Supplementary Material

CCS Concepts

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

Keywords

Multiple Appropriate Facial Reaction Generation (MAFRG); Personalisation; Weight editing; Diffusion

ACM Reference Format:

. 2024. Supplementary Material. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3664647.3680752

1 Notations and Abbreviations

Table 1 lists the abbreviations of **definitions** used in the paper. Table 2 provides the abbreviations of **different components in PerFRDiff**. Table 3 and 4 list the notations of **variables and latent representations** used in the paper. Table 5 lists the notations of the used **loss functions**.

Table 1: Abbreviations for definitions

Abbreviations	Descriptions
PMAFRG	Personalised Multiple Appropriate Reaction Generation
MAFRG	Multiple Appropriate Reaction Generation
ML	Machine learning
GT	Ground truth
AFRs	Appropriate facial reactions
3DMM	3D Morphable Model
MFCC	Mel-frequency Cepstral Coefficients

Table 2: Abbreviations for PerFRDiff components

Abbreviations	Components
SBE	Speaker Behaviour Encoding
GAFRG	Generic Appropriate Facial Reaction Generator
GAFRG^{θ^l}	An personalised instance of GAFRG
GAFRG_k	The k_{th} layer in the GAFRG
GAFRG^l_k	The k_{th} layer in the GAFRG^{θ^l}
PCSM	Personalised Cognitive Style Modelling
PSSL	Personalised Style Space Learning
PWSG	Personalised Weight Shift Generation
Enc_{aud}	Speaker audio behaviour semantics encoder
Enc_{app}	Speaker facial appearance semantics encoder
Enc_{emo}	Speaker facial emotional semantics encoder
Enc_{p}	Transformer encoder

2 Implementation Details

Network details: In this study, we adopt: (1) the torchaudio package[1] to extract MFCC features from the raw speaker audio signal; (2) the FaceVerse [7] to extract 58 3DMM coefficients (i.e., 52 coefficients describing facial expression, 3 coefficients describing the transition and 3 coefficients rotation) from the speaker face video; and (3) the GraphAU [4, 5] to extract facial emotional representations, including occurrences of 15 action units (i.e., AU1, AU2, AU4, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU23, AU24, AU25, AU26), 8 facial expression probabilities (i.e., Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger and Contempt), valence and arousal intensities from every frame of the input speaker facial behaviour. The speaker audio semantic encoder $\textbf{Enc}_{\text{aud}}$ and the speaker appearance semantic encoder $\textbf{Enc}_{\text{app}}$ within **the SBE module** are implemented as a fully-connected layer (linear layer), while the speaker emotional semantics encoder $\textbf{Enc}_{\text{emo}}$ is an RNN-based VAE [1]. The **PCSM module** comprises a PSSL block and a PWSG block. The PSSL block adopts FaceVerse [7] as the identity-free attribute extractor to extract 3DMM coefficients from each frame of listeners' historical face video. The transformer encoder $\textbf{Enc}_{\text{p}}$ in PSSL consists of four transformer encoder layers, where the number of heads in the multi-head attention block of each encoder layer is set to 4. Finally, a fully-connected layer is attached at the top of the transformer encoder to output the personalised cognitive style representation. Meanwhile, the PWSG block is a multi-branch network with the number of branches depending on the number of layers in the GAFRG module (i.e., the number of layers that need to be edited). The PWSG block starts with two fully-connected layers followed by multiple branches. Each branch is implemented as a fully-connected layer to produce the weight shift matrices corresponding to the layers of the GAFRG module. The transformer decoder within the **GAFRG module** consists of seven transformer decoder layers, where the number of heads in the multi-head attention block of each decoder layer is also set to 4. The cross-attention operation in each decoder layer takes the concatenation of the speaker audio behavioural semantics, facial appearance semantics and facial emotional semantics encoded by the SBE module as the key and value, which is considered as a condition to guide the reverse (denoising) process for appropriate facial reaction generation. Unless specifically noted, the three types of speaker behavioural semantics are used by default. We employ

[1] https://pytorch.org/audio/stable/index.html

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680752

Table 3: Notations of variables - Part 1

Notations	Descriptions
B^s	A speaker behaviour
$B^s_{[1:t-w]}$	The speaker behaviour expressed at the time interval $[1:t-w]$
$A^s_{[1:t-w]}$	The speaker audio behaviour expressed at the time interval $[1:t-w]$
$F^s_{[1:t-w]}$	The speaker facial behaviour expressed at the time interval $[1:t-w]$
$\hat{R}^l(m)$	The m-th generated AFR for the l-th listener in response to the speaker behaviour B^s
$\hat{R}^l_{[t-w+1:t]}(m)$	The m-th generated AFR at the time interval $[t-w+1:t]$ for the l-th listener in response to the speaker behaviour $B^s_{[1:t-w]}$
$\mathbb{R}^l_{[t-w+1:t]}$	A set of predicted personalised AFRs $\{\hat{R}^l_{[t-w+1:t]}(1),\cdots,\hat{R}^l_{[t-w+1:t]}(m)$ at the time interval $[t-w+1:t]$ for the l-th listener in response to the speaker behaviour $B^s_{[1:t-w]}$
$\mathbb{R}^l$	A set of predicted personalised AFRs $\{\hat{R}^l(1),\cdots,\hat{R}^l(M)\}$ for the l-th listener in response to the speaker behaviour B^s
F^l_h	A historical facial behaviour of the l-th listener
$F^l(n)$	The n_{th} real AFR expressed by the l-th in response to the speaker behaviour B^s
$F_{([t-w+1:t],d)}$	A noisy version of an AFR segment expressed at the time interval $[t-w+1:t]$ in response to $B^s_{[1:t-w]}$ obtained at the d-th diffusion step
$F_{([t-w+1:t],0)}$	Original clean real AFR segment expressed at the time interval $[t-w+1:t]$ in response to $B^s_{[1:t-w]}$ at the 0-th diffusion step
$\hat{F}_{([t-w+1:t],0)}$	Predicted original clean AFR segment expressed at the time interval $[t-w+1:t]$ in response to $B^s_{[1:t-w]}$ in the reverse process
$\mathbb{F}^l$	A set of real AFRs $\{F^l(1),\cdots,F^l(N)\}$ expressed by the l-th listener in response to the speaker behaviour B^s
$\mathbb{F}$	A union set of real AFRs $\{\mathbb{F}^1\cup\mathbb{F}^2\cup\cdots\cup\mathbb{F}^L\}$ expressed by L listeners in response to the B^s
p^l	Personalised cognitive style of the l-th listener
$p^l(F^l(n))$	The personalised cognitive style inferred from the n-th facial reaction $F^l(n)$ expressed by the l-th listener

Table 4: Notations of variables - Part 2

Notations	Descriptions
$\mathcal{Z}$	A latent space
ψ^l_h	Identity-free facial behaviour attributes extracted from the historical facial behaviour of the l-th listener
E_p	The encoded embedding of ψ^l_h
$E^{\text{aud}}_{[1:t-w]}$	MFCC features of the the speaker audio behaviour $A^s_{[1:t-w]}$
$\bar{E}^{\text{aud}}_{[1:t-w]}$	Speaker audio behavioural semantics at the time interval $[1:t-w]$
$E^{\text{app}}_{[1:t-w]}$	3DMM coefficients extracted from the speaker facial behaviour $F^s_{[1:t-w]}$
$\bar{E}^{\text{app}}_{[1:t-w]}$	Speaker facial appearance semantics at the time interval $[1:t-w]$
$E^{\text{emo}}_{[1:t-w]}$	Facial emotional representations of the speaker facial behaviour $F^s_{[1:t-w]}$
$\bar{E}^{\text{emo}}_{[1:t-w]}$	Speaker facial emotional semantics at the time interval $[1:t-w]$
d	The d-th step in the forward diffusion process
ϵ_θ	Predicted noise added in the forward diffusion process
c	Conditions of the diffusion model
θ	The weights of GAFRG
$\Delta\theta^l$	Personalised weight shifts for the l-th listener
θ^l	Personalised weights for the l-th listener
W_k	The weight matrix of k-th layer GAFRG_k
ΔW^l_k	Personalised weight shift applied to the k-th layer GAFRG_k in the GAFRG for the l-th
z^l_k	The output of the target layer $\text{GAFRG}^{\theta^l}_k$
b_k	the biases at the target layer $\text{GAFRG}^{\theta^l}_k$

Table 5: Notations of loss functions

Loss	Descriptions
$\mathcal{L}_1$	MSE loss for training the GAFRG and PWSG
$\mathcal{L}_2$	Contrastive loss for training the PSSL

learnable positional embeddings to maintain the positional information within the transformer decoder. Then, a fully-connected mapping layer is finally employed after the transformer decoder to map the latent representations to facial reactions.

Training details: The first training stage individually trains the GAFRG and PSSL, where the maximum training epochs for both GAFRG and PSSL are set to 500 and 50, respectively. During GAFRG training, the batch size is set to 4 (speaker behaviors). To generate generic appropriate facial reactions, we randomly select n real AFRs that can be expressed by various listeners in response to each speaker behavior, where n is empirically set to 10. If an input speaker behavior is associated with fewer than 10 available real AFRs in the training set, we randomly replicate the provided real AFRs until the total reaches 10. We set the batch size to 3 (speaker behaviours) to train PSSL which models personalised cognitive style. Differently, for each speaker behaviours, all of its corresponding real AFRs are used in the training. The parameters of GAFRG and PSSL are optimised by two separate AdamW optimiser [3] with

an initial learning rate of 10^{-4} and weight decay of 10^{-4} using a cosine-annealing schedule. The temperature parameter τ in the contrastive loss [2] for training PSSL is set to 0.07. In the experiments, unless specifically noted, classifier-free guidance which enhances the balance between the quality and diversity of the generated facial reactions is employed in GAFRG's training by default. The second training stage trains the PWSG block with the pre-trained GAFRG. In particular, the PSSL block is kept frozen at this stage, while the weights of the GAFRG are only updated (edited) by the PWSG block in the forward propagation process. The SGD optimiser with an initial learning rate of 10^{-3}, weight decay of 10^{-4} and momentum of 0.9 is employed to optimise the weights of the PWSG block. The batch size is set to 1 (speaker audio-visual behaviour). The maximum number of training epochs is set to 100. The default length of the listener's historical facial behaviour used for personalised cognitive style modelling is 30 seconds. All experiments are conducted on Nvidia A100 GPUs using PyTorch.

3 Influence of different speaker behavioural modalities

Table 6 evaluates the importance of different speaker behavioural semantics contributing to multiple diverse and appropriate facial reaction generation. It is worth noting that the results are obtained by the GAFRG without weight editing. The best performance over the appropriateness of the generated facial reactions in terms of correlations with the real AFRs is achieved when all of speaker audio, appearance and facial emotional behavioural semantics are considered, as indicated by the highest FRCorr (0.35 and 0.31) on the MAFRG and PMAFRG tasks. Merely considering the speaker audio semantics leads to a larger DTW distance (FRdist) between the generated facial reactions and real AFRs, indicated by the largest FRdist values on both MAFRG and PMAFRG tasks. In contrast, speaker facial behavioural semantics (i.e., speaker facial appearance semantics and facial emotional semantics) are more reliable for predicting facial reactions that have a lower distance to real AFRs. The results reveal that considering multi-modalities of the speaker behaviour helps to generate more appropriate facial reactions in response to each speaker behaviour. In addition, multi-modalities help to improve the diversity among multiple generated facial reactions in response to the same speaker behaviour, as indicated by the highest FRDiv.

4 Results achieved by directly generating personalised weights for the GAFRG and generating personalised weight shifts for editing a pre-trained GAFRG

Table 7 compares the performance achieved by directly generating personalised weights (weight generation) and personalised weight shifts (weight editing) for producing a personalised instance of GAFRG module. It can be observed that the personalised instance of GAFRG obtained by weight editing outperforms the other way over appropriateness (FRCorr and FRdist), diversity (FRDiv, FRDvs and FRVar) and synchrony (FRSyn). The reason is that directly generating personalised weights focuses more on the personalised aspects of facial reaction generation. In contrast, editing a pre-trained GAFRG allows to take into account both of personalised cognitive processes of the target listener and the commonly shared cognitive processes among different listeners. The superior performance achieved by weight editing indicates the importance of pre-training a GAFRG in facilitating personalised facial reaction generation.

5 Results achieved by predicting the added noise and predicting the original clean facial reactions in the reverse (denoising) process

Table 8 compares the performance between predicting the added noise and predicting original clean facial reaction segments by GAFRG$^{\theta^l}$ in the reverse (denoising) process to generate facial reactions. As reported, predicting the original clean facial reaction segments enables the GAFRG$^{\theta^l}$ to generate more appropriate facial reactions in response to each speaker behaviour on both MAFRG and PMAFRG tasks (higher correlations and lower distances to the real AFRs). In contrast, predicting noise to be removed allows the GAFRG$^{\theta^l}$ to generate more diverse and synchronised facial reactions in response to speaker behaviours.

6 Influence of the classier-free guidance in model training

Table 9 evaluates the importance of the employment of classifier-free guidance in the training of GAFRG and GAFRG$^{\theta^l}$ in multiple diverse and appropriate facial reaction generation. The comparison in Table 9 shows that adding classifier-free guidance generally helps to generate facial reactions with significantly higher correlations to real AFRs, but it slightly worsens distances and the synchrony (measured by FRSyn), with a mixed impact on different diversity metrics (FRDiv, FRDvs and FRVar). For example, the diversity among multiple facial reactions generated in response to the same speaker behaviour (measured by FRDiv) is decreased, while the diversity among facial reactions generated for different speaker behaviours (measured by FRDvs) is improved. We

7 Statistical Difference Analysis

We conducted a two-tailed test with 95% confidence to compare our proposed PerFRDiff and three state-of-the-art MAFRG approaches on both MAFRG and PMAFRG tasks. The comparative results are reported in Table 10. It can be observed that PerFRDiff significantly outperforms the compared approaches in generating facial reactions that are highly correlated to the real AFRs on both MAFRG and PMAFRG tasks. In addition, the facial reactions generated by the PerFRDiff are significantly more diverse and realistic than those generated by the compared approaches. In contrast, the facial reactions generated by REGNN and Trans-VAE are significantly more synchronised with the speaker behaviours compared to facial reactions generated by PerFRDiff.

8 Model Complexity Analysis

The intricacies of PerFRDiff are clearly revealed through multiple metrics that capture its performance and the computational effort it requires, as shown in Table. 11. For inference purposes, it handles

Table 6: Influence of different speaker behavioural modalities.

Behavioural semantics			FRCorr ↑		FRdist ↓		FRDiv ↑	FRDvs ↑	FRVar ↑	FRRea ↓	FRSyn ↓
$\mathbf{Enc}_{\text{app}}$	$\mathbf{Enc}_{\text{emo}}$	$\mathbf{Enc}_{\text{aud}}$	MAFRG	PMAFRG	MAFRG	PMAFRG		($\times 10^{-2}$)			
✓			0.32	0.28	109.99	112.18	15.39	25.63	10.79	47.41	45.34
	✓		0.21	0.16	103.33	**104.51**	3.39	6.19	2.54	56.51	45.32
		✓	0.32	0.28	171.47	174.71	0.01	**39.24**	**11.10**	48.68	**45.05**
	✓	✓	0.33	0.29	158.31	160.65	0.01	32.39	11.03	46.44	45.40
✓		✓	0.34	0.30	105.34	108.10	15.75	26.78	10.33	48.92	45.80
✓	✓		0.32	0.27	**101.13**	109.87	16.31	26.06	10.21	49.32	45.52
✓	✓	✓	**0.35**	**0.31**	105.57	107.46	**17.14**	23.46	9.38	**45.94**	45.41

Table 7: Results achieved by directly generating personalised weights for the GAFRG module and generating personalised weight shifts for editing a pre-trained GAFRG module in personalised facial reaction generation.

Method	FRCorr ↑		FRdist ↓		FRDiv ↑	FRDvs ↑	FRVar ↑	FRRea ↓	FRSyn ↓
	MAFRG	PMAFRG	MAFRG	PMAFRG		($\times 10^{-2}$)			
Weight generation	0.36	0.32	97.28	100.48	8.44	13.95	5.44	**46.89**	45.31
Weight editing	**0.38**	**0.36**	**94.72**	**98.43**	**13.68**	**21.91**	**8.79**	47.62	**45.28**

Table 8: Results achieved by predicting the added noise and predicting the original clean facial reactions in the reverse (denoising) process.

Prediction	FRCorr ↑		FRdist ↓		FRDiv ↑	FRDvs ↑	FRVar ↑	FRRea ↓	FRSyn ↓
	MAFRG	PMAFRG	MAFRG	PMAFRG		($\times 10^{-2}$)			
Added noise	0.36	0.31	110.24	112.84	**18.63**	**27.12**	**10.84**	47.83	**45.06**
Facial reaction	**0.38**	**0.36**	**94.72**	**98.43**	13.68	21.91	8.79	**47.62**	45.28

Table 9: Results achieved by the PerFRDiff with/without classifier-free guidance.

Setting	FRCorr ↑		FRdist ↓		FRDiv ↑	FRDvs ↑	FRVar ↑	FRRea ↓	FRSyn ↓
	MAFRG	PMAFRG	MAFRG	PMAFRG		($\times 10^{-2}$)			
w/o Classifier-free guidance	0.31	0.27	**93.88**	**96.27**	**17.93**	20.00	**8.97**	**45.72**	**44.92**
w Classifier-free guidance	**0.38**	**0.36**	94.72	98.43	13.68	**21.91**	8.79	47.62	45.28

Table 10: Statistical analysis (T-test with 95% confidence) results comparing our PerFRDiff with three state-of-the-art approaches, where different levels of statistically significant difference are indicated by * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$, and N.A. indicates no statistically significant difference. + indicates that PerReactor has outperformed the comparative approach, whereas - indicates that PerFRDiff has under-performed the comparative approach.

Comparison	FRCorr		FRdist		FRDiv	FRDvs	FRVar	FRRea	FRSyn
	MAFRG	PMAFRG	MAFRG	PMAFRG					
PerFRDiff VS. Trans-VAE [6]	+(***)	+(***)	N.A.	+(**)	+(***)	+(***)	+(***)	+(***)	-(***)
PerFRDiff VS. BeLFusion [1]	+(***)	+(***)	+(*)	N.A.	+(***)	+(***)	+(***)	+(***)	N.A.
PerFRDiff VS. REGNN [8]	+(***)	+(***)	-(***)	-(***)	+(***)	+(***)	+(***)	+(***)	-(***)

a 30-second sample of speaker behavior and a 30-second historical facial video of the listeners, and it outputs tailored 30-second facial reactions all within 5 seconds. This demonstrates the model's capability for real-time usage. Examining the details of the GARFG

Table 11: Model complexity analysis

Metrics	GAFRG	PSSL	PWSG
Single iteration time (s) in training	0.16	0.28	0.12
Single epoch time (s) in training	139	300	360
Inference time (s) for a 30s facial reaction	5	-	-
Parameters	28.93 (M)	4.76 (M)	1.63 (B)
FLOPs	1.78 (T)	3.19 (G)	1.63 (G)

training process, it takes 0.16 seconds to complete a single iteration and 2 minutes and 19 seconds for a full epoch. On the other hand, the PWSG's training involves 0.12 seconds per iteration and just 6 minutes per epoch. During PSSL training, a single iteration takes 0.28 seconds, whereas an entire epoch is completed in 5 minutes. Structurally, the GAFRG consists of 28.929 million parameters, and its computational load, measured in FLOPs (Floating Point Operations per Second), is 1.783T when generating ten facial reactions (i.e., action units (AUs), facial expression probabilities, valence and arousal intensities) of the size $10 \times 750 \times 25$. The PSSL comprises 4.76 million parameters and its FLOPs stands at 3.185G when processing a listener's historical facial behaviour of size $1 \times 750 \times 58$ (in the form of 3DMM coefficients). The PWSG comprises 1.627 billion parameters and its FLOPs stand at 1.625G when processing a personalised cognitive style representation of the size 1×512 to generate personalised weight shifts. The above computation of the FLOPs and the number of model parameters is done by using the PyTorch-OpCounter package[2]. Altogether, these features offer a comprehensive insight into the model's complexity.

9 Visualisation of facial reactions generated for different speaker behaviours

Fig. 1 to 4 display facial reactions generated by different approaches in response to different speaker behaviours. It can be clearly observed that ours show more head movements and diverse facial expressions in response to different speaker behaviours, compared to facial reactions generated by other approaches. A video (named 'demo.mp4') demonstrating speaker behaviours and the corresponding generated facial reactions in response to the speaker behaviours is provided in the Supplementary Materials folder.

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] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[3] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

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

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

[6] 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).

[7] 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.

[8] 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).

[2] https://pypi.org/project/thop/

Figure 1: Visualisation of facial reactions generated by different approaches in response to the speaker behaviour #1, where ours successfully captures the listener's personalised smile style, characterised by a slight exposure of teeth during smiling.

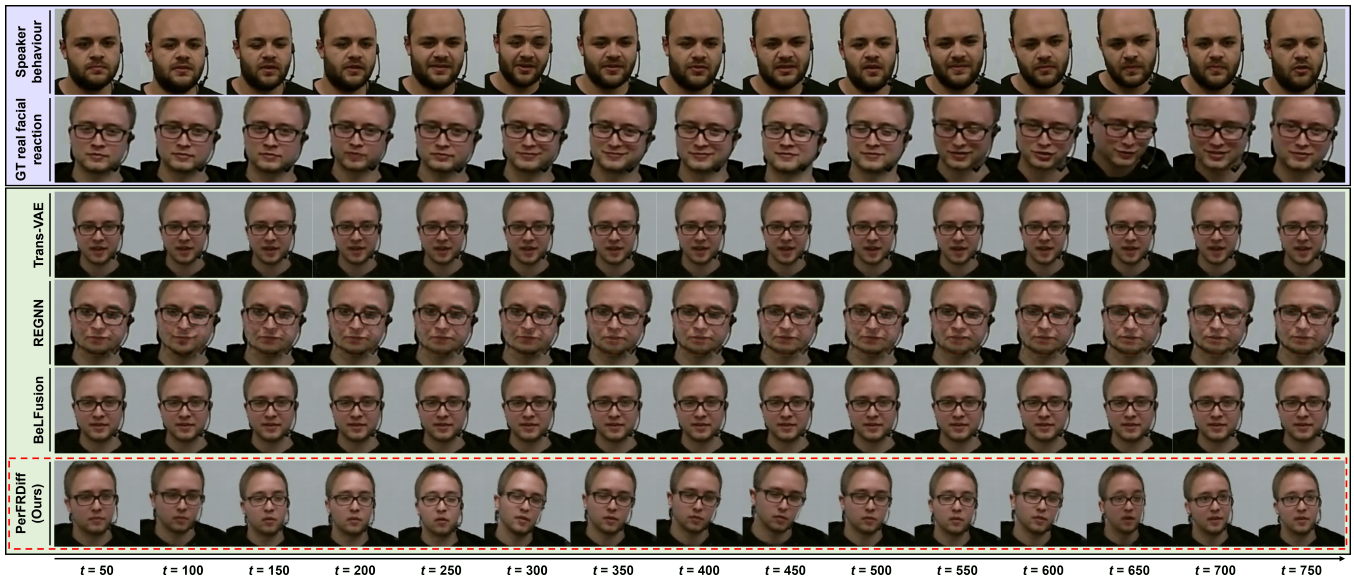

Figure 2: Visualisation of facial reactions generated by different approaches in response to the speaker behaviour #2, where ours clearly show more head movements in the interaction.

Figure 3: Visualisation of facial reactions generated by different approaches in response to the speaker behaviour #3, where ours aligns with the real AFR that the facial expressions of the listener turn to smile in response to the speaker behaviour at the 700-th and 750-th frames.

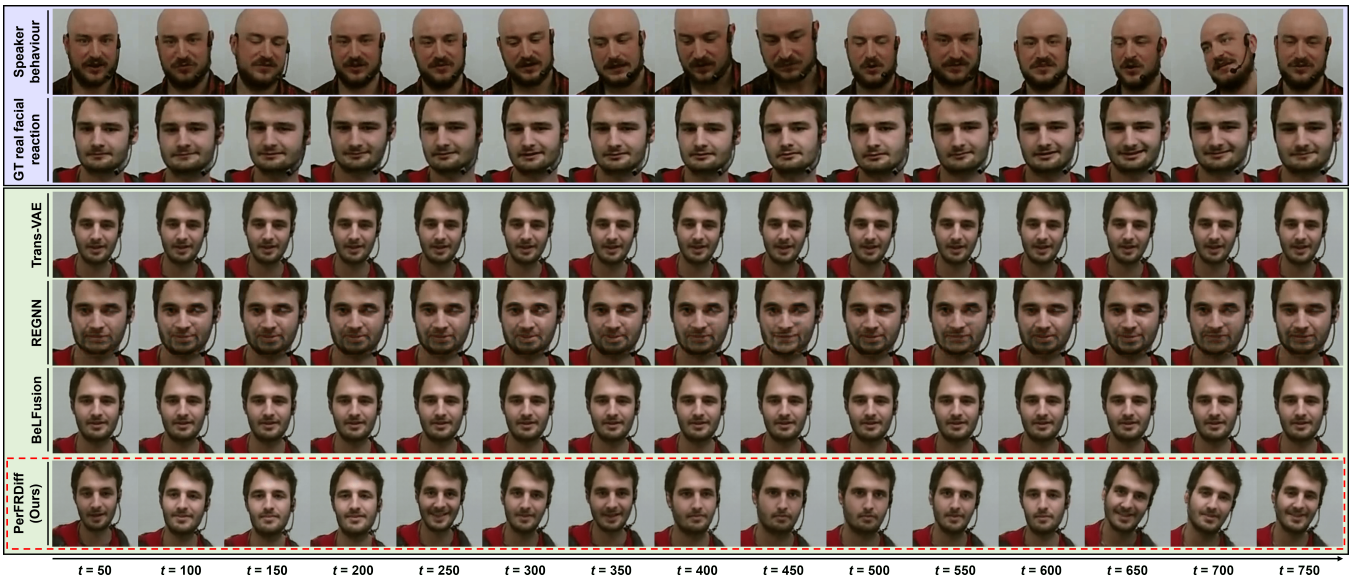

Figure 4: Visualisation of facial reactions generated by different approaches in response to the speaker behaviour #4, where ours shows more facial expressions in response to the speaker behaviour.