-Supplementary Material-
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

1. Additional ablation experiments

Here we include further experiments we performed to assess the validity of various design choices for the proposed Dyadformer. First, we evaluate an alternative design for the cross-attentional modules. Second, we explore the usefulness of the self-attentional modules at different stages of our model.

Cross-attention versus bidirectional encoding. Besides cross-attention, we also tried to follow the approach of bidirectional encoding from BERT [1] (discussed in Sec. 2 of the main paper). This alternative was implemented through two stages. First, two parallel multi-modal BERT encoders (which share weights among them and within them), each performing video-audio joint attention on its corresponding subjects. Then, their outputs are fed to a second stage with one BERT encoder, effectively attending over the two subjects. For a fair comparison with our $\text{DF}_{\text{xm, xs}}$ with $L_{\text{xm}}, L_{\text{xs}} \in \{1, 2\}$, we tried with different number of layers for the encoders of this BERT-like architecture such that the number of MHA blocks in both was similar. In particular, BERT with $L_{\text{bm}}, L_{\text{bs}}$, where $L_{\text{bm}}, L_{\text{bs}} \in \{3, 6\}$ are, respectively, the number of layers in the multi-modal BERT encoders and the multi-subject one. The BERT configuration $L_{\text{bm}} = L_{\text{bs}} = 3$ corresponds to the same number of attention layers included in our model with $L_{\text{xm}} = L_{\text{xs}} = 1$ and $L_{\text{bm}} = L_{\text{bs}} = 6$ corresponds to $L_{\text{xm}} = L_{\text{xs}} = 2$. Moreover, regardless of the combination of $(L_{\text{bm}}, L_{\text{bs}})$, the number of parameters of the architecture is 17.1M, which is comparable to either DF_{xm} or DF_{xs} (both with 19.4M). We set $T = 12$ for these experiments. We show the results on Tab. 1, where the other results are the same as the ones reported in Tab. 1 of the main paper. This variant resulted slightly worse than the equivalent Dyadformer variants ($\text{DF}_{\text{xm,xs}}$) for all metrics and combinations of layers. These results highlight the effectiveness of the used cross-attentional modules. One possible reason for this to happen is that our cross-attentional design helps decouple self-attention from accesses to the external memory (through separate MHA operations). The bidirectional encoding, however, emulates accesses to internal and external representations through a single multi-head attention, which may hinder learning to attend differently to one and the other.

Self-attention before cross-attention. In preliminary experiments, the Dyadformer included self-attention modules before every cross-attention module. However, motivated by the observation of an overfitting trend for overly complex models, we considered discarding all self-attention modules so as to reduce the number of parameters. As a result, for our model in Fig. 1 on the main document, we removed the self-attention encoder between the video embedding and the cross-modal encoder. The self-attention after the audio embeddings was kept to give the audio features a chance to evolve (as video embeddings do during the cross-modal attention), especially given the fact that audio embeddings were extracted from a model not fine-tuned on the personality prediction task – differently from video ones. Regarding the self-attention encoders prior to cross-subject encoders, we experimentally found the impact was negative

*These authors contributed equally to this work.

Arch.	L		MSE_{seq}		MSE_{part}		Params
			$T=6$	$T=12$	$T=6$	$T=12$	
TF_{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_{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_{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	
$\text{DF}_{\text{xm,xs}}$	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	
	L_{bm}	L_{bs}	$T=6$	$T=12$	$T=6$	$T=12$	
BERT	3	3	-	0.818	-	0.784	17.1M
	3	6	-	0.820	-	0.780	
	6	3	-	0.814	-	0.766	
	6	6	-	0.800	-	0.761	

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

when removing those layers in our best cross-subject models, i.e., DF_{xs} and $\text{DF}_{\text{xm, xs}}$. Without those layers, MSE_{part} increases from 0.738 and 0.722 (reported in Tab. 1) to, respectively, 0.758 and 0.740.

2. Correlation analysis

In order to complement Tab.3 from the main text, we also report the Pearson correlation metric among the per-trait/per-task predictions and the self-reported personality ground truth for the participants in the test partition in Tab. 2.

By looking at this metric, TF_{v} displayed the worst average ("Avg") results, mostly correlating negatively with the ground truth. A notable exception is, however, that it obtained the highest correlation (over 0.8) for the *Agreeableness* ("A") trait in *Animals* and *Ghost*.

In contrast, it can be observed that all of our Dyadformer variants correlated positively with the ground truth scores (except for DF_{xm} in *Open-mindedness* ("O"), for which correlation is usually close to zero). DF_{xm} was less accurate for *Conscientiousness* ("C"), *Extraversion* ("E") and *Negative emotionality* ("N") than DF_{xs} when looking at the Pearson correlation, despite the opposite trend was observed looking at MSE-based metrics. DF_{xs} correlated best with "N", although it showed poor correlation with "A" and "O". $\text{DF}_{\text{xm,xs}}$ obtained the best "Avg" performance in terms of correlation for all the tasks, followed by DF_{xs}.

Arch. \ Trait	O	C	E	A	N	Avg
Animals (A)						
TF_{v}	**0.186**	0.722	**0.659**	**0.049**	1.511	**0.626**
	0.455	1.062	**1.283**	**0.054**	0.975	**0.766**
	-0.533	0.440	-0.638	**0.894**	0.110	0.055
DF_{xm}	0.206	**0.691**	0.677	0.050	1.658	0.656
	0.515	**1.008**	1.328	**0.054**	1.041	0.789
	-0.020	0.524	0.458	0.406	0.339	0.342
DF_{xs}	0.242	0.927	0.672	0.123	**1.367**	0.666
	0.628	1.227	1.433	0.134	**0.889**	0.862
	0.267	0.490	0.494	0.353	**0.599**	0.441
$\text{DF}_{\text{xm,xs}}$	0.263	0.920	0.670	0.115	1.520	0.698
	0.674	1.239	1.448	0.134	0.947	0.888
	0.373	**0.592**	**0.705**	0.341	0.283	**0.459**
Ghost (G)						
TF_{v}	1.217	0.609	0.665	0.595	0.783	0.774
	0.858	0.633	0.723	**0.589**	**0.988**	0.758
	-0.535	**0.608**	-0.693	**0.896**	0.137	0.083
DF_{xm}	1.231	**0.563**	**0.629**	0.615	0.778	0.763
	0.889	**0.584**	**0.707**	0.617	0.989	**0.757**
	-0.028	0.565	0.470	0.387	0.343	0.347
DF_{xs}	1.156	0.619	0.778	**0.564**	0.786	0.781
	0.808	0.707	0.781	0.604	1.039	0.788
	0.251	0.517	0.496	0.353	**0.588**	0.441
$\text{DF}_{\text{xm,xs}}$	**1.122**	0.582	0.733	0.577	**0.775**	**0.758**
	0.771	0.691	0.754	0.616	1.029	0.772
	0.363	0.603	**0.706**	0.334	0.277	**0.457**
Lego (L)						
TF_{v}	0.925	0.806	0.514	0.614	**0.534**	0.679
	0.808	0.657	0.755	0.710	0.866	0.759
	-0.588	-0.042	-0.741	-0.212	0.193	-0.278
DF_{xm}	0.916	0.753	**0.488**	0.647	0.537	0.668
	0.827	**0.616**	0.743	0.732	**0.844**	**0.752**
	0.103	0.427	0.381	0.382	0.282	0.315
DF_{xs}	0.847	0.801	0.575	0.555	0.567	0.669
	0.749	0.663	0.789	**0.709**	0.975	0.777
	0.351	0.495	0.512	0.354	**0.511**	0.445
$\text{DF}_{\text{xm,xs}}$	**0.808**	**0.727**	0.517	**0.527**	0.555	**0.627**
	0.741	0.635	**0.736**	0.747	0.908	0.753
	0.510	**0.580**	**0.714**	**0.388**	0.215	**0.481**
Talk (T)						
TF_{v}	1.107	0.472	0.561	0.846	1.074	0.812
	0.736	0.513	0.462	0.708	**1.076**	0.699
	-0.573	0.114	-0.726	-0.020	0.213	-0.198
DF_{xm}	1.117	0.467	**0.526**	0.862	**1.057**	0.806
	0.735	**0.488**	0.440	0.719	1.081	0.693
	0.193	0.452	0.419	**0.404**	0.312	0.356
DF_{xs}	0.896	0.454	0.707	**0.771**	1.095	0.785
	0.632	0.529	0.479	**0.671**	1.124	0.687
	0.401	0.542	0.529	0.370	**0.525**	0.473
$\text{DF}_{\text{xm,xs}}$	**0.861**	**0.450**	0.617	0.794	1.082	**0.761**
	0.574	0.504	**0.419**	0.683	1.135	**0.663**
	0.585	**0.597**	**0.743**	0.403	0.229	**0.511**

Table 2. Results per trait and task. For each model, first row is MSE_{seq}, second row is MSE_{part}, and third row is Pearson Correlation also at participant level (ranging in $[-1, 1]$, closer to 1 is better). The "Avg" column depicts the average performance per row (over all the traits). Best result per task, trait, and metric in bold.

This shows that explicitly modeling cross-subject interactions helps better approximate the distributions of the traits. The former achieved the highest correlation when predicting "O" and "E", even for *Animals*, where MSE_{part} was very high. More concretely, its highest correlations were found for the latter trait ($\sim$0.7). $\text{DF}_{\text{xm,xs}}$ was also the best correlating with "C", except for *Ghost*, where it ranked second. Nevertheless, and opposite to DF_{xs}, it correlated very poorly with "N", while obtaining reasonably good results in "A" for *Lego* and *Talk*.

References

[1] 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. 1