Smooth Online Multiple Appropriate Facial Reaction Generation

Weicheng Xie
School of Computer Science & Software Engineering, Shenzhen University
Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
Shenzhen, China
wcxie@szu.edu.cn

Chunlin Yan
Shenzhen University
Shenzhen, China
2310273020@email.szu.edu.cn

Siyang Song*
School of Computer Science, University of Exeter
Exeter, United Kingdom
s.song@exeter.ac.uk

Zitong YU
Department of Computing and Information Technology, Great Bay University
Dongguan, China
zitong.yu@ieee.org

Linlin Shen
Computer Vision Institute, Shenzhen University
School of Artificial Intelligence, Shenzhen University
Shenzhen, China
llshen@szu.edu.cn

Laizhong Cui
School of Computer Science & Software Engineering, Shenzhen University
Shenzhen, China
cuilz@szu.edu.cn

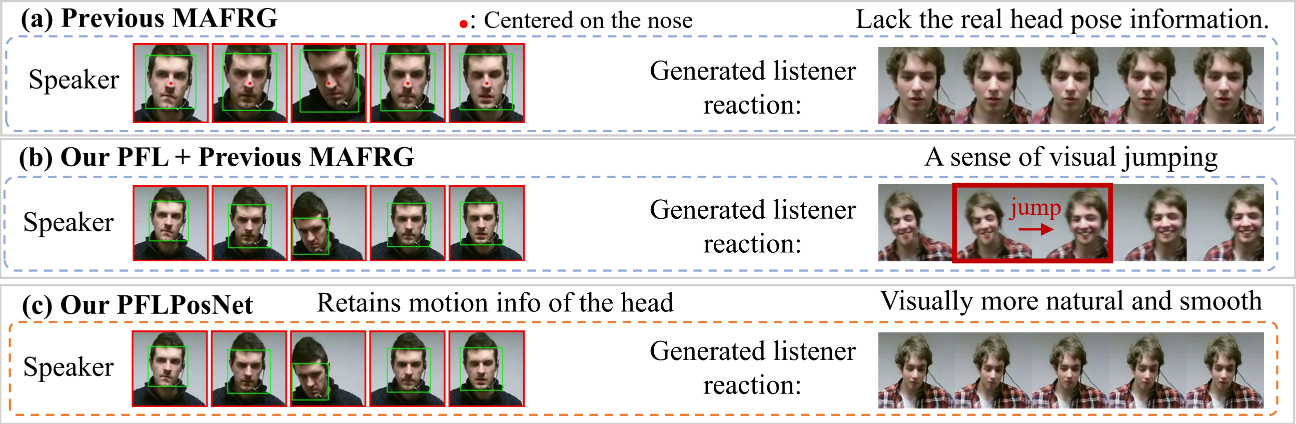

(a) Previous MAFRG
●: Centered on the nose
Lack the real head pose information.
Speaker
Generated listener reaction:
(b) Our PFL + Previous MAFRG
A sense of visual jumping
Speaker
Generated listener reaction:
jump
(c) Our PFLPosNet
Retains motion info of the head
Visually more natural and smooth
Speaker
Generated listener reaction:

Figure 1: Comparison of our PFLPosNet with previous MAFRG tasks: (a) Previous MAFRG tasks utilized a floating window cropping method (centered on the nose) to obtain data for training, which lacked real head displacement information, leading to output results also missing such information. (b) Our Pose-aware Face Behavior Localization (PFL) acquires data retaining real head displacement information, but the generated results exhibit visual jitter. (c) Our PFLPosNet adjusts head displacement information in real-time in the generated results, making the visualization outcomes smoother.

Abstract

In dyadic interactions, facial reactions are crucial for conveying an individuals' responses to their conversational partners. Individuals may exhibit varied but appropriate facial reactions (AFRs) when perceiving the same behavioral expression. Although some recent methods can already respond multiple appropriate facial reactions to the given human speaker behaviors, the AFRs generated by these methods often fail to adequately preserve crucial head motions, leading to visual jitter and unnatural transitions between generated AFR segments. In this paper, we propose a novel and generic PFLPosNet framework which addresses the aforementioned problems at both pre-processing and post-processing stages, where a new pose-aware face behavior localization method PFL is introduced to retain the head pose displacement information from the source data. In addition, the framework proposes a real-time head pose adjustment method, PosNet, to ensure continuity and smoothness in the visual output of the model when using data with correct head pose displacement. Experimental results demonstrate that our approach not only generates more coherent and natural facial reaction sequences but also significantly outperforms existing online MAFRG methods in terms of continuity and smoothness. Our code is made available at https://github.com/rainforcetime/PFLPosNet.

CCS Concepts

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

[*] Corresponding author.

MM '25, Dublin, Ireland.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/XXXXXX.XXXXXX

Keywords

Facial Reaction Generation, generation sequence smoothness, Head posture, real time reaction

ACM Reference Format:
Weicheng Xie, Chunlin Yan, Siyang Song, Zitong YU, Linlin Shen, and Laizhong Cui. 2025. Smooth Online Multiple Appropriate Facial Reaction Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 Introduction

Human convey essential intentions and emotional states through their facial behaviors during interpersonal communication, where various facial reactions can be exhibited in response to a specific speaker behavior due to individual differences (e.g., personality [37]) and varied contextual factors [26]. Therefore, the capability of automatically generating appropriate human-style facial reactions is crucial for developing virtual agents and assistants for social interactions [20, 22], educational guidance [5, 34], and human-computer interaction applications [6, 43].

Early automatic facial reaction generation (FAG) methods [13–15, 28, 31, 35, 37, 50] focused on reproducing the specific real facial reactions (called 'GT' facial reaction) triggered by the given speaker's behavior, where conditional generative adversarial networks (GANs) [14, 15, 27, 31], CNNs explored by Neural Architecture Search [35, 37] and Variational Auto-encoder (VAE) [28] have been frequently extended to generate a 'GT' facial reaction from the given speech and visual behaviors (i.e., called speaker behavior). While these approaches treat FAG as a 'one-to-one mapping' task by reproducing GT real facial reactions, in real-world dyadic interactions, the same speaker behavior can elicit different real facial reactions expressed by different human listeners [26], making facial reaction generation being naturally a 'one-to-many mapping' task [40]. Consequently, the Multiple Appropriate Facial Reaction Generation (MAFRG) task that aims to produce a variety of different but appropriate facial reactions (AFRs) for responding to each external stimuli (e.g., human speaker behavior) have been formally defined [40], with a certain number of MAFRG solutions [12, 23–25, 29, 30, 47, 51] developed in the past two years.

In both the early FAG and recent MAFRG methods, the implementation process can be divided into two primary stages: facial reaction representation generation and facial video synthesis. Specifically, the first stage typically involves predicting facial reactions in the form of facial landmarks [35, 37], facial sketches [13, 14], 3DMM (3D Morphable Model) coefficients [25, 28, 50] or facial action units (AUs) [46], and then additionally GAN [13, 14, 28, 50] or 3DMM-based semantic neural rendering for faces [25] are trained/applied in the second stage to synthesize face frames accordingly. However, facial landmarks fall short in accurately representing complex facial texture variations, especially when dealing with occlusions, varying poses, and lighting conditions. They primarily focus on key points of the face, which may not suffice for applications requiring detailed facial analysis or reconstruction. Although 3DMM coefficients excel in these areas by offering a comprehensive representations describing facial displays, enabling better handling of the aforementioned challenges, the use of them in previous MAFRG frameworks includes facial expression, head rotation and translation coefficients, while these frameworks mainly use cropped face region from the source image during training and inference, where head rotation and translation information are typically ignored/distorted. As a result, previous MAFRG studies have been conducted under the absence of true head pose information (**Problem 1**).

More importantly, although there are already well-defined quantitative indicators for evaluating various aspects of facial reactions generated by MAFRG models [40], visualizing them remains a highly intuitive quality reflection. This is determined by not only the facial reaction generation network but also the separate face image rendering tool (e.g., GAN [4, 27] or 3DMM-based semantic neural rendering [32]) mentioned above. In previous online MAFRG solutions [23, 25, 38, 39], facial reactions are considered as dynamic processes containing rich temporal information, while human behavior responses are always experiencing a short delay caused by the related cognitive process [28, 35, 37], thus window-based strategies have been frequently employed to continuously generate facial reactions, i.e., they generate a small set of frames representing facial reactions over a short period rather than frame-by-frame generation. Although the generation of the current facial reaction segment is conditioned on previously generated facial reactions in these approaches, data discontinuity still frequently occurred at window boundaries due to facial displacement caused by the lack of specific optimization for displacement variations at window boundaries during training, which are more visually sensitive than expression changes, leading to visual jitter and unnatural transitions in facial reaction generation outcomes (**Problem 2**).

In this paper, we propose simple yet effective AND generic online MAFRG pre-processing and post-processing strategies to enhance the visual continuity and smoothness of the generated AFRs in dyadic human-computer conversational scenarios. The proposed Pose-aware Facial Behavior Localization pre-processing strategy specifically retains head motions of face videos while removing unnecessary backgrounds, allowing the MAFRG models not only additionally considers speaker head motions compared to existing online MAFRG approaches but also enforces the MAFRG models to generate facial reactions with head motions rather than facial behaviours only (**addressing Problem 1**). Furthermore, to ensure the visual quality of the generated AFRs, the pose-aware post-processing dynamically adjusts the head poses from the generated AFRs, ensuring the temporal coherence and natural head transitions between adjacent AFR segments (**addressing Problem 2**). Our main contributions and novelties are summarized as follows:

- We propose a novel generic Pose-aware Face Behavior Localization (PFL) which defines a fixed-position pose-aware facial region across all frames within a video, with the region location and size are adaptively decided by the head poses of the video.
- We propose a novel and lightweight MLP-based PosNet which refines head poses of the generated AFRs using spatio-temporal context provided by previously adjusted poses, ensuring the continuity and smoothness of the pose variations in online AFR generation.

- Experimental results demonstrate that our proposed PFLPosNet is capable of generating reasonable and smoothly continuous listener responses with head poses. By employing PosNet to refine the outputs of existing online MAFRG methods, the final results appear more continuous and smooth to human visual perception.

2 Related Work

Facial reaction generation: Early deterministic FAG methods [13–15, 31] aim to reproduce the ground truth (GT) facial reactions of interlocutors based on speaker behaviors. These methods typically consist of two main stages: (i) predicting a representation describing the listener's facial reaction via GAN [13–15, 31], or Convolutional Neural Networks (CNNs) [35, 37] in response to the speaker behaviour, which is required to exactly match the GT facial reaction; and (ii) visualizing these facial reaction representations as face images/videos using GAN [13–15, 31], and 3DMM-based semantic neural rendering for faces [32]. Since human FAG is a non-deterministic process, various recent methods have been proposed to generate multiple diverse yet appropriate facial reactions from the input speaker behavior. Some methods learn an explicit reaction distribution (e.g., Gaussian distribution) from the input speaker's behavior and sample multiple AFRs from this distribution [23, 46]. Luo et al. [25] employed this strategy, incorporating a behavioral synchronization module to achieve alignment between speakers and listeners, with other works [12, 38, 46] also utilizing this approach. Other methods [30, 47, 51] leverage a Diffusion Model to learn a latent distribution, obtaining multiple AFRs during the denoising process. However, these approaches still overlook the preservation of head displacement dynamics and temporal continuity in segment-based online generation, leading to visual jitter between adjacent windows.

Face Animation: The key for existing MAFRG approaches to visualize their generated facial reactions is through facial animation, a task that has garnered significant attention in recent years [3, 7, 18, 21, 41, 45, 48, 49]. Among them, studies on facial representation and controllable generation [19, 32, 44, 49] have been continuously explored, where some methods [32, 44] specifically use facial expressions and head pose representations as conditional signals to enable manual control of the generated facial sequences. The existing face animation techniques are mainly realized based on methods such as Generative Adversarial Networks (GAN) [4, 17], 3D Morphable Models (3DMM) [10, 32, 42], Landmark-driven [9], Neural Rendering [8, 11], Motion Transfer [36], etc.

3 Task and problem definition

Task definition: In dyadic interaction, the goal of the Online Multiple Appropriate Facial Reaction Generation (Online MAFRG) task is to develop a machine learning model $\mathcal{H}^{\text{MAFRG}}$ that considers only the γ-th frame (i.e., $\gamma \in [t_1, t_2]$) along with behaviours $B^s_{[t_1,\gamma-1]}$ previously expressed by the corresponding speaker to generate each appropriate facial reaction (AFR) frame $\hat{R}^l_\gamma(i)$ as:

$$\hat{R}^l_\gamma(i) = \mathcal{H}^{\text{MAFRG}}\left(B^s_{[t_1,\gamma-1]}\right) \tag{1}$$

where $\hat{R}^l_\gamma(i)$ denotes the γ-th frame of the predicted i-th facial reaction $\left(\hat{R}^l(i), i \in (1, M)\right)$. Ultimately, the generated M AFRs can be represented as $\mathbb{R}^l = \{\hat{R}^l(1), \hat{R}^l(2), \cdots, \hat{R}^l(M)\}$. For more details, please refer to [40].

Head motion leakage problem: Existing online MAFRG approach typically generate AFR segments continuously, where the head pose transition at the boundaries of these segments are usually unnatural and discontinuous, which can be attributed to two problems:

(1) **Data:** Previous online MAFRG studies employ a floating window strategy (i.e., face detection) to obtain face regions, where the centers of the detected faces are used to define the centers of the corresponding cropped face images. However, such methods neglect the spatial displacement of the head within the frame sequence. Assuming that a rigid head motion is described by a translation vector $T(t) \in \mathbb{R}^3$ and a rotation matrix $Q(t) \in \text{SO}(3)$, where SO(3) denotes the special orthogonal group representing all 3D rotations without reflection, previous floating window strategies track a reference point (e.g., nose tip $\mathbf{X}_{\text{ref}}$) by dynamically adjusting the window center $\mathbf{f}_{\text{float}}(t)$ as:

$$\mathbf{X}_{\text{float}}(t) = \mathbf{Q}(t)\mathbf{X}_{\text{ref}} + \mathbf{T}(t) \tag{2}$$

where $\mathbf{X}_{\text{float}}(t)$ represents the position of the floating window center $\mathbf{f}_{\text{float}}(t)$ in the 3D global coordinate system at the time t, while $\mathbf{X}_{\text{ref}}$ is the position of the reference point (e.g., the tip of the nose) used to determine the center of the floating window. The cropped 3D local coordinates $\mathbf{X}'_{\text{float}}$ of any head keypoint X_i can be represented as:

$$\mathbf{X}'_{\text{float}} = \mathbf{X}_i^{\text{global}} - \mathbf{X}_{\text{float}}(t) = \underbrace{\mathbf{Q}(t)(\mathbf{X}_i - \mathbf{X}_{\text{ref}})}_{\text{Keep only rotation}} \tag{3}$$

where $\mathbf{X}_i^{\text{global}}$ denotes the position vector of keypoint X_i in the 3D global coordinate system. This way, the head displacement information $T(t)$ is entirely discarded, preventing the cropped result from reflecting the overall head movement in physical space (e.g., lateral displacement, moving closer to or away from the camera). Tasks relying on displacement information (such as conversational interaction analysis) suffer from performance degradation due to data distortion. For instance, head motion between time instants t_1 and t_2 can be represented as:

$$\Delta\mathbf{X}'_{\text{float}} = \mathbf{Q}(t_2)(\mathbf{X}_i - \mathbf{X}_{\text{ref}}) - \mathbf{Q}(t_1)(\mathbf{X}_i - \mathbf{X}_{\text{ref}}) \tag{4}$$

At this point, the resulting head motion $\Delta\mathbf{X}'_{\text{float}}$ only includes rotational changes ΔQ, with the actual displacement $\Delta T = T(t_2) - T(t_1)$ completely lost, leading to distorted head displacement information in the outcome. More details on the data analysis can be found in the supplementary material.

(2) **Omission due to data loss:** Due to the lack of real head movement information, previous MAFRG studies have not considered the constraints on the abruptness of head movements between generated segments. When the magnitude of movement at the junction of adjacent segments in the generated sequence exceeds the visual jump perception threshold, it results in significant motion discontinuity.

4 Methodology

Overview: In this section, we propose our generic PFLPosNet to improve the visual continuity and smoothness of facial reaction generation results in dynamic conversational environments (As shown in Fig. 2). It consists of a novel **Pose-aware Face Behavior Localization (PFL) module** which is a generic pre-processing strategy that crops the face region from every frame while retaining undistorted crucial head motions (including both head rotation and location) across frames (**addressing Problem 1**), as well as a **PosNet** which is a generic post-processing strategy that mitigates unreasonable facial behaviour jittering and jumping to achieve visual discontinuities for the generated facial reactions (**addressing Problem 2**). Specifically, the PFL crops fixed-location image regions containing complete head displacement information from all frames of the given original video V. It first detects the face location for every frame $I_n \in V$ of the original video $V = \{I_1, I_2, \cdots, I_N\}$ (N represents the number of video frames), and subsequently obtains all initial frame-level face regions $F = \{f_1, f_2, \cdots, f_N\}$. Then, the PFL statistically analyzes the face locations F of all video frames in the current single video to adaptively determine a fixed spatial region f_{fix} in all frames of V, where faces appear most frequently and completely. This ensures to maximally capture head displacements and facial behaviour cues provided in the V as:

$$f_{\text{fix}} = \text{PFL}(V) \tag{5}$$

Based on the obtained fixed spatial region f_{fix}, a cropped face video V_f is obtained, which contains not only facial behaviours but also undistorted head motions across frames. This step allows the employed MAFRG generator to: (1) directly access both facial behaviours and head motions expressed by speaker at the inference stage; and (2) learn to generate facial reactions with natural head motions at the training phase. As detailed in sec. 4.1. Then, we introduce a generic PosNet to process the head poses of all frames in the facial reactions generated by any MAFRG Generator MAFRGen (i.e., the MAFRGen can be any existing developed online MAFRG model). It adjusts the current facial reaction segment $\hat{R}^l_{[t-w+1:t]}$ based on the last frame $\hat{R}^l_{t-w}$ of the previously generated facial reaction segment $\hat{R}^l_{[t-2w+1:t-w]}$ as:

$$\begin{aligned}\hat{R}^l_{[t-w+1:t]} &= \mathcal{H}^{\text{MAFRG}}(\hat{B}^s_{1:t-w}) \\ &= \text{PosNet}\left(\hat{R}^l_{t-w},\ \text{MAFRGen}(\hat{B}^s_{1:t-w})\right)\end{aligned} \tag{6}$$

where $\hat{B}^s_{1:t-w}$ denotes the speaker facial behaviors represented by the fixed spatial window (defined by PFL), while w accounting for the time delay caused by executing human cognitive process [28, 37]. This step ensures the smoothness and naturalness when transitioning from one facial reaction segment to the next, while retaining all characteristics and information of the originally generated facial reaction.

4.1 Pose-aware Facial Behavior Localization

The Pose-aware Facial Behavior Localization(PFL) module conducts a fixed-position frame region localization strategy to retain complete head motions (i.e., head rotation and translation) in the given video V for subsequent processing Unlike traditional floating window (i.e., Cropping only the detected facial region with the tip of the nose as the center, where the cropping box size varies according to the face size.) strategies used in previous MAFRG studies (compared in sec. 3), our PFL module not only addresses head displacement data leakage to facilitate the MAFRG model to capture essential facial behaviours, but also removes irrelevant background cues to reduce the extra computational burden. To obtain an image region of fixed size and location in all frames of V, which maximally retain head motions, our PFL starts with the face detection for all frames, where the initially detected 2D face region f_i in the i-th frame is represented as $f_i = (x_i, y_i, w_i, h_i)$, with x_i, y_i being the coordinates of the 2D face region center while w_i, h_i denoting the width and height of the detected face region f_i. Then, the mean center of all frame-level face regions in V can be computed as:

$$x_V = \frac{1}{N}\sum_{i=1}^{N} x_i, y_V = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{7}$$

While $F = \{f_1, f_2, \cdots, f_N\}$ contain all frame-level facial behaviours, the PFL further learns w_c and h_c to adaptively define a fixed location spatial window $f_{\text{fix}} = (x_V, y_V, w_c, h_c)$ to crop all frames of V, aiming to additionally retain head motions across all frames, which are computed as:

$$\begin{aligned} w_c &= \min\left(\omega_w, \left[\max_{i=1}^{N}\left(x_i + \frac{w_i}{2}\right) - \min_{i=1}^{N}\left(x_i - \frac{w_i}{2}\right)\right]\right), \\ h_c &= \min\left(\omega_h, \left[\max_{i=1}^{N}\left(y_i + \frac{h_i}{2}\right) - \min_{i=1}^{N}\left(y_i - \frac{h_i}{2}\right)\right]\right). \end{aligned} \tag{8}$$

where ω_w and ω_h define the maximum allowable width and height of the fixed image region of the given video V, respectively.

To rigorously validate the head motions retained by our PFL, we model the relationship between 2D cropping and 3D head kinematics below, similar to the one described in sec. 3. Let $\mathbf{X}_i$ denote the 3D coordinates of a head keypoint in the global reference system. The fixed cropping window f_{fix} defines a local coordinate system centered at x_V, y_V, with spatial extents w_c, h_c. By projecting $\mathbf{X}_i$ into this local system, we decouple rigid head motions as:

$$\mathbf{X}'_{\text{pose}} = \underbrace{\mathbf{Q}(t)(\mathbf{X}_i - \mathbf{X}_{\text{ref}})}_{\text{rotation}} + \underbrace{\mathbf{T}(t) - \mathbf{T}_{\text{ref}}}_{\text{translation}} \tag{9}$$

Here, $\mathbf{X}_{\text{ref}} = (x_V, y_V, 0)$ and $\mathbf{T}_{\text{ref}}$ represent the invariant origin of the cropped window in 3D space. Unlike floating windows where $\mathbf{X}_{\text{ref}}$ drifts with facial landmarks, our fixed f_{fix} ensures: (a) Rotation Isolation: $\mathbf{Q}(t)$ purely encodes head orientation changes, as $\mathbf{X}_{\text{ref}}$ remains static.(b) Translation Faithfulness: $\mathbf{T}(t) - \mathbf{T}_{\text{ref}}$ directly measures global displacement, unobscured by cropping shifts. For example, head motion between time instants t_1 and t_2 can be expressed as:

$$\Delta\mathbf{X}'_{\text{pose}} = \underbrace{\mathbf{T}(t_2) - \mathbf{T}(t_1)}_{\Delta\mathbf{T}} + \underbrace{\mathbf{Q}(t_2)\mathbf{X}_i - \mathbf{Q}(t_1)\mathbf{X}_i}_{\Delta\mathbf{Q}} \tag{10}$$

Here, the translation $\Delta\mathbf{T}$ is independent of rotational change $\Delta\mathbf{Q}$, directly reflecting the head displacement in physical space.

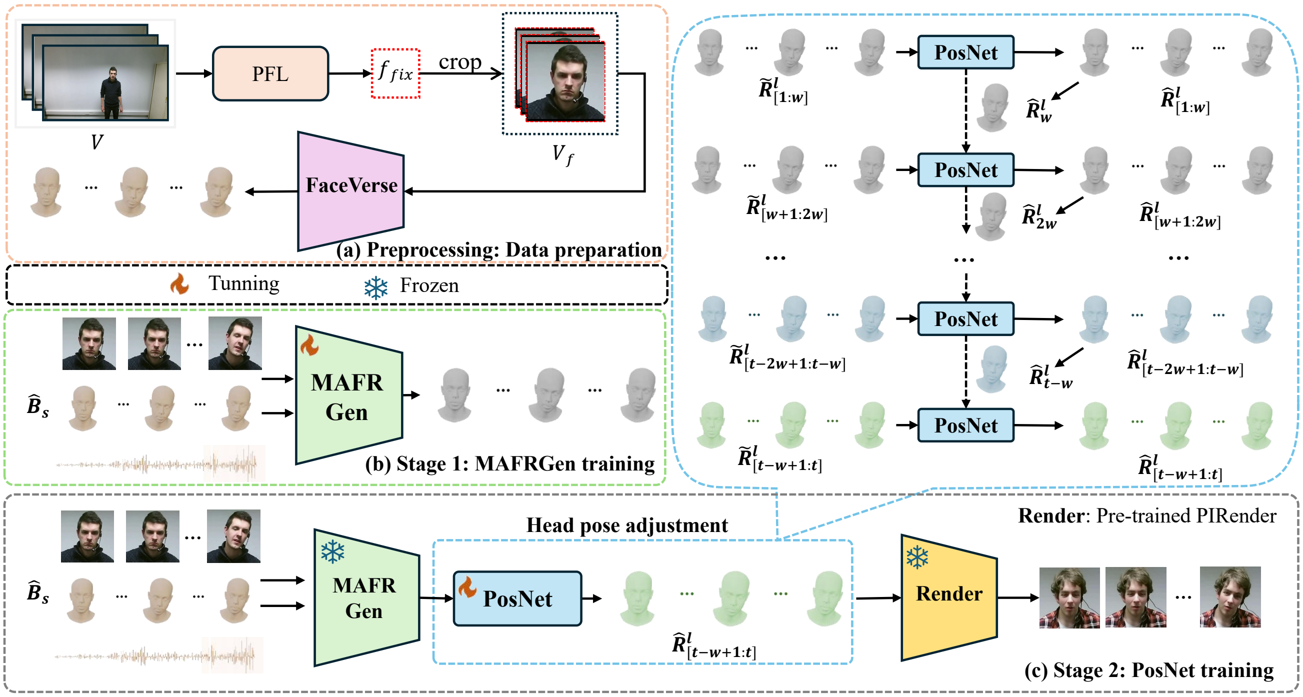


Figure 2: The pipeline of the proposed PFLPosNet. (a) The data preprocessing procedure of the PFL module. (b) Phase One: B^s represents the speaker's past behavioral data processed by the PFL module, which is input into MAFRGen for training. (c) The corresponding initial facial response sequence is obtained through the pre-trained MAFRGen (MAFRGen can be any existing online MAFRG model that has been developed). At time t, the PosNet module takes the initial facial response $\tilde{R}^l_{[t-w+1:t]}$ and the last frame $\hat{R}^l_{t-w}$ of the previously adjusted segment $\hat{R}^l_{[t-2w+1:t-w]}$ as inputs to obtain the adjusted facial response sequence $\hat{R}^l_{[t-w+1:t]}$.

4.2 Pose-aware Appropriate Facial Reaction generation

This section presents our PosNet which can post-process the head poses of facial frames $\tilde{R}^l_{[1:kw]}$ generated by any MAFRG generator (represented by MAFRGen), ensuring the coherence of head motions between adjacent facial reaction segments, while keeping facial expressions undistorted. This facilitates the data continuity of head motions at the window boundaries. This module is motivated by: **(a) Task Separation:** we employ an independent PosNet module for post-smoothing of head poses in the generated facial reactions instead of integrating temporal smoothing within MAFRGen, avoiding the drop of diversity and temporal continuity caused by the introduction of temporal smoothing; and **(b) Plug-and-Play:** The pose optimization functionality is isolated into a independent lightweight PosNet, enabling it to be generic to any pre-trained MAFRGen architecture without redesigning the smoothing mechanism for different generators.

As demonstrated in Eq. 6, at the time t, we refine the generated current facial reaction segment $\tilde{R}^l_{[t-w+1:t]}$ using PosNet based on the last frame $\hat{R}^l_{t-w}$ of the previously adjusted facial reaction segment $\hat{R}^l_{[t-2w+1:t-w]}$ as:

$$\hat{R}^l_{[t-w+1:t]} = \text{PosNet}(\hat{R}^l_{t-w}\,,\ \tilde{R}^l_{[t-w+1:t]}) \tag{11}$$

In this process, our PosNet applies FaceVerse [44] to decompose each facial reaction frame $\hat{R}^l_t$ into facial expression $\tilde{E}^l_t$ and pose $\tilde{P}^l_t$ as:

$$\left\{\tilde{E}^l_{[t-w+1:t]}\,,\ \hat{P}^l_{[t-w+1:t]}\right\} = \text{FaceVerse}(\hat{R}^l_{[t-w+1:t]}) \tag{12}$$

Then, our PosNet only adjusts the decomposed head poses, while keeping other non-head-pose attributes (e.g., facial expressions) unchanged as:

$$\hat{R}^l_{[t-w+1:t]} = \left\{\tilde{E}^l_{[t-w+1:t]}\,,\ \text{MLP}\left(\hat{P}^l_{t-w} \oplus \text{Enc}(\tilde{P}^l_{[t-w+1:t]})\right)\right\} \tag{13}$$

where MLP is a three-layer multilayer perceptron (MLP) aiming to achieve precise correction of head poses through nonlinear mapping of input features; Enc$(\cdot)$ represents the feature encoder that specifically extracts latent features related to head poses $\tilde{P}^l_{[t-w+1:t]}$ from the current facial reaction segment $\tilde{R}^l_{[t-w+1:t]}$; $\oplus$ denotes the vector concatenation, which combines the head pose $\hat{P}^l_{t-w}$ corresponding to the last frame $\hat{R}^l_{t-w}$ of the previous facial reaction segment with the latent head poses feature Enc$(\tilde{P}^l_{[t-w+1:t]})$, thereby providing comprehensive spatio-temporal contexts for MLP. Consequently, the MLP generates a natural and smooth sequence of head poses $\hat{P}^l_{[t-w+1:t]}$ based on detailed historical and current information $(\hat{P}^l_{t-w} \ \oplus\ \text{Enc}(\tilde{P}^l_{[t-w+1:t]}))$, which is used as a foundational condition for ensuring perceptual continuity of head motions across

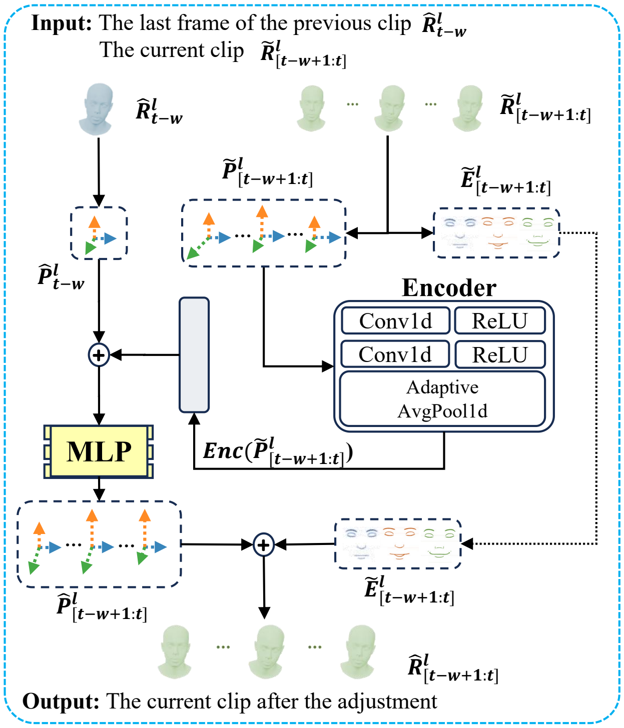

Input: The last frame of the previous clip $\widehat{R}^l_{t-w}$
The current clip $\widetilde{R}^l_{[t-w+1:t]}$
$\widehat{R}^l_{t-w}$
$\widetilde{R}^l_{[t-w+1:t]}$
$\widetilde{P}^l_{[t-w+1:t]}$
$\widetilde{E}^l_{[t-w+1:t]}$
$\widehat{P}^l_{t-w}$
Encoder
Conv1d
ReLU
Conv1d
ReLU
Adaptive AvgPool1d
MLP
$Enc(\widetilde{P}^l_{[t-w+1:t]})$
$\widehat{P}^l_{[t-w+1:t]}$
$\widetilde{E}^l_{[t-w+1:t]}$
$\widehat{R}^l_{[t-w+1:t]}$
Output: The current clip after the adjustment

Figure 3: Architecture of the PosNet Module for Real-Time Head Pose Adjustment.

adjacent facial segments, thereby eliminating visually abrupt transitions at window boundaries. This process is also depicted in Fig. 3.

4.3 Training Strategy and Loss Functions

Pre-training MAFRGen: The training of our PosNet is divided into two stages. The first stage involves pre-training a general appropriate facial reaction generator MAFRGen, and the second stage focuses on training the pose adjustment module PosNet. The parameters of MAFRGen are not updated in Stage 2.

Typically, the loss function of MAFRGen includes the MSE loss between the true facial reaction $F_{[t-w+1:t]}$ and the predicted facial reaction AFR $\hat{F}_{[t-w+1:t]}$:

$$\mathcal{L}_{\text{general}} = \mathbb{E}\left[\|F_{[t-w+1:t]} - \hat{F}_{[t-w+1:t]}\|^2\right] \tag{14}$$

where $\|\cdot\|$ denotes the L2 norm, and $\mathbb{E}[\cdot]$ represents the expected value over the distribution of input data. Moreover, different MAFRGen methods have their customized loss functions $\mathcal{L}_{\text{custom}}$, such as $\mathcal{L}_{\text{kl}}$ and $\mathcal{L}_{\text{smo}}$ in [25], among other functional losses.

Training PosNet for MAFRGen: To achieve smoother generation results, we introduce various loss functions for the constraint. Similarly, taking $\tilde{R}^l_{[t-2w+1:t-w]}$ and $\tilde{R}^l_{[t-w+1:t]}$ as examples, our PosNet adjusts the current head pose $\tilde{P}^l_{[t-w+1:t]}$ using the head pose $\tilde{P}^l_{t-w}$ of the last frame $\tilde{R}^l_{t-w}$ of the previous segment. Firstly, we should ensure the smoothness at the connection:

$$\mathcal{L}_{\text{connect}} = \left|\tilde{P}^l_{t-w+1} - \tilde{P}^l_{t-w}\right| \tag{15}$$

where $|\cdot|$ denotes the absolute value.

Secondly, the head poses within the current segment $\tilde{P}^l_{[t-w+1:t]}$ should be smooth, meaning the change between adjacent time steps should be small:

$$\mathcal{L}_{\text{sm1}} = \frac{1}{w-1}\sum_{i=1}^{w-1}\left|\tilde{P}^l_{t-w+i+1} - \tilde{P}^l_{t-w+i}\right| \tag{16}$$

Finally, it is necessary to control the overall change magnitude of the current segment's head pose to be similar:

$$\mathcal{L}_{\text{sm2}} = \frac{1}{w-2}\sum_{i=1}^{w-2}\left[\left||\tilde{P}^l_{t-w+i+2} - \tilde{P}^l_{t-w+i+1}| - |\tilde{P}^l_{t-w+i+1} - \tilde{P}^l_{t-w+i}|\right|\right] \tag{17}$$

The total loss is as follows:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{general}} + \mathcal{L}_{\text{custom}} + \lambda_1\mathcal{L}_{\text{connect}} \\ &+ \lambda_2\mathcal{L}_{\text{sm1}} + \lambda_3\mathcal{L}_{\text{sm2}}\end{aligned} \tag{18}$$

The pseudocode of our algorithm is provided in the supplementary material.

5 Experiments

5.1 Experimental Settings

Dataset: In our work, PFLPosNet utilizes two datasets: NoXI [2], and RECOLA [33].The original data contains 2962 pairs of human speaker-listener dyadic interactions recorded under various contexts (including 1593 pairs for training, 562 pairs for validation, and 806 pairs for testing), each lasting 30 seconds. After processing with our new cropping method and removing outliers, our dataset comprises 2534 pairs (including 1318 pairs for training, 519 pairs for validation, and 697 pairs for testing). We allocated the data according to the data indices provided in the REACT2024 challenge [39].
Implementation Details: We first crop the face region from each frame of the given segment and resize it to 256×256 ($\omega_w = 256, \omega_h = 256$). Then, we pretrain the Multiple Appropriate Facial Reaction Generator (MAFRGen) using an Adam optimizer with an initial learning rate of 0.0001. Additionally, we train our PosNet using an SGD optimizer with lr = 0.001, and the loss weights are set as follows: $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.2$. All experiments were conducted on Nvidia V100 GPUs using PyTorch. Further details regarding implementation and datasets can be found in the supplementary material. Moreover, we pretrain the image rendering model PIRender using fixed-window cropping.
Evaluation Metrics: As our primary objective is to enhance video smoothness as perceived by the human eye, and given that previous studies on facial reaction generation have not targeted this specific goal, we adopt the VBench [16] evaluation benchmark. This benchmark provides metrics such as 'subject consistency', 'background consistency', 'motion smoothness', 'dynamic degree', and 'imaging quality' for our customized videos. Please refer to the supplementary material for more details. In the ablation study section, we used the metrics from previous challenges [38, 39] to measure the impact on aspects such as appropriateness, diversity, and realism.

5.2 Comparison with existing approaches

Table. 1 presents the quantitative comparison results of various existing MAFRG methods before and after PosNet adjustments in the online MAFRG task. When "Dynamic degree" is high and

Table 1: Quantitative comparisons of various MAFRG methods on the dataset processed by the PFL module, before and after the application of the PosNet module. The best results for each metric on each method are marked in bold.

Method	Subject consistency (↑)	Background consistency (↑)	Motion smoothness (↑)	Imaging quality (↑)	Temporal flickering (↑)	Dynamic degree (↓)
Trans-VAE[38]	97.3114%	96.5110%	99.0939%	50.7164%	98.6968%	95.3372%
Trans-VAE + PosNet	**97.9097%**	**97.3707%**	**99.5345%**	**51.7964%**	**99.4224%**	**0.0000%**
Belfusion[1]	97.3758%	97.0828%	99.3669%	51.4821%	99.2623%	0.0079%
Belfusion + PosNet	**98.1602%**	**97.6105%**	**99.5031%**	**52.5414%**	**99.4910%**	**0.0000%**
Reactface[25]	98.8673%	98.1330%	99.5976%	52.4293%	99.6001%	0.0000%
Reactface + PosNet	**98.9415%**	**98.3419%**	**99.6094%**	**52.8171%**	**99.6781%**	0.0000%
PerFRDiff[51]	96.2711%	95.9321%	98.4633%	50.6099%	98.1133%	86.5136%
PerFRDiff + PosNet	**97.4501%**	**97.0771%**	**99.5008%**	**52.5685%**	**99.3552%**	**6.0976%**

Table 2: Ablation study results on the Reactface method.

Float Crop	PFL Crop	PosNet	Subject consistency (↑)	Background consistency (↑)	Motion smoothness (↑)	Imaging quality (↑)	Temporal flickering (↑)	Dynamic degree (↓)
✓			**99.0058%**	97.6767%	**99.6162%**	40.7967%	99.5376%	0.0000%
	✓		98.8673%	98.1330%	99.5976%	52.4293%	99.6001%	0.0000%
	✓	✓	98.9415%	**98.3419%**	99.6094%	**52.8171%**	**99.6781%**	0.0000%

Table 3: Ablation study results on the PerFRDiff method.

Float Crop	PFL Crop	PosNet	Subject consistency (↑)	Background consistency (↑)	Motion smoothness (↑)	Imaging quality (↑)	Temporal flickering (↑)	Dynamic degree (↓)
✓			95.5255%	95.6970%	99.2403%	41.3867%	98.4462%	91.5352%
	✓		96.2711%	95.9321%	98.4633%	50.6099%	98.1133%	86.5136%
	✓	✓	**97.4501%**	**97.0771%**	**99.5008%**	**52.5685%**	**99.3552%**	**6.0976%**

Table 4: The Impact of PosNet Adjustments on MAFRG Outcome Metrics.

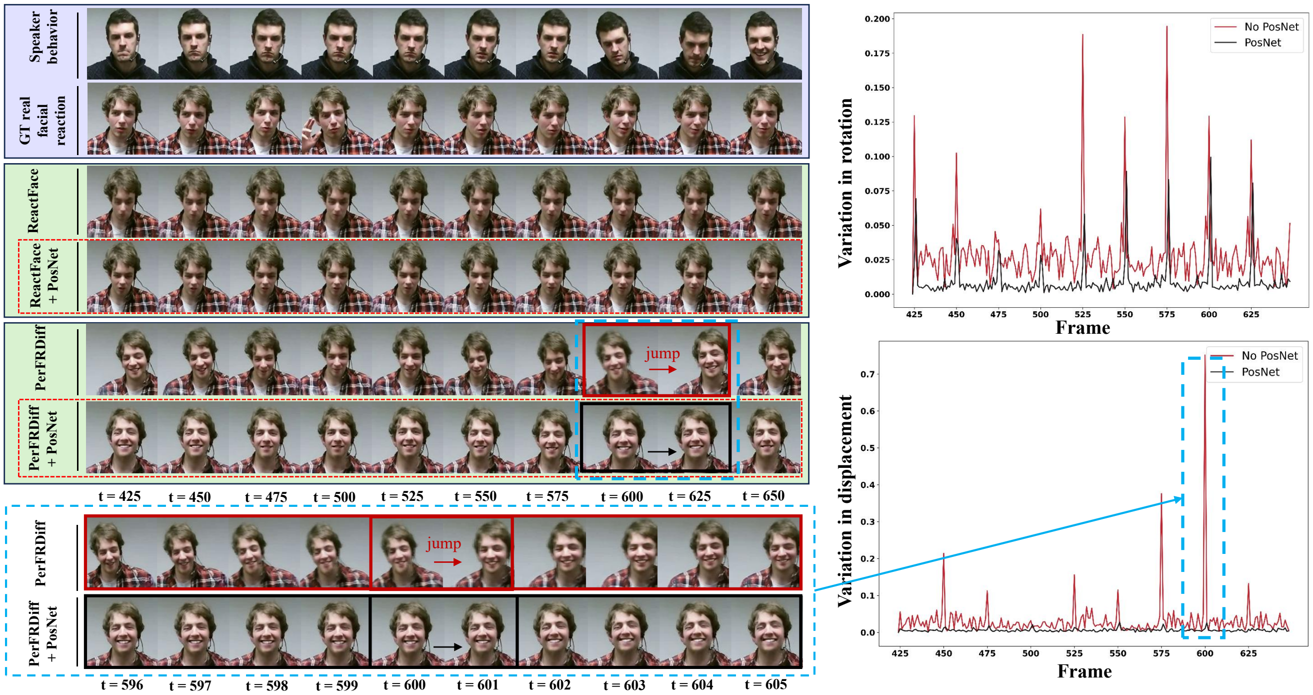
Method	FRDvs (↑)	FRVar (↑)	FRDiv (↑)	FRCorr (↑)	FRSyn (↓)
ReactFace[25]	**0.0397**	**0.0064**	0.0341	**0.1927**	46.29
ReactFace + PosNet	0.0368	0.0063	0.0341	0.1921	**45.74**
PerFRDiff[51]	**0.1261**	**0.0612**	**0.0486**	0.2318	45.45
PerFRDiff + PosNet	0.1251	0.0607	0.0481	**0.2320**	**45.39**

"Motion smoothness" is low, the generated results visually exhibit abnormal jittering and jumping phenomena. Experimental data show that methods adjusted by PosNet demonstrate significant advantages at the level of human visual perception. Specifically, the motion smoothness metrics for Trans-VAE, Belfusion, ReactFace, and PerFRDiff are improved by 0.44% (99.0939%→99.5345%), 0.14%(99.3669%→99.5031%), 0.01%(99.5976%→99.6094%), and 1.04% (98.4633%→99.5008%), respectively, while their dynamic degree metrics significantly decreased, with PerFRDiff's dynamic degree sharply dropping from 86.5136% to 6.0976% (a reduction of 80.4%). It is noteworthy that due to inherent generation mechanism limitations, the facial reactions produced by ReactFace and Belfusion methods have minimal motion magnitude visually, with their original dynamic degrees nearly zero. Since this method does not modify the core structure of the general appropriate facial reaction generator (MAFRGen), the adjusted visual effects retain the original expression characteristics. However, by dynamically adjusting head pose data in real-time, the final visual output notably enhances action coherence while maintaining content consistency. As shown in Fig. 4, particularly in the 5th and 6th frames of the 7th and 8th rows (images depicted in Fig. 4), the head displacement coefficient curves after PosNet adjustment exhibit smoother transitions, verifying the effectiveness of this method in eliminating visual discontinuities at window boundaries. More details can be found in the supplementary material.

5.3 Ablation Study

In this section, we conducted ablation studies on data and methods for the ReactFace method and the PerFRDiff method to evaluate the effectiveness of our PFLPosNet. This includes: (i) the effectiveness of the fixed cropping scheme; (ii) the effectiveness of PosNet in enhancing visual effects; (iii) the impact of PosNet adjustments on the generation results of MAFRGen. Additional ablation studies on the loss function and cropping window size are provided in the supplementary material.

Effectiveness of the PFL Cropping Scheme: To evaluate the performance difference between the PFL cropping scheme and the floating window cropping scheme, this study conducted systematic ablation experiments. As shown in Table. 2, under the ReactFace framework, although the PFL scheme exhibited minor gaps in "Subject consistency" (98.87% vs 99.01%) and "Motion smoothness" (99.60% vs 99.62%), it improved background consistency by 0.46% (97.68%→98.13%) and imaging quality by 11.63% (40.80%→52.43%) through the fixed window mechanism described in 4.1. The floating



(a) Visualisation of generated facial reactions

(b) Variation in head rotation and displacement

Figure 4: (a) Visualization of facial responses generated by various methods, with the visualization results from our PosNet showing more natural head movements and visual effects in response to the speaker's behavior. (b) Changes in head position before and after adjustment by our PosNet.

cropping mechanism, due to its reliance on local facial tracking, resulted in the loss of head displacement information, leading to motion trajectory fragmentation in generated sequences (Sec. 3). Cross-model validation further indicated (Table. 3) that within the PerFRDiff framework, the PFL cropping reduced the dynamism index by 5% (91.54%→86.51%), confirming the universal advantages of this scheme across various application scenarios.

Effectiveness of PosNet in Enhancing Visual Effects: By comparing the changes in metrics before and after the activation of PosNet in Table. 2 and Table. 3, this study confirms the network's critical role in enhancing visual quality. In the ReactFace framework, PosNet improves motion smoothness by 0.0118% (99.5976%→99.6094%), achieving a state-of-the-art performance of 99.6781% in temporal flicker metrics; in the PerFRDiff framework, it leads to an 80.4% sharp decrease in dynamism index (86.51%→6.10%) and enhances motion smoothness by 1.04% (98.46%→99.50%). This quantitative enhancement across models indicates that PosNet (Sec. 4.2), through fine-tuning the dynamic details of head pose (Fig. 3), not only optimizes technical parameters but also effectively improves motion continuity and visual immersion.

Impact of PosNet Adjustments on the Generation Results of MAFRGen: We evaluated the impact of our PosNet adjustments on the generation appropriateness using the evaluation metrics from previous challenges [38, 39], as shown in Table. 4. Experiments were conducted on both the ReactFace and PerFRDiff methods. By comparing the various metrics results of MAFRG, it was found that although most metrics showed a slight decline, the magnitude was extremely minor. Observing the overall metrics, the application of PosNet did not significantly affect the results of the original methods. MAFRG metrics indicated that the overall error margin between using PosNet and not using PosNet was merely between 0.001 and 0.005, demonstrating that our PosNet, when only adjusting head pose without altering the original expression coefficients, had a negligible impact on the appropriateness of the generated reaction outcomes.

6 Conclusion

This paper proposes a general PFLPosNet, which can improve the visual continuity and smoothness of facial reaction generation in dynamic dialogue environments. The results show that, compared with previous methods, our approach not only retains the authentic key head pose information from the original video but also provides a more comprehensive and realistic data foundation for subsequent facial reaction analysis and interactive applications. Our solution also mitigates unreasonable facial behavior jitter and jumps caused by human vision's sensitivity to visual discontinuities at window boundaries, achieving better visualization effects. A critical limitation is that when there are excessively large pose reactions, the training difficulty of the rendering tools used for visualization becomes extremely high, which will be addressed in future work.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under grants no. 62276170, 82261138629, 62306061, the Science and Technology Project of Guangdong Province under grants no. 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030, Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No. GML-KF-24-11, and Guangdong Provincial Key Laboratory under grant no. 2023B1212060076.

References

[1] German Barquero, Sergio Escalera, and Cristina Palmero. 2023. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2317–2327.

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

[3] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics* 35, 4 (2016).

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8789–8797.

[5] Harry Barton Essel, Dimitrios Vlachopoulos, Akosua Tachie-Menson, Esi Eduafua Johnson, and Papa Kwame Baah. 2022. The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *International Journal of Educational Technology in Higher Education* 19, 1 (2022), 57.

[6] Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*. 76–83.

[7] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[8] Lin Gao, Feng-Lin Liu, Shu-Yu Chen, Kaiwen Jiang, Chunpeng Li, Yukun Lai, and Hongbo Fu. 2023. SketchFaceNeRF: Sketch-based facial generation and editing in neural radiance fields. *ACM Transactions on Graphics* 42, 4 (2023).

[9] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. 2020. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 10861–10868.

[10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*. Springer, 152–168.

[11] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[12] Ximi Hoque, Adamay Mann, Gulshan Sharma, and Abhinav Dhall. 2023. BEAMER: Behavioral Encoder to Generate Multiple Appropriate Facial Reactions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9536–9540.

[13] Yuchi Huang and Saad Khan. 2018. A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 437–445.

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

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

[16] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.

[17] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[18] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM transactions on graphics (TOG)* 37, 4 (2018), 1–14.

[19] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. 2020. Config: Controllable neural face image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 299–315.

[20] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality* 2 (2022), 786665.

[21] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. 2021. Lipsync3d: 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*. 2755–2764.

[22] Sun Kyong Lee, Pavitra Kavya, and Sarah C Lasser. 2021. Social interactions and relationships with an intelligent virtual agent. *International Journal of Human-Computer Studies* 150 (2021), 102608.

[23] Cong Liang, Jiahe Wang, Haofan Zhang, Bing Tang, Junshan Huang, Shangfei Wang, and Xiaoping Chen. 2023. Unifarn: Unified transformer for facial reaction generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9506–9510.

[24] Zhenjie Liu, Cong Liang, Jiahe Wang, Haofan Zhang, Yadong Liu, Caichao Zhang, Jialin Gui, and Shangfei Wang. 2024. One-to-many appropriate reaction mapping modeling with discrete latent variable. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–5.

[25] Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Zongyuan Ge, Linlin Shen, and Hatice Gunes. 2024. ReactFace: Online Multiple Appropriate Facial Reaction Generation in Dyadic Interactions. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[26] Albert Mehrabian. 1974. An approach to environmental psychology. *Massachusetts Institute of Technology* (1974).

[27] Mehdi Mirza. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[28] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20395–20405.

[29] Dang-Khanh Nguyen, Prabesh Paudel, Seung-Won Kim, Ji-Eun Shin, Soo-Hyung Kim, and Hyung-Jeong Yang. 2024. Multiple facial reaction generation using gaussian mixture of models and multimodal bottleneck transformer. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–5.

[30] Minh-Duc Nguyen, Hyung-Jeong Yang, Ngoc-Huynh Ho, Soo-Hyung Kim, Seung-won Kim, and Ji-Eun Shin. 2024. Vector quantized diffusion models for multiple appropriate reactions generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–5.

[31] Behnaz Nojavanasghari, Yuchi Huang, and Saad Khan. 2018. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092* (2018).

[32] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13759–13768.

[33] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[34] Kyoungwon Seo, Joice Tang, Ido Roll, Sidney Fels, and Dongwook Yoon. 2021. The impact of artificial intelligence on learner–instructor interaction in online learning. *International journal of educational technology in higher education* 18 (2021), 1–23.

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

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[37] Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. 2022. Learning person-specific cognition from facial reactions for automatic personality recognition. *IEEE Transactions on Affective Computing* 14, 4 (2022), 3048–3065.

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

[39] Siyang Song, Micol Spitale, Cheng Luo, Cristina Palmero, German Barquero, Hengde Zhu, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, et al. 2024. React 2024: the second multiple appropriate facial reaction generation challenge. *arXiv preprint arXiv:2401.05166* (2024).

[40] Siyang Song, Micol Spitale, Yiming Luo, Batuhan Bal, and Hatice Gunes. 2023. Multiple Appropriate Facial Reaction Generation in Dyadic Interaction Settings: What, Why and How? *arXiv preprint arXiv:2302.06514* (2023).

[41] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. 2024. Diffused heads: Diffusion models beat gans on

talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5091–5100.

[42] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183–1.

[43] Isaac Wang and Jaime Ruiz. 2021. Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction* 37, 17 (2021), 1648–1673.

[44] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20333–20342.

[45] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. 2022. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2531–2539.

[46] Tong Xu, Micol Spitale, Hao Tang, Lu Liu, Hatice Gunes, and Siyang Song. 2023. Reversible graph neural network-based reaction distribution learning for multiple appropriate facial reactions generation. *arXiv preprint arXiv:2305.15270* (2023).

[47] Jun Yu, Ji Zhao, Guochen Xie, Fengxin Chen, Ye Yu, Liang Peng, Minglei Li, and Zonghong Dai. 2023. Leveraging the latent diffusion models for offline facial multiple appropriate reactions generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9561–9565.

[48] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. 2023. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[49] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4176–4186.

[50] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. 2022. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*. Springer, 124–142.

[51] Hengde Zhu, Xiangyu Kong, Weicheng Xie, Xin Huang, Linlin Shen, Lu Liu, Hatice Gunes, and Siyang Song. 2024. Perfrdiff: Personalised weight editing for multiple appropriate facial reaction generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9495–9504.