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

One-to-Many Appropriate Reaction Mapping Modeling with Discrete Latent Variable

Zhenjie Liu, Cong Liang, Jiahe Wang, Haofan Zhang, Yadong Liu, Caichao Zhang, Jialin Gui and Shangfei Wang*
University of Science and Technology of China

***Abstract*— In dyadic interaction, listener reaction generation can be treated as a one-to-many mapping problem since multiple listener reactions can correspond to a given speaker action. The existing methods have not modeled the diversity of contextual factors well and fail to generate diverse appropriate listener reactions. In response, we introduce discrete latent variables to tackle this one-to-many mapping problem. We conducted experiments on the datasets provided by the REACT2024 Challenge, and the results demonstrated that our approach is capable of generating appropriate listening reactions with higher diversity. Our method achieved first place in the offline track and second in the online track.**

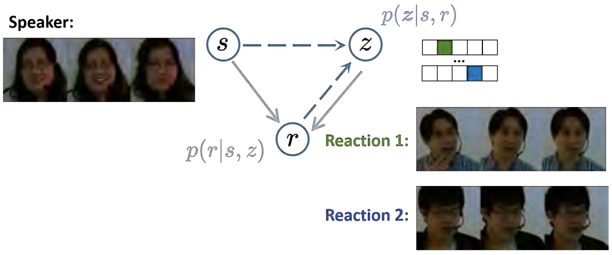

Fig. 1. Graphical illustration of listener reaction generation (gray lines) and posterior latent distribution estimation (dashed blue lines).

I. INTRODUCTION

Dyadic interaction, characterized by one-to-one conversations, has a considerable role in human society. Two roles are always involved in a single dyadic interaction: a speaker and a listener. Research on listener behaviors is significantly less prominent than speaker-centric generation, such as audio-driven talking head generation. Generally, dyadic interaction can be considered a one-to-many mapping problem, where *multiple appropriate listener reactions (e.g. facial expressions) can correspond to a given speaker action*[23].

In light of this, the REACT2023 challenge [21] first formalized the Multiple Appropriate Reaction Generation (MARG) problem and gained widespread attention. Following the successful organization of that, the second REACT Challenge [22] was proposed, focusing on generating multiple appropriate, diverse, realistic, and synchronized facial reactions under both online and offline settings.

Some previous studies have investigated this MARG task by now. They model one-to-many relationships by leveraging the inherent non-determinism in the model generation process[4][13][28], while their results were far from satisfactory. Yu et al.[28] used a diffusion model to generate listener features, and Liang et al.[13] proposed a UniLM-based[8] model that unifies both online and offline tasks through a shared fast-forward layer.

Other methods [16][1][27] summarized appropriate facial reaction distributions during the training phase to represent all facial reactions considered appropriate in response to the speaker's behavior. Based on this, they can sample a set of embeddings representing different appropriate facial reactions for generation for inference. However, the facial reaction of a listener to a speaker's behavior depends not solely on the stimulus presented by the speaker but also on contextual factors such as the conversational environment and the listener's disposition. Thus, simply utilizing the speaker's actions can not capture the appropriateness of the one-to-many mapping.

Following the success in the verbal diverse conversation flow generation [6][2][3], we introduce discrete latent variables to tackle this one-to-many mapping problem. Each value of the latent variable corresponds to the particular reaction intent of one response. Apart from the aforementioned approaches, the latent variable is generated through a posteriori estimation process. As shown in Fig. 1, given the speaker actions and one selected appropriate listener reactions, the underlying latent reaction intents can be estimated as $p(z|s,r)$. We can generate multiple listener reactions conditioned on the speaker actions and latent variables. We conducted experiments on the datasets provided by the REACT2024 challenge, and the results demonstrated that our approach is capable of generating listener reactions with higher diversity. As a result, we achieved first place in the offline track and second in the online track.

II. METHODOLOGY

Let $\boldsymbol{V} = \{v_1, \cdots, v_T\}$ denotes the speaker video with T frames, we divide the corresponding audio $\boldsymbol{A}$ into windows of 40ms length each for aligning to video frames, denoted as $\boldsymbol{A} = \{a_1, \cdots, a_T\}$. Given $\boldsymbol{V}$ and $\boldsymbol{A}$, the task is to predict non-verbal listener facial reaction sequence $\boldsymbol{R} = \{r_1, \cdots, r_T\}$, in which the reaction frame $r_t = \{au_t^1, \cdots, au_t^{15}, valence_t, arousal_t, exp_t^1, \cdots, exp_t^8\}$ consists of 15 Facial Action Units (AUs), Valence and Arousal (VA), and eight facial expressions.

The architecture of our model is illustrated in Fig. 3, utilizing the same Transformer Encoder-Decoder backbone for both online and offline tasks.

[*] The corresponding author.

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

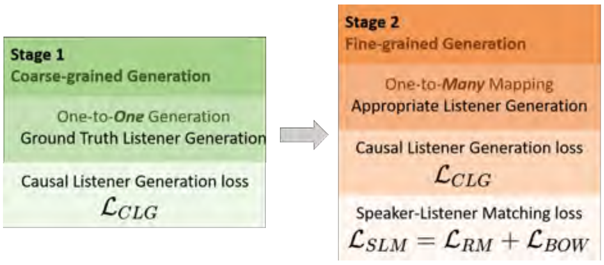
Stage 1
Coarse-grained Generation
One-to-*One* Generation
Ground Truth Listener Generation
Causal Listener Generation loss
$\mathcal{L}_{CLG}$
Stage 2
Fine-grained Generation
One-to-*Many* Mapping
Appropriate Listener Generation
Causal Listener Generation loss
$\mathcal{L}_{CLG}$
Speaker-Listener Matching loss
$\mathcal{L}_{SLM} = \mathcal{L}_{RM} + \mathcal{L}_{BOW}$

Fig. 2. Curriculum learning process.

A. Data Processing

We leverage off-the-shelf tools to extract reaction-relevant features, limiting video input to facial features. Since the target is to generate AUs, VA, and expression distribution, We use extracted features of these three as part of the encoder input. Following [13], we use MEGraphAU [14], ResMaskNet [17] and FaceTorch[1] to extract these three features. In addition, to generate high-fidelity listener videos, we employ the 3D Morphable Model (3DMM) [26] as an intermediary representation and part of the speaker feature encoder.

For the audio, We extract the Mel-frequency cepstral coefficients (MFCC) feature with the corresponding MFCC Delta and Delta-Delta features. Furthermore, we incorporate deep speech features extracted by a fine-tuned Wav2Vec2 [7] model[2].

To adapt the transformer model, we further converted the listener's non-verbal reaction label into tokens. Expression distribution and 3DMM coefficients are tokenized using K-means to assign class center indexes. The AUs are tokenized by identifying all combinations of 15 AUs' activation per frame and assigning these combinations of unique identifiers in the AU codebook. And VA values are tokenized by mapping them to the nearest 80×80 grid point and assigning identifiers in the VA codebook.

B. Transformer Block

In the speaker encoder, after extracting the features, we stack the output hidden states of a linear layer over the feature dimension. And in the listener decoder, the same treatment is applied to the embeddings. Unlike methods that generate tokens of different phases in the temporal dimension, such as [13] and [12], our treatment improves computational efficiency and has better temporal consistency. Also, since hidden states are stacked and transformed from features of different phases, we use the gated attention unit (GAU) [10] instead of multi-head attention (MHA) in the vanilla Transformer[25] to prevent this relationship from breaking. GAU is a *single-head* gated attention mechanism and generally uses a small key width and a large value width, with two GAUs replacing a single transformer layer.

C. Training

Since the experimental results show that it is more difficult for the model to learn one-to-many relationships directly, we leverage curriculum learning to learn the generation of listener reactions gradually. As shown in Fig. 2, the learning process involves two stages: during Stage 1, a coarse-grained baseline model is trained under the simplified one-to-one mapping relationship; during Stage 2, the model for fine-grained generation is further trained to generate diverse listener reactions. In the following, we will provide the details about the two-stage curriculum learning processes.

1) Coarse-grained Generation: We first train a coarse-grained baseline model under the simplified relationship of one-to-one mapping, in which the training is achieved by directly pairing the input speaker actions with the corresponding listener's ground truth facial reactions. Given one training pair of speaker and listener (s, r), we need to minimize the following **Causal Listener Generation (CLG) loss**:

$$\mathcal{L}_{CLG} = \sum_{t=1}^{T}\sum_{c\in\mathcal{C}} CrossEntropy\Big(p(\hat{r}_{t,c}|s, r_{<t}),\ r_{t,c}\Big) \quad (1)$$

where T is the length of listener reactions and $r_{<t}$ donates the previously generated tokens. And c refers to four types of tokens to generate: 3DMMs, AUs, VA, and expressions, respectively.

2) Fine-grained Generation: Based upon the coarse-grained baseline model, we further train the model under the relationship of one-to-many mapping. We randomly sample one reaction sequence from all appropriate reactions to the speaker as the label in each iteration. A K-way categorical variable z is introduced for modeling one-to-many relationships. Each value of the latent variable corresponds to the particular reaction intent of one response. The latent variable is generated through a posteriori estimation process:

$$\begin{aligned} z_1, z_2 &\sim p(\boldsymbol{z}|s, r) \\ &= \text{Gumbel-Softmax}(W_z h_z + b_z) \in \mathbb{R}^K \end{aligned} \quad (2)$$

where $h_z \in \mathbb{R}^d$ is the final decoder hidden state of the input special mask token `[M]`. We use Gumbel-Softmax[11] to sample from the categorical distribution as it is differentiable.

We then employ three loss functions: **Causal Listener Generation (CLS) loss, Reaction matching (RM) loss, and Bag of Words (BOW) loss**.

Based on (1) and conditioned on the latent variable and the speaker, the CLG loss in stage 2 is defined as:

$$\mathcal{L}_{CLG} = \mathbb{E}_{z\sim p(\boldsymbol{z}|s,r)} \sum_{t=1}^{T}\sum_{c\in\mathcal{C}} \Big(p(\hat{r}_{t,c}|z, s, r_{<t}),\ r_{t,c}\Big) \quad (3)$$

and c refers to four types of tokens to generate: 3DMMs, AUs, VA, and expressions, respectively.

Following previous works [13][2], a Reaction matching (RM) loss is introduced to help distinguish whether the reaction is appropriate to the speaker's action and consistent with the context. The positive and anchor training samples

[1] https://pypi.org/project/facetorch/

[2] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-french

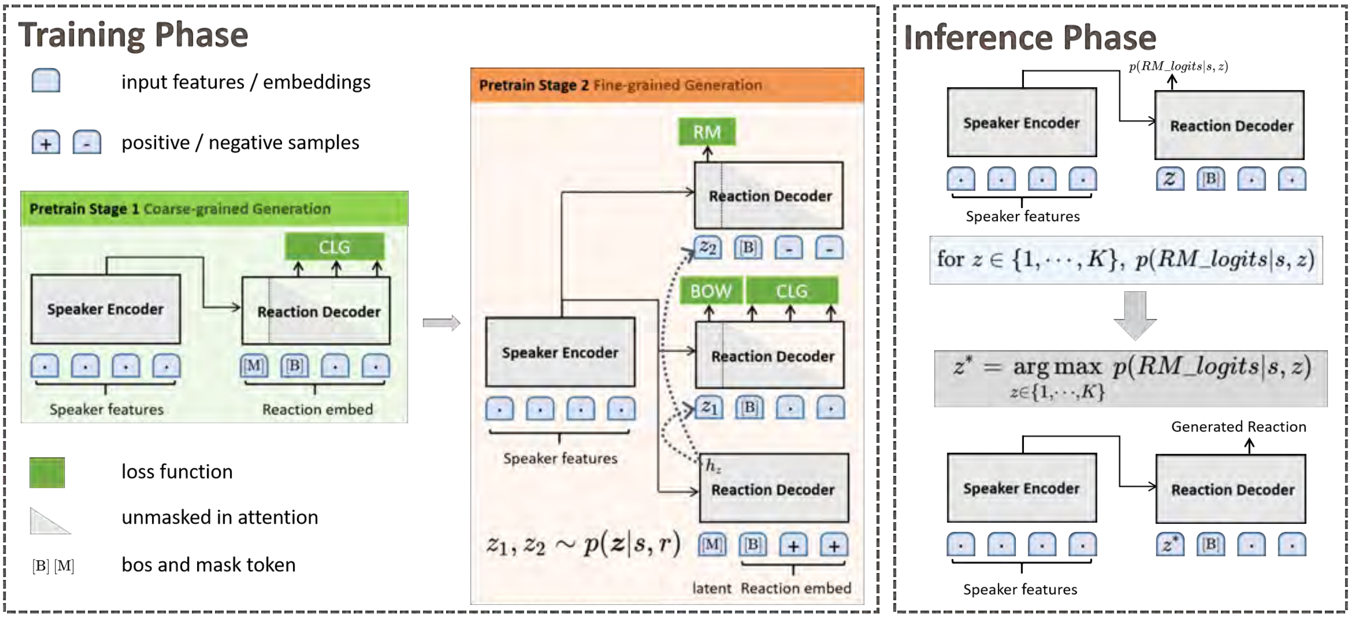

Fig. 3. Architecture of appropriate listener reaction generation with discrete latent variable.

are randomly sampled from all appropriate reactions, and the negative samples r^- are created by randomly selecting non-appropriate responses from the dataset.

$$\begin{aligned}\mathcal{L}_{RM} = &-\log p\Big(linear(h_{z_1}) = 1|s,r,z_1\Big)\\ &-\log p\Big(linear(h_{z_2}) = 0|s,r^-,z_2\Big)\end{aligned} \tag{4}$$

Besides the RM loss, the bag of words (BOW) loss [29] is also employed:

$$\mathcal{L}_{BOW} = -\mathbb{E}_{z_1\sim p(z|s,r)} \sum_{t=1}^{T}\sum_{c\in\mathcal{C}} \log \frac{\exp f(h_{z_1})_{t,c}}{\sum\limits_{v\in\mathcal{V}_c} \exp f(h_{z_1})_{v,c}} \tag{5}$$

where $\mathcal{V}_c$ refers to the whole vocabulary. The function f is defined as follows:

$$f(x) = softmax(Wx+b) \in \mathbb{R}^{|\mathcal{V}_c|} \tag{6}$$

The BOW loss discards the order of reaction tokens and forces the latent variable to capture the global information of the target listener response. The final loss for stage 2 is a summation of the above losses:

$$\mathcal{L} = \lambda_c\mathcal{L}_{CLG} + \lambda_r\mathcal{L}_{RM} + \lambda_b\mathcal{L}_{BOW} \tag{7}$$

D. Inference

We use auto-regression for generation. In the offline case, we utilize the whole temporal dimension of speaker features. In the online case, only information up to the timestep $t-1$ is utilized to predict the reaction at timestep t for both the speaker and the listener. And the inference is carried out with the second stage's models as follows:

1. Conditioned on *each* latent value $z \in \{1,\cdots,K\}$, generate K corresponding listener reactions and compute RM logits $p(RM\_logits|s,z)$.

2. Preform ranking and select the latent with the highest coherence value:

$$z^* = \arg\max_{z\in\{1,\cdots,K\}} p(RM\_logits|s,z) \tag{8}$$

and based on which generate multiple listener reactions.

In generating the listener tokens, we adopt language model (LM)[18] hyperparameters such as temperature, top-k, and top-p for sampling control, which can empirically balance appropriateness and diversity, similar to [13].

III. EXPERIMENTS

A. Dataset

We evaluate our method on the official dataset[20][15][24][9] of the REACT2024 challenge, which consists of 5924 clips of 30 seconds each (3188 training examples, 1124 validation examples, and 1612 test examples). The video clips are selected from the existing RECOLA[19] and NoXI[5] datasets.

B. Evaluation Metrics

We follow the baseline papers [21][22][23] to evaluate our method using six metrics: facial reaction correlation (FRCorr), appropriate facial reaction distance (FRDist), diverseness of facial reactions (FRDiv), facial reaction variance (FRVar), diversity among facial reactions generated from different speaker behaviors (FRDvs), and synchrony between generated facial reactions and speaker behaviors (FRSyn). The detailed formulations can be found in the theory paper [23]. The appropriateness of reaction is measured by FRCorr and FRDist, while diversity is measured by FRDiv, FRVar, and FRDvs. Other metrics are auxiliary.

TABLE I

RESULTS ACHIEVED ON THE TEST SET

Method	**Appropriateness**		**Diversity**			**Synchrony**
	FRCorr (↑)	**FRDist** (↓)	**FRDiv** (↑)	**FRVar** (↑)	**FRDvs** (↑)	**FRSyn** (↓)
B_Random	0.05	237.21	0.1667	0.0833	0.1667	43.84
B_Mime	0.38	92.94	0.0000	0.0724	0.2483	38.54
			Offline Results			
Trans-VAE	0.03	92.81	0.0008	0.0002	0.0006	43.75
BeLFusion (k=1)	0.10	[92.32]	0.0068	0.0073	0.0094	44.94
BeLFusion (k=10)	0.12	91.60	0.0105	0.0082	0.0116	44.87
BeLFusion (k=10) + Binarized AUs	0.12	94.16	[0.0360]	[0.0249]	[0.0384]	49.00
REGNN	**0.19**	**84.54**	0.0007	0.0061	0.0342	**41.35**
Ours (w/o latent)	0.1664	93.97	0.1018	0.0325	0.1	44.71
Ours (w/ latent)	[0.139]	140.5	**0.3059**	**0.1409**	**0.2881**	[44.19]
			Online Results			
Trans-VAE	0.07	**90.31**	0.0064	0.0012	0.0009	[44.65]
BeLFusion (k=1)	[0.12]	91.11	0.0083	0.0079	0.0103	45.17
BeLFusion (k=10)	[0.12]	[91.45]	0.0112	0.0082	0.0120	44.89
BeLFusion (k=10) + Binarized AUs	[0.12]	94.09	[0.0379]	[0.0248]	[0.0397]	49.00
Ours (w/o latent)	**0.1587**	92.05	0.1029	0.0387	0.1065	44.52
Ours (w/ latent)	0.1436	135.5	**0.303**	**0.139**	**0.2878**	**44.05**

TABLE II

FINAL RESULTS AND RANKINGS

Rank	**Team Name**	**Appropriateness**		**Diversity**			**Synchrony**
		FRCorr (↑)	**FRDist** (↓)	**FRDiv** (↑)	**FRVar** (↑)	**FRDvs** (↑)	**FRSyn** (↓)
				Offline Results			
1	**USTC-AC (Our Team)**	0.2172	100.43	0.1675	0.0535	0.1385	44.54
				Online Results			
1	AISLAB	0.3104	84.94	0.1167	0.0349	0.1165	47.43
2	**USTC-AC (Our Team)**	0.2186	88.32	0.1029	0.0387	0.1065	44.41
3	CNU_SCLAB	0.0322	11.68	0.0000	0.1006	0.196	45.29

C. *Results*

We compare our approach with the baseline methods and present the results in terms of appropriateness, diversity, and synchrony metrics. The overall scores on the test set are provided in Table I. Where bolded numbers (**x**) indicate the best results, underlining (x) indicates the second-best results, and brackets ([x]) indicate the third-best results.

Our approach outperforms the baseline methods in most metrics. In addition, the model with hidden variables achieves the highest diversity, while one without hidden variables shows better appropriateness measured by FRCorr and FRDist. These results indicate that utilizing hidden variables significantly enhances generation diversity while maintaining high appropriateness.

As shown in Table II, we achieved first place in the offline track and second in the online track.

IV. CONCLUSION

In this paper, we utilize discrete latent variables to model the one-to-many appropriate reaction relationship to generate diverse and appropriate reactions in dyadic interactions. Experimental results show that our method achieves impressive diversity with competitive appropriateness. As a result, we achieved first place in the offline track and second in the online track.

V. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 92048203 and under Grant 62376255, and in part by Anhui Science Technology Agency under Grant 202104h04020011.

References

[1] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022.

[2] S. Bao, H. He, F. Wang, H. Wu, and H. Wang. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*, 2019.

[3] S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Guo, Z. Liu, and X. Xu. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*, 2020.

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

[5] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. T. Torres, C. Pelachaud, E. André, and M. F. Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017.

[6] C. Chen, J. Peng, F. Wang, J. Xu, and H. Wu. Generating multiple diverse responses with multi-mapping and posterior mapping selection. *arXiv preprint arXiv:1906.01781*, 2019.

[7] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[8] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

[9] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[10] W. Hua, Z. Dai, H. Liu, and Q. V. Le. Transformer quality in linear time. In *International Conference on Machine Learning*, 2022.

[11] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[12] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[13] C. Liang, J. Wang, H. Zhang, B. Tang, J. Huang, S. Wang, and X. Chen. Unifarn: Unified transformer for facial reaction generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9506–9510, 2023.

[14] 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.

[15] 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. In *International Joint Conference on Artificial Intelligence*, 2022.

[16] C. Luo, S. Song, W. Xie, M. Spitale, L. Shen, and H. Gunes. Reactface: Multiple appropriate facial reaction generation in dyadic interactions. *arXiv preprint arXiv:2305.15748*, 2023.

[17] L. Pham, T. H. Vu, and T. A. Tran. Facial expression recognition using residual masking network. In *2020 25Th international conference on pattern recognition (ICPR)*, pages 4513–4519. IEEE, 2021.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[19] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[20] S. Song, Y. Song, C. Luo, Z. Song, S. Kuzucu, X. Jia, Z. Guo, W. Xie, L. Shen, and H. Gunes. Gratis: Deep learning graph representation with task-specific topology and multi-dimensional edge features. *ArXiv*, abs/2211.12482, 2022.

[21] 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.

[22] 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.

[23] 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.

[24] 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*, 3:42 – 50, 2021.

[25] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[26] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022.

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

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

[29] T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.