REACT2023: The First Multiple Appropriate Facial Reaction Generation Challenge

Siyang Song*†
University of Leicester & Cambridge
Leicester & Cambridge, UK
ss1535@leicester.ac.uk

Micol Spitale*
University of Cambridge
Cambridge, United Kingdom
ms2871@cam.ac.uk

Cheng Luo
Shenzhen University
Shenzhen, China
Chengluo@cn

Germán Barquero
Universitat de Barcelona & CVC
Barcelona, Spain
germanbarquero@ub.edu

Cristina Palmero
Universitat de Barcelona & CVC
Barcelona, Spain
crpalmec7@alumnes.ub.edu

Sergio Escalera
Universitat de Barcelona & CVC
Barcelona, Spain
sergio@maia.ub.es

Michel Valstar
University of Nottingham
Nottingham, United Kingdom
michel@blueskeye.com

Tobias Baur
University of Augsburg
Augsburg, Germany
tobias.baur@uni-a.de

Fabien Ringeval
Université Grenoble Alpes
Grenoble, France
fabien.ringeval@imag.fr

Elisabeth André
University of Augsburg
Augsburg, Germany
andre@uni-a.de

Hatice Gunes
University of Cambridge
Cambridge, United Kingdom
hatice.gunes@cl.cam.ac.uk

ABSTRACT

The Multiple Appropriate Facial Reaction Generation Challenge (REACT2023) is the first competition event focused on evaluating multimedia processing and machine learning techniques for generating human-appropriate facial reactions in various dyadic interaction scenarios, with all participants competing strictly under the same conditions. The goal of the challenge is to provide the first benchmark test set for multi-modal information processing and to foster collaboration among the audio, visual, and audio-visual behaviour analysis and behaviour generation (a.k.a generative AI) communities, to compare the relative merits of the approaches to automatic appropriate facial reaction generation under different spontaneous dyadic interaction conditions. This paper presents: (i) the novelties, contributions and guidelines of the REACT2023 challenge; (ii) the dataset utilized in the challenge; and (iii) the performance of the baseline systems on the two proposed sub-challenges: Offline Multiple Appropriate Facial Reaction Generation and Online Multiple Appropriate Facial Reaction Generation, respectively. The challenge baseline code is publicly available at https://github.com/reactmultimodalchallenge/baseline_react2023.

*Both authors contributed equally to this research.
†Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612832

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Computer vision problems**; *Learning latent representations.*

KEYWORDS

facial reaction generation, challenge event, multi-modal behaviour

ACM Reference Format:
Siyang Song, Micol Spitale, Cheng Luo, Germán Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, and Hatice Gunes. 2023. REACT2023: The First Multiple Appropriate Facial Reaction Generation Challenge. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3612832

1 INTRODUCTION

The Multiple Appropriate Facial Reaction Generation Challenge (REACT2023) is the first competition aimed at the comparison of multimedia processing and machine learning methods for automatic human appropriate facial reaction generation under different dyadic interaction scenarios, with all participants competing under the same conditions.

As discussed in [24], the generation of human facial reactions in dyadic interactions poses uncertainties, as various (non-verbal) reactions may be deemed appropriate in response to specific speaker behaviours. Although some prior studies [4, 5, 10, 11, 15, 20, 21, 24] have already explored the task of automatically generating human-style facial and bodily reactions based on the conversational partner's behaviours, their models are mainly trained to reproduce a specific *real* facial reaction corresponding to the input speaker's

behaviour, which introduces challenges due to the potential divergence of non-verbal reaction labels for similar speaker behaviours at the training stage. Just a very recent work [14] presents a non-deterministic approach to generate multiple listener reactions from a speaker behaviour but without evaluating the appropriateness of the generated reactions. Please refer to [24] for a theoretical background, formulation and an in-depth discussion on multiple appropriate facial reaction generation.

The main goal of the REACT2023 challenge is to facilitate collaboration among multiple research communities representing different disciplines, in particular, the face and gesture, affective computing, multimedia, and graphics communities, as well as researchers from the psychological and social sciences specializing in expressive facial behaviours. The challenge aims to encourage **the initial development and benchmarking** of Machine Learning (ML) models capable of generating *appropriate* facial reactions in response to a given stimulus, using three state-of-the-art datasets for dyadic interaction research, namely, RECOLA [19], NOXI [6], and UDIVA [15, 16]. As part of the challenge, we provided the challenge participants with the REACT2023 Challenge Dataset, comprising segmented 30-second interaction audio-visual clips (clip pairs) from the aforementioned three datasets, annotated with challenge-specific labels indicating the appropriateness of the facial reactions. We then invited the participating groups to submit their developed / trained ML models for evaluation, which we benchmarked in terms of appropriateness, diversity, and synchrony of the generated facial reactions. The main contributions and novelties are introduced under two separated sub-challenges focusing on *online* and *offline* appropriate facial reaction generation as follows.

Offline Multiple Appropriate Facial Reaction Generation (Offline MAFRG). task focuses on generating multiple appropriate facial reaction videos from the input speaker behaviour (i.e., audio-visual clip). Specifically, this task aims to develop a machine learning model $\mathcal{H}$ that takes the entire speaker behaviour sequence $B_S^{t_1,t_2}$ as the input, and generates multiple (M) appropriate and realistic / naturalistic spatio-temporal facial reactions $p_f(b_S^{t_1,t_2})_1, \cdots, p_f(b_S^{t_1,t_2})_M$; where $p_f(b_S^{t_1,t_2})_m$ is a multi-channel time-series (consisting of AUs, facial expressions, valence and arousal state) which represent the m_{th} predicted appropriate facial reaction in response to $B_S^{t_1,t_2}$. Based on the predicted facial attributes, the challenge participants have had to generate M appropriate and realistic / naturalistic spatio-temporal facial reactions (2D face image sequences) given each input speaker behaviour.

Online Multiple Appropriate Facial Reaction Generation (Online MAFRG). task focuses on the continuous generation of facial reaction frames based on the current and previous speaker behaviours. This task aims to develop a machine learning model $\mathcal{H}$ that estimates multiple facial attributes (AUs, facial expressions, valence and arousal state) representing each appropriate facial reaction frame (i.e., $\gamma_{\text{th}} \in [t_1, t_2]$ frame) by only considering the γ_{th} frame and its previous frames of the corresponding speaker behaviour (i.e., $t_{1\text{th}}$ to γ_{th} frames in $B_S^{t_1,t_2}$), rather than taking all frames from t_1 to t_2 into account. The model is expected to gradually generate multiple multi-channel facial attribute time-series to represent all face frames of multiple appropriate and realistic / naturalistic spatio-temporal facial reactions $p_f(b_S^{t_1,t_2})_1, \cdots, p_f(b_S^{t_1,t_2})_M$, where $p_f(b_S^{t_1,t_2})_m$, where $p_f(b_S^{t_1,t_2})_m$ is a multi-channel time-series (consisting of AUs, facial expressions, valence and arousal state) representing the m_{th} predicted appropriate facial reaction in response to $B_S^{t_1,t_2}$. Based on the predicted facial attributes, the challenge participants have had to generate M appropriate and realistic / naturalistic spatio-temporal facial reactions (2D face image sequences) given each input speaker behaviour.

Both sub-challenges allowed participants to explore their own features and machine learning algorithms, along with the baseline system scripts made available in a public repository[1], to facilitate the reproducibility of the baseline facial reaction generation systems (Sec. 4). All participants were required to report their results achieved on the test partition. The REACT2023 Challenge adopts the metrics defined in [24] to evaluate the performance of the submitted approaches in terms of generated facial reactions, namely: appropriateness, diversity, realism and synchrony. Participants were required to report their results and submit the developed model(s) and checkpoints. The ranking of the submitted model competing in the Challenge relies on the two metrics: Appropriate facial reaction distance (FRDist) and facial reactions' diverseness FRDiv, for both sub-challenges.

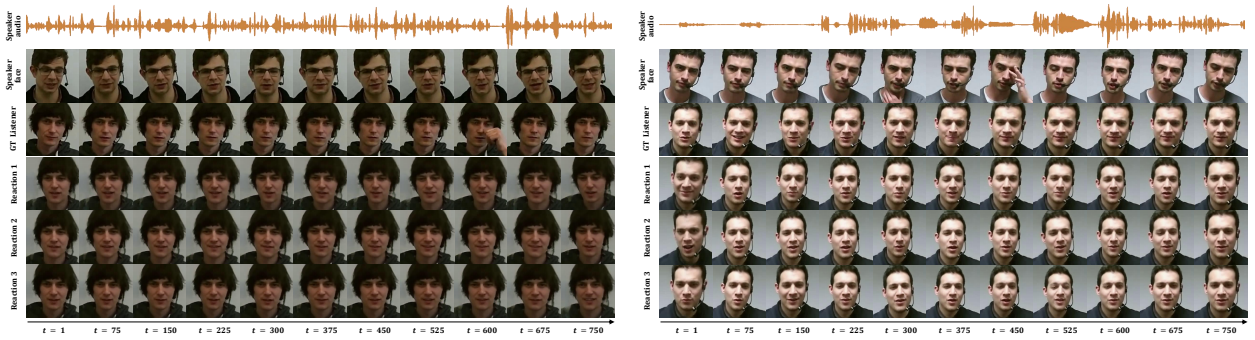
2 CHALLENGE CORPORA

The REACT2023 Challenge relies on three corpora: NoXi [6], UDIVA [16], and RECOLA [19]. Specifically, we first segmented each audio-video clip in three datasets into a 30-seconds long clip as in [1]. Then, we cleaned the dataset by selecting only the dyadic interactions with complete data of both conversational partners (where both faces were within the frame of the camera). This resulted in 8616 clips of 30 seconds each (71,8 hours of audio-video clips), specifically: 5870 clips (49 hours) from the NoXi dataset, 54 clips (0,4 hour) from the RECOLA dataset, and 2692 clips (22,4 hours) from the UDIVA dataset. We divided the datasets into training, test and validation sets. We split the datasets with a subject-independent strategy (i.e., the same subject was never included in the train and test sets).

3 EVALUATION METRICS

In this challenge, the participants were required to develop models that can generate two types of outputs for representing each facial reaction: (i) 25 facial attribute time-series (explained in Sec. 4.1); and (ii) a 2D facial image sequence. We followed [24] to comprehensively evaluate four aspects of the facial reactions generated by each participant model, including (i) **Appropriateness** based on two metrics, **FRDist**: Dynamic Time Warpping (DTW) and **FRCorr**: Concordance Correlation Coefficient (CCC); (ii) **Diversity** Metrics: **FRVar**, **FRDiv**, and **FRDvs**; (iii) **Synchrony**, where the Time Lagged Cross Correlation (TLCC) is employed, referred to as **FRSyn** in this challenge; and (iv) **Realism** of the generated facial reactions, which is assessed using the Fréchet Inception Distance (FID), denoted as **FRRea**.

[1] https://github.com/reactmultimodalchallenge/baseline_react2023



(a) Offline visualizations.

(b) Online visualizations.

Figure 1: Examples of generated multiple listener reactions to a given speaker behaviour (including the speaker's audio and face frames). These reactions are generated by an online Trans-VAE model.

4 BASELINE SYSTEMS

This section presents the baseline systems and results achieved for the REACT23 Challenge. Please refer to [23] for more details (e.g., training strategies for baselines and results on the validation set).

4.1 Behavioural features

Visual features. We followed [24] to provide three widely-used frame-level facial attribute features for each video frame as the baseline facial features. This included the occurrence of 15 facial action units (AUs), 2 facial affect – valence and arousal intensities – and the probabilities of 8 categorical facial expressions. Specifically, 15 AU occurrences (AU1, AU2, AU4, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU23, AU24, AU25 and AU26) were predicted by the state-of-the-art GraphAU model [12, 22], while the facial affect (i.e., valence and arousal intensities) and 8 facial expression probabilities (i.e., Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger and Contempt) were predicted using the approach proposed by [25].

Audio features. We also applied OpenSmile [8] to extract clip-level audio descriptors, including GEMAP and MFCC features. Consequently, we represented each speaker behaviour by combining all frame-level descriptors as a multi-channel audio-visual time-series behavioural signal.

4.2 Baseline systems

Naive baselines. we first established a set of naive baselines, namely B_Random, B_Mime, B_MeanSeq, and B_MeanFr. Specifically, B_Random randomly samples $\alpha = 10$ facial reaction sequences from a Gaussian distribution. B_Mime generates facial reactions by mimicking the corresponding speaker's facial expressions. For B_MeanSeq and B_MeanFr, the generated facial reactions are decided by the sequence- and frame-wise average reaction in the training set, respectively. Despite their simplicity, these baselines illustrate the bounds of the metrics.

Trans-VAE. The Trans-VAE baseline has a similar architecture to the TEACH proposed in [2], which consists of: (i) a **CNN encoder** that encodes the speaker facial image sequence (i.e., a short video) as a sequence-level embedding; (ii) a **transformer encoder** that first combines learned facial embeddings and audio embeddings (78-dimensional MFCC features) extracted by Torchaudio library [28], and then predicts a pair of tokens μ_{token} and σ_{token} representing the Gaussian Distribution of multiple appropriate facial reactions for the corresponding input speaker behaviour (i.e., this distribution learning strategy is inspired by [27]), based on not only the combined audio-visual embedding but also a pair of learnable tokens; and (iii) a **transformer decoder** that samples a set of representations describing an appropriate facial reaction based on the predicted distribution tokens, which include 58 3D Morphable Model (3DMM) coefficients [26] and a 25-channel emotion time-series (15 AU occurrences, 8 facial expression probabilities as well as valence and arousal intensities). Based on the learned 3DMM coefficients and the corresponding listener's portrait, FaceVerseV2 [26] is finally employed to translate the learned 3DMM coefficients to the facial reaction image sequence. Specifically, the offline Trans-VAE baseline takes the entire sequence of speaker audio-visual behaviours (i.e., 30s clip in this challenge) as the input and generates a sequence of facial reactions consisting of 750 frames. The online Trans-VAE baseline follows [13] to iteratively predict a short segment consisting of w facial reaction frames corresponding to the time $[t-w+1:t]$, where causal mask [7, 9, 13, 18] is employed to avoid future speaker behaviours to be used for the facial reaction prediction. Particularly, the τ_{th} facial reaction frame is predicted based on: (i) $t-w$ frames ($[1:t-w]$) of past speaker behaviours; (ii) $t-w$ frames ($[1:t-w]$) of previously predicted facial reactions; and (iii) τ frames ($[t-w+1:\tau]$) of the current speaker behaviour.

BeLFusion. We used BeLFusion (without behavioural disentanglement) as our second baseline [3]. It is trained in two stages. First, a variational autoencoder (VAE) is trained to learn a lower representation from visual features (e.g., AUs, facial affects, and expressions) of w frames. On the VAE's head, we include a regressor that transforms the decoded reaction to a sequence of 3DMM coefficients. Then, a latent diffusion model (LDM) learns to, given the speaker's reaction, predict the lower-dimensional representation of the listener's appropriate facial reaction. Similarly to Trans-VAE, this baseline also adopts a window-based approach where T/w reactions are predicted independently. Then, the w-frames-long reactions are stacked to build the full reaction. For the online sub-challenge, the generation of the listener's visual features for the window $[t, t+w)$ is conditioned on the past speaker's features at

Table 1: Baseline offline and online facial reaction generation results achieved on the test set.

Method	Appropriateness		Diversity			Realism	Synchrony
	FRC (↑)	FRD (↓)	FRDiv (↑)	FRVar (↑)	FRDvs (↑)	FRRea (↓)	FRSyn (↓)
GT	8.74	0.00	0.0000	0.0723	0.2474	47.50	47.72
B_Random	0.04	237.62	0.1667	0.0833	0.1667	-	43.99
B_Mime	0.38	92.95	0.0000	0.0723	0.2474	-	38.66
B_MeanSeq	0.01	98.39	0.0000	0.0000	0.0000	-	45.39
B_MeanFr	0.00	99.04	0.0000	0.0000	0.0000	-	49.00
Offline Results							
Trans-VAE w/o visual modality	0.08	99.03	0.0229	0.0029	0.0255	65.18	44.47
Trans-VAE w/o audio modality	0.09	96.83	0.0088	0.0013	0.0094	63.77	45.24
Trans-VAE	0.10	98.48	0.0242	0.0040	0.0263	69.24	44.88
BeLFusion (k=1)	0.12	90.21	0.0085	0.0056	0.0103	-	44.95
BeLFusion (k=10)	0.13	89.84	0.0137	0.0078	0.0149	-	45.02
BeLFusion (k=10) + Binarized AUs	0.12	92.58	0.0322	0.0170	0.0337	-	49.00
Online Results							
Trans-VAE w/o visual modality	0.13	134.78	0.1121	0.0581	0.1166	69.20	44.24
Trans-VAE w/o audio modality	0.13	134.77	0.1087	0.0564	0.1095	74.54	44.33
Trans-VAE	0.13	135.57	0.1168	0.0604	0.1202	71.15	44.31
BeLFusion (k=1)	0.12	89.56	0.0086	0.0058	0.0103	-	45.09
BeLFusion (k=10)	0.13	89.42	0.0133	0.0077	0.0143	-	44.80
BeLFusion (k=10) + Binarized AUs	0.12	92.13	0.0306	0.0164	0.0317	-	49.00

$[t-w, t)$. It predicts all zeroes for segment $[0, w)$. For the offline sub-challenge, such generation is conditioned on the speaker's features on the same time period: $[t, t+w)$. The LDM's loss is the average of the MSE in the latent space and the MSE in the reconstructed space. The denoising chain has 10 steps, and every denoising step is implemented with a sequence of residual MLPs as in [17].

4.3 Baseline results

Offline facial reaction generation sub-challenge: Table 1 shows that both baselines can generate facial reactions that positively correlate to the appropriate real facial reactions, where BeLFusion outperforms Trans-VAE in terms of the distance between the predictions and real appropriate facial reactions (FRD), as well as intra-sequence diversity (FRVar), but Trans-VAE achieved better intra- and inter-subject diversities (FRDiv and FRDvs). Moreover, the results achieved by Trans-VAE suggest that both visual and audio modalities positively contribute to the diversity of generated facial reactions (FRDiv, FRVar, and FRDvs). Randomly sampled facial reaction (i.e., B_Random) are diverse but not appropriate (in terms of FRC and FRD), whereas deterministic baselines (i.e., B_Mime, B_MeanSeq and B_MeanFr) achieved better appropriateness but much lower diversity. Unlike above baselines, deep learned probabilistic models (i.e., Trans-VAE and BeLFusion) can make a trade-off between the these two, suggesting that they can generate multiple different but appropriate facial reactions.

Online facial reaction generation sub-challenge: As demonstrated in Table 1, the results confirmed that both baselines can generate real-time facial reactions that are positively correlated with the appropriate face reactions. Also, Trans-VAE outperforms the BeLFusion in terms of diversity (FRDiv, FRVar, and FRDvs), synchrony (FRSync), and while BeLFusion achieved better results in terms of DTW distances (FRD). Similarly to the offline task, the randomly sampled facial reactions (i.e., B_Random) are diverse but not appropriate, while the deterministic naive approaches (i.e., B_Mime, B_MeanSeq and B_MeanFR) achieved better results in terms of appropriateness but not diversity. Again, the proposed Trans-VAE and BeLFusion baselines can have a better trade-off between appropriateness and diversity. In this sub-challenge, the differences are magnified for the four metrics.

The results achieved for both sub-challenges suggest the existence of a trade-off between the appropriateness and diversity of the generated facial reactions, which is observed when binarizing the predicted AUs: while the FRD worsens, all diversity metrics are doubled. Compared to the offline setting, online generation is more challenging and causes jitters and inconsistency between windows, which are the main reasons for the decrease in Realism (i.e., FRRea metric). Figure 1 visualises the 2D facial reactions generated by Trans-VAE baseline under both offline and online scenarios.

5 PARTICIPATION AND CONCLUSION

This paper introduced REACT2023, the First Multiple Appropriate Facial Reaction Generation Challenge, which provides the very first attempt to bring together researchers contributing to a new challenging research direction at the intersection of face and gesture research, affective computing, and generative AI. The call for participation attracted registration of 11 teams from 6 countries, with 10 teams participating in the Offline and Online MAFRG sub-challenges, respectively. The top 3 teams have successfully submitted valid models, results and papers for the challenge, with each paper submission being assigned two reviewers. We hope that both the challenge data and code, as well as the systems and results of the competing teams, will serve as a valuable stepping stone for researchers and practitioners interested in the area of generative AI and automatic facial reaction generation.

Acknowledgements. Funding: M. Spitale and H. Gunes are supported by the EPSRC/UKRI ref. EP/R030782/1. This work has been supported by the Spanish project PID2019-105093GB-I00 and ICREA under the ICREA Academia programme. **Data access:** Data can be accessed upon request following the terms and conditions of the dataset owners.

REFERENCES

[1] Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin* 111, 2 (1992), 256.

[2] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEACH: Temporal Action Composition for 3D Humans. In *International Conference on 3D Vision 2022*.

[3] German Barquero, Sergio Escalera, and Cristina Palmero. 2022. BeLFusion: Latent Diffusion for Behavior-Driven Human Motion Prediction. *arXiv preprint arXiv:2211.14304* (2022).

[4] German Barquero, Johnny Núnez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. 2022. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 139–178.

[5] German Barquero, Johnny Núñez, Zhen Xu, Sergio Escalera, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. 2022. Comparison of Spatio-Temporal Models for Human Motion and Pose Forecasting in Face-to-Face Interaction Scenarios. In *Understanding Social Behavior in Dyadic and Small Group Interactions (Proceedings of Machine Learning Research, Vol. 173)*, Cristina Palmero, Julio C. S. Jacques Junior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera (Eds.). PMLR, 107–138.

[6] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350–359.

[7] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems* 32 (2019).

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[9] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18770–18780.

[10] Yuchi Huang and Saad M Khan. 2017. Dyadgan: Generating facial expressions in dyadic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 11–18.

[11] Yuchi Huang and Saad M Khan. 2018. Generating Photorealistic Facial Expressions in Dyadic Interactions.. In *BMVC*. 201.

[12] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 1239–1246.

[13] Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Linlin Shen, and Hatice Gunes. 2023. ReactFace: Multiple Appropriate Facial Reaction Generation in Dyadic Interactions. *arXiv preprint arXiv:2305.15748* (2023).

[14] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to Listen: Modeling Non-Deterministic Dyadic Facial Motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20395–20405.

[15] Cristina Palmero, German Barquero, Julio C. S. Jacques Junior, Albert Clapés, Johnny Núñez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, David Gallardo-Pujol, Georgina Guilera, David Leiva, Feng Han, Xiaoxue Feng, Jennifer He, Wei-Wei Tu, Thomas B. Moeslund, Isabelle Guyon, and Sergio Escalera. 2022. ChaLearn LAP Challenges on Self-Reported Personality Recognition and Non-Verbal Behavior Forecasting During Social Dyadic Interactions: Dataset, Design, and Results. In *Understanding Social Behavior in Dyadic and Small Group Interactions (Proceedings of Machine Learning Research, Vol. 173)*, Cristina Palmero, Julio C. S. Jacques Junior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera (Eds.). 4–52.

[16] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. 2021. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1–12.

[17] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10619–10629.

[18] Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409* (2021).

[19] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. 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)*. IEEE, 1–8.

[20] Zilong Shao, Siyang Song, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. 2021. Personality recognition by modelling person-specific cognitive processes using graph representation. In *proceedings of the 29th ACM international conference on multimedia*. 357–366.

[21] Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. 2022. Learning Person-specific Cognition from Facial Reactions for Automatic Personality Recognition. *IEEE Transactions on Affective Computing* (2022).

[22] Siyang Song, Yuxin Song, Cheng Luo, Zhiyuan Song, Selim Kuzucu, Xi Jia, Zhijiang Guo, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. GRATIS: Deep Learning Graph Representation with Task-specific Topology and Multi-dimensional Edge Features. *arXiv preprint arXiv:2211.12482* (2022).

[23] Siyang Song, Micol Spitale, Cheng Luo, German Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, et al. 2023. REACT2023: the first Multi-modal Multiple Appropriate Facial Reaction Generation Challenge. *arXiv preprint arXiv:2306.06583* (2023).

[24] Siyang Song, Micol Spitale, Yiming Luo, Batuhan Bal, and Hatice Gunes. 2023. Multiple Appropriate Facial Reaction Generation in Dyadic Interaction Settings: What, Why and How? *arXiv e-prints* (2023), arXiv–2302.

[25] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 3, 1 (2021), 42–50.

[26] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20333–20342.

[27] Tong Xu, Micol Spitale, Hao Tang, Lu Liu, Hatice Gunes, and Siyang Song. 2023. Reversible Graph Neural Network-based Reaction Distribution Learning for Multiple Appropriate Facial Reactions Generation. *arXiv preprint arXiv:2305.15270* (2023).

[28] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Artyom Astafurov, Caroline Chen, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z Yang, et al. 2022. Torchaudio: Building blocks for audio and speech processing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6982–6986.