REFINING PLANE PRIORS FOR QUERY-CONDITIONED FETAL ULTRASOUND SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-guided fetal ultrasound segmentation requires locating the structure specified by a clinical query among multiple structures in the same image. Existing multimodal large language model (MLLM) approaches rely primarily on internal representations that lack fetal-ultrasound domain knowledge. A pretrained fetal-ultrasound vision-language model can provide domain-specific visual features and plane-level activation maps. However, the visual features do not directly provide an explicit spatial prior, while the activation maps remain coarse and unstable for query-specific localization. We introduce PRISM (Plane-prior Refinement and Integration for Segmentation with MLLMs), a framework that converts plane-level evidence into query-conditioned spatial prompts. PRISM uses the pretrained model through two complementary pathways: its domain-specific visual features are supplied to the MLLM, while its plane activation maps provide spatial evidence for refinement. Activation Prior Refinement (APR) learns a residual correction of the activation map from visual context and query semantics, producing a target-specific spatial prior. Patch-Adaptive Prior Composition (PAPC) then constructs experts from the refined prior, visual features, and query representation, selects an expert subset with a Set Router, and fuses their predictions spatially with a Patch Router. Extensive experiments on public and private fetal ultrasound benchmarks demonstrate that PRISM consistently outperforms the compared methods across model scales, reaching 80.55% and 71.41% cIoU on the public and private benchmarks, respectively, with a 13B language model.

1 INTRODUCTION

Fetal ultrasound is a non-invasive, convenient, and widely used imaging modality in prenatal care Salomon et al. (2022); Whitworth et al. (2015). Clinical examinations typically rely on multiple standard planes to visualize different fetal organs Krishna & Kokil (2024); He et al. (2026), as illustrated in Fig. 1(a). Accurate segmentation of anatomical structures supports biometric measurements, structural assessment, and the localization of abnormal regions Lee et al. (2023); Slimani et al. (2023). A single ultrasound frame can also support different segmentation requests. For example, given the same abdominal circumference plane, a clinician may request the fetal abdomen, an internal structure such as the stomach, or another structure such as the spine, as shown in Fig. 1(b). Although the image remains unchanged, the target structure and its spatial extent vary with the textual query. The desired mask is therefore determined jointly by the image and the clinical query.

MLLM-based segmentation provides a natural framework for this query-dependent setting by connecting language reasoning with pixel-level prediction. LISA introduces a special segmentation token, `[SEG]`, whose hidden embedding is decoded into a segmentation mask, establishing the embedding-as-mask paradigm (Lai et al., 2024). Subsequent methods extend this paradigm in different directions. SESAME improves query understanding through false-premise handling and query correction (Wu et al., 2024). READ examines the correspondence between the `[SEG]` embedding and image-token representations, and converts their similarity responses into point prompts for SAM (Qian et al., 2025). Sim4Seg further derives region-aware similarity masks from internal vision-language representations to provide dense prompts for medical segmentation (Song et al., 2026). These methods establish an effective route from text-conditioned reasoning to mask predic-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

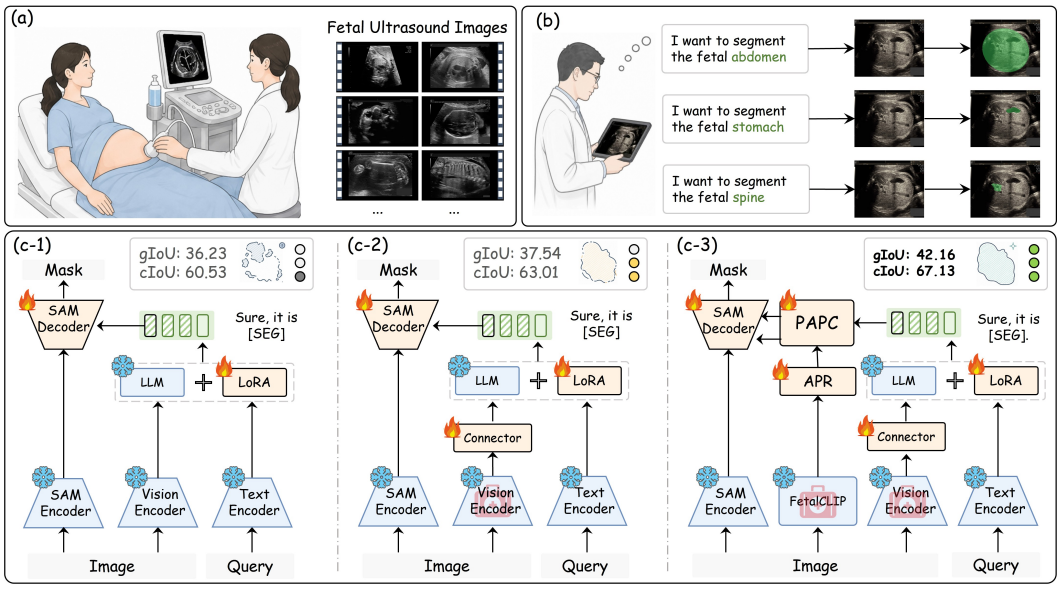

Figure 1: **Motivation for text-guided fetal ultrasound segmentation.** (a) Fetal ultrasound examinations capture multiple anatomical views. (b) Different text queries specify distinct targets within the same frame. (c-1) The standard text-guided segmentation paradigm uses MLLM representations to predict masks with SAM. (c-2) A medical-domain encoder replaces the MLLM visual tower to introduce domain knowledge. (c-3) PRISM refines fetal-plane priors and integrates them with MLLM representations. The displayed scores report performance on the private dataset.

tion, but their spatial evidence is primarily derived from MLLM-internal representations and does not explicitly encode fetal-plane knowledge.

Domain knowledge can be introduced by replacing the general-purpose visual tower with a medical-domain pretrained encoder (Huang et al., 2025; Bai et al., 2026). This changes the visual features supplied to the MLLM, but does not directly expose domain knowledge as an explicit spatial prior for target localization. A complementary route uses patch-text similarity to spatialize domain knowledge for localization and mask prediction (Rao et al., 2022; Zhou et al., 2022; Cho et al., 2024; Song et al., 2026). FetalCLIP learns image-plane alignment for fetal ultrasound (Maani et al., 2026). We extract plane-conditioned activation maps from its frozen visual encoder using ScoreCAM. These maps provide useful plane-level evidence, but remain coarse and unstable for query-specific segmentation, which poses a challenge: ❶ ***refining external plane knowledge into spatial evidence aligned with the queried target***. Moreover, the existing `[SEG]`-based interface uses the `[SEG]` representation as a semantic prompt and MLLM visual representations as image-conditioned features for mask prediction (Lai et al., 2024; Qian et al., 2025; Song et al., 2026). Introducing an external plane prior adds plane-level spatial evidence to this query-conditioned interface, which poses a second challenge: ❷ ***adaptively integrating external plane evidence with query-conditioned MLLM representations***.

To address these challenges, we develop PRISM(**P**lane-prior **R**efinement and **I**ntegration for **S**egmentation with **M**LLMs), a plane-prior refinement framework for query-conditioned fetal ultrasound segmentation. PRISM uses FetalCLIP through two complementary pathways: its domain visual tower introduces fetal-ultrasound representations into the MLLM, while a frozen external branch predicts the imaging plane and provides a ScoreCAM-based plane activation map. The Activation Prior Refinement module (APR) combines this plane evidence with MLLM visual and query representations to produce a query-conditioned spatial prior. The Patch-Adaptive Prior Composition module (PAPC) then adaptively composes the refined prior, visual tokens, and query representation into a dense prompt for SAM. During training, a prior surrogate evaluates candidate compositions, and a KL loss aligns its loss-derived distribution with the Set Router distribution, updating both branches. To summarize, our contributions are as follows:

❶ We present PRISM as a framework that injects pretrained fetal-ultrasound knowledge into multimodal large language models through domain-specific visual representations and plane-level spatial evidence for query-conditioned mask prediction.

❷ We introduce APR to refine coarse plane activation responses into target-specific spatial priors using visual and query context, and PAPC to select and spatially fuse expert predictions from the refined prior, visual representations, and query semantics.

❸ Experiments show that explicit fetal-plane evidence complements MLLM representations across public and private benchmarks and model scales. PRISM consistently improves over the compared methods, achieving 80.55% and 71.41% cIoU on the public and private benchmarks, respectively, with a 13B language model.

2 Related Work

2.1 Medical Image Segmentation

U-Net combines contextual features with spatial detail through encoder-decoder skip connections (Ronneberger et al., 2015), while nnU-Net automates pipeline configuration for different datasets (Isensee et al., 2021). TransUNet further integrates convolutional features with Transformer representations to capture global context (Chen et al., 2024b). Beyond these advances in image representation and task adaptation, promptable models allow users to specify segmentation targets at inference time. SAM separates image encoding from prompt-conditioned mask decoding (Kirillov et al., 2023), and MedSAM adapts this framework to medical images through large-scale training with bounding-box prompts (Ma et al., 2024). Language-guided approaches extend target specification to natural-language queries. LISA connects multimodal language representations to a segmentation decoder (Lai et al., 2024), while Sim4Seg uses region-aware vision-language similarity masks to support medical segmentation (Song et al., 2026). Within this setting, our work incorporates fetal-plane activation evidence as an anatomical prior and refines it according to the query to guide segmentation of the requested structure.

2.2 Fetal Ultrasound Analysis

Fetal ultrasound interpretation depends strongly on imaging plane, because the same anatomical structure can appear with different visibility, geometry, and boundary quality across views He et al. (2026); Pu et al. (2026). FetalCLIP and SonoCLIP learn image-text alignment for fetal ultrasound images and provide representations of plane-level and anatomical semantics (Maani et al., 2026; Su et al., 2026). Text-derived priors have also supported fetal-ultrasound segmentation within video summarization and measurement pipelines (Guo et al., 2024). FetalAgents further incorporates structured plane knowledge into a multi-agent analysis system (Hu et al., 2026). Our work focuses on converting plane-level activation evidence into a target-specific dense prompt through query-conditioned residual refinement and adaptive composition with MLLM representations.

2.3 Language-Guided Semantic Segmentation

Language-guided segmentation models connect a multimodal language model to a pixel decoder so that users can specify regions with natural-language instructions. LISA introduces a segmentation token and a SAM decoder, while PixelLM and GLaMM extend language-to-mask interaction with pixel-level grounding and region-aware representations (Lai et al., 2024; Kirillov et al., 2023; Ren et al., 2024; Rasheed et al., 2024). SEEM unifies visual and text prompts for open-vocabulary segmentation (Zou et al., 2023), and READ improves the spatial grounding of the segmentation token by converting image-token similarity responses into point prompts (Qian et al., 2025). SESAME addresses query validity and false premises jointly with language feedback and segmentation (Wu et al., 2024); subsequent methods explore stronger multimodal backbones, reasoning supervision, and verifier or reinforcement signals (Zhang et al., 2024b;a; Chen et al., 2024c; Liu et al., 2025a;b; Huang et al., 2026; You & Wu, 2025). Our work follows the language-conditioned mask-decoding paradigm but separates two sources of information: the query-driven `[SEG]` representation specifies the target, while the fetal-plane CAM supplies an auxiliary spatial signal that is refined and adaptively composed with MLLM visual and semantic representations.

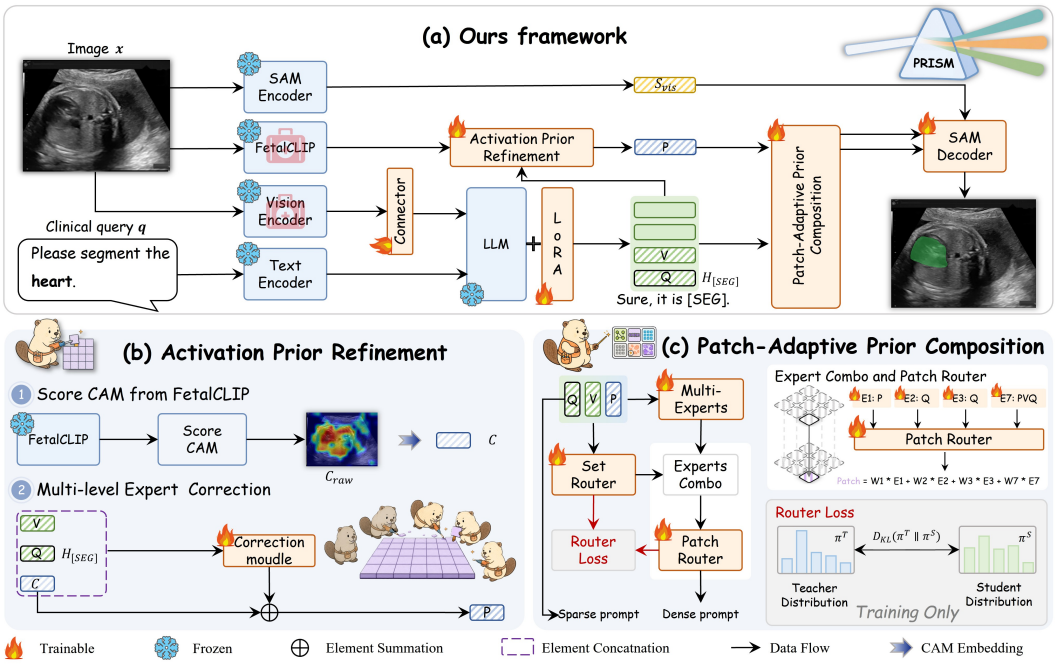

Figure 2: **PRISM framework.** (a) PRISM maps an ultrasound image and a clinical query to the requested segmentation mask. The sparse prompt is projected directly from Q, while PAPC supplies the dense prompt. (b) APR refines the FetalCLIP plane activation map using MLLM visual context and query semantics. The refined response provides a spatial prior tailored to the requested anatomy. (c) PAPC uses the Set Router to select an expert subset and the Patch Router to fuse expert predictions across spatial locations.

3 Methodology

Figure 2(a) shows the three image-processing paths in PRISM: a frozen SAM image encoder produces S_{vis}, an external FetalCLIP branch extracts plane activation evidence, and a visual encoder supplies features to the MLLM. A trainable connector maps the visual features into the language embedding space, where a LoRA-adapted LLM processes them with the clinical query to produce contextualized visual tokens V and the `[SEG]` representation Q. APR uses V and Q to refine the external plane evidence into prior P, and PAPC composes P, V, and Q into a dense prompt. The SAM decoder combines S_{vis} with this dense prompt and a sparse prompt projected from Q to predict the requested mask.

3.1 Multimodal and Domain Representations

Given an ultrasound image x and a clinical query q specifying the imaging plane and target structure, PRISM predicts a binary mask for the requested structure. Following LISA (Lai et al., 2024), the MLLM produces contextualized visual tokens $V = [v_1, \ldots, v_N]^\top \in \mathbb{R}^{N \times d}$ and the hidden state $Q \equiv H_{\texttt{[SEG]}} \in \mathbb{R}^d$ of the generated `[SEG]` token. With a 224×224 visual input and 14×14 patches, $N = 256$ tokens form a 16×16 grid. The frozen SAM image encoder separately produces $S_{\text{vis}} = E_{\text{SAM}}(x)$. The external plane-prior construction and its query-conditioned refinement are described below.

3.2 Activation Prior Refinement

Plane activation maps encode regions associated with the fetal ultrasound image plane, but they do not identify which of several structures should be selected by the clinical query. Using the raw map directly as a dense prompt can therefore activate anatomically irrelevant regions. APR converts this plane-level evidence into a query-specific spatial prior by treating the CAM as an initialization

and learning a residual correction conditioned on visual tokens V and query representation Q. The residual formulation preserves the original plane evidence while adapting its spatial response to the requested structure.

We first identify the plane whose activation responses will initialize the prior. Using frozen FetalCLIP (Maani et al., 2026), we construct a text prototype t_p for each plane in the dataset-specific candidate set $\mathcal{P}$ by averaging its normalized description embeddings and renormalizing the result. The normalized image embedding v_x selects the plane $c = \arg\max_{p\in\mathcal{P}} v_x^\top t_p$, while softmax over the scaled image–text similarities gives plane probabilities $\pi_p(x)$. For the selected plane, ScoreCAM scores each activation channel using $s_k(x) = \pi_c(x \odot M_k)$. A softmax over channels gives weights $w_k(x) = \mathrm{softmax}_k(s(x))$ for constructing the spatial map:

$$C_{\mathrm{raw}} = \mathrm{Norm}_{[0,1]}\left[\mathrm{ReLU}\left(\sum_{k=1}^{D} w_k(x)A_k(x)\right)\right]. \tag{1}$$

Here $A_k(x)$ is one of the $D = 1024$ spatial activation channels from the last visual Transformer block, and M_k is its upsampled, min–max normalized mask. The softmax is taken over the channel index, and the selected plane c remains fixed across masked forward passes. To refine this map using the image-query representations, APR first aligns it with the visual-token grid. We clamp C_{raw} to $[10^{-4}, 1-10^{-4}]$ and bilinearly resize it, obtaining probabilities a_i at each token position. These values enter two paths: a bypass converts them to base logits $c_i = \log(a_i/(1-a_i))$, while a correction branch combines their embeddings with visual and query features.

For the correction branch, projections ϕ_C, ϕ_V, and ϕ_Q map the three inputs to 256-dimensional features. We use an MLP for the prior and LayerNorm followed by a linear layer for V and Q. Broadcasting the projected query across positions allows each local prior and visual feature to be conditioned on the requested structure. A shared position-wise correction function f_{APR} predicts residual logits, which are added to the base logits after spatial upsampling:

$$r_i = f_{\mathrm{APR}}([\phi_C(a_i)\|\phi_Q(Q)\|\phi_V(v_i)]), \qquad P = \mathcal{U}_{256}(C) + \mathcal{U}_{256}(R). \tag{2}$$

Here C and R are the reshaped base-logit and residual grids, $\|$ denotes concatenation, and $\mathcal{U}_{256}$ interpolates to 256×256. The residual formulation preserves a direct path from the initial prior, with zero correction recovering the interpolated base logits. Applying sigmoid to P yields the refined probabilities passed to PAPC. Further implementation details for prior construction and refinement are provided in Appendices A.1 and A.2.

3.3 Patch-Adaptive Prior Composition

To adaptively integrate external plane evidence with query-conditioned MLLM representations, PAPC constructs experts from the refined prior, visual tokens V, and query representation Q. A Set Router selects the expert subset for each image–query pair, while a shared Patch Router fuses the selected predictions at each spatial location. We first align the inputs by resizing $\sigma(P)$ to the token grid, projecting each source, and broadcasting the query feature across positions. Their concatenation is $h_i = [h_i^P\|h_i^Q\|h_i^V]$. The three sources define seven nonempty input combinations: P, V, Q, PV, PQ, QV, and PVQ. Each combination feeds a position-wise MLP expert that produces a spatial logit $z_{i,e}$, forming the expert pool $\mathcal{J}$.

PAPC selects a combination of k experts (Top-k) from the candidate family $\mathcal{A} = \{A \subseteq \mathcal{J} : |A| = k\}$, which contains $\binom{7}{k}$ subsets. To score these combinations, the Set Router uses pooled context $\bar{h} = N^{-1}\sum_i h_i$ to predict individual-expert scores u_e and pairwise scores b_{ef}. The pairwise terms make the subset score depend on which experts are included together:

$$s(A) = \sum_{e\in A} u_e + \sum_{\{e,f\}\subseteq A} b_{ef}, \qquad A^* = \arg\max_{A\in\mathcal{A}} s(A). \tag{3}$$

Once an expert combination is specified, the Patch Router determines how its predictions are fused at each position. This shared router maps local context h_i to logits $r_{i,e}$, then normalizes over the members of a candidate subset A:

$$w_{i,e}^A = \frac{\exp(r_{i,e}/\tau_R)}{\sum_{f\in A}\exp(r_{i,f}/\tau_R)} \quad (e\in A), \qquad z_i^A = \sum_{e\in A} w_{i,e}^A z_{i,e}. \tag{4}$$

Reshaping and upsampling the fused logits give $L^A = \mathcal{U}_{256}(\mathrm{reshape}(z^A))$. The selected candidate supplies the dense prompt $M_{\mathrm{PAPC}} = \sigma(L^{A^*})$ in both training and inference. Thus expert membership is shared across the image, while fusion weights are computed at each position.

To supervise subset selection, a prior-proxy teacher evaluates each candidate against the training mask Y, resized to 256×256. Its loss is $\ell_A = \mathcal{L}_{\mathrm{BCE}}(L^A, Y) + \mathcal{L}_{\mathrm{Dice}}(\sigma(L^A), Y)$, with BCE applied to logits. Softmax over the negative losses defines teacher probabilities π^{T}, assigning more weight to lower-loss candidates. Softmax over the Set Router scores $s(A)$ defines student probabilities π^{S}. At temperatures $\tau_T = 0.5$ and $\tau_R = 1.0$, respectively, we align these distributions through

$$\mathcal{L}_{\mathrm{router}} = D_{\mathrm{KL}}\left(\pi^{\mathrm{T}} \| \pi^{\mathrm{S}}\right). \tag{5}$$

Gradients from this KL loss propagate through both distributions, jointly updating the candidate-prior branch and the Set Router. The teacher evaluates priors directly, without decoding each candidate with SAM. Following the MLLM-to-SAM interface (Lai et al., 2024; Kirillov et al., 2023), SAM decodes the projected query Q and M_{PAPC} as sparse and dense prompts together with S_{vis} to predict the target mask. Further implementation details are provided in Appendix A.2.

3.4 Training Objectives

We jointly optimize language generation, final mask prediction, and expert-subset routing. Following LISA (Lai et al., 2024), $\mathcal{L}_{\mathrm{txt}}$ is the autoregressive cross-entropy loss for the response containing `[SEG]`. The final mask loss combines binary cross-entropy and Dice loss:

$$\mathcal{L}_{\mathrm{mask}} = \lambda_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}}(\hat{y}, y) + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}}(\hat{y}, y). \tag{6}$$

The router loss $\mathcal{L}_{\mathrm{router}}$ in Eq. 5 aligns the distribution derived from candidate prior losses with the Set Router distribution. The overall objective is

$$\mathcal{L} = \lambda_{\mathrm{txt}}\mathcal{L}_{\mathrm{txt}} + \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{router}}\mathcal{L}_{\mathrm{router}}. \tag{7}$$

We set $\lambda_{\mathrm{router}} = 0.1$. The final mask loss propagates through the selected dense prompt to train APR and PAPC. The MLLM visual tower, external FetalCLIP branch, SAM image encoder, and SAM prompt encoder remain frozen. We train the projection layers, APR, PAPC, SAM mask decoder, and language-model adaptation parameters. Additional module and adaptation settings are provided in Appendix A.2.

4 Experiments

4.1 Experimental Setting

4.1.1 Network Architecture

Following the MLLM-to-SAM architecture of LISA (Lai et al., 2024), we use LLaVA-1.5-7B and LLaVA-1.5-13B for the respective model scales, with ViT-H SAM (Kirillov et al., 2023) as the segmentation backbone. We evaluate CLIP and FetalCLIP (Maani et al., 2026) as alternative MLLM visual encoders, both using a ViT-L/14 architecture with an input resolution of 224×224.

4.1.2 Implementation Details

We initialize PRISM from the pretrained LISA checkpoint (Lai et al., 2024). All models are trained for four epochs using the AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 0.01. We use a batch size of 1 and accumulate gradients over 20 optimization steps. The text and mask objectives are weighted by $\lambda_{\mathrm{txt}} = \lambda_{\mathrm{mask}} = 1.0$. Within the mask objective, the binary cross-entropy and Dice losses are weighted by $\lambda_{\mathrm{bce}} = 2.0$ and $\lambda_{\mathrm{dice}} = 0.5$, respectively. All experiments are conducted on four NVIDIA A100 GPUs, each with 40 GB of memory.

4.1.3 Dataset

We evaluate on two fetal-ultrasound benchmarks, denoted as FetalSeg-Pub and FetalSeg-Priv. FetalSeg-Pub is assembled from FASSD, FOCUS, HC18, Large-Scale Fetal Head, and

Table 1: Query-conditioned segmentation results on the private and public benchmarks. Metrics are reported as percentages. Best, second-best, and third-best results are marked by bold, underline, and wavy underline, respectively.

Methods	Private				Public			
	7B		13B		7B		13B	
	gIoU	*cIoU*	*gIoU*	*cIoU*	*gIoU*	*cIoU*	*gIoU*	*cIoU*
LISA(CVPR2024)	36.23	60.53	44.65	66.06	38.41	70.01	49.39	76.58
SESAME(CVPR2024)	39.16	64.37	42.40	67.82	49.38	76.30	48.52	72.98
READ(CVPR2025)	21.62	36.44	28.53	47.10	37.49	50.95	44.03	63.50
Sim4Seg(AAAI2026)	35.74	64.13	38.86	67.21	37.74	66.80	48.49	76.67
Ours(CLIP)	41.89	66.56	45.24	69.73	50.43	77.09	51.72	78.78
Ours(FetalCLIP)	**42.16**	**67.13**	**47.58**	**71.41**	**51.01**	**78.82**	**52.97**	**80.55**

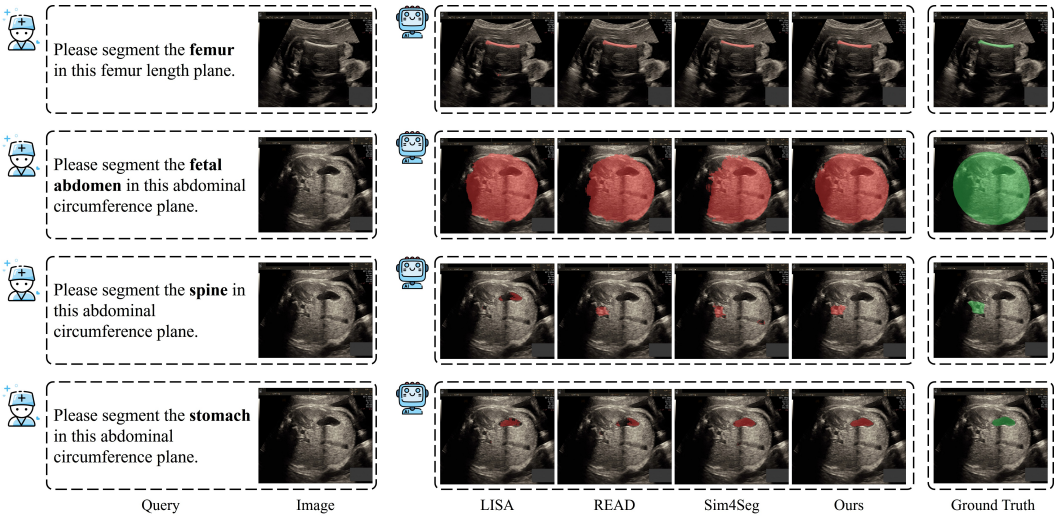

Figure 3: Visual comparison among LISA, READ, Sim4Seg, and our PRISM.

PSFHS (Da Correggio et al., 2023; Songxiong et al., 2025; van den Heuvel et al., 2018; Alzubaidi et al., 2023; Chen et al., 2024a), and contains 12 segmentation categories. Each non-empty image–target annotation is converted into one text query and one binary mask. FETALSEG-PRIV contains 8,753 source images, 44 segmentation categories, and 17 plane categories. Its split is performed at the source-image level and stratified by plane category, with approximately 7:1:2 assigned to training, validation, and testing. All annotations and queries derived from one source image inherit the same split. Dataset-level statistics are summarized in Appendix A.5.

4.1.4 EVALUATION METRICS.

We report two evaluation metrics: gIoU and cIoU. The gIoU metric is calculated as the mean of the Intersection-over-Union (IoU) values across all individual images, and cIoU is computed by the cumulative intersection across the cumulative union.

4.2 MAIN RESULTS

4.2.1 COMPARISON WITH THE STATE-OF-THE-ART.

Performance on the Private Benchmark. Table 1 reports query-conditioned segmentation performance using average query-mask overlap (gIoU) and pooled foreground overlap (cIoU). On the private benchmark, PRISM with FetalCLIP achieves the best results at both model scales. At 7B, it

obtains 42.16%/67.13% in gIoU/cIoU, exceeding the strongest prior result by 3.00 and 2.76 percentage points, respectively. At 13B, it reaches 47.58%/71.41%, improving over the best prior results by 2.93 and 3.59 points. PRISM with CLIP ranks second in all four private-benchmark metrics, indicating that the proposed segmentation framework remains effective across visual-tower configurations.

Performance on the Public Benchmark. The same trend holds on the public benchmark. The FetalCLIP variant achieves 51.01%/78.82% at 7B and 52.97%/80.55% at 13B. These results improve over the strongest prior methods by 1.63/2.52 and 3.58/3.88 percentage points in gIoU/cIoU, respectively. The consistent gains across both datasets and model scales show that PRISM improves query-conditioned mask prediction beyond the choice of language-model size alone.

4.3 Qualitative Results

Figure 3 shows qualitative results on four query–image cases. In the first row, the query involves a single target, and most methods recover the target region. In the second row, PRISM produces a mask for the fetal abdomen that more closely follows the target, whereas Sim4Seg yields an incomplete segmentation. In the third row, the query requests the spine, while LISA predicts the stomach; PRISM more closely follows the reference spine mask. In the fourth row, READ recovers only part of the stomach, whereas PRISM covers more of the target region.

4.4 Ablation Study

In this section, we conduct an ablation study to analyze the contribution of each component. We report the gIoU and cIoU performance on the public dataset using the 7B model. For all configurations, the MLLM visual tower is initialized from the pretrained general-CLIP model.

Table 2: Ablation study on the contributions of CAM, APR, and PAPC.

CAM	APR	PAPC	gIoU	cIoU
✗	✗	✗	38.41	70.01
✓	✗	✗	46.57	74.11
✓	✓	✗	48.08	75.64
✓	✗	✓	49.25	76.82
✓	✓	✓	**50.43**	**77.09**

4.4.1 Component Ablation

Table 2 evaluates the base model, CAM, APR, and PAPC under a common experimental setting. The CAM configuration tests the direct use of the external plane prior, while the APR and PAPC configurations isolate prior refinement and evidence composition, respectively. The results show a consistent improvement as the external plane evidence is introduced, refined with query and visual context, and adaptively composed with the MLLM representations. The complete configuration performs best, reaching 50.43% gIoU and 77.09% cIoU, compared with 38.41% and 70.01% for the base model. These results support the complementary roles of prior refinement and adaptive composition.

4.4.2 Analysis of APR

Table 3 examines APR inputs, correction formulation, and CAM source. Combining the CAM, visual tokens, and `[SEG]` representation gives the highest reported values, 48.08% gIoU and 75.64% cIoU. Visual conditioning alone changes the CAM-only result from 46.57%/74.11% to 46.50%/74.24%, while query conditioning reaches 47.78%/74.82%. The *Direct Predictor* uses the same CAM, visual, and query inputs as APR, but predicts the prior directly without the additive bypass from the original plane prior. APR exceeds this predictor by 0.24 and 0.87 percentage points in gIoU and cIoU, respectively, in the reported point estimates. With CLIP and SonoCLIP CAMs, the corresponding configurations also exceed the base

Table 3: Ablation of APR inputs, predictor formulation, and CAM sources.

CAM	Visual	`[SEG]`	gIoU	cIoU
✗	✗	✗	38.41	70.01
✓	✗	✗	46.57	74.11
✓	✓	✗	46.50	74.24
✓	✗	✓	47.78	74.82
✓	✓	✓	**48.08**	**75.64**
Direct Predictor			47.84	74.77
CAM from CLIP			39.67	71.12
CAM from SonoCLIP			45.76	72.43

model, reaching 39.67%/71.12% and 45.76%/72.43%. These comparisons establish compatibility with alternative CAM sources; a source-specific refinement gain requires a corresponding raw-CAM comparison.

4.4.3 SET-ROUTING SUPERVISION AND UTILITY MODELING

Table 4 examines teacher supervision and subset utility modeling. The no-distillation variant sets $\lambda_{\text{router}} = 0$ while retaining hard subset selection. With singleton and pairwise scoring enabled, soft prior-proxy supervision achieves 50.43% gIoU and 77.09% cIoU, exceeding the no-distillation variant by 1.47 and 0.95 percentage points and the hard teacher by 0.30 and 0.67 points, respectively. Unlike the hard teacher, which retains only the lowest-loss subset, the soft teacher assigns probabilities to all candidates according to their losses. Under soft supervision, adding pairwise utilities improves gIoU and cIoU by 0.78 and 0.22 points, showing that pairwise utilities improve subset selection.

Table 4: Set-router supervision and utility ablation. "Soft prior-proxy" constructs a teacher distribution from the losses of all candidate priors. Metrics are reported as percentages; higher is better.

Variant	Teacher	Singleton	Pairwise	gIoU	cIoU
No route distillation	✗	✓	✓	48.96	76.14
Hard prior teacher	*hard*	✓	✓	50.13	76.42
Soft prior-proxy teacher	*soft*	✓	✗	49.65	76.87
Soft prior-proxy teacher	*soft*	✓	✓	**50.43**	**77.09**

4.4.4 TOP-k SENSITIVITY

PAPC selects a subset of experts before patch-adaptive composition. We vary the subset cardinality while keeping the seven-expert pool and routing formulation fixed. For each k, the router considers $\binom{7}{k}$ candidate subsets. This experiment examines how the number of experts affects segmentation performance. As shown in Fig. 4, Top-4 achieves the highest gIoU and cIoU of 50.43% and 77.09%, respectively, improving both metrics over the single-expert setting. Increasing k beyond four reduces both metrics. Scores partially recover from $k = 6$ to $k = 7$, but remain below Top-4, supporting this intermediate choice over using all experts.

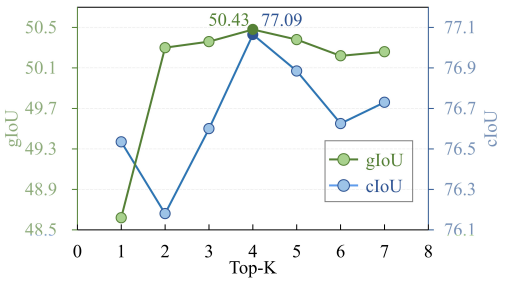

Figure 4: Top-k sensitivity of PAPC. The left and right axes report gIoU and cIoU, respectively.

5 CONCLUSION

We introduced PRISM to incorporate pretrained fetal-ultrasound knowledge into query-conditioned segmentation. Experiments on public and private benchmarks show that explicit plane-level evidence complements MLLM representations, yielding consistent improvements over the compared methods across model scales. With a 13B language model, PRISM achieves 80.55% and 71.41% cIoU on the public and private benchmarks, respectively. The current framework operates on individual 2D images, leaving temporal anatomical context unexploited and consistency across successive frames unexplored. Future work will investigate temporally informed anatomical priors to support consistent query-conditioned segmentation during continuous ultrasound scanning.

ETHICS STATEMENT

This study received ethics approval for the research use of the private fetal ultrasound data. Identifying approval details are withheld for anonymous review. The work concerns anatomical segmentation and does not make diagnostic claims.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the experimental setup, data processing steps, and key implementation details. Specifically, we build on the LISA framework, use LoRA for language-model adaptation, and implement the proposed APR and PAPC modules in PyTorch. The datasets used in this work are derived from real clinical applications.

AI USE STATEMENT

We did not use generative AI tools for any of the required-disclosure tasks listed in the ICLR 2027 AI policy, including developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, generating or processing datasets, or interpreting results.

We used generative AI tools for the following recommended-disclosure tasks: improving the grammar, wording, and readability of portions of the manuscript; generating or editing icons and schematic illustrations unrelated to experimental results; assisting with LaTeX formatting and layout; and normalizing the formatting, naming, and organization of existing source code. The AI tools were not used to create algorithmic components, modify the experimental procedure, generate or alter datasets, produce quantitative results, or draw scientific conclusions.

All AI-assisted text, figures, LaTeX code, and source-code edits were reviewed by the authors. The source-code changes were checked to ensure that they did not alter the intended functionality. We take full responsibility for the final content of this work, including all text, claims, figures, and code.

REFERENCES

Mahmood Alzubaidi, Marco Agus, Michel Makhlouf, Fatima Anver, Khalid Alyafei, and Mowafa Househ. Large-scale annotation dataset for fetal head biometry in ultrasound images. *Data in Brief*, 51:109708, 2023.

Yujie Bai, Juanjuan Zhao, Yapeng Zhao, Bing Gu, Yan Qiang, Qi Chen, and Huilan Zhao. Medmove-hrmllm: A mixture-of-vision-experts mllm for fine-grained high-resolution medical image understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7577–7581. IEEE, 2026.

Gaowen Chen, Jieyun Bai, Zhanhong Ou, Yaosheng Lu, and Huijin Wang. Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head. *Scientific data*, 11(1):436, 2024a.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical image analysis*, 97:103280, 2024b.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. pp. 323–340, 2024c.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. pp. 4113–4123, 2024.

Karine Souza Da Correggio, Roberto Noya Galluzzo, Luís Otávio Santos, Felipe Soares Muylaert Barroso, Thiago Zimmermann Loureiro Chaves, Alexandre Sherlley Casimiro Onofre, and Aldo von Wangenheim. Fetal abdominal structures segmentation dataset using ultrasonic images, 2023.

Xiaoqing Guo, Qianhui Men, and J Alison Noble. Mmsummary: multimodal summary generation for fetal ultrasound video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 678–688. Springer, 2024.

Xiao He, Huangxuan Zhao, Guojia Wan, Jiancheng Pan, Yanxing Liu, Xin Zou, Yong Luo, Yongchao Xu, Juhua Liu, Wei Zhou, et al. Epistemic-aware vision-language foundation model for fetal ultrasound interpretation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11061–11072, 2026.

Xiaotian Hu, Junwei Huang, Mingxuan Liu, Kasidit Anmahapong, Yifei Chen, Yitong Luo, Yiming Huang, Xuguang Bai, Zihan Li, Yi Liao, et al. Fetalagents: A multi-agent system for fetal ultrasound image and video analysis. *arXiv preprint arXiv:2603.09733*, 2026.

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. volume 38, pp. 138362–138383, 2026.

Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3779–3787, 2025.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Thunakala Bala Krishna and Priyanka Kokil. Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. *Expert Systems with Applications*, 238:122153, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Lok Hin Lee, Elizabeth Bradburn, Rachel Craik, Mohammad Yaqub, Shane A Norris, Leila Cheikh Ismail, Eric O Ohuma, Fernando C Barros, Ann Lambert, Maria Carvalho, et al. Machine learning for accurate estimation of fetal gestational age based on ultrasound images. *NPJ digital medicine*, 6(1):36, 2023.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025a.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv e-prints*, pp. arXiv–2505, 2025b.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Fadillah Maani, Numan Saeed, Tausifa Jan Saleem, Zaid Farooq, Hussain Alasmawi, Werner Diehl, Ameera Mohammad, Gareth Waring, Saudabi Valappil, Leanne Bricker, et al. Fetalclip: A visual-language foundation model for fetal ultrasound image analysis. *npj Digital Medicine*, 2026.

Bin Pu, Siyu Wang, Rongbin Li, Xinpeng Ding, Lei Zhao, Chaoqi Chen, Shengli Li, and Kenli Li. Organ-aware routing mixture-of-retrieval augmented generation for fetal ultrasound reporting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8448–8456, 2026.

Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how¡ seg¿ token works. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24722–24731. IEEE, 2025.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. pp. 18061–18070, 2022.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018. IEEE, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

LJ Salomon, Z Alfirevic, V Berghella, CM Bilardo, GE Chalouhi, F Da Silva Costa, E Hernandez-Andrade, G Malinger, H Munoz, D Paladini, et al. Isuog practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics and Gynecology*, 59(6):840–856, 2022.

Saad Slimani, Salaheddine Hounka, Abdelhak Mahmoudi, Taha Rehah, Dalal Laoudiyi, Hanane Saadi, Amal Bouziyane, Amine Lamrissi, Mohamed Jalal, Said Bouhya, et al. Fetal biometry and amniotic fluid volume assessment end-to-end automation using deep learning. *Nature Communications*, 14(1):7047, 2023.

Lingran Song, Yucheng Zhou, and Jianbing Shen. Sim4seg: Boosting multimodal multi-disease medical diagnosis segmentation with region-aware vision-language similarity masks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9079–9087, 2026.

Wu Songxiong, Zhang Hongyuan, Ye Tingting, X Haoyu, Z Ping, S Qingjun, W Panying, H Bingsheng, D Lei, and W Guangyao. Focus: four-chamber ultrasound image dataset for fetal cardiac biometric measurement, 2025.

Hang Su, Chao Sun, Zhaofan Li, Wei Hu, Juhua Liu, and Bo Du. Sonoclip: Mask-guided region-aware vision-language pretraining for fetal ultrasound analysis. *arXiv preprint arXiv:2606.29586*, 2026.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8): e0200412, 2018.

Melissa Whitworth, Leanne Bricker, and Clare Mullan. Ultrasound for fetal assessment in early pregnancy. *Cochrane database of systematic reviews*, (7), 2015.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. See, say, and segment: Teaching lmms to overcome false premises. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13459–13469. IEEE, 2024.

Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024a.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. pp. 74–91, 2024b.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. volume 36, pp. 19769–19782, 2023.

A APPENDIX

Appendix A.1 describes the plane-conditioned CAM. Appendix A.2 provides APR/PAPC implementation details and language-model adaptation settings. Appendices A.5 and A.6 give the dataset composition and query template.

A.1 FETALCLIP CAM CONSTRUCTION

We use the frozen FetalCLIP model to construct class-specific activation maps for the candidate plane categories of each dataset. The set of plane labels $\mathcal{P}$ is defined separately for each dataset. Each input image is zero-padded symmetrically to a square image and then processed using the OpenCLIP preprocessing pipeline at 224×224.

For each plane $p \in \mathcal{P}$, the prompt file provides $K = 5$ text descriptions. Let $q_{p,k}$ denote the k-th prompt for plane p. With $\mathrm{Norm}(z) = z/\|z\|_2$, the text prototype, image embedding, class logit, and class probability are computed as

$$\begin{aligned} t_p &= \mathrm{Norm}\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{Norm}(f_{\text{text}}(q_{p,k}))\right), \\ v_x &= \mathrm{Norm}(f_{\text{image}}(x)), \\ \ell_p(x) &= 100\, v_x^\top t_p, \qquad \pi_p(x) = \frac{\exp(\ell_p(x))}{\sum_{p'\in\mathcal{P}}\exp(\ell_{p'}(x))}. \end{aligned} \tag{8}$$

The factor 100 is a fixed logit scale, equivalent to a softmax temperature of 0.01. It controls the concentration of the plane probabilities used for ScoreCAM channel weighting while preserving the highest-scoring plane. The predicted plane is $c = \arg\max_{p\in\mathcal{P}} \pi_p(x)$.

For CAM generation, we use the gradient-free ScoreCAM implementation from `pytorch-grad-cam`. The target layer is the `ln_1` layer of the last visual Transformer block,

```
model.visual.transformer.resblocks[-1].ln_1.
```

Its output contains one class token followed by $16 \times 16 = 256$ patch tokens. The class token is removed, and the patch tokens are reshaped into a spatial activation tensor $A(x) \in \mathbb{R}^{1024\times16\times16}$. For a target plane c, ScoreCAM upsamples each channel A_d to the input resolution, normalizes it to form a mask, and obtains a target-class score from a masked forward pass. It then applies a softmax over the channel scores. The channel masks, scores, and raw CAM are computed as

$$\begin{aligned} &M_d = \mathrm{Norm}_{[0,1]}\left(\mathrm{Upsample}(A_d)\right), \qquad s_d = \pi_c(x \odot M_d), \qquad w_d = \frac{\exp(s_d)}{\sum_{d'=1}^{1024}\exp(s_{d'})}, \\ &C_c^{\text{raw}}(x) = \mathrm{Norm}_{[0,1]}\left[\mathrm{ReLU}\left(\sum_{d=1}^{1024} w_d A_d\right)\right]. \end{aligned} \tag{9}$$

The resulting map is returned on the input spatial grid and is overlaid on the min–max normalized input image using an image weight of 0.5.

For segmentation, the ScoreCAM target is the predicted plane c. For visualization, the ScoreCAM target is the annotated plane label for the corresponding image group. The model prediction and its probability are recorded separately, and prediction–annotation mismatches are reported for diagnostic purposes. No parameter updates or gradient-based optimization are used during CAM construction.

Prompt Template for Plane Prototype Construction

- "Ultrasound image of the p plane."
- "Fetal ultrasound image showing the p view."
- "Ultrasound imaging in the fetal p plane."
- "A fetal ultrasound examination in the p view."
- "The p plane in a fetal ultrasound image."

A.2 APR AND PAPC IMPLEMENTATION

Training uses bfloat16 precision. We apply LoRA to the language model's query and value attention projections with rank 8, scaling 16, and dropout 0.05. The language embedding and output head are also trainable, alongside the projection layers, APR/PAPC modules, and SAM mask decoder described in Section 3.4.

A.3 FEATURE AND SPATIAL ALIGNMENT

The MLLM visual tower supplies 1,024-dimensional patch features, which a linear projector maps to the language-model width d ($d = 4096$ for 7B). The contextualized MLLM outputs are $V \in \mathbb{R}^{256\times d}$ and $Q \in \mathbb{R}^d$ for each query. APR and PAPC each have their own feature projections: LayerNorm followed by a $d \to 256$ linear layer for V and Q, and a $1 \to 256 \to 256$ MLP with an intermediate GELU for the scalar prior. The query feature is broadcast to all 256 positions.

For SAM prompting, a $d \to d \to 256$ projector with an intermediate ReLU and no dropout maps Q to a sparse prompt. The SAM prompt encoder embeds this prompt and the 256×256 probability map M_{PAPC}; the mask decoder combines them with S_{vis}, and SAM postprocessing maps the prediction back to the original image size.

APR clamps the raw CAM to $[10^{-4}, 1 - 10^{-4}]$ and bilinearly resizes it to 16×16. Its correction MLP takes features in the order CAM, query, visual, with widths $768 \to 256 \to 256 \to 1$ and GELU after the first two linear layers. The correction function in Eq. 2 includes the learnable output scale: $f_{\mathrm{APR}}(z) = \alpha g_{\mathrm{APR}}(z)$, where g_{APR} is this MLP and α is initialized to 0.1. Thus R denotes the effective residual grid including this scale. Bilinear interpolation commutes with scalar multiplication, so the base and effective residual grids are independently interpolated to 256×256 before addition. A sigmoid converts the corrected logits to probabilities before PAPC, which bilinearly resizes them to the visual-token grid before embedding them. All bilinear interpolation in APR/PAPC uses `align_corners=False`.

Let $h_i^P, h_i^Q, h_i^V \in \mathbb{R}^{256}$ denote the PAPC projections and $h_i = [h_i^P \| h_i^Q \| h_i^V]$. The seven expert inputs, in implementation order, are $\mathsf{Q}, \mathsf{V}, \mathsf{PQ}, \mathsf{QV}, \mathsf{PV}, \mathsf{PVQ}, \mathsf{P}$; $\mathcal{J}$ indexes these experts. Each expert concatenates its designated inputs and applies an MLP with one 256-dimensional hidden layer and GELU to produce a scalar $z_{i,e}$. Because Q is broadcast and its expert is pointwise, the Q-only expert supplies a spatially constant, query-dependent logit. The P experts use a prior that has already been conditioned on V and Q by APR; expert labels therefore describe input paths rather than independent modalities.

A.4 SHARED PATCH ROUTER AND SET ROUTER

The Patch Router is a single $768 \to 256 \to 7$ MLP with GELU, producing logits $r_{i,e}$. For a candidate subset A, the weights and candidate spatial logits are

$$w_{i,e}^A = \frac{\mathbf{1}[e \in A] \exp(r_{i,e}/\tau_R)}{\sum_{f\in A} \exp(r_{i,f}/\tau_R)}, \qquad z_i^A = \sum_{e\in A} w_{i,e}^A z_{i,e}. \tag{10}$$

Excluded experts are masked before softmax. All 35 candidates share the expert outputs and Patch Router parameters. We reshape each candidate to 16×16 and bilinearly interpolate it to 256×256, giving $L^A = \mathcal{U}_{256}(\mathrm{reshape}(z^A))$.

The Set Router applies a $768 \to 256 \to 28$ MLP with GELU to $\bar{h} = N^{-1} \sum_i h_i$. Its outputs are seven singleton utilities u_e and 21 pairwise utilities b_{ef}. For $\mathcal{A} = \{A \subseteq \mathcal{J} : |A| = 4\}$, it computes

$$s(A) = \sum_{e \in A} u_e + \sum_{\{e,f\} \subseteq A} b_{ef}, \qquad \pi_A^{\mathrm{S}} = \frac{\exp(s(A)/\tau_R)}{\sum_{B \in \mathcal{A}} \exp(s(B)/\tau_R)}. \tag{11}$$

Under this parameterization, the predicted marginal utility of adding e to a set S is $u_e + \sum_{f \in S} b_{ef}$. These utilities are learned from route supervision, rather than measured from separate singleton or pairwise segmentation runs. Both training and inference use $A^* = \arg\max_{A \in \mathcal{A}} s(A)$ and the dense prompt $M_{\mathrm{PAPC}} = \sigma(L^{A^*})$. The student probabilities supervise routing through the KL loss; they do not form a soft mixture in the forward pass.

A.5 Dataset Details

FetalSeg-Pub combines five released 2D fetal-ultrasound segmentation datasets. The public benchmark contains 12 segmentation categories in total. Each non-empty image–target annotation is converted into one query–mask pair. The split is fixed and group-stratified, with all target annotations from the same source image assigned to one split.

Table 5: Source composition of FetalSeg-Pub. Counts are non-empty query–mask pairs; one source image may contribute more than one pair when multiple targets are annotated.

Source	Segmentation targets	Train	Val	Test	Total
FASSD	Abdominal artery, liver, stomach, vein	4,497	615	1,147	6,259
FOCUS	Fetal cardiac region, fetal thoracic region	420	60	120	600
HC18	Derived fetal head region	699	100	200	999
Large-Scale Fetal Head	Fetal brain, CSP, lateral ventricle	3,345	547	1,150	5,042
PSFHS	Fetal head, pubic symphysis	1,900	272	544	2,716
Total	**12 categories**	**10,861**	**1,594**	**3,161**	**15,616**

The source abbreviations in Table 5 denote the Fetal Abdominal Structures Segmentation Dataset (FASSD) (Da Correggio et al., 2023), the four-chamber ultrasound image dataset (FOCUS) (Songxiong et al., 2025), the fetal head circumference dataset (HC18) (van den Heuvel et al., 2018), the Large-Scale Fetal Head dataset (Alzubaidi et al., 2023), and the pubic symphysis and fetal head segmentation dataset (PSFHS) (Chen et al., 2024a). HC18 head regions are derived by filling the closed head-circumference annotations; they are distinguished from the source datasets' region-mask annotations. Here, CSP denotes the cavum septi pellucidi.

FetalSeg-Priv contains 8,753 source images across 17 plane categories and 44 segmentation categories. The source-image split is stratified by plane category and follows an approximately 7:1:2 train–validation–test allocation. The category-level image counts are listed below.

In Table 6, the plane labels are defined as follows: ac, abdominal-circumference plane; medium, trans-thalamic plane; fl, femur plane; hl, humerus plane; cereb, trans-cerebellar plane; vavp, transventricular plane; kidney, bilateral renal plane; eye, bilateral orbital transverse plane; sideface, midsagittal facial plane; lvot, left ventricular outflow tract plane; bladder, abdominal-wall transverse plane through the bladder; 3vv, three-vessel-trachea plane; chestc, thoracic coronal plane; rvot, right ventricular outflow tract plane; forearm, radius-ulna plane; leg, tibia-fibula long-axis plane; and 4ch, four-chamber plane.

A.6 Prompt Template

Queries specify both the imaging plane and the target structure. In the training data, plane names are taken from the dataset's plane annotations and written into the `text` field, which contains wording variants. For example, a record for the aortic isthmus in the three-vessel and trachea view contains:

- "In the fetal ultrasound three-vessel and trachea view, segment the aortic isthmus."
- "Please segment the aortic isthmus in this three-vessel and trachea view."

Table 6: Category-level composition of FETALSEG-PRIV. Counts are source images.

Plane/category	Train	Val	Test	Total
3vv	315	45	90	450
4ch	310	44	90	444
ac	403	57	116	576
bladder	392	56	112	560
cereb	370	52	107	529
chestc	592	84	170	846
eye	387	55	112	554
fl	363	51	105	519
forearm	355	50	103	508
hl	369	52	107	528
kidney	392	56	112	560
leg	350	50	101	501
lvot	381	54	110	545
medium	349	49	101	499
rvot	112	16	32	160
sideface	383	54	111	548
vavp	298	42	86	426
Total	**6,121**	**867**	**1,765**	**8,753**

- "Identify and segment the aortic isthmus on the three-vessel and trachea view."

At inference, the query specifies the imaging plane and requested structure. The external FetalCLIP branch selects the ScoreCAM target from $\mathcal{P}$ by image-text matching, as defined in Appendix A.1.