MCAT: Visual Query-Based Localization of Standard Anatomical Clips in Fetal Ultrasound Videos Using Multi-Tier Class-Aware Token Transformer

Divyanshu Mishra[1], Pramit Saha[1], He Zhao[3], Netzahualcoyotl Hernandez-Cruz[1], Olga Patey[2], Aris 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]Institute of Life Course and Medical Sciences, University of Liverpool
{divyanshu.mishra, pramit.saha, netzahualcoyotl.hernandez-cruz, alison.noble}@eng.ox.ac.uk,
{olga.patey, aris.papageorghiou}@wrh.ox.ac.uk, he.zhao@liverpool.ac.uk

Abstract

Accurate standard plane acquisition in fetal ultrasound (US) videos is crucial for fetal growth assessment, anomaly detection, and adherence to clinical guidelines. However, manually selecting standard frames is time-consuming and prone to intra- and inter-sonographer variability. Existing methods primarily rely on image-based approaches that capture standard frames and then classify the input frames across different anatomies. This ignores the dynamic nature of video acquisition and its interpretation. To address these challenges, we introduce Multi-Tier Class-Aware Token Transformer (MCAT); a visual query-based video clip localization (VQ-VCL) method to assist sonographers by enabling them to capture a quick US sweep. By then providing a visual query of the anatomy they wish to analyze, MCAT returns the video clip containing the standard frames for that anatomy, facilitating thorough screening for potential anomalies. We evaluate MCAT on two ultrasound video datasets and a natural image VQ-VCL dataset based on Ego4D. Our model outperforms state-of-the-art methods by 10% and 13% mtIoU on the ultrasound datasets and by 5.35% mtIoU on the Ego4D dataset, using 96% fewer tokens. MCAT's efficiency and accuracy have significant potential implications for public health, especially in low- and middle-income countries (LMICs), where it may enhance prenatal care by streamlining standard plane acquisition, simplifying US based screening, diagnosis and allowing sonographers to examine more patients.

Introduction

Fetal ultrasound is essential for monitoring prenatal development, detecting potential abnormalities, and ensuring the health of both the fetus and the expectant mother. In routine pregnancy assessments, a sonographer scans different fetal anatomies to assess fetal development and identify anomalies. Selecting standard frames (Salomon et al. 2022; Hernandez-Cruz et al. 2025; Mishra et al. 2023) that meet clinical guidelines (*e.g.*, ISUOG) is a time-consuming process, and a typical fetal ultrasound scan can take up to an hour. Multiple studies have attempted to streamline this process by automatically identifying standard planes in 2D fetal ultrasound using deep learning-based classification models (Rahmatullah, Papageorghiou, and Noble 2011; Cai et al. 2018; Lee, Gao, and Noble 2021; Baumgartner et al. 2016; Schlemper et al. 2018). Other recent works have looked into leveraging temporal information for more complex tasks, such as anomaly detection in ultrasound videos (Zhao et al. 2022, 2023) and generative modeling of standard planes (Men et al. 2023), although these approaches do not explicitly localize standard frames. Integrating a video-clip localization model could enhance sonographer workflow by allowing them to focus on detailed video reviews and anomaly detection. However, automatically detecting standard frames in video is challenging due to the high similarity of frames before and after the standard ones, making it difficult to determine temporal anatomical boundaries, as shown in Fig. 1. Additionally, even human experts may find it hard to agree on standard frame selection, as evidenced by our study showing a kappa score of only 66% between two fetal cardiologists annotating the same fetal heart videos (see Supp. Fig. 1), highlighting the complexity and inherent noise in annotations. Text query-based localization tasks, such as video-temporal grounding (Yang et al. 2022; Zeng et al. 2020; Lin et al. 2023), video moment retrieval (Liu et al. 2022; Moon et al. 2023), and highlight detection, have shown promising performance in natural video understanding. However, textual data often falls short of providing the dense video understanding required for some applications. In the medical domain, reports traditionally rely on static images and text to convey diagnostic information (Moon et al. 2022; Saha et al. 2024a,b,c). While image-based methods are informative, video-based analysis can offer a significant advancement in diagnostic capabilities. For instance, a dynamic ultrasound video of a beating heart provides a more detailed and holistic assessment of cardiac function compared to a single static frame (Scott et al. 2013). Similarly, in fetal anatomy examinations, video clips allow practitioners to measure biometry more accurately and review the entire sequence for optimal plane selection, thereby enhancing diagnostic precision. Despite the advantages, paired video-textual data is typically scarce in the medical field. When available, it usually includes sparse class labels or radiology reports providing a diagnosis for the entire video rather than detailed clip-level information. This is where image-based queries, or visual queries (VQs), are potentially valuable. VQs allow for intuitive and direct identification of ob-

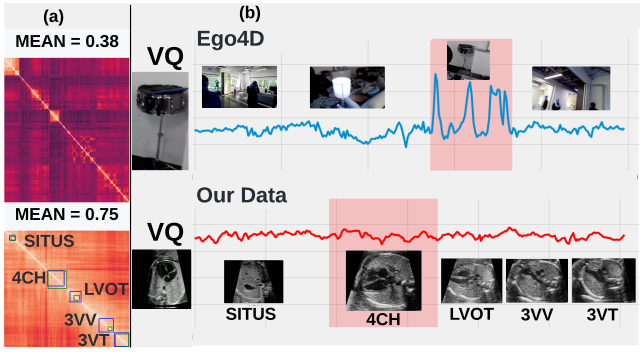

Figure 1: (a) Self-similarity matrix for a randomly chosen video from Ego4D (**top**, mean=0.3) (Grauman et al. 2022) and our clinical video dataset (**bottom**, mean=0.75), which reveals higher task difficulty for our video clip localization task. The uncertainty in the annotations of two expert cardiologists is shown in green and blue boxes. (b) Cosine similarity of the visual query with the video for both Ego4D (top) and our data (bottom). Our clinical data obtains similar scores along the video emphasizing the challenge, whereas Ego4D exhibits high scores only within region of interest.

jects or similar images, reducing language barriers and effectively expressing complex concepts that might be difficult to articulate through text. For example, describing a medical anomaly can be challenging with a text query, whereas an example frame containing the anomaly can provide a more effective query for model training. In the context of ultrasound videos, retrieving a video clip rather than a single frame is more challenging due to the motion of the ultrasound probe and the scanned object, leading to various deformations, occlusions, and motion blur, making it harder to localize all instances of the object as shown in Fig. 1. To reduce the time to conduct a full scan assessment, we introduce the Visual-Query-based Video Clip Localization (VQ-VCL) task. In this approach, a sonographer performs a quick sweep to capture all relevant anatomies. With a visual query depicting the required anatomy, our method can automatically select the relevant standard-frame clips from this video sweep. This significantly reduces manual effort, enhances efficiency, and allows sonographers to scan more patients while focusing more on analyzing the standard video clips.

To tackle the challenges of the VQ-VCL task, we introduce MCAT, a Multi-Tier transformer-based model with class-specific tokens. It consists of three primary components: a Multi-Tier Class-Aware Spatio-Temporal Transformer for modeling spatial and temporal interactions and learning class-specific features through class-specific tokens, a Temporal Uncertainty Localization Loss to mitigate label noise, and a Multi-Tier, Dual Anchor Contrastive Loss for addressing complex event boundaries.

Our contributions are as follows:

1. We introduce the VQ-VCL task and propose MCAT, a spatio-temporal video Transformer model for automatic standard-plane video clip retrieval.
2. We propose a multi-tier feature extraction module to learn spatio-temporal features in a coarse-to-fine manner. A query-aware Transformer captures spatial information, while temporal information is condensed into class-specific learnable tokens. These tokens disentangle class-specific features into distinct tokens, improving video clip localization and significantly boosting model efficiency by reducing the number of tokens by 96%. This makes the approach potentially suitable for applications in resource-constrained public health including low- and middle-income country (LMIC) settings.
3. We propose a hybrid loss function comprising Multi-Tier, Dual Anchor Contrastive Loss, and Temporal Uncertainty-Aware Localization Loss to handle complex event boundaries and noisy labels.
4. We assess model performance on two real-world clinical datasets for standard-plane detection with limited data and annotations which naturally contain a high degree of noise. Additionally, we create and evaluate our model on an open-source VQ-VCL natural videos dataset based on Ego4D (Grauman et al. 2022).

Methods

Video Clip Localization Task Formulation

The visual query-based video clip localization (VQ-VCL) task is formulated as a temporal localization task. Formally, given a video v and an exemplar frame q from a separate exemplar database $\mathcal{Q}$, the model is trained to predict the start (t_s) and end (t_e) frame number of a clip v_q where $v_q \subset v$ and contains frames semantically similar to q.

MCAT Overall Architecture

Our proposed model, illustrated in Fig. 2, processes a video v and a visual query q as inputs. These inputs are fed through a shared encoder $\mathcal{E}$, generating K tier video features $f_{v_k} \in \mathbb{R}^{T\times H_k\times W_k\times C_k}$ and visual query features $f_{q_k} \in \mathbb{R}^{H_k\times W_k\times C_k}$, where k iterates over the K tiers and T, H_k, W_k, and C_k represent the number of frames, height, width, and channel dimensions at tier k. The features extracted at each tier from a visual backbone $\mathcal{E}$ are enriched with scale-aware learnable embeddings and spatially fused using a Multi-Tier Query-Guided Spatial Transformer, resulting in multi-tier VQ aware features as shown in Fig. 2. A multi-tier temporal fusion transformer is proposed to learn a series of tokens (CLS, $\{C_i|i = 1, \cdots, N\}$) from the VQ aware features where N is the number of classes. The tokens are further utilized to predict the start and end frames.

Multi-Tier Spatio-Temporal Transformer

Multi-Tier Query Guided Spatial Transformer The design of the encoder to fuse the video and visual query features is crucial, especially in fine-grained video localization settings where the classes are highly similar. Previous work on visual grounding (Yang et al. 2022) and moment retrieval (Lei, Berg, and Bansal 2021) naively concatenates the features from the video and query together. This approach can diminish the relevance of visual queries and result in features with low information about the visual query (Moon et al. 2023). Moreover, these works are designed for text

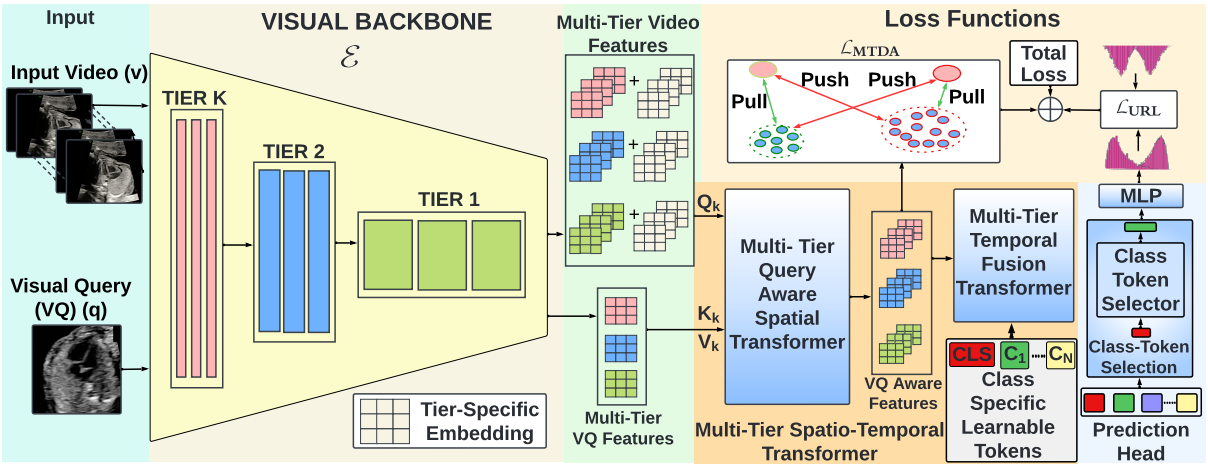

Figure 2: Main architecture of MCAT. The input video v and visual query q are passed to the visual backbone to give multi-Tier features. These features are fused spatially using the Multi-Tier Query Aware Spatial Transformer. The Tier-specific features are passed to a) $\mathcal{L}_{MTDA}$ to learn the separation between classes, b) the Multi-Tier Temporal Fusion transformer to learn Tier-Aware Spatio-Temporal Embedding, which is further passed to an MLP to make final prediction and calculate $\mathcal{L}_{URL}$ loss.

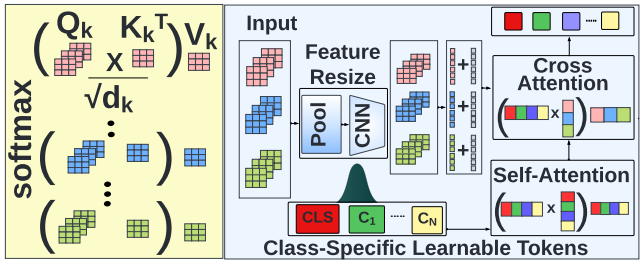

Figure 3: Fig (left) shows the spatial feature fusion mechanism where Tier-specific video and VQ features are spatially fused to give Tier-specific query-aware features. Figure 3 (right) shows how the Tier-specific query-aware features are first resized, flattened and enriched with positional information. The resulting features are concatenated and fused to learn the Tier-Aware Spatio-Temporal Embedding.

query-based video retrieval where modality features are only extracted in a single hierarchy. In contrast, features from videos and images can be extracted at multiple tiers, each tier containing coarse to fine-grained information. This variability in information can be beneficial for retrieval, especially in scenarios where the classes are highly similar with some local variations. To ensure video features at Tier k (f_{v_k}) are contextualized by visual query features (f_{q_k}) from the respective tier, we designed a Multi-Tier Query Guided Spatial Transformer where $k = 1, 2, 3, \ldots, K$. We achieve this by extracting features from K tiers of the shared visual backbone for the video and the visual query. Tier-specific learnable embeddings (emb_K) are added to each tier video feature to ensure optimal learning and fusion of tier-specific information from the video and visual query. The resulting video features and visual query features for each tier are then fused using cross-attention (Vaswani et al. 2017) to learn these tier-specific embeddings (emb_K). Formally, given the video feature f_{v_k} and visual query feature f_{q_k} at tier k where $k = 1, 2, 3, \ldots, K$ and $k = 1$ means the features from the last layer of the visual backbone. We first add to each tier video feature the tier-specific learnable embedding (emb_K) such that $f_{v_k} = f_{v_k} + emb_k$. We project the video feature to get query (Q_{v_k}), whereas key (K_{q_k}) and value (V_{q_k}) are obtained from the visual query feature. The attention mechanism (Vaswani et al. 2017) is applied to Q_{v_k}, K_{q_k}, and V_{q_k}, and the output is passed to a feed-forward network as shown in Eq. 1 to produce the tier-specific query-aware video features (QV_{f_k}) for tier k. This process is performed in parallel for all K tiers to obtain tier-specific query-aware features QV_{f_K} for each tier.

$$QV_{f_K} = FFN\left(\text{softmax}\left(\frac{Q_{v_k}K_{q_k}^T}{\sqrt{d_k}}\right)V_{q_k}\right) \quad (1)$$

Multi-Tier Temporal Fusion Transformer To incorporate temporal information into the Tier-specific query-aware video features QV_{f_k} and to fuse these spatio-temporal features across Tiers for learning the class-specific Tier-aware spatio-temporal tokens (CLS, $\{C_i | i = 1, \cdots, N\}$) where N is number of classes, we designed a multi-Tier temporal fusion transformer. Formally, given the Tier-specific query-aware features for each Tier QV_{f_k}, and randomly initialized class-selection token CLS, Class-Specific Tier-Aware Spatio-Temporal learnable tokens $C_N \in \mathbf{R}^{(\mathbf{N})\times\mathbf{H_M}\times\mathbf{W_M}\times\mathbf{C_M}}$ where k= $1, 2, 3 \ldots K$. We first perform self-attention between the $E_T = CLS + C_N$ tokens as in Eq. 2.

$$E_T = FFN\left(\text{softmax}\left(\frac{Q_v\left(K_v^T\right)}{\sqrt{d_k}}\right)V_v\right) \quad (2)$$

Cross-attention is performed between the resulting vector and the Tier-specific query-aware video features QV_{f_k} as

formulated in Eq.3 and shown in Supp. Fig 3.

$$E_T = FFN\left(\text{softmax}\left(\frac{Q_{E_T}\left(K^T_{QV_f}\right)}{\sqrt{d_k}}\right)V^T_{QV_f}\right) \quad (3)$$

This helps fuse the spatial and temporal information available across the Tiers into the class-specific Tier-Aware Spatio-Temporal tokens. The spatio-temporal information-rich E_T tokens are fed to the token selection block that helps select the token corresponding to the VQ and updates only the selected token with the current spatio-temporal class-specific information.

Class-Specific Token Selection and Learning The block is designed to select the class-specific token (C_S) corresponding to the visual query and to enable class-specific token learning. During training, the class-selection token (CLS) obtained after spatio-temporal fusion is passed through a multi-layer perceptron (MLP) to predict the class to which the visual query (VQ) belongs. This is formulated as an N-class classification problem, and cross-entropy loss is used to train the MLP. Since the class of the VQ is known during training, we use this information to select the class-specific token (C_S) and only update the token for the specific VQ class. During inference, the prediction from the trained MLP is used to select C_S and predict the start and end frames of the ground-truth video clip.

Loss Functions

Multi-Tier, Dual Anchor Contrastive Loss In settings with high spatial similarity between the video frames, as seen in Fig. 1, estimating the correct event boundary is challenging. Moreover, in such a case, object appearance can significantly vary from that of the visual query as the objects of interest and the data acquisition device are both in motion. To mitigate the above challenges and learn subtle differences between the classes, we propose a Multi-Tier, Dual Anchor Contrastive Loss, where the anchors and samples are selected from different tiers. The loss function has two main components: 1. **Multi-Tier Positive Anchor Contrastive Loss (L_{PAC})**, which aims to bring the tier-specific visual query-aware features in the ground-truth clip together while pushing away features belonging to other classes. 2. **Multi-Tier Negative Anchor Contrastive Loss (L_{NAC})**, which utilizes a negative anchor to further push the positive tier-specific query-aware features away from the negative ones. Formally, given Tier-Specific Query Aware features f_{vq_k} for each tier, we project the features to a shared feature space to ensure that only rich-semantic features from each tier are captured. This is achieved through a CNN projection layer $P_{\theta k}$, resulting in projected features f'_{vq_k}. Subsequently, we extract the video features belonging to the ground-truth clip and define them as positive features ($f'^{+}_{vq_k}$) for each tier. The video features of the frames lying outside the ground-truth clip are defined as negative features ($f'^{-}_{vq_k}$). We randomly sample a tier and utilize its features as anchors. A tier's positive features serve as the positive ($f^{+}_{vq_a}$), while the negative features serve as the negative anchor ($f'^{-}_{vq_a}$) for the remaining tiers. For the Positive Anchor Contrastive Loss, we calculate the cosine similarity between ($f'^{+}_{vq_a}$, $f'^{+}_{vq_{k,i}}$) and ($f'^{+}_{vq_a}$, $f'^{-}_{vq_{k,j}}$) as stated in Eq. 4, where sim(.) denotes the cosine similarity function and i, j iterate over M_1 positive and M_2 negative samples, while k iterates over $K-1$ tiers.

$$\mathcal{L}_{PAC} = -\log\frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_1}\exp\left(sim(f'^{+}_{vq_a}, f'^{+}_{vq_{k,i}})/\tau^{+}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_2}\exp\left(sim(f'^{+}_{vq_a}, f'^{-}_{vq_{k,j}})/\tau^{+}\right)} \quad (4)$$

Finally, we optimize the loss function to pull positive features $f'^{+}_{vq_{k,i}}$ closer to the positive anchor feature $f'^{+}_{vq_a}$ while pushing all M_2 negative features $f'^{-}_{vq_{k,j}}$ away, as formulated in Eq. 4, where τ^{+} is the positive temperature.
On the other hand, for $\mathcal{L}_{NAC}$, we consider the negative features of the randomly selected tier as the negative anchor ($f'^{-}_{vq_a}$). We calculate the cosine similarity between ($f'^{-}_{vq_a}$, $f'^{-}_{vq_{k,i}}$) and ($f'^{-}_{vq_a}$, $f'^{+}_{vq_{k,j}}$), where i and j iterate over M_2 negative and M_1 positive features, respectively, while k iterates over $K-1$ tiers, as stated in Eq. 5.

$$\mathcal{L}_{NAC} = -\log\frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_2}\exp\left(sim(f'^{-}_{vq_a}, f'^{-}_{vq_{k,i}})/\tau^{-}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_1}\exp\left(sim(f'^{-}_{vq_a}, f'^{+}_{vq_{k,j}})/\tau^{-}\right)} \quad (5)$$

Finally, we optimize the loss to pull the negative features ($f'^{-}_{vq_{k,i}}$) closer to the negative anchor features ($f'^{-}_{vq_a}$) while pushing all M_1 positive features ($f'^{+}_{vq_{k,j}}$) away, as shown in Eq. 5, where τ^{-} is the temperature parameter for $\mathcal{L}_{NAC}$. The final loss $\mathcal{L}_{MTDA}$ is given in Eq. 6 where w_p and w_n are tunable weights for $\mathcal{L}_{PAC}$ and $\mathcal{L}_{NAC}$ respectively.

$$\mathcal{L}_{MTDA} = w_p * \mathcal{L}_{PAC} + w_n * \mathcal{L}_{NAC} \quad (6)$$

Temporal Uncertainty Aware Localization Loss The VQ-VCL task becomes more challenging when there is a high similarity between the frames belonging to different classes and the event boundaries are not well defined. This leads to noisy manual annotations. To reduce the effect of noisy annotations, we introduce a Temporal Uncertainty Aware 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. 7.

$$T_s(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu_s)^2/2\sigma^2},\ T_e(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu_e)^2/2\sigma^2} \quad (7)$$

Finally, we optimize 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 as shown in Eqs. 8 and 9 respectively.

$$KL_s(P_s||T_s) = \sum_x P_s(x)\log(\frac{P_s(x)}{T_s(x)}),$$
$$KL_e(P_e||T_e) = \sum_x P_e(x)\log(\frac{P_e(x)}{T_e(x)}) \quad (8)$$

$$\mathcal{L}_{URL} = KL_s + KL_e \quad (9)$$

Finally, we combine Eqs 6 and 9 to give the total loss $\mathcal{L}$ used to train our model as formulated in Eq. 10.

$$\mathcal{L} = \mathcal{L}_{MTDA} + \mathcal{L}_{URL} \quad (10)$$

Experiments and Results

Dataset and Implementation We evaluated MCAT on two distinct fetal ultrasound video datasets following (Mishra et al. 2024) and one egocentric computer vision dataset, Ego4D VQ-VCL, which we created based on the Ego4D dataset (Grauman et al. 2022). The first dataset consists of fetal heart video sweeps from CAIFE (Development of Clinical Artificial Intelligence Models in Fetal Echocardiogra- phy for the Detection of Congenital Heart Defects). It includes 10-second transversal heart sweeps over the fetal heart (see Supp. Fig 3), 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. Unlike routine heart scans, where the sonographer pauses to capture the perfect plane for each anatomical view, these sweeps continuously scan across the heart. The VQ-VCL task retrieves a standard heart-view clip given a visual query of the standard heart view. We used 200 healthy heart sweep videos for training and 47 videos for testing. The visual query for the heart sweep data consisted of 2804 standard frames extracted from 12 held-out videos. Further details about our unique heart sweep data are in Ultrasound dataset details section of Supplementary. Our second dataset is derived from the PULSE (Drukker et al. 2021) fetal ultrasound anomaly scan video dataset. We extracted clips for 8 fetal anatomical planes utilized for clinical anomaly detection, including Transventricular and Transcerebellar Views of the fetal head, Abdomen, Femur, and the 4CH, LVOT, 3VV, and 3VT views of the fetal heart. We trained the MCAT model on 200 videos and tested it on 30 videos. The visual query comprised 4378 standard frames extracted from 30 videos. As we introduce the VQ-VCL task, we acknowledge the lack of open-source datasets for model reproducibility. Therefore, we utilized the existing Ego4D (Grauman et al. 2022) dataset to create the Ego4D VQ-VCL dataset. We plan to release the dataset creation script along with our code to ensure the reproducibility of our work. Video and visual query frames were resized to dimensions of 224×224. During training, we augmented the dataset by sampling clips with varying start and end frames, each containing 150 frames. All models were trained for 200 epochs in PyTorch version 1.8 using a Tesla V100 32 GB GPU. We employed AdamW optimizer with a StepLR learning scheduler, utilizing cosine annealing with a step-size of 75. Our visual encoder was ResNet101, and both our multi-tier feature fusion transformers consisted of 2 layers each.

Metrics To evaluate the performance of MCAT, we follow previous works on temporal video grounding (Wang et al. 2023; Moon et al. 2023) and our baselines (Goyal et al. 2023; Jiang, Ramakrishnan, and Grauman 2023). Hence, we compute the mean temporal intersection-over-union (mtIoU) and "R @ t", where R represents recall measured at predefined temporal IoU (tIoU) thresholds (t). For our experiments, we report recall at thresholds $t = 0.1, 0.3, 0.5$ and 0.7.

Heart Sweep Data					
Method	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**R@0.1**
CS Sup CNN	5.03	0.00	0.00	4.00	16.22
TubeDETR	12.72	2.00	2.00	10.22	20.00
MomentDETR	14.89	0.00	8.00	25.00	39.72
Resnet 3D	19.79	6.00	6.00	23.22	47.17
VQLOC	24.05	2.50	13.50	34.50	62.17
MCAT (Ours)	**34.1**	**11.00**	**30.17**	**56.17**	**66.17**
PULSE (Drukker et al. 2021) Data					
Method	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**R@0.1**
CS Sup CNN	2.6	2.04	2.04	2.04	2.04
TubeDETR	7.07	4.76	4.76	4.76	14.42
MomentDETR	10.34	2.04	6.93	16.60	21.50
Resnet 3D	17.89	14.29	17.14	22.04	28.16
VQLOC	12.62	0.00	14.29	14.29	22.04
MCAT (Ours)	**30.63**	**26.80**	**31.70**	**34.56**	**39.32**
Ego4D (Grauman et al. 2022) VQ-VCL Dataset					
Method	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**R@0.1**
CS Sup CNN	4.89	0.00	0.00	7.93	15.87
Resnet 3D	10.72	3.57	8.33	12.70	20.24
VQLOC	25.35	7.14	19.44	32.94	44.84
MomentDETR	38.44	15.08	25.40	52.78	66.67
TubeDETR	38.59	18.65	32.94	61.51	71.03
MCAT (Ours)	**43.94**	**32.54**	**39.68**	**64.68**	**71.83**

Table 1: Quantitative comparison of MCAT

Quantitative Results We compare, MCAT with ResNet3D CNN (Hara, Kataoka, and Satoh 2017), Cosine Similarity Supervised 2D CNN, TubeDETR (Yang et al. 2022), VQLOC (Jiang, Ramakrishnan, and Grauman 2023), and MomentDETR (Lei, Berg, and Bansal 2021), as in Table 1. Further details about baselines are in Supp.

From Table 1, we observe that the cosine similarity supervised baseline performs worst, achieving an mtIoU of 5.03%, R @ 0.7 = 0.00% , R @ 0.5 = 0.0% . This indicates the model's inability to effectively extract, fuse, and model long-range features from both the input video and the visual query. TubeDETR demonstrates improved performance with an mtIoU of 12.72%, R @ 0.3 = 10.22%. This improvement may be attributed to the spatio-temporal transformer in TubeDETR, facilitating the extraction and fusion of video features in both spatial and temporal dimensions. However, the model still struggles with longer

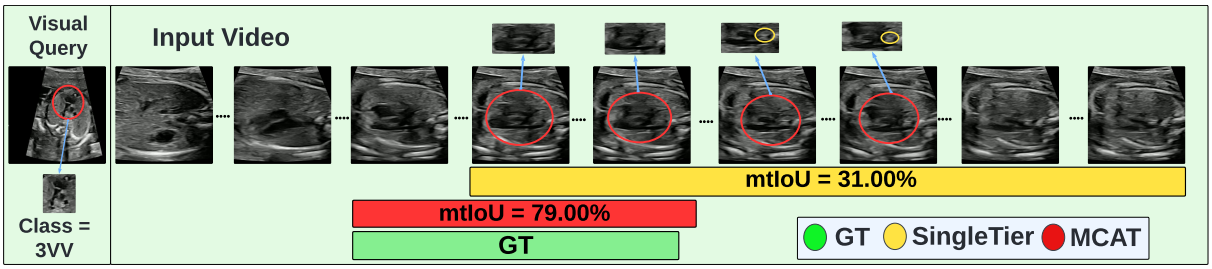

Figure 4: This figure compares the predictions of a single-tier model with our multi-tier model for an LVOT visual query.

interactions, as indicated by R @ 0.7 and R @ 0.3 both being 2.00%, suggesting that the features from the visual query and the video are insufficient for modeling extended interactions. This limitation may be due to the direct concatenation of video and visual query features in the model. A similar pattern is observed with MomentDETR, where mtIoU is 14.89%. The model handles short-range interactions well with R @ 0.3 = 25.00% and R @ 0.1 = 39.72%, but it performs poorly in capturing longer-range interactions (R @ 0.7 = 0.00% and R @ 0.5 = 8.00%), possibly due to the concatenation of visual query and video features. The ResNet3D baseline outperforms TubeDETR and MomentDETR, achieving an mtIoU of 19.79% and demonstrating better modeling of longer interactions with R @ 0.7 = 6.00% and R @ 0.5 = 6.00%. ResNet3D also performs well in modeling shorter interactions with R @ 0.1 = 47.17%. This improvement may be attributed to the equal interaction of the visual query with each frame of the video, achieved by concatenating them together for each frame. VQLOC achieves an mtIoU of 24.05%, R @ 0.5 = 13.50%, R @ 0.1 = 62.17%, indicating its ability to model both short and long-range interactions. MCAT outperforms all baselines with a mtIoU of 34.10%, 10.05% higher than VQLOC. Its performance in modeling long-range and short-range dependencies is significantly better, with R @ 0.7 = 11.00%, R @ 0.5 = 30.17%, R @ 0.3 = 56.17%, and R @ 0.1 = 66.17%. MCAT effectively models the relationship between the visual query and the video in both spatial and temporal dimensions due to the Multi-Tier Spatio-Temporal Fusion Transformer and disentanglement of class-specific spatio-temporal features through class-specific tokens. Additionally, the incorporation of boundary losses ($\mathcal{L}_{MTDA}$ and $\mathcal{L}_{URL}$) enhances its ability to detect boundaries, making it robust to noisy annotations and resulting in improved performance. A similar trend is seen in the PULSE (Drukker et al. 2021) data (refer to Table 1) and Ego4D VQ-VCL (refer to Table 1), with our model MCAT outperforming the best-performing baseline by 12.74% and 5.35%, respectively.

Qualitative Results In the qualitative comparison, we analyze the single-tier model, which only utilizes features from the last layer (Tier=1), against our multi-tier model (Tiers=1, 2, 3). As shown in Fig. 4, the single-tier model struggles to differentiate between the 3VV and 3VT views in the video, as both views display three vessels. The critical distinction is the appearance of the trachea in the 3VT view, as highlighted by the yellow circle. Our multi-tier model successfully identifies this subtle change by leveraging features from multiple tiers, leading to significantly improved performance. Additional qualitative results are provided in the supplementary.

Ablation Study

This section reports ablation experiments to justify the inclusion of the key components in the MCAT model.

Importance of Tiers First we show the importance of Tiers to model performance. The model utilizing features only from the last layer of the visual backbone is referred to as Tier 1, while that utilising from Tier 1 and some layer before that is referred to as Tier 2, and so on. Experiments were performed for Tier = 1,2,3 where the feature sizes were $T \times 2048 \times 7 \times 7$, $T \times 1024 \times 14 \times 14$ and $T \times 512 \times 28 \times 28$ respectively. Table 2 shows that the model utilising only Tier 1 features performs the worst with mtIoU = 28.23 %. This can be explained because Tier 1 features are insufficient to capture the fine-grained detail necessary to determine event boundaries and to distinguish between highly similar classes. When features from Tier 1 and Tier 2 are utilized, mtIoU increases by 3.1%. The highest performance is seen with Tier 3, which surpasses the single Tier results by almost 6% and Tier 2 by 2.77% respectively stressing the advantage of using multi-Tier features to capture both low- and high-level details to distinguish fine-grained classes.

Tier	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
1	28.23	10.5	19.72	38.72	**66.61**
1, 2	31.33	**13.0**	28.22	42.44	**66.61**
1, 2, 3	**34.10**	11.00	**30.17**	**56.17**	**66.17**

Table 2: Showing importance of multi-Tier features.

Importance of different Loss functions In Table 3, we investigate the impact of each loss function on model performance. Our baseline loss is Cross-Entropy Loss, as shown in the first row of Table 3. Replacing it with $\mathcal{L}_{URL}$ improved performance by 14% mtIoU, highlighting the significance of incorporating uncertainty in loss functions when the ground truth annotation contains a high degree of noise. Additionally, we ablate the Multi-Tier, Dual Anchor Contrastive loss ($\mathcal{L}_{MTDA}$). Including this loss, which consists of our dual-anchor losses ($\mathcal{L}_{PAC}$ and $\mathcal{L}_{NAC}$), alongside $\mathcal{L}_{URL}$, further enhances performance. Specifically, mtIoU increases by

4.46%, R @ 0.5 by 3.95%, R @ 0.3 by 11.23%, and R @ 0.1 by 9.78%. These improvements indicate the importance of both positive and negative anchors to distinguish highly similar classes and to reduce confusion at event boundaries.

$\mathcal{L}_{URL}$	$\mathcal{L}_{MTDA}$	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
✗	✗	15.93	2.00	12.00	22.22	42.44
✓	✗	29.64	11.50	26.22	44.94	56.39
✓	✓	**34.10**	**11.00**	**30.17**	**56.17**	**66.17**

Table 3: Analysis of contribution of different loss functions.

Sequential vs Parallel Fusion In existing works utilizing multi-scale features (Wang et al. 2021; Fan et al. 2021), sequential feature fusion is employed. In sequential fusion, multi-scale features are projected into a common feature space and fused sequentially from coarse to fine. In contrast, our work employs parallel fusion in the encoder to exploit the implicit bias of image/video modalities and capture minute variations. In parallel fusion, features from the video and visual query at each tier are fused separately, maintaining the original resolution across tiers. As shown in Table 4, parallel fusion outperforms sequential fusion, improving R@0.3 by 11.45%, R@0.5 by 6.67%, and mtIoU by 2.33%, demonstrating its superiority in capturing short-term and long-term interactions between the video and visual query.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
Sequential Fusion	31.77	**11.00**	23.50	44.72	**68.39**
Parallel Fusion (Ours)	**34.10**	**11.00**	**30.17**	**56.17**	66.17

Table 4: Effect of sequential and parallel fusion.

Video Query Fusion We examined the impact of Concat Self-Attention (SA) method, which is popular for multi-modality fusion (Yang et al. 2022; Devlin et al. 2018; Lei, Berg, and Bansal 2021), with Cross-Attention feature fusion for fusing visual query features with video features. As shown in Table 5, cross-attention fusion significantly outperforms Concat SA fusion by 20.63% mtIoU. This improvement is because, in cross-attention fusion, the Key/Value is derived from the visual query while the Query comes from the video. This ensures a substantial contribution from the visual query features, resulting in query-aware fused representations. In contrast, Concat SA involves directly concatenating the visual query features with the video features and performing self-attention on the entire sequence. This approach reduces the contribution of visual query features in the fused representation, as the VQ feature becomes just one element within the sequence.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
Concat Self-Attention	14.47	6.00	8.00	20.00	**34.22**
Parallel Fusion (Ours)	**34.10**	**11.00**	**30.17**	**56.17**	**66.17**

Table 5: Comparing methods for video-visual query fusion.

Class-Specific Tokens vs Generic Embedding In existing works such as (Yang et al. 2022; Jiang, Ramakrishnan, and Grauman 2023), the temporal transformer learns a generic embedding that is shared across classes and corresponds to the number of frames in the video. While this approach may be effective for coarse-grained videos, it results in sub-optimal performance for fine-grained videos. As shown in Table 6, our class-specific embedding outperforms the generic embedding by 3.1% with 96% fewer tokens.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
Generic Embedding	31.00	**13.22**	**27.72**	40.44	65.67
Class-Specific Embedding (Ours)	**34.10**	11.00	**30.17**	**56.17**	**66.17**

Table 6: Generic Embedding vs Class-Specific Token

Tier-Specific Embedding We demonstrate the importance of using scale-specific embedding in the Query-Guided Spatial Transformer. As shown in Table 7, Tier-Specific Embedding improves model performance by 3.52%, highlighting its crucial role in capturing tier-specific features essential for fine-grained video retrieval.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
W/O Tier-Specific Embedding	30.58	**15.50**	25.72	42.67	62.61
W/ Tier-Specific Embedding	**34.10**	11.00	**30.17**	**56.17**	**66.17**

Table 7: Importance of Tier-Specific Embedding

Conclusion

This paper introduces an visual-query based solution for detecting standard anatomy video clips in fetal US videos. Our model MCAT, is a video-based transformer that leverages multi-tier features and class-specific token learning to understand the video with the visual query. This significantly improves video-clip localization compared to models that use single-tier features with 96% less tokens.This enables the model to retrieve the relevant video clip based on a visual query in just 2.69 seconds while using only 4.62 GB of memory during inference, allowing it to run effectively on affordable GPUs, even in resource-limited settings. Additionally, we introduce a temporal uncertainty-aware loss to handle the inherent noise in real-world annotations. Furthermore, to differentiate highly similar classes in fine-grained videos, we propose a contrastive loss that utilizes multi-tier features and multi-anchor guidance to learn subtle class-discriminative features. We apply MCAT to real-world standard plane video-clip detection task with limited data and fine-grained classes and validate its effectiveness through comparisons with SOTA baselines, demonstrating significant improvements in localization accuracy and resource efficiency. These traits make our model beneficial for prenatal care in LMICs, where access to advanced diagnostic tools and skilled health professionals is limited.

Acknowledgments

This work was supported in part by the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1 (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, and the UKRI grant EP/X040186/1 (Turing AI Fellowship).

References

Baumgartner, C. F.; Kamnitsas, K.; Matthew, J.; Smith, S.; Kainz, B.; and Rueckert, D. 2016. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 203–211. Springer.

Cai, Y.; Sharma, H.; Chatelain, P.; and Noble, J. A. 2018. Multi-task sonoeyenet: detection of fetal standardized planes assisted by generated sonographer attention maps. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, 871–879. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Drukker, L.; Sharma, H.; Droste, R.; Alsharid, M.; Chatelain, P.; Noble, J. A.; and Papageorghiou, A. T. 2021. Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. *Scientific Reports*, 11(1): 14109.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6824–6835.

Goyal, R.; Mavroudi, E.; Yang, X.; Sukhbaatar, S.; Sigal, L.; Feiszli, M.; Torresani, L.; and Tran, D. 2023. MINOTAUR: Multi-task Video Grounding From Multimodal Queries. *arXiv preprint arXiv:2302.08063*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

Hara, K.; Kataoka, H.; and Satoh, Y. 2017. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, 3154–3160.

Hernandez-Cruz, N.; Patey, O.; Teng, C.; Papageorghiou, A. T.; and Noble, J. A. 2025. A comprehensive scoping review on machine learning-based fetal echocardiography analysis. *Computers in Biology and Medicine*, 186: 109666.

Jiang, H.; Ramakrishnan, S. K.; and Grauman, K. 2023. Single-Stage Visual Query Localization in Egocentric Videos. *arXiv preprint arXiv:2306.09324*.

Lee, L. H.; Gao, Y.; and Noble, J. A. 2021. Principled ultrasound data augmentation for classification of standard planes. In *International Conference on Information Processing in Medical Imaging*, 729–741. Springer.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2794–2804.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3042–3051.

Men, Q.; Zhao, H.; Drukker, L.; Papageorghiou, A. T.; and Noble, J. A. 2023. Towards Standard Plane Prediction of Fetal Head Ultrasound with Domain Adaption. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Mishra, D.; Saha, P.; Zhao, H.; Patey, O.; Papageorghiou, A. T.; and Noble, J. A. 2024. STAN-LOC: Visual Query-based Video Clip Localization for Fetal Ultrasound Sweep Videos . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15004. Springer Nature Switzerland.

Mishra, D.; Zhao, H.; Saha, P.; Papageorghiou, A. T.; and Noble, J. A. 2023. Dual Conditioned Diffusion Models for Out-of-Distribution Detection: Application to Fetal Ultrasound Videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 216–226. Springer.

Moon, J. H.; Lee, H.; Shin, W.; Kim, Y.-H.; and Choi, E. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12): 6070–6080.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23023–23033.

Rahmatullah, B.; Papageorghiou, A.; and Noble, J. A. 2011. Automated selection of standardized planes from ultrasound volume. In *Machine Learning in Medical Imaging: Second International Workshop, MLMI 2011, Held in Conjunction with MICCAI 2011, Toronto, Canada, September 18, 2011. Proceedings 2*, 35–42. Springer.

Saha, P.; Mishra, D.; Wagner, F.; Kamnitsas, K.; and Noble, J. A. 2024a. Examining Modality Incongruity in Multimodal Federated Learning for Medical Vision and Language-based Disease Detection. *arXiv preprint arXiv:2402.05294*.

Saha, P.; Mishra, D.; Wagner, F.; Kamnitsas, K.; and Noble, J. A. 2024b. FedPIA–Permuting and Integrating Adapters

leveraging Wasserstein Barycenters for Finetuning Foundation Models in Multi-Modal Federated Learning. *arXiv preprint arXiv:2412.14424*.

Saha, P.; Wagner, F.; Mishra, D.; Peng, C.; Thakur, A.; Clifton, D.; Kamnitsas, K.; and Noble, J. A. 2024c. F ̃3 OCUS–Federated Finetuning of Vision-Language Foundation Models with Optimal Client Layer Updating Strategy via Multi-objective Meta-Heuristics. *arXiv preprint arXiv:2411.11912*.

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. 2022. 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.

Schlemper, J.; Oktay, O.; Chen, L.; Matthew, J.; Knight, C.; Kainz, B.; Glocker, B.; and Rueckert, D. 2018. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*.

Scott, T. E.; Jones, J.; Rosenberg, H.; Thomson, A.; Ghandehari, H.; Rosta, N.; Jozkow, K.; Stromer, M.; and Swan, H. 2013. Increasing the detection rate of congenital heart disease during routine obstetric screening using cine loop sweeps. *Journal of Ultrasound in Medicine*, 32(6): 973–979.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Mittal, G.; Sajeev, S.; Yu, Y.; Hall, M.; Boddeti, V. N.; and Chen, M. 2023. ProTeGe: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6575–6585.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16442–16453.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10287–10296.

Zhao, H.; Zheng, Q.; Teng, C.; Yasrab, R.; Drukker, L.; Papageorghiou, A. T.; and Noble, J. A. 2022. Towards unsupervised ultrasound video clinical quality assessment with multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 228–237. Springer.

Zhao, H.; Zheng, Q.; Teng, C.; Yasrab, R.; Drukker, L.; Papageorghiou, A. T.; and Noble, J. A. 2023. Memory-based unsupervised video clinical quality assessment with multi-modality data in fetal ultrasound. *Medical Image Analysis*, 90: 102977.