breakdown. Therefore, we provide an estimated cost to reproduce this project using 1) the estimated time to completion for each task and 2) Modal's contemporary price per hour for all types of hardware used (i.e. CPU and GPU). [6] For reference, all tasks were done on a starter account with $30/month free credit, limited to 10 and 100 concurrent GPUs and containers, respectively.

- *ETL*: 2 minutes * 0.125 cores * ~25 containers * $0.047/core/hour = $.0049
    - *Note*: although 100 containers can be spun up at once, this doesn't occur here because each container completes its task quickly enough such that it can start another task before another container is assigned.
- *Quantization*: 15 minutes * (0.125 cores * 1 container * $0.047/core/hour + 1 L4 * $0.80/hour) = $0.2015
- *Evaluation*: 10 minutes * (0.125 cores * 1 container * $0.047/core/hour + 1 L4 * $0.80/hour) = $0.1343
- *Training*: 10 minutes * (0.125 cores * 1 container * $0.047/core/hour + 2 H100s * $3.95/hour) = $1.3176
- *API (per request)*: ~(1 + 5 minutes) * (0.125 cores * 1 container * $0.047/core/hour + 1 L4 * $0.80/hour) = $0.0806
    - *Note*: estimating the time spent on 1 request to be 1 minute with a 5-minute idle period before shutdown.
- *Website (per request)*: ~(5 + 5 minutes) * 0.125 cores * 1 containers * $0.047/core/hour = $.0010
    - *Note*: estimating the time 1 user spends on the website to be 5 minutes with a 5-minute idle period before shutdown.
- *Total*: $1.6583 + (~3 API calls * $0.0806) + (~3 website visits * $.0010) = **$1.90**

Team and project management

The main issue was finding a meeting time that worked for each of our team members as well as our community partner. This was solved by creating a when2meet between all group members. The times each of us was available was then shared with our partner and when meetings outside of those times had to be scheduled it was clear who could and could not make it.

For team management, each team member took charge of a specific phase of that project. During that phase, they were responsible for communicating with our community partner and the team about expectations and goals, setting up meetings, and ensuring everything remained on the correct timeline. We did not encounter any issues with the system as it gave each person a chance to have a leadership role, playing to our strengths while still allowing us to work as a team to accomplish our work as a whole.