PerReactor: Offline Personalised Multiple Appropriate Facial Reaction Generation

Hengde Zhu[1], Xiangyu Kong[2,3], Weicheng Xie[4], Xin Huang[5], Xilin He[4], Lu Liu[2], Linlin Shen[4], Wei Zhang[6,3], Hatice Gunes[7], Siyang Song[2*]

[1]School of Computing and Mathematical Sciences, University of Leicester, Leicester, UK
[2]Department of Computer Science, University of Exeter, Exeter, UK
[3]Affect AI, Anhui, China
[4]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China
[5]School of Electronic Science and Engineering (School of Microelectronics), South China Normal University, Foshan, China
[6]School of Software Technology, Zhejiang University, Hangzhou, China
[7]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK.
hz204@le.ac.uk, {xk219, l.liu3}@exeter.ac.uk, {wcxie, llshen}@szu.edu.cn, huangxin@m.scnu.edu.cn, 2020152115@email.szu.edu.cn, cstzhangwei@zju.edu.cn, {hg410, ss2796}@cam.ac.uk

Abstract

In dyadic human-human interactions, individuals may express multiple different facial reactions in response to the same/similar behaviours expressed by their conversational partners depending on their personalised behaviour patterns. As a result, frequently-employed reconstruction loss-based strategies lead the training of previous automatic facial reaction generation (FRG) models to not only suffer from the 'one-to-many mapping' problem, but also fail to comprehensively consider the quality of the generated facial reactions. Besides, none of them considered such personalised behaviour patterns in generating facial reactions. In this paper, we propose the first adversarial FRG model training strategy which jointly learns appropriateness and realism discriminators to provide comprehensive task-specific supervision for training the target facial reaction generators, and reformulates the 'one-to-many (facial reactions) mapping' training problem as a 'one-to-one (distribution) mapping' training task, i.e., the FRG model is trained to output a distribution representing multiple appropriate/plausible facial reaction from each input human behaviour. In addition, our approach also serves as the first offline FRG approach that considers personalised behaviour patterns in generating of target individuals' facial reactions. Experiments show that our PerReactor not only largely outperformed all existing offline solutions for generating appropriate, diverse and realistic facial reactions, but also is the first offline approach that can effectively generate personalised appropriate facial reactions.

Code — https://github.com/AffectAI/PerReactor

Introduction

In human-human dyadic interactions, facial reaction is defined as human non-verbal facial behaviours expressed in response to behaviours expressed by their conversational partners, which plays a key role for people to convey their intentions and emotional states (Sagliano et al. 2022). Therefore, developing reliable automatic human-style facial reaction generation (FRG) systems would promote humanoid virtual agents/robots with the capability of reacting human interpretable emotional behaviours for various human-computer interaction applications (Tang et al. 2023). While FRG is a challenging 'one-to-many mapping' problem (Song et al. 2023b), where various human facial reactions could be triggered by the same/similar human behaviours (called speaker behaviour) (Mehrabian and Russell 1974; Pandita, Mishra, and Chib 2021), it is almost impossible to develop a generic model that can accurately reproduce real facial reactions expressed by different individuals (called listeners) under various contexts. Specifically, previous studies found that gener-

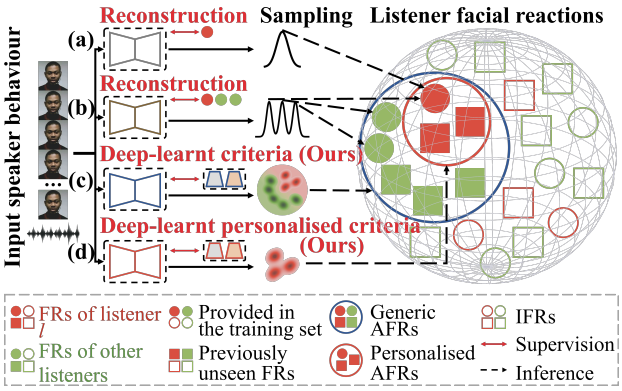


Figure 1: Comparison between the distributions learnt by PerReactor and existing MAFRG approaches, which pair the input speaker behaviour with: (a) the corresponding 'GT' real appropriate facial reaction (AFR); (b) a finite set of generic real AFRs (expressed by different listeners) defined in the training set; (c) a infinite set of AFRs including generic real AFRs defined in the training set and previously unseen generic AFRs; and (d) a infinite set of personalised real AFRs (expressed by the target listener) provided in the training set and previously unseen personalised AFRs.

*Corresponding author.

ation of a human listener's facial reaction not only depends on external stimulus (e.g., speaker behaviours and environments), but also is influenced by the listener's personalised factors (Song et al. 2022; Pandita, Mishra, and Chib 2021).

However, early FRG approaches (Huang and Khan 2018b; Nojavanasghari, Huang, and Khan 2018; Woo, Pelachaud, and Achard 2021) were mainly developed for the purpose of reproducing a single real facial reaction expressed by the corresponding listener in response to the input speaker behaviour (*called 'GT' real facial reaction*). These approaches directly pair each speaker behaviour (input) with its 'GT' real facial reaction (label) for simple reconstruction loss-based training, enforcing the well-trained models to reproduce a single 'GT' real facial reaction from each input. This means that the same/similar inputs (speaker behaviours) might be paired with different labels (real facial reactions) in the training phase due to the 'one-to-many mapping' nature of human facial reactions, leading models to learn unreliable hypothesis. Consequently, a new research topic - Multiple Appropriate Facial Reaction Generation (MAFRG) - has been proposed (Song et al. 2023b,a), which only require machine learning (ML) models to output multiple appropriate (plausible) human facial reactions (AFRs) rather than the 'GT' facial reaction in response to each input speaker behaviour. Consequently, recent MAFRG approaches frequently reformulate this 'one-to-many mapping' training problem as a 'one-to-one mapping' task by pairing each input speaker behaviour with a specific distribution (e.g., Gaussian Mixture Model (GMM) (Xu et al. 2023; Nguyen et al. 2024a), Gaussian distribution (Luo et al. 2024; Nguyen et al. 2024b) and codebook (Liang et al. 2023; Liu et al. 2024)) representing multiple AFRs for responding to it, where reversible network, transformer and diffusion models have been proposed to learn AFR distributions.

These MAFRG models are trained by either still directly pairing each input speaker behaviour with its 'GT' real facial reaction (Liang et al. 2023; Yu et al. 2023), or minimising the reconstruction loss (e.g., Mean Square Error (MSE)) between each generated facial reaction with its corresponding multiple real AFRs provided in the training set (Xu et al. 2023; Luo et al. 2024; Zhu et al. 2024). As illustrated in Fig. 1, these training strategies typically suffer from three problems: **(i)** each learnt distribution only cover one or a finite set of real AFRs defined by the training set; **(ii)** a single manually-defined loss (e.g., MSE) cannot comprehensively measure the appropriateness and realism of the generated facial reactions; and **(iii)** the optimal component number of their manually-defined distribution representations (e.g., codebook and GMM) is hard to be determined. Besides, while facial reactions expressed by different listeners are largely influenced by their personalised factors, previous MAFRG approaches can only generate person-agnostic (generic) AFRs without considering behaviour differences caused by such factors (**Problem (iv)**). Consequently, the quality (e.g., appropriateness and realism) of the facial reactions generated by existing MAFRG models is limited.

In this paper, we propose a the first offline personalised MAFRG (PMAFRG) approach (called **PerReactor**) which deep learns comprehensive appropriateness and realism criteria via two appropriateness discriminators and a realism discriminator. These enforce the generator to not only predict a comprehensive distribution that covers infinitely diverse AFRs from each input speaker behaviour, i.e., including AFRs that are not defined in the training set, but also generate high-quality AFRs by comprehensively interpreting and considering critical appropriateness and realism aspects that may not be considered by manually-defined loss functions (**addressing/avoiding Problem (i)-(iii)**). In addition, our PerReactor specifically learns personalised behaviour patterns for each target listener to generate more appropriate personalised AFRs (**addressing Problem (iv)**). The main contributions and novelties are summarised below:

- We propose the first Generative Adversarial Network (GAN)-based multiple appropriate facial reaction distribution learning strategy to address the 'one-to-many mapping' problem occurring in MAFRG models' training, allowing previously unseen but appropriate facial reactions to be generated in response to the given speaker behaviour. As compared in Fig. 1, it is the only training strategy that comprehensively considers generic AFRs, personalised AFRs, and inappropriate facial reactions (IFRs) for evaluating the generated facial reactions.
- We propose a novel strategy to disentangle personalised behaviour patterns from an available historical facial video of the target listener, which are further applied to generate personalised AFRs in response to the given speaker behaviour. This is the first MAFRG approach that specifically integrates personalised behaviour patterns for generating personalised AFRs.
- Our approach brings significant appropriateness (FRCorr) improvements to previous state-of-the-art (SOTA) method with more than 71%, with diversity (FRDiv and FRDvs) and realism (FRRea) also being the best among all competitors for both MAFRG and PMAFRG tasks.

Related Work

Early facial reaction generation approaches have been frequently developed for the purpose of reproducing the 'GT' real facial reaction expressed by the corresponding listener under a specific context. For example, Huang et al. (Huang and Khan 2017, 2018b) trained a conditional Generative Adversarial Network (GAN) (Mirza and Osindero 2014) to generate real facial reaction sketch from the input speaker facial action units (AUs). Similar frameworks (Huang and Khan 2018a; Song et al. 2022; Shao et al. 2021; Nojavanasghari, Huang, and Khan 2018; Woo, Pelachaud, and Achard 2021; Woo et al. 2023; Ng et al. 2022) have been extended for the same purpose, with more modalities (e.g., audio or language) are employed as the input. However, the training processes of such deterministic approaches usually face the ill-posed 'one-to-many mapping' problem, making their model to convergence for generating mean facial reactions.

Two recent pioneering non-deterministic MAFRG approaches (Xu et al. 2023; Luo et al. 2024) proposed to reformulate the 'one-to-many mapping' problem to a 'one-to-one mapping' task by learning a distribution from each input speaker behaviour, which describes all of its AFRs. Fol-

lowed the recently organized REACT2023 and REACT2024 challenge (Song et al. 2023a, 2024), Liang et al. (Liang et al. 2023) proposed to learn a codebook to represent all facial reactions in response to different speaker behaviours. The BEAMER (Hoque et al. 2023) directly compares semantic similarity between the speaker behaviour and generated facial reaction representations to decide their appropriateness. In (Yu et al. 2023), a diffusion-based model generates facial reaction frames conditioned on both corresponding speaker behaviour and previously predicted facial reaction frames. Meanwhile, the winner (Dam et al. 2024) of REACT2024 challenge employed the Finite Scalar Quantization strategy (Mentzer et al. 2024) to tokenize speaker facial behaviours, which are further incorporated with speaker audio features for online facial reaction prediction. Liu et al. (Liu et al. 2024) addressed the 'one-to-many mapping' task using a K-way categorical latent variable with each of its value indicating a reaction intent of one response. Nguyen et al. (Nguyen et al. 2024a) extended the Trans-VAE baseline (Song et al. 2023a) by incorporating the audio-visual features with emotional attributes for better speaker behaviour understanding.

Task Definition

Offline Personalised Multiple Appropriate Facial Reaction Generation (PMAFRG) task: In dyadic interaction scenarios, the offline PMAFRG task aims to learn a ML model $\mathcal{H}^{\textbf{PMAFRG}}$ that can generate multiple (M) personalised AFRs $\mathbb{R}_i^l = \{\hat{R}_i^l(1), \cdots, \hat{R}_i^l(M)\}$ (i.e., face videos) from each speaker behaviour B_i^s (e.g., audio-visual clips):

$$\hat{R}_i^l(m) \in \mathbb{R}_i^l = \mathcal{H}^{\textbf{PMAFRG}}\left(B_i^s, p^l\right), \tag{1}$$

where p^l denotes the personalised behaviour patterns of the target listener l. Here, each generated facial reaction $\hat{R}_i^l(m) \in \mathbb{R}_i^l$ is expected to be similar to at least one **personalised** real AFR $F_i^l(n)$ expressed by the listener l as:

$$\hat{R}_i^l(m) \approx F_i^l(n) \in \mathbb{F}_i^l, \tag{2}$$

where $\mathbb{F}_i^l$ denotes all real AFRs of the B_i^s in the training set, which are expressed by the target listener l. The PMAFRG task differs from the **offline MAFRG task** that only requires each generated facial reaction $\hat{R}_i^G(m) \in \mathbb{R}_i^G$ to be similar to one of the generic real AFRs $\mathbb{F}_i^G$ in response to the B_i^s, which can be expressed by different listeners rather than listener l only. Please refer to (Song et al. 2023b) for detailed definition of the MAFRG task.

The Proposed PerReactor

Pipeline: As shown in Fig. 2, given a speaker audio-visual behaviour $B_i^s = \{A_i^s, F_i^s\}$ consisting of T frames and a historical face video F_h^l previously expressed by the target listener l, our PerReactor starts with a **Generic Facial Reaction Generation (GFRG)** module whose Generic AFR Generator **GAG** learns a distribution $\mathcal{Q}_i^G$ representing a set of ***generic AFRs*** $\mathbb{R}_i^G$ in response to the B_i^s. As a result, multiple different AFRs $\mathbb{R}_i^G = \{\hat{R}_i^G(1), \ldots, \hat{R}_i^G(M)\}$ can be sampled from the $\mathcal{Q}_i^G$ as:

$$\mathbb{R}_i^G = \text{GAG}\left(B_i^s\right). \tag{3}$$

Then, a **Personalised Behaviour Pattern Modelling (PBPM)** module is introduced to further generate ***personalised AFRs***. Specifically, its personalised AFR Generator **PAG** first captures personalised behaviour patterns p^l from historical personalised facial behaviours F_h^l (i.e., a face video) of the target listener l, and then integrates p^l into the generic AFR distribution $\mathcal{Q}_i^G$ to generate multiple personalised AFRs $\mathbb{R}_i^l$ as:

$$\mathbb{R}_i^l = \text{PAG}\left(\mathcal{Q}_i^G, F_h^l\right). \tag{4}$$

Generic Appropriate Facial Reaction Generation

The GFRG module consists of a generator **GAG** aiming to output multiple (M) generic AFRs $\mathbb{R}_i^G = \{\hat{R}_i^G(1), \cdots, \hat{R}_i^G(M)\}$ in response to each input speaker behaviour B_i^s, as well as two discriminators $\{D_A^G, D_R\}$ to provide a comprehensive supervision for training **GAG**. Specifically, the D_A^G evaluates the appropriateness of the generated AFRs $\mathbb{R}_i^G$ in response to B_i^s, while D_R evaluates whether each $\hat{R}_i^G(m)$ is a realistic human facial behaviour.

Generator: Given an input audio-visual speaker behaviour $B_i^s = \{A_i^s, F_i^s\}$, the video encoder $\textbf{Enc}_\text{f}$ and audio encoder $\textbf{Enc}_\text{a}$ of the **GAG** first encode it as a pair of embeddings, which are concatenated and processed by a linear layer to represent B_i^s in a latent space as E_i^{Ltn}. Then, we employ the same Trans-VAE architecture (Song et al. 2023a) to learn a distribution $\mathcal{Q}_i^G \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ representing multiple generic AFRs of B_i^s based on $\mathbb{E}_i^{\text{Ltn}}$ and two learnable distribution tokens. This way, a latent generic AFR embedding $E_i^G(m) \in \mathbb{E}_i^G$ can be sampled from $\mathcal{Q}_i^G$ as:

$$E_i^G(m) \sim \mathcal{Q}_i^G = \text{GAG}\left(B_i^s\right) \tag{5}$$

where $\mathbb{E}_i^G = \{E_i^G(1), \cdots, E_i^G(M)\}$ denote M sampled AFR embeddings for generating M generic AFRs. Then, a transformer decoder decodes each $E_i^G(m)$ along with positional encodings (Vaswani et al. 2017) to produce a AFR embedding sequence $\tilde{R}_i^G(m) = \{\tilde{r}_i^G(m,t) | t = 1, \cdots, T\}$ consisting of T frame-level embeddings, based on which an AFR $\hat{R}_i^G(m)$ is produced. This way, multiple diverse generic AFRs $\mathbb{R}_i^G = \{\hat{R}_i^G(1), \ldots, \hat{R}_i^G(M)\}$ can be obtained by repeatedly sampling from the learnt generic AFR distribution $\mathcal{Q}_i^G$. In this paper, we represent each generated AFR by a multivariate facial behaviour attribute time-series (i.e., AUs, facial expressions, affects and 3DMM coefficients).

Discriminators: To enforce the GAG learning a distribution that can comprehensively represent diverse and realistic AFRs of each input speaker behaviour, we propose two discriminators: (i) **realism discriminator** D_R evaluates whether the t_{th} facial reaction frame $r(m,t) \in R(m)$ is synthesised or expressed by a human being (binary classification); and (ii) **appropriateness discriminator** D_A^G evaluates whether the given facial reaction $R(m)$ is appropriate for responding to the paired speaker behaviour B_i^s. Since the appropriateness of each facial reaction is conditioned on the given speaker behaviour, the D_A^G takes both B_i^s and a facial reaction $R(m) = \{r(m,t) | t = 1, \cdots, T\}$ (i.e., either real or synthesised) as the input. It maps $R(m)$ and B_i^s into a latent

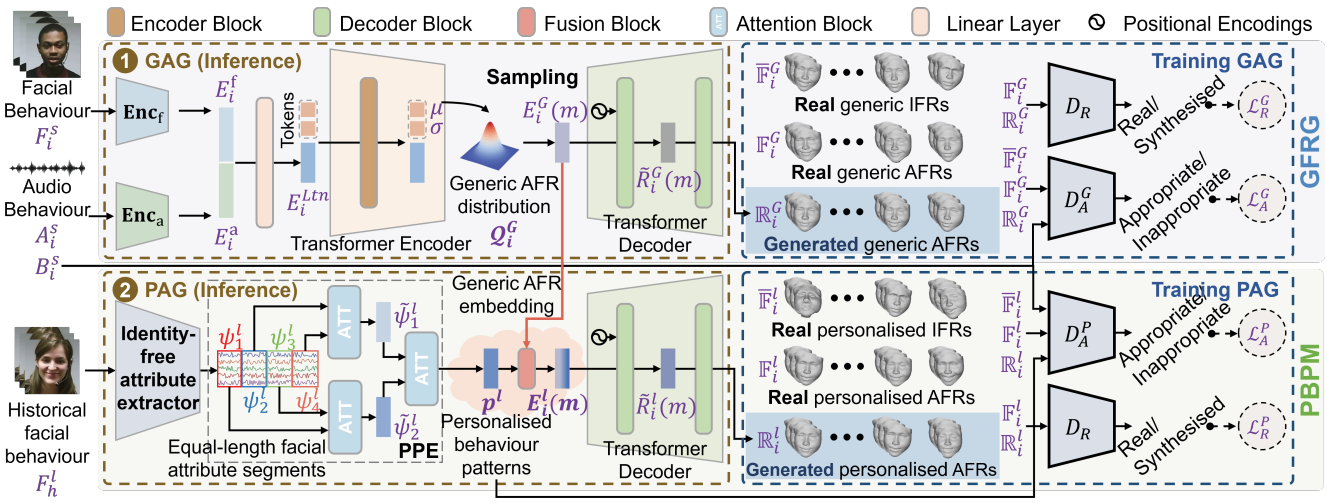


Figure 2: PerReactor pipeline. **Inference:** (1) the generator **GAG** (**GFRG** module) takes an audio-visual speaker behaviour $B_i^s = \{F_i^s, A_i^s\}$ and predicts a distribution $\mathcal{Q}_i^G$ representing various generic AFRs in response to it. Then, a generic AFR embedding $E_i^G(m)$ is sampled from $\mathcal{Q}_i^G$, based on which ***a generic AFR*** $\hat{R}_i^G(m)$ is generated (multiple generic AFRs $\mathbb{R}_i^G$ can be repeatedly sampled). (2) the generator **PAG** (**PBPM** module) first disentangles the personalised behaviour pattern p^l from the target listener l's historical facial behaviour F_h^l, and integrates it into the obtained generic AFR embedding $E_i^G(m)$ to generate a personalised AFR embedding $E_i^l(m)$, and finally decodes ***a personalised AFR*** $\hat{R}_i^l(m) \in \mathbb{R}_i^l$. **Training:** the learnt discriminator D_R comprehensively evaluates the realism of the generated generic/personalised AFRs, while D_A^G and D_A^P provide deep-learnt task-specific criteria to independently evaluate the appropriateness of the generated generic and personalised AFRs.

space via a pair of learnable fully-connected (FC) layers, and then applies a cross-attention to model their interaction. This results in a latent appropriateness representation $z_{i,m}$ describing the appropriateness of the facial reaction $R(m)$ for responding to B_i^s. Finally, the obtained $z_{i,m}$ is fed to a classifier composed of a Long Short-Term Memory (LSTM) network and a FC layer to determine weather $R(m)$ is appropriate in response to B_i^s. These processes are defined as:

$$D_R(r(m,t)) = \begin{cases} 1, & r(m,t) \text{ is expressed by human} \\ 0, & r(m,t) \text{ is synthesised} \end{cases}$$
$$D_A^G(R(m), B_i^s) = \begin{cases} 1, & R(m) \text{ is an AFR of } B_i^s \\ 0, & R(m) \text{ is not an AFR of } B_i^s \end{cases} \quad (6)$$

Personalised Behaviour Pattern Modelling

While the well-trained **GAG** can only predict generic AFRs from the given speaker behaviour B_i^s, our **Personalised Behaviour Pattern Modelling (PBPM)** module additionally considers personalised behaviour patterns of each target listener, aiming to generate more appropriate, realistic and personalised AFRs. Our **PBPM** also consists of a personalised AFR generator **PAG** and two discriminators D_R and D_A^P. The **PAG** starts with learning personalised behaviour patterns p^l from a historical facial behaviour F_h^l previously expressed by the target listener l, which is then integrated into a generic AFR embedding $E_i^G(m)$ sampled by **GAG** to generate a personalised AFR $\hat{R}_i^l(m)$.

Learning Personalised Behaviour Pattern. Our approach hypothesises that a historical facial behaviour F_h^l is potentially made up of the listener's: (i) identity; (ii) the required personalised behaviour patterns; and (iii) temporal emotions caused by contextual factors (e.g., past experiences). Importantly, we assume that the identity and personalised behaviour patterns remain consistent in each listener's historical behaviour, while temporal emotions could be changed over time (i.e., emotions expressed in non-adjacent facial behaviour segments can be different). To remove identity information, a set of identity-free facial attributes ψ_h^l (e.g., AUs) are first extracted from each frame of F_h^l. Then, the **PAG** removes temporal emotions from ψ_h^l based on a Personalised Pattern Encoder **PPE**. This encoder first splits ψ_h^l into 2^K equal-length segments $\{\psi_1^l, \cdots, \psi_{2^K}^l\}$ ($K \geq 2$ and $K = 2$ used in this paper). Subsequently, K pairs of non-adjacent segments $(\psi_k^l, \psi_{k+2^{K-1}}^l)$ with $k = 1, 2, \cdots, 2^{K-1}$ are obtained. To emphasise their commonly shared information (i.e., personalised behaviour patterns p^l) while suppressing their differences (i.e., temporal emotions), each pair of segments is processed by a shared cross-attention layer. This results in 2^{K-1} attention representations $\{\tilde{\psi}_1^l, \ldots, \tilde{\psi}_{2^{K-1}}^l\}$, which are further grouped as 2^{K-2} non-adjacent pairs and processed by the same cross-attention layer. By repeating the above process K times, an attention representation p^l emphasising identity and emotion-free personalised behaviour patterns of the target listener l is obtained. The pseudocode of this process is provided in the Supplementary Material.

Personalised Facial Reaction Generation. The **PAG** then integrates the obtained personalised behaviour patterns p^l into a sampled generic AFR embedding $E_i^G(m)$ for producing a personalised AFR $\hat{R}_i^l(m)$ by computing the

weighted sum of p^l and $E_i^G(m)$ via a factor β, resulting in a personalised latent facial reaction embedding $E_i^l(m)$ as:

$$E_i^l(m) = \beta \cdot p^l + E_i^G(m), \tag{7}$$

Then, the decoder of the **PAG**, which has a similar architecture to the decoder of the **GAG**, further maps $E_i^l(m)$ as a personalised AFR $\hat{R}_i^l(m) = \{\hat{r}_i^l(m,t)|t = 1,\cdots,T\}$ consisting of T frames. This way, multiple diverse personalised AFRs $\mathbb{R}_i^l$ for the listener l can be generated by integrating p^l into repeatedly sampled generic AFR embeddings.

Discriminators. The D_R from **GFRG** module is re-employed here to evaluate the realism of the facial reactions generated by **PAG**. Meanwhile, an appropriateness discriminator D_A^P that has a similar architecture as D_A^G is introduced to evaluate whether the input $R(m)$ is a personalised AFR of the target listener l in response to the input speaker behaviour B_i^s, which considers not only B_i^s and $R(m)$ but also the corresponding personalised behaviour patterns p^l as:

$$D_A^P(B_i^s, R(m)|p^l) = \begin{cases} 1, & R(m) \text{ is a personalised AFR} \\ 0, & R(m) \text{ is not a personalised AFR} \end{cases} \tag{8}$$

Training and Regularisation Strategies

We train **GAG** and **PAG** in an two-stage adversarial manner. Different from existing reconstruction loss-based training strategies that enforce each generated AFR to be similar to a real generic AFR in terms of a single metric (e.g., L2 distance), *our strategy enforces the generators to output AFRs meeting the comprehensive appropriateness and realism criteria defined by deep-learnt discriminators*.

Training Generic Generator GAG: As shown in Fig. 2, this stage trains the generic AFR generator **GAG** by pitting it and two discriminators (D_R and D_A^G) against each other. Specifically, the **GAG** and realism discriminator D_R are jointly trained by optimising an adversarial loss as:

$$\begin{aligned} \mathcal{L}_R^G = & \underbrace{\mathbb{E}_{f_i^G(m,t)}\left[\log\left(D_R\left(f_i^G(m,t)\right)\right)\right]}_{\text{Evaluating the realism over each frame of a real AFR}} \\ & + \underbrace{\mathbb{E}_{B_i^s}\left[\log\left(1 - D_R\left(\text{GAG}\left(B_i^s\right)\right)\right)\right]}_{\text{Evaluating the realism over frames of a generated facial reaction}}, \end{aligned} \tag{9}$$

where $f_i^G(m,t)$ is the t_{th} frame of a generic real AFR $F_i^G(m)$ (provided in the training set) for responding to B_i^s, while $\text{GAG}\left(B_i^s\right)$ denote an AFR generated by **GAG**. With this loss, the **GAG** is enforced to generate realistic face frames expressed by human beings, thereby 'deceiving' D_R that is trained to distinguish between synthesised and real face frames. Meanwhile, the **GAG** and D_A^G are also jointly trained by optimising an adversarial loss $\mathcal{L}_A^G$ as:

$$\begin{aligned} \mathcal{L}_A^G = & \underbrace{\mathbb{E}_{F_i^G(m)}\left[\log\left(D_A^G\left(F_i^G(m), B_i^s\right)\right)\right]}_{\text{Evaluating the appropriateness of a real AFR}} \\ & + \underbrace{\mathbb{E}_{B_i^s}\left[\log\left(1 - D_A^G\left(\text{GAG}\left(B_i^s\right), B_i^s\right)\right)\right]}_{\text{Evaluating the appropriateness of a generated facial reaction}} \\ & + \underbrace{\mathbb{E}_{\bar{F}_i^G(m)}\left[\log\left(1 - D_A^G\left(\bar{F}_i^G(m), B_i^s\right)\right)\right]}_{\text{Evaluating the appropriateness of a real IFR}}, \end{aligned} \tag{10}$$

where $\bar{F}_i^G(m)$ denotes a real but *inappropriate* facial reaction (IFR) for responding to B_i^s. This loss enforces the **GAG** to generate AFRs that can deceive D_A^G trained for distinguishing real AFRs $\mathbb{F}_i^G = \{F_i^G(1),\cdots,F_i^G(M)\}$ from not only synthesised AFRs/IFRs $\text{GAG}\left(B_i^s\right)$ but also real IFRs $\bar{\mathbb{F}}_i = \{\bar{F}_i^G(1),\cdots,\bar{F}_i^G(M)\}$ for responding to B_i^s.

In addition, two regularisation losses are employed to ensure that **GAG** can generate smooth and diverse facial reactions, including: a **(i) Motion Smoothness Loss** $\mathcal{L}_{\text{mot}}$ that evaluates sharp changes between adjacent frames of each generated AFR; and a **(ii) Diversity Loss** $\mathcal{L}_{\text{div}}$ which encourages the diversity of the M AFRs $\mathbb{R}_i = \{\hat{R}_i(1),\cdots,\hat{R}_i(M)\}$ generated for responding to the same speaker behaviour. Details of $\mathcal{L}_{\text{mot}}$ and $\mathcal{L}_{\text{div}}$ are provided in Supplementary Material. Consequently, the training of the generic generator **GAG** is supervised by combining of all above losses as:

$$\mathcal{L}_G = \min_{\text{GAG}} \max_{D_R, D_A^G} (\mathcal{L}_R^G + \mathcal{L}_A^G + \lambda_m \mathcal{L}_{\text{mot}} + \lambda_d \mathcal{L}_{\text{div}}), \tag{11}$$

where λ_m and λ_d are balancing coefficients for adjusting the importance of the motion smoothness loss and diversity loss.

Training Personalised Generator PAG: The **PAG** is also jointly trained with two discriminators D_R and D_A^P under a similar adversarial setting as **GAG**. Different from the D_A^G's training where generic real AFRs $\mathbb{F}_i^G$ expressed by different listeners are paired as correct labels for the input B_i^s, when training D_A^P conditioned on the p^l, only personalised real AFRs $\mathbb{F}_i^l$ expressed by the listener l are treated as correct labels, i.e., generic real AFRs expressed by other listeners are not correct labels in this case. Accordingly, adversarial losses $\mathcal{L}_R^P$ (same as the Eq. 9) and $\mathcal{L}_A^P$ ensuring the realism and appropriateness of the facial reactions generated by our **PAG** are employed, where $\mathcal{L}_A^P$ is formulated as:

$$\begin{aligned} \mathcal{L}_A^P = & \underbrace{\mathbb{E}_{F_i^l(m)}\left[\log\left(D_A^P\left(B_i^s, F_i^l(m)|p^l\right)\right)\right]}_{\text{Evaluating the appropriateness of a real personalised AFR}} \\ & + \underbrace{\mathbb{E}_{B_i^s}\left[\log\left(1 - D_A^P\left(B_i^s, \text{PAG}\left(\mathcal{Q}_i^G, F_h^l\right)|p^l\right)\right)\right]}_{\text{Evaluating the appropriateness of a generated facial reaction}} \\ & + \underbrace{\mathbb{E}_{\bar{F}_i^l(m)}\left[\log\left(1 - D_A^P\left(B_i^s, \bar{F}_i^l(m)|p^l\right)\right)\right]}_{\text{Evaluating the appropriateness of a real personalised IFR}}, \end{aligned} \tag{12}$$

where $\bar{F}_i^l(m)$ denotes a real but inappropriate facial reaction expressed by the listener l in response to B_i^s. Again, the same **Motion Smoothness Loss** $\mathcal{L}_{\text{mot}}$ and **Diversity Loss** $\mathcal{L}_{\text{div}}$ are used. Consequently, the final loss function for training **PAG** can be defined as:

$$\mathcal{L}_P = \min_{\text{PAG}} \max_{D_R, D_A^P} (\mathcal{L}_R^P + \mathcal{L}_A^P + \lambda_m \mathcal{L}_{\text{mot}} + \lambda_d \mathcal{L}_{\text{div}}). \tag{13}$$

At this stage, the GAG is frozen but only samples latent generic AFR embeddings $E_i^G(m) \sim \mathcal{Q}_i^G$ for PAG's training.

Experiments

Experimental Settings

Datasets: Our approach is evaluated on a publicly available MAFRG dataset provided by REACT2024 Challenge [1] (RE-

[1] https://sites.google.com/cam.ac.uk/react2024/home

Method	Appropriateness				Diversity ($\times 10^{-2}$)			Realism	Synchrony
	FRCorr ↑		FRdist ↓		FRDiv ↑	FRDvs ↑	FRVar ↑	FRRea ↓	FRSyn ↓
	MAFRG	PMAFRG	MAFRG	PMAFRG					
Trans-VAE (Song et al. 2023a)	0.09	0.08	98.31	105.27	3.04	3.16	0.37	67.74	44.86
REGNN (Xu et al. 2023)	0.19	0.17	**84.54**	**91.33**	0.07	3.42	0.61	-	**41.35**
Unifarn (Liang et al. 2023)	0.19	0.15	98.51	103.21	8.19	7.60	2.59	-	46.11
Beamer (Hoque et al. 2023)	0.11	0.10	97.33	105.09	5.08	3.74	1.96	-	48.12
FRDiff (Yu et al. 2023)	0.14	0.13	91.05	98.28	6.90	4.43	1.08	69.37	47.66
BeLFusion (Barquero, Escalera, and Palmero 2023)	0.12	0.11	94.16	98.95	3.60	3.84	2.49	78.96	49.00
USTC-AC (Liu et al. 2024)*	0.17	0.14	104.04	107.70	16.75	13.85	**5.35**	-	44.54
PerReactor (Ours)	**0.34**	**0.29**	107.60	109.04	**17.87**	**17.89**	4.84	**65.15**	46.12

Table 1: Comparison between PerReactor and existing MAFRG methods on REACT2024 test set for both tasks. The **best** results are marked in bold. *: Results computed using predictions provided by authors (the reported MAFRG FRCorr is 0.22).

ACT2023 dataset is not available). It contains 2962 dyadic interaction audio-visual clip pairs (5924 clips in total), including 1594 pairs for training, 562 pairs for validation and 806 for test. These clips are originally recorded by NoXI (Cafaro et al. 2017) and RECOLA (Ringeval et al. 2013).

Implementation details: In our experiments, we utilise the AdamW optimizer (Loshchilov and Hutter 2017) to train our PerReactor. The default coefficients λ_d and λ_m for balancing the regularisation losses are set to 10 and 10^{-2}, respectively. More details are provided in Supplementary Material.

Metrics: We follow (Song et al. 2023b,a) to evaluate four aspects of the generated facial reactions, including **appropriateness** (FRCorr and FRdist), **diversity** (FRDiv, FRDvs and FRVar), **realism** (FRRea) and **synchrony** (FRSyn). Please refer to (Song et al. 2023b) for more details.

Comparison With Existing Approaches

Table 1 compares our PerReactor with MAFRG competitors on both offline MAFRG and PMAFRG tasks, where facial reactions generated by our PerReactor achieved the highest correlation (FRCorr) with real AFRs (0.34 and 0.29), showing more than 79% and 71% improvements over previous SOTA (i.e., REGNN and Unifarn) for MAFRG and PMAFRG tasks, respectively. The reason that the appropriateness results of the same systems for MAFRG task are consistently superior to their results for PMAFRG task is that the generic real AFRs of each generated facial reaction include not only its personalised real AFRs but also real AFRs expressed by other listeners. Compared to existing approaches (except USTC-AC), our PerReactor achieved around 200% improvements in generating diverse facial reactions (all three diversity metrics), as well as the best realism performance. This validates that our adversarial training strategy allows the model to learn distributions representing more diverse, realistic and appropriate facial reactions compared to other MAFRG training strategies. Also, the results suggests that the higher diversities may be at the cost of enlarging distances (FRdist) between the generated facial reactions and their AFRs. Fig. 3 further illustrates that our PerReactor generate more diverse generic/personalised AFRs.

Ablation Studies

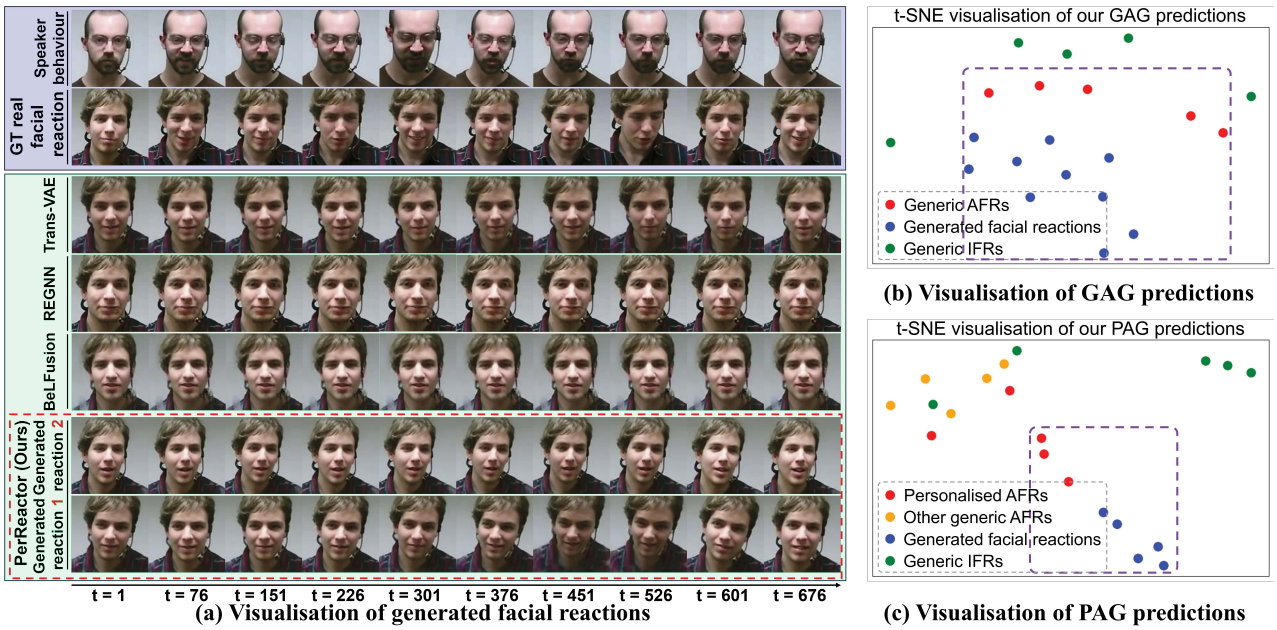
Several ablation studies are conducted below to deeply investigate our approach. Besides, the analysis for: the number of attribute segments, fusion strategies, regularisation coefficients, weighting factor β, model complexity, and statistical differences are provided in Supplementary Material.

GAG	PAG	D_R	D_A^G	D_A^P	FRCorr ↑		FRDiv ↑	FRRea ↓
					MAFRG	PMAFRG		
✓					0.06	0.04	0.22	92.63
✓		✓			0.02	0.01	**18.81**	72.00
✓			✓		0.24	0.20	18.31	251.14
✓		✓	✓		0.31	0.26	17.78	75.01
✓	✓	✓	✓		0.03	0.02	0.14	73.90
✓	✓		✓	✓	0.30	0.24	9.83	264.35
✓	✓	✓	✓	✓	**0.34**	**0.29**	8.93	**65.15**

Table 2: Ablation study results achieved by various PerReactor settings, where **reconstruction loss** is employed to train GAG and PAG when their corresponding discriminators are not used. The FRDiv are scaled up by 10^2 as Table 1.

Results Achieved by Various PerReactor Settings Table 2 comprehensively evaluates the contributions of different modules/blocks in our PerReactor to the generated facial reactions. It can be observed that the introduction of the appropriateness discriminators D_A^G and D_A^P significantly improved the appropriateness (FRCorr) of facial reactions generated by GAG and PAG for MAFRG and PMAFRG tasks, respectively. Specifically, D_A^G/D_A^P increased the FRCorr from 0.06 to 0.24/0.03 to 0.34 for MAFRG task and 0.04 to 0.20/0.02 to 0.29 for PMAFRG task. We also found that simply employing a realism discriminator D_R to train GAG/-PAG without using appropriateness discriminators D_A^G/D_A^P lead the generated facial reactions to have very poor appropriateness performance, suggesting that appropriateness discriminators indeed learnt effective and comprehensive appropriateness metrics for training generators. Meanwhile, the use of the realism discriminator D_R also consistently ensured the generated facial reactions to be more realistic. Importantly, our PBPM can effectively model personalised behaviour patterns to improve the appropriateness and realism of the generated facial reactions.

Contributions of Different Modalities Table 3 evaluates the results achieved by inputting different speaker behaviour modalities, indicating that using both speaker audio and visual behaviours yielded the highest FRCorr values on both MAFRG and PMAFRG tasks compared to using unimodal



(a) Visualisation of generated facial reactions

(b) Visualisation of GAG predictions

(c) Visualisation of PAG predictions

Figure 3: **(a)** Visualisation of facial reactions generated by different approaches, where ours have clearly more head movements and diverse facial expressions. **(b)** Given an speaker behaviour, most generic facial reactions (blue dots) sampled from the distribution learnt by our GAG are similar to the corresponding real generic AFRs (red dots). **(c)** Given an speaker behaviour, most personalised facial reactions (blue dots) generated from our PAG are more similar to the corresponding personalised AFRs (red dots) than generic AFRs expressed by others (orange dots) or IFRs (green dots).

Modality		FRCorr ↑		FRdist ↓		FRDiv ↑
Audio	Visual	MAFRG	PMAFRG	MAFRG	PMAFRG	
✓		0.21	0.19	109.18	110.48	17.25
	✓	0.26	0.23	**106.93**	**108.48**	15.17
✓	✓	**0.31**	**0.26**	114.28	115.50	**17.87**

Table 3: Results of different speaker behaviour modalities.

Length	FRCorr ↑		FRdist ↓		FRDiv ↑	FRSyn ↓
	MAFRG	PMAFRG	MAFRG	PMAFRG		
9.6s	0.32	0.27	**107.18**	**108.51**	7.73	**45.61**
19.2s	**0.34**	0.29	107.60	109.04	**8.93**	46.12
28.8s	**0.34**	**0.30**	107.65	109.28	5.05	45.86

Table 4: Results of varying historical video lengths.

speaker behaviour, despite that excluding speaker audio behaviour results in an improved FRdist. Additionally, the multi-modal setting results in higher diversity (FRDiv) as the model benefits from the complementary nature of the information provided by each modality, which enriches the facial reaction distribution. While both speaker audio and visual behaviours can positively predict AFRs, visual behaviours provided more informative cues for both tasks.

Influences of the Historical Facial Videos Table 4 shows that utilising longer historical face videos for personalised behaviour pattern modelling facilitates generating facial reactions with clearly higher FRCorr with their personalised

Video	FRCorr↑		FRdist ↓		FRDiv ↑	FRSyn ↓
	MAFRG	PMAFRG	MAFRG	PMAFRG		
Video 1	0.34	0.29	107.60	109.04	8.93	46.12
Video 2	0.33	0.30	107.59	108.77	8.84	46.07
Video 3	0.33	0.30	107.63	108.86	8.85	45.79
Video 4	0.33	0.30	107.63	108.99	8.95	45.88

Table 5: Results of different historical video contents.

AFRs. We explain this as longer videos provide more hints for one's personalised behaviour patterns. Additionally, Table 5 demonstrates that our PPE block can effectively capture consistent personalised behaviour patterns across face videos of varying contents, as evidenced by the stable performance over appropriateness, diversity and synchrony.

Conclusion

This paper proposes the first GAN-based MAFRG approach which deep learns a comprehensive criteria to train generators that can predict a distribution representing multiple AFRs from each speaker behaviour, which also includes a novel listener personalised factor modelling strategy. Experiments show that both the proposed adversarial training and personalised factor modelling strategies positively contribute to the achieved promising performance. As a pioneer personalised MAFRG approach, our model still has relatively limited FRDist and visualisation performances, which will be our future research directions.

Acknowledgments

This research used the ALICE High Performance Computing facility at the University of Leicester and the Sulis Tier 2 HPC platform hosted by the Scientific Computing Research Technology Platform at the University of Warwick. Sulis is funded by EPSRC Grant EP/T022108/1 and the HPC Midlands+ consortium. This work was supported by the National Natural Science Foundation of China under Grant 62001173 and 62171188. This work was also supported by the National Natural Science Foundation of China under Grant 82261138629, Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010688, Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

References

Barquero, G.; Escalera, S.; and Palmero, C. 2023. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2317–2327.

Cafaro, A.; Wagner, J.; Baur, T.; Dermouche, S.; Torres Torres, M.; Pelachaud, C.; André, E.; and Valstar, M. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 350–359.

Dam, Q. T.; Nguyen, T. T. N.; Tran, D. T.; and Lee, J.-H. 2024. Finite Scalar Quantization as Facial Tokenizer for Dyadic Reaction Generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–5. IEEE.

Hoque, X.; Mann, A.; Sharma, G.; and Dhall, A. 2023. BEAMER: Behavioral Encoder to Generate Multiple Appropriate Facial Reactions. In *Proceedings of the ACM International Conference on Multimedia*, 9536–9540.

Huang, Y.; and Khan, S. 2018a. A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 437–445.

Huang, Y.; and Khan, S. M. 2017. Dyadgan: Generating facial expressions in dyadic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 11–18.

Huang, Y.; and Khan, S. M. 2018b. Generating Photorealistic Facial Expressions in Dyadic Interactions. In *BMVC*, 201.

Liang, C.; Wang, J.; Zhang, H.; Tang, B.; Huang, J.; Wang, S.; and Chen, X. 2023. UniFaRN: Unified Transformer for Facial Reaction Generation. In *Proceedings of the ACM International Conference on Multimedia*, 9506–9510.

Liu, Z.; Liang, C.; Wang, J.; Zhang, H.; Liu, Y.; Zhang, C.; Gui, J.; and Wang, S. 2024. One-to-Many Appropriate Reaction Mapping Modeling with Discrete Latent Variable. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–5. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, C.; Song, S.; Xie, W.; Spitale, M.; Ge, Z.; Shen, L.; and Gunes, H. 2024. ReactFace: Online Multiple Appropriate Facial Reaction Generation in Dyadic Interactions. *IEEE Transactions on Visualization and Computer Graphics*, 1–18.

Mehrabian, A.; and Russell, J. A. 1974. *An approach to environmental psychology.* the MIT Press.

Mentzer, F.; Minnen, D.; Agustsson, E.; and Tschannen, M. 2024. Finite Scalar Quantization: VQ-VAE Made Simple. In *The Twelfth International Conference on Learning Representations*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ng, E.; Joo, H.; Hu, L.; Li, H.; Darrell, T.; Kanazawa, A.; and Ginosar, S. 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.

Nguyen, D.-K.; Paudel, P.; Kim, S.-W.; Shin, J.-E.; Kim, S.-H.; and Yang, H.-J. 2024a. Multiple Facial Reaction Generation Using Gaussian Mixture of Models and Multimodal Bottleneck Transformer. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–5. IEEE.

Nguyen, M.-D.; Yang, H.-J.; Ho, N.-H.; Kim, S.-H.; Kim, S.; and Shin, J.-E. 2024b. Vector Quantized Diffusion Models for Multiple Appropriate Reactions Generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–5. IEEE.

Nojavanasghari, B.; Huang, Y.; and Khan, S. 2018. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092*.

Pandita, S.; Mishra, H. G.; and Chib, S. 2021. Psychological impact of covid-19 crises on students through the lens of Stimulus-Organism-Response (SOR) model. *Children and Youth Services Review*, 120: 105783.

Ringeval, F.; Sonderegger, A.; Sauer, J.; and Lalanne, D. 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)*, 1–8. IEEE.

Sagliano, L.; Ponari, M.; Conson, M.; and Trojano, L. 2022. The interpersonal effects of emotions: The influence of facial expressions on social interactions. *Frontiers in Psychology*, 13: 1074216.

Shao, Z.; Song, S.; Jaiswal, S.; Shen, L.; Valstar, M.; and Gunes, H. 2021. Personality recognition by modelling person-specific cognitive processes using graph representation. In *proceedings of the 29th ACM international conference on multimedia*, 357–366.

Song, S.; Shao, Z.; Jaiswal, S.; Shen, L.; Valstar, M.; and Gunes, H. 2022. Learning Person-specific Cognition from Facial Reactions for Automatic Personality Recognition. *IEEE Transactions on Affective Computing*.

Song, S.; Spitale, M.; Luo, C.; Barquero, G.; Palmero, C.; Escalera, S.; Valstar, M.; Baur, T.; Ringeval, F.; André, E.; et al. 2023a. React2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9620–9624.

Song, S.; Spitale, M.; Luo, C.; Palmero, C.; Barquero, G.; Zhu, H.; Escalera, S.; Valstar, M.; Baur, T.; Ringeval, F.; André, E.; and Gunes, H. 2024. REACT 2024: the Second Multiple Appropriate Facial Reaction Generation Challenge. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–5.

Song, S.; Spitale, M.; Luo, Y.; Bal, B.; and Gunes, H. 2023b. Multiple Appropriate Facial Reaction Generation in Dyadic Interaction Settings: What, Why and How? *arXiv e-prints*, arXiv–2302.

Tang, B.; Cao, R.; Chen, R.; Chen, X.; Hua, B.; and Wu, F. 2023. Automatic generation of robot facial expressions with preferences. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 7606–7613. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Woo, J.; Fares, M.; Pelachaud, C.; and Achard, C. 2023. AMII: Adaptive Multimodal Inter-personal and Intra-personal Model for Adapted Behavior Synthesis. *arXiv preprint arXiv:2305.11310*.

Woo, J.; Pelachaud, C. I.; and Achard, C. 2021. Creating an interactive human/agent loop using multimodal recurrent neural networks. In *WACAI 2021*.

Xu, T.; Spitale, M.; Tang, H.; Liu, L.; Gunes, H.; and Song, S. 2023. Reversible Graph Neural Network-based Reaction Distribution Learning for Multiple Appropriate Facial Reactions Generation. *arXiv preprint arXiv:2305.15270*.

Yu, J.; Zhao, J.; Xie, G.; Chen, F.; Yu, Y.; Peng, L.; Li, M.; and Dai, Z. 2023. Leveraging the Latent Diffusion Models for Offline Facial Multiple Appropriate Reactions Generation. In *Proceedings of the ACM International Conference on Multimedia*, 9561–9565.

Zhu, H.; Kong, X.; Xie, W.; Huang, X.; Shen, L.; Liu, L.; Gunes, H.; and Song, S. 2024. Perfrdiff: Personalised weight editing for multiple appropriate facial reaction generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9495–9504.