Dyadformer: A Multi-modal Transformer for Long-Range Modeling of Dyadic Interactions

David Curto[1,2][*], Albert Clapés[3,4][*], Javier Selva[1,3][*], Sorina Smeureanu[1,3],
Julio C. S. Jacques Junior[3], David Gallardo-Pujol[1], Georgina Guilera[1], David Leiva[1],
Thomas B. Moeslund[4], Sergio Escalera[1,3,4], and Cristina Palmero[1,3]

[1]Universitat de Barcelona, [2]Universitat Politècnica de Catalunya,
[3]Computer Vision Center, [4]Aalborg University

david.curto@estudiantat.upc.edu, alcl@create.aau.dk, jaselvaca@ub.edu,
{crpalmec7, ssmeursm28}@alumnes.ub.edu, jjacques@cvc.uab.cat,
{david.gallardo, gguilera, dleivaur}@ub.edu, tbm@create.aau.dk, sergio@maia.ub.es

Abstract

Personality computing has become an emerging topic in computer vision, due to the wide range of applications it can be used for. However, most works on the topic have focused on analyzing the individual, even when applied to interaction scenarios, and for short periods of time. To address these limitations, we present the Dyadformer, a novel multi-modal multi-subject Transformer architecture to model individual and interpersonal features in dyadic interactions using variable time windows, thus allowing the capture of long-term interdependencies. Our proposed cross-subject layer allows the network to explicitly model interactions among subjects through attentional operations. This proof-of-concept approach shows how multi-modality and joint modeling of both interactants for longer periods of time helps to predict individual attributes. With Dyadformer, we improve state-of-the-art self-reported personality inference results on individual subjects on the UDIVA v0.5 dataset.

1. Introduction

In the past years, human interaction and, in particular, dyadic interaction, have been deeply studied by both psychological and artificial intelligence research communities [9, 8, 17]. This rising trend has led to remarkable development not only in data collection [10, 6, 12], but also in defining methods for automatically understanding and modeling interpersonal social signals [63, 52]. The way people adapt and react to such signals and behaviors during a conversation depends not only on their individual characteristics (*e.g.*, personality) but also on the specific situation and their shared history [9]. For example, one might behave more relaxed during a conversation with a friend than in a meeting with their foreman. When analyzing social interactions from a computational perspective, all these influential factors should be taken into consideration to truly under-

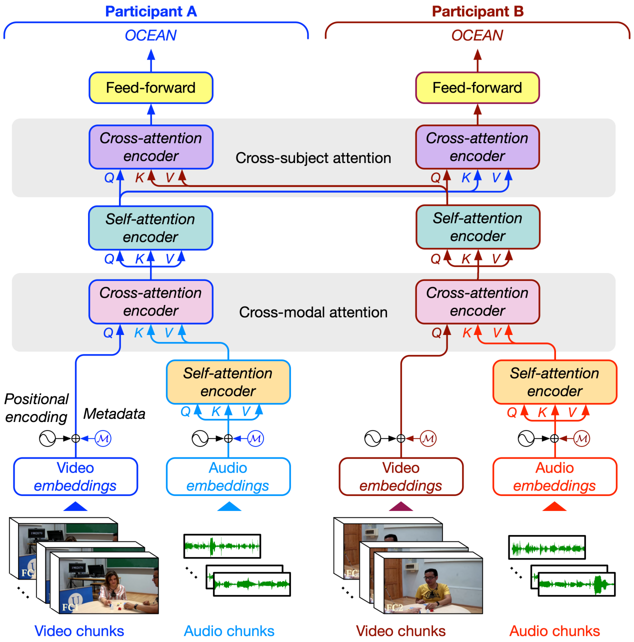

Figure 1. Proposed Dyadformer including different kinds of attention (self, cross-modal, and cross-subject). The model jointly infers the self-reported personality of both participants (A and B). Model complexity is reduced by sharing weights between parallel encoders (as illustrated by their colors) and across layers within each encoder. $\mathcal{M}$ are the corresponding metadata embeddings of each participant added to both their video and audio embeddings.

[*] These authors contributed equally to this work.

stand human behavior, even when the focus is on predicting individual attributes such as personality traits [72]. However, this is still not the norm throughout the literature [58].

Despite the growing interest in this area, current computational approaches for social interaction understanding present some other shortcomings. For instance, long-term modeling is crucial in interaction settings, as more complex dynamics emerge at different time scales, and an event may unchain effects that take time to be observed [9]. In the case of personality computing in such scenarios, the need for long-term modeling is heightened, as behavioral manifestations of certain traits may not be fully observed in short periods of time. Hence, more time is needed to find salient patterns arising during the interaction that can be associated to given traits [24]. Most existing works that deal with longer-term modeling have generally been addressed through single frame descriptors averaged over whole sequences [46], missing to represent the temporal evolution of features. Another aspect that fails to be properly modeled when assessing individual attributes in dyadic interactions is the joint modeling of both interlocutors. Despite its importance for triggering individual behaviors that provide insights on individual features [4], most of the works that do model it are focused on analyzing interaction attributes.

In this work, we propose a novel architecture to leverage long-term information for joint modeling of both interlocutors in dyadic scenarios. More precisely, we predict the personalities of both interactants by considering not only the audio-visual information and contextual factors (referred to as metadata) independently for each one but also by explicitly modeling their interaction. The proposed model, *Dyadformer*, mainly consists of two stages: (1) a *cross-modal* stage where cross-attention encoders fuse multi-modal information, and (2) a *cross-subject* stage which aims to shape the interaction by performing double cross-attention (see Fig. 1). Our contributions are summarized as follows:

- To our knowledge, this method is the first one to jointly model (and infer) self-reported personality in dyadic interactions using time windows of up to ∼30 seconds.
- Inspired by the classical decoder block of the Transformer network [69], we leverage a cross-attention mechanism to both fuse modalities and allow information to flow between subjects.
- Dyadformer obtained state-of-the-art results on the large-scale UDIVA v0.5 [58] dataset, by reducing previous participant-level error by a 12.5% (from 0.812 to 0.722) when predicting self-reported personality.

2. Related Work

Social signals and behaviors in context. Dyadic and small group interactions are a rich source of overt behavioral cues. They can provide insight into our personal attributes and cognitive/affective inner states dependent upon the context in which they are situated. Context can take many forms, from the interaction partner's attributes and behaviors to spatio-temporal and multi-view information. Joint modeling of both interlocutors and/or other sources of context have been extensively considered when trying to measure interpersonal constructs [13, 82], individual social behaviors [14, 81] and impressions [81, 59], and even empathy [59]. When considering individual attributes instead, context has often been misrepresented, in spite of extensive claims on its importance [4, 74, 70, 56].

In recent years, interlocutor-aware approaches have started to gain more attention, especially for emotion recognition in conversation [60, 50]. Richer contexts have been captured by explicitly modeling the temporal dimension, which was traditionally done through recurrent approaches [53, 77, 26]. However, recent works have started using BERT/Transformer-like architectures [83, 49]. Beyond text, using additional modalities has also been proposed, e.g., raw audiovisual data [79, 31, 76, 36], or speech cues and personality of the target speaker [47]. Regarding context-aware personality recognition (the focus of this work), a similar trend is seen, but the literature is even scarcer, as discussed next.

Automatic personality recognition. Personality is defined as the manifestation of individual nuances in patterns of thought, feeling, and behavior, that remain relatively stable over time [65]. In the personality computing field [71], it is usually characterized by the basic Big Five traits [30] (*Open-mindedness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Negative emotionality*), often referred to as OCEAN, based on self-reported assessments. Most works focus on personality recognition from the individual point of view [32, 75], even in a dyadic or small group conversational context [1], using only features from the target person. Initial studies tended to use handcrafted features from gestures and speech [57], while more recent works rely on deep learning approaches from raw data [55].

Few methods propose interlocutor- or context-aware methods for self-reported personality recognition in small group interactions. Most recent works focus on personality analysis on social media, generally limited to the textual modality (see [54] for a complete review), involving much more people while missing useful cues from face-to-face interactions. The work of [66] was one of the firsts to model conversation in small group settings, leveraging turn-taking temporal evolution from transcript features but focusing on apparent personality recognition (i.e., personality reported by external observers [38]). Other works have also focused on modeling transcribed interviews [35], but disregarding the interviewer. In [18], authors regressed individual and dyadic features of personality and social impressions utilizing handcrafted descriptors of prosody, speech, and visual activity. [51] proposed an interlocutor-modulated recurrent attention model with turn-based acoustic features. [80] predicted personality and performance labels by correlation analysis of co-occurrent key action events, which were extracted from head and hand pose, gaze, and motion intensity features. The use of person metadata (*e.g.*, gender, age, ethnicity, and perceived attractiveness) together with audiovisual data has only been applied in [61]. However, their goal was to better approximate the crowd biases for apparent personality inference in one-person videos.

Long-term modeling in personality computing. The need for longer-term modeling in personality regression tasks is highlighted in [64]. The authors proposed a model based on facial features for individual apparent and self-reported personality, but limited to 3-second time windows. Others have attempted long-term modeling of features for personality inference, but most are limited to compute sequence representations by averaging small clips or individual frames features [46, 40], which miss temporal relationships. As far as we know, only one previous work has used up to 1 minute without aggregating across clips [67]. However, they focused on first-impressions regression, which does not benefit from longer temporal windows.

In the past years, a new family of architectures has risen to address some limitations of traditional recurrent methods [41], i.e., the Transformer [69]. Originally designed for machine translation, it has shown impressive results for many sequence modeling tasks in a plethora of modalities [15, 16, 5]. These models are capable of attending to long-range data dependencies with few layers, allowing them to learn very useful representations. Recently, some works have started using Transformer-like architectures to model personality. However, these works tend to focus on the apparent personality of individuals alone [29] by only modeling text features, generally on social media posts [45, 78]. We focus on self-reported personality on real face-to-face dyadic interactions, which our proposed architecture explicitly exploits, and use the Transformer to model video, audio, and metadata modalities altogether.

In this line, [58] is, to the best of our knowledge, the only work on self-reported personality regression in dyadic scenarios that takes multi-modal context into account. However, this approach has some limitations. First, participant personality was regressed from just 3-second chunks, which may not be enough to properly model long-term interactions. In this work, we input longer clips (up to 30 seconds), allowing the model to learn useful longer-range relationships. Second, multi-modal fusion was done simply by concatenating the information from the video and audio modalities. In this work, we leverage multi-modal Transformers that exploit useful features from each source by looking at interdependencies, and fuse them in a shared representation space. Finally, despite [58] combining information from both individuals in the interaction, only the personality of the target subject was regressed. Our Dyadformer explicitly models the behavior from both individuals simultaneously, through our proposed two-stream cross-attentional Transformer, to eventually predict their personalities jointly.

Multi-modal Transformers. Our work is related to the recent use of Transformers [69] to learn multi-modal representations. The most common approach employs contrastive losses to bring paired samples (such as video and caption [27] or subtitles [48]) closer together. This is generally used for captioning or retrieval tasks, where both modalities provide similar information and the aim is to *translate* between them. However, in our setting we expect audio and video to convey different complementary information, for which we explore two better suited Transformer families. The first one uses a BERT-like [15] stream which concatenates modalities along the temporal dimension [44, 21] before input, effectively doubling sequence length. Nevertheless, as Transformers scale quadratically with input length, these methods incur in memory efficiency limitations. The second one solves this by using separate cross-attention streams [39, 37], replacing self-attention to allow both modalities to attend and enrich each other (see Sec. 3.2 for details), while the separate streams allow for independent modeling and maintain sequence length. This design has generally been used to fuse two modalities, as is our case, but it can easily be extended further [84]. In preliminary experiments, we tested using a BERT-like approach but, when compared with the latter alternative and in our setting, we found it to underperform. For this reason, we opted for cross-attentional streams to fuse multi-modal information and go one step further by also using this technique to model cross-subject interactionss.

3. Methodology

In this section, we present the Dyadformer (Fig. 1), composed of a set of *attentional encoder modules*. Each of these is a stack of *Transformer layers* [69]. A complete transformer layer is composed of two or more *sub-layers*. Each sub-layer executes a core *block*, followed by a residual connection and layer normalization [2]. We define these four elements in Sec. 3.1, and describe the cross-modal and cross-subject attention that form the full architecture in Sec. 3.2.

3.1. The Transformer

The core of the Transformer layer is a non-local operation [73], which allows every element in the input sequence to access information from any other. This is achieved through a special form of attention. To compute it, the input representation $J \in \mathbb{R}^{T\times d_\mathrm{w}}$ is mapped to a set of queries $Q \in \mathbb{R}^{T\times d_\mathrm{k}}$, and a memory $M \in \mathbb{R}^{T\times d_\mathrm{w}}$ is mapped to a set of paired keys $K \in \mathbb{R}^{T\times d_\mathrm{k}}$ and values $V \in \mathbb{R}^{T\times d_\mathrm{k}}$, where

$d_{\text{k}} = d_{\text{w}}/h$ and h is the number of heads (defined below). In the Transformer, the non-local operation is instantiated as the dot-product between Q and K in order to generate an affinity (attention) matrix that weights how much each value should contribute to the augmented representation of every other value. In general, J will be equal to M, hence this attention is called self-attention. The output of the self-attention operation is computed as

$$\text{Att}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_{\text{k}}}})V. \tag{1}$$

However, in order to build a full transformer layer, attention is not all you need. A complete transformer layer is composed by sub-layers that can be defined as $\text{SubLayer}^n_{\text{Block}}(x) = \text{LayerNorm}(x + \text{Block}(x))$, where the output of one sub-layer is fed as input to the next, n references the attentional encoder module that the sub-layer belongs to and $Block$ will either be Multi-Head self-Attention (MHA) or *position-wise Feed-Forward Network* (FFN) [69] which we define next.

Sub-Layer blocks. In order to allow for the model to attend to different information in a single sub-layer, [69] proposed the Multi-Head self-Attention (MHA). Similar to the multiple filters of a single convolutional layer, the multi-head self-attention maps J and M to h different representation sub-spaces to perform different attention operations. Then, the output of each head is concatenated and mapped back to a common d_{w}-dimensional representation through a linear transformation $W^O \in \mathbb{R}^{(h*d_{\text{k}})\times d_{\text{w}}}$. More formally,

$$\begin{gathered}\text{MHA}^n_f(J, M) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O, \\ \text{where head}_i = \text{Att}(JW^Q_{n,f,i}, MW^K_{n,f,i}, MW^V_{n,f,i}),\end{gathered} \tag{2}$$

$W^{\cdot}_{n,f,i} \in \mathbb{R}^{d_{\text{w}}\times d_{\text{k}}}$ and f is the sub-layer within which the MHA block is, detailed next.

In practice, every sub-layer in a Transformer layer contains a MHA block except for the last one, which contains a FFN block. It is composed of two linear layers with GELU [33] activation function in between, i.e., $\text{FFN}(x) = \text{GELU}(xW_{F1})W_{F2}$, where $x \in \mathbb{R}^{d_{\text{w}}}$ is the embedding corresponding to one timestep, and $W_{F1} \in \mathbb{R}^{d_{\text{w}}\times(4*d_{\text{w}})}$ and $W_{F2} \in \mathbb{R}^{(4*d_{\text{w}})\times d_{\text{w}}}$ are matrices of weights. In practice, point-wise FFN are equivalent to applying a fully connected (FC) layer repeatedly and independently to each timestep (these can also be seen as two 1D convolutional layers with kernel size 1). Note that FFN layers are also dependent on n and f, but we have omitted denoting those for simplicity.

Attentional Encoder modules. In our design, we use two main modules to build the complete architecture: the self-attention encoder $\text{SA}(J)$, which is used to enhance features by attending to themselves, and the cross-attention encoder $\text{CA}(J, M)$, which is used to allow for a set of features to attend to a different source. The former is composed of Transformer layers with a single $\text{SubLayer}^{\text{SA}}_{\text{MHA}}(J)$, while the latter is composed by two, $\hat{J}_l = \text{SubLayer}^{\text{CA}}_{\text{MHA}_1}(J_l)$ and $\text{SubLayer}^{\text{CA}}_{\text{MHA}_2}(\hat{J}_l, M)$. As mentioned earlier, in both cases ($\text{SA}(J)$ and $\text{CA}(J, M)$), the described sub-layers are followed by a last $\text{SubLayer}^n_{\text{FFN}}$ sub-layer. It is important to note that, in the cross-attention encoder, while the input J_l of layer l is the output from the previous cross-encoder layer, M is the same for all layers. This allows J_l features to iteratively attend to M to be progressively augmented.

Positional Encoding. Finally, as the self-attention operation is agnostic to relative position among input elements, [69] proposed using positional encodings to indicate the order of the input sequence by a composition of sine and cosine functions at varying frequencies. We followed the same procedure to indicate the ordering of the timesteps in the sequences from both input modalities.

3.2. The Dyadformer

Dyadformer (depicted in Fig. 1) is a multi-subject multi-modal architecture that follows the aforementioned transformer layers. The Dyadformer receives as input a sequence of T small, consecutive and temporally aligned video/audio chunks and infers the personality traits for both subjects in a dyadic interaction. It is composed of two main streams, each of which simultaneously processes a single subject.

As discussed in Sec. 2, context and interpersonal features are crucial to predict individual features in dyadic and small group interaction scenarios. For this reason, we propose a model which is capable of (a) fusing information from multiple sources (video, audio, and contextual metadata), and (b) allowing per-subject streams to access each other, in order to consider crossed influence during the interaction. To satisfy both, we go beyond self-attention, where $J = M$, and also use cross-attention, where $J \neq M$. Cross-attention works similarly to encoder-decoder attention in [69], where the input and memory come from different sources. For a transformer focusing on dyadic interactions, the target (J) will be from the subject of interest, while the memory (M) will be from the other one. The intuition behind this is to allow information from a given subject to *query* for useful information from the other. But first, each stream will create an individual representation for each subject. In order to do so, we draw inspiration from multi-modal transformer models [84, 39, 37], and use this same cross-attentional mechanism to fuse data coming from video and audio modalities. In this cross-modal module, J is from the video modality, while M is from the audio one, thus enriching video information with the audio signal. Finally, personality scores for both individuals are predicted jointly.

Input. We temporally divide videos and audios into small chunks first. Next, we precompute per-chunk feature representations using pre-trained networks (see Sec. 4.1 for details). Doing so, we can then feed our model with two pairs of sequences (X^p, U^p), where $p \in \{A, B\}$ denote the participants, $X^p = [x_1^p, \ldots, x_T^p]$ is a sequence of precomputed video features and $U^p = [u_1^p, \ldots, u_T^p]$ is the corresponding sequence of precomputed audio features. Note that X^A, U^A, X^B, and U^B are all temporally aligned. Apart from these, the model also receives the metadata handcrafted features, namely m^p. Then, the precomputed video and audio features, as well as metadata, are linearly projected into d_w-dimensional embeddings via three independent linear layers. Next, for each participant, positional encodings and their respective metadata embeddings are summed to their video and audio embeddings. Given m^p has no temporal dimension, before the summation, m^p is replicated T times using the outer product operation: $\mathcal{M} = \mathbf{1} \otimes m^p$, where $\mathbf{1}$ is a T-sized vector of ones.

Cross-modal and cross-subject attention. In order to build the multi-modal representation for each subject, we first feed the audio features U^p to an audio encoder module SA_aud composed by L_aud layers, such that $\hat{U}^p = \text{SA}_\text{aud}(U^p)$ where $\hat{U}^p \in \mathbb{R}^{T \times d_\text{w}}$. Then, we use a cross-encoder CA_vid with L_xm layers to enhance video features X^p with the new audio features, such that $\hat{X}^p = \text{CA}_\text{vid}(X^p, \hat{U}^p)$, where $\hat{X}^p \in \mathbb{R}^{T \times d_\text{w}}$.

The enhanced video features of each subject $\hat{X}^p$ are transformed through a subject encoder SA_sbj with L_sbj layers, such that $S^p = \text{SA}_\text{sbj}(\hat{X}^p)$, in order to learn rich relationships within individual subject features. This subject encoder is followed by a cross-encoder with L_xs layers as to allow the features from each subject to draw relevant information from each other, such that $\hat{S}^A = \text{CA}_\text{sbj}(S^A, S^B)$ and $\hat{S}^B = \text{CA}_\text{sbj}(S^B, S^A)$, where S^p and $\hat{S}^p \in \mathbb{R}^{T \times d_\text{w}}$.

Inference. For a given sequence, to infer the personality of the participant p in the dyad, we feed the output subject representations $\hat{S}^p = \{\hat{s}_1^p, \ldots, \hat{s}_T^p\}$ through an average pooling and two FC layers in order to regress the final OCEAN values for p, i.e. $\hat{o}^p = ((z^p W_\text{FC1}) W_\text{FC2})$, where $z^p = \frac{1}{T} \sum_{t=1}^{T} \hat{s}_t^p$, $z^p \in \mathbb{R}^{d_\text{w}}$, $W_\text{FC1} \in \mathbb{R}^{d_\text{w} \times 4 * d_\text{w}}$ and $W_\text{FC2} \in \mathbb{R}^{4 * d_\text{w} \times 5}$.

4. Experimental evaluation

Next, we experimentally evaluate a set of variants of the Dyadformer architecture for the task of self-reported personality traits regression and discuss the obtained results.

4.1. Data

UDIVA v0.5 dataset. All the experiments were based on UDIVA v0.5[1], a preliminary subset of the UDIVA dataset [58]. This subset is a highly varied multi-modal, multi-view dataset of zero- and previous-acquaintance, face-to-face dyadic interactions. It consists of 145 interaction sessions, where 134 participants (17-75 years old, 55.2% male) arranged in dyads performed a set of four tasks in a lab setting: *Talk*, *Lego*, *Ghost*, and *Animals*. Contained speech data is multi-lingual, with 73.1% of the sessions in Spanish, 17.25% in Catalan, and 9.65% in English. The UDIVA v0.5 dataset provides one frontal camera view per participant, the audio streams from each participant's lapel microphone, and participants' and sessions' metadata, as well as other modalities and annotations. Personality trait ground truth values were obtained from a self-reported BFI-2 questionnaire [22], and are provided as z-scores.

In our experiments, we used the audio-visual data and the metadata from [58]. The latter consists of 21 features encoding information of an individual (age, gender, cultural background, session index of participant, and pre-session mood/fatigue), session (task order within the session and its difficulty), and dyadic information (relationship between interactants). The dataset is divided into three subject-independent splits: 116/18/11 sessions and 99/20/15 participants in training, validation, and test, respectively.

Pre-segmented chunks and feature extraction. For the sake of comparison, we utilized the same set of video and audio chunks used in [58], provided by the authors. In their work, chunk availability was limited by a face detection algorithm, such that chunks with no detected face were discarded. Given also the difference in duration throughout sessions and tasks, the final number of chunks per task was uniformly subsampled based on the median. Then, to sample contiguous sequences of T chunks, some of them have been further discarded. Given these limitations, some tasks do not contain many chunks and, to avoid losing more data, we limited our experiments to $T \leq 12$. After all, we end up with a substantially smaller dataset: resulting train, validation and test splits contain, respectively, 94,960/15,350/7,870 pre-segmented chunks (equivalent to 67.5/10.9/5.6 hours). As Transformers are known to be data hungry [16], we follow other works [28, 42] who have successfully trained Transformers on smaller datasets by leveraging backbones pre-trained on Kinetics [11].

Each video chunk is composed of 32 frames at 12.5 fps ($\sim$2.56 seconds) at a spatial resolution of 224 × 224 pixels (normalized between $[0, 1]$), whereas each audio chunk was 3 seconds long, acquired at 44.1 kHz, and time-synchronized to its respective video chunk (i.e., the centers of corresponding video/audio chunks are aligned). Video, audio, and contextual metadata features are generated for each subject individually. Visual features are computed with R(2+1)D [68] pre-trained on IG-65M [25] and Kinetics [11]. We also fine-tuned its 5th block on the training set of UDIVA v0.5 during 13 epochs (after having replaced the last fully connected layer by another one of size 5 to predict OCEAN). Once trained, all the pre-segmented chunks

[1] https://chalearnlap.cvc.uab.es/dataset/41/description/.

of UDIVA v0.5 were reprocessed and the 512-dimensional feature representations output by the second to last layer of R2+1D were saved. Analogously, for audio, we used a VGGish [34] pre-trained on AudioSet [23] to compute a 128-dimensional representation for each audio chunk. Sequences of T such video/audio precomputed features were used as input for each subject in our method.

4.2. Parameters and implementation details

Following [15], we fixed $d_{\mathrm{w}} = 768$ and $h = 12$, and hence $d_{\mathrm{k}} = 64$. We set $L_{\mathrm{aud}} = L_{\mathrm{xm}}$ and $L_{\mathrm{sbj}} = L_{\mathrm{xs}}$ for our experiments. To maximize the number of consecutive T-length training sequences, they were sampled with a stride of 1 chunk. Metadata was included for all the experiments if not otherwise stated, based on the findings of [58].

Transformer models quickly grow in number of parameters. In our simplest model (see TF_{v} in Tab. 1) one Transformer layer accounted for ∼7.1M, whereas 8 layers accounted for ∼56.8M parameters (disregarding the backbones and final linear layers). Nevertheless, recent studies on Transformer models in NLP [3, 43], later extended to the audio-visual domain with similar results [44], have shown that weight sharing does not hurt representational power nor performance, while allowing for lighter and faster-to-train models. For this reason, in this work we always shared weights between all equivalent layers of both subject's streams. In other words, both streams were exactly the same. Also, for experiments where layers for any given module $L_{\cdot} \geq 1$, we shared parameters across them (*e.g.*, all cross-modal Transformer layers share weights).

Our model was trained by minimizing a MSE loss measuring the error of the inferred personality traits at sequence level versus its associated ground truth: $\mathcal{L} = \sum_{p \in \mathcal{P}} \sum_{i=0}^{5} (o_i^p - \hat{o}_i^p)^2$, o^p is the ground truth of self-reported personality and $\mathcal{P} \subseteq \{A, B\}$ (depending on the experiment). Model weights were trained by minimizing $\mathcal{L}$ via SGD optimization with weight decay $5e^{-3}$. Training was early stopped after 6 epochs if no improvement was observed on the validation loss. The learning rate was initially set to $5e^{-4}$ and reduced by a factor of 2 after 3 epochs without improvement. The dropout rate throughout all the layers in the architecture was set to 0.2.

4.3. Evaluation metrics

For the following experiments, we report the average per-trait Mean Squared Error (MSE) at two levels: (a) *sequence-level* (MSE$_{\mathrm{seq}}$), where the error was computed for every T-length sequence by comparing the predictions against ground truth personality of the subject appearing in them. The MSE$_{\mathrm{seq}}$ reported is the mean over all the T-length sequences in the test set; and (b) *participant-level* (MSE$_{\mathrm{part}}$), for which we first aggregated the predictions over all the sequences of a given participant by the median, and then compared it to that participant's personality ground truth. In contrast to MSE$_{\mathrm{seq}}$, MSE$_{\mathrm{part}}$ removes bias towards participants that appear more in the test set, hence being a more balanced metric for this problem. We choose to report both in this work to compare the effect of the different aggregation mechanisms.

Arch.	L		**MSE$_{\mathrm{seq}}$**		**MSE$_{\mathrm{part}}$**		**Params**
			$T = 6$	$T = 12$	$T = 6$	$T = 12$	
TF$_{\mathrm{v}}$	2		0.807	0.771	0.742	0.732	10.0M
	4		0.857	0.792	0.781	0.744	
	6		0.919	0.856	0.837	0.807	
	8		0.948	0.860	0.867	0.804	
	L_{xm}	L_{xs}	$T = 6$	$T = 12$	$T = 6$	$T = 12$	
DF$_{\mathrm{xm}}$	1	-	**0.797**	0.767	**0.738**	0.732	19.4M
	2	-	0.845	0.767	0.777	**0.722**	
	3	-	0.880	0.802	0.824	0.762	
DF$_{\mathrm{xs}}$	-	1	0.802	0.768	0.763	0.745	19.4M
	-	2	0.831	0.760	0.778	0.738	
	-	3	0.843	0.767	0.794	0.743	
DF$_{\mathrm{xm,xs}}$	1	1	0.831	0.760	0.794	0.741	36.0M
	1	2	0.847	0.765	0.802	0.748	
	2	1	0.854	**0.738**	0.809	**0.722**	
	2	2	0.894	0.758	0.842	0.737	

Table 1. Ablation of different architectures and sequence lengths (T chunks) in terms of average sequence- and participant-level mean squared errors: TF$_{\mathrm{v}}$, a Transformer on each subject's sequence separately; DF$_{\mathrm{xm}}$ or DF$_{\mathrm{xs}}$, the Dyadformer with only cross-modal ("xm") or cross-subject ("xs") attention respectively; and DF$_{\mathrm{xm,xs}}$ with both. $L_{\cdot}$ are the number of layers in the encoders. Best result per column in bold.

4.4. Ablation

Here we evaluate our two main contributions: (1) the use of multi-modal information and joint modeling of both participants against vanilla self-attention (using only video and one participant at a time); and (2) the inclusion of longer-range temporal context ($T = 6$ and $T = 12$ chunk sequences, corresponding to 15.36 and 30.72 seconds, respectively) with respect to [58] ($T = 1$, i.e., 2.56 seconds). In order to mitigate the stochasticity introduced by the random initialization of the network weights, we repeated each experiment 4 times (or 8 for models with $T = 12$) and report the average of their results[2].

Cross-modal and cross-subject attentions. To assess the cross-attention's contribution we test four variants of our model: (1) a self-attention Transformer (TF$_{\mathrm{v}}$) on the visual modality only and for each participant separately, i.e., attention is applied within each subject's sequence and neither cross-modal nor cross-subject attention are considered; (2) the Dyadformer with either only cross-modal attention (DF$_{\mathrm{xm}}$) or (3) cross-subject attention (DF$_{\mathrm{xs}}$); and (4) the full architecture with both cross-attentions (DF$_{\mathrm{xm,xs}}$). As shown in Tab. 1, the two strongest variants were DF$_{\mathrm{xm}}$ and DF$_{\mathrm{xm,xs}}$. Although TF$_{\mathrm{v}}$ was already a strong baseline model, it did not obtain the best result in any metric, suggesting that in-

[2] Further ablations are included in the supplementary material.

	O	C	E	A	N
Training (ground truth)	0.255 ±1.136	0.160 ±1.020	−0.053 ±0.969	−0.006 ±0.957	−0.346 ±1.085
TF_v ($L = 2$) wo/ metadata	−0.008 ±0.256	0.057 ±0.112	−0.186 ±0.062	−0.178 ±0.086	−0.431 ±0.064
TF_v ($L = 2$) w/ metadata	−0.053 ±0.323	0.126 ±0.313	−0.321 ±0.364	−0.134 ±0.345	−0.238 ±0.317

Table 2. Ablation on the regression to the mean problem. Mean and standard deviations of personality trait predictions by one run of the simpler TF_v ($L = 2$) without and with metadata and the same values over the training ground truth for comparison.

volving multiple modalities and explicitly modeling interaction among subjects is indeed beneficial for this task. The diminishing trend we observed on the performance of the models when further increasing their depth (number of encoder layers) discouraged us from trying further combinations and/or increasing their capacity with more parameters.

Temporal context. We then evaluated different temporal context lengths, i.e., $T \in \{6, 12\}$, for the aforementioned combinations. As shown in Tab. 1, $T = 12$ achieves better results (lower MSE_seq and MSE_part) throughout all the ablation. Interestingly, the Dyadformer variants with cross-subject attention, DF_xs $\text{DF}_\text{xm,xs}$, benefited more from longer sequences. This is aligned to the fact that interpersonal dynamics can span very different temporal ranges. That is, the behavior of one interlocutor could be considerably delayed in time. Hence, using $T = 12$ allows such long-term interdependencies to emerge and be further leveraged.

Use of metadata. Preliminary experiments showed the benefits of their use at a marginal computational cost. Tab. 2 shows, for the simplest ablated model TF_v ($L = 2$), that using only video results in very low values for the standard deviation. This *regression to the mean* problem is alleviated by allowing the model to access metadata information. Note that the lack of metadata especially hurts *Extraversion* ("E"), *Agreeableness* ("A"), and *Negative emotionality* ("N"). If we compute the mean of the two sets of standard deviations (with and without metadata, from Tab. 2), we obtain 0.332 versus 0.116, respectively. This indicates the models are more willing to deviate the personality trait predictions from a mean value when incorporating the extra context provided by metadata. This is in line with current state-of-the-art research in personality psychology, which states that personality needs to be expressed in *situations* [62], i.e., taking the interaction context into account.

4.5. Analysis across personality traits and tasks

Here, we analyze the results obtained in the ablation studies described in Sec. 4.4. First, we evaluate the results from the four different tasks present in the UDIVA v0.5 dataset, as each of them was designed to elicit different behaviors. Then, we study how different tested variants of the Dyadformer model the different OCEAN traits, given that not all traits are equally expressed nor captured. We compare our results to the two best-performing models of [58]. Note that such models were trained per task, whereas our tested models were trained on all tasks jointly.

Per-task analysis. As in [58], we analyzed the performance of the different model variations predicting the OCEAN traits separately depending on the task at hand. The results are shown in Tab. 3. As we can observe, among our models there is not a clear winner[3]. For *Animals*, TF_v is the one which provided more accurate results on average ("Avg") both in terms of MSE_seq and MSE_part, although DF_xm did equally well for "A". DF_xs outperformed the rest for the "N" trait in this task. Both for *Ghost* and *Lego*, $DF_\text{xm, xs}$ and DF_xm got the lowest error in terms of MSE_seq and MSE_part, respectively. Finally, for *Talk*, $DF_\text{xm, xs}$ outperformed the rest of the models on average, doing better than the rest for *Open-mindedness* ("O") and *Conscientiousness* ("C") measuring MSE_seq and also for "O" and "E" measuring MSE_part instead. Some of the findings diverge from the ones reported in [58]. For instance, whereas they found *Animals* to benefit more from audio than *Lego*, we see a contrary trend. However, note that our models are not trained in a task-specific fashion, thus the network has been able to learn from a wider range of behaviors encountered across tasks, which might impact the relative importance of each modality.

Per-trait analysis. Transversely to all tasks except for *Animals*, $DF_\text{xm, xs}$ is the most accurate model predicting "O" at participant-level. It is also the best at predicting "E" at participant-level and "C" at sequence-level, whereas DF_xm does a better job at participant-level for the latter across all tasks. For "A", $DF_\text{xm, xs}$ is a close second after DF_xs. Interestingly, for "A", both variants incorporating cross-subject attention improved results."A" is positively correlated with kindness, consideration, and cooperativeness, pro-social behaviors that are more clearly understood when the network attends to both interactants. In contrast, "N" does not usually benefit from cross-subject attention as this trait is more associated to the individual's inner context (i.e., stress, mood changes) [30]. Surprisingly though, we find opposite trends for *Animals*, for which "N" does highly benefit from cross-subject whereas "A" does not.

Per-trait vs. per-task discussion. While, on average, *Talk* is the task obtaining the lowest MSE_part error, that is not the case per trait. If we focus on participant-level, the *Talk* scenario does allow to better predict "C", and "E", but *Animals* is more informative for "O" and "A", and *Lego* for "N". At sequence-level, "E" is better predicted with *Lego* and "N" with *Ghost*. These findings are consistent with those reported in [58]. This can be useful for psychological research, because it provides evidence that different sit-

[3] Due to lack of space, Pearson correlation results (typically used in personality psychology [7]) are provided in the supplementary material to further validate our contributions.

uations actually enact different traits [20]. For the case of *Animals*, we can observe a strikingly low error for "A" followed by "O". This suggests that these two traits are likely enacted by this task. Trait-enactment refers to the idea that some situations enact, or activate, certain levels of traits required for this situation [19]. This pattern is confirmed when we look at *Talk*. Extraverted individuals are generally more talkative, but conscientious participants, even though they are not particularly extraverted, will engage in active talking when they are demanded to.

Comparison to state of the art. There exists only one previous work that published results on UDIVA v0.5 dataset [58]. The authors evaluated different model variants to complement the information from the "target person", the one whose personality was being predicted. Their simplest video-based model, namely "L", was enriched with either metadata ("m"), extended context ("E") – that is, the view from the other interlocutor – , and/or the audio ("a"). For their best model, namely "LEam", they reported a MSE_{part} of 0.812, which is largely reduced by our best model by 12.5% (0.722 in Tab. 1). Moreover, for the different architecture alternatives, they report per-task/per-trait MSE_{part}. Interestingly, they noted that including the different sources of information could benefit or worsen the performance in different scenarios. Their two best-performing models were "LEm" and "LEam". The latter achieved generally better results, except for *Lego*, where the noise in the audio signal caused by the bricks might have hurt performance. Our different proposed models outperform their two best variations in 15/20 cases, as can be seen in Tab. 3.

5. Conclusion

We presented the Dyadformer, a novel multi-modal multi-subject transformer architecture for modeling individual and interpersonal features in dyadic interactions using variable time windows, thus allowing the capture of long-term interdependencies. We thoroughly ablate our model in the UDIVA v0.5 dataset for the task of self-reported personality prediction to demonstrate the contributions of each attentional module, as well as the modeling of longer timesteps. Experimental results demonstrated the reliability of our approach by surpassing previous results in UDIVA v0.5, reducing the error by 12.5% with respect to [58]. Results also showed that context (or situations) matters in personality computing. Recently, situations have been put in the forefront of personality research to understand and predict real behavior [62]. In this sense, a promising extension of this work into the psychological realm would be to extract situational perceptions as we compute personality scores, since considering both features would undoubtedly improve behavior forecasting.

In addition to audio/video-based personality computing, our model allows for straightforward adaptations to other modalities, as well as extending our analysis to other individual and dyadic features. Future work will include the validation of the architecture for longer time windows, using other interaction datasets applied to different tasks, and exploring end-to-end learning which would allow for better coordination between backbones and the Dyadformer.

Arch. \ Trait	O	C	E	A	N	Avg
Animals (A)						
[58] (LEm)	- 0.736	- 0.834	- 0.968	- 0.669	- 1.192	- 0.880
[58] (LEam)	- 0.737	- **0.756**	- **0.887**	- 0.580	- 1.023	- 0.797
TF_v	**0.186** **0.455**	0.722 1.062	**0.659** 1.283	**0.049** **0.054**	1.511 0.975	**0.626** **0.766**
DF_xm	0.206 0.515	**0.691** 1.008	0.677 1.328	0.050 **0.054**	1.658 1.041	0.656 0.789
DF_xs	0.242 0.628	0.927 1.227	0.672 1.433	0.123 0.134	**1.367** **0.889**	0.666 0.862
$\text{DF}_\text{xm,xs}$	0.263 0.674	0.920 1.239	0.670 1.448	0.115 0.134	1.520 0.947	0.698 0.888
Ghost (G)						
[58] (LEm)	- 0.743	- 0.944	- 0.868	- 0.657	- 1.153	- 0.873
[58] (LEam)	- **0.741**	- 0.893	- 0.844	- 0.667	- 1.139	- 0.857
TF_v	1.217 0.858	0.609 0.633	0.665 0.723	0.595 **0.589**	0.783 **0.988**	0.774 0.758
DF_xm	1.231 0.889	**0.563** **0.584**	**0.629** **0.707**	0.615 0.617	0.778 0.989	0.763 **0.757**
DF_xs	1.156 0.808	0.619 0.707	0.778 0.781	**0.564** 0.604	0.786 1.039	0.781 0.788
$\text{DF}_\text{xm,xs}$	**1.122** 0.771	0.582 0.691	0.733 0.754	0.577 0.616	**0.775** 1.029	**0.758** 0.772
Lego (L)						
[58] (LEm)	- **0.727**	- 0.763	- 0.826	- **0.611**	- 1.037	- 0.793
[58] (LEam)	- 0.745	- 0.839	- 0.953	- 0.659	- 1.099	- 0.859
TF_v	0.925 0.808	0.806 0.657	0.514 0.755	0.614 0.710	**0.534** 0.866	0.679 0.759
DF_xm	0.916 0.827	0.753 **0.616**	**0.488** 0.743	0.647 0.732	0.537 **0.844**	0.668 **0.752**
DF_xs	0.847 0.749	0.801 0.663	0.575 0.789	0.555 0.709	0.567 0.975	0.669 0.777
$\text{DF}_\text{xm,xs}$	**0.808** 0.741	**0.727** 0.635	0.517 **0.736**	**0.527** 0.747	0.555 0.908	**0.627** 0.753
Talk (T)						
[58] (LEm)	- 0.825	- 0.718	- 0.878	- **0.639**	- 1.047	- 0.821
[58] (LEam)	- 0.773	- 0.790	- 0.869	- 0.670	- **0.985**	- 0.817
TF_v	1.107 0.736	0.472 0.513	0.561 0.462	0.846 0.708	1.074 1.076	0.812 0.699
DF_xm	1.117 0.735	0.467 **0.488**	**0.526** 0.440	0.862 0.719	**1.057** 1.081	0.806 0.693
DF_xs	0.896 0.632	0.454 0.529	0.707 0.479	**0.771** 0.671	1.095 1.124	0.785 0.687
$\text{DF}_\text{xm,xs}$	**0.861** **0.574**	**0.450** 0.504	0.617 **0.419**	0.794 0.683	1.082 1.135	**0.761** **0.663**

Table 3. Results per trait and task. For each model, first row is MSE_{seq} and second row is MSE_{part}. The "Avg" column depicts the average performance per row (over all the traits). We compare to the best two models of [58]. Best result per task, trait, and metric in bold. Also, best result among our ablations underlined.

Acknowledgments. This work has been partially supported by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya, and ICREA under the ICREA Academia programme.

References

[1] Oya Aran and Daniel Gatica-Perez. Cross-domain personality prediction: from video blogs to small group meetings. In *International Conference on Multimodal Interaction*, pages 127–130, 2013. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[4] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. 2

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 3

[6] Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. The corpus of interactional data: A large multimodal annotated resource. In *Handbook of linguistic annotation*, pages 1323–1356. Springer, 2017. 1

[7] Wiebke Bleidorn and Christopher James Hopwood. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2):190–203, 2019. 7

[8] Dawn O Braithwaite and Paul Schrodt. *Engaging theories in interpersonal communication: Multiple perspectives*. Sage Publications, 2014. 1

[9] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 2007. 1, 2

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 1

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[12] Huili Chen, Yue Zhang, Felix Weninger, Rosalind Picard, Cynthia Breazeal, and Hae Won Park. Dyadic speech-based affect recognition using dami-p2c parent-child multimodal interaction dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 97–106, 2020. 1

[13] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012. 2

[14] Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. In *2019 International Conference on Multimodal Interaction*, pages 440–445, 2019. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186, 2019. 3, 6

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3, 5

[17] Valentín Escudero, Minsun Lee, and Myrna L. Friedlander. *Dyadic Interaction Analysis*, page 45–67. Cambridge Handbooks in Psychology. Cambridge University Press, 2018. 1

[18] Sheng Fang, Catherine Achard, and Séverine Dubuisson. Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 225–232, 2016. 2

[19] William Fleeson and Mary Kate Law. Trait enactments as density distributions: The role of actors, situations, and observers in explaining stability and variability. *Journal of personality and social psychology*, 109(6):1090, 2015. 8

[20] David C Funder. Taking situations seriously: The situation construal model and the riverside situational q-sort. *Current Directions in Psychological Science*, 25(3):203–208, 2016. 8

[21] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[22] David Gallardo-Pujol, Victor Rouco, Anna Cortijos-Bernabeu, Luis Oceja, Christopher Soto, and Oliver John. Factor structure, gender invariance, measurement properties and short forms of the spanish adaptation of the big five inventory-2 (bfi-2). (Accepted). 5

[23] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 6

[24] Katharina Geukes, Simon M Breil, Roos Hutteman, Steffen Nestler, Albrecht CP Küfner, and Mitja D Back. Explaining the longitudinal interplay of personality and social relationships in the laboratory and in the field: The pils and the connect study. *PloS one*, 14(1):e0210424, 2019. 2

[25] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:12038–12047, 5 2019. 5

[26] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: Findings, pages 2470–2481, 2020. 2

[27] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. In *Advances on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[28] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 5

[29] Dersu Giritlioğlu, Burak Mandira, Selim Firat Yilmaz, Can Ufuk Ertenli, Berhan Faruk Akgür, Merve Kınıklıoğlu, Aslı Gül Kurt, Emre Mutlu, Şeref Can Gürel, and Hamdi Dibeklioğlu. Multimodal analysis of personality traits on videos of self-presentation and induced behavior. *Journal on Multimodal User Interfaces*, pages 1–22, 2020. 3

[30] Lewis R Goldberg. The structure of phenotypic personality traits. *American psychologist*, 48(1):26, 1993. 2, 7

[31] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2122–2132, 2018. 2

[32] Daniel Helm and Martin Kampel. Single-modal video analysis of personality traits using low-level visual features. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 2

[33] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[34] Shawn Hershey, Sourish Chaudhuri, Daniel P W Ellis, Jort F Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. 6

[35] Louis Hickman, Rachel Saef, Vincent Ng, Sang Eun Woo, Louis Tay, and Nigel Bosch. Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal*, 2021. 2

[36] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, Online, Aug. 2021. Association for Computational Linguistics. 2

[37] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*, 2020. 3, 4

[38] J. C. S. Jacques Junior, Y. Güçlütürk, M. Perez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. Van Gerven, R. Van Lier, and S. Escalera. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. 2

[39] Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. Sbat: Video captioning with sparse boundary-aware transformer. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 630–636. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track. 3, 4

[40] Jyoti Joshi, Hatice Gunes, and Roland Goecke. Automatic prediction of perceived traits using visual cues under varied situational context. In *2014 22nd International Conference on Pattern Recognition*, pages 2855–2860. IEEE, 2014. 3

[41] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019. 3

[42] Myeongjun Kim, Taehun Kim, and Daijin Kim. Spatio-temporal slowfast self-attention network for action recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2206–2210. IEEE, 2020. 5

[43] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. 6

[44] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *International Conference on Learning Representations*, 2021. 3, 6

[45] Simone Leonardi, Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. Multilingual transformer-based personality traits estimation. *Information*, 11(4):179, 2020. 3

[46] Bruno Lepri, Nadia Mana, Alessandro Cappelletti, Fabio Pianesi, and Massimo Zancanaro. Modeling the personality of participants during group interactions. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 114–125. Springer, 2009. 2, 3

[47] Jeng-Lin Li and Chi-Chun Lee. Attention learning with retrievable acoustic embedding of personality for emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 171–177. IEEE, 2019. 2

[48] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 3

[49] Qingbiao Li, Chunhua Wu, Zhe Wang, and Kangfeng Zheng. Hierarchical transformer network for utterance-level emotion recognition. *Applied Sciences*, 10(13):4447, 2020. 2

[50] Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. Context-dependent domain adver-

sarial neural network for multimodal emotion recognition. *Proc. Interspeech 2020*, pages 394–398, 2020. 2

[51] Yun-Shao Lin and Chi-Chun Lee. Using interlocutor-modulated attention blstm to predict personality traits in small group interaction. In *International Conference on Multimodal Interaction*, pages 163–169, 2018. 3

[52] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133, 2018. 1

[53] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019. 2

[54] Davide Marengo, Christian Montag, et al. Digital phenotyping of big five personality via facebook data mining: a meta-analysis. *Digital Psychology*, 1(1):52–64, 2020. 2

[55] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27, 2019. 2

[56] Philip Moore. Do we understand the relationship between affective computing, emotion and context-awareness? *Machines*, 5(3):16, 2017. 2

[57] Laurent Son Nguyen, Alvaro Marcos-Ramiro, Martha Marrón Romera, and Daniel Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *International Conference on Multimodal Interaction (ICMI)*, pages 437–444, 2013. 2

[58] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapes, Alexa Mosegui, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 1–12, January 2021. 2, 3, 5, 6, 7, 8

[59] Hye Jeong Park and Jae Hwa Lee. Looking into the personality traits to enhance empathy ability: A review of literature. In *International Conference on Human-Computer Interaction*, pages 173–180, 2020. 2

[60] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019. 2

[61] Ricardo Darío Pérez Principi, Cristina Palmero, Julio C Junior, and Sergio Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, 2019. 3

[62] John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. The situational eight diamonds: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, 107(4):677, 2014. 7, 8

[63] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *ACL/IJCNLP*, 2021. 1

[64] Siyang Song, Shashank Jaiswal, Enrique Sanchez, Georgios Tzimiropoulos, Linlin Shen, and Michel Valstar. Self-supervised learning of person-specific facial dynamics for automatic personality recognition. *IEEE Transactions on Affective Computing*, 2021. 3

[65] Christopher Soto and Oliver John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113:117–143, 07 2017. 2

[66] Ming-Hsiang Su, Chung-Hsien Wu, and Yu-Ting Zheng. Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):733–744, 2016. 2

[67] Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez. What your face vlogs about: expressions of emotion and big-five traits impressions in youtube. *IEEE Transactions on Affective Computing*, 6(2):193–205, 2014. 3

[68] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 11 2017. 5

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3, 4

[70] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7(4):397–413, 2015. 2

[71] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291, 2014. 2

[72] Seth A. Wagerman and David C. Funder. *Personality psychology of situations*, page 27–42. Cambridge Handbooks in Psychology. Cambridge University Press, 2009. 2

[73] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[74] Aidan GC Wright. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, 5(3):292–296, 2014. 2

[75] Liangqing Wu, Dong Zhang, Qiyuan Liu, Shoushan Li, and Guodong Zhou. Speaker personality recognition with multi-

modal explicit many2many interactions. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[76] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors*, 21(14):4913, 2021. 2

[77] Songlong Xing, Sijie Mai, and Haifeng Hu. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 2020. 2

[78] Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. Multi-document transformer for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14221–14229, 2021. 3

[79] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *AAAI Conference on Artificial Intelligence*, volume 2018, page 5642, 2018. 2

[80] Lingyu Zhang, Indrani Bhattacharya, Mallory Morgan, Michael Foley, Christoph Riedl, Brooke Welles, and Richard Radke. Multiparty visual co-occurrences for estimating personality traits in group meetings. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2085–2094, 2020. 3

[81] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 2

[82] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218–233, 2016. 2

[83] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, 2019. 2

[84] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020. 3, 4