Table of Contents

[2.2 1](#_Toc208775436)

[2.3 Zero-shot gestational age estimation 2](#_Toc208775437)

[2.4**:** 3](#_Toc208775438)

[📖 2.4.1 Probing FetalCLIP for fetal views classification 3](#_Toc208775439)

[question 3](#_Toc208775440)

[2.4.2 Probing FetalCLIP for video-based CHD detection 4](#_Toc208775441)

[2.4.3 Probing FetalCLIP for segmenting fetal structures 4](#_Toc208775442)

[2.5 FetalCLIP interpretability 6](#_Toc208775443)

[4.2 FetalCLIP architecture and pretraining 6](#_Toc208775444)

[4.3 Zero-shot view classification 8](#_Toc208775445)

# 2.2

**2. Dataset used**

* **Planes DB** dataset (fetal ultrasound images from 2 hospitals).
* The task: classify **5 main anatomical planes**:
  + Abdomen
  + Brain
  + Cervix
  + Femur
  + Thorax
* Plus **3 brain subplanes**:
  + Transcerebellum
  + Transthalamic
  + Transventricular
* So, the model had to say: *“This is abdomen / this is transthalamic view,” etc.*

**3. Results**

* FetalCLIP got **87.1% F1 score** (a metric that balances accuracy & precision).
* Comparison:
  + **+17.2% better than SonoNet**
  + **+37.6% better than UniMed-CLIP**
  + **+40.5% better than BiomedCLIP**
  + **+60.1% better than CLIP**

# 2.3 Zero-shot gestational age estimation

**2. What they tested**

* Can FetalCLIP **predict gestational age directly from ultrasound images** (without extra training = zero-shot)?
* Dataset used: **HC18** (fetal brain ultrasound images).
  + Includes **head circumference (HC)** + pixel spacing.
  + True GA was not given → so they used HC as a proxy.

**3. How they judged correctness**

* A prediction was **valid** if the *true head circumference (HC)* fell within the **normal WHO growth range** (2.5th–97.5th percentile) for the predicted GA.
* In other words → if the predicted GA matched realistic fetal growth curves.

**4. Results**

* **FetalCLIP validity rate: 83.5%**
* Other models much worse:
  + CLIP: **11%**
  + BiomedCLIP: **24%**
  + UniMed-CLIP: **9%**

✅ Shows FetalCLIP is far better at estimating GA.

**5. Extra finding**

* When looking at fetuses between **20 weeks 0 days – 21 weeks 6 days (middle range of training data)**, FetalCLIP did even better → **89% validity rate**.
* This means the **distribution of training data** (which had more samples around 20–22 weeks) influenced how well the model performed.
* 👉 **In short:**  
  FetalCLIP can estimate gestational age from ultrasound images without retraining, and it’s much more accurate than other vision-language models, especially in the gestational weeks where it had more training data.

# 2.4**:**

FetalCLIP’s image encoder produces strong features, making it useful for different fetal ultrasound tasks (classification, CHD detection, segmentation).

# 📖 2.4.1 Probing FetalCLIP for fetal views classification

 Dataset: **Planes DB** (same as before).

 Task: classify **6 fetal views + 3 brain subplanes**.

 Method:

* They **froze FetalCLIP’s image encoder** (didn’t retrain it).

4. **Data-efficient training test**

 Example: only **32 patients’ data**.

 Even with so little data, **FetalCLIP performed as good as or better than UniMed-CLIP trained on the full dataset (717 patients)**.

 This proves FetalCLIP is **data-efficient** → you don’t need huge amounts of training data to get good results if you use its pretrained features.

Two common strategies:

1. **Fine-tuning** → update *all or most* of the pretrained model’s weights.
2. **Feature extraction (linear probing)** → **freeze** the pretrained model, only train a small new layer on top.

## question

so other than 6 fetal views + 3 brain subplanes.i can basically expand to more

if yes then how can in know with which i can expand more

**🔹 3. Practical workflow for expansion**

1. Start with **zero-shot testing** → see if FetalCLIP already distinguishes new views with just prompts.
2. If performance is **decent**, you can:
   * Freeze the encoder.
   * Collect a small dataset for the new views.
   * Train a linear classifier to add them.
3. If performance is **poor**, you might need:
   * More labelled data.
   * Or partial fine-tuning of the encoder.

# 2.4.2 Probing FetalCLIP for video-based CHD detection

* Still, its **frozen image encoder** gave such strong features that even with limited data, the linear layer could classify CHD vs normal reliably.
* This shows **transfer learning power**: pretrained representations work beyond the exact task they were trained for.

✅ In short:  
They froze FetalCLIP’s encoder, extracted per-frame features, combined them for each video, and trained a small classifier for **CHD detection**. The model **outperformed all baselines** by a wide margin.

# 2.4.3 Probing FetalCLIP for segmenting fetal structures

**🔹 Problem Setup**

* **Task:** Segment (outline) fetal anatomical structures in ultrasound images → pixel-level labeling.
* **Why:** Needed for **biometry** (measuring fetus growth accurately).
* Example: Brain → outline head, Abdomen → outline stomach/spine, Heart → outline different chambers.

**🔹 What They Did**

1. **Intermediate features from FetalCLIP**
   * Instead of just using the final image embedding, they took **intermediate feature maps** from the FetalCLIP encoder (ViT-B or ViT-L).
   * These contain **fine-grained spatial information** needed for pixel-level segmentation.
2. **Lightweight decoder**
   * Added a **small decoder network** (~1.3M to 1.6M parameters).
   * Decoder takes FetalCLIP features → upsamples → outputs a segmentation mask (pixel-by-pixel class labels).
3. **Datasets & Views**  
   They tested segmentation on three important ultrasound views:
   * **Brain view (head)** → Segment head.
   * **Abdomen view** → Segment abdomen, stomach, and spine.
   * **4-chamber heart view** → Segment 9 structures (different chambers and vessels).

**🔹 Results**

Metric: **Dice Similarity Coefficient (DSC)** → measures overlap between predicted segmentation and ground truth (higher = better).

* **Brain view:** 97.92%
* **Abdomen view:** 81.82%
* **4-chamber heart view:** 72.91%

📊 FetalCLIP outperformed UniMed-CLIP by:

* +0.08% (Brain)
* +1.7% (Abdomen)
* +3.64% (Heart)

⚡ In **data-efficient settings** (few training samples), FetalCLIP still stayed ahead → showing it generalizes better with less data.

**🔹 Why This Matters**

* Segmentation is **harder than classification** → needs detailed pixel-level understanding.
* General-purpose models (CLIP, UniMed-CLIP, etc.) are not specialized enough.
* But FetalCLIP, trained with **fetal-specific text + images**, learns features that are **robust and fine-grained**, making it a **strong feature extractor for segmentation tasks**.

✅ **In short:**  
They froze FetalCLIP’s encoder, added a lightweight decoder, and tested segmentation on brain, abdomen, and heart views. FetalCLIP achieved **high DSC scores**, beating other models, and worked well even with limited data.

# 2.5 FetalCLIP interpretability

CAM → what part of image is used, UMAP → how views cluster)

# 4.2 FetalCLIP architecture and pretraining

**🔹 What They Did in FetalCLIP Architecture & Pretraining**

**1. Base Model (CLIP-inspired)**

* **Image encoder** → ViT-L (Vision Transformer, Large version)
  + Input size: **224 × 224**
  + Patch size: **14 × 14**
  + Layers: **24 transformer layers**
* **Text encoder** → Transformer (like GPT-style but smaller)
  + Layers: **12**
  + Max tokens: **117** (more than CLIP’s 77, since medical captions are longer).

👉 Both encoders project into a **shared 768-dimensional embedding space**.  
This way, an ultrasound image and its text caption can be compared directly.

**2. Training Objective**

* Goal: **Match image embeddings with their correct caption embeddings**.
* If image–caption pair is correct → **maximize similarity**.
* If it’s a mismatch → **minimize similarity**.

✅ This is **contrastive learning** → same as CLIP, but specialized for fetal ultrasound.

**3. Data Augmentation**

To make the model robust against ultrasound variability, they applied:

* Random **rotation** (−7° to +7°)
* Random **translation** (small shifts, up to 5%)
* Random **color jittering** (brightness, contrast, saturation ∈ [0.85, 1.15])

This simulates real-world variations between hospitals/machines/operators.

**4. Training Setup**

* **20 epochs** total.
* **Learning rate:** 5e-6.
* **Warmup steps:** 2,000 (gradually increase LR at start).
* **Scheduler:** cosine decay (smooth LR reduction).
* **Weight decay:** 0.1 (to prevent overfitting).
* **Hardware:** 4× RTX A6000 GPUs.
* **Batch size:** 140 per GPU (~560 total).
* **Precision:** mixed precision (for speed + memory efficiency).
* **Checkpointing:** best model = highest F1 score in **zero-shot view classification**.

**5. Initialization Trick............**

* Starting from **general CLIP** → not optimal (natural images).
* Starting from **medical-domain CLIP (UniMed-CLIP / BiomedCLIP)** → better.
* They found fine-tuning CLIP from the **medical domain** gave:
  + Zero-shot F1 improved **from 85.2% → 87.1%**.

👉 Lesson: **The closer your pretraining domain is to the target domain, the better the initialization.**

**🔹 Big Picture**

FetalCLIP = CLIP adapted to **fetal ultrasound** by:

1. Using ViT-L + longer-text encoder.
2. Pretraining on paired ultrasound image–caption datasets.
3. Applying smart augmentations for robustness.
4. Initializing from **medical CLIP weights**, not natural CLIP.

# 4.3 Zero-shot view classification

**. How Classification Works**

* Each prompt → converted to a **768-dim text embedding**.
* For each class, they **average the embeddings across its 5 prompts** → class representation.
* Input ultrasound image → encoded into a **768-dim image embedding**.
* Classification = find the class whose text embedding has the **highest cosine similarity** with the image embedding.

👉 This is **text-guided classification** without explicit retraining.