ReactFace: Online Multiple Appropriate Facial Reaction Generation in Dyadic Interactions

Cheng Luo, Siyang Song*, Weicheng Xie*, Micol Spitale, Zongyuan Ge, Linlin Shen and Hatice Gunes

Abstract—In dyadic interaction, predicting the listener's facial reactions is challenging as different reactions could be appropriate in response to the same speaker's behaviour. Previous approaches predominantly treated this task as an interpolation or fitting problem, emphasizing deterministic outcomes but ignoring the diversity and uncertainty of human facial reactions. Furthermore, these methods often failed to model short-range and long-range dependencies within the interaction context, leading to issues in the synchrony and appropriateness of the generated facial reactions. To address these limitations, this paper reformulates the task as an extrapolation or prediction problem, and proposes an novel framework (called ReactFace) to generate multiple different but appropriate facial reactions from a speaker behaviour rather than merely replicating the corresponding listener facial behaviours. Our ReactFace generates multiple different but appropriate photo-realistic human facial reactions by: (i) learning an appropriate facial reaction distribution representing multiple different but appropriate facial reactions; and (ii) synchronizing the generated facial reactions with the speaker verbal and non-verbal behaviours at each time stamp, resulting in realistic 2D facial reaction sequences. Experimental results demonstrate the effectiveness of our approach in generating multiple diverse, synchronized, and appropriate facial reactions from each speaker's behaviour. The quality of the generated facial reactions is intimately tied to the speaker's speech and facial expressions, achieved through our novel speaker-listener interaction modules. Our code is made publicly available at https://github.com/lingjivoo/ReactFace.

Index Terms—Multiple Appropriate Facial Reaction Generation (MAFRG), Generative Model, Video Generation

✦

1 INTRODUCTION

HUMAN facial reactions play a crucial role in conveying significant non-verbal cues that aid in expressing intentions during human-human interactions [3], [4]. In dyadic speaker-listener interactions, the facial reaction of a listener to a speaker's behaviour is depending on not solely the stimulus presented by the speaker, but also the contextual factors such as the conversational environment and the listener's individual disposition (*e.g.*, personality [5], [6]) [7]–[9]. As a result, facial behaviours expressed for responding to a specific speaker's behaviour might be varied across different listeners and even within the same listener across diverse contextual settings. In other words, *multiple different facial reactions of listeners may be appropriate in response to a specific stimulus, and thus facial reaction generation can be treated as a 'one-to-many mapping' problem* [10]. To this end, it is much more challenging compared with the 'one-
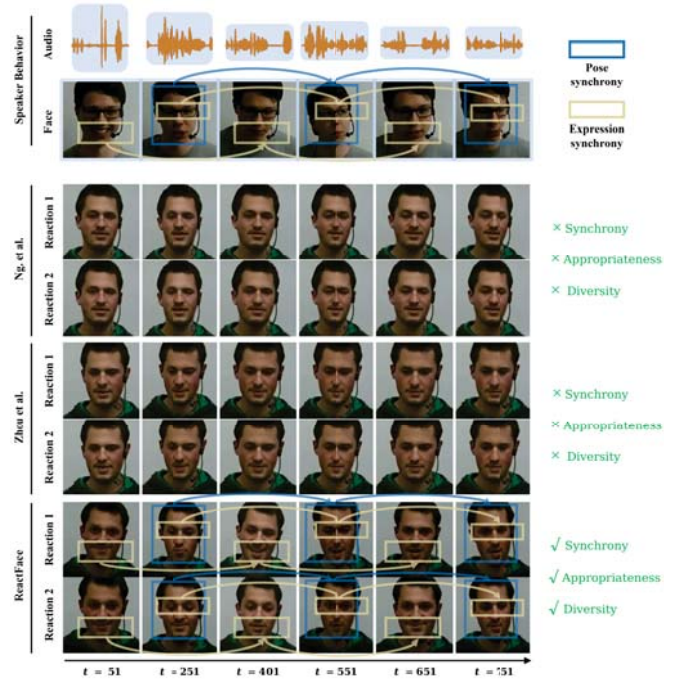

Fig. 1. Comparison between our ReactFace and facial reaction generation methods proposed by Ng. et al [1] and Zhou et al. [2]. The uppermost two rows delineate the auditory, and visual behaviours of a speaker's behavioural clip. The subsequent two rows display a pair of facial reaction sequences generated by [1], in response to the speaker's input. The fifth and sixth rows show a pair of facial reaction sequences generated by [2]. Finally, the last two rows present diverse, appropriate and synchronised facial reactions generated by our ReactFace.

- *Cheng Luo, Weicheng Xie and Linlin Shen are with Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. Weicheng Xie and Linlin Shen are also with Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China. Linlin Shen is also with the Department of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China.*
- *Siyang Song is with Computer Sciences, University of Exeter, Exeter, EX4 4PY, United Kingdom.*
- *Siyang Song, Micol Spitale and Hatice Gunes are with the Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FT, United Kingdom. E-mail: {ss2796, ms2871, hg410}@cam.ac.uk.*
- *Cheng Luo and Zongyuan Ge are with the Airdoc-Monash Research Centre and Monash University Faculty of IT, Monash University, Melbourne, Australia. E-mail: {cheng.luo, zongyuan.ge}@monash.edu.*
- *Corresponding Author: * Dr Siyang Song and Dr Weicheng Xie. E-mail: ss2796@cam.ac.uk and wcxie@szu.edu.cn.*

to-one' mapping or fitting human behaviour generation problems (with deterministic results), such as speech-driven 3D lip animation [11], next sentence prediction [12] and deterministic behaviour prediction [2], [13].

Only a limited number of previous studies have investigated the online facial reaction generation task. Online facial reaction generation is the task of producing a listener's facial reaction sequence iteratively during an ongoing conversation with a speaker, in real-time. This approach is fundamentally different from offline generation, where the entire sequence of the speaker's video is processed at once, and the corresponding listener's video is output as a complete sequence. The online method is designed to generate reactions promptly, either autoregressively or segment-by-segment, making it suitable for real-time applications where immediate response is essential.

Early studies [14]–[16] commonly employed conditional generative adversarial networks (CGANs) [17], [18], which leverage a facial expression image or static facial attributes (*e.g.,* Facial Action Units (AUs)) of the speaker, as conditioning cues. These approaches aimed to reproduce each static facial reaction frame of the corresponding listener. Nevertheless, these approaches overlooked the temporal correlation spanning consecutive frames of the speaker's facial dynamics. Besides, they omitted other speaker behaviour modalities such as audio and text, which are crucial in triggering the listener's facial reactions. Recent facial reaction generation approaches address these limitations by incorporating temporal networks (*e.g.,* Long-short-term-memory (LSTM) architecture [2], [5], [6]), or coupling with supplemental cues extracted from the speaker's auditory [5], [6] or textual [1] signals. However, these models remain remiss in capturing the *synchrony* between the speaker's behaviour and the corresponding listener's facial response (**Problem 1**) – a crucial element for creating authentic facial expressions during face-to-face interactions.

Moreover, almost all previous facial reaction generation approaches [1], [5], [6], [15], [16], [19] shared a problematic training strategy, *i.e.,* their training are achieved by directly pairing the input speaker behaviour with the corresponding listener's real facial reaction (called GT facial reaction in this paper). Since facial reaction generation is a 'one-to-many mapping' task, such 'one-to-one mapping' training strategies would cause an ill-posed machine learning problem where similar speaker inputs could paired with different listener facial reaction labels within a given dataset. This would theoretically undermine the feasibility of building reliable models (**Problem 2**) [10]. Consequently, most of these frameworks [5], [6], [15], [16], [19] are constrained to generating a single facial reaction in response to each speaker behaviour. Although some recent models [1], [20] have partially overcame this limitation by generating different facial reactions from a single speaker behaviour input, these approaches failed to consider the *appropriateness* of the generated facial reactions in a given conversational context (**Problem 3**).

In this paper, we propose **ReactFace**, an novel online multi-modal framework for generating multiple different yet appropriate, realistic, and temporally synchronized facial reactions in response to audio-visual speaker behaviours. ReactFace pioneers a solution to the MAFRG problem discussed above, with the introduction of two novel strategies. Firstly, the **speaker-listener behaviour synchronization (BS)** strategy synchronises the generated facial reactions with both the speaker speech and visual behaviours across the temporal dimension. This iterative process aligns the facial reactions with the speaker behaviours at each time stamp, ensuring that the generated facial reactions are continuous as well as synchronised with the corresponding speaker behaviours (**Addressing Problem 1**). Additionally, a **multiple-appropriate facial reaction generation (AFRG)** mechanism is proposed to tackle the 'one-to-many mapping' training problem by re-formulating it as a 'one-to-one mapping' problem. This is achieved by training the model to map a input single speaker behaviour to an distribution representing all appropriate facial reactions (**Addressing Problem 2**). Consequently, a range of different yet appropriate facial reactions can be sampled based on the obtained distribution learned from the input speaker behaviour (**Addressing Problem 3**). In summary, the main contributions and novelties of this paper are listed as follows:

- We propose the first approach for online multiple appropriate facial reaction generation (online MAFRG), to facilitate the generation of diverse, appropriate, realistic, and temporally aligned facial reactions in response to each speaker behaviour. The main advantage of this online approach lies in its capacity to craft individual facial reaction frames while taking into account the historical and present speaker behaviours. To the best of our knowledge, our ReactFace is the first and only online approach for generating multiple appropriate facial reactions.
- We propose a novel speaker-listener behaviour synchronization (BS) module to temporally synchronize the generated facial reactions with the speaker speech and visual behaviours. To the best of our knowledge, this is the first multi-modal synchronization module specifically proposed for the task of facial reaction generation.
- The experimental results demonstrate that ReactFace, as posited in this study, outperforms existing state-of-the-art methods in terms of achieving superior interaction synchronization, enhanced diversity, and heightened realism.

2 RELATED WORK

This section begins with a recap of previous approaches to facial reaction generation and facial behavior generation in Sec. 2.1, followed by a review of the literature on other non-verbal behaviors, such as gesture and body motion generation, in Sec. 2.2. Next, it discusses conditional generative models in Sec. 2.3 and examines the modality alignment techniques utilized in Sec. 2.4.

2.1 Automatic Facial Reaction Generation

Facial Reaction Generation. Existing approaches attempt to reproduce the listener's GT facial reactions from the input speaker behaviour. For example, early studies frequently [14]–[16], [19] extended the conditional GAN [17],

Notations	Descriptions
$\mathbf{B}_{1:t}^{s}$	Speaker audio-facial behaviour
$\boldsymbol{\mathcal{X}}_{1:t}^{s}$	Raw speaker audio signal
$\mathbf{F}_{1:t}^{s} = \{\mathbf{f}_1^s, \ldots, \mathbf{f}_t^s\}$	2D speaker face image sequence
$\mathbf{R}_{1:t}^{s} = \{\mathbf{r}_1^s, \ldots, \mathbf{r}_t^s\}$	3D speaker face image sequence
$\boldsymbol{A}_{1:kt}^{s} = \{\boldsymbol{a}_1^s, \ldots, \boldsymbol{a}_{kt}^s\}$	Extracted speaker speech feature sequence
$\tilde{\boldsymbol{F}}_{1:t}^{s} = \{\tilde{\boldsymbol{f}}_1^s, \ldots, \tilde{\boldsymbol{f}}_t^s\}$	Extracted speaker facial feature sequence
$\bar{\boldsymbol{F}}_{1:t}^{s} = \{\bar{\boldsymbol{f}}_1^s, \ldots, \bar{\boldsymbol{f}}_t^s\}$	Aligned speaker facial feature sequence
$\hat{\mathbf{R}}_{1:t}^{s} = \{\hat{\mathbf{r}}_1^s, \ldots, \hat{\mathbf{r}}_t^s\}$	3D speaker face image sequence prediction
$\mathbf{F}_{1:t}^{l} = \{\mathbf{f}_1^l, \ldots, \mathbf{f}_t^l\}$	2D facial reaction image sequence
$\mathbf{R}_{1:t}^{l} = \{\mathbf{r}_1^l, \ldots, \mathbf{r}_t^l\}$	3D facial reaction image sequence
$\bar{\boldsymbol{F}}_{1:t-w}^{l}\{\bar{\boldsymbol{f}}_1^l, \ldots, \bar{\boldsymbol{f}}_{t-w}^l\}$	Aligned listener facial feature sequence
$\boldsymbol{R}_{t-w+1:t}^{l} = \{\boldsymbol{r}_{t-w+1}^l, \ldots, \boldsymbol{r}_t^l\}$	Raw 3D facial reaction sequence prediction
$\bar{\boldsymbol{R}}_{t-w+1:t}^{l} = \{\bar{\boldsymbol{r}}_{t-w+1}^l, \ldots, \bar{\boldsymbol{r}}_t^l\}$	Visually synchronised 3D facial reaction sequence prediction
$\hat{\mathbf{R}}_{t-w+1:t}^{l} = \{\hat{\mathbf{r}}_{t-w+1}^l, \ldots, \hat{\mathbf{r}}_t^l\}$	Finally synchronised 3D facial reaction sequence prediction
$\boldsymbol{P}_{t-w+1:t}^{l} = \{\boldsymbol{p}_{t-w+1}^l, \ldots, \boldsymbol{p}_t^l\}$	Positional encodings
$\boldsymbol{Z}_{t-w+1:t} = \{\boldsymbol{z}_{t-w+1}, \ldots, \boldsymbol{z}_t\}$	Sampled latent vectors representing an appropriate facial reaction

TABLE 1
Notations used in this paper.

[21] to predict each listener's GT facial reaction under a one-on-one virtual interview scenario, which takes each speaker's facial image as the input and aim to reproduce a frame-level GT facial reaction (sketches) expressed by the corresponding listener. Recent facial reaction studies [1], [2] frequently represent facial motions using 3DMM (3D Morphable Model) coefficients (*e.g.*, expression and pose coefficients) to better visualize the facial muscle movements. Meanwhile, speaker's speech and audio behaviours were also additionally employed as the input to deliver richer verbal and non-verbal speaker behavioural cues [1], [5], [6], which are crucial in triggering listeners' facial reactions. For example, Ng et al. [1] proposed a VQ-VAE-based framework that considers the nondeterministic nature of human facial reactions by learning a set of codebook vectors. Song et al. [5], [6] search for a network whose architecture and weights are specifically adapted to each target listener's person-specific facial reactions in response to the speaker's audio and facial non-verbal behaviour. As discussed in [8], [10], human facial reaction has an uncertain nature, *i.e.*, the same/similar behaviours expressed by speakers could trigger different facial reactions across not only various subjects but also the same subject under different contexts. Thus, it is problematic to train models that aims to reproduce the corresponding listener's GT facial reaction only from each speaker behaviour sequence. As a result, a publicly Grand Challenge [22] focusing on Multiple Appropriate Facial Reaction Generation (MAFRG) has been organised. It is important to note that since the submission of our paper, numerous MAFRG approaches have been developed [23]–[32]. However, it is not fair to directly compare these methods with ours, as some of them [24]–[26], [28] are offline MAFRG approaches, while others [28], [32] are built upon our ReacFace framework.

Facial Animation. The proposed facial reaction generation approach also involves facial animation, which has attracted considerable attention in recent years [33]–[38]. In particular, speech-driven face generation models aim to automatically animate vivid facial expressions of a 2D portrait [36], [38]–[40] or 3D avatar face sequence [11], [37] from a human speech signal. For example, StyleGAN-V [41] and StyleFaceV [42] are designed to generate a temporally coherent facial motion sequence with the same identity. Moreover, facial expression [43]–[45] or head pose [39], [44] representations are additionally employed as the condition signal in some approaches, in order to manually control the generated face sequence. However, given a subject's speech, all these works can only generate a face sequence of the corresponding subject rather than the conversational partner's facial reactions.

2.2 Non-verbal Human body/gesture Behaviour Generation

Besides the facial behaviour generation, a large number of studies also have been devoted to the domain of the neural body and gesture motion generation models [13], [13], [46]–[51] over the preceding decades. They frequently represent the generated human motions in the form of 3D skeletons [52]–[54], video frames [55], or 3D parameters (*e.g.*, SMPL [46], [48], [56]). This evolution encompasses a spectrum of generative models such as Generative Adversarial Networks (GANs) [50], [54], variational autoencoder (VAE) [56], and normalizing flows [52] as well as the recent entrant of diffusion models [46].

These motion synthesis approaches can be divided into two primary categories: unconstrained generation and conditioned generation. Unconstrained approaches [57]–[59] are usually achieved by networks that model a latent space of all possible motions. For example, Yan et al. [57] proposed a convolutional sequence generation framework that can produce long actions transformed from a sequence of latent vectors sampled from a Gaussian process. In contrast, conditioned networks are defined to yield different outputs depending on conditional signals, including music [53], [60], speech [61], [62], action [63], and text [54], [56], [64]. However, the samples generated by these deterministic generative models [13], [65] and non-deterministic generative models [56]–[58] frequently suffer from insufficient diversity problem. In this paper, we consider generation diversity in interaction scenarios, where we offer a fresh
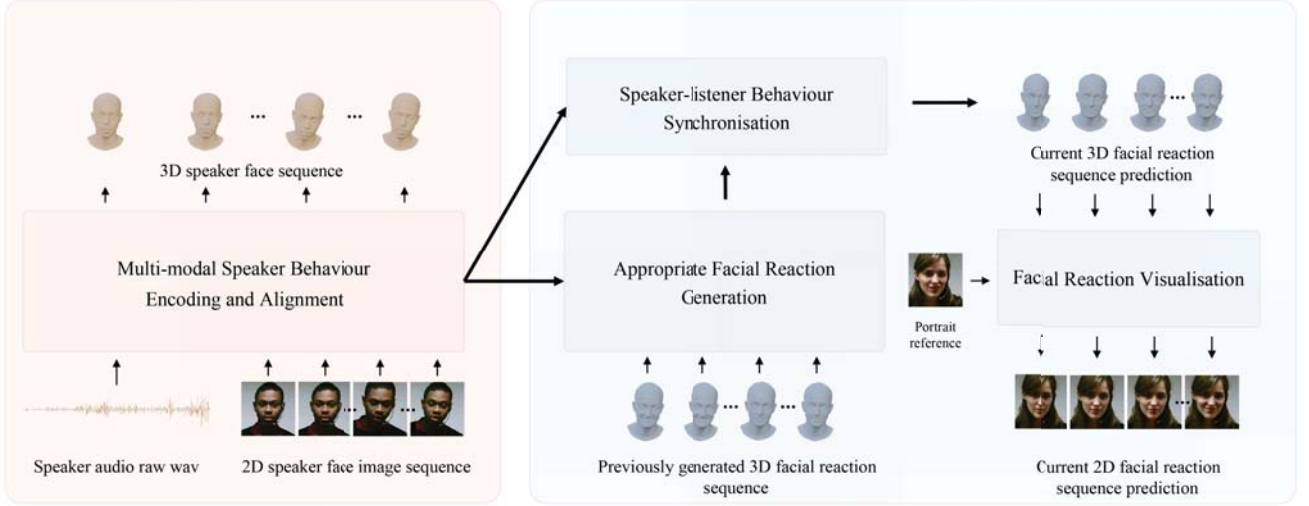
Fig. 2. The pipeline of the proposed ReactFace model.

idea by transforming from one-to-one supervision to one-to-many supervision, based on which the well-trained model achieved high diversity in its outputs (illustrated in Fig. 6).

2.3 Conditional Generative Models

Recent techniques for conditional generation have been developed to incorporate various modalities, ensuring that synthesized results align with the instructions provided by these modality signals [66]. Early approaches, such as conditional variational autoencoders (CVAEs) [66] and conditional generative adversarial networks [66], leverage class labels to distinguish image domains and guide generated image samples to possess domain-specific properties. For example, CycleGAN was proposed to translate images from a source domain to a target domain, while the StarGAN series [18], [67] further succeeded in translating images from a source domain to multiple target domains. Some techniques embed discrete class labels as conditional signals in the generation process, either by directly concatenating class labels with the input [18], [66] or by using conditional normalization techniques [67], [68]. Building on these successful practices, conditional diffusion models incorporate class information (*e.g.*, class labels or text) into normalization layers and guide the generation process using classifier gradients. Further research has shown that guidance can be derived from the generative model itself without a classifier, a method known as classifier-free guidance. This advancement unlocks various forms of conditional signals, leading to conditional generative models [69], [70] that control synthesized samples with multiple modalities, such as text, edge maps, human pose skeletons, segmentation maps, depth, and normals.

2.4 Modality Alignment in Generative Models

Modality alignment is crucial for generative models processing diverse data types like text, images, audio, and video. This alignment ensures coherent content generation and manipulation across modalities, but faces challenges due to semantic and dimensional differences. To address the semantic disparity, researchers have turned to contrastive learning techniques. Works such as Zhang et al. [69] and Kim et al. [71] have built upon the foundational approach introduced by Radford et al. [72] in CLIP. This method utilizes positive and negative pairs to learn aligned representations, proving particularly effective in text-to-vision synthesis tasks. The dimensionality mismatch between modalities presents another significant hurdle. Recent advancements, exemplified by Zhang et al. [69] and Rombach et al. [70], have leveraged cross-modal attention mechanisms. These allow models to focus on relevant features within each modality, effectively reducing dimensionality while preserving crucial information. A straightforward yet effective approach to unifying different modalities has been the adoption of multimodal transformers, as demonstrated by Tsai et al. [73]. This method, which represents different modalities as unified tokens, has found wide application across various generative tasks. It has been successfully employed in text-to-image generation [69], text-to-motion synthesis [46], [48], and even multi-modal generation [74], showcasing its versatility and effectiveness in bridging the gap between diverse data types. For temporal alignment, innovative attention variants have emerged. Press et al. [75] introduced Attention with Linear Biases (ALiBi), while Fan et al. [11] proposed Biased Cross-Modal Attention with periodic position encoding. These approaches incorporate temporal alignment biases into the attention mechanism, ensuring meaningful matches between query and key modalities in the temporal dimension. However, these temporal alignment methods fail to model the temporal interaction between two conversation participants. For instance, the point-to-point audio-visual alignment in Fan et al. [11] does not account for audio-visual influences from previous frames. In this paper, we propose a novel approach to modality interaction over time, which can effectively model the multi-modal signal exchange between two subjects engaged in an ongoing dyadic interaction.

3 PROBLEM FORMULATION

At the time t, the goal of the **online MAFRG task** is to develop a machine learning (ML) model $\mathcal{M}$ that can generate

multiple, different, but appropriate spatio-temporal facial reactions in response to the given speaker behaviour $\mathbf{B}^s_{1:t}$, which can be formulated as:

$$\hat{\mathbf{F}}^l_{t-w+1:t} = \mathcal{M}(\hat{\mathbf{F}}^l_{1:t-w}, \mathbf{B}^s_{1:t}) \tag{1}$$

where $\hat{\mathbf{F}}^l_{1:t-w}$ denotes the previously predicted/generated 2D facial reaction sequence, and w represents the facial reaction delay caused by the execution of the human cognitive processes [76]. Importantly, the well-trained model $\mathcal{M}$ is expected to be able to generate multiple different but appropriate facial reaction sequences from each input speaker behaviour. For more details, please refer to [10].

4 METHODOLOGY

This section introduces our online MAFRG model (ReactFace), which aims to generate multiple, different, yet appropriate, realistic, and synchronized facial reactions from each input spatio-temporal speaker behavior. To account for the time delay caused by human cognitive processes in facial reaction generation, we follow [1], [5], [6] in defining a small time window w (with $w = 8$ used in this paper). Subsequently, the speaker's audio-visual behavior $\mathbf{B}^s_{1:t}$ can be split into two parts as:

$$\mathbf{B}^s_{1:t} = \{\mathbf{B}^s_{1:t-w}, \mathbf{B}^s_{t-w+1:t}\}, \tag{2}$$

where $\mathbf{B}^s_{1:t-w} = \{\boldsymbol{\mathcal{X}}^s_{1:t-w}, \mathbf{F}^s_{1:t-w}\}$ denotes the previous audio-facial speaker behavior expressed over the time period $[1, t-w]$, and $\mathbf{B}^s_{t-w+1:t} = \{\boldsymbol{\mathcal{X}}^s_{t-w+1:t}, \mathbf{F}^s_{t-w+1:t}\}$ denotes the current audio-facial speaker behavior expressed over the time period $[t-w+1, t]$. Thus, $\mathbf{B}^s_{1:t}$ can also be denoted as:

$$\mathbf{B}^s_{1:t} = \{\boldsymbol{\mathcal{X}}^s_{1:t}, \mathbf{F}^s_{1:t}\}, \tag{3}$$

where $\boldsymbol{\mathcal{X}}$ and $\mathbf{F}$ represent the original audio signal and 2D facial image sequence, respectively. Then, our ReactFace model generates a 3D facial reaction segment ($\hat{\mathbf{R}}^l_{t-w+1:t}$) consisting of w frames (corresponding to the time period $[t-w+1, t]$) based on: (1) the previously generated/predicted facial reaction sequence $\hat{\mathbf{R}}^l_{1:t-w}$; (2) the previous speaker behavior $\mathbf{B}^s_{1:t-w}$; and (3) the current speaker behavior $\mathbf{B}^s_{t-w+1:t}$, which can be formulated as:

$$\hat{\mathbf{R}}^l_{t-w+1:t} = \mathcal{M}(\hat{\mathbf{R}}^l_{1:t-w}, \mathbf{B}^s_{1:t-w}, \mathbf{B}^s_{t-w+1:t}). \tag{4}$$

In the following subsections, we first present the entire framework of the ReactFace in Sec. 4.1, providing a high-level overview for the pipeline. Next, we provide a detailed explanation of the main technical contributions of the React-Face model: the multiple facial reaction generation strategy (Sec. 4.2) and the speaker-listener behaviour synchronisation module (Sec. 4.3).

4.1 The ReactFace Framework

The entire framework of the proposed ReactFace is illustrated in Fig. 2, which is made up of four main modules: (i) a **Multi-modal speaker behaviour encoding and alignment (MSBEA) module** that encodes the input speaker audio-facial behaviour as a set of aligned and semantic meaningful speaker behaviour embeddings; (ii) an **Appropriate**

facial reaction generation (AFRG) module, which predicts a distribution of appropriate facial reactions based on the speaker behaviour embeddings, allowing for the sampling of multiple appropriate facial reaction embeddings; (iii) a **Speaker-listener behaviour synchronisation (SLBS) module** that ensures the generated facial reaction embeddings to be synchronised with the corresponding speaker behaviour; and (iv) a **Facial reaction visualisation (FRV) module** that further decodes a set of 3D and 2D facial frames representing the corresponding sampled appropriate facial reaction. The full pipeline is described as follows:

- **(i) Multi-modal speaker behaviour encoding and Alignment (MSBEA):** Given a speaker behaviour $\mathbf{B}^s_{1:t}$, the MSBEA module first encodes it as a set of verbal speech and non-verbal facial behaviour embeddings. To achieve this, a speech encoder A-Enc (*i.e.*, a state-of-the-art pre-trained speech model wav2vec2.0 [77]) is leveraged to encode the raw audio signal $\boldsymbol{\mathcal{X}}^s_{1:t}$ as a set of speaker speech embeddings $\boldsymbol{A}^s_{1:kt}$, where k denotes that sampling rate for obtaining speech frames is k times over the sampling rate for obtaining facial frames, which depends on the audio encoder's setting. Meanwhile, a 2D facial sequence encoder F-Enc (*i.e.*, a network consisting of 3D convolutional layers and transformer encoder layers) encodes all speaker facial frames $\mathbf{F}^s_{1:t}$ as a set of speaker facial embeddings $\tilde{\boldsymbol{F}}^s_{1:t}$. This can be formulated as:

$$\begin{aligned} \boldsymbol{A}^s_{1:kt} &= \text{A-Enc}(\boldsymbol{\mathcal{X}}^s_{1:t}) \\ \tilde{\boldsymbol{F}}^s_{1:t} &= \text{F-Enc}(\mathbf{F}^s_{1:t}) \end{aligned} \tag{5}$$

Then, a modality alignment module (*i.e.*, a cross-attention transformer with an specific alignment bias explained in [11]) is proposed to align the speaker facial embeddings with the corresponding speech embeddings in a point-to-point manner (please refer to Fig. 5 for the explanation of this term), resulting in a set of aligned facial embeddings $\bar{\boldsymbol{F}}^s_{1:t}$ as:

$$\bar{\boldsymbol{F}}^s_{1:t} = \text{PMA}(\boldsymbol{A}^s_{1:kt}, \tilde{\boldsymbol{F}}^s_{1:t}) \tag{6}$$

where PMA denotes the 'Point-to-Point Modality Alignment', *i.e.*, the operation that aligns facial feature at each time stamp only with the corresponding speech features. To ensure the aligned speaker facial embeddings being semantic meaningful, the MSBEA module also enforces the $\bar{\boldsymbol{F}}^s_{1:t}$ to contain information that allows the corresponding 3D human-style facial reaction sequence $\hat{\mathbf{R}}^s_{1:t}$ can be reconstructed from it. Please refer to Fig. 3 (a) for the visual illustration of the MSBEA module.

- **(ii) Appropriate facial reaction generation (AFRG):** The AFRG involves applying a Conditional Interaction Encoder (CIE) to predict a distribution (*i.e.*, Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^l_t, \boldsymbol{\sigma}^l_t)$) representing multiple appropriate facial reaction segments (*i.e.*, each segment contains w frames) at the current time $[t-w+1, t]$ based on: (1) aligned previous speaker speech and facial behaviour embeddings ($\boldsymbol{A}^s_{1:k(t-w)}$ and $\bar{\boldsymbol{F}}^s_{1:t-w}$); (2) previously predicted and aligned facial reaction embeddings $\hat{\mathbf{R}}^l_{1:t-w}$; and (3) two
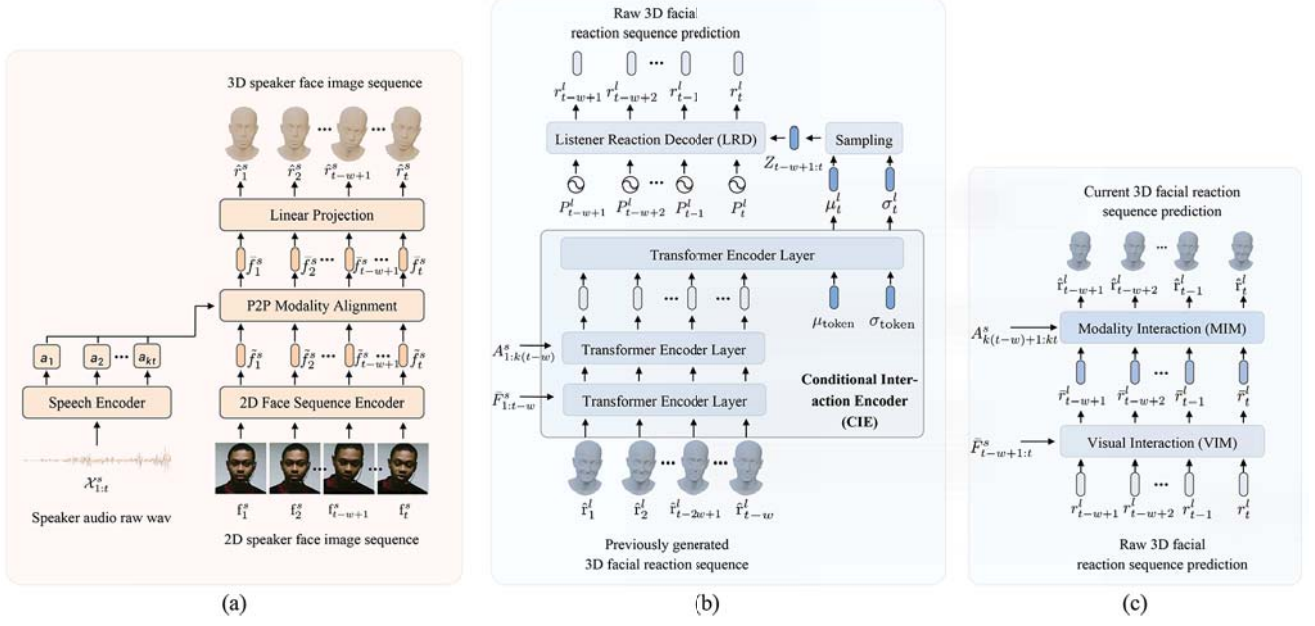
Fig. 3. Illustration of: (a) multi-modal speaker behaviour encoding and alignment (MSBEA) module; (b) appropriate facial reaction generation (AFRG) module; and (c) speaker-listener behaviour synchronisation (SLBS) module.

learnable tokens (*i.e.*, $\boldsymbol{\mu}_{\text{token}}$ and $\boldsymbol{\sigma}_{\text{token}}$). This can be formulated as:

$$\mathcal{N}(\boldsymbol{\mu}_t^l, \boldsymbol{\sigma}_t^l) = \text{CIE}(\boldsymbol{A}_{1:k(t-w)}^s, \bar{\boldsymbol{F}}_{1:t-w}^s, \\ \hat{\mathbf{R}}_{1:t-w}^l, \boldsymbol{\mu}_{\text{token}}, \boldsymbol{\sigma}_{\text{token}}) \tag{7}$$

Consequently, an embedding $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_t^l, \boldsymbol{\sigma}_t^l)$ can be sampled from the learned distribution and we then use a temporal linear interpolation to get a set of embeddings $\boldsymbol{Z}_{t-w+1:t} = \{\boldsymbol{z}_{t-w+1}, \dots, \boldsymbol{z}_t\}$ representing an appropriate facial reaction segment at the time $[t - w + 1, t]$, where each component $\boldsymbol{z}_\tau \in \boldsymbol{Z}_{t-w+1:t}$ denotes an appropriate facial reaction frame of the time stamp τ. **Importantly, this process allows multiple different embeddings that represent various appropriate facial reactions to be sampled, while the use of previously predicted facial reaction embeddings enables the generated current facial reaction to be continuous with previously generated facial reactions.** Then, a listener reaction decoder (LRD) decodes the sampled $\boldsymbol{Z}_{t-w+1:t}$ as w 3D facial reaction frames, which can be formulated as:

$$\boldsymbol{R}_{t-w+1:t}^l = \text{LRD}(\boldsymbol{Z}_{t-w+1:t}, \boldsymbol{P}_{t-w+1:t}^l) \tag{8}$$

where $\boldsymbol{P}_{t-w+1:t}^l = \{\boldsymbol{P}_{t-w+1}^l, \dots, \boldsymbol{P}_t^l\}$ are positional encodings representing temporal positions in the whole generated clip and $\boldsymbol{R}_{t-w+1:t}^l = \{\boldsymbol{r}_{t-w+1}^l, \dots, \boldsymbol{r}_t^l\}$ denote the w sets of 3D facial reaction features representing w current 3D facial reaction frames. These generated 3D frames can be further used to render 2D listener identity-aware face image sequence $\hat{\mathbf{F}}_{1:t}^l$. Please refer to Fig. 3 (b) for the visual illustration of the AFRG module.

- **(iii) Speaker-listener behaviour synchronisation (SLBS).** the SLBS module synchronises the generated facial reactions for every w frames. At the time t, it synchronises the generated w current 3D facial reaction frames $\boldsymbol{R}_{t-w+1:t}^l$ with the corresponding aligned current speaker facial embeddings $\bar{\boldsymbol{F}}_{t-w+1:t}^s$ and speech embeddings $\boldsymbol{A}_{k(t-w)+1:kt}^s$, which can be formulated as:

$$\hat{\mathbf{R}}_{t-w+1:t}^l = \text{SLBS}(\boldsymbol{R}_{t-w+1:t}^l, \bar{\boldsymbol{F}}_{t-w+1:t}^s, \boldsymbol{A}_{k(t-w)+1:kt}^s) \tag{9}$$

  where $\hat{\mathbf{R}}_{t-w+1:t}^l = \{\hat{\mathbf{r}}_{t-w+1}^l, \dots, \hat{\mathbf{r}}_t^l\}$ denotes the synchronised current 3D facial reaction (w frames). Please check Fig. 3 (c) for the visual illustration of the SLBS module.

- **(iv) Facial reaction visualisation.** To translate from generated 3D facial reactions (represented by 3DMM coefficients) to 2D facial frames, we re-train the PIRender ([78]) to make it adapt to our used Face-Verse [79] 3DMM. With the produced 3DMM coefficients and a reference portrait of a specific facial identity, this render is able to output corresponding 2D frames of a human listener facial reaction clip.

4.2 Appropriate Facial Reactions Generation

As described above, the AFRG module consists of three main blocks: (i) a Conditional Interaction Encoder (CIE) that predicts an appropriate facial reaction distribution $\mathcal{N}(\boldsymbol{\mu}_t^l, \boldsymbol{\sigma}_t^l)$; (ii) a sampling block that outputs a latent embedding $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_t^l, \boldsymbol{\sigma}_t^l)$ and gets a set of latent samples $\boldsymbol{Z}_{t-w+1:t} = \{\boldsymbol{z}_{t-w+1}, \dots, \boldsymbol{z}_t\}$ describing an appropriate facial reaction from the predicted distribution; and (iii) a listener facial reaction decoder (LRD) that decodes the sampled facial reaction embeddings as w raw 3D facial reaction frames $\boldsymbol{R}_{t-w+1:t}^l = \{\boldsymbol{r}_{t-w+1}^l, \dots, \boldsymbol{r}_t^l\}$.

The **CIE** is a variational encoder that is composed of three transformer encoder layers. As illustrated in Fig. 3 (b), it takes: (i) $\hat{\mathbf{R}}^l_{1:t-w}$, which represents previously generated facial reaction features that are further synchronized with previous speech ($\boldsymbol{A}^s_{1:k(t-w)}$) and facial ($\bar{\boldsymbol{F}}^s_{1:t-w}$) features of the corresponding speaker behaviour, and (ii) a pair of learnable tokens ($\boldsymbol{\mu}_{token}$ and $\boldsymbol{\sigma}_{token}$) representing reaction distribution prior, as the input. Then, it outputs a pair of vectors $\boldsymbol{\mu}^l_t$ and $\boldsymbol{\sigma}^l_t$, which are respectively the mean and standard deviation of a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^l_t, \boldsymbol{\sigma}^l_t)$ describing multiple appropriate facial reactions in response to the input speaker behaviour. In this process, $\bar{\boldsymbol{F}}^l_{1:t-w}$, $\boldsymbol{\mu}_{token}$ and $\boldsymbol{\sigma}_{token}$ are jointly input as transformer tokens to the CIE. Consequently, three types of tokens can be produced, where two of them ($\boldsymbol{\mu}^l_t$ and $\boldsymbol{\sigma}^l_t$) are employed for the sampling. Note that the positions of $\boldsymbol{\mu}^l_t$ and $\boldsymbol{\sigma}^l_t$ are the same as their prior $\boldsymbol{\mu}_{token}$ and $\boldsymbol{\sigma}_{token}$, where an attention mask is utilized to only calculate attention scores between distribution tokens ($\boldsymbol{\mu}^l_t$ and $\boldsymbol{\sigma}^l_t$) and the corresponding facial reaction features $\bar{\boldsymbol{F}}^l_{1:t-w}$, respectively. This makes the CIE generating each facial reaction by considering the temporal relationship between it and its previous facial reactions (*i.e.*, previously generated facial reactions).

Once the distribution $\mathcal{N}(\boldsymbol{\mu}^l_t, \boldsymbol{\sigma}^l_t)$ is obtained, the **sampling block** first samples a latent embedding $\boldsymbol{z}^*_t$ that represents the t_{th} frame of the predicted facial reaction. In particular, the embedding $\boldsymbol{z}^*_t$ is further updated as $\boldsymbol{z}_t$ based on an embedding $\boldsymbol{z}_{t-w}$ that represents a previously predicted facial reaction frame (*i.e.*, $(t-w)_{th}$ frame). This can be defined as:

$$\boldsymbol{z}_t = \alpha \cdot \boldsymbol{z}_{t-w} + (1 - \alpha) \cdot \boldsymbol{z}^*_t \qquad (10)$$

where α is a momentum parameter to avoid the generated facial reaction frames to be dramatically changed from adjacent previously predicted facial reaction frames. Then, linear interpolation is employed to produce w smooth and continuous latent vectors $\boldsymbol{Z}_{t-w+1:t} = \{\boldsymbol{z}_{t-w+1}, \dots, \boldsymbol{z}_t\}$ corresponding to the facial reaction predictions at the time $[t-w+1, t]$, which can be formulated as:

$$\boldsymbol{z}_\tau = \boldsymbol{z}_{t-w} + (\tau - (t-w))(\boldsymbol{z}_t - \boldsymbol{z}_{t-w})/w, \tau \in (t-w, t] \quad (11)$$

This process further allows the smooth transition for the w generated facial reaction frame from the facial display of the $(t-w)_{th}$ frame to that of the t_{th} frame. In short, the w facial reaction frames are all generated based on the previously predicted $(t-w)_{th}$ facial reaction frame, which ensures the facial reactions generated for the time $[t-w+1, t]$ to be smooth and continuous with the previously generated facial reactions.

Finally, the **LRD** is introduced to generate w 3D facial reaction frames from the sampled embeddings $\boldsymbol{Z}_{t-w+1:t} = \{\boldsymbol{z}_{t-w+1}, \dots, \boldsymbol{z}_t\}$. It performs cross-attention between a sequence of positional encodings $\boldsymbol{P}^l_{t-w+1:t}$ (*i.e.*, treated as the keys and values) and the $\boldsymbol{Z}_{t-w+1:t}$ (*i.e.*, treated as the queries), which is computed as:

$$\boldsymbol{R}^l_{t-w+1:t} = \text{softmax}\left(\frac{\boldsymbol{Z}_{t-w+1:t}(\boldsymbol{P}^l_{t-w+1:t})^T}{\sqrt{d_p}}\right)\boldsymbol{P}^l_{t-w+1:t} \qquad (12)$$

where $\sqrt{d_p}$ denotes the channel number of the key $\boldsymbol{P}_{t-w+1:t}$. These positional encodings are obtained based on the sinusoidal functions proposed in [80] as:

$$\begin{aligned} \boldsymbol{P}^l_{(t,2i)} &= \sin(t/10000^{2i/d_p}) \\ \boldsymbol{P}^l_{(t,2i+1)} &= \cos(t/10000^{2i/d_p}) \end{aligned} \qquad (13)$$

where each positional encoding $\boldsymbol{P}^l_t$ embeds a specific temporal cue into its corresponding frame token, which makes the generation of each facial reaction frame considering its global position at the time $(0, t]$. This way, each facial reaction embedding in $\boldsymbol{R}^l_{t-w+1:t}$ would be aware of its temporal position within the entire facial reaction sequence.

4.3 Speaker-listener Behaviour Synchronisation

Since human facial reaction is a spatio-temporal signal, which is highly correlated with the corresponding speaker behaviour in the temporal dimension, properly synchronising the generated facial reactions with the corresponding speaker behaviours is a key to ensure them to be appropriate and realistic.

As shown in Fig. 3 (c), the **SLBS** module is made up of two models: a **Visual Interaction Model (VIM)** and a **Modality Interaction Model (MIM)**. Specifically, the VIM first synchronises the generated current raw 3D facial reactions (*i.e.*, the embeddings $\boldsymbol{R}^l_{t-w+1:t}$) with the corresponding aligned speaker facial embeddings $\bar{\boldsymbol{F}}^s_{t-w+1:t}$ produced by the MSBEA module as:

$$\bar{\boldsymbol{R}}^l_{t-w+1:t} = \text{VIM}(\boldsymbol{R}^l_{t-w+1:t}, \bar{\boldsymbol{F}}^s_{t-w+1:t}) \qquad (14)$$

Then, the MIM is introduced to further synchronise these visually synchronised facial reactions $\bar{\boldsymbol{R}}^l_{t-w+1:t}$ with the corresponding speaker speech embeddings (*i.e.*, speaker speech embeddings $\boldsymbol{A}^s_{k(t-w)+1:kt}$), respectively. This process can be formulated as:

$$\hat{\mathbf{R}}^l_{t-w+1:t} = \text{MIM}(\bar{\boldsymbol{R}}^l_{t-w+1:t}, \boldsymbol{A}^s_{k(t-w)+1:kt}) \qquad (15)$$

This way, the final generated facial reaction frames $\hat{\mathbf{R}}^l_{t-w+1:t}$ would be synchronised with both non-verbal facial behaviours and verbal speech behaviours expressed by the corresponding speaker.

In this paper, both VIM and MIM are cross-attention operations, where a novel alignment bias $\boldsymbol{M}^b$ is proposed under the assumption that the speaker behaviours that are temporally closer to the time i would have more influence on the facial reaction $\boldsymbol{R}^l_i$ generated at the time i. In particular, this alignment bias is proposed for two purposes: (i) synchronising the generated facial reactions with the corresponding speaker behaviours (both non-verbal and verbal behaviours) in the temporal dimension; and (ii) ensuring the synchronisation of each facial reaction frame $\boldsymbol{r}^l_i/\bar{\boldsymbol{r}}^l_i$ is only influenced by historical speaker behaviours as well as the previously generated facial reactions, *i.e.*, frames previous to the position i. This novel cross-attention operation is formulated as:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{M}^b) = \text{softmax}\left(\frac{\boldsymbol{Q}(\boldsymbol{K})^T}{\sqrt{d_k}} + \boldsymbol{M}^b\right)\boldsymbol{V} \quad (16)$$

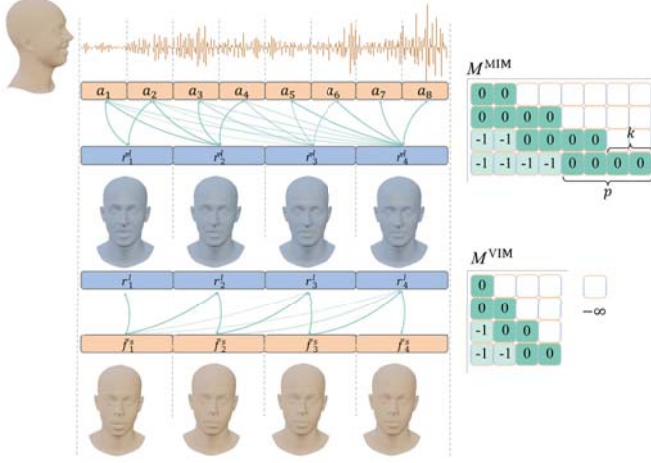where the speech/facial behaviours embedding expressed by the speaker are input as the keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ in
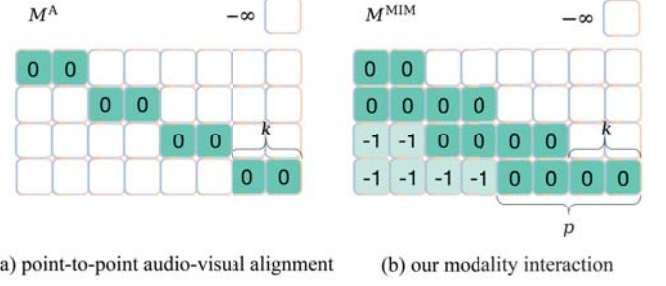
Fig. 4. Illustration of visual interaction and modality interaction

(a) point-to-point audio-visual alignment

(b) our modality interaction

Fig. 5. Comparison of: (a) point-to-point (P2P) audio-visual alignment proposed in Faceformer [11] and (b) our proposed modality interaction. P2P audio-visual alignment primarily establishes corrections solely between a given visual frame and its corresponding audio frames at the identical timestamp (achieved through the assignment of zero attention weights) while effectively precluding the visual frame from attending to any other frames (achieved through the assignment of negative infinity attention weights). Our proposed modality interaction mechanism establishes both long-range and short-range relationships between a visual frame at a specific timestamp and audio frames occurring at both the current and previous time instances. This approach significantly enriches the capacity to capture intermodal dependencies and contextual nuances across disparate frames within the audio-visual data stream.

VIM/MIM, while the generated facial reactions are treated as the queries Q.

The M^b is a matrix that has similar forms in VIM and MIM, which can be denoted as M^{VIM} and M^{MIM}. Here, the component in i_{th} row and j_{th} column (*i.e.*, $M_{i,j}^{\mathrm{VIM}}$ and $M_{i,j}^{\mathrm{MIM}}$) of M^{VIM} and M^{MIM} can be computed as:

$$M_{i,j}^{\mathrm{VIM}} = \begin{cases} -\left\lfloor \frac{i-j}{p} \right\rfloor, & j \leq i, \\ -\infty, & j > i \end{cases}$$
$$M_{i,j}^{\mathrm{MIM}} = \begin{cases} -\left\lfloor \frac{ki-j}{kp} \right\rfloor, & j \leq k \times i, \\ -\infty, & j > k \times i \end{cases} \tag{17}$$

where i and j denote that the attention is conducted between the i_{th} facial reaction frame and the j_{th} speaker facial frame/speech frame, *i.e.*, k represents that the sampling rate for obtaining speech frames is k times over the sampling rate for obtaining facial frames. Here, p is a parameter that divides all processed speaker facial/speech frames and the corresponding facial reaction frames into several units, where each unit contains p facial frames and $k \times p$ consecutive speech frames. This is because nearby frames (within a window) usually share similar behavior patterns. Setting parameter p considers this human behaviour nature and prevents excessively small attention weights to distant previous context, ensuring sufficient attention to long-range information. Subsequently, the attention operation treats all frames in each unit equally. Fig. 4 further illustrates our temporal bias matrix, where the upper triangle of M^b contains negative infinity to ensure that each facial reaction frame is synchronised only based on previous and current speaker behaviours, without considering future information, which is not known in real-world online applications. More importantly, Eqa. 17 demonstrated that in both VIM and MIM, the synchronisation of the i_{th} facial reaction frame only takes the speaker speech and facial behaviour frames that prior or equal to it into consideration. This is because if $j > i$ / $j > ki$, $M_{i,j}^{\mathrm{VIM}} = -\infty$ / $M_{i,j}^{\mathrm{MIM}} = -\infty$, and subsequently the output of the attention operation defined in Eqa. 16 would be zero. In summary, the proposed parameter M^b determines the influence of speaker speech and facial frames on the synchronization of the i_{th} facial reaction frame, based

on their temporal proximity. Specifically, the speaker speech and facial frames that exhibit closer temporal distances to the i_{th} facial reaction frame would have more substantial impacts on its synchronization.

Different from previous work [11] that only individually conducts point-to-point (p2p) audio-visual alignment at each time stamp, aiming at matching human's verbal signals to its visual lip movements synchronously in talking head generation task, the proposed SLBS module aligns each generated facial reaction frame with both current and past speaker speech/facial behaviours. This allows our SLBS module to capture long-range dependencies between speaker audio-visual behaviour and the generated facial reactions, as listener's facial reactions are always influenced by both current and past interactions in dyadic interaction.

4.4 Loss functions and training strategy

To allow the proposed framework to be easily reproduced, a simple end-to-end training strategy is proposed in Sec. 4.4.1, where several loss functions are jointly employed to ensure the well-trained model to generate appropriate, diverse, realistic and synchronised facial reactions (Sec. 4.4.2).

4.4.1 Training Strategy

We propose an end-to-end strategy to train our React-Face model, which is achieved by jointly optimising five loss functions as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}}^a + \mathcal{L}_{\mathrm{rec}}^s + \lambda_{\mathrm{kl}} \mathcal{L}_{\mathrm{kl}} + \lambda_{\mathrm{smo}} \mathcal{L}_{\mathrm{smo}} + \lambda_{\mathrm{div}} \mathcal{L}_{\mathrm{div}} \tag{18}$$

where $\mathcal{L}_{\mathrm{rec}}^a$ represents an *appropriate facial reaction reconstruction loss* that focuses on enforcing the model to generate appropriate and realistic (semantic meaningful) facial reactions; $\mathcal{L}_{\mathrm{rec}}^s$ is a reconstruction loss that supervises the 3D speaker facial reaction sequence generation achieved by our MSBEA module; $\mathcal{L}_{\mathrm{kl}}$ is a *Kullback-Leibler (KL) divergence loss* that constrains the learned appropriate facial reaction distribution; $\mathcal{L}_{\mathrm{smo}}$ is a *temporal smooth loss* that avoids jitters in the

Algorithm 1 Training Strategy

Input: 2D speaker facial image sequence $\mathbf{F}^s_{1:T}$, 3D speaker facial image sequence $\mathbf{R}^s_{1:T}$, Raw speaker audio signal $\boldsymbol{\mathcal{X}}^s_{1:t}$, the number of sampling M, window size w

$NT \leftarrow \lceil T/w \rceil$

for $i = 1$ to NT **do**

  $t \leftarrow i \times w$;

  Obtain $\hat{\mathbf{R}}^s_{1:t}$, $\boldsymbol{A}^s_{1:t}$, and $\bar{\boldsymbol{F}}^s_{1:t}$ though MSBEA module;

  Obtain $\boldsymbol{R}^l_{t-w+1:t}$ though AFRG module;

  Synchronize $\boldsymbol{R}^l_{t-w+1:t}$ with $\bar{\boldsymbol{F}}^s_{t-w+1:t}$ and $\boldsymbol{A}^s_{k(t-w+1):kt}$

  through SLBS module and obtain $\hat{\mathbf{R}}^l_{t-w+1:t}$;

end for

Calculate speaker face reconstruction loss as in Eq. (19).

Calculate appropriate facial reaction reconstruction loss, KL divergence loss and temporal smooth loss as in Eq. (20), Eq. (22) and Eq. (23), respectively.

for $j = 2$ to M **do**

  Sample new reaction predictions $\hat{\mathbf{R}}^l_{j,1:T}$

end for

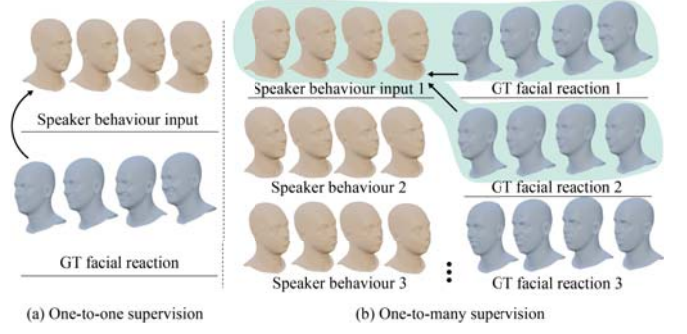Calculate diversity loss as in Eq. (21)

Fig. 6. Illustration of appropriate facial reaction reconstruction loss. We replace (a) one-to-one GT reaction supervision with (b) supervision by appropriate reactions (GT 1, GT 2, and so on).

generated facial frames; and $\mathcal{L}_{\text{div}}$ denotes the *energy-based diversity loss* that enforces the model to generate diverse facial reactions. Meanwhile, a set of weights $\lambda_{\text{kl}}, \lambda_{\text{smo}}, \lambda_{\text{div}}$ are also employed to decide the relative importance of the aforementioned five loss functions, respectively, where λ_{kl} and λ_{div} are used to balance the trade-off between realism/appropriateness and diversity of the generated facial reactions. For clarity, we present the pseudo-code in Algorithm 1 to outline our training strategy.

4.4.2 Loss Functions

To ensure that MSBEA module encodes human interpretable cues from the input speaker facial behaviour, we first employ a reconstruction loss to supervise its outputs $\hat{\mathbf{R}}^s_{1:t}$, which is defined as:

$$\mathcal{L}^s_{\text{rec}} = \mathscr{L}_1(\hat{\mathbf{R}}^s_{1:T}, \mathbf{R}^s_{1:T}) \tag{19}$$

where $\mathscr{L}_1$ denotes the smooth L1 loss [81] and T denotes the total number of frames in the sequence.

To enforce the model to learn a distribution $\mathcal{N}(\mu^l_t, \sigma^l_t)$ that can represent multiple appropriate and realistic facial reactions in response to an input speaker behaviour, the **appropriate facial reaction reconstruction loss ($\mathcal{L}^a_{\text{rec}}$)** is computed between the generated 3D facial reaction sequence (*i.e.,* $\hat{\mathbf{R}}^l_{1:T}$) and the 3D facial sequence corresponding to its most similar appropriate real facial reaction $\mathbf{R}^l_{1:T}(i)$ in the training set as:

$$\mathcal{L}^a_{\text{rec}} = \min(\{\mathscr{L}_1(\hat{\mathbf{R}}^l_{1:T}, \mathbf{R}^l_{1:T}(i)) | \mathbf{R}^l_{1:T}(i) \in \mathbb{N}(\mathbf{B}^s_{1:T})\}) \tag{20}$$

where $\mathbf{R}^l_{1:T}(i)$ denotes an appropriate real facial reaction in response to the input speaker behaviour, which has the highest similarity to the generated facial reaction among all appropriate real facial reactions of the input speaker behaviour; $\mathbf{B}^s_{1:T}$ is the input speaker behaviour and $\mathbb{N}(\mathbf{B}^s_{1:T})$ represents a set of real facial reactions that are appropriate for responding to $\mathbf{B}^s_{1:T}$. Here, the appropriate real facial

reactions for each speaker is obtained based on the automatic appropriate labelling strategy proposed in [10]. While each speaker behaviour may lead to multiple different facial reactions, the proposed appropriate facial reaction reconstruction loss $\mathcal{L}^a_{\text{rec}}$ does not propagate a large loss value if the generated facial reaction is similar to any of the appropriate real facial reactions, *i.e.,* even the generated facial reaction is not similar to the corresponding listener's GT facial reaction. Please refer to Fig. 6 and [10] for more detailed definition of appropriate facial reactions.

Since one goal of the developed model is to generate multiple different facial reactions from each input speaker behaviour, we employ an **energy-based diversity loss ($\mathcal{L}_{\text{div}}$)** [52] to compute the difference between each pair of generated facial reactions (*i.e.,* we sample M facial reactions for each speaker behaviour) as:

$$\mathcal{L}_{\text{div}} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i}^{M} \exp(-\frac{\|\hat{\mathbf{R}}^l_{i,1:T} - \hat{\mathbf{R}}^l_{j,1:T}\|_2^2}{\sigma_d}) \tag{21}$$

where an RBF kernel with scale σ_d is used for measuring the distance between a pair of facial reactions ($\hat{\mathbf{R}}^l_{T,i}$ and $\hat{\mathbf{R}}^l_{T,j}$), and $M = 3$ at the training stage (we follow [22] to set $M = 10$ at the testing stage, *i.e.,* generating 10 facial reactions from each speaker behaviour). This way, minimizing the diversity loss $\mathcal{L}_{\text{div}}$ allows the model to generate facial reactions toward a lower-energy (**higher diversity**). In summary, the use of the above two loss functions achieves two goals: (1) making the learned distribution $\mathcal{N}(\mu^l_t, \sigma^l_t)$ to well represent corresponding appropriate real facial reactions; and (2) enabling model to generate diverse facial reactions from each speaker behaviour.

To constrain distribution parameters μ^l and σ^l for the purpose of maintaining realism of the generated facial reactions, we apply the canonical **Kullback-Leibler (KL) divergence loss ($\mathcal{L}_{kl}$)** to make the learned $\mathcal{N}(\mu^l_t, \sigma^l_t)$ approaching the standard normal distribution $\mathcal{N}(0, I)$ as:

$$\mathcal{L}_{\text{kl}} = \text{KL}(\phi, \boldsymbol{\omega}) \tag{22}$$

This strategy has been widely used in [47], [56], [64], [82] to regularize the learned Gaussian distribution $\phi = \mathcal{N}(\mu^l_t, \sigma^l_t)$ with the standard normal distribution $\boldsymbol{\omega} = \mathcal{N}(0, I)$.
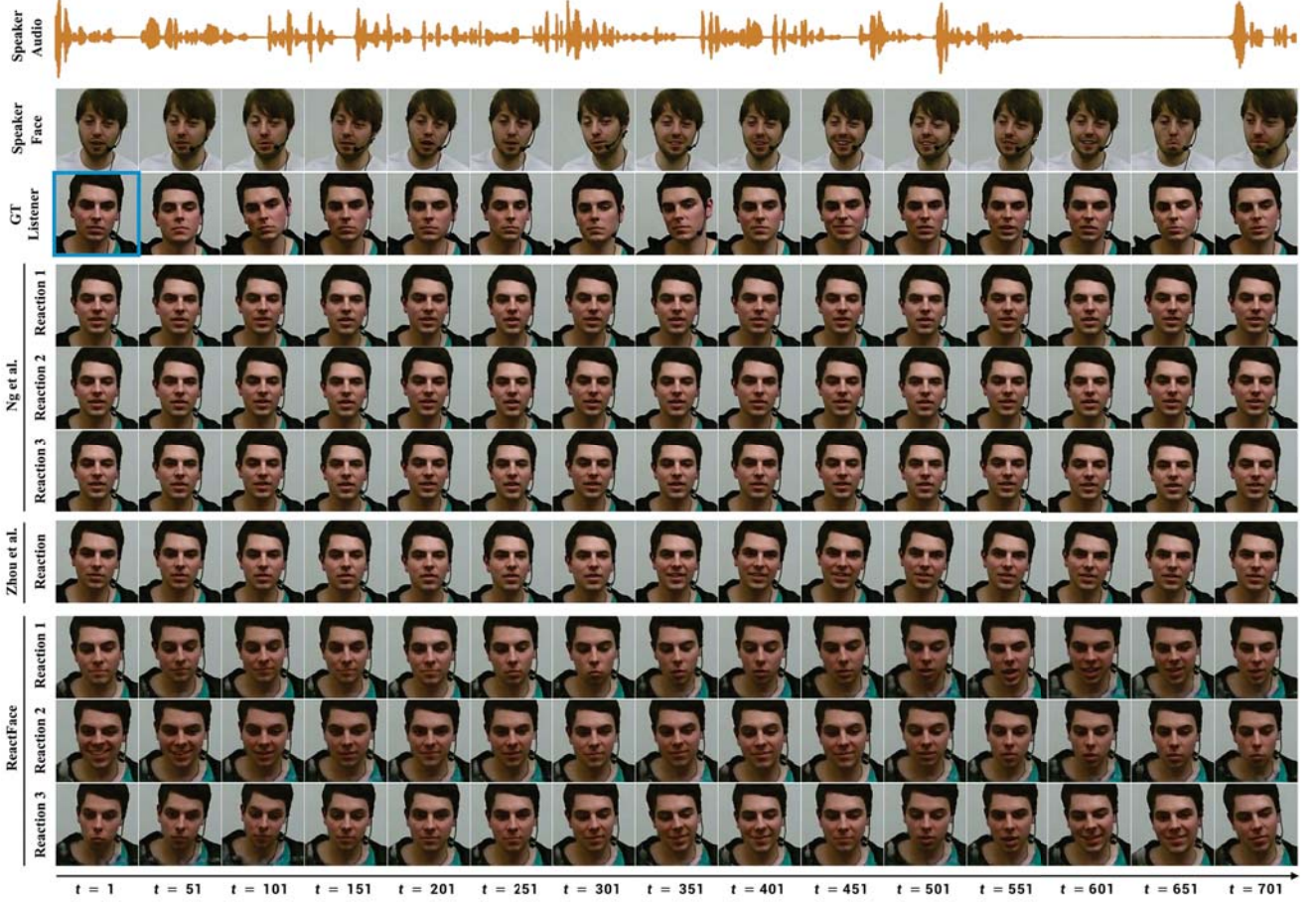
Fig. 7. Qualitative results: Our approach (ReactFace) can generate multiple different but appropriate and synchronised facial reactions in response to each audio-visual speaker behaviour (including its face frames and audio). The portrait reference (the 1st GT frame) is framed in blue.

Finally, to avoid jitters in the generated facial frames, we utilize a **temporal smooth loss ($\mathcal{L}_{\mathbf{smo}}$)** to constrain facial motion variations between adjacent frames as:

$$\mathcal{L}_{\text{smo}} = \frac{1}{(T-2)} \sum_{t=2}^{T} \mathcal{L}_1((\hat{\mathbf{r}}_t - \hat{\mathbf{r}}_{t-1}), (\hat{\mathbf{r}}_{t-1} - \hat{\mathbf{r}}_{t-2})) \quad (23)$$

This loss term constrains the change velocities ($\hat{\mathbf{v}}_t = (\hat{\mathbf{r}}_t - \hat{\mathbf{r}}_{t-1})$ and $\hat{\mathbf{v}}_{t-1} = (\hat{\mathbf{r}}_{t-1} - \hat{\mathbf{r}}_{t-2})$) of adjacent frames to be close, making translation between adjacent frames to be more smooth.

5 EXPERIMENTS

In this section, we first present the details of the employed datasets in Sec. 5.1. All experimental setups are provided in Sec. 5.2. We then compare qualitative and quantitative performances between our ReactFace and several re-produced baselines in Sec. 5.3 and Sec. 5.4. Finally, ablation and perceptual study results, and failure cases are discussed in Sec. 5.5, Sec. 5.7, and Sec. 5.9, respectively.

5.1 Datasets

We evaluate our ReactFace on a hybrid video conference dataset provided by REACT2023 challenge [22], which is made up of 2,962 dyadic interaction sessions (1,594 training examples, 562 validation examples and 806 test examples) coming from two video conference datasets: RECOLA [83] and NOXI [84]. Each session is 30s long and contains a pair of audio-visual clips describing two subjects' interaction behaviours. The objective appropriate facial reaction annotations are also obtained using the strategy proposed in [22]. Due to the license issue, this paper does not include UDIVA [85] dataset. Please refer to the challenge site [1] for more detailed description of the employed dataset, license and appropriateness labels used in our experiments.

5.2 Experimental Setup

5.2.1 Implementation Details

In our experiments, the resolution of the input speaker image sequence is set as 224×224, based on which the ReactFace model is trained with an AdamW optimizer [86] with a fixed learning rate of $2e^{-5}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Our minibatch size is set to 4. For hyper-parameters balancing loss terms in Sec. 4.4.1, we set $\lambda_{\text{kl}} = 1e^{-5}$, $\lambda_{\text{smo}} = 10$, and $\lambda_{\text{div}} = 100$. The momentum parameter α in Sec. 4.2 is empirically set to 0.999. Our code is implemented in PyTorch [87] using a single Tesla A100 GPU with 80G memory and runs for total 200 epochs.

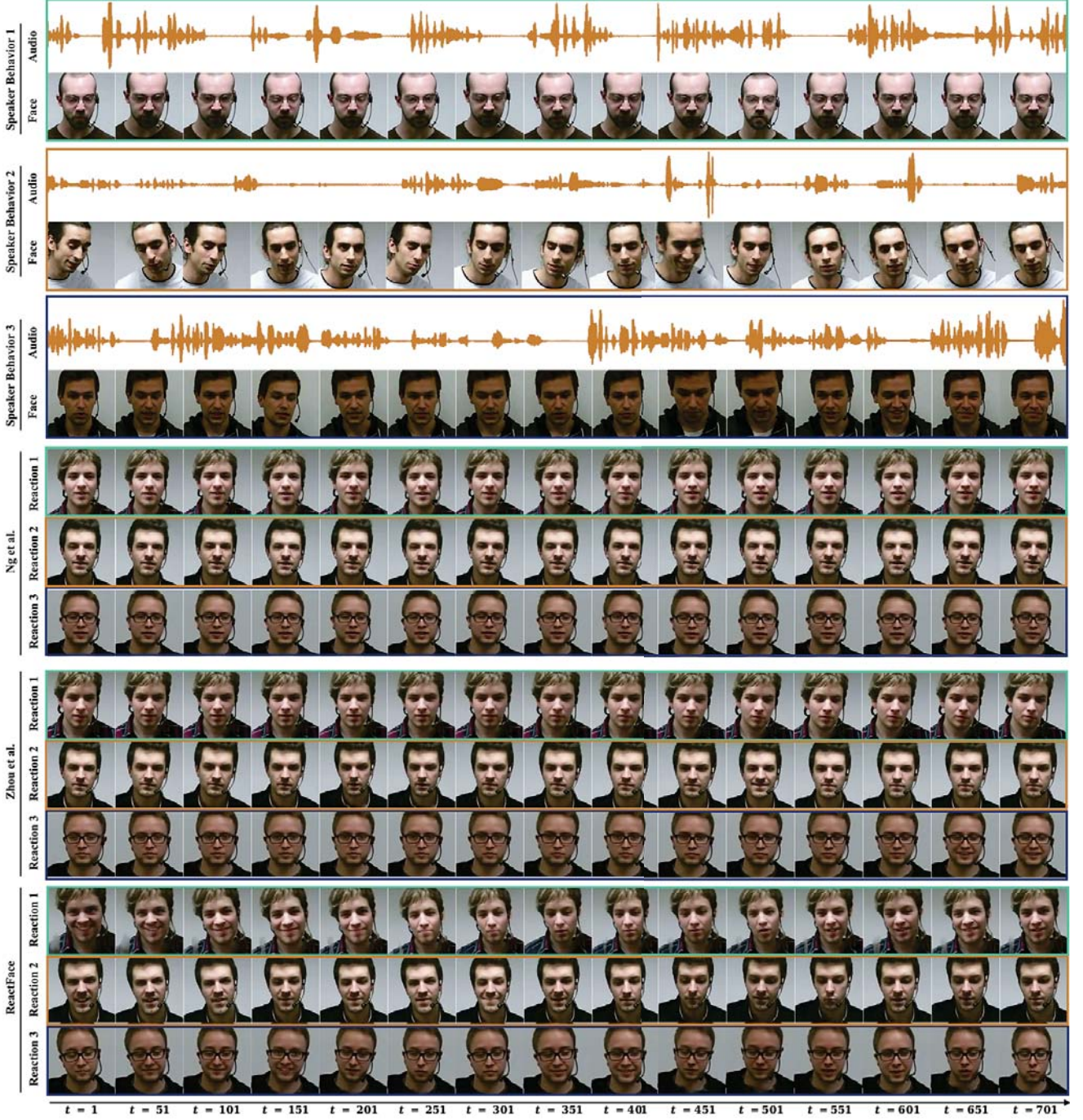1. https://sites.google.com/cam.ac.uk/react2023/home

Fig. 8. Qualitative results: Our approach (ReactFace) can generate facial reactions with high diversity in response to different speaker behaviours.

To extract 3D expression and pose coefficients from videos, we leverage the state-of-the-art 3DMM models Face-Verse [79], which estimates the pose, expression, shape, texture and lighting parameters, and defines expression coefficients ψ ($|\psi| = 52$) along with pose coefficients θ (i.e., 3-dimensional translation and the 3-dimensional rotation), resulting in totally 58 coefficients for each face frame. The reason for choosing FaceVerse 3DMM system [79] is that each of its expression coefficient corresponds to an ARKit blendshape, which has an explicitly and human interpretable definition such as 'browInnerUp', 'eyeLookDown-

Right ', ' jawOpen', ' mouthFunnel', 'noseSneerRight' and 'tongueOut'. In contrast to the principal component analysis (PCA) blendshapes employed in conventional 3D Morphable Model (3DMM) systems, the ARKit face blendshapes [2] used in FaceVerse constitute discrete micro-expression components. These components afford a finer-grained representation of facial muscle movements, enabling a more precise and nuanced characterization of facial expressions. In short, the use of ARKit blendshapes enhances the degree

2. https://arkit-face-blendshapes.com

TABLE 2
Quantitative results comparison between the ReactFace, baselines and four state-of-the-art method. The best result in each column is marked in bold.

Method	Diversity			Realism	Appropriateness	Synchrony	Speed
	FRDvs ($\uparrow$)	FRVar ($\uparrow$)	FRDiv ($\uparrow$)	FRRea ($\downarrow$)	FRCorr ($\uparrow$)	FRSyn ($\downarrow$)	FPS ($\uparrow$)
GT	0.0374	0.0120	-	282.03	9.480	48.46	-
Mirror	0.0374	0.0120	0	282.03	0.936	42.65	-
Random	0.0415	0.0202	0.0041	477.49	0.127	45.82	-
NN motion	0.0420	0.0199	0	452.38	0.334	46.90	15.36
NN audio	0.0464	0.0218	0	496.25	0.017	47.67	3.32
Trans-AE	0.0063	0.0003	0	599.35	**0.245**	45.01	119.73
LFI [20]	0.0247	0.0109	0.0138	734.06	0.048	45.92	24.15
Ng et al. [1]	0.0015	0.0102	0.0003	691.24	0.059	45.70	91.75
Zhou et al. [2]	0.0106	0.0039	0	527.47	0.104	45.24	**150.14**
ReactFace (Ours)	**0.0409**	**0.0159**	**0.0395**	**424.46**	0.197	**43.94**	54.96

of controllability in the process of facial expression generation. Both expression and pose coefficients are then used for image rendering. In this paper, we fine-tune the PIRender [78] to fit the target 3DMM system and utilize it to translate predicted 3DMM coefficients to a portrait of the listener.

5.2.2 Baselines

We leverage five baselines (*e.g.,* Mirror, Random, NN motion, NN audio and Trans-AE) and reproduce three recently proposed facial reaction generation methods (*e.g.,* LFI [20], Ng et al. [1] and Zhou et al. [2]) as the baseline facial reaction generation models, where we strictly followed the training settings given in the corresponding papers:

- **Mirror:** Mirror refers to replicating the facial movements of the speaker.
- **Random:** Random involves sampling reactions from Gaussian distributions.
- **NN motion:** NN motion searches for the nearest neighbor (NN) of the current speaker motion segment and returns the corresponding listener segment, a method commonly used in graphics synthesis.
- **NN audio:** NN audio searches for the NN through the speaker's auditory signals.
- **Trans-AE:** An autoencoder model with transformer architecture, which has the same speaker behaviour encoder and alignment module as that in our ReactFace but a straightforward decoder outputting reactions.
- **LFT [20]:** a flow-based generative model that has demonstrated strong performance in generating listener head motions.
- **Ng et al. [1]:** A VQ-VAE [88] based model with a motion-audio cross attention transformer branch encoding speaker's 3DMM coefficients and audio signals.
- **Zhou et al. [2]:** a simple LSTM model which takes speaker's 3DMM coefficients, diverse audio features and initial listener 3DMM coefficients at first frame as input and outputs listener 3DMM coefficients in a sequence-to-sequence manner.

To ensure a fair and comprehensive comparison, we additional present our results on the RLD dataset, which has been employed in Zhou et al. [2]. To produce final facial reactions, we followed the same protocol as our approach, including the utilization of the identical 3D Morphable Model (3DMM) system [89] as employed in [2].

5.2.3 Evaluation Metrics

In this paper, we follow [10] to evaluate four aspects (*e.g.,* appropriateness, diversity, realism, and synchrony) of the generated facial reactions, and also assess the generation speed. Firstly, we use the Facial Reaction Correlation (**FRCorr**) metric proposed in to assess the **appropriateness** by measuring the DTW distance and correlations between the generated facial reactions and their most similar appropriate real facial reactions.

Then, diversity is evaluated by (1) inter-condition diversity **FRDvs**: the average MSE distance between all pairs of generated reactions conditioned on different speaker behaviours (N different speaker sequences) to measure diversity across speaker behaviour conditions, which is computed as:

$$\text{FRDvs} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N-1}\sum_{k=i+1}^{N}\text{MSE}(\hat{\mathbf{R}}_{i,j,1:T}^{l}, \hat{\mathbf{R}}_{k,j,1:T}^{l})}{N(N-1)M}; \tag{24}$$

where N, M, and T denotes the number of sequence pairs, sampling times for diverse facial reaction generation and the length of a video clip, respectively. (2) inter-sample diversity **FRDiv**: the average $L2$ distance between all pairs of generated reactions in response to the same speaker behaviour, which is computed as:

$$\text{FRDiv} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}\|\hat{\mathbf{R}}_{i,j,1:T}^{l} - \hat{\mathbf{R}}_{i,k,1:T}^{l}\|_{2}^{2}}{NM(M-1)}; \tag{25}$$

and (3) inter-frame diversity **FRVar**: average variation of generated reaction frames within a clip (T frames) to measure diversity within one generated facial reaction sequence, which is computed as:

$$\text{FRVar} = \frac{1}{N \times M}\sum_{i=1}^{N}\sum_{j=1}^{M}\text{var}(\hat{\mathbf{R}}_{i,j,1:T}^{l}); \tag{26}$$

Thirdly, realism **FRRea** is measured using Fr'echet Video Distance (FVD) [90] metric. Finally, we employ the Time Lagged Cross Correlation (TLCC) metric to assess the **synchrony (FRSyn)** between speaker behaviours and the corresponding generated facial reactions. Please refer to [10] for the more details of these metrics. We utilize Frames Per Second (FPS) as a key metric to assess the generation speed of each generative model.

5.3 Qualitative Results

We present a qualitative comparison in Fig. 7 to visually compare the appropriateness, diversity, realism, and synchrony of the facial reactions generated by our ReactFace model with two state-of-the-art methods [1], [2]. The result shows that facial reactions generated by our approach have **higher diversity** and are better synchronised with the speaker behaviour compared to competitors that can only produce perturbations based on a single facial expression. For example, at time $t = 601$, our ReactFace generated a smiling facial reaction that corresponds to the happy facial expression of the speaker. Also, both speaker facial expression and the generated facial reaction are transferred to different states at the time $t = 701$. In contrast, the baselines produced almost identical facial expressions throughout the sequence, and not synchronised with speaker behaviours. In summary, our ReactFace produced appropriate, diverse, continuous and synchronised facial reactions, which are superior to baselines.

5.4 Quantitative Results

We report the quantitative results in Table 2. It is clear that the proposed ReactFace generated facial reactions with the best diversity and realism and synchrony among all competitors (*i.e.,* recently proposed state-of-the-art facial reaction generation approaches) [1], [2], [20] as well as the baseline Trans-AE (*i.e.,* a transformer-based AutoEncoder that has the same architecture of our speaker behaviour encoding network, which directly maps the speaker behaviour features to facial reaction predictions). We explain such results as the training process of these competitors only pairs a single real facial reaction label with the input speaker behaviour, which lead to their models being optimised to generate the average facial behaviour of multiple appropriate facial reactions in response to each speaker behaviour, *i.e.,* this results in suboptimal diversity and appropriateness of the generated facial reactions. Also, the novel SLBS module also ensures the ReactFace to always generate facial reactions synchronised with speaker behaviours, making our ReactFace largely outperform competitors in terms of the synchrony. More importantly, we also found that the facial reactions generated by ReactFace are also appropriate, *i.e.,* it achieved competitive results in term of appropriateness metric with 0.197 FRCorr, which is ranked at the second among all compared approaches. Although Trans-AE achieved the best appropriateness, it sacrifices the diversity of generated results, with 0.0063 FRDvs and 0.0003 FRVar.

Additionally, our ReactFace model achieves an impressive performance of over 54.96 FPS on a single Tesla A100 GPU, making it highly suitable for real-time interactive scenarios that require a minimum of 30 FPS. It is important to note that ReactFace handles raw audio and video frame processing directly, without the need for pre-processing steps such as mapping video frames to 3DMM coefficients. In contrast, the approaches by Zhou et al. [2], Ng et al. [1], LFI [20], and NN motion require an additional pre-processing step to extract 3DMM coefficients from video frames. The FPS evaluation also accounts for the entire Facial Reaction Visualization process.

5.5 Ablation Studies

In this section, we provide a set of ablation study results to systematically investigate: (i) contributions of different modalities; (ii) effectiveness of the proposed components in AFRG and SLBS modules; and (iii) contributions of different losses. In addition, Fig. 9 further provides visualisation results to explain the temporal alignments achieved by our proposed MIM and VIM.

Contributions of different modalities: Table 3 evaluates the importance of different modality contributing to the final generated facial reactions. Our analysis yielded three key findings: (i) both speaker facial and speech behaviours are important in evoking listener's facial reactions; (ii) excluding facial modality resulted in low synchrony of the generated facial reactions; and (iii) in our framework, speaker speech behaviours made more contributions than facial behaviours in generating more diverse, realistic, and appropriate facial reactions.

Effectiveness of MIM and VIM in the SLBS module: Table 4 evaluates the importance of different components of our ReactFace. The results show that MIM (interaction between speaker audio signals and listener reactions) made large contributions to the Diversity (FRDiv) and Appropriateness (FRCorr) of the generated facial reactions. Meanwhile, our VIM (interaction between speaker facial behaviour and listener reactions) plays a key role in improving the diversity, realism, and appropriateness performances. While employing either VIM or MIM improves the performance, combining the two leads to substantial gains in synchrony, appropriateness, and diversity of generated reactions. This improvement may come from enriched interaction features derived from multiple perspectives, including both speaker's verbal and non-verbal cues.

Comparison between MIM in the SLBS module and p2p audio-visual alignment [11]: The comparative analysis presented in Table 5 shows different results achieved by the modality interaction module (MIM) and the p2p audio-visual alignment method used in Faceformer [11] performed. These results clearly show that MIM outperforms the p2p method in many aspects. Specifically, the p2p audio-visual alignment technique that has been leveraged to synchronize visual lip movements with human audio at individual timestamps lacks consideration for comprehensive conversation context, leading to poorer appropriateness (FRCorr) and less varied reaction patterns (in terms of FRVar and FRDiv) and reaction realism (FRRea).

TABLE 3
Ablation study results achieved for different components.

Speech	Face	FRVar	FRDiv	FRRea	FRCorr	FRSyn
✓		0.0173	0.0351	450.33	0.168	46.33
	✓	0.0155	0.0217	597.35	0.123	44.38
✓	✓	0.0159	0.0395	424.46	0.197	43.94

TABLE 4
Ablation study results achieved for different components.

MIM	VIM	FRVar	FRDiv	FRRea	FRCorr	FRSyn
		0.1824	0.0001	526.09	0.102	45.42
✓		0.0107	0.0180	668.38	0.123	46.30
	✓	0.0190	0.0315	460.75	0.184	45.36
✓	✓	0.0159	0.0395	424.46	0.197	43.94

TABLE 5
Ablation study results achieved for different audio-visual alignment strategies.

	FRVar	FRDiv	FRRea	FRCorr	FRSyn
P2P audio-visual alignment	0.0129	0.0321	594.14	0.185	44.40
MIM	0.0159	0.0395	424.46	0.197	43.94

TABLE 6
Ablation study results achieved for different audio-visual alignment strategies.

	FRVar	FRDiv	FRRea	FRCorr	FRSyn
sampling w/o momentum	0.0035	0.1002	614.36	0.146	45.56
sampling w/ momentum	0.0159	0.0395	424.46	0.197	43.94

Effectiveness of the momentum employed in the AFRG module (in sampling block): As shown in Table 6, it is clear that momentum operation is essential for the sampling process, ensuring the appropriateness (FRCorr) and realism (FRRea) of the generated facial reactions. The reason is that the absence of momentum in reaction sampling leads to an abrupt transition between the previous and current frames' windows, making the generated facial reaction videos lacking the temporal continuity inherent in human behaviour. While momentum operation reduces the facial reaction diversity (FRDiv), it prevents abrupt changes and jitters, ensuring a more realistic and stable video output.

Contributions of different losses: Table 7 then demonstrates the contributions of the loss functions employed in our training. Our observations are presented as follows: (i) adding a 3D speaker face reconstruction loss $\mathcal{L}_{rec}^{s}$ allows MSBEA to extract semantic meaningful features from speaker facial behaviours, which results in improved realism, appropriateness and synchrony of the generated facial reactions; (ii) Kullback-Leibler divergence loss $\mathcal{L}_{kl}$ and energy-based diversity loss $\mathcal{L}_{div}$ are critical in generating diverse facial reactions, while $\mathcal{L}_{div}$ additionally facilitates our ReactFace to learn different appropriate facial reactions in response to each speaker behaviour; and (iii) the smooth loss $\mathcal{L}_{smo}$ enhancing the appropriateness (in terms of the FRDist) of the generated facial reactions by smoothing variations between adjacent frames and avoiding jitter.

5.6 Visualization Analysis

To provide more insightful explanation of the proposed two speaker-listener interaction modules (*i.e.*, MIM and VIM), Fig. 9 visualises the attention weights within the transformer decoder layers of these modules. Specifically, we calculate the average attention weights across all attention heads, with each representing the correlation between a pair of frames. For the MIM, Fig. 9(a) and Fig. 9(c) correspond to the correlation between speaker speech frames and the generated facial reaction frames, while for VIM, Fig. 9(b) and Fig. 9(d) represent the correlation between speaker facial frames and the generated facial reaction frames. We perform this visualisation at both the inference stage (Fig. 9(a) and Fig. 9(b)) and the training stage (Fig. 9(c) and Fig. 9(d)), with
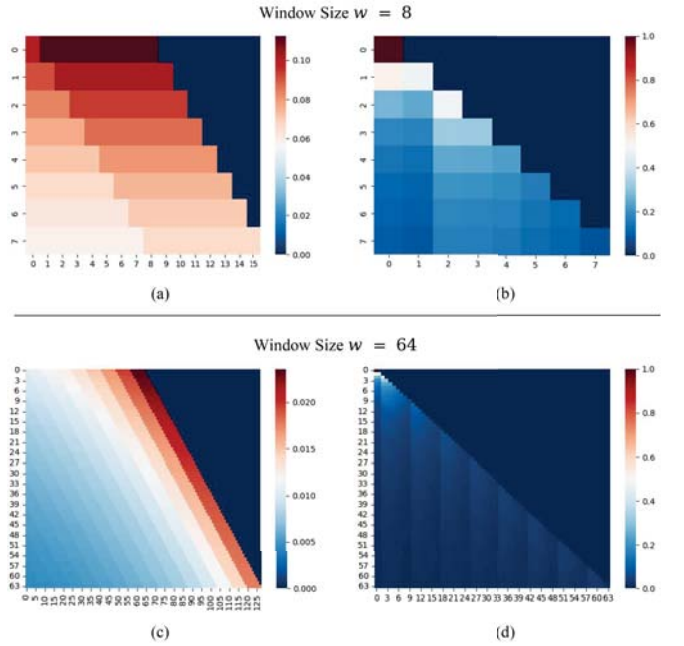

Fig. 9. Visualizations of attention weights in MIM and VIM. Attention weights of the (a) modality interaction module and (b) visual interaction module of the decoder, when setting window size $w = 8$ at the inference stage. Attention weights of the (c) modality interaction module and (d) visual interaction module of the decoder, when setting window size $w = 64$ at the training stage.

window sizes set to 8 and 64, respectively. As illustrated in Fig. 9, several key observations can be found: (i) later frames have zero influences on earlier facial reaction frames' generation, preventing the generation of existing facial reaction frames to be influenced by future speaker behaviours; (ii) the attention mechanism establishes both long-range and short-range context dependencies between the speaker speech/visual feature frames and facial reactions during the reasoning, emphasising the importance of contextual information; and (iii) Notably, a facial reaction frame tends to prioritize nearby speaker auditory/visual frames in its attention mechanism, while attenuating its focus on more distant speaker behaviour frames.

TABLE 7
Ablation study results achieved for applying different loss functions at the training stage.

$\mathcal{L}_{rec}^a$	$\mathcal{L}_{rec}^s$	$\mathcal{L}_{kl}$	$\mathcal{L}_{div}$	$\mathcal{L}_{smo}$	FRVar	FRDiv	FRRea	FRCorr	FRSyn
	✓	✓	✓	✓	0.0220	0.0433	412.14	0.166	45.96
✓		✓	✓	✓	0.0054	0.0831	596.22	0.143	46.26
✓	✓		✓	✓	0.0156	0.0238	479.77	0.186	45.36
✓	✓	✓		✓	0.0001	0.0000	590.24	0.112	43.39
✓	✓	✓	✓		0.0171	0.0248	420.61	0.139	44.16
✓	✓	✓	✓	✓	0.0159	0.0395	424.46	0.197	43.94

TABLE 8
User preference results between the facial reactions generated by our ReactFace and competitors.

Ours vs. Competitor	Realism	Diversity	Appropriateness	Sync
Ours vs. Ng et al. [1]	90.24%	95.12%	87.80%	92.68%
Ours vs. Zhou et al. [2]	80.49%	90.24%	82.93%	87.80%
Ours vs. GT	19.51%	9.76%	17.07%	14.63%

TABLE 9
Comparison of quantitative results on RLD dataset.

Methods	FRRea	FRVar	FRDiv	FRSyn
GT	168.24	1.8439	-	29.61
Trans-AE	250.09	0.0145	0	32.52
Ng et al. [1]	460.48	**1.1032**	0	**31.00**
Zhou et al. [2]	**180.56**	0.9314	0	32.62
ReactFace w/o $\mathcal{L}_{rec}^a$	271.09	0.3539	**0.3015**	31.12

Fig. 10. Failure cases of generated listener facial reaction by ReactFace.

5.7 Perception Survey

We conducted user studies on the Tencent Questionnaire platform to evaluate the facial reactions generated by our ReactFace in comparison to these generated three competitors: Ng et al [1], Zhou et al. [2] and the GT real facial reactions. Specifically, a total of 41 volunteers from Shenzhen University are recruited, comprising 12 females and 29 males, aged between 22 and 45. These volunteers, primarily from the college of computer science and software engineering and possessing relevant machine learning or deep learning expertise, were instructed with an online survey. This survey aimed to determine their preferences (*i.e.,* choosing a video containing more appropriate, realistic and synchronised facial reaction from a pair of compared videos) between the facial reaction videos generated by ReactFace and competitors. Specifically, In total, each volunteer watched 8 video clips (24 video pairs), with each video clip containing three pairs of videos, *i.e.,* comparing our ReactFace with Ng et al [1], Zhou et al. [2] and the GT real facial reactions, respectively. The platform displayed randomized video pairs for each comparison, and volunteers were instructed to evaluate the quality of the generated facial reactions in terms of realism, diversity, appropriateness, and synchronization. Results displayed in Tab. 8 indicate that the facial reactions generated by our ReactFace were generally favored by more than 80% human participants over the competitor in all cases. Interestingly, ReactFace was even preferred over ground truth samples in 17.07% and 19.51% of the evaluated cases in terms of appropriateness and realism, respectively.

5.8 Additional Results on RLD Dataset

In this section, we extend our experimental analysis to further include results on the RLD dataset [2]. This dataset comprises data from 92 subjects, consisting of 67 speakers and 76 listeners, with a total of 483 video clips sourced from YouTube. Notably, the RLD dataset lacks 'appropriate facial reaction' labels. Consequently, we present results obtained using our ReactFace model, trained without the utilization of the appropriate facial reaction reconstruction loss $\mathcal{L}_{rec}^a$. In this particular configuration, we replace $\mathcal{L}_{rec}^a$ with a standard reconstruction loss $\mathcal{L}_{rec}$, which solely assesses the dissimilarity between each generated facial reaction and its corresponding Ground Truth (GT) facial reaction. It's important to acknowledge that the limited scale of the RLD dataset, comprising 483 video clips, presents challenges for training models based on Transformer architecture. This is because Transformer-based models typically require more data for optimization as compared to LSTMs [91], and CNNs [92]–[94]. Consequently, both our ReactFace and the Trans-AE models struggle in this scenario. However, despite these challenges, as demonstrated in Table 9, our ReactFace model manages to achieve competitive performance in terms of realism, diversity, and synchrony when compared to state-of-the-art approaches.

5.9 Failure Cases

In this section, we discuss several failure cases observed in our generated samples. Fig. 10 presents the failure cases of our method on REACT2023 dataset. The qualitative results indicate that these failures primarily fall into two categories: (1) the rendering network for facial reaction visualisation may introduce image distortions, affecting both the face and background, as illustrated in the first row of Fig. 10; (2) ReactFace struggles to capture the dynamics of eye blinking, resulting in a slower blink speed (over one second) compared to natural human blinks, which typically occur within 100–150 milliseconds.

6 CONCLUSION

This paper proposes the first online MAFRG model (called ReactFace) capable of generating multiple appropriate facial reactions in response to each speaker behaviour. Our ReactFace starts with processing the input speaker behaviour

using a multi-modal speaker behaviour encoding and alignment (MSBEA) module, and then produces multiple facial reactions based on an appropriate facial reaction generation (AFRG) module, where a speaker-listener behaviour synchronisation (SLBS) module is introduced to model interactions between speaker behaviours and the corresponding generated facial reactions (*i.e.,* synchronising speaker behaviours and the corresponding generated facial reactions).

We systematically assess the efficacy of our ReactFace through a rigorous evaluation conducted on the combination of two dyadic interaction datasets provided by an open appropriate facial reaction generation challenge [22], leveraging various metrics encompassing diversity, realism, appropriateness, and synchrony that also have been employed by the challenge [22]. The obtained results demonstrate the effectiveness of our multiple facial reaction generation strategy. This strategy enables our model to produce a spectrum of diverse and contextually appropriate facial reactions. Furthermore, our introduced synchronization strategy significantly contributes to enhancing the realism of the generated reactions. Note that these reactions exhibit a high degree of continuity and synchrony with the corresponding audio-visual behaviours of the speaker. In-depth ablation analysis of the individual modules in our framework further demonstrates their effectiveness and indispensability. These modules, meticulously designed and integrated, play pivotal roles in augmenting the overall performance and cohesiveness of the proposed system.

Nevertheless, given the pioneering nature of our approach in the field of multiple appropriate facial reaction generation, certain modules employed may not yet represent the optimal configurations, leaving room for refinement in future investigations. Furthermore, there is a pressing need for the model to undergo further refinement to ensure its efficiency and practicality when applied in real-world scenarios. Finally, due to the limited resources, we were unable to reproduce all existing generative methods for comparison. Thus, our future work will focus on not only exploring more effective and efficient algorithms for online MAFRG task but also benchmarking existing generative methods (*e.g.,* GANs and diffusion models) for both offline and online MAFRG tasks.

ACKNOWLEDGMENTS

REFERENCES

[1] E. Ng, H. Joo, L. Hu, H. Li, T. Darrell, A. Kanazawa, and S. Ginosar, "Learning to listen: Modeling non-deterministic dyadic facial motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 395–20 405.

[2] M. Zhou, Y. Bai, W. Zhang, T. Yao, T. Zhao, and T. Mei, "Responsive listening head generation: a benchmark dataset and baseline," in *European Conference on Computer Vision*, 2022, pp. 124–142.

[3] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[4] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen, "Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features," in *Proceedings of the International Conference on Multimodal Interfaces*, 2005, pp. 175–182.

[5] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2022.

[6] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Personality recognition by modelling person-specific cognitive processes using graph representation," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 357–366.

[7] X. Zhai, M. Wang, and U. Ghani, "The sor (stimulus-organism-response) paradigm in online learning: an empirical study of students' knowledge hiding perceptions," *Interactive Learning Environments*, vol. 28, no. 5, pp. 586–601, 2020.

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

[9] S. Pandita, H. G. Mishra, and S. Chib, "Psychological impact of covid-19 crises on students through the lens of stimulus-organism-response (sor) model," *Children and Youth Services Review*, vol. 120, p. 105783, 2021.

[10] S. Song, M. Spitale, Y. Luo, B. Bal, and H. Gunes, "Multiple appropriate facial reaction generation in dyadic interaction settings: What, why and how?" *arXiv preprint arXiv:2302.06514*, 2023.

[11] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 770–18 780.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] D. Pavllo, D. Grangier, and M. Auli, "Quaternet: A quaternion-based recurrent model for human motion," in *British Machine Vision Conference*, 2018.

[14] Y. Huang and S. M. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–18.

[15] Y. Huang and S. Khan, "Generating photorealistic facial expressions in dyadic interactions," in *British Machine Vision Conference*, 2018.

[16] Y. Huang and S. M. Khan, "A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2018, pp. 437–445.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[18] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

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

[20] P. Jonell, T. Kucherenko, G. E. Henter, and J. Beskow, "Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings," in *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, 2020, pp. 1–8.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

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

[23] S. Song, M. Spitale, C. Luo, C. Palmero, G. Barquero, H. Zhu, S. Escalera, M. Valstar, T. Baur, F. Ringeval *et al.*, "React 2024: the second multiple appropriate facial reaction generation challenge," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2024, pp. 1–5.

[24] T. Xu, M. Spitale, H. Tang, L. Liu, H. Gunes, and S. Song, "Reversible graph neural network-based reaction distribution learning for multiple appropriate facial reactions generation," *arXiv preprint arXiv:2305.15270*, 2023.

[25] M.-D. Nguyen, H.-J. Yang, N.-H. Ho, S.-H. Kim, S. Kim, and J.-E. Shin, "Vector quantized diffusion models for multiple appropriate reactions generation," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2024, pp. 1–5.

[26] Q. T. Dam, T. T. N. Nguyen, D. T. Tran, and J.-H. Lee, "Finite scalar quantization as facial tokenizer for dyadic reaction generation," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2024, pp. 1–5.

[27] H. Zhu, X. Kong, W. Xie, X. Huang, L. Shen, L. Liu, H. Gunes, and S. Song, "Perfrdiff: Personalised weight editing for multiple appropriate facial reaction generation," in *Proceedings of the ACM International Conference on Multimedia*, 2024.

[28] D.-K. Nguyen, P. Paudel, S.-W. Kim, J.-E. Shin, S.-H. Kim, and H.-J. Yang, "Multiple facial reaction generation using gaussian mixture of models and multimodal bottleneck transformer," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2024, pp. 1–5.

[29] Z. Liu, C. Liang, J. Wang, H. Zhang, Y. Liu, C. Zhang, J. Gui, and S. Wang, "One-to-many appropriate reaction mapping modeling with discrete latent variable," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2024, pp. 1–5.

[30] C. Liang, J. Wang, H. Zhang, B. Tang, J. Huang, S. Wang, and X. Chen, "Unifarn: Unified transformer for facial reaction generation," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 9506–9510.

[31] J. Yu, J. Zhao, G. Xie, F. Chen, Y. Yu, L. Peng, M. Li, and Z. Dai, "Leveraging the latent diffusion models for offline facial multiple appropriate reactions generation," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 9561–9565.

[32] X. Hoque, A. Mann, G. Sharma, and A. Dhall, "Beamer: Behavioral encoder to generate multiple appropriate facial reactions," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 9536–9540.

[33] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, "Real-time facial animation with image-based dynamic avatars," *ACM Transactions on Graphics*, vol. 35, no. 4, 2016.

[34] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–14, 2019.

[35] B. H. Le, X. Ma, and Z. Deng, "Live speech driven head-and-eye motion generators," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 11, pp. 1902–1914, 2012.

[36] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–14, 2018.

[37] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "Lip-sync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2755–2764.

[38] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives." *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 42–1, 2013.

[39] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4176–4186.

[40] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic audio-driven video portraits," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3457–3466, 2020.

[41] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny, "Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3626–3636.

[42] H. Qiu, Y. Jiang, H. Zhou, W. Wu, and Z. Liu, "Stylefacev: Face video generation via decomposing and recomposing pretrained stylegan3," *arXiv preprint arXiv:2208.07862*, 2022.

[43] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European Conference on Computer Vision*, 2020, pp. 700–717.

[44] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3387–3396.

[45] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 080–14 089.

[46] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[47] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 985–10 995.

[48] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "Teach: Temporal action compositions for 3d humans," in *International Conference on 3D Vision*, September 2022.

[49] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura, "A recurrent variational autoencoder for human motion synthesis," in *British Machine Vision Conference*, 2017.

[50] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1418–1427.

[51] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5614–5623.

[52] Y. Yuan and K. Kitani, "Dlow: Diversifying latent flows for diverse human motion prediction," in *European Conference on Computer Vision*, 2020, pp. 346–364.

[53] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[54] X. Lin and M. R. Amer, "Human motion modeling using dvgans," *arXiv preprint arXiv:1804.10652*, 2018.

[55] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3332–3341.

[56] M. Petrovich, M. J. Black, and G. Varol, "Temos: Generating diverse human motions from textual descriptions," in *European Conference on Computer Vision*, 2022, pp. 480–497.

[57] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4394–4402.

[58] Y. Zhang, M. J. Black, and S. Tang, "Perpetual motion: Generating unbounded human motion," *arXiv preprint arXiv:2007.13886*, 2020.

[59] R. Zhao, H. Su, and Q. Ji, "Bayesian adversarial human motion synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6225–6234.

[60] A. Aristidou, A. Yiannakidis, K. Aberman, D. Cohen-Or, A. Shamir, and Y. Chrysanthou, "Rhythm is a dancer: Music-driven motion synthesis with global structure," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[61] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, "Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 2027–2036.

[62] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3497–3506.

[63] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2021–2029.

[64] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5152–5161.

[65] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7144–7153.

[66] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[67] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188–8197.

[68] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[69] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[70] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[71] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.

[72] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[73] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[74] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," *arXiv preprint arXiv:2402.12226*, 2024.

[75] O. Press, N. A. Smith, and M. Lewis, "Train short, test long: Attention with linear biases enables input length extrapolation," *arXiv preprint arXiv:2108.12409*, 2021.

[76] S. Card, T. MORAN, and A. Newell, "The model human processor-an engineering model of human performance," *Handbook of Perception and Human Performance.*, vol. 2, no. 45–1, 1986.

[77] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[78] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 759–13 768.

[79] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu, "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*, June 2022.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[81] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[82] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "Teach: Temporal action composition for 3d humans," in *International Conference on 3D Vision 2022*, 2022.

[83] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[84] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2017, pp. 350–359.

[85] C. Palmero, J. Selva, S. Smeureanu, J. Junior, C. Jacques, A. Clapés, A. Moseguí, Z. Zhang, D. Gallardo, G. Guilera *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1–12.

[86] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[87] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[88] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[89] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 157–164.

[90] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[91] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[93] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[94] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.