Audio-visual modelling in a clinical setting

Jiao, Jianbo; Alsharid, Mohammad; Drukker, Lior; Papageorghiou, Aris T.; Zisserman, Andrew; Noble, J. Alison

[Link to publication on Research at Birmingham portal](#)

scientific reports

Check for updates

OPEN

Audio-visual modelling in a clinical setting

Jianbo Jiao[1,2,6]✉, Mohammad Alsharid[1,3,6], Lior Drukker[4,5], Aris T. Papageorghiou[4], Andrew Zisserman[1] & J. Alison Noble[1]✉

Auditory and visual signals are two primary perception modalities that are usually present together and correlate with each other, not only in natural environments but also in clinical settings. However, audio-visual modelling in the latter case can be more challenging, due to the different sources of audio/video signals and the noise (both signal-level and semantic-level) in auditory signals—usually speech audio. In this study, we consider audio-visual modelling in a clinical setting, providing a solution to learn medical representations that benefit various clinical tasks, without relying on dense supervisory annotations from human experts for the model training. A simple yet effective multi-modal self-supervised learning framework is presented for this purpose. The proposed approach is able to help find standard anatomical planes, predict the focusing position of sonographer's eyes, and localise anatomical regions of interest during ultrasound imaging. Experimental analysis on a large-scale clinical multi-modal ultrasound video dataset show that the proposed novel representation learning method provides good transferable anatomical representations that boost the performance of automated downstream clinical tasks, even outperforming fully-supervised solutions. Being able to learn such medical representations in a self-supervised manner will contribute to several aspects including a better understanding of obstetric imaging, training new sonographers, more effective assistive tools for human experts, and enhancement of the clinical workflow.

As two distinct sensory modalities, the auditory sense and visual sense are widely and commonly used for humans to perceive and interact with the world. Visual signals are often accompanied by auditory signals which provide essential assistance for scene perception and understanding. Such multi-modal scenarios are not only widely exist in natural environments, but also in clinical settings. For instance, in the clinical setting we study in this work, obstetric ultrasound scanning, the routine scans are performed by a sonographer (*i.e.* operators) sitting next to a pregnant woman, controlling the movement of a probe (transducer) on the abdomen to acquire a series of standard imaging planes displayed in real-time on the display screen of the ultrasound machine; in the meantime, the sonographer is describing to the patient the anatomies being visualised (*e.g.* the planes shown and interpretation of the fetus's movement) in a conversational manner. In this case, the visual and auditory signals are the ultrasound videos shown on the screen and the speech audio data, respectively. Our goal is to be able to learn visual capabilities (such as localising anatomies including the *fetal brain*, or detecting standard planes in an ultrasound scan) in clinical scenarios, simply by listening to sonographers describing what they are doing, aiming to assist clinical applications.

Audio-visual modelling has attracted interest in the recent literature[1–3], in which the basic assumption is that the underlying video and audio signals are densely correlated with each other. Based on this assumption, correspondence modelling is leveraged, *i.e.* examining if the input audio and video signals correspond or not. Such a binary discrimination model is simple, but has been shown to be effective in learning audio-visual representations. Some work[4,5] has proposed contrastive learning for audio-visual modelling by referring to negative samples. This helps model discrimination ability and has been shown to result in a better representation quality[5].

Promising progress has been made in the above-mentioned literature for audio-visual modelling in natural scenes. However, in clinical scenarios the setting is different and consequently it is more challenging to model (see Fig. 1a). Different from the natural scenario, in a clinical setting the auditory signal and visual signal usually arise

[1]Department of Engineering Science, University of Oxford, Oxford, UK. [2]School of Computer Science, University of Birmingham, Birmingham, UK. [3]Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates. [4]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK. [5]Rabin Medical Center, Tel-Aviv University Faculty of Medicine, Tel Aviv, Israel. [6]These authors contributed equally: Jianbo Jiao and Mohammad Alsharid. ✉email: jianbo.jiao@eng.ox.ac.uk; alison.noble@eng.ox.ac.uk
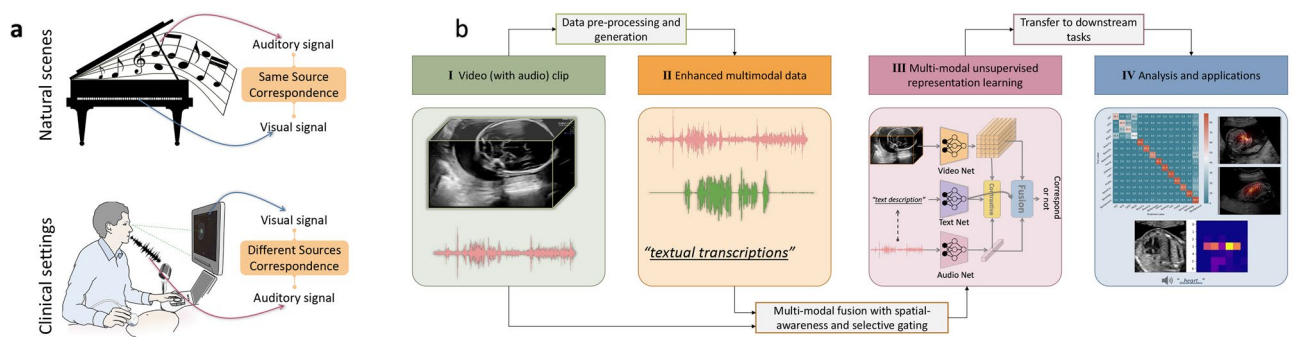
Figure 1. Multi-modal representation learning in clinical setting. (**a**) Illustration of audio-visual modelling in a natural scene setting (top) and in a clinical setting (bottom). (**b**) Pipeline for audio-video modelling and analysis in clinical settings from raw video footage. Video frames with the corresponding speech audio are extracted from the raw footage (I). After audio data pre-processing (the illustrated red waveform to the green one) and text data generation from the audio signal, the enhanced multi-modal data (II) are fed into a joint fusion framework (each data modality is encoded via a network to the corresponding features, more detail please refer to Fig. 2) to learn multi-modal representations without human annotation (III). The whole system can be transferred to several downstream tasks and used for large-scale analysis and support human experts (IV).

from different sources, and have sparse correlation, *e.g.* the conversational speech audio and the corresponding captured medical images/videos. This prevents the direct application of existing methods that were designed for natural scenes. Similar to natural scenes, the modelling of audio-visual signals is helpful - in this case for learning medical representations without relying on supervisory training signals acquired from human expert manual annotations, which are costly and hard to scale. During real-time ultrasound (US) scans, narrations from clinicians describe the visual signals and as a result they have a strong underlying correspondence with the anatomy seen on the screen of the ultrasound machine. Our hypothesis is that modelling such correlations in a self-supervised manner will be beneficial for automating downstream clinical tasks, which is the aim of this work.

To this end, in this study, we consider audio-visual modelling in a clinical setting and propose a simple framework (Fig. 1b) to address the challenges faced in doing this. Here, the *clinical setting* we work with is fetal US scans, where audio recordings containing the *speech* uttered by the sonographers during the scan are considered, together with the US video shown on the screen. We can do this because sonographers speak about what they see on the screen during an examination, as an explanation to the pregnant individual being scanned. More specifically, we start with modelling the direct correspondence between the speech audio and its corresponding video signal. Cross-modal contrastive learning is then introduced to strengthen the discrimination between the current sample and unrelated negative samples. Clinical audio recordings are noisy (*e.g.* background noise and unrelated sounds and conversations). Therefore, we propose to pre-process the audio data to enhance the related speech content before feeding it to the deep model. To enrich the semantic understanding of the audio data and bootstrap the audio-visual modelling, an additional textual modality with a text branch is introduced to the proposed framework, forming a multi-modal modelling framework. The text is automatically transcribed from the speech audio signal, and we propose a learnable gating module towards the end of this text branch to spot the text embeddings of the relevant anatomies. Note that in our clinical ultrasound setting, the originally captured data are the fetal *ultrasound video* from the ultrasound machine and the *speech audio* from the sonographers. The text information is not available and accurate transcription is infeasible to achieve. Thus, our work aims at audio-visual modelling and takes textual data as intermediate information to help the modelling.

In addition to learning self-supervised cross-modal representations, we demonstrate audio-guided visual anatomy localisation by directly leveraging the learned representations. This could be used in various clinical applications, *e.g.* better user experience for patients during scanning, more effectively training new sonographers and low-cost data annotation for region-of-interest by only using audio, to name a few. An example demo video was shown in the "supplementary information". Note that during the inference stage (*i.e.* in real applications), no text information is needed and only the audio signal is sufficient to help the visual localisation. To show how the designed approach could benefit human experts and clinical applications, we perform extensive experimental analysis on a large-scale multi-modal clinical US dataset to validate its effectiveness by transferring the learned representations to downstream clinical tasks. Analysis of three downstream clinical tasks shows that the proposed approach learns good transferable representations, which surpass the performance of fully-supervised solutions. Instead of replacing humans, we aim more towards helping and boosting the capability of human experts, *e.g.* alleviate workload, improve efficiency, and help ultrasound imaging understanding. To summarise, the main contributions and aims of this work are as follows:

- We propose, to our knowledge, a unique audio-visual modelling framework (with generated text as an additional modality) for representation learning, in a clinical setting where the correlation modelling for audio-visual data is more challenging than non-clinical natural scenarios.
- We introduce several strategies to address the challenges in clinical settings, including audio pre-processing, selective gating module for text embedding refinement and multi-modal joint optimisation.
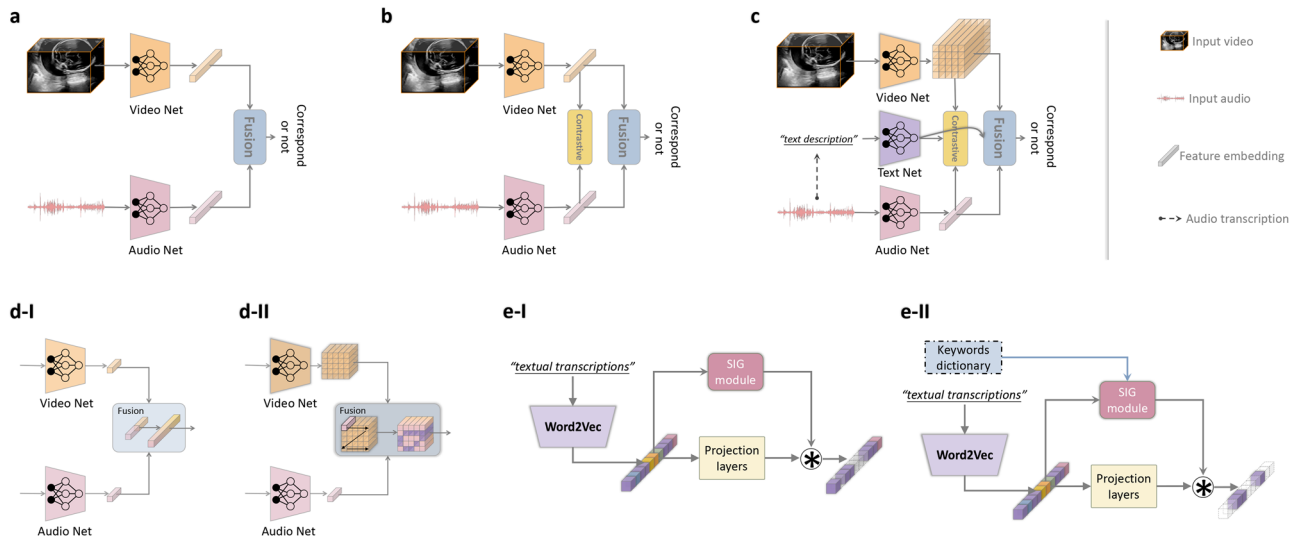
Figure 2. Audio-video modelling architectures. (**a**) Baseline model of correspondence modelling. (**b**) Correspondence modelling with contrastive learning. (**c**) The proposed multi-modal modelling approach with additional textual information and visual-spatial awareness. (**d-I**) Basic fusion strategy by concatenating multi-modal features. (**d-II**) Spatial-aware fusion by element-wise product. (**e-I**) The proposed selective information gating (SIG) module, with the coarse level of only filtering outliers. (**e-II**) The SIG module with fine-grained level of keywords spotting by only focusing on specific keywords of interest. The ⊛ means element-wise product.

- We present experimental analysis to validates the effectiveness of the proposed framework, with better performance than alternative solutions and superior to fully-supervised settings. We also introduce an approach to visually localise anatomical regions of interest according to audio guidance, and discuss the impact and potential of the study.

Methods

We propose a framework to model the audio-visual representations by leveraging intermediate text information. The proposed approach has three original components: audio concentration for audio data processing, a selective gating module for embedding refinement, and a multi-modal correlation optimisation loss. Figure 2 illustrates the framework of audio-video modelling with the main ideas of the proposed method shown in Fig. 2c.

Large-scale multi-modal clinical dataset

The proposed approach is validated on a large-scale multi-modal clinical US dataset called the PULSE (Perception Ultrasound by Learning Sonographer Experience) dataset which is described in[6]. Ethics approval was granted by the West of Scotland Research Ethics Service, UK Research Ethics Committee (Reference 18/WS/0051). Informed consent was obtained from all participants, and all methods were performed in accordance with the relevant guidelines and regulations. Briefly, the PULSE dataset consists of full-length US video recorded with concurrent speech audio, eye-gaze tracking data and probe motion data, as sonographers perform routine scans on pregnant women. A typical scan setting is shown in Fig. 1a bottom. In this work, we focus on modelling the video and audio data for the second-trimester data (*i.e.* the 20-week anomaly scan). The video data was frame-grabbed from the US machine in real-time to a computer in a fully anonymised fashion. The original full-length video was stored at 30 FPS using lossless compression. The speech audio data was recorded in synchronisation with the video data by using microphones placed in the scanning room.

Multi-modal medical representation learning

Baselines

We first present a baseline to model the correlation between the US video and audio data modalities. Specifically, the most intuitive approach is to predict if the given video and audio samples correspond or not, as usually done in natural scenes. To this end, we use a binary classification framework to model audio-visual correspondence. As illustrated in Fig. 2a, the US video clip and audio segment are input to the video and audio sub-networks to obtain corresponding feature embeddings, followed by a fusion operation to produce the final feature for correspondence prediction. The original recorded video with the accompanying synchronised audio signals are assumed to correspond with each other, while shifted video-audio data pairs or pairs from different scans are considered to be non-corresponding. Let the video signal be V and the audio signal A. The correspondence modelling task can be simply modelled as a binary classification problem:

$$L_{base}(V_i, A_i) = -\frac{1}{N}\sum_{i=1}^{N} c_i \cdot \log\left(p(A_i, V_i)\right) + (1 - c_i) \cdot \log\left(1 - p(A_i, V_i)\right), \tag{1}$$

where c is the correspondence label (0 or 1) and p is the predicted probability of the audio-video pair being in correspondence. N is the number of samples. Note here the label c does not need any human annotations, but instead naturally encoded in the data. Specifically, when the audio and video signals are aligned with each other (by default naturally as they are in the dataset), $c = 1$, otherwise, $c = 0$ if they have not been matched with the signal that they originally correspond to in the dataset.

Cross-modal contrastive learning
The above baseline model only describes the alignment of the audio and video data, being a very strict constraint that may not be sufficient to model the relationship between these two modalities due to the challenges in the clinical setting. We observe that in our speech audio data, there are some medical-unrelated contents (*e.g.* random talk), which deteriorate the correlation learning between the two modalities. Contrastive learning aims to pull the similar data pairs closer while pushing dissimilar ones away from each other, in the embedding space during optimisation. In addition to modelling the relationship of the positive samples that are aligned at the same timestamp, it also considers dissimilar sample pairs, which act as negative samples for the modelling. As a result, built upon the baseline model, cross-modal contrastive learning is introduced to model the audio-video representation, and encourage a soft constraint towards the audio-visual relationship modelling. Specifically, consider a positive pair of audio-video data as (A_i, V_i), and other sample pairs (A_i, V_j) that are misaligned (*i.e.* $i \neq j$) as the negative pairs. Then the loss function for contrastive learning L_{contra} is defined by:

$$L_{contra}(V_i, A_i) = -\log \frac{\text{sim}(V_i, A_i)}{\sum_{j=1}^{N} \mathbb{1}_{[j \neq i]} \text{sim}(V_j, A_i)}, \tag{2}$$

where $\mathbb{1}_{[j \neq i]}$ equals 1 iff $j \neq i$ and otherwise 0. The sim() function is a similarity measurement over the exponential of the arguments. Though intra-modal contrastive learning may be able to boost the representation learning ability if further included, here we focus on cross-modal contrastive learning. And we found in a recent work[7] that it may lead to some issues in clinical settings, where additional anatomy annotations may be required. As a result, in this paper we focus on the cross-modal learning approach.

Spatial awareness
Although contrastive learning provides an adaptive correlation modelling solution, the conventional similarity measure, *i.e.* a one-dimensional feature vector (as shown in Fig. 2a,b), ignores spatial context in the feature embedding and loses local spatial awareness. To address this, instead of compressing the output bottleneck feature of the video network to be a 1-D vector, we preserve the spatial dimensions of the feature before fusing it with the audio feature (Fig. 2c). Specifically, the spatial dimensions are achieved by keeping the spatial feature map from the second last layer in the video network instead of the 1-D feature vector after the last pooling layer as in the above settings. In this way, the merged features would account for not only global information, but also local structural patterns (anatomies). Here, the audio-video fusion is achieved by element-wise dot product between the audio feature vector and each of the visual feature vectors in the visual feature map, followed by global average pooling and fully-connected layers in an attention-like aggregation manner into a single vector before being fed into the loss functions. The element-wise dot product is used to measure the similarity between the audio feature and the visual features, aiming to activate the high-response regions on the spatial visual feature map. As illustrated in Fig. 2d, by using the spatial-aware fusion strategy, the multi-modal features are able to capture local spatial information instead of global semantics only (as in Fig. 2d-I).

In this case, the final fused feature is a three-dimensional tensor (instead of one-dimensional vector) with spatial information that represents the spatial relationship between the visual and audio signals. The corresponding loss function used for training is then defined as:

$$L_{base}(V_i^s, A_i) + L_{contra}(V_i^s, A_i), \tag{3}$$

where V_i^s is the 3-D visual feature with spatial information.

Adding textual information
The speech audio data modality provides auxiliary supportive information to assist in video understanding and representation learning, but according to our observations, in our application, it can also bring ambiguity in correlation modelling. Specifically, the audio signal is not generated from the visual source and as a result, may have a weak (or even no) correlation with video content. For example, the content of the audio is often a conversation between the sonographer and the subject while the video being shown on the screen is that of fetal US content. There are even different kinds of background sounds that can exist in the audio data (we have observed ambulance sirens, people coming into the scanning room, etc). These challenges prevent the model from learning accurate and robust multi-modal representations. To this end, we propose to leverage additional information along with the audio signal, by including a third data modality – text. Here, the text modality refers to the transcriptions from the speech audio data, *i.e.* a textual representation of the speech content. The textual transcriptions are acquired by applying an automatic speech recognition (ASR) model to the speech audio data. ASR is used here as most of the meaningful content in the speech audio is conversational natural language speech data. In this way, we can use the textual feature embedding to help the multi-modal US representation learning, and this also allows us to focus on key information (*e.g.* anatomy-related terminology) while eliminating unrelated information by keyword spotting. The basic idea is illustrated in Fig. 2c. Specifically, we use the word2vec model[8] followed by two fully-connected layers to project textual information into word embeddings. The main role of the text modality aims to provide a more focused view on the content within the speech audio

and highlight those US-related context information, so that strengthen the representation learning. More detailed validation of the effectiveness of the text modality is presented in the Results section below.

However, due to the poor performance of the ASR model when applying to unseen US data, the ASR transcribed text cannot be directly leveraged as input for the model training. Apart from the existing challenges in audio, the ASR transcription has many errors (*e.g.* non-meaningful English), which may further confuse the model during the joint representation learning. Therefore, we propose approaches to address these issues (*i.e.* noise in audio, ambiguity in transcribed text, *etc.*), as will be elaborated on in the following sections.

Selective information gating

As mentioned above, the transcribed raw textual data is inappropriate to be directly fed into the text embedding model. In this section, we introduce a new selective information gating (SIG) module to filter out the undesired information and only preserve the key information that we are interested in in a selective manner. The proposed SIG module is designed in two variants: outlier filtering and keywords spotting, as illustrated in Fig. 2e and detailed below.

Filtering outliers

The first SIG approach is at a relatively coarse level, where only outliers are removed. Here we consider those wrongly transcribed words *e.g.* invalid English words and blank text segments (zero padding) as outliers. Whenever such outliers are detected in the input text, the corresponding embeddings are removed and thus do not contribute to the final correlation modelling. The wrongly transcribed words are effectively identified before being fed into the SIG module, and the module simply filters them out to remove their possible influence on model training. These words have been identified through the tokenizer used with the word2vec embedding layer, and such 'unknown words' that do not exist in the word2vec dictionary would be represented with a ⟨unk⟩ token. Then it is the word2vec embedding of this token that gets filtered out by the SIG module. More specifically, the SIG module consists of a convolutional layer with 1×1 kernel size followed by a fully-connected layer, which outputs a weighting vector to combine with the word embeddings, as shown in Fig. 2e-I. The projection layers are fully-connected layers to project the word2vec embeddings to combine with the SIG module output. The final filtered text embedding is used to optimise the multi-modal learning objective.

Keywords spotting

Even when the ASR model is able to correctly recognise all the remaining speech from the SIG output, there can be some content that is not related to the video. For instance in some scans, another medical professional enters the room and has a conversation with the sonographer that is unrelated to the US scan (*e.g.* conversations like 'how are you?', 'did you go to the room...'). In this case, modelling the text (as well as the audio) with the video would confuse the model. To reduce this potential source or error, we introduce a *keywords dictionary* to the SIG module, so that the SIG module is aware of the keywords and select the corresponding information (Fig. 2e-II). Specifically, the SIG module only focuses on the keywords within the dictionary when predicting the weights for the gating operation, in a compare and matching manner by filtering out and only selecting those words that exist in the dictionary while depressing the features of other words. The keywords in the dictionary are constructed empirically with domain knowledge from the clinicians. Keywords spotting, as a result, highlights the US-related words/content and encourages the model to focus more on learning high-quality US representations. The effectiveness of the SIG module and the keyword spotting is validated in the ablation study below.

Multi-modal optimisation

To optimise the multi-modal joint representation learning, we propose to use a combined loss function that consists of correspondence modelling and contrastive learning for the video-audio-text modalities. More specifically, according to the definitions in Eqs. (1) and (2), the final joint loss function for multi-modal optimisation is defined as

$$\mathscr{L} = \alpha \cdot L_{base}(V_i, A_i) + \beta \cdot L_{base}(V_i, T_i) + \gamma \cdot L_{contra}(V_i, A_i) + \delta \cdot L_{contra}(V_i, T_i), \qquad (4)$$

where T_i is the text sample and $\{\alpha, \beta, \gamma, \delta \mid \alpha + \beta + \gamma + \delta = 1\}$ are weighting parameters to balance the contribution of each loss term.

Results

In this section, we first describe the detailed experimental settings including the data and evaluation strategies. We will also compare with other alternative or state-of-the-art solutions, with isolated ablation studies to validate the effectiveness of the proposed approach.

Data

Full-length routine fetal second-trimester US videos were acquired by expert sonographers as part of the PULSE study. As there is no public dataset available with the same clinical setting as ours to our knowledge, here the experiments are performed on our PULSE data. Sonographer audio recordings that include audible vocal descriptions of the performed scans were collected with the videos, creating an audio-visual dataset. In total, there are 81 scans with such audio and video data and each scan contains around 55,000 frames. Each whole scan was divided into smaller video clips to feed into the networks, resulting in a total of 73,681 clips. The speech audio data was transcribed via ASR, allowing for the inclusion of the third modality, text. As mentioned above, the text was further processed by filtering unrelated words and focusing on the keywords for the text embedding extraction.

The keywords dictionary consists of anatomical terminologies within the fetal US scenario. When sampling the positive and negative training pair, for each video sample at a particular timestamp t, the corresponding audio and text are sampled in a neighbour region with the time range of 0.6 s (*i.e.* $t \pm 0.3$ s of the video sample at t). Before feeding into the audio network, the original audio data was converted to the 2-dimensional log-spectrogram format with size 256×256 using a short-time Fourier transform with frequency bands of 256, window length 10ms and hop length of 5ms. For the aforementioned downstream tasks, we used an additional set of 135 scans and conducted three-fold cross-validation (by randomly selecting 90 scans for training in each fold).

Audio data cleaning

As mentioned earlier, the original audio data contains noise and unrelated signals in addition to the relevant speech content. Therefore, the signal to noise ratio is very low, and poor performance (more details please refer to the "supplementary information") was observed by directly working with such audio data. To this end, an audio cleaning step was implemented as illustrated in Fig. 3a. The raw audio data contains unwanted background signals (including the sound of US machine keyboard actions and the room air conditioner fan among other background sounds) as well as irrelevant, undesired speech. The first step to reduce undesired audio is to apply audio denoising followed by voice activity detection, in which the former is achieved by speech enhancement[9] and the latter by applying VADNet[10]. That makes it possible to differentiate speech segments from noise segments as well as seconds of silence.

After background noise reduction, we further separate the audio signal into different speakers using speaker diarization. Speaker diarization is a speech processing technique that aims to separate an audio recording into two (or more) speech segments where each segment contains a single unique speaker. Specifically, we employ the pyannote-audio[11] framework to perform speaker diarization and assume that the speaker with the most speech segments is the sonographer. By only preserving the sonographer's speech and disposing of the speech of other people, we have a "clean" speech dataset suitable for analysis. Examples of the audio data before the pre-processing and after the pre-processing are presented in Fig. 3b for reference.

Finally, transcribing the sonographer speech segments with a fine-tuned wav2vec-based automatic speech recognition (ASR) model[12], we identify the speech segments where anatomically relevant information is being spoken by comparing the transcribed text with an existing dictionary of anatomically relevant keywords.

Assessment and evaluation criteria

After learning the multi-modal representations, we need an assessment tool to evaluate the quality of the learned representations and their applications. Following prior works[1,13], we use several downstream tasks to assess the quality of these representations, in a transfer learning manner. Here, in our fetal US scan setting, three downstream tasks are chosen: standard plane detection, eye-gaze saliency prediction, and audio-guided visual localisation. Through these three tasks, we not only evaluate the learned visual representations (as in prior works) but also evaluate the quality of the joint representation (by the third task). The standard plane detection is a classification problem, where given an US image, the task aims to classify it into an anatomical category. The eye-gaze saliency prediction is a regression task that tries to predict the eye-gaze attention regions that the sonographer is focusing on. The visual localisation task aims to localise the anatomy regions in the US video according to the indication from the corresponding audio data. The first two downstream tasks focus more on the visual representations, while the third one needs the model to have a thorough multi-modal understanding so that it can successfully localise the audio-guided visual anatomies. For quantitative assessment, following prior works[13–17] on this task, we leverage the *precision, recall* and *F1* score to evaluate the standard plane detection task; and for the saliency prediction task, the Kullback-Leibler divergence (*KL*), normalised scanpath saliency (*NSS*), area under ROC curve (*AUC*), Pearson's correlation coefficient (*CC*) and similarity (*SIM*)[18] are used following previous studies when evaluating saliency prediction[13,15–17,19,20]; as there is no localisation annotations available in our dataset and it is hard to acquire accurate pixel-level annotations, it is infeasible to present a quantitative evaluation on the third task. Therefore, we consider eye-gaze saliency to be a close enough approximation to expert-made annotations as the eye gaze data comes from the experts themselves. Eye-gaze saliency has been used in other work[21,22] as well in an effort to tie language with visual regions of interest. We provide a corresponding quantitative evaluation in this case.
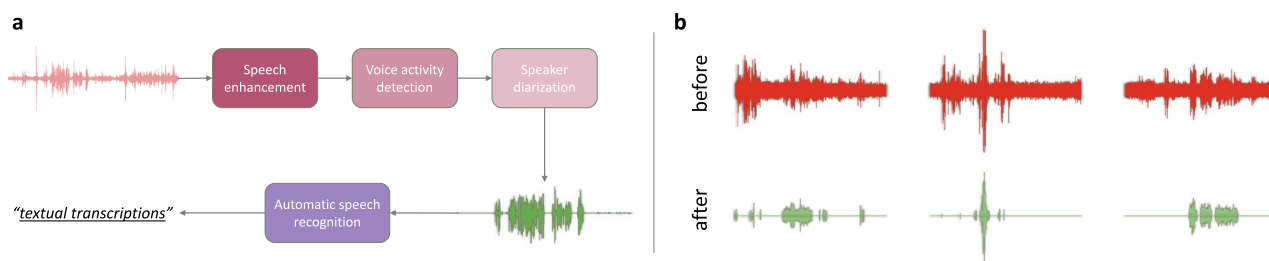

Figure 3. Audio data cleaning. (**a**) Pipeline for the proposed audio data cleaning. The input audio signal (the left red waveform) results in an enhanced audio signal (the bottom-right green waveform) from the sonographer accompanied by a corresponding text segment (the bottom-left text). (**b**) Example audio data samples *before* (top) and *after* (bottom) the proposed audio data pre-processing.

We also found that the saliency prediction in the clinical setting is different from that in natural vision scenarios[18], as we focus on more fine-grained anatomical local structures in medical imaging. As a result, in addition to the aforementioned metrics, we further design a new metric tailored for eye-gaze saliency assessment in the clinical setting. This metric is designed to model the compactness of the predicted saliency region, as well as the intensity of the highly-confident response. Specifically, assuming the area of the whole saliency regions to be $\alpha_S = \|I_S\|_0$ (I_S is the normalised saliency map) and the highly-confident response region as $\alpha_H = \|I_{HS}\|_0$ where $I_{HS} = I_S > thres$, the high response intensity is defined as $\eta_H = \alpha_H/\alpha_S$. Additionally, we also consider the correctness of the predicted saliency when measuring its compactness, by using the intersection-over-union (IoU) metric (IoU=intersection($I_S, \hat{I}_S$)/union($I_S, \hat{I}_S$)). The proposed compactness metric is then defined as

$$Comp = \text{IoU} \cdot \frac{\eta_H}{\alpha_S}. \tag{5}$$

Implementation details

We follow our preliminary work[17] for the baseline network design and the parameter settings. The ResNeXt50[23] was used as the backbone with squeeze-and-excitation module and dilated convolutions for the video branch. To take images/frames as input for the downstream tasks, following[17], within an interval of a video clip, we randomly sample a few frames (2 in this work due to the memory limitation) to represent the video clip, and each of them is fed into the encoder for the feature extraction. The corresponding features are concatenated subsequently as the video feature. Following[17], the audio branch shares the same network architecture as the video branch but is optimised separately and with different inputs. The speech audio was extracted with a 0.6s interval and resampled to 24kHz before being converted to a 2D log-spectrogram representation (of size 256×256) using a short-time Fourier transform. Then this log-spectrogram is fed into the audio net for the following representation learning. The text branch and audio branch are designed in a similar manner except that we used the word2vec model[8] for text embedding extraction, while the remaining joint modelling with the video embedding is performed in the same way. We also tried a more advanced text embedding model (BERT[24]) in replacement of the word2vec model, but did not observe any noticeable improvement. We suspect the reason lies in the relatively simpler context of our text data, while the BERT model was designed to address the ambiguous context situation where the same word can have more than one meaning depending on the context. Most of our words are high-level anatomical (*e.g.* 'bones', 'heart', 'arms', 'legs', 'blood', 'kidneys', and 'elbow'), and are adequately covered by a general corpus. BioBERT is an alternative to BERT that has been trained on biomedical corpora. BioBERT has been shown to work well for specialist biomedical terms such as 'antimicrobial' and 'transcriptional'[25]. However, despite the medical context in which our data is acquired, our vocabulary is more conversational because it is spoken between a clinical professional and the patient. It did not seem necessary to use such a specialised embedding model. In terms of optimisation, we use a similar combined objective as in the aforementioned audio-video modelling, *i.e.* correspondence modelling with contrastive learning. The weighting parameters in the final loss (Eq. (4)) are empirically set to be equal after normalising to the same scale. We also did an analysis on the weighting parameters by setting different values to each of the parameters, and found that the model performs the best when they are set to be equal. The whole model was trained end-to-end with the stochastic gradient descent (SGD) optimiser and the learning rate was initialised as 10^{-3} and decayed by the scale of 10 for every 20 epochs. The batch size was set to 40. The visual signal was scaled to 256×256 and centre cropped to 224×224 to feed into the GPU. We implemented with the PyTorch deep learning framework on a workstation equipped with NVIDIA GPU cards (Titan V and Tesla V100). The implementation code will be made publicly available.

Representation quality assessment

Standard plane detection

As mentioned above, the quality of the learned multi-modal representations is assessed by transfer learning to downstream tasks. First we perform evaluation on the standard plane detection (SPD) task. The 14 categories used in this task are: three-vessel tracheal view of the heart (3VT), four-chamber view of the heart (4CH), left ventricular outflow tract of the heart (LVOT), right ventricular outflow tract of the heart (RVOT), transventricular plane of the brain (BrainTv.), transcerebellar plane of the brain (BrainCb.), abdomen, femur, kidneys, lips, profile, spine in the coronal plane (SpineCor.), spine in the sagittal plane (SpineSag.) and background. The performance of this classification task is reported in Table 1, where we include the results of using random initialisation, only video modality, video-audio baseline[17], video-text baseline and our multi-modal approach (*Ours*) as initialisation for the SPD model. Here the *Video* refers to the self-supervised representation learning approach (with a frame

	Rand.Init.	Video	V-A Base	V-T Base	Ours	ImageNet Init.	US-Sup
Precision	70.4±2.3	71.9±2.0	72.7±1.8	75.6±1.4	**77.2**±1.4	74.6±1.8	87.1±1.9
Recall	64.9±1.6	71.7±3.5	73.3±2.4	75.4±1.6	**77.3**±1.8	71.2±1.9	81.0±2.7
F1-score	67.0±1.3	71.5±2.4	72.6±1.7	75.1±0.2	**76.8**±1.0	72.5±1.8	82.7±1.6

Table 1. Evaluation results on standard plane detection (mean ± std.[%]). Best performance is marked in bold. Note the methods on the right side are fully-supervised using external manual annotations. All the three metrics are the higher the better.

order prediction pretext task) using only video data for pre-training; and the *V-A Base* refers to the video-audio self-supervised representation learning approach proposed in[17]. The *V-T Base* is the baseline using only the video and text branch (without the proposed keywords spotting SIG module) as shown in Fig. 2c. Additionally, we also report the results of two fully-supervised models which were trained with manual annotated ground-truth labels: *ImageNet Init.* for using the ImageNet[26] dataset as training data, and *US-Sup* for using US data with expert annotations as supervision for the same standard plane detection task during pre-training.

Note that all these different methods (*i.e.* columns in Table 1) are just different initialisations for the model weights before the training on SPD task. Then the models are trained (fine-tuned) with the corresponding labels for standard planes. The results were shown to the human experts for validation and an agreement was reached. From the quantitative results shown in Table 1, we can see that by using the additional audio data, the performance of SPD can be improved compared to using video only. When using the proposed multi-modal approach, the performance is further improved by a large margin (*e.g.* 72.7→77.2 for the precision). It can be seen that when only using the video and text branches, the model also performs quite well, suggesting the effectiveness of the data processing and text-related modules as well as the joint modelling. It is interesting to see that even though the *ImageNet Initialisation* was trained with human annotations in a fully-supervised manner, the proposed approach still performs much better. As *US-Sup* was pre-trained with the exact same task of SPD and with human annotations, it can be considered as an upper bound. To better understand how the representation learning methods perform in more detail, we also present the confusion matrix over all the 14 categories, as shown in Fig. 4. From the confusion matrix, it can be seen that the cardiac-related categories (*i.e. 3VT, 4CH, LVOT, RVOT*) are relatively poorly performing; it is well known that these are hard even for a human expert. However, when we compare the *Video-Audio Baseline* with the *Video only*, the performance of these categories is greatly improved, suggesting the effectiveness of the auxiliary audio modality. Building upon that, the proposed multi-modal spatial aware approach further improved almost all the categories significantly. For instance, *LVOT*: 46.7→54.5 and *Kidneys*: 79.1→92.9. This fine-grained analysis further validates the effectiveness of the proposed method.
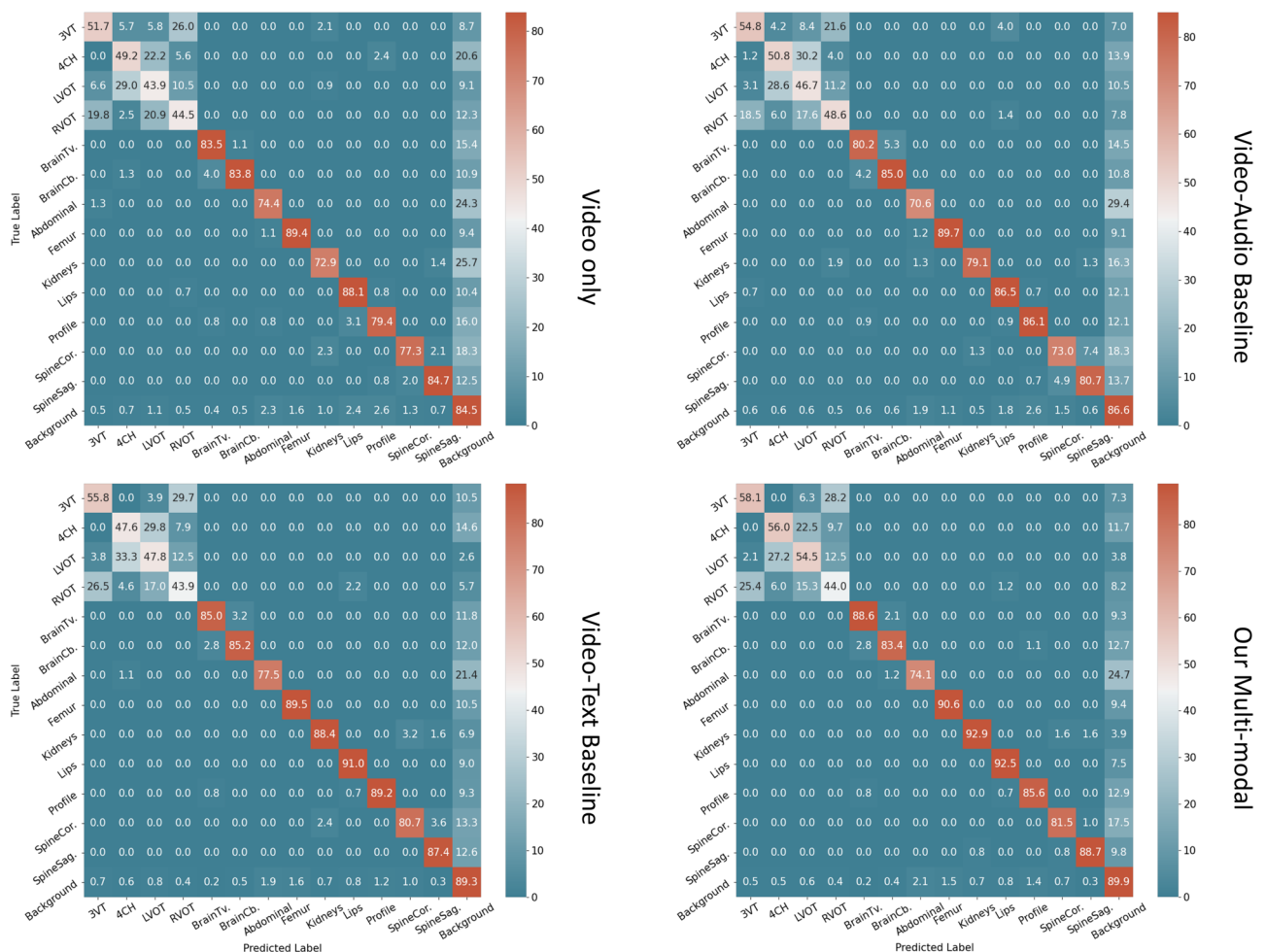

Figure 4. Confusion matrices of the representation learning methods on the SPD task.

Eye-gaze saliency prediction

As in our PULSE dataset, the eye-gaze tracking information of the sonographers was captured while performing US, in addition to the above SPD task; this allows us to take the eye-gaze saliency prediction as an additional downstream task to evaluate the quality of the learned representations. Similar to the SPD task, the pre-trained model was loaded to the saliency prediction model followed by fine-tuning. The only difference is that the last layers were modified so that the model is able to predict a spatial saliency map instead of a single category. The quantitative performance is reported in Fig. 5, with the evaluation metrics mentioned above. Note this task aims to predict the real human gaze (last column in Fig. 5c) from a given US image, including both the location and focus. Again, we can see that with the addition of the audio modality, the model performs much better when compared to using only video data. The textual information also helps the representation learning, with similar performance as the video-audio baseline and outperforming the single-modal approach. When jointly trained with video, audio and text, the proposed multi-modal approach further boosts the performance by a large margin. Different from the results for the standard plane detection task (Table 1), in this regression task we can see that our method outperforms the supervised settings (*ImageNet Init.* and *US-Sup*), suggesting that our approach has better generalisability. We also present qualitative examples in Fig. 5c, in which the predicted saliency map is overlaid to the original US image. This qualitative comparison further validates the effectiveness of the proposed method, and the results were validated by human experts. An interesting observation is that, when comparing our approach with the video-audio baseline, we can see (in the qualitative results) that our predicted saliency is more compact and closer to the real human expert performance, though for some of the quantitative metrics shown in Fig. 5a, the difference is not that significant (*Video-audio baseline* vs. *Ours*). As clarified above, this is mainly due to the fact that the metrics in Fig. 5a are proposed for natural images instead of medical imaging. When we conduct a further evaluation by using the newly proposed *Comp* metric (as in Fig. 5b), we can see that the proposed approach has a much better compactness performance, aligning well with the qualitative results.

Ablation study

In order to analyse the effectiveness of each component in the proposed framework, we conducted a controlled experiment (*i.e.* ablation study), as shown in Fig. 6. Performance on the above two downstream tasks is reported with the corresponding evaluation metrics. Here we chose the baseline as the very basic audio-video correspondence modelling (Fig. 2a). Then we compared this to the performance of adding cross-modal contrastive learning (*CM. Contra.*), the aforementioned video-audio baseline, the spatial awareness, the additional textual information, only using visual and text path for training (*w/o Audio*) and the SIG module.
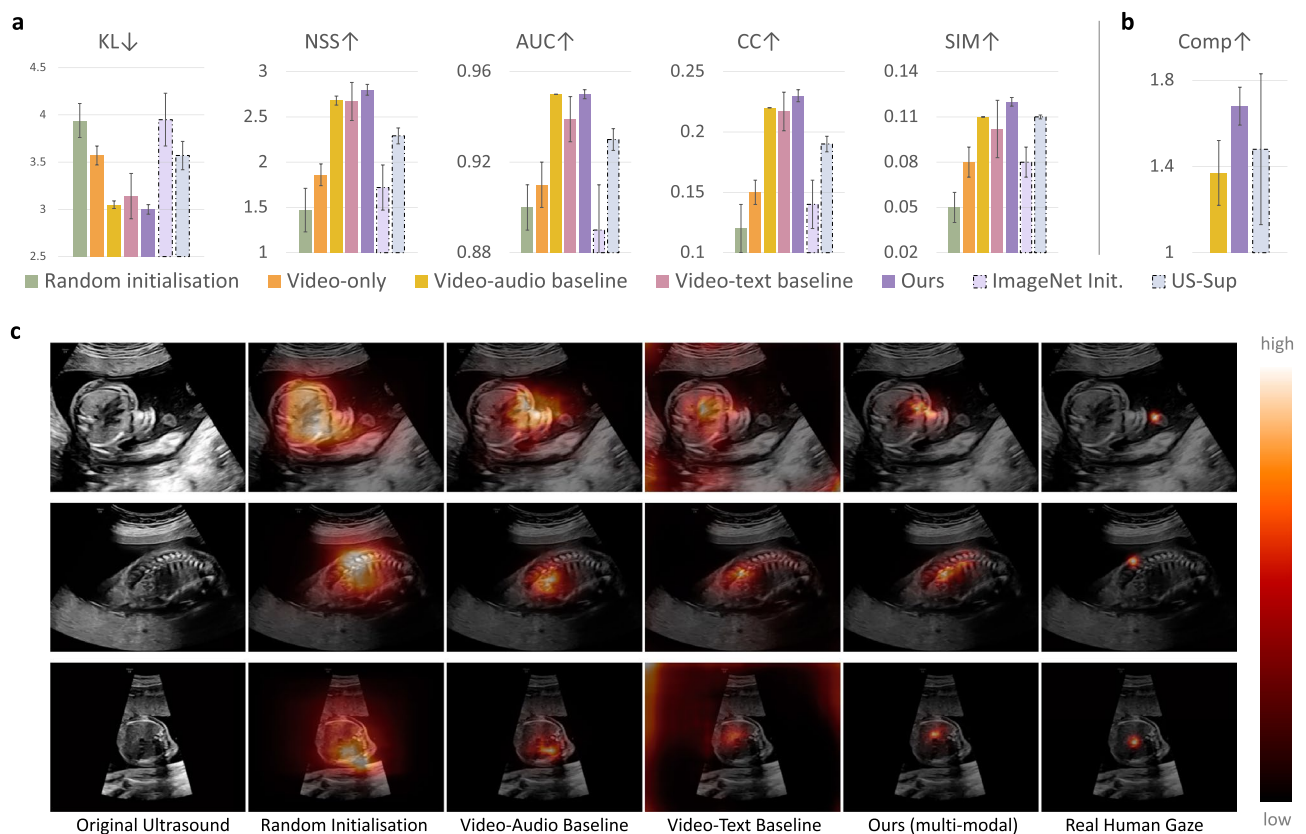

Figure 5. Evaluation on eye-gaze saliency prediction. Quantitative performance (**a**) evaluated by conventional metrics (KL, NSS, AUC, CC, and SIM), and (**b**) with the newly proposed *Comp* metric. ↓ indicates the lower the better while ↑ means the higher the better. (**c**) Qualitative performance on eye-gaze saliency prediction.
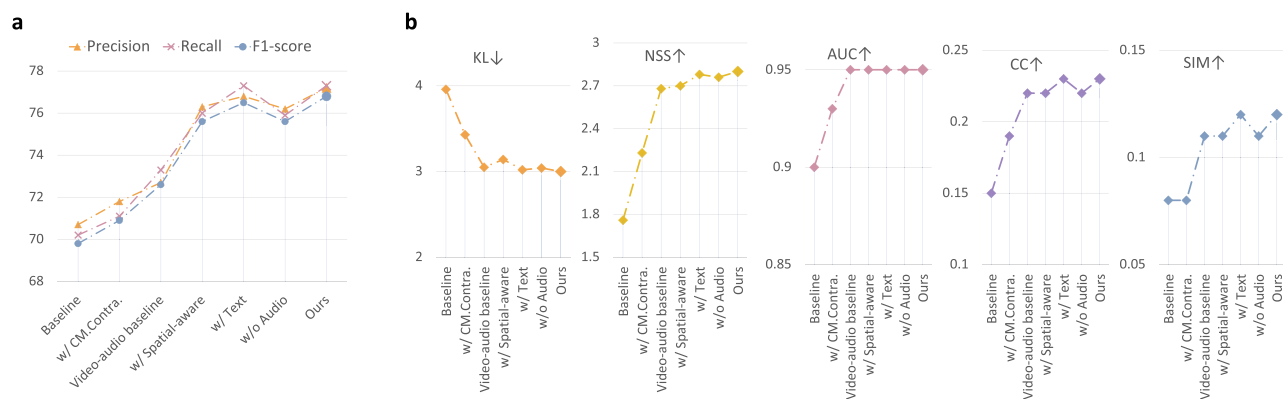
Figure 6. Ablation study on each of the proposed strategies for two downstream tasks. (**a**) Standard plane detection. (**b**) Eye-gaze saliency prediction, where ↓ indicates the lower the better while ↑ means the higher the better.

For further validation as part of a more detailed ablation study, please refer to the "supplementary information". We can see that each component contributes to the performance gain of our proposed approach, validating their effectiveness.

Audio-guided visual anatomy localisation

We have shown that the proposed multi-modal audio-visual modelling framework is able to perform audio-guided visual anatomy localisation with no extra human interventions. In this section, we showcase how this can be achieved. During the training of the multi-modal representation learning framework, audio (as well as derived text) embedding perceive the spatial information of the visual data well, by the fusion module. As a result, after the training, given input audio the model is able to highlight the high-response visual regions and such regions are usually related to anatomies due to the training target. Note that instead of visualising landmarks, our model shows the association between the visual signal and the audio input, trying to highlight the corresponding anatomical structures. In Fig. 7a, we showcase examples of this audio-guided visual anatomy localisation task. We can see that our model is able to correctly localise the anatomies that the speech audio contains, *e.g. heart, kidneys etc.*. This demonstrates the quality of the learned joint multi-modal representations and also illustrates a useful application towards anatomy localisation. The strategy for the audio-visual anatomy localisation task is the same setting as in our work mentioned above, *i.e.* the speech audio data is fed into the pretrained audio net together with the video data captured from the ultrasound machine being fed into the video net. Then the model is able to highlight the high-response visual anatomical regions accordingly. Under our experimental settings, the time to process a 244 × 244 frame and generate the corresponding response map is 0.08s, *i.e.* over 12 FPS (frames per second), which is acceptable for applications. In this application, the text data is not explicitly used, but has a role in the model pre-training stage, where the high-quality US representations are learned. It is important to note that if the text modelling component was removed during pre-training, a performance drop was observed for the downstream localisation task. As mentioned above, it is infeasible to acquire accurate clinical annotations for this task and we instead seek help from the eye-gaze saliency data as a reference to provide an approximate quantitative evaluation here. Specifically, we compare the localisation map of the prediction from our model and the eye-gaze saliency from the captured gaze data, and measure their similarity on a separate unseen test set with both audio and gaze data. The results are shown in Fig. 7c, using the evaluation metrics of
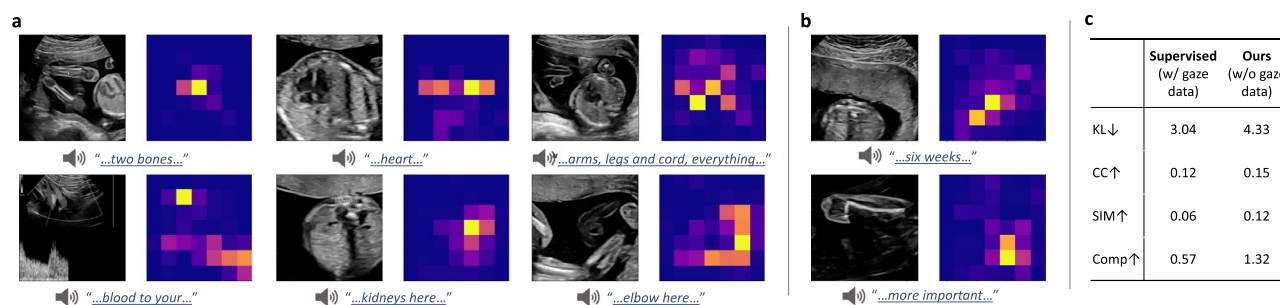

Figure 7. Performance on audio-guided visual anatomy localisation. (**a**) Performance on normal cases where audio is related to the visual anatomy, while in (**b**) the audio is unrelated to anatomy. In each sample, the left shows the input US with the corresponding speech audio (transcription) below; the right side shows the estimated localisation map from the proposed method. Higher response is in yellower color while lower in bluer color. (**c**) Quantitative evaluation of the audio-guided visual localisation task.

KL, CC, SIM and Comp (NSS and AUC are discarded as in this task we are not measuring human fixations). For comparison, we also include a saliency prediction method that was trained in a fully-supervised manner with ground-truth supervision from eye-gaze saliency data, and applied to the same test data. Note our method in this audio-guided visual localisation task never saw any gaze/saliency data but was only guided by the audio signal. From the results we can see that even without seeing any eye gaze data, our method still performs well and better (3 out of 4 metrics) than the solution that was trained with eye-gaze data as ground-truth supervision. As described above, the content of the speech audio is not always related to anatomical regions. In the case where the audio does not relate to any anatomy, the model may fail to give meaningful localisation. Examples of such "failure cases" are shown in Fig. 7b for reference. Interestingly, though the audio content does not relate to any anatomy, the model still tries to reason about some localisation, *e.g.* the model is looking at the anatomy and its neighbours to try to 'figure out' the "six weeks" narration.

Discussion
Limitations
Although this work made the first attempt towards learning multi-modal representations in real-world clinical settings and demonstrated meaningful representations, there are some limitations that may prevent the proposed approach from being applied more generally. In this work, we only validated the proposed model on clinical fetal US data. The proposed framework is general and there are no assumptions which make it obstetrics specific, however it has not been validated in other clinical settings as yet to prove generalisability. Another limitation is that the current model is designed to localise *anatomy-related* regions and not more general concepts (as shown in Fig. 7b). The main reason is that the most consistent audio-video pairs are related to anatomy in our case; if other well-paired training data for other concepts were available, this limitation could be addressed accordingly.

Potential impact and benefits
The study was conducted using data from one hospital site in Oxford, thus the corresponding conclusions may not identical for data from other sites. However, data was collected in a real-world setting; this also explains the presence of audio signals that are not relevant to anatomical regions, for example conversations between the sonographer and prospective parents. Thus, we believe that the model should apply to most similar settings, though including more data from other clinical environments could strengthen the model. There are several strengths of this work. We have made the first attempt towards multi-modal clinical representation learning with a particular focus on video-audio data modalities. This enables us to have a better and more thorough understanding of the clinical data and eliminates the requirement of human expert annotations. Yet this does not mean that the need for human experts has in been abated, instead the proposed approach provides a better way for models to work together with human experts, with the aim of getting more robust results. In some clinical settings, including the UK where this study was conducted, sonographers will communicate their findings and provisional information to the patient, thereby providing an additional modality to the data science of sonography. And sonographers usually impart such information during the scan, and confirming the normality of views or absence of abnormalities. The proposed model can also be utilised to help localise anatomical regions of interest, by only using speech audio, *i.e. talk to the screen, and show from tell*. This has both scientific and clinical contributions. Data annotation can be made much easier by only verbally describing the regions that need to be labelled. The proposed model may be useful to help the training of new sonographers.

Conclusion
In conclusion, in this work, we presented a study considering audio-visual modelling in a clinical setting, providing a solution to learn multi-modal medical representations that benefit various clinical tasks, without human expert annotation. A new audio-visual modelling framework has been presented that also incorporates textual data transcribed from the speech audio. Extensive experimental analysis on a large-scale clinical dataset showed that the proposed approach can learn meaningful representations, and such representations were shown to be effective in assisting several downstream clinical tasks. We also introduced a new clinical application, audio-guided visual localisation, which may be useful to both clinicians and patients in video interpretation. The specific tasks we looked at are all examples of assistive technologies that can support clinicians during clinical workflow.

Data availability
The datasets analysed during the current study are not publicly available due to the patient data governance policy. But the generated results and analyses performed during the current study are available from the corresponding author on reasonable request.

References
1. Arandjelovic, R. & Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, 609–617 (2017).
2. Arandjelovic, R. & Zisserman, A. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, 435–451 (2018).
3. Kazakos, E., Nagrani, A., Zisserman, A. & Damen, D. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF international conference on computer vision (ICCV)* (2019).
4. Korbar, B., Tran, D. & Torresani, L. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in neural information processing systems (NeurIPS)* (2018).

5. Morgado, P., Vasconcelos, N. & Misra, I. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12475–12486 (2021).
6. Drukker, L. *et al.* Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. *Sci. Rep.* **11**, 1–12 (2021).
7. Fu, Z. *et al.* Anatomy-aware contrastive representation learning for fetal ultrasound. In *European conference on computer vision workshop* (2022).
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inform. Process. Syst.* **26**, 3111–3119 (2013).
9. Defossez, A., Synnaeve, G. & Adi, Y. Real time speech enhancement in the waveform domain. In *Interspeech* (2020).
10. Wagner, J., Schiller, D., Seiderer, A. & André, E. Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant? In *Interspeech* (2018).
11. Bredin, H. *et al.* Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 7124–7128 (IEEE, 2020).
12. Baevski, A., Zhou, H., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint* arXiv:2006.11477 *(2020)*.
13. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A. T. & Noble, J. A. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, 1847–1850 (IEEE, 2020).
14. Baumgartner, C. F. *et al.* Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Trans. Med. Imaging* **36**, 2204–2215 (2017).
15. Cai, Y., Sharma, H., Chatelain, P. & Noble, J. A. 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, 2018).
16. Droste, R. *et al.* Ultrasound image representation learning by modeling sonographer visual attention. In *Information processing in medical imaging: 26th international conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pp. 592–604 (Springer, 2019).
17. Jiao, J. *et al.* Self-supervised contrastive video-speech representation learning for ultrasound. In *International conference on medical image computing and computer-assisted intervention*, 534–543 (Springer, 2020).
18. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. & Durand, F. What do different evaluation metrics tell us about saliency models?. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 740–757 (2018).
19. Borji, A. & Itti, L. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 185–207 (2012).
20. Wang, W., Shen, J., Guo, F., Cheng, M.-M. & Borji, A. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4894–4903 (2018).
21. Sugano, Y. & Bulling, A. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint* arXiv:1608.05203 (2016).
22. Alsharid, M. *et al.* Gaze-assisted automatic captioning of fetal ultrasound videos using three-way multi-modal deep neural networks. *Med. Image Anal.* **82**, 102630 (2022).
23. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500 (2017).
24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805 (2018).
25. Lee, J. *et al.* Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
26. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

Acknowledgements

Author contributions

J.J.: planning, developing the pipeline, carrying out, analysing and writing. M.A.: developing audio cleaning pipeline, carrying out, analysing, planning, data acquisition, and writing. L.D. and A.T.P.: data acquisition, analysing and writing. A.Z. and J.A.N.: planning, analysing and writing. All authors reviewed the manuscript.

Competing interests

J.A.N. and A.T.P. are Senior Scientific Advisors of Intelligent Ultrasound Ltd. J.J., M.A., L.D., A.Z. declare that they have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-66160-4.

Correspondence and requests for materials should be addressed to J.J. or J.A.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.