STAN-LOC: Visual Query-based Video Clip Localization for Fetal Ultrasound Sweep Videos

Divyanshu Mishra[1], Pramit Saha[1], He Zhao[3], Olga Patey[2], Aris T. Papageorghiou[2], J. Alison Noble[1]

[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

Abstract. Detecting standard frame clips in fetal ultrasound videos is crucial for accurate clinical assessment and diagnosis. It enables healthcare professionals to evaluate fetal development, identify abnormalities, and monitor overall health with clarity and standardization. To augment sonographer workflow and to detect standard frame clips, we introduce the task of Visual Query-based Video Clip Localization in medical video understanding. It aims to retrieve a video clip from a given ultrasound sweep that contains frames similar to a given exemplar frame of the required standard anatomical view. To solve the task, we propose STAN-LOC that consists of three main components: (a) a Query-Aware Spatio-Temporal Fusion Transformer that fuses information available in the visual query with the input video. This results in visual query-aware video features which we model temporally to understand spatio-temporal relationship between them. (b) a Multi-Anchor, View-Aware Contrastive loss to reduce the influence of inherent noise in manual annotations especially at event boundaries and in videos featuring highly similar objects. (c) a query selection algorithm during inference that selects the best visual query for a given video to reduce model's sensitivity to the quality of visual queries. We apply STAN-LOC to the task of detecting standard-frame clips in fetal ultrasound heart sweeps given four-chamber view queries. Additionally, we assess the performance of our best model on PULSE [2] data for retrieving standard transventricular plane (TVP) in fetal head videos. STAN-LOC surpasses the state-of-the-art method by 22% in mtIoU.

1 Introduction

In a routine pregnancy ultrasound assessment of the fetus, the sonographer scans through different fetal anatomies to evaluate fetal development and identify potential anomalies. For each anatomy, the sonographer reviews each frame meticulously and selects standard frames which are frames that contain all the anatomical landmarks in the correct anatomical orientation, size and position as defined by clinical guidelines (such as ISUOG [1, 15]). This process is time-consuming. Integrating a video-clip localization model has the potential to augment the sonographer's workflow allowing the sonographer to focus on detailed analysis and

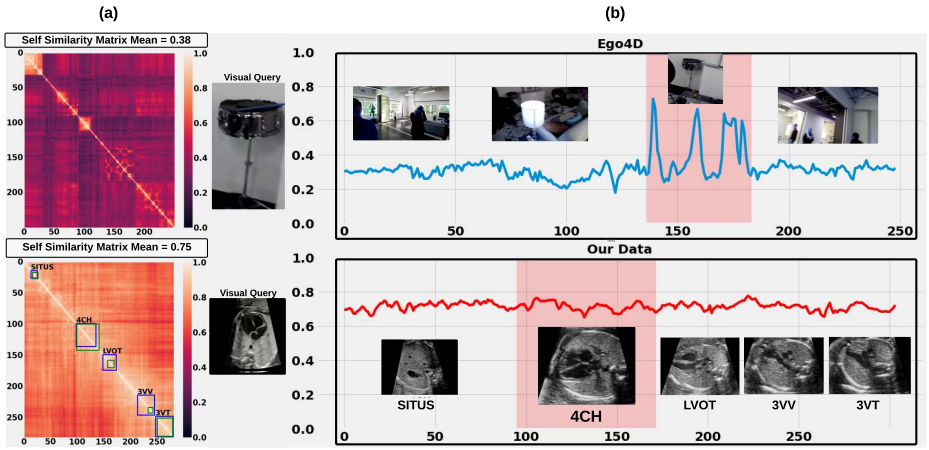

Fig. 1: (a) Self-similarity matrix for a randomly chosen video from Ego4D (top, mean=0.38) [4] and our clinical video dataset (bottom, mean=0.75), which reveals higher task difficulty for our video clip localization task. The uncertainty in annotations of two expert cardiologists are shown in green and blue boxes, respectively. (b) Cosine similarity of the visual query with the video for both Ego4D (top) and our data (bottom). Compared to Ego4D, our clinical data obtains similar scores along the video emphasizing the challenge whereas Ego4D exhibits high scores only within the region of interest.

anomaly detection. However, automatically detecting standard frames is challenging as the frames before/after the standard frames are highly similar, with often small misalignment of anatomical landmarks. Moreover, most of the views have high global structural similarity with only minor local variations, thereby making the detection of their temporal boundaries challenging as shown in Fig.1. As the task is complex, even experts can find it difficult to agree on what they refer to as a standard or non-standard frame as shown in Suppl. Fig. 1 that depicts a study where two cardiologists were asked to annotate the same 10 fetal heart videos. The kappa score [11] between the two experts was only 66% in this case, verifying the complexity of the task. This results in noisy annotations further complicating the issue. Existing works utilizing visual queries mainly comprise image retrieval [3, 8, 14, 13] and the recently defined task of visual query-based 2D localization (VQ2D) [17, 18, 7] in the Ego4D [4] dataset. However, both lines of work output a single image and typically utilise coarser-grained datasets. In scenarios like surgical procedure planning, disease screening/diagnosis, and procedure/process demonstration, users often need a video clip of the object rather than just a single image. Retrieving a video clip in our context is more challenging because along with the object, the ultrasound probe is in motion, leading to various deformations, occlusions, and motion blur. These factors deviate the object's appearance from the original query, making it harder for the model to accurately locate all its instances within the video clip.

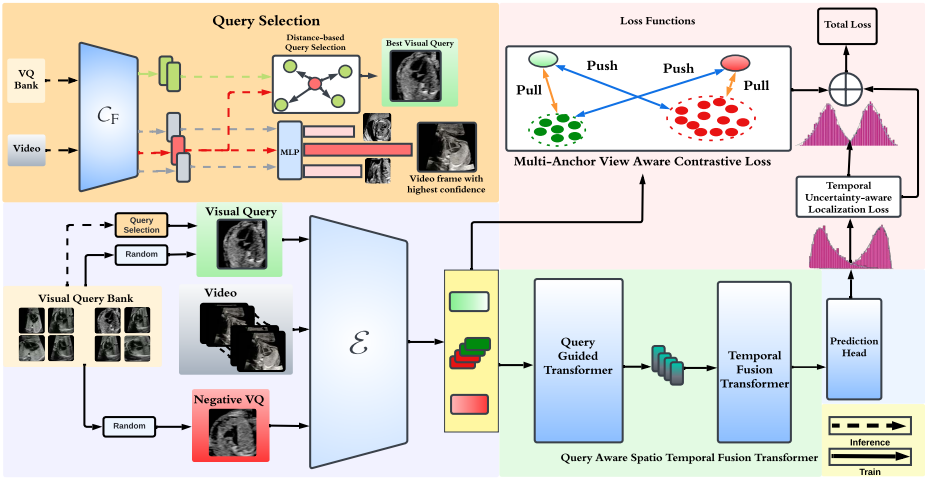

Fig. 2: Main architecture for STAN-LOC where input video, visual query (VQ), and negative VQ are passed through the backbone to extract features during training. The features from the visual query and the video are passed to the query-aware spatio-temporal fusion transformer and the resultant fused features are fed to the prediction head to predict the distributions of start and end frames. During inference, we select the best VQ from the visual query bank using the Query Selection algorithm.

In this paper, we introduce the Visual Query-based Video Clip Localization (VQ-VCL) task where, given an ultrasound scan, a sonographer provides an exemplar frame representing the anatomical view they wish to review. The model's objective is to retrieve a clip from the scan that contains the corresponding standard frames depicting the anatomy. We develop a query-aware spatio-temporal transformer model (STAN-LOC) that retrieves the clip-containing frames similar to the visual query from a given video.

Our contributions are: (a) We introduce the task of Visual Query-based Video Clip Localization (VQ-VCL) and propose a query-aware spatio-temporal transformer model, STAN-LOC, to automate this task. STAN-LOC includes a Query Aware Spatio-Temporal Fusion transformer to model the spatial and temporal relationship between the video and visual query. (b) To deal with noisy labels and challenging event boundaries, we include a Multi-Anchor, View-Aware Contrastive Loss and a Temporal Uncertainty Robust Localization Loss. (c) We propose a VQ selection module to guide the model during inference to select the best query candidate for a given input video. (d) We demonstrate STAN-LOC performance for two different real-world tasks of standard-frame detection with limited data availability and noisy labels.

2 Methods

Video Clip Retrieval Task Description: Visual query-based video clip localization (VQ-VCL) is formulated as a temporal localization task. Given a video v

and an exemplar query frame q from a separate exemplar database $\mathcal{Q}$, the model is trained to predict the start (t_s) and end (t_e) frames of a clip v_q where $v_q \subset v$ contains frames semantically similar to q.

STAN-LOC Overall Architecture: Our proposed architecture, as depicted in Fig. 2, takes a video v and a visual query q as inputs. These inputs are passed through a shared ResNet101 [6] encoder $\mathcal{E}$, resulting in video features $f_v \in \mathbb{R}^{\mathbf{TxHxWxC}}$ and visual query features $f_q \in \mathbb{R}^{\mathbf{HxWxC}}$. The extracted features are then fed to the Query-Aware Spatio-Temporal Fusion Transformer which first fuses the visual query with the video features and then models the resulting features temporally to yield spatio-temporal features. Finally, the spatio-temporal features are passed through an MLP responsible for predicting the distribution of start and end frames. During training, a Multi-Anchor View-Aware Contrastive Loss is proposed to make the model more sensitive to the query frame, which is further elucidated in Section 2.2. At inference, we integrate a query selection algorithm detailed in Section 2.3 to choose the most suitable query for the input video, enhancing the model's overall performance.

2.1 Query-Aware Spatio-Temporal Fusion Transformer

Query-Guided Spatial Transformer: The design of the encoder to fuse the video and the visual query features is crucial. Previous works for visual grounding [19], and moment retrieval [9] naively concatenate the features from video and query together. This reduces the relevance of visual queries and results in features possessing less information about the visual query [12]. To ensure that the video features (f_v) are contextualised by the information contained within the visual query features (f_q), we designed a Query-Guided Spatial Transformer. Specifically. we introduce cross-attention [16] layers to fuse the video and visual query features. Formally, given the video features f_v and visual query features f_q, we project video features to obtain query Q_v whereas key K_q and value V_q are obtained from the visual query feature f_q. The attention operation [16] is performed between Q_v, K_q and V_q to obtain query-guided feature QV_f, which can be formulated as:

$$QV_f = FFN\left(softmax(\frac{Q_v K_q^T}{\sqrt{d_k}})V_q\right), \tag{1}$$

where FFN is a feed-forward network and d_k is the dimensionality of the query and key vectors.

Temporal Fusion Transformer: To incorporate temporal information in the query-aware video features QV_f and fuse the spatio-temporal features, we designed a temporal fusion transformer. Formally, given QV_f, we first add fixed sinusoidal positional encoding to enrich the features with positional information. Then we perform self-attention [16] by projecting the resulting video features to Q_{v_q}, K_{v_q}, and V_{v_q} as shown in Eq. 2. This helps in modelling the temporal

interactions within the visual query-aware video features and generates spatio-temporal features F_T.

$$F_T = FFN\left(softmax(\frac{Q_{v_q}K_{v_q}^T}{\sqrt{d_k}})V_{v_q}\right) \tag{2}$$

2.2 Loss Functions

Multi-Anchor, View-Aware Contrastive Loss: In settings with high spatial similarity between the video frames as seen in Fig. 1, estimating the correct event boundary is an extremely challenging task. Moreover, as the objects of interest and the data acquisition device are both in motion, object appearance can strongly deviate from the visual query. To mitigate the above issues, we introduce a Multi-Anchor, View-Aware Contrastive Loss. The loss has two components: a) **Positive View-Aware Contrastive Loss ($\mathcal{L}_{PVAC}$)** which aims to bring the visual query features and the ground-truth clip features together while pushing away frames belonging to other classes; b) **Negative View-Aware Contrastive Loss ($\mathcal{L}_{NVAC}$)** which utilises a negative query (frame belonging to other classes) and aims to push the feature representation of positive frames in the video away from negative frames. Formally, given video-features f_v, visual-query features f_q and negative visual-query features f_{q^-}, we extract the video features belonging to the ground truth clip and consider them as positive features ($f_{v_i}^+$) while the video features of the frames lying outside the ground-truth clip are considered as negative features ($f_{v_j}^-$).

For $\mathcal{L}_{PVAC}$, we consider the visual query features f_q as the anchor and calculate the cosine similarity of f_q with $f_{v_i}^+$ and f_q with $f_{v_j}^-$ where i, j iterate over K_1 positive and K_2 negative features as shown by Eq. 3, where $sim(\cdot)$ indicates the cosine similarity function. Finally, we optimize the loss to pull positive features $f_{v_i}^+$ closer to the visual query feature f_q while pushing all K_2 negative $f_{v_j}^-$ away as formulated in in Eq. 3 where τ^+ is the positive temperature.

$$\mathcal{L}_{PVAC} = -\log \frac{\sum_{i=0}^{K_1} \exp\left(sim(f_q, f_{v_i}^+)/\tau^+\right)}{\sum_{j=0}^{K_2} \exp\left(sim(f_q, f_{v_j}^-)/\tau^+\right)} \tag{3}$$

On the other hand, for $\mathcal{L}_{NVAC}$ the negative visual query features f_{q^-} are considered as the anchor and we calculate the cosine similarity of f_{q^-} with $f_{v_i}^-$ and f_{q^-} with $f_{v_j}^+$ where i, j iterate over K_2 negative and K_1 positive features as shown in Eq. 4. Finally, we optimize the loss to pull the negative features $f_{v_i}^-$ closer to the negative visual query feature f_{q^-} while pushing all K_1 positive $f_{v_j}^+$ away as stated in Eq. 4 where τ^- is temperature for $\mathcal{L}_{NVAC}$.

$$\mathcal{L}_{NVAC} = -\log \frac{\sum_{i=0}^{K_2} \exp\left(sim(f_{q^-}, f_{v_i}^-)/\tau^-\right)}{\sum_{j=0}^{K_1} \exp\left(sim(f_{q^-}, f_{v_j}^+)/\tau^-\right)} \tag{4}$$

The final loss $\mathcal{L}_{MVAC}$ is denoted in Eq. 5 where w_p and w_n are tunable weights for $\mathcal{L}_{PVAC}$ and $\mathcal{L}_{NVAC}$ respectively.

$$\mathcal{L}_{MVAC} = w_p * \mathcal{L}_{PVAC} + w_n * \mathcal{L}_{NVAC} \tag{5}$$

Temporal Uncertainty Robust Localization Loss: The task of VQ-VCL becomes more challenging when there is a high similarity between the frames belonging to different classes and the event boundaries are unclear. This leads to noisy annotations available to train the model. To reduce sensitivity to noisy annotations, we introduce a Temporal Uncertainty Robust Localization Loss ($\mathcal{L}_{URL}$). Instead of using binary ground truth, we generate two Gaussian distributions $T_s(x)$ and $T_e(x)$ corresponding to the true start frame (t_s) and true end frame (t_e) of the target video clip, with means $\mu_s = t_s$ and $\mu_e = t_e$ and standard deviation ($\sigma = 1$) respectively as shown in Eq. 6 . Finally, we optimise the KL-divergence loss between the predicted ($P_s(x)$, $P_e(x)$) and true ($T_s(x)$, $T_e(x)$) start and end distribution and combine them together as shown in Eqs. 7 and 8 respectively.

$$T_s(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_s)^2/2\sigma^2},\ T_e(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_e)^2/2\sigma^2} \tag{6}$$

$$KL_s(P_s||T_s) = \sum_x P_s(x)\log(\frac{P_s(x)}{T_s(x)}),\ KL_e(P_e||T_e) = \sum_x P_e(x)\log(\frac{P_e(x)}{T_e(x)}) \tag{7}$$

$$\mathcal{L}_{URL} = KL_s + KL_e \tag{8}$$

Finally, we combine $\mathcal{L}_{MAC}$ and $\mathcal{L}_{URL}$ to give total loss ($\mathcal{L}$) which we use to train the model as expressed in Eq. 9.

$$\mathcal{L} = \mathcal{L}_{MAC} + \mathcal{L}_{URL} \tag{9}$$

2.3 Inference Query Selection

During inference, a user might provide queries which are low quality or extremely different from the VQ database $\mathcal{Q}$ used in training. To ensure that STAN-LOC is agnostic to the quality of visual queries, we introduce a classifier-based query selection module as shown in Fig. 2. The idea of the query selection is to provide a related query according to the input video, where the query frame is dynamically selected during inference. Given a video v, a visual query database $\mathcal{Q}_N$, where N is the number of visual queries, a reference frame F_v^{ref} from the video v is selected by a pre-trained classifier $\mathcal{C}_F$ with the highest confidence. Subsequently, we select M visual queries most similar to our reference frame F_v^{ref} by a distance function D (*e.g.,* Euclidean distance) between corresponding feature vectors. The M visual queries are averaged in the feature space to get the query feature for further retrieval process.

3 Experiments and Results

Dataset and Implementation: We evaluate STAN-LOC on two different fetal ultrasound video datasets. The first dataset gathered as part of the XX project, comprises fetal heart sweep videos for standard 4CH clip retrieval. The second

dataset is sourced from the PULSE [2] dataset and contains fetal head videos for standard fetal head TV clip retrieval. The fetal heart video sweep dataset comprised 10-second transversal heart sweeps (TS) over the fetal heart. A TS sweep is obtained by scanning from the cardiac situs (Situs) to the four-chamber view (4CH), through the left ventricular outflow tract (LVOT), the three-vessel view (3VV), and finally, the three-vessel trachea view (3VT) of the fetal heart. We utilised 96 videos for training and 10 videos for testing the model. The visual queries for heart data include 609 4CH frames extracted from 11 held-out videos (In-Distribution (ID)) and standard 4CH heart frames extracted from the PULSE [2] data where sonographers freeze the video and only capture standard frames (Out-of-Distribution (OOD)). The fetal head dataset comprises fetal head frames in Transventricular (TV) and Transcerebellar views (TC). The visual queries for this dataset comprise standard TV frames from 8 videos and we utilise 159 videos to train and 23 to test. For all datasets, the visual query and video frames were resized to 224×224 dimensions. We sampled clips with different start and end frames during training to augment the dataset. Further details are given in Dataset and Training Details section of the supplementary.

Results: STAN-LOC is compared with five different baseline models on two different datasets as shown in Table 1, where the models are ResNet3D[5], cosine-similarity supervised 2D CNN, TubeDETR [19],VQLOC [7] and MomentDETR [9], respectively. The chosen comparison metrics are Mean temporal intersection-over-union (mtIoU) and R @ t where R is recall, calculated at temporal IoU thresholds t. ResNet3D [5] exhibits the worst performance, with a mtIoU of 13.89 ± 3.67. Its R @ 0.7 is 0.02 and R @ 0.5 is 0.06 showing the model's inability to model longer-range interactions. TubeDETR [19], performs significantly better than ResNet3D with mtIoU of 27.85 ± 2.70 and R@ 0.5 = 0.22. However, R @ 0.7 of the model is 0.00, implying the model's failure to extract long-range features. Surprisingly, a simple cosine similarity supervised CNN baseline, outperforms the transformer-based TubeDETR with mtIoU of 29.43 ± 5.65 and R @ 0.5 of 0.28. This suggests that the model can learn the spatial correspondence between the video frame and the visual query but struggles with longer interactions (R@0.7=0), possibly due to the absence of temporal information in a 2D-CNN. MomentDETR has the best mtIoU (35.09 ± 3.27) and R@0.3 (0.58) across baselines, however, VQLOC surpasses it in R@0.7 (0.18) and R@0.5 (0.34), demonstrating superior performance in capturing longer interactions. STAN-LOC, with and without query selection, outperforms all baselines with mtIoU of 46.54, 55.04 and 57.42 respectively which is almost 22% more than MomentDETR. Its performance in modelling long-range dependencies is exceptional with R @ 0.7 = 0.50, R @ 0.5 = 0.60 and R @ 0.3 = 0.80 respectively.

Ablation Study: We performed an ablation study to evaluate the importance of each of the key STAN-LOC components on overall model performance. Refer to Table 2. In loss functions ablation, the first row displays the model with only Focal loss [10]. We observe that utilising $\mathcal{L}_{URL}$ instead of Focal loss in STAN-LOC boosts the performance by 14.22 % mtIoU indicating the importance of

Table 1: Quantitative Results. We test each baseline 5 times with different visual queries and report mean, and standard deviation. For STAN-LOC, we show the performance with and without Query selection (QS) where M is the number of best queries selected.

		Our Data				PULSE Data [2]			
Method		**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**
Resnet 3D [5]		13.89 ± 3.67	0.02 ± 0.04	0.06 ± 0.09	0.20 ± 0.10	43.45 ± 2.62	0.18 ± 0.04	0.43 ± 0.04	0.60 ± 0.04
TubeDETR[19]		27.85 ± 2.70	0.00 ± 0.00	0.22 ± 0.08	0.48 ± 0.08	55.91 ± 1.41	0.36 ± 0.04	0.57 ± 0.05	0.77 ± 0.02
Cosine Similarity Sup CNN		29.43 ± 2.38	0.00 ± 0.00	0.28 ± 0.08	0.50 ± 0.00	23.01 ± 0.15	0.17 ± 0.00	0.21 ± 0.02	0.26 ± 0.03
VQLOC [7]		30.87 ± 5.65	0.18 ± 0.04	0.34 ± 0.11	0.44 ± 0.11	42.83 ± 2.57	0.14 ± 0.02	0.34 ± 0.06	0.62 ± 3.64
MomentDETR [9]		35.09 ± 3.27	0.04 ± 0.05	0.32 ± 0.08	0.58 ± 0.11	57.20 ± 0.92	0.26 ± 0.06	0.64 ± 0.06	0.83 ± 0.02
STAN-LOC	W/O QS	46.54 ± 5.53	0.38 ± 0.08	0.50 ± 0.07	0.60 ± 0.10	58.35 ± 2.96	0.51 ± 0.06	0.59 ± 0.09	0.77 ± 0.06
	QS (M=1)	55.04 ± 0.00	0.50 ± 0.00	0.60 ± 0.00	0.70 ± 0.00	$\mathbf{58.67 \pm 0.00}$	0.57 ± 0.00	0.61 ± 0.00	0.83 ± 0.00
	QS (M=5)	$\mathbf{57.42 \pm 0.00}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.80 \pm 0.00}$	58.36 ± 0.00	$\mathbf{0.57 \pm 0.00}$	$\mathbf{0.61 \pm 0.00}$	$\mathbf{0.83 \pm 0.00}$

soft ground truth for noisy labels. Incorporating $\mathcal{L}_{PVAC}$ to STAN-LOC further improves the performance for mtIoU (+ 9.16%) and recall, demonstrating the importance of positive anchors and their role in pushing positive samples away from negative ones. Further, adding $\mathcal{L}_{NVAC}$ to STAN-LOC boosts the mtIoU to 57.42 showing the importance of a negative anchor and its role in pulling negative samples closer in the feature space and away from positive samples. In Query Selection ablation, we observed variability in performance when selecting random queries during inference with standard deviation (S.D) of 5.53% and 2.90% in mtIoU for ID and OOD VQ databases. We show our Query selection algorithm improves performance significantly. We also ablate different distance functions for query selection and the number of queries selected during inference. We find KL divergence to work well across datasets and visual queries for M=5 to work best. In the Architecture ablation, we observe that both query-guided and temporal fusion transformers are essential for best performance.

4 Conclusion

This paper introduces a novel task of Visual Query-based Video-Clip Localization and proposes a video-based transformer model STAN-LOC. STAN-LOC has two architectural components: Query-Guided and temporal-fusion transformers to fuse the query features with the video and further model interactions between these features in the temporal dimension respectively. To deal with noise at temporal class boundaries, a Multi-Anchor View-Aware contrastive loss and Temporal Uncertainty Robust Localization loss are introduced. Finally, to reduce model sensitivity to the quality of visual queries during inference, a test-time query selection algorithm is introduced to select the best query for the input video. The model is evaluated for two ultrasound video cases, where the video frames are highly similar and a low amount of training data is available. The effectiveness of the approach is demonstrated by comparing it with SOTA baselines and ablating different components.

Acknowledgments. This work was supported in part by the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1

Table 2: Ablation study showing effect of loss functions, query selection and architecture components on our model's performance where M is the number of VQ selected.

Loss Functions Ablation							
$\mathcal{L}_{\text{URL}}$	$\mathcal{L}_{\text{PVAC}}$	$\mathcal{L}_{\text{NVAC}}$	**mtIoU**	**R@0.7**	**R@0.5**	**R@ 0.3**	
✗	✗	✗	31.77	0.10	0.20	0.40	
✓	✗	✗	45.97	0.30	0.50	0.70	
✓	✓	✗	55.13	0.40	0.60	0.80	
✓	✓	✓	**57.42**	**0.50**	**0.60**	**0.80**	

Query Selection Ablation							
VQ Database	**QS**	**Distance Function**	**M**	**mtIoU**	**R@0.7**	**R@0.5**	**R@ 0.3**
In Distribution	✗	N/A	Random 5	46.54 ± 5.53	0.38 ± 0.08	0.50 ± 0.07	0.60 ± 0.10
	✓	Euclidean	1	48.27 ± 0.00	0.40 ± 0.00	0.50 ± 0.00	0.70 ± 0.00
	✓	Cosine Similarity	1	51.86 ± 0.00	0.40 ± 0.00	0.50 ± 0.00	0.80 ± 0.00
	✓	KL Divergence	1	55.04 ± 0.00	0.50 ± 0.00	0.60 ± 0.00	0.70 ± 0.00
	✓	KL Divergence	5	$\mathbf{57.42 \pm 0.00}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.80 \pm 0.00}$
Out Of Distribution	✗	N/A	Random 5	42.96 ± 2.90	0.34 ± 0.05	0.50 ± 0.0	0.54 ± 0.05
	✓	KL Divergence	1	52.23 ± 0.00	0.40 ± 0.00	0.50 ± 0.00	$\mathbf{0.80 \pm 0.00}$
	✓	KL Divergence	5	$\mathbf{55.07 \pm 0.00}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	0.70 ± 0.00

Architecture Ablation						
Query-guided Fusion	**Spatio-Temporal**	**mtIoU**	**R @ 0.7**	**R @ 0.5**	**R @ 0.3**	
✗	✓	42.36	0.20	0.40	0.60	
✓	✗	37.53	0.20	0.40	0.50	
✓	✓	**57.42**	**0.50**	**0.60**	**0.80**	

(Cardiovascular risks in early life and fetal echocardiography), the UK EPSRC (Engineering and Physical Research Council) Programme Grant EP/T028572/1 (VisualAI), and a UK EPSRC Doctoral Training Partnership award.

Disclosure of Interests. We have no competing interests.

References

1. Carvalho, J.S., Allan, L., Chaoui, R., Copel, J., DeVore, G., Hecher, K., Lee, W., Munoz, H., Paladini, D., Tutschek, B., et al.: Isuog practice guidelines (updated): sonographic screening examination of the fetal heart (2013)
2. Drukker, L., Sharma, H., Droste, R., Alsharid, M., Chatelain, P., Noble, J.A., Papageorghiou, A.T.: Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Scientific Reports **11**(1), 14109 (2021)
3. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 241–257. Springer (2016)
4. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
5. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 3154–3160 (2017)

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Jiang, H., Ramakrishnan, S.K., Grauman, K.: Single-stage visual query localization in egocentric videos. arXiv preprint arXiv:2306.09324 (2023)
8. Lee, S., Lee, S., Seong, H., Kim, E.: Revisiting self-similarity: Structural embedding for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23412–23421 (2023)
9. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
11. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)
12. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23023–23033 (2023)
13. Pan, T., Xu, F., Yang, X., He, S., Jiang, C., Guo, Q., Qian, F., Zhang, X., Cheng, Y., Yang, L., et al.: Boundary-aware backward-compatible representation via adversarial learning in image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15201–15210 (2023)
14. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2765–2775 (2023)
15. Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Chalouhi, G., Costa, F.D.S., Hernandez-Andrade, E., Malinger, G., Munoz, H., Paladini, D., et al.: Isuog practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology **59**(6), 840–856 (2022)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
17. Xu, M., Fu, C.Y., Li, Y., Ghanem, B., Perez-Rua, J.M., Xiang, T.: Negative frames matter in egocentric visual query 2d localization. arXiv preprint arXiv:2208.01949 (2022)
18. Xu, M., Li, Y., Fu, C.Y., Ghanem, B., Xiang, T., Pérez-Rúa, J.M.: Where is my wallet? modeling object proposal sets for egocentric visual query localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2593–2603 (2023)
19. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16442–16453 (2022)