Dialogue Cross-Enhanced Central Engagement Attention Model for Real-Time Engagement Estimation

Jun Yu[1,2], **Keda Lu**[1,3*], **Ji Zhao**[1], **Zhihong Wei**[1], **Iek-Heng Chu**[4] and **Peng Chang**[4]
[1]University of Science and Technology of China
[2]Jianghuai Advance Technology Center
[3]Ping An Technology Co., Ltd, China
[4]PAII Inc.
harryjun@ustc.edu.cn, {lukeda, jzhao_tco, weizh588}@mail.ustc.edu.cn,
{zhuyixing276, changpeng805}@paii-labs.com

Abstract

Real-time engagement estimation has been an important research topic in human-computer interaction in recent years. The emergence of the NOvice eXpert Interaction (NOXI) dataset, enriched with frame-wise engagement annotations, has catalyzed a surge in research efforts in this domain. Existing feature sequence partitioning methods for ultra-long videos have encountered challenges including insufficient information utilization and repetitive inference. Moreover, those studies focus mainly on the target participants' features without taking into account those of the interlocutor. To address these issues, we propose the center-based sliding window method to obtain feature subsequences. The core of these subsequences is modeled using our innovative Central Engagement Attention Model (CEAM). Additionally, we introduce the dialogue cross-enhanced module that effectively incorporates the interlocutor's features via cross-attention. Our proposed method outperforms the current best model, achieving a substantial gain of 1.5% in coordination correlation coefficient (CCC) and establishing a new state-of-the-art result. Our source codes and model checkpoints are available at https://github.com/wujiekd/Dialogue-Cross-Enhanced-CEAM.

1 Introduction

Engagement is the process by which two (or more) participants establish, maintain, and end their perceived connection [Sidner and Dzikovska, 2002]. To understand the changing process of engagement we are studying human-to-human engagement interaction. The related studies provide essential capabilities for human-robot interaction. However, while it is easy for humans to recognize each other's engagement, it is difficult for machines to apply it in real time [Pellet-Rostaing *et al.*, 2023]. Therefore, automatic real-time engagement estimation has become an important problem in the machine learning field. Recently, researchers have increasingly recognized the importance of this task, mainly due to its wide range of applications in fields such as education [Nomura *et al.*, 2019], human-computer interaction [Sidner and Dzikovska, 2002; Yao *et al.*, 2023a], social interaction [Rajagopalan *et al.*, 2015; Lu and Churchill, 2014], and healthcare [Lo Presti *et al.*, 2019; Zhang *et al.*, 2022].

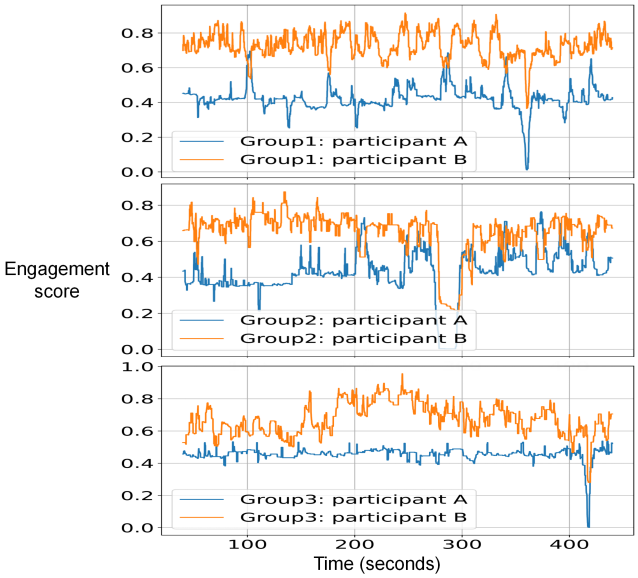

Figure 1: **The dynamic process of engagement during a conversation.** Three groups of conversations are randomly sampled from the NOXI dataset. Engagement scores range between 0 and 1, with 0 representing disengaged and 1 representing highly engaged.

Müller et al. [2023] proposed the first dataset suitable for real-time engagement estimation, named NOXI. Each video in NOXI contains 10,938 frames or more and frame-wise annotations. Many excellent works have been able to extract corresponding features from participants' video data, including Open Face 2.0 [Baltrusaitis *et al.*, 2018], Open Pose [Cao *et al.*, 2017], soundnet [Aytar *et al.*, 2016], Geneva Minimum Acoustic Parameter Set (GeMAPS) [Eyben *et al.*, 2016] and so on. Previous study [Müller *et al.*, 2023] had shown that

*Corresponding author

multi-modal features outperform uni-modal features on this task. To apply ultra-long video data to real-time engagement estimation, it is necessary to partition the complete video or the corresponding feature sequences. The normal data partitioning method shown in Figure 2(a) is to directly partition the feature sequences into multiple subsequences. After partitioning the raw sequences, most methods used the sequence-to-sequence (Seq2seq) model for modeling, including LSTM, transformer based on self-attention (SA), and others. In previous studies, SA-based models did not perform as well as LSTM-based models for this task [Yu *et al.*, 2023b]. In addition, the normal data partitioning method encountered a significant loss of contextual information. The sliding window method proposed by Yu et al. [2023b] for acquiring overlapping subsequences can utilize fuller contextual information. However, it required repetitive inference, leading to a decrease in real-time inference performance.

Engagement is not only reflected in the target participant but also influenced by their interlocutors [Rudovic *et al.*, 2019; Presti *et al.*, 2013]. As shown in Figure 1, the changing process in engagement during the conversation is highly correlated between pairs. However, most existing methods usually ignore information from their interlocutor.

To address the above issues, we explore how to use participants' multi-modal feature sequences for real-time engagement estimation. We study the effectiveness of sliding windows and overcome its drawback of requiring repetitive inference. We propose the center-based sliding window method, which uses the target prediction sequence as the central focus of the window (referred to as the core). This method extracts subsequences from the participant's ultra-long video multi-modal feature sequence by sliding this window. Then, we propose the Central Engagement Attention Model based on SA, which emphasizes attention to the core of the subsequence by reducing attention to the edges of the subsequence. In addition, we propose the dialog cross-enhanced module based on cross-attention (CA), which uses the encoded feature subsequence of the target participant and their interlocutor for interaction, further enhancing performance. This marks the first application of CA to real-time engagement estimation in dialogue interactions. Finally, our proposed model achieves CCC of 0.835 and 0.704 on the validation set and test set for real-time engagement estimation, respectively, establishing new state-of-the-art (SOTA) results.

Overall, our main contribution lies in the following aspects.

- We propose the center-based sliding window to obtain subsequences from the multi-modal feature sequences, which outperforms previous data partitioning methods.
- We propose a novel real-time engagement estimation model, CEAM, in combination with the center-based sliding window. The model is mainly implemented using SA, surpassing the previous best models.
- On the basis of CEAM, we propose the dialogue cross-enhanced module based on CA, which uses the feature subsequence of interlocutors to augment the target participant's feature subsequence, further improving the performance and establishing new SOTA results.

2 Related Work

2.1 Engagement Estimation

Dataset of Engagement Estimation. A large number of datasets existed from previous engagement estimation tasks, such as RECOLA [Ringeval *et al.*, 2013], MHHRI [Celiktutan *et al.*, 2017], and the dataset [Bednarik *et al.*, 2012] from annotated by Hradis et al. [2012]. However, these datasets were mainly used for estimating engagement at the video level or for simple classification tasks (like categorizing engagement as positive or negative). Using them to estimate constantly changing human engagement is difficult. It was not until the emergence of NOXI [Müller *et al.*, 2023] that the first dataset suitable for real-time estimation of continuous engagement became available. This dataset provides ultra-long videos of each participant with their interlocutor, with frame-wise annotated engagement.

Real-Time Engagement Estimation. As shown in Figure 2, for the engagement estimation task of videos with frame-wise annotation, previous research initially extracts uni-modal or multi-modal feature sequences from ultra-long videos. For these ultra-long feature sequences, recurrent neural networks (RNNs) [Elman, 1990] or SA-based model had found it difficult to directly model the complete feature sequence. Simultaneously considering real-time engagement estimation, these feature sequences must be partitioned and then subsequently trained. The mainstream partitioning method is shown in Figure 2(a), cutting into multiple subsequences for training and inference. Yu et al. [2023b] introduced the sliding window to obtain overlapping subsequences to ensure the correlation between the upper and lower subsequences, as shown in Figure 2(b).

Once the data was partitioned, it needed to be modeled by choosing an appropriate model. Müller et al. [2023] based on the multilayer perceptual network (MLP), discusses the effects of head, pose, and voice unimodal as well as multimodal features. Tu et al. [2023] addressed the problem by designing dilated convolution combined with LSTM or transformer (DCTM). Yu et al. [2023b] designed Seq2seq models based on BiLSTM and Self-Attention (SA) to predict engagement, respectively. Furthermore, low engagement may occur when participants are distracted by their interlocutor or other unrelated tasks, such as an unexpected phone call or microphone malfunction. Yang et al. [2023] reduced the dimensionality of the feature subsequences for participants and their interlocutors using Principal Component Analysis (PCA), concatenated them, and then employed an MLP to predict the target engagement. Those model structures are also categorized into three types: MLP-based, LSTM-based, and SA-based (including Transformer). In these previous works, the BiLSTM-based model combined with the sliding window method performs the best, while the SA-based and MLP-based model are slightly less effective.

2.2 Attention Mechanism

Self-Attention. Self-attention (SA) is a mechanism used to process sequence data, especially widely used in the transformer model [Vaswani *et al.*, 2017]. This mechanism also

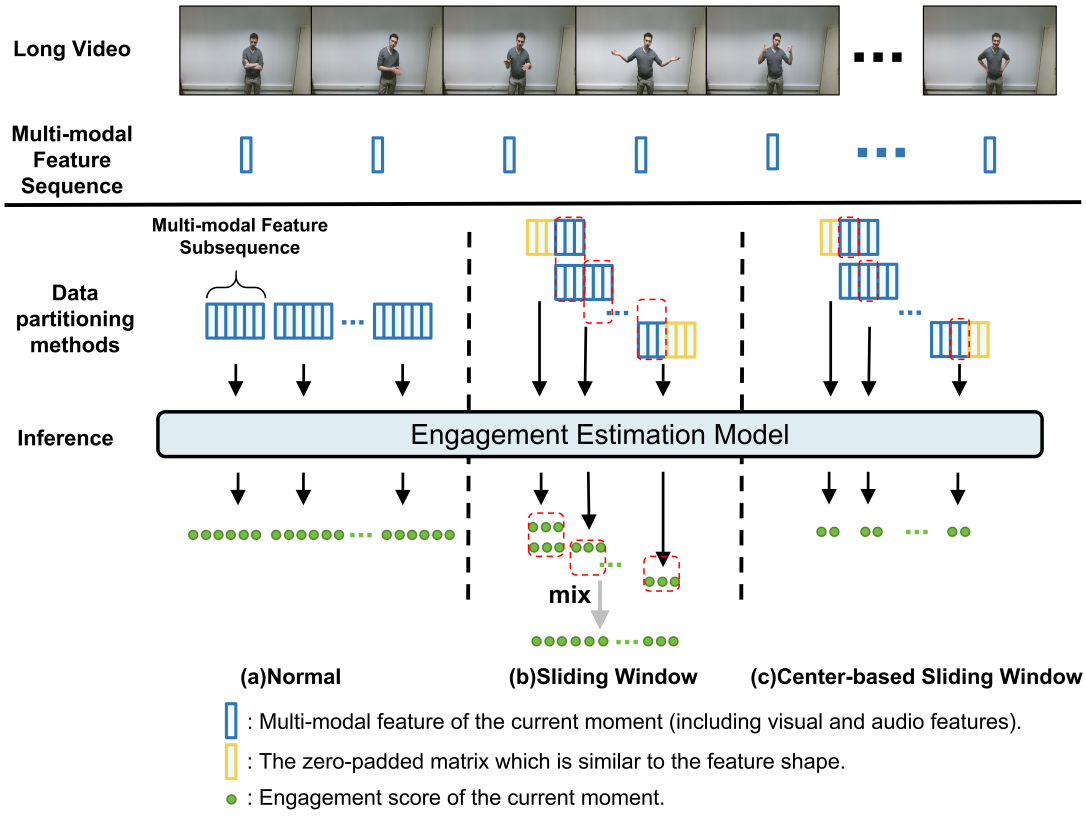

Figure 2: **Summary and Comparison of feature sequence partitioning methods for ultra-long videos.** (a) The existing mainstream partition method to get non-overlapping subsequences for training and inference. (b) Existing sliding window based partitioning method to generate overlapping subsequences, which needs repetitive inference. Here, "Mix" refers to average fusion, the fusion of the resultant levels of multiple inference on overlapping subsequences. (c) Our proposed center-based sliding window method focuses on the center (core) of the subsequence and uses the core extended window for sliding to get the subsequence (see Section 3.1). In the inference phase, frame-level scores are only computed once to output the core's engagement.

plays an important role in multi-modal sentiment analysis, such as multiview task [Zadeh *et al.*, 2018; Yu *et al.*, 2023a] and emotion recognition [Chudasama *et al.*, 2022; Le *et al.*, 2023]. When using SA to deal with sequence data, especially for data like video, audio, or text that has an obvious sequence structure but no explicit positional information, to help the model understand the positional relationships of the elements in the sequence, it is necessary to introduce the positional embedding [Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Yao *et al.*, 2023b].

Cross-Attention. Cross-attention (CA) is also an attention mechanism for modeling relationships between multiple sequences or different patterns. This mechanism has a wide range of applications in tasks such as processing multi-modal information [Li *et al.*, 2021; Lu *et al.*, 2019], emotion recognition [Zhou *et al.*, 2023; Praveen *et al.*, 2023], disease diagnosis [Praveen *et al.*, 2023]. CA is not limited to feature interactions between different modalities, it can also be applied to feature interactions within the same modality, e.g., CrossViT [Chen *et al.*, 2021]. However, the application of CA in conversational videos is nearly non-existent, particularly in real-time engagement estimation tasks.

3 Methodology

In this section, we detail our proposed dialogue cross-enhanced CEAM for real-time engagement estimation, as shown in Figure 3. We will first summarize previous data partitioning methods of ultra-long videos, then introduce our proposed center-based sliding window. Next, we will introduce the CEAM based on SA. Finally, we will introduce the dialogue cross-enhanced module based on CA.

3.1 Center-based Sliding Window

We summarize two previous data partitioning methods before presenting our proposed method. The normal data partitioning method, as shown in Figure 2(a), is fast in inference but is ineffective because it cannot fully utilize the contextual information of long videos. As proposed by Yu et al. [2023b], the sliding window method is shown in Figure 2(b). During the inference phase, however, it is necessary to compute predictions repeatedly for each subsequence. Then the results of these overlapping subsequences are mixed to generate the corresponding engagement scores. This approach ensures that the target subsequence obtains semantic information from both previous and subsequent features, respectively.

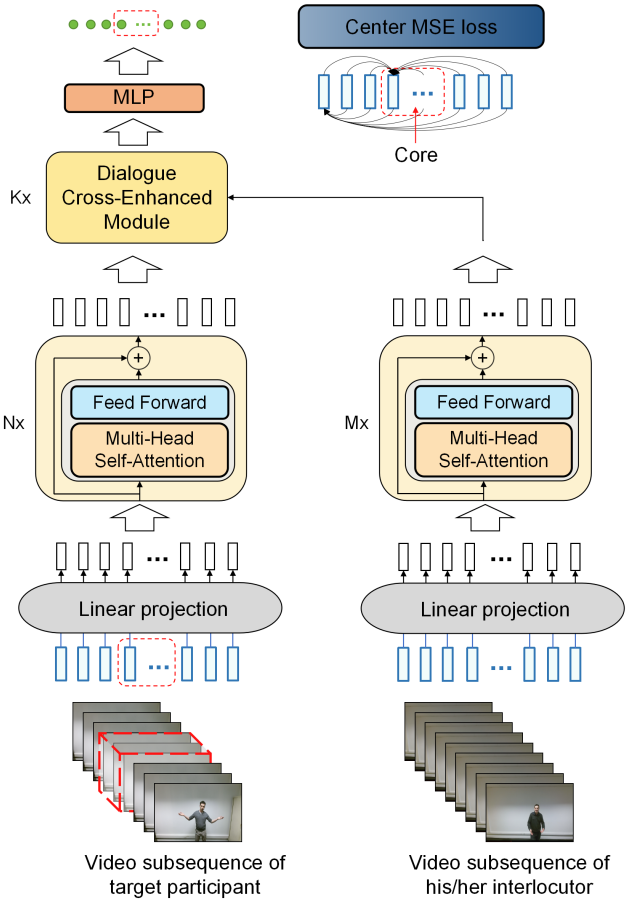

Figure 3: **An illustration of our proposed dialogue cross-enhanced Central Engagement Attention Model for real-time engagement estimation.** The Linear projection and self-attention encoder for the target participant and their interlocutor have the same structure, but the weights are not shared.

However, repetitive inference prolongs the overall inference time, posing a challenge to real-time processing.

We analyze the partitioned feature subsequence, where each node in the subsequence is affected differently. Specifically, edge features can not consider front or back influences, while the middle features can bridge the past and the future. When human beings express their thoughts, the changes in movements are continuous rather than discrete, despite the momentary changes. Therefore, based on the strong correlation between participants' video contexts, we propose the center-based sliding window method, as shown in Figure 2(c). The method focuses on modeling the middle sequences of each subsequence and using the middle sequences as the core of the sliding window to obtain the subsequences for training. The training strategy of our proposed data partitioning method is consistent with the previous two methods, but slightly different in the inference phase. Due to the insufficient information received at the front and back ends of the subsequence, we only infer about the middle sequence (core) of the subsequence. In addition, this method requires only one inference to obtain the engagement scores of the target participant, thus outperforming the previous sliding window in terms of inference efficiency. The training process of this center-based sliding window approach can be formalized as follows

$$\begin{aligned} &\hat{y}_{t-w}, \ldots, \hat{y}_{t-1}, \\ &\hat{y}_{t}, \ldots, \hat{y}_{t+l-1}, \\ &\hat{y}_{t+l}, \ldots, \hat{y}_{t+l+w-1} = f(x_{t-w}, \ldots, x_{t-1}, \\ &\qquad\qquad x_{t}, \ldots, x_{t+l-1}, \\ &\qquad\qquad x_{t+l}, \ldots, x_{t+l+w-1}), \end{aligned} \tag{1}$$

where $[x_t, \ldots, x_{t+l-1}]$ represents the multi-modal feature from t to $t+l-1$ moment, $[\hat{y}_t, \ldots, \hat{y}_{t+l-1}]$ represents the estimated target engagement subsequence, $[x_{t-w}, \ldots, x_{t-1}]$ and $[x_{t+l}, \ldots, x_{t+l+w-1}]$ represent contextual subsequence extended around estimated subsequence, $[\hat{y}_{t-w}, \ldots, \hat{y}_{t-1}]$ and $[\hat{y}_{t+l}, \ldots, \hat{y}_{t+l+w-1}]$ represent extended predicted engagement scores, l represents the length of the target subsequence, w represents the length of the extended window, and f is the engagement estimation model.

The inference process of the center-based sliding window method can be formalized as

$$\begin{aligned} \hat{y}_{t}, \ldots, \hat{y}_{t+l-1} = &f(x_{t-w}, \ldots, x_{t-1}, \\ &x_{t}, \ldots, x_{t+l-1}, \\ &x_{t+l}, \ldots, x_{t+l+w-1}), \end{aligned} \tag{2}$$

where the representations here are consistent with Equation 1.

3.2 Central Engagement Attention Model

Next, we introduce our proposed Central Engagement Attention Model based on SA. As shown in Figure 3, this is the structure of our proposed model. The cross-enhanced module will be introduced in the next subsection, and we first focus on the main branch, i.e., when M=0 and K=0. We take the input multi-modal feature subsequences and pass them through a linear projection, which includes embedding and sinusoidal positional embedding [Vaswani *et al.*, 2017]. Subsequently, they are passed through a stacked SA encoder, and finally, we use the MLP to predict engagement scores at the corresponding moments. The SA encoder and transformer share similar configurations, consisting of a series of blocks. Each block contains multi-head self-attention (MSA) along with a feed-forward network (FFN). The FFN comprises 2 multi-layer perceptrons, with a GELU activation function applied after the first linear layer. Layer normalization (LN) is applied after each subblock, and the residual shortcut is applied after each subblock. The input of the CEAM, x_0, and the processing of the k-th block can be expressed as

$$\begin{aligned} \mathrm{x}_0 &= \mathrm{x}_{emb} + \mathrm{x}_{pos} \\ \mathrm{z}_k &= \mathrm{LN}(\mathrm{x}_{k-1} + \mathrm{MSA}(\mathrm{x}_{k-1}))) \\ \mathrm{x}_k &= \mathrm{LN}(\mathrm{z}_k + \mathrm{FFN}(\mathrm{z}_k)), \end{aligned} \tag{3}$$

where $\mathrm{x}_{emb} \in \mathbb{R}^{N \times C}$ is the embedding subsequence after embedding the original multi-modal feature subsequence and $\mathrm{x}_{pos} \in \mathbb{R}^{N \times C}$ is the positional embedding. N and C are the length of the input feature subsequence and the dimension of the embedding, respectively.

In addition, we need to mitigate the network degradation problem caused by the deep network hierarchy. Inspired by the skip connection proposed by He et al. [2016], we design a weighted block skip connection, which is regulated by the control factor α and can be expressed as follows

$$\mathrm{x}_k = \alpha \cdot \mathrm{x}_k + (1-\alpha) \cdot \mathrm{x}_{k-1}, \tag{4}$$

where α represents a hyperparameter from 0 to 1 that controls each block of the skip connection.

Center MSE loss. When the attention mechanism computes the attention score matrix, it treats features from various time steps in the input subsequence equally. However, the information obtained from the features at each moment in the subsequence is unequal, and there is an information gap, as shown in the upper right corner of Figure 3. To address this issue, we design an optimized loss function for the output target engagement scores. This loss function enhances the predictive power of middle features and reduces the influence of front and end features. We propose the center mean squared error (MSE) loss function, combined with the center-based sliding window method, to improve the engagement attention model's capability to focus on the core of subsequences. It can be formalized explicitly as

$$\begin{aligned}\text{Center MSE} = \frac{1}{l}\sum_{i=t}^{t+l-1}(y_i-\hat{y}_i)^2 + \frac{\beta}{2w}[\sum_{i=t-w}^{t-1}(y_i-\hat{y}_i)^2 \\ + \sum_{i=t+l}^{t+l+w-1}(y_i-\hat{y}_i)^2],\end{aligned} \tag{5}$$

where $\hat{y}$ and y are predicted engagement score and true label, respectively. β represents a coefficient from 0 to 1 that controls the level of attention of the extended subsequence.

3.3 Dialogue Cross-enhanced Module

Finally, we propose the dialogue cross-enhanced module based on CA. As illustrated in Figure 3, this module seamlessly integrates with our proposed CEAM. Specifically, the SA encoder independently encodes the feature subsequences of the participants and their interlocutors. Then, these encoded feature subsequences interact in the dialogue cross-enhanced module, finally being processed through the MLP to predict engagement scores at the corresponding moments.

The internal details of the module are shown in Figure 4. The module performs CA between x_t and x_p, where x_t and x_p are the encoded feature subsequences of the participant and their interlocutor, respectively. Mathematically, the CA can be expressed as

$$\begin{aligned}\mathbf{q} = \mathrm{x}_p\mathbf{W}_q, \mathbf{k} = \mathrm{x}_t\mathbf{W}_k, \mathbf{v} = \mathrm{x}_t\mathbf{W}_v, \\ \mathbf{A} = \mathrm{softmax}(\mathbf{q}\mathbf{k}^T/\sqrt{\mathrm{C/h}}), \mathrm{CA}(\mathrm{x}^t, \mathrm{x}^p) = \mathbf{A}\mathbf{v},\end{aligned} \tag{6}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are learnable parameters, C and h are the embedding dimension and number of heads. Note that we use the interlocutor's features to compute the query to enhance the target features. Because the two input feature sequences are highly aligned, the output feature sequence shape is identical to the input. In a conversational video with two participants, it is possible to use each other's features to mutually enhance them. Moreover, as in SA, we also use multiple heads in the CA and represent it as (MCA), which is realized by the linear mapping function $g(\cdot)$. We also apply the same FFN as SA after the CA. Specifically, this module is also composed of multiple blocks stacked together, which also introduces a weighted block skip connection. The input of the dialogue cross-enhanced module, x_0, and the processing of the k-th block can be expressed as

$$\begin{aligned}\mathrm{x}_0 &= \mathrm{x}_t \\ \mathrm{z}_k &= \mathrm{LN}(\mathrm{x}_{k-1} + \mathrm{MCA}(\mathrm{x}_{k-1}, \mathrm{x}_p)) \\ \mathrm{x}_k &= \mathrm{LN}(\mathrm{z}_k + \mathrm{FFN}(\mathrm{z}_k)) \\ \mathrm{x}_k &= \alpha \cdot \mathrm{x}_k + (1-\alpha) \cdot \mathrm{x}_{k-1},\end{aligned} \tag{7}$$

where $\mathrm{x}_t \in \mathbb{R}^{N\times C}$ and $\mathrm{x}_p \in \mathbb{R}^{N\times C}$ are the encoded feature subsequences of the target participant and their interlocutor, respectively. N and C are the length of the input feature sequence and the dimension of the embedding, respectively. α represents the block skip coefficient from 0 to 1.

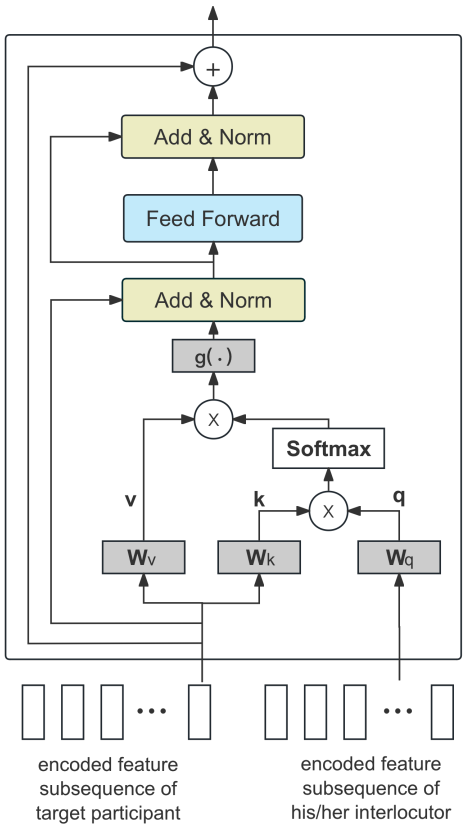

Figure 4: Dialogue Cross-enhanced module Internal Details.

4 Experiments

In this section, we experimentally demonstrate that our proposed dialogue cross-enhanced CEAM is more effective compared to existing methods. First, we introduce the dataset and evaluation metrics. Then, we present the experimental setup and the main results. Finally, we conduct ablation studies to analyze the necessity of each component in the architecture.

Model	Method (Time Window Size)	Interlocutor	Val CCC ↑	Test CCC ↑
MLP-based [Müller *et al.*, 2023]	Normal (0.1s)		0.710	0.590
DCTM (LSTM) [Tu *et al.*, 2023]	Normal (2.5s)		0.750	0.630
DCTM (Transformer) [Tu *et al.*, 2023]	Normal (2.5s)		0.750	0.660
PCA+MLP-based [Yang *et al.*, 2023]	Normal (10s)	✓	0.745	0.695
SA-based [Yu *et al.*, 2023b]	Normal (2.5s)		0.775	-
BiLSTM-based [Yu *et al.*, 2023b]	Normal (2.5s)		0.799	-
SA-based [Yu *et al.*, 2023b]	Sliding window (2.5s)		0.796	-
BiLSTM-based [Yu *et al.*, 2023b]	Sliding window (2.5s)		0.818	0.689
BiLSTM-based [Yu *et al.*, 2023b]	Center-based Sliding window (2.5s)		0.820	-
CEAM (Ours)	Center-based Sliding window (2.5s)		0.821	0.691
CEAM (Ours)	Center-based Sliding window (10s)		0.834	**0.711**
Dialogue Cross-Enhanced CEAM (Ours)	Center-based Sliding window (2.5s)	✓	**0.835**	0.704

Table 1: Comparison of validation and test results for engagement estimation under different methods. The time window size refers to the maximum duration for utilizing the video subsequence. Specifically, 0.1s, 2.5s, and 10s correspond to 1 frame, 64 frames, and 256 frames of images, respectively. The above models all use the same multi-modal features, including visual and audio features.

4.1 Dataset and Evaluation Metrics

Benchmark Dataset. The NOXI for Engagement Estimation dataset was obtained by Müller et al. [2023] using the published NOvice eXpert Interaction database (NOXI) [Cafaro *et al.*, 2017] for re-labeling. The NOXI contains interactions recorded in eight languages (English, French, German, Spanish, Indonesian, Arabic, Dutch, and Italian) at three locations (France, Germany, and the United Kingdom) on a wide range of topics. The dataset provides over 25 hours (x2) of recorded two-person interactions in natural environments, as well as synchronized audio, video (25fps), and motion capture data (using Kinect 2.0). Each participant's video is continuously annotated, meaning that each video frame has an engagement score from 0 to 1. The dataset, which is currently the longest recorded and the only dataset with continuous annotated engagement scores, is divided into a training and validation set. Additionally, there is an unpublished labeled test set available for online testing.

Evaluation Metrics. Human change is continuous, the authors of the benchmark dataset [Müller *et al.*, 2023] suggest using the concordance correlation coefficient (CCC) [Lin, 1989] to assess the similarity between the complete sequence of predicted scores, $\hat{y}^L$, and the complete sequence of true labels, y^L, for each of the target participants in the validation or test set, where L represents the length of the sequence. The CCC is defined as

$$\rho_c = \frac{2\rho\sigma_{y^L}\sigma_{\hat{y}^L}}{\sigma_{y^L}^2 + \sigma_{\hat{y}^L}^2 + (\mu_{y^L} - \mu_{\hat{y}^L})^2}, \tag{8}$$

where ρ is the pearson correlation coefficient, σ_{y^L} and $\sigma_{\hat{y}^L}$ are the standard deviations of y^L and $\hat{y}^L$, and μ_{y^L} and $\mu_{\hat{y}^L}$ are the means of y^L and $\hat{y}^L$.

4.2 Experimental Setup

Similar to previous studies [Müller *et al.*, 2023; Yao *et al.*, 2024], we extract multi-modal features from each frame of all ultra-long videos and their corresponding audio clips. Then, we concatenate these features, as done in Yu et al. [2023b], to form 1704**D** multi-modal features at the current moment. Finally, we obtain multi-modal feature sequences for each video and use the center-based sliding window to partition them into subsequences for subsequent training. The core length is set to 32 with an extended window length of 32, i.e., using the time window size of approximately 2.5 seconds.

The linear projection layer maps the multi-modal feature subsequence of length l to $\mathbf{x}_{emb}$, where $\mathbf{x}_{emb} \in \mathbb{R}^{l\times 768}$, and then adds sinusoidal positional embeddings to obtain the embedded vectors. When the dialogue cross-enhanced module is not used, we set N = 3, M = 0, and K = 0. The SA block comprises MSA with 8 heads. The FFN of the SA encoder consists of 2 linear layers with dimensions of 768×4 and 768, respectively. When using the dialogue cross-enhanced module, we set N = 1, M = 1, K = 2. The dialogue cross-enhanced module repeatedly stacks K blocks, each block comprises MSA with 8 heads. The setting of each block is similar to the block of the SA encoder. In both the SA encoder and the dialogue cross-enhanced module, the block skip connection coefficient α is set to 0.5. The MLP output block has one hidden layer consisting of 128 neurons (with SELU activation function) and one output layer.

We train all our models for 100 epochs on 1 Nvidia V100 GPU with a batch size of 32. Other setups include a learning rate scheduler, specifically utilizing the Reduce Learning Rate On Plateau algorithm, with a reduction factor of 0.5 and a patience of 10 epochs. Additionally, we use an Adam optimizer with a learning rate of $1e^{-3}$ and our proposed center MSE loss function with the β of 0.5.

4.3 Main Results

Comparisons with SOTA models. As shown in Table 1, we compare our method with all previous methods. The feature sequence partitioning methods are classified into normal, sliding window, and center-based sliding window. Considering the specific applications of real-time estimation, we temporarily ignore the method proposed by Yang et al. [2023], which uses feature subsequences of 10 seconds as input. As shown in Table 1, Our proposed CEAM, in combination with

Model (Method)	Params (M)	Speed ↑ (FPS)
SA-based model (SW)	22.67	4,537
BiLSTM-based model (SW)	36.17	1,310
BiLSTM-based model (CSW)	36.17	2,029
CEAM (CSW)	23.98	**6,455**
Dialogue Cross-Enhanced CEAM (CSW)	31.07	6,185

Table 2: Comparison of the inference speed between our proposed method and the previous best model in the real-time engagement estimation task. FPS: Frames Per Second, refers to the number of frames inferred per second, obtaining the corresponding engagement score. SW refers to the sliding window, and CSW refers to our proposed center-based sliding window.

N	M & K	Val CCC ↑	N & M	K	Val CCC ↑
1	0	0.819	1	1	0.819
2	0	0.820	2	1	0.825
3	0	**0.821**	1	2	**0.835**
4	0	0.815	2	2	0.818
5	0	0.814	1	3	0.812

Table 3: Ablation study with the different architecture of Dialogue Cross-Enhanced CEAM.

our proposed center-based sliding window method, performs excellently. The CCC on the validation set improves by 2.5% compared to the previous best SA-based model [Yu *et al.*, 2023b] and outperforms the previous SOTA BiLSTM-based model [Yu *et al.*, 2023b] with sliding window.

Efficient use of interlocutor's information. When incorporating informative interactions from the interlocutor, the previous method, as demonstrated by Yang et al. [2023], is rudimentary. The long subsequence (10 seconds) brings more stable results on both validation and test sets but is not suitable for real-time engagement estimation. In addition, PCA resulted in a loss of subsequence feature information, causing underfitting on the validation set.

As shown in Table 1, we also test the time window size of 10 seconds on CEAM, and the results show that sufficiently long feature subsequences lead to significant improvement, and the same applies to our proposed model. Our proposed dialogue cross-enhanced module, which can be directly applied to our base model to fully use interlocutor information and enhance attention capabilities, finally achieves the CCC of 0.835 on the validation set and 0.704 on the test set, further enhancing the ability for real-time engagement estimation.

Comparison of inference speed. As shown in Table 2, the inference speed of our proposed method is significantly improved by nearly 400% compared to the previous best BiLSTM-based model. Due to the superiority of parallel computation in SA, it substantially outperforms the serial computation of BiLSTM overall. Moreover, the overall speed of both CSW-based methods is better than SW-based methods. In addition, our model has an advantage in terms of the number of parameters, and the smaller number of parameters is easier to deploy to mobile devices.

Heads	4	6	8	12
Val CCC ↑	0.790	0.831	**0.835**	0.796

Table 4: Ablation study on the multi-head attention.

α	Val CCC ↑
0.3	0.817
0.5	**0.835**
0.7	0.821

Table 5: Ablation results of block skip coefficient α.

β	Val CCC ↑
0	0.807
0.5	**0.835**
1.0	0.810

Table 6: Ablation results of weighting coefficient β.

4.4 Ablation Study

We perform some ablation studies for the important parameters in our proposed dialogue cross-enhanced CEAM.

Depth of SA encoder and dialogue cross-enhanced module. When the dialogue cross-enhanced module is not used, i.e., M = 0 and K = 0, we conduct ablation studies on the number of blocks in the SA encoder of CEAM, as illustrated in the left column of Table 3. As the number of blocks increases from 3 upwards, performance will start to degrade. To enhance the fusion frequency across two branches (SA encoder of targeted participants and interlocutors), CA modules (K) can be stacked (by reducing N and M to maintain the same total model depth). The results are presented in the right column of Table 3. Fusing branches too often does not improve performance but introduces more parameters.

Number of heads in multi-head attention. As shown in Table 4, we test the effect of different numbers of heads in multi-head attention and find the best performance. In our models, an excessive number of heads may lead to overfitting, while too few heads could impede the capture of intricate input relationships, thereby impacting performance.

Effectiveness of weighted block skip connection. Skip connection is introduced in each block in the SA encoder and dialogue cross-enhanced module, and the circulation information is controlled through α. Table 5 shows the necessity of using weighted skip connections between blocks, which are optimized when set to 0.5.

Effectiveness of Center MSE loss. As shown in Table 6, center MSE loss is better than MSE loss, and the model cannot miss the attention of the edge of feature subsequences.

5 Conclusions

To summarize, we introduce a novel center-based sliding window method for partitioning feature sequences in real-time engagement estimation. Then, we propose the dialogue cross-enhanced CEAM that effectively incorporates the features of the interlocutor via cross-attention. Our experimental results validate the effectiveness of our approach and highlight the potential of the attention mechanism in real-time engagement estimation. In the future, we aspire to further enhance the model performance by fostering closer interaction with both the target participants and their interlocutors.

Acknowledgments

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), Dreams Foundation of Jianghuai Advance Technology Center (2023-ZM01Z001), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Z231100005923035).

Contribution Statement

Jun Yu and Keda Lu contributed equally to this work.

References

[Aytar *et al.*, 2016] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[Baltrusaitis *et al.*, 2018] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face amp; Gesture Recognition (FG 2018)*, May 2018.

[Bednarik *et al.*, 2012] Roman Bednarik, Shahram Eivazi, and Michal Hradis. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, pages 1–6, 2012.

[Cafaro *et al.*, 2017] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Nov 2017.

[Cao *et al.*, 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[Celiktutan *et al.*, 2017] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497, 2017.

[Chen *et al.*, 2021] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[Chudasama *et al.*, 2022] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661, 2022.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[Elman, 1990] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[Eyben *et al.*, 2016] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, page 190–202, Apr 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[Hradis *et al.*, 2012] Michal Hradis, Shahram Eivazi, and Roman Bednarik. Voice activity detection from gaze in video mediated communication. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, Mar 2012.

[Le *et al.*, 2023] Hoai-Duy Le, Guee-Sang Lee, Soo-Hyung Kim, Seungwon Kim, and Hyung-Jeong Yang. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access*, 11:14742–14751, 2023.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[Lin, 1989] Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, page 255, Mar 1989.

[Lo Presti *et al.*, 2019] Letizia Lo Presti, Mario Testa, Vittoria Marino, and Pierpaolo Singer. Engagement in healthcare systems: Adopting digital tools for a sustainable approach. *Sustainability*, 11(1):220, 2019.

[Lu and Churchill, 2014] Jie Lu and Daniel Churchill. The effect of social interaction on learning engagement in a social networking environment. *Interactive learning environments*, 22(4):401–417, 2014.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[Müller *et al.*, 2023] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, et al. Multimediate'23: Engagement estimation and bodily behaviour recognition in social interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9640–9645, 2023.

[Nomura *et al.*, 2019] Kazuaki Nomura, Motoi Iwata, Olivier Augereau, and Koichi Kise. Estimation of student's engagement based on the posture. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 164–167, 2019.

[Pellet-Rostaing *et al.*, 2023] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science*, 5:1062342, 2023.

[Praveen *et al.*, 2023] R Gnana Praveen, Patrick Cardinal, and Eric Granger. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.

[Presti *et al.*, 2013] Liliana Presti, Stan Sclaroff, and Agata Rozga. Joint alignment and modeling of correlated behavior streams. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 730–737, 2013.

[Rajagopalan *et al.*, 2015] Shyam Sundar Rajagopalan, OV Ramana Murthy, Roland Goecke, and Agata Rozga. Play with me—measuring a child's engagement in a social interaction. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

[Ringeval *et al.*, 2013] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr 2013.

[Rudovic *et al.*, 2019] Ognjen Rudovic, Hae Won Park, John Busche, Bjorn Schuller, Cynthia Breazeal, and Rosalind W. Picard. Personalized estimation of engagement from videos using active learning with deep reinforcement learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun 2019.

[Sidner and Dzikovska, 2002] Candace L Sidner and Myrosia Dzikovska. Human-robot interaction: Engagement between humans and robots for hosting activities. In *Proceedings. fourth ieee international conference on multimodal interfaces*, pages 123–128. IEEE, 2002.

[Tu *et al.*, 2023] Vu Ngoc Tu, Van Thong Huynh, Hyung-Jeong Yang, Soo-Hyung Kim, Shah Nawaz, Karthik Nandakumar, and M Zaigham Zaheer. Dctm: Dilated convolutional transformer model for multimodal engagement estimation in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9521–9525, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems*, Jun 2017.

[Yang *et al.*, 2023] Chunxi Yang, Kangzhong Wang, Peter Q Chen, MK Michael Cheung, Youqian Zhang, Eugene Yujun Fu, and Grace Ngai. Multimediate 2023: Engagement level detection using audio and video features. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9601–9605, 2023.

[Yao *et al.*, 2023a] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9455–9465, October 2023.

[Yao *et al.*, 2023b] Jiawei Yao, Tong Wu, and Xiaofeng Zhang. Improving depth gradientcontinuity in transformers: A comparative study on monocular depth estimation with cnn. *arXiv preprint arXiv:2308.08333*, 2023.

[Yao *et al.*, 2024] Jiawei Yao, Xiaochao Pan, Tong Wu, and Xiaofeng Zhang. Building lane-level maps from aerial images. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3890–3894. IEEE, 2024.

[Yu *et al.*, 2023a] Jun Yu, Mohan Jing, Weihao Liu, Tongxu Luo, Bingyuan Zhang, Keda Lu, Fangyu Lei, Jianqing Sun, and Jiaen Liang. Answer-based entity extraction and alignment for visual text question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9487–9491, 2023.

[Yu *et al.*, 2023b] Jun Yu, Keda Lu, Mohan Jing, Ziqi Liang, Bingyuan Zhang, Jianqing Sun, and Jiaen Liang. Sliding window seq2seq modeling for engagement estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9496–9500, 2023.

[Zadeh *et al.*, 2018] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Zhang *et al.*, 2022] Zhijie Zhang, Jianmin Zheng, and Nadia Magnenat Thalmann. Engagement estimation of the elderly from wild multiparty human–robot interaction. *Computer Animation and Virtual Worlds*, 33(6):e2120, 2022.

[Zhou *et al.*, 2023] Siwei Zhou, Xuemei Wu, Fan Jiang, Qionghao Huang, and Changqin Huang. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. *International Journal of Environmental Research and Public Health*, 20(2):1400, 2023.