2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Multiple Facial Reaction Generation using Gaussian Mixture of Models and Multimodal Bottleneck Transformer

Dang-Khanh Nguyen[1], Prabesh Paudel[1], Seung-won Kim[1], Ji-eun Shin[1], Soo-hyung Kim[1] and Hyung-Jeong Yang[1*]

[1] Department of Artificial Intelligence, Chonnam National University, Gwangju, Republic of Korea

Abstract— **Facial reaction generation has gained prominence in recent years. However, while there has been extensive research on synthesizing facial expressions from the perspective of the speaker, the generation of reactions from the listener's standpoint remains relatively unexplored. Predicting the facial reactions of the listener in a conversational setting presents a challenge due to the diverse range of reactions that can be elicited by the behavior of a single speaker. In this study, we introduce a Multimodal Transformer-based Variational Autoencoder designed to learn the distribution of listener facial reactions based on speaker audiovisual cues. Our proposed approach incorporates the Multimodal Bottleneck Token mechanism to capture interactions between acoustic and visual speaker features and utilizes the Variational Autoencoder framework to generate latent representations of multiple listener reactions. Additionally, we employ Gaussian Mixture Models to enhance the generative capabilities of the Autoencoder. Experimental results demonstrate that our method surpasses baseline models and previous approaches on the REACT24 benchmark dataset.**

I. INTRODUCTION

In our everyday lives, conversations play a crucial role, involving dynamic exchanges where individuals take turns speaking and listening to convey and receive information during face-to-face interactions. While the speaker communicates verbally, the listener typically responds through non-verbal cues, providing immediate feedback. According to Song et al. [12], the listener's reactions to the same information from the speaker can vary depending on different contexts.

Despite significant research efforts focused on synthesizing speech from the speaker's perspective, there has been a lack of emphasis on generating the listener's reaction. Song et al. [9] introduce a benchmark for generating multiple appropriate facial reactions, which includes a multimodal dataset and three baseline non-deterministic models: TransVAE utilizes the Transformer-based Variational Autoencoder, the BeLFusion, and a graph neural network approach [5]. Meanwhile, Luo et al. propose ReactFace [13], an encoder-decoder architecture designed to address the challenging task of synchronizing the generated listener's reaction with the visual and acoustic features of the speaker in the temporal dimension.

In this study, our objective is to enhance the interaction between the audio and visual modalities of the speaker by employing the Multimodal Bottleneck Transformer (MBT) [6] for the encoder component. By leveraging the informative interaction features extracted from the MBT, we aim to enhance the effectiveness of the generative function performed by TransVAE and the cross-modal transformer in the decoder. In addition, we utilize the effectiveness of the Gaussian Mixture of Models [8] to learn the distribution of listener reaction. Our proposed model demonstrates a notable improvement in terms of appropriateness and diversity evaluation metrics compared to existing methodologies [11], [10], [4], [13].

Our contributions in this paper are summarized as follows:

- We utilize the Multimodal Bottleneck Transformer to perform cross-modality learning between audio and visual features extracted from the speaker's video. The audio and visual attributes can exchange their meaningful information with each other via a set of learnable tokens called bottleneck tokens. Therefore, we can improve the output of the speaker behavior multimodal feature extractor.
- We use the sequence of the speaker's emotional features to align the generated listener's reaction. By exploiting this particular information of the speaker, our model can work with 3 modalities including acoustic, visual, and emotional input.
- We attempt to model more complex distributions of the listener's reactions by applying the Gaussian Mixtures of Models. The Transformer VAE can generate more sophisticated features to describe the listener's behavior rather than features simply sampled from a single Gaussian distribution.

II. PROPOSED METHOD

Our proposed model adapts the encoder-decoder architecture receiving the video of a speaker as input and generating 3D facial features and facial reactions of the listener. The 3D facial feature is a sequence of 3D Morphable Model (3DMM) [1] coefficients used to render the sequence of frames of the listener based on a static reference image. On the other hand, the facial reaction includes three widely-used facial descriptors: the probabilities of eight emotions, 15 well-defined facial movements also known as the action units, and the facial affect consisting of valence and arousal levels [14]. The model comprises two main functional blocks: the speaker behavior multimodal feature extractor and the multiple appropriate listener reaction generator. A detailed illustration of our model is shown in Figure II.

[*] Corresponding author: hjyang@jnu.ac.kr

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

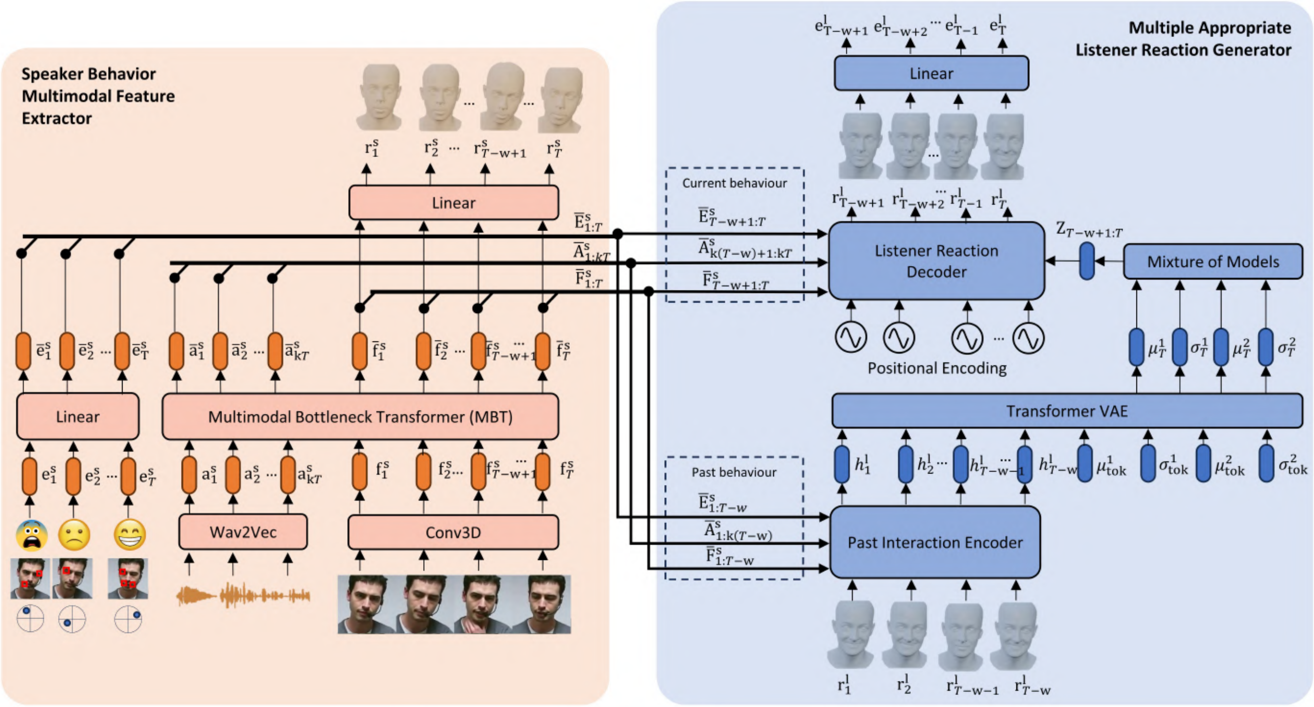

Fig. 1. Block diagram of the proposed method.

A. Speaker Behavior Multimodal Feature Extractor

Initially, the speaker encoder extracts the visual and acoustic features from the video by exploiting the dedicated neural networks. Particularly, we use pre-trained Wav2Vec for generating audio modalities feature $A^s_{1:kT} = \left\{a^s_1, a^s_2, ..., a^s_{kT} | a^s_t \in R^D\right\}$ and Conv3D neural network to obtain visual attributes $F^s_{1:T} = \left\{f^s_1, f^s_2, \dots, f^s_T | f^s_t \in R^C\right\}$. Sequentially, we propose the Multimodal Bottleneck Transformer for cross-modal learning. It is a low-cost transformer approach for fusing two time-series inputs with long sequence lengths. It adapts the idea of bottleneck tokens [12] to force the model to extract the most meaningful information from each modality. We define a sequence of M learnable vectors $B_{1:M} = \left\{b_1, b_2, b_M | b_m \in R^C\right\}$ called bottleneck tokens where M is significantly less than T. We will use them as a means to transfer information between audio and visual features. The illustration of MBT in Figure II-Aa is formulated by these below functions:

$$\begin{aligned}
\tilde{A}^s_{1:kT}||\tilde{B}^A_{1:M} &= TransEncLayer^A_1(A^s_{1:kT}||B_{1:M})\\
\tilde{F}^s_{1:T}||\tilde{B}^F_{1:M} &= TransEncLayer^F_1(F^s_{1:T}||B_{1:M})\\
\tilde{B}_{1:M} &= Average(\tilde{B}^A_{1:M}, \tilde{B}^F_{1:M}) \qquad (1)\\
\bar{A}^s_{1:kT}||\bar{B}^A_{1:M} &= TransEncLayer^A_2(\tilde{A}^s_{1:kT}||\tilde{B}_{1:M})\\
\bar{F}^s_{1:T}||\bar{B}^F_{1:M} &= TransEncLayer^F_2(\tilde{F}^s_{1:T}||\tilde{B}_{1:M})
\end{aligned}$$

Using bottleneck tokens will prevent the audio transformer encoder layer ($TransEncLayer^A_l$) from paying attention to the entire visual feature $F^s_{1:T}$, and vice versa. The audio and visual encoders only exchange their useful information via the bottleneck tokens $B_{1:M}$. Moreover, this method is more efficient than feeding the audio and visual sequence into a common transformer encoder layer. As we know the attention mechanism is sensitive to the length of the input sequence, concatenating the audio and visual features and feeding them to a self-attention layer will consume a significant computation. Because the number of bottleneck tokens M is noticeably less than the sequence length T, it decreases the input sequence length of the transformer encoder layer compared to the naïve concatenating approach. As a result, utilizing bottleneck tokens can reduce the computational cost and improve the quality of cross-modal features $\bar{A}^s_{1:T}$ and $\bar{F}^s_{1:T}$ in video understanding.

Besides video information, our network also leverages the emotional attributes of the speaker $E^s_{1:T} = \left\{e^s_1, e^s_2, ..., e^s_T | e^s_t \in R^{25}\right\}$ to gain more knowledge about the speaker's behavior. The input emotional feature is transformed to $\bar{E}^s_{1:T} = \left\{\bar{e}^s_1, \bar{e}^s_2, ..., \bar{e}^s_T | \bar{e}^s_t \in R^C\right\}$ by a linear layer so that it is projected to the same dimensional space as the audio and visual features. As a result, we obtain three latent sequences of features from three corresponding modalities. Each sequence is split into two segments including the past behavior $(\bar{A}^s_{1:k(T-w)}, \bar{F}^s_{1:T-w}, \bar{E}^s_{1:T-w})$. and the current behavior sequence of feature $(\bar{A}^s_{k(T-w)+1:kT}, \bar{F}^s_{T-w+1:T}, \bar{E}^s_{T-w+1:T})$. The length of the current behavior is also the number of timesteps computed in one inference called the window size w. Both current and past behavior features are fed into the multiple appropriate listener reaction generator. Additionally, in the training process, we also regenerate the 3D facial attributes of the speakers $R^s_{1:T} = \{r^s_1, r^s_2, ..., r^s_T\}$ from the latent visual features and used as an auxiliary output of the model.

B. Multiple Appropriate Listener Reaction Generator

Multiple Appropriate Listener Reaction Generator is a nondeterministic functional block that generates the sequence of listener's aligned emotional features from the speaker's behavior features. Firstly, a past interaction encoder leverages the speaker behavior features in the past including audio, visual, and emotional features to extract the enhanced interaction features $H^l_{1:T-w} = \{h^l_1, h^l_2, ..., h^l_{T-w}\}$ from the past predictions of the 3D listener face $R^l_{1:T-w} = \{r^l_1, r^l_2, \ldots, r^l_{T-w}\}$ by transformer-based cross-attention. The process of Past Interaction Encoder is illustrated in Figure II-Ab and formulated by below formula:

$$\begin{aligned} X^1_{1:T-w} &= Linear(R_(1:T-w)^l) \\ X^2_{1:T-w} &= TransDecLayer_1(X^1_{1:T-w}, \bar{F}^s_{1:T-w}) \\ X^3_{1:T-w} &= TransDecLayer_2(X^2_{1:T-w}, \bar{A}^s_{1:k(T-w)}) \\ H^l_{1:T-w} &= TransDecLayer_3(X^3_{1:T-w}, \bar{E}^s_{1:T-w}) \end{aligned} \quad (2)$$

Where $X^1_{1:T-w}, ..., X^3_{1:T-w}$ are temporary variables and $TransDecLayer$ is used to perform cross-attention. Afterward, we define 2 pairs of learnable tokens: $^1_{tok} - ^1_{tok} and ^2_{tok} - ^2_{tok}$ aiming to construct two Gaussian distributions. A conventional transformer autoencoder is employed to learn two suitable facial reaction distributions for the listener based on the enhanced interaction features $H^l_{1:T-w}$. Two sequences of vectors $Z^1_{t-w+1:t}$ and $Z^2_{t-w+1:t}$ are sampled from two corresponding distributions and aggregated by:

$$Z_{t-w+1:t} = \alpha_1 Z^1_{t-w+1:t} + \alpha_2 Z^2_{t-w+1:t} \quad (3)$$

Where α_1 and α_2 are two trainable scalars. The aggregated sequence of vectors $Z_{T-w+1:T}$ represents the facial reaction of the listener. Based on these latent features, a sequence of 3D listener facial coefficients is synthesized by the Listener Reaction Decoder using the audio, visual, and emotional features of the speaker's current behavior. The operation of the Listener Reaction Decoder is shown in Figure II-Ad and formulated by the equations below:

$$\begin{aligned} Y^1_{T-w+1:T} &= TransDecLayer_1(PE, Z_{t-w+1:t}) \\ Y^2_{T-w+1:T} &= TransDecLayer_2(Y^1_{T-w+1:T}, F^s_{T-w+1:T}) \\ Y^3_{T-w+1:T} &= TransDecLayer_3(Y^2_{T-w+1:T}, A^s_{k(T-w)+1:kT}) \\ Y^4_{T-w+1:T} &= TransDecLayer_4(Y^3_{1:T-w}, E^s_{T-w+1:T}) \\ R^l_{T-w+1:T} &= Linear(Y^4_{T-w+1:T}) \end{aligned} \quad (4)$$

Where $Y^1_{T-w+1:T}, ..., Y^4_{T-w+1:T}$ are temporary variables and PE is positional encoding. Finally, the prediction of the listener's facial reaction $E^l_{T-w+1:T} = \{e^l_{T-w+1}, e^l_{T-w+2}, \ldots, e^l_T | e^l_t \in R^2 5\}$ is obtained from a linear mapping layer with the predicted 3D features of the listener $R_(T-w+1:T)^l$.

$$E^l_{T-w+1:T} = Linear(R^l_{T-w+1:T}) \quad (5)$$

III. EXPERIMENTS

A. Datasets

Our experiments were conducted on the REACT24 dataset, which is a compilation of two multimodal datasets NoXI [2] and RECOLA [9]. There are 2962 pairs of speaker's and listener's clips. To construct a multiple appropriate facial reaction dataset, Song et al. [12] applied the automatic appropriate facial reaction labeling strategy to define the correct facial reactions corresponding to each speaker in the dataset.

Regarding the assessment, we follow Song et al. [12] using a well-defined set of evaluation metrics to measure the appropriateness, diversity, realism, and synchrony of our proposed model's output. Concisely, we use Facial reaction distance (FRDis) and Facial reaction correlation (FRC) for appropriateness measurement. To evaluate diversity, we adapt Facial reaction variance (FRVar), Diverseness among generated facial reactions (FRDiv), and Diversity among facial reactions generated from different speaker behavior (FRDvs). For synchrony evaluation, the Time Lagged Cross Correlation (TLCC) is computed to measure the leader-follower relationship between the speaker and listener (FRSyn). Lastly, they use the Frechet Inception Distance (FID) as the realism score (FRRea). The higher value of the mentioned evaluation metrics is better, except FRDis, FRRea, and FRSyn.

B. Experiment settings

We followed the pre-processing steps of the baseline framework. Facial images from all frames were cropped and resized to 224x224. The EMOCA [3] models were utilized to extract 3D Morphable Model (3DMM) coefficients including the pose and expression parameters following the FLAME 3DMM. We used PIRender [7] to render the sequence of facial frames from the 3DMM coefficients. We implemented our solution using the Pytorch library and executed it on the Linux machine with the NVIDIA A100 GPU. Regarding the hyperparameters of MBT, we used 2 layers of transformer encoder with 4 heads of self-attention and a latent dimension of 128. The number of bottleneck tokens should be significantly less than the length of sequence input so we used 4 bottleneck tokens in our experiments. For the GMM, we used two Gaussian distributions with corresponding learnable weights to describe the listener reaction distribution. Discussing the training settings, we clipped the input videos at random positions with a length of 256 frames. We trained our model on these clips using the AdamW optimizer with a batch size of 4 for 20 epochs. Meanwhile, in the evaluation, we used full clips with 750-frame lengths and decreased the batch size to 1. In both training and evaluation, the window size for online mode is 8 frames per prediction.

C. Experimental results

Firstly, we evaluated variations of our proposed method in terms of appropriateness and diversity on the validation set. Due to the similarity in model structure, we consider ReactFace [13] as our main competitor. We reproduced the results of the ReactFace [13] model and Trans-VAE[11]

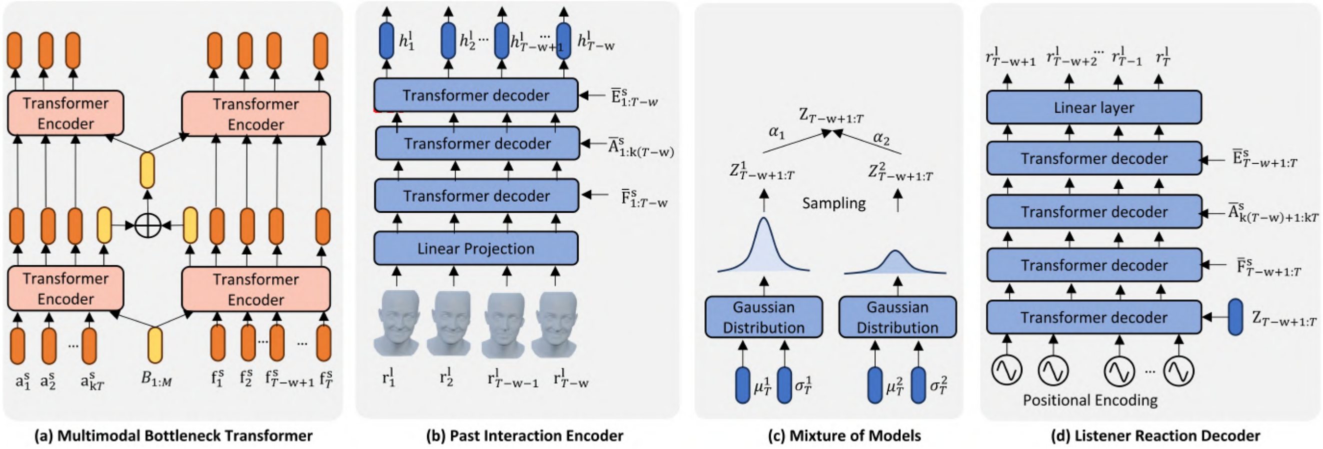
Transformer Encoder
Transformer Encoder
Transformer Encoder
Transformer Encoder
a_1^s a_2^s a_{kT}^s
$B_{1:M}$
f_1^s f_2^s f_{T-w+1}^s f_T^s
(a) Multimodal Bottleneck Transformer
h_1^l h_2^l h_{T-w+1}^l h_{T-w}^l
Transformer decoder
Transformer decoder
Transformer decoder
Linear Projection
$\overline{E}_{1:T-w}^s$
$\overline{A}_{1:k(T-w)}^s$
$\overline{F}_{1:T-w}^s$
r_1^l r_2^l r_{T-w-1}^l r_{T-w}^l
(b) Past Interaction Encoder
$Z_{T-w+1:T}$
α_1 α_2
$Z_{T-w+1:T}^1$ $Z_{T-w+1:T}^2$
Sampling
Gaussian Distribution
Gaussian Distribution
μ_T^1 σ_T^1 μ_T^2 σ_T^2
(c) Mixture of Models
r_{T-w+1}^l r_{T-w+2}^l r_{T-1}^l r_T^l
Linear layer
Transformer decoder
Transformer decoder
Transformer decoder
Transformer decoder
$\overline{E}_{T-w+1:T}^s$
$\overline{A}_{k(T-w)+1:kT}^s$
$\overline{F}_{T-w+1:T}^s$
$Z_{T-w+1:T}$
Positional Encoding
(d) Listener Reaction Decoder

Fig. 2. Illustration of submodules in our proposed method.

TABLE I

THE RESULTS ON VALIDATION SET. (1) DENOTES OUR MODEL WITHOUT USING SPEAKER EMOTION AND GMM. (2) DENOTES OUR MODEL WITHOUT USING THE GMM. (3) DENOTE OUR MODEL USING SPEAKER EMOTION AND GMM.

Method	Appropriateness		Diversity		
	FRCorr	FRDist	FRDiv	FRVar	FRDvs
TransVAE	0.17	127.07	0.0024	0.0013	0.0024
ReactFace	3.74	50.68	0.1293	0.0579	0.1293
Ours $^{(1)}$	4.25	55.18	0.1561	0.0773	0.1578
Ours $^{(2)}$	4.99	93.77	0.1166	0.0703	0.1191
Ours $^{(3)}$	4.52	72.77	0.1910	0.0988	0.1926

model and compared them with our method. As shown in Table I, the ReactFace model outperformed the baseline Trans-VAE across all evaluation metrics. Our model, using MBT and ignoring the speaker's emotion and GMM, achieved higher results than the ReactFace model, except for the FRDis score. By leveraging the speaker's emotion, our model can improve the appropriateness but trade-off with the decrement of diversity metrics. Eventually, when the GMM was applied, it significantly boosted the diversity of our models' output. As a result, our full model generates more diverse listener reactions compared to the ReactFace model. Concerning the appropriateness, our model achieves a better correlation metric while ReactFace attains a lower distance to the ground truth.

Next, we chose our model with the best diversity scores in the validation set and conducted the assessment and comparison on the test set. According to Table II, our proposed model outperforms the Trans-VAE and BeLFusion baselines in FRDist, FRVar, and FRDvs. These baselines have relatively low variations when our proposed method can accomplish approximately similar diversity scores to the grouth truth. Regarding the appropriateness, our FRDist exhibits a significant improvement compared to previous studies. As a result, our model successfully synthesizes multiple appropriate facial reactions in a dyadic conversation setting.

IV. CONCLUSIONS

Bottleneck Transformer to improve the extracted interaction feature between speaker and listener in conversation. Based on the informative video feature and the emotion of the speaker, we use the Transformer Variational Autoencoder combined with the Gaussian Mixture of Models to generate multiple appropriate reactions from the listener. This framework accomplishes a noticeable enhancement compared to prior methods in multiple appropriate facial reaction generation. Our future research will focus on balancing the objective functions including the reconstruction loss, energy-based diversity, and the Kullback-Leibler loss in the training process.

V. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219107). This work was also supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP- 2023-RS-2023-00256629) grant funded by the Korean government (MSIT).

REFERENCES

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

[2] 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 19th ACM International Conference on Multimodal Interaction*, pages 350–359, 2017.

[3] R. Daněček, M. J. Black, and T. Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[4] X. Hoque, A. Mann, G. Sharma, and A. Dhall. Beamer: Behavioral encoder to generate multiple appropriate facial reactions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9536–9540, 2023.

TABLE II

The results on Test set

Method	Appropriateness		Diversity			Realism	Synchorny
	FRCorr	FRDist	FRDiv	FRVar	FRDvs	FRRea	FRSyn
GT	8.73	0.00	0.0000	0.0724	0.2483	53.96	47.69
TransVAE	0.07	90.31	0.0064	0.0012	0.0009	69.19	44.65
BeLFusion	0.12	91.45	0.0112	0.0082	0.0120	-	44.89
Ours	0.03	11.68	0.0000	0.1006	0.1960	51.28	45.29

[5] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022.

[6] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[7] 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*, pages 13759–13768, 2021.

[8] D. A. Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[9] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. 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)*, pages 1–8. IEEE, 2013.

[10] 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 31st ACM International Conference on Multimedia*, pages 9620–9624, 2023.

[11] 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. *arXiv preprint arXiv:2401.05166*, 2024.

[12] 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.

[13] 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.

[14] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.