

Supplementary Material

Divyanshu Mishra¹, Prमित Saha¹, He Zhao³, Olga Patey², Aris T. Papageorghiou², J.Alison Noble¹

¹ Department of Engineering Science, University of Oxford

² Nuffield Department of Women's and Reproductive Health, University of Oxford

³ Department of Eye and Vision Science, University of Liverpool

Table 1: Table showing the effect of selecting top K queries using our query selection algorithm and averaging them to get the final visual query during inference for both ID and OOD VQ Bank.

Num Queries	VQ Data = In-Distribution				VQ Data = Out-of-Distribution			
	mtIoU	R @ 0.7	R @ 0.5	R @ 0.3	mtIoU	R @ 0.7	R @ 0.5	R @ 0.3
1	55.04	0.5	0.6	0.7	52.23	0.4	0.5	0.8
5	57.42	0.5	0.6	0.8	55.75	0.5	0.6	0.7
7	57.42	0.5	0.6	0.8	51.08	0.4	0.5	0.8
9	52.55	0.4	0.5	0.8	51.08	0.4	0.5	0.8

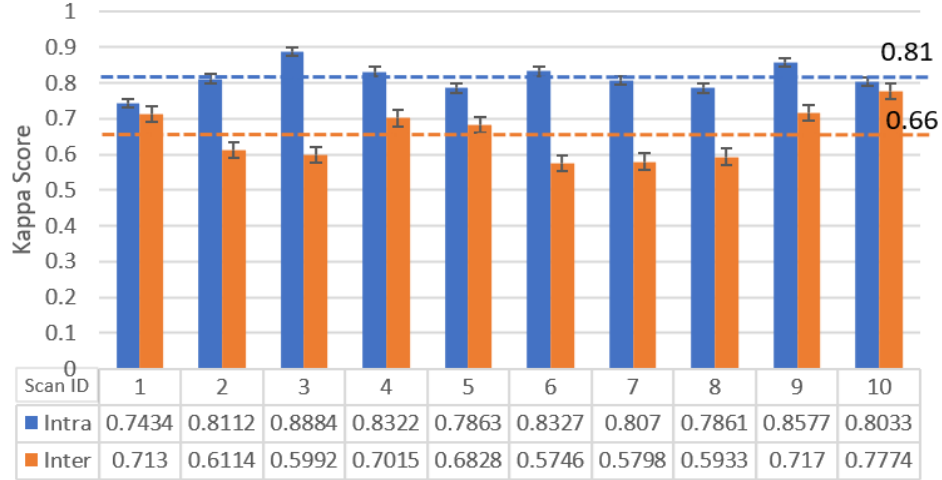


Fig. 1: Figure showing the inter and intra-annotation agreement of annotators for the task of standard frame detection in Transversal heart sweep(TS). We can see that the kappa score between annotators is only 66% highlighting the difficulty of the problem

Dataset and Training Details

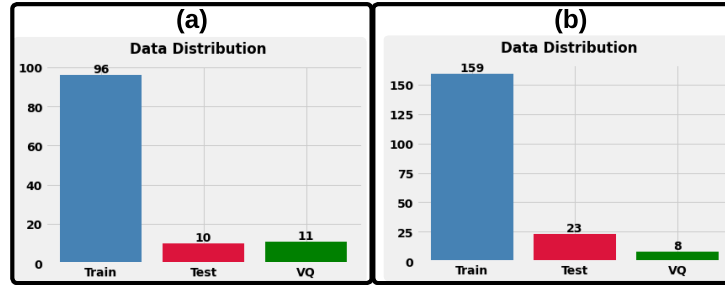


Fig. 2: Data Distribution of (a) Our Heart Sweep Data for the task of standard 4CH clip retrieval (b) PULSE data for the task of standard TV frame retrieval in fetal head video clips.

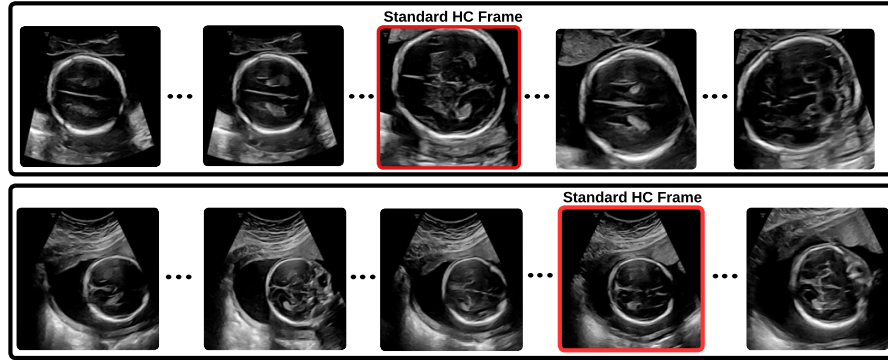


Fig. 3: Example fetal head ultrasound video frames depicting the task of standard fetal TV clip retrieval and showing similarity between standard and non-standard frames.

Table 2: Training Details

Component	Value
Framework	Pytorch
Pytorch Version	1.8
Optimizer	AdamW
Epochs	200
Number of Frames	150
Learning Rate	1e−05
LR Scheduler	Step
LR Scheduler Step Size	75
τ^+	0.7
τ^-	0.2
w1	1
w2	0.4
Query Guided Transformer Layers	1
Spatio-Temporal Transformer Layers	6
Classifier (\mathcal{C}_F) Architecture	ConvNext small
Visual Encoder (\mathcal{E}) Architecture	ResNet101
GPU	Tesla V100 32 GB