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

REACT 2024: the Second Multiple Appropriate Facial Reaction Generation Challenge

Siyang Song[1,3*†], Micol Spitale[2,3*], Cheng Luo[4*], Cristina Palmero[5], German Barquero[5], Hengde Zhu[1], Sergio Escalera[5], Michel Valstar[6], Tobias Baur[7], Fabien Ringeval[8], Elisabeth André[7], and Hatice Gunes[3]

[1] University of Leicester, UK; [2] Politecnico di Milano, Italy

[3] University of Cambridge, UK; [4] Monash University, Australia

[5] Universitat de Barcelona, Spain; [6] University of Nottingham, UK

[7] University of Augsburg, Germany; [8] Université Grenoble Alpes, France

***Abstract*— In dyadic interactions, humans communicate their intentions and state of mind using verbal and non-verbal cues, where multiple different facial reactions might be *appropriate* in response to a specific speaker behaviour. Then, how to develop a machine learning (ML) model that can automatically generate multiple appropriate, diverse, realistic and synchronised human facial reactions from an previously unseen speaker behaviour is a challenging task. Following the successful organisation of the first REACT challenge (REACT 2023), this edition of the challenge (REACT 2024) employs a subset used by the previous challenge, which contains segmented 30-secs dyadic interaction clips originally recorded as part of the NOXI and RECOLA datasets, encouraging participants to develop and benchmark Machine Learning (ML) models that can generate multiple *appropriate* facial reactions (including facial image sequences and their attributes) given an input conversational partner's stimulus under various dyadic video conference scenarios. This paper presents: (i) the guidelines of the REACT 2024 challenge; (ii) the dataset utilized in the challenge; and (iii) the performance of the baseline systems on the two proposed sub-challenges: Offline Multiple Appropriate Facial Reaction Generation and Online Multiple Appropriate Facial Reaction Generation, respectively. The challenge baseline code is publicly available at `https://github.com/reactmultimodalchallenge/baseline_react2024`.**

I. INTRODUCTION

Recent years have seen an increasing number of studies targeting human-human dyadic interaction analysis [4]. Previous studies [9], [17], [7] have investigated the problem of automatically generating a specific reaction that resembles the ground-truth (real) response or reaction for a given input. Most of these studies proposed deterministic approach that aims to reproduce the ground-truth reaction without considering the non-verbal aspects that enrich the message conveyed. Few studies have looked into the generation of appropriate reactions as non-verbal behaviours, with a main focus on generating a *single appropriate* reaction, e.g., hand gesture [7], facial reaction [9], [18], [16], [13], [28], or full-body postures [5].

As discussed in [21], given a human behaviour (called speaker behaviour), multiple appropriate facial reactions could be expressed by not only different individuals but also the same individual under different situations in response to it. Consequently, a Multiple Appropriate Facial Reaction Generation (MAFRG) task has been proposed. The REACT 2024 Challenge is the second competition event aimed at comparison of machine learning methods for MAFRG tasks, with all participants competing under strictly the same conditions. The REACT 2024 Challenge follows the similar purpose and form as the REACT 2023 challenge [20], focusing on two MAFRG tasks: offline and online Multiple Appropriate Facial Reaction Generation (offline and online MAFRG).

Although the organization of the REACT 2023 challenge facilitated the creation of several successful solutions [25], [12], [27], [10], [8], [24], [3] for both online and offline MAFRG tasks, most of them were not able to provide realistically generated facial reaction sequences but only focussed on generating facial attributes of the predicted facial reactions. Hence, this edition aims to promote the submission of results that include realistic facial reaction video clips. To assist in addressing the challenge of generating facial reaction video clips, this edition focuses specifically on video-conference settings and therefore includes only the NoXI [6] and RECOLA [15] datasets, due to the more noisy data of in-person settings (i.e., a major reason for excluding the UDIVA dataset [14] in this edition that was used in REACT 2023 challenge).

The REACT 2024 Challenge adopts the metrics defined in [21] to evaluate four aspects of the submitted models in terms of their generated facial reactions, namely: appropriateness, diversity, realism and synchrony. Participants are required to submit their developed model, checkpoints and well-explained source code, accompanied by a paper submitted to the REACT 2024 Challenge describing their proposed methodology and the achieved results. Only contributions that meet the pre-determined requirements, terms and conditions [1] are eligible for participation. The organisers do not engage in active participation themselves, but instead undertake a re-evaluation of the findings of the systems submitted to both sub-challenges. Differently from the previous edition, the ranking of the submitted models in this challenge depend on two metrics: Appropriate

∗ Equal Contribution; † Corresponding Author

[1] https://sites.google.com/cam.ac.uk/react2024/home

facial reaction correlation (FRCorr) of the generated facial reaction attributes and facial reaction realism (FRRea) of the generated facial reaction video clips, for both sub-challenges. In addition, participants should also report Facial reaction distance (FRDist), facial reaction diverseness (FRDiv), Facial reaction variance (FRVar), Diversity among facial reactions generated from different speaker behaviours (FRDvs) and Synchrony between generated facial reactions and speaker behaviours (FRSyn).

II. Challenge Corpora

The REACT 2024 challenge employs two video conference corpora: NoXi [6] and RECOLA [15]. Specifically, we first segmented each audio-video clip in two datasets into a 30-seconds long clip as in [1], [20]. Then, we cleaned the dataset by selecting only the dyadic interactions with complete data of both conversational partners (where both faces were within the frame of the camera). This resulted in 5919 clips of 30 seconds each (71,8 hours of audio-video clips), specifically: 5870 clips (49 hours) from the NoXi dataset and 54 clips (0,4 hour) from the RECOLA dataset. We divided the datasets into training (1,585 video clips from NoXI and 9 video clips from RECOLA), test (797 video clips from NOXI and 9 video clips from RECOLA) and validation (553 from NOXI and 9 from RECOLA) sets. We split the datasets with a subject-independent strategy (i.e., the same subject was never included in the training/validation and test sets). In this challenge, 25 frame-level facial attributes are provided for each facial frame, namely 15 AUs' occurrence (AU1, AU2, AU4, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU23, AU24, AU25 and AU26) predicted by the state-of-the-art GraphAU model [11], [19], as well as 2 facial affects (i.e., valence and arousal intensities) and 8 facial expression probabilities (i.e., Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger and Contempt) predicted by [22].

III. Evaluation Metrics

In this challenge, the submitted models are expected to generate two types of outputs for representing each facial reaction: (i) 25 facial attribute time-series; and (ii) a 2D facial image sequence. We followed [21], [20] to comprehensively evaluate three aspects of the generated facial reaction attributes: (i) **Appropriateness** based on two metrics, **FRCorr**: Concordance Correlation Coefficient (CCC) and **FRDist**: Dynamic Time Warping (DTW); (ii) **Diversity**: **FRVar**, **FRDiv**, and **FRDvs**; and (iii) **Synchrony**: the Time Lagged Cross Correlation (TLCC), called **FRSyn** in this challenge. Also, the **Realism** of the generated facial reaction video clips is assessed using the Fréchet Inception Distance (FID), denoted as **FRRea**.

IV. Baseline systems

Trans-VAE: We re-employ the same Trans-VAE baseline used in previous challenge [20] to this challenge. This baseline is inspired by [12], which follows the similar architecture as the TEACH [2]. As shown in Fig. 1, it is made up

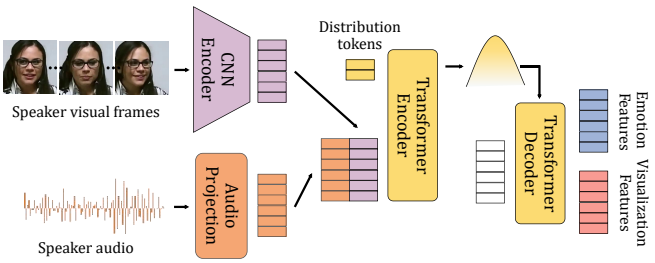

Fig. 1. Overview of the Trans-VAE baseline.

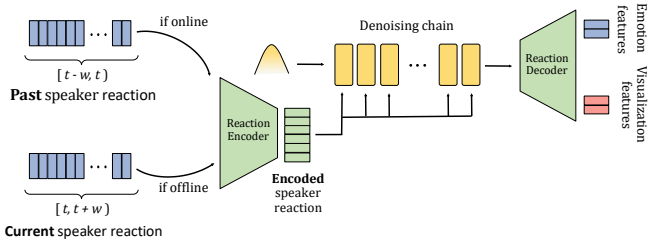

Fig. 2. Overview of the BeLFusion baseline.

of (i) a **CNN encoder** that extract facial reaction-related features from the input speaker facial image sequence; (ii) a **transformer encoder** that combines the learned facial embeddings and baseline audio embeddings (78-dimnesional MFCC features) extracted from the speaker audio behaviours using Torchaudio library [26], based on which an Gaussian Distribution is predicted to describe multiple appropriate facial reactions of the input speaker behaviour; and (iii) a **transformer decoder** that samples two types of facial reaction representations from the predicted distribution: 1) a set of 3D Morphable Model (3DMM) coefficients (i.e., 52 facial expression coefficients, 3 pose coefficients and 3 translation coefficients defined by [23]) and 2) an multi-channel facial attribute time-series (i.e., 25-channel time-series including 15 frame-level AUs' occurrence, 8 frame-level facial expression probabilities as well as frame-level valence and arousal intensities). Please refer to [20] for the detailed description of applying this baseline for online and offline MAFRG tasks.

BeLFusion. We also re-use BeLFusion without be-

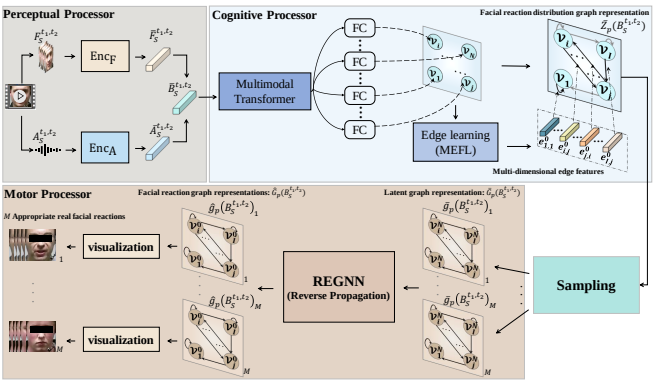

Fig. 3. Overview of the REGNN baseline [25].

havioural disentanglement as our second baseline [3], see Fig. 2. Its training involves two steps. First, a variational autoencoder (VAE) learns a lower representation of the sequence of visual features (e.g., AUs, facial affects, and expressions) for w frames. A regressor is incorporated after the VAE's decoder to transform the decoded reaction into a sequence of 3DMM coefficients. Second, a latent diffusion model (LDM) is optimized to, when given the speaker's reaction, predict the lower-dimensional representation of the listener's appropriate facial reaction. BeLFusion employs a window-based approach where the T/w reactions are predicted independently. Afterwards, the w-frames-long T/w reactions are arranged to construct the complete reaction. As in [20], the listener's visual features for the window $[t, t+w)$ is conditioned on the past speaker's features at $[t-w, t)$. Features for the segment $[0, w)$ are all set to zeroes. In the offline subchallenge, the generation is conditioned on the speaker's features within the same time period: $[t, t+w)$. The LDM's loss is the average of the latent and reconstructed MSE losses, and the denoising chain length is set to 10 steps.

Reversible Graph Neural Network (REGNN): We also employ the REGNN-based MAFRG approach [25] as the second baseline. As illustrated in Fig. 3, it consists of three main modules: (i) a **Perceptual Processor** that encodes the input speaker audio-facial behaviour as a pair of latent audio and facial representations; (ii) a **Cognitive Processor (Cog**) that predicts a Gaussian Mixture Graph Distribution describing all appropriate facial reactions in response to the input speaker behaviour; and (iii) an Reversible GNN-based **Motor Processor** that samples an appropriate facial reaction from the predicted appropriate facial reaction distribution. During the training, the Reversible GNN employed in this approach encodes all appropriate facial reactions of each input speaker behaviour as an ground-truth appropriate facial reaction distribution, enforcing the cognitive processor to predict the same distribution from the speaker representations obtained by the perceptual processor. As a result, the *one-to-many mapping* training problem is re-formulated as a *one-to-one mapping* problem. Please refer to [25] for more implementation details of its offline MAFRG system.

V. Baseline Results

It is clear that all three baselines outperformed the B_Random, B_Mime, B_MeanSeq and B_MeanFr, suggesting that they can predict meaningful appropriate human facial reactions from different speaker behaviours despite that the predicted facial reactions performances are not very solid.

Trans-VAE baseline: The Trans-VAE baseline model serves as a fundamental baseline for comparison purposes. This baseline demonstrates the capability to generate facial reactions that exhibit a modest level of diversity, as measured by metrics such as FRDiv, FRVar, and FRDvs, alongside a moderate level of appropriateness and comparable synchronization in both offline and online scenarios. In contrast to random facial reactions (B_Random), the Trans-VAE model achieved higher appropriateness (FRD). Furthermore, it surpasses in generating more diverse samples (FRDiv) compared to replicating facial sequences mirroring the speaker's facial behaviour. We visualise example facial reactions generated by this baseline in Fig. 4.

BeLFusion baseline: While BeLFusion shows a performance similar to Trans-VAE in terms of accuracy (FRD), it generates more diverse reactions. The competitive performance of such baseline without access to raw audio or video data highlights the need for better multimodal approaches tailored for this application. We also observe that binarizing the action units predicted greatly improves the diversity, but penalizes the accuracy and synchrony. The similar results in both online and offline scenarios suggest that a window-based approach might be insufficient to exploit all the information available in the visual features.

Reversible Graph Neural Network (REGNN) baseline: In the offline task evaluation, REGNN demonstrates clear advantages over Trans-VAE and BeLFusion baselines in terms of the appropriateness metrics, as indicated by the highest FRCorr and lowest FRD. In addition, the facial reactions generated by REGNN are more synchronised with the speaker behaviour. When it comes to the diversity, REGNN is less effective in generating diverse facial reactions, compared to the BeLFusion baseline.

VI. Participation and Conclusion

This paper introduced REACT 2024 Challenge in conjunction with the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2024, which focuses on multiple appropriate facial reaction generation under various video conference-based dyadic interactions scenarios. A total of 13 teams from 6 countries registered for this challenge, with 12 teams participating in the offline MAFRG sub-challenge and 13 teams participating in the online MAFRG sub-challenge. Our evaluation protocol strictly will rank all participant models under the same settings by comprehensively considering two aspects of their generated facial reactions: appropriateness, diversity, realism and synchrony. We hope that both the challenge data and code, as well as the systems and results of the competing teams, will serve as a valuable stepping stone for researchers and practitioners interested in the area of generative AI and automatic facial reaction generation. Our future efforts will be directed at continuing to organize REACT challenges in conjunction with well-known conferences while introducing new datasets and new modalities.

Acknowledgements

M. Spitale is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program. H. Gunes is supported by the EPSRC/UKRI under grant ref. EP/R030782/1 (ARoEQ).

References

[1] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

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

TABLE I

BASELINE OFFLINE AND ONLINE FACIAL REACTION GENERATION RESULTS ACHIEVED ON THE TEST SET.

Method	Appropriateness		Diversity			Realism	Synchrony
	FRCorr (↑)	**FRDist** (↓)	**FRDiv** (↑)	**FRVar** (↑)	**FRDvs** (↑)	**FRRea** (↓)	**FRSyn** (↓)
GT	8.73	0.00	0.0000	0.0724	0.2483	53.96	47.69
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
B_MeanSeq	0.01	97.13	0.0000	0.0000	0.0000	-	45.28
B_MeanFr	0.00	97.86	0.0000	0.0000	0.0000	-	49.00
Offline Results							
Trans-VAE	0.03	92.81	0.0008	0.0002	0.0006	67.74	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
Online Results							
Trans-VAE	0.07	90.31	0.0064	0.0012	0.0009	69.19	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

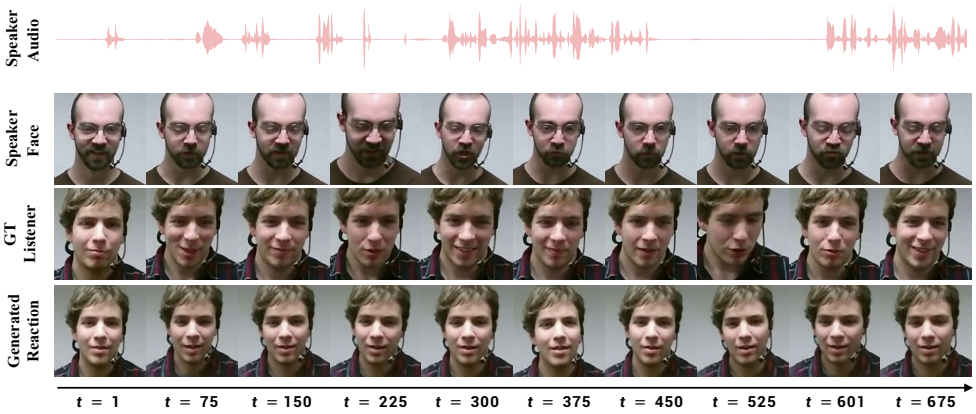

Fig. 4. Examples of generated listener reactions to a given speaker behaviour (including the speaker's audio and face frames). These reactions are generated by an offline Trans-VAE model.

[3] 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.

[4] G. Barquero, J. Núnez, S. Escalera, Z. Xu, W.-W. Tu, I. Guyon, and C. Palmero. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 139–178. PMLR, 2022.

[5] G. Barquero, J. Núñez, Z. Xu, S. Escalera, W.-W. Tu, I. Guyon, and C. Palmero. Comparison of spatio-temporal models for human motion and pose forecasting in face-to-face interaction scenarios. In C. Palmero, J. C. S. Jacques Junior, A. Clapés, I. Guyon, W.-W. Tu, T. B. Moeslund, and S. Escalera, editors, *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research*, pages 107–138. PMLR, 16 Oct 2022.

[6] 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.

[7] Y. et al. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. *arXiv preprint arXiv:2208.10441*, 2022.

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

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

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

[11] 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 *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1239–1246, 2022.

[12] 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.

[13] E. Ng, H. Joo, L. Hu, H. Li, T. Darrell, A. Kanazawa, and S. Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022.

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

[15] 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.

[16] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes. Personality recognition by modelling person-specific cognitive processes using graph representation. In *proceedings of the 29th ACM international conference on multimedia*, pages 357–366, 2021.

[17] H. Song, W.-N. Zhang, Y. Cui, D. Wang, and T. Liu. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188*, 2019.

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

[19] 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 preprint arXiv:2211.12482*, 2022.

[20] 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.

[21] 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 e-prints*, pages arXiv–2302, 2023.

[22] 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.

[23] 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.

[24] J. Xu, C. Luo, W. Xie, L. Shen, X. Liu, L. Liu, H. Gunes, and S. Song. Mrecgen: Multimodal appropriate reaction generator. *arXiv preprint arXiv:2307.02609*, 2023.

[25] 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.

[26] Y.-Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, et al. Torchaudio: Building blocks for audio and speech processing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6982–6986. IEEE, 2022.

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

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