MMSummary: Multimodal Summary Generation for Fetal Ultrasound Video

Xiaoqing Guo[1][0000−0002−9476−521X], Qianhui Men[1][0000−0002−0059−5484], and J. Alison Noble[1][0000−0002−3060−3772]

Department of Engineering Science, University of Oxford, Oxford, UK
xiaoqing.guo@eng.ox.ac.uk

Abstract. We present the first automated multimodal summary generation system, MMSummary, for medical imaging video, particularly with a focus on fetal ultrasound analysis. Imitating the examination process performed by a human sonographer, MMSummary is designed as a three-stage pipeline, progressing from *keyframe detection* to *keyframe captioning* and finally anatomy *segmentation and measurement*. In the *keyframe detection* stage, an innovative automated workflow is proposed to progressively select a concise set of keyframes, preserving sufficient video information without redundancy. Subsequently, we adapt a large language model to generate meaningful captions for fetal ultrasound keyframes in the *keyframe captioning* stage. If a keyframe is captioned as fetal biometry, the *segmentation and measurement* stage estimates biometric parameters by segmenting the region of interest according to the textual prior. The MMSummary system provides comprehensive summaries for fetal ultrasound examinations and based on reported experiments is estimated to reduce scanning time by approximately 31.5%, thereby suggesting the potential to enhance clinical workflow efficiency.

Keywords: Multimodal summary · Video analysis · Fetal ultrasound

1 Introduction

Ultrasound (US) examinations are essential for monitoring fetal development and maternal health throughout pregnancy. Conducting these examinations requires skill to carefully *manipulate the US probe to locate anatomies, interpret images, and perform biometry* [25]. However, learning to scan well is hard and can take years of training [14]. As a result, there is a worldwide shortage of highly-skilled qualified sonographers, including in low-and-middle-income countries, point-of-care environments, and emergency medical situations [24]. Fetal screening is also time-consuming [26], often taking around 28 minutes per second-trimester examination, as observed in our dataset (left of Fig. 1). Moreover, the redundancy in US videos complicates retrospective analysis and documentation. These challenges necessitate an automated summarizing system capable of efficiently highlighting useful content in examinations and ensuring accurate assessments regardless of expertise, even when examinations need to be conducted quickly.
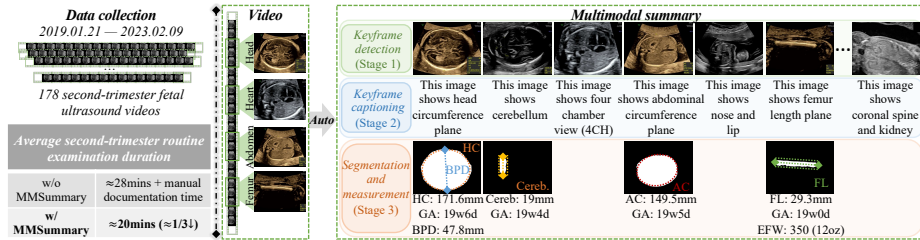
Fig. 1. Multimodal summary for fetal US video to support sonography workflow.

Such an automated summarizing system differs from video summarization in computer vision [2], posing its unique challenges. First, US videos exhibit frame redundancy in appearance, and frames of the same anatomy may have different appearances due to the probe positioning. Hence, it is crucial to detect a subset of representative keyframes rather than potentially redundant video clips as in traditional video summarization. Second, this system should not only identify useful content from the video, but also interpret anatomies and findings within the identified content, and even measure important parameters (right of Fig. 1).

In this paper, we propose MMSummary, a multimodal summary generation system for medical imaging video, with a particular focus on fetal US. Mimicking the examination process, our method comprises a three-stage pipeline, consisting of *keyframe detection*, *keyframe captioning*, and *segmentation and measurement*. In the ***keyframe detection*** stage, a transformer-based network mines temporal information within the video to detect representative frames with anatomical structures of interest. Then a diverse keyframe detection algorithm is designed to eliminate redundant keyframes, forming a concise subset of keyframes. In the ***keyframe captioning*** stage, following ClipCap [20], we leverage pre-trained foundation models, namely BiomedCLIP [33] and GPT2 [23], which possess a broad understanding of biomedical visual and textual data. By only fine-tuning a lightweight mapping network to bridge the gap between visual and textual features, the large language model GPT2 is adapted to generate meaningful and user-friendly captions for detected keyframes. If a keyframe is captioned as fetal biometry, the ***segmentation and measurement*** stage segments a region of interest for automated biometry. Unlike existing fetal anatomy segmentation methods [4, 34, 21], we incorporate prior textual information as prompts to guide anatomy segmentation. By automating the process of US video summarization, we show in our experiments that MMSummary saves around 31.5% scanning time, expedites and simplifies the scan documentation process, and can achieve consistent and accurate assessments independent of operator expertise.

Related Work: (1) Automated ***medical report/summary generation*** has been explored in several medical image analysis applications including for chest X-ray images [28, 31, 35], breast US images [11, 12, 30], and fetal US images [1]. Typically, it is framed as a visual captioning task, where medical image features extracted from convolutional neural networks are input into recurrent
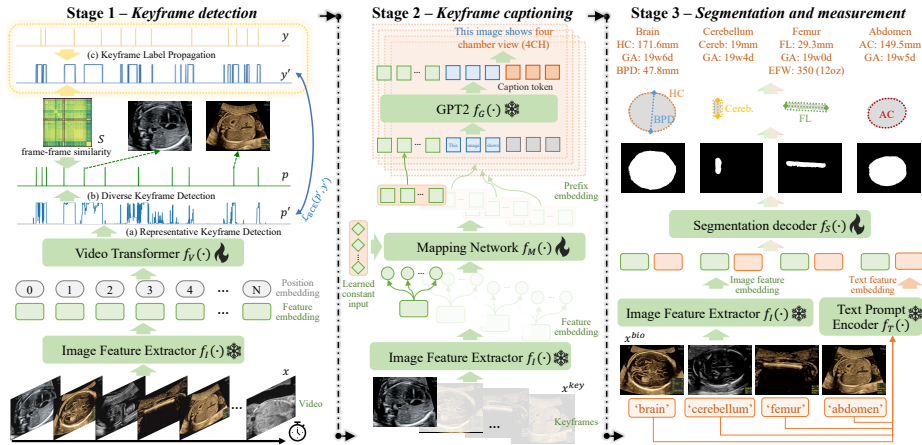
Fig. 2. Illustration of the three-stage MMSummary pipeline.

neural networks to produce word sequences, forming medical reports, and the recurrent neural network is usually trained from scratch [1, 30, 35]. The recent emergence of large language models has led to significant advancement in report generation [12, 28]. These methods are primarily tailored to 2D medical images and text-based summarization, except [27] which provides multimodal summarization of clinical conversation. To the best of our knowledge, we are the first to generate multimodal summaries for medical imaging video. (2) **Video summarization** is extensively studied in computer vision, aiming to generate a concise representation of video content defined as a set of keyframes or clips [2]. Early non-deep learning-based methods focused on selecting keyframes [6, 10]. Recent methods prioritize video clips [8, 15, 16, 29, 32] since the clips incorporate audio and motion for a more natural narrative and enriched summarization. In fetal US video, one prior work employed reinforcement learning to extract video clips [17]. However, frames within clips may be too numerous to include in a multimodal summary. Moreover, and importantly, in fetal US video, keyframes often exhibit a similar anatomical appearance to neighboring frames, leading to frame redundancy. Thus, our approach focuses on extracting keyframes sufficient to summarize the whole examination with minimal frame redundancy. Additionally, unlike prior video summarization methods, our MMSummary generates text descriptions for the extracted keyframes.

2 Method

The proposed three-stage MMSummary pipeline is illustrated in Fig. 2. Given an input video, MMSummary generates a multimodal summarization with keyframes, text descriptions, and biometry parameters. In **keyframe detection** (stage 1), the method takes the video as input $\mathcal{V} = \{x_1, ..., x_T\}$, which spans over T time steps. The keyframe label for the input video is denoted as $y = \{y_1, ..., y_T\}$, where

$y_t = 1$ indicates a keyframe, and otherwise, it is a non-keyframe. Keyframes extracted in stage 1 are separately fed into stage 2 for **keyframe captioning**. Among these keyframes, $\{x_1^{key}, ..., x_k^{key}\}$ are annotated with caption labels $\{y_1^{key}, ..., y_k^{key}\}$. Additionally, some keyframes $\{x_1^{bio}, ..., x_b^{bio}\}$ can be used for biometry estimation and are equipped with region-of-interest mask labels $\{y_1^{bio}, ..., y_b^{bio}\}$. In **segmentation and measurement** (stage 3), a unified model is learned that automatically segments regions of interest, and estimates biometric parameters, gestational age, and fetal weight.

Keyframe detection. This stage detects a concise subset of representative and diverse keyframes that are distinct from each other, summarizing the useful anatomical content in the video. As in Figure 2 (left side), frames of the video are passed through a pre-trained image feature extractor ($f_I(\cdot)$, BiomedCLIP ViT-B [33]) followed by a learnable linear layer, yielding a sequence of feature embeddings. These embeddings are added to vanilla sine-cosine temporal position embeddings, and fed into a 4-layer video transformer $f_V(\cdot)$ (each containing a multi-head self-attention and MLP), outputting the probability of each frame being a keyframe $p' \in [0, 1]^T$. To optimize the learnable linear layer and video transformer, binary cross-entropy loss $\mathcal{L}_{BCE}(p', y) = -y \cdot log(p') - (1 - y) \cdot log(1 - p')$ is computed w.r.t. the keyframe label y.

When minimizing $\mathcal{L}_{BCE}(p', y)$, we encounter extreme class imbalance in our application with the keyframe label y exhibiting a class ratio over 1600:1. Indeed, due to fine-grained probe adjustments, image annotation, and biometry measurements performed by sonographers, keyframe neighbors often share a similar feature representation. Based on this observation, we propose a *keyframe label propagation* algorithm to balance the non-keyframe and keyframe classes in the label distribution. Specifically, we introduce a frame-frame similarity matrix computed via the cosine similarity $S \in [0, 1]^{T \times T}$ between frame features derived from [5] within the video to propagate keyframe labels. If a frame exhibits a high similarity score (above 0.99) compared with any keyframes within the video, it is relabeled as a keyframe, resulting in an updated keyframe label y'. Thus, by minimizing $\mathcal{L}_{BCE}(p', y')$, the model is optimized to *detect representative frames* that preserve as much essential information across the entire video as possible.

To mitigate redundancy in detected frames, we further propose an *diverse keyframe detection* algorithm (*supp.* Fig. 1), which iteratively selects keyframes to ensure exclusiveness in the resulting subset. In each iteration, the frame with the highest probability in the video is selected as the keyframe, and then we mask out other frames that share similar features with the selected keyframe (i.e., have a similarity above τ). This process continues iteratively until the highest probability falls below τ'. Finally, we obtain a concise set of keyframes.

It is worth noting that choosing good values for τ and τ' is crucial (*supp.* Fig. 2). A too large τ cannot thoroughly filter out similar frames, leading to frame redundancy. Conversely, a too small τ may result in a risk of information loss as some representative frames are discarded. As for τ', a too small value may lead to the inclusion of non-informative frames, while choosing a too large value may

result in information loss due to discarding potentially useful frames. Hence, well chosen τ and τ' values can ensure the selected keyframes effectively capture the essential information while minimizing redundancy and information loss.

Keyframe captioning. To automatically generate text descriptions for detected keyframes, we use BiomedCLIP (i.e., image feature extractor $f_I(\cdot)$) [33] and GPT2 [23] ($f_G(\cdot)$), which possess a broad understanding of biomedical visual and textual data, respectively. Then we introduce a lightweight mapping network $f_M(\cdot)$ to bridge the gap between visual and textual features. It takes two inputs, visual feature embeddings and a learned constant input. The constant retrieves meaningful information from visual feature embeddings through multi-head attention, resulting in $m = 50$ prefix embeddings, i.e., $p_1^{emb}, ..., p_m^{emb} = f_M(f_I(x^{key}))$. Additionally, we use the tokenizer of GPT2 to generate a sequence of embeddings for text prompt '*this image shows*', i.e., $t_1^{emb}, t_2^{emb}, t_3^{emb}$, which, together with previous prefix embeddings, serve as input prompts for GPT2.

During the training phase, GPT2 model outputs are constrained with the caption label y^{key}, i.e., a sequence of tokens $y^{key,1}, ..., y^{key,l}$. Our training objective is to predict caption tokens conditioned on both visual and textual prompts in an autoregressive manner. Specifically, we optimize the mapping network using the cross-entropy loss $\mathcal{L}_{CE}^{caption} = -\sum_{i=1}^{l} logp(y^{key,i} \mid p_1^{emb}, ..., p_m^{emb}, t_1^{emb}, ..., t_3^{emb}, y^{key,1}, ..., y^{key,i-1})$. By optimizing the mapping network to refine prefix embeddings, GPT2 is effectively adapted to generate meaningful and user-friendly captions for detected keyframes. In the inference phase, we extract the visual prefix embeddings for detected keyframes from the previous stage and obtain text prompt embeddings as well. Then we start generating a caption conditioned on these prompts by, iteratively, predicting the next token. At each step, the language model outputs probabilities for all vocabulary tokens, and we select the token with the highest probability as the next token in the caption sequence.

Segmentation and measurement. If a keyframe x^{key} is recognized as a biometry image x^{bio}, it is fed into the segmentation model to segment the region of interest for biometric parameter measurement. Given that semantic information embedded in text can enhance medical image segmentation [7], we incorporate prior textual information generated from the *keyframe captioning* stage as prompts to guide segmentation of the corresponding anatomy.

Specifically, input image x^{bio} is fed into the image feature extractor (ViT-B) pre-trained from [18], yielding a feature size of 64×64 with a 256-dimensional embedding vector $f_I(x^{bio})$. Simultaneously, the corresponding caption is input to the text prompt encoder $f_T(\cdot)$ (pre-trained from CLIP [22]) followed by a trainable projection head to generate a 256-dimensional text feature embedding. Then a lightweight mask decoder, comprising two transformer layers and two convolutional layers, enhances the image embedding with text prompt features. The resulting embedding is upsampled and processed by a dynamic mask prediction head to derive the final mask foreground probability p^{bio} at each image

location. The projection heads and mask decoder are optimized by the combination of binary cross-entropy loss and dice loss w.r.t. the mask label y^{bio}.

Specific to our video analysis application, there are four biometry planes: head circumference (HC), abdominal circumference (AC), femur length (FL), and head cerebellum (Cereb). Empirically the head and abdomen exhibit elliptical shapes, while the cerebellum and femur are measured by lines. Therefore, we employ ellipse fitting on the segmented head and abdomen masks using a least-squares-fit algorithm [9]. The perimeter of the fitted ellipse represents HC and AC respectively. For the head, the minor axis of the fitted ellipse is the biparietal diameter (BPD) estimate. For the cerebellum and femur, we fit a minimum external rectangle onto the predicted masks. The long side of the rectangle is used as the cerebellum and femur length estimate, respectively. Lengths measured in pixels are scaled to millimeters using a caliper visible on the left-hand side of the US image [4]. This allows for direct comparison with clinically obtained measurements. Finally, the estimated biometry is converted to gestational age and estimated fetal weight using clinical equations [13, 19].

3   Experiments

Dataset. The dataset utilized in this study is from [anonymized project], with approval from [anonymized Ethics Committee]. Clinical fetal US scans were conducted using a GE Voluson E8 scanner. The videos were captured at 30 Hz and downsampled to 5 Hz. Our study included 178 second-trimester routine examination videos conducted by 3 newly qualified (less than two years of experience) and 1 expert (14 years of experience) sonographers. These videos were randomly split into training (145), validation (11), and test (22) sets. The ground-truth keyframes containing meaningful anatomical structures were obtained from screenshots captured by sonographers during clinical scanning. This resulted in 4596, 329, 679 keyframes in the three sets, of which 3481, 254, 531 frames had caption labels describing anatomies. For biometry estimation, we annotated bounding boxes to identify regions of interest and used MedSAM [18] for mask label inference. Due to the poor quality of generated mask labels for the cerebellum and femur classes, manual reannotation was performed, resulting in 574, 43, 75 mask labels. Ground-truth biometric parameters were manually annotated by sonographers during the scanning process.

Evaluation metrics. In keyframe detection (stage 1), we used four metrics to evaluate the performance of predicted keyframes p compared with the ground truth y, as detailed in Fig. 1 of *supplementary*. 1) *'Cosine simi.'* and 2) *'absolute time err.'* measure keyframe feature similarity and absolute timestamp error in a set-to-set manner. As some predicted keyframes exhibiting large time errors share the same appearance as the ground truth, 3) *'correct time err.'* measures the absolute keyframe timestamp error only when the maximum feature similarity is below a certain threshold (e.g., 0.96). 4) *'Keyframe num. err.'* indicates the absolute error in the number of predicted keyframes. In the keyframe captioning (stage 2), two conventional metrics *'BLEU'* and *'ROUGE-L'* were utilized,

Table 1. Quantitative experimental results for each of the three stages.

	Scan time saved (%)	Cosine simi.(%) ↑	Absolute time err.(s) ↓	Correct time err.(s) ↓	Keyframe num. err.↓
Stage 1	33.25 (35%×95%)	94.12 ±11.29	11.96 ±5.18	9.46 ±4.58	9 ±5
	31.50 (35%×90%)	94.19 ±11.50	10.57 ±4.81	8.50 ±4.53	6 ±4
	26.25 (35%×75%)	96.89 ±3.75	**10.37** ±6.29	7.98 ±5.90	**4** ±3
	17.50 (35%×50%)	96.12 ±2.76	11.36 ±6.55	8.05 ±5.53	**4** ±3
	0 (35%×0%)	**97.02** ±2.74	13.32 ±6.15	**7.66** ±5.31	5 ±3
	0 (NQ/expert)	97.19/96.47	13.61/12.00	8.14/6.17	5/4
	0 (*p*-value)	0.627	0.629	0.505	0.747

	Method	Bleu-1 (%) ↑	Bleu-2 (%) ↑	Bleu-3 (%) ↑	Bleu-4 (%) ↑	ROUGE-L (%) ↑
Stage 2	Alsharid et. al [1]	39.00 ±38.18	19.22 ±34.47	8.94 ±20.44	5.02 ±17.70	44.70 ±39.15
	ClipCap [20]	75.80 ±36.26	68.31 ±40.01	46.33 ±43.87	41.53 ±42.75	80.17 ±34.19
	Ours w/ audio data	**80.77** ±32.96	**74.40** ±37.19	**57.42** ±42.42	**50.75** ±42.99	**85.31** ±30.49
	Ours	79.04 ±33.69	71.65 ±37.66	48.68 ±42.85	45.40 ±41.67	83.05 ±31.04
	Ours (NQ/expert)	78.57/80.31	71.08/73.20	50.87/42.68	45.23/45.86	82.70/83.99
	Ours (*p*-value)	0.626	0.719	0.255	0.941	0.699

	Method	AC(mm) ↓	HC(mm) ↓	BPD(mm) ↓	Cereb(mm) ↓	FL(mm) ↓
Stage 3	Reference (mask label)	4.23 ±3.28 (2.82%)	4.46 ±2.72 (2.58%)	1.02 ±0.67 (2.15%)	0.84 ±0.87 (4.13%)	0.57 ±0.36 (1.83%)
	Bano et. al [4]	7.65 ±5.84 (5.10%)	6.77 ±3.70 (3.93%)	2.74 ±2.45 (5.81%)	3.22 ±2.21 (15.89%)	1.86 ±1.89 (5.97%)
	Nagabotu et. al [21]	5.26 ±3.96 (3.51%)	7.38 ±6.44 (4.28%)	1.97 ±2.23 (4.18%)	3.40 ±3.04 (16.77%)	**0.95** ±1.46 (3.06%)
	Ours w/ rand prompt	6.54 ±4.84 (4.35%)	8.46 ±4.29 (4.90%)	2.42 ±2.35 (5.13%)	1.82 ±1.11 (8.98%)	3.40 ±2.96 (10.94%)
	Ours	**5.23** ±3.81 (3.48%)	**5.25** ±3.44 (3.04%)	**1.25** ±0.99 (2.65%)	**1.05** ±1.10 (5.16%)	2.78 ±2.64 (8.99%)
	Ours (NQ/expert)	5.57/4.50	4.77/6.20	1.27/1.22	1.15/0.86	2.38/3.98
	Ours (*p*-value)	0.972	0.214	0.186	0.227	0.740

following [1]. For the final stage 3, absolute error and relative error of AC, HC, BPD, Cereb, FL biometry w.r.t. the clinical measurement were computed.

Implementation. MMSummary is implemented with PyTorch on RTX 8000. In the pipeline, $f_I(\cdot)$, $f_G(\cdot)$, $f_T(\cdot)$ are frozen, while other parameters are updated with AdamW optimizer and a learning rate of 2e-5. τ and τ' are set as 0.96, 0.8.

Quantitative results. According to the frame-frame similarity matrix S, frames exhibiting a high similarity score above 0.97 compared to any keyframes within the video account for approximately 35% of the entire video. For these frames, we observed that the sonographer was typically performing fine-grained probe adjustments, interpretations, or biometry. By randomly dropping frames during these activities (i.e., from these 35% frames) in various proportions (i.e., 0%, 50%, 75%, 90%, or 95%), we obtained several shortened videos. We then fed these videos into the keyframe detection model learned in **stage 1** for evaluation. The shortened video, with less frame redundancy, was found to simplify the keyframe detection process, leading to the lowest *keyframe num. err.* when dropping 50%, 75% frames, as in Table 1. Moreover, our results demonstrate that keyframe detection maintains a good performance with only a 0.84s increase in *correct time err.*, even for the case of saving 31.5% scanning time. This indicates the potential of MMSummary to significantly reduce the time required for fetal US examinations without compromising keyframe detection accuracy.

In **stage 2**, our method outperforms the fetal US image captioning method [1] and baseline [20] with increments of 40.04%, 3.24% in Bleu-1, respectively. Owing to the flexibility and reasoning capability of the language model, the sonographer's speech transcribed from audio data using [3] can also be incorporated as an additional input prompt to GPT2 for caption reasoning. After fine-tuning GPT2 for 10 epochs, our model achieved 5.35% improvement in Bleu-4.
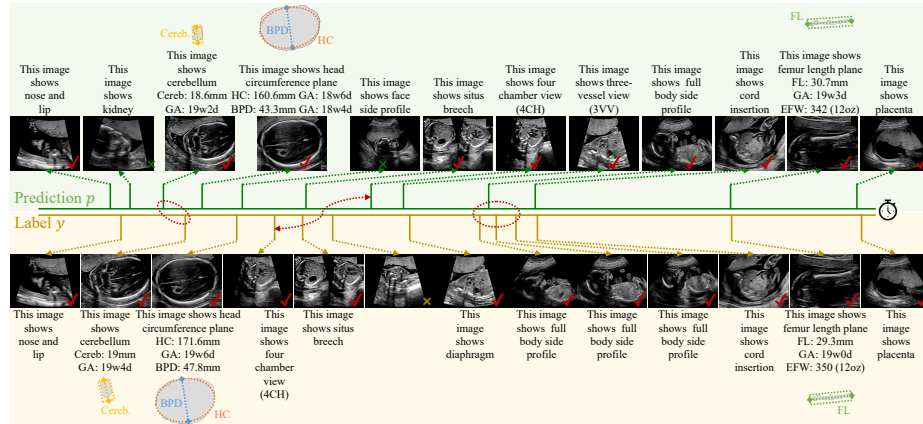
Fig. 3. MMSummary prediction (upper part) and ground truth (lower part).

For **stage 3**, compared with fetal biometry estimation methods [4, 21], our method obtained the best performance in AC, HC, BPD, Cereb biometry. Moreover, we also investigated replacing the textual prompt with a random prompt that lacked any prior information. This negatively impacted the biometry estimation performance, demonstrating the importance of semantic relations embedded in textual prompts to enhance biometry estimation performance.

Finally, we report quantitative results of our method on US videos acquired by sonographers with two skill levels, namely newly qualified (NQ) and expert, shown in gray cells of Table 1. An independent two-tailed t-test was performed to compute p-values. The analysis indicates that there is no significant difference (p-value>0.1) in the performance of our method on NQ and expert videos across all metrics, suggesting that our method is robust to different operator expertise.

Qualitative results.

Fig. 3 (upper part) presents a multimodal summary of an example video generated by our pipeline. Compared with the ground truth (lower part), most keyframes are correctly detected. However, some predicted keyframes exhibit large time errors (left two red dotted indicators). Additionally, our method only detects one 'full body side profile' keyframe (right red dotted circle), while three similar frames are shown in the ground truth. Despite these discrepancies, the accuracy of the multimodal summary is not affected, as the image content of the keyframe is still correct. The predicted keyframes marked with green crosses are redundant or non-informative. In practice, these would be easily removed by a sonographer. The keyframe with a yellow cross is present in the ground truth but not predicted. This may be due to a sonographer incorrectly recording the screenshot since it is of poor image quality. The predicted captions and biometry estimations are reasonably consistent with the ground truth. Captioning errors mainly arise for keyframes with low image quality and in distinguishing cardiac views, which are visually similar in appearance.

4   Conclusion

This paper focuses on the automated generation of a multimodal summary for a given medical video, using a fetal US video as a potential clinical use case. Imitating the real examination process, we propose a three-stage MMSummary system that automatically generates a concise subset of keyframes summarizing the useful video content, accompanied by corresponding text descriptions and biometry estimations, for a fetal US video. This system has the potential to save fetal US scanning time and reduce the expertise needed to scan.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alsharid, M., Sharma, H., Drukker, L., Papageorgiou, A.T., Noble, J.A.: Weakly supervised captioning of ultrasound images. In: MIUA. pp. 187–198 (2022)
2. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video summarization using deep neural networks: A survey. Proceedings of the IEEE **109**(11), 1838–1863 (2021)
3. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio. arXiv preprint arXiv:2303.00747 (2023)
4. Bano, S., Dromey, B., Vasconcelos, F., Napolitano, R., David, A.L., Peebles, D.M., Stoyanov, D.: Autofb: Automating fetal biometry estimation from standard ultrasound planes. In: MICCAI. pp. 228–238 (2021)
5. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE transactions on medical imaging **36**(11), 2204–2215 (2017)
6. De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., de Albuquerque Araújo, A.: Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern recognition letters **32**(1), 56–68 (2011)
7. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. arXiv preprint arXiv:2311.13385 (2023)
8. Dvornik, N., Hadji, I., Zhang, R., Derpanis, K.G., Wildes, R.P., Jepson, A.D.: Stepformer: Self-supervised step discovery and localization in instructional videos. In: CVPR. pp. 18952–18961 (2023)
9. Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. IEEE Transactions on pattern analysis and machine intelligence **21**(5), 476–480 (1999)
10. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: Stimo: Still and moving video storyboard for the web scenario. Multimedia Tools and Applications **46**, 47–69 (2010)

11. Ge, S., Ye, Q., Xie, W., Sun, D., Zhang, H., Zhou, X., Yuan, K.: Ai-assisted method for efficiently generating breast ultrasound screening reports. Current Medical Imaging **19**(2), 149–157 (2023)
12. Huh, J., Park, H.J., Ye, J.C.: Breast ultrasound report generation using langchain. arXiv preprint arXiv:2312.03013 (2023)
13. Kiserud, T., Piaggio, G., Carroli, G., Widmer, M., Carvalho, J., Neerup Jensen, L., et al.: The world health organization fetal growth charts: a multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. PLoS medicine **14**(1), e1002220 (2017)
14. Le Lous, M., Despinoy, F., Klein, M., Fustec, E., Lavoué, V., Jannin, P.: Impact of physician expertise on probe trajectory during obstetric ultrasound: a quantitative approach for skill assessment. Simulation in Healthcare **16**(1), 67–72 (2021)
15. Li, H., Ke, Q., Gong, M., Drummond, T.: Progressive video summarization via multimodal self-supervised learning. In: WACV. pp. 5584–5593 (2023)
16. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: ICCV. pp. 2794–2804 (2023)
17. Liu, T., Meng, Q., Vlontzos, A., Tan, J., Rueckert, D., Kainz, B.: Ultrasound video summarization using deep reinforcement learning. In: MICCAI. pp. 483–492 (2020)
18. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
19. Maraci, M.A., Yaqub, M., Craik, R., Beriwal, S., Self, A., von Dadelszen, P., Papageorghiou, A., Noble, J.A.: Toward point-of-care ultrasound estimation of fetal gestational age from the trans-cerebellar diameter using cnn-based ultrasound image analysis. Journal of Medical Imaging **7**(1), 014501–014501 (2020)
20. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
21. Nagabotu, V., Namburu, A.: Precise segmentation of fetal head in ultrasound images using improved u-net model. ETRI Journal (2023)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
24. Self, A., Chen, Q., Desiraju, B.K., Dhariwal, S., Gleed, A.D., Mishra, D., Thiruvengadam, R., Chandramohan, V., Craik, R., Wilden, E., et al.: Developing clinical artificial intelligence for obstetric ultrasound to improve access in underserved regions: protocol for a computer-assisted low-cost point-of-care ultrasound (calopus) study. JMIR Research Protocols **11**(9), e37374 (2022)
25. Selvathi, D., Chandralekha, R.: Fetal biometric based abnormality detection during prenatal development using deep learning techniques. Multidimensional Systems and Signal Processing pp. 1–15 (2022)
26. Sharma, H., Drukker, L., Chatelain, P., Droste, R., Papageorghiou, A.T., Noble, J.A.: Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. Medical Image Analysis **69**, 101973 (2021)
27. Tiwari, A., Saha, A., Saha, S., Bhattacharyya, P., Dhar, M.: Experience and evidence are the eyes of an excellent summarizer! towards knowledge infused multimodal clinical conversation summarization. In: ACM CIKM. pp. 2452–2461 (2023)

28. Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023)
29. Wu, G., Lin, J., Silva, C.T.: Intentvizor: Towards generic query guided interactive video summarization. In: CVPR. pp. 10503–10512 (2022)
30. Yang, S., Niu, J., Wu, J., Wang, Y., Liu, X., Li, Q.: Automatic ultrasound image report generation with adaptive multimodal attention mechanism. Neurocomputing **427**, 40–49 (2021)
31. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis **86**, 102798 (2023)
32. Zala, A., Cho, J., Kottur, S., Chen, X., Oguz, B., Mehdad, Y., Bansal, M.: Hierarchical video-moment retrieval and step-captioning. In: CVPR. pp. 23056–23065 (2023)
33. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
34. Zhao, L., Tan, G., Pu, B., Wu, Q., Ren, H., Li, K.: Transfsm: Fetal anatomy segmentation and biometric measurement in ultrasound images using a hybrid transformer. IEEE Journal of Biomedical and Health Informatics (2023)
35. Zhou, Y., Huang, L., Zhou, T., Fu, H., Shao, L.: Visual-textual attentive semantic consistency for medical report generation. In: ICCV. pp. 3985–3994 (2021)