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

Vector Quantized Diffusion Models for Multiple Appropriate Reactions Generation

Minh-Duc Nguyen, Hyung-Jeong Yang*, Ngoc-Huynh Ho, Soo-Hyung Kim, Seungwon Kim, Ji-eun Shin
* Corresponding author: hjyang@jnu.ac.kr

Chonnam National University, South Korea

***Abstract*— In the realm of dyadic interactions, the ability to generate appropriate facial reactions is paramount for the conveyance of empathy and understanding. This paper introduces a novel framework that leverages the strengths of a diffusion model architecture, underpinned by a vector quantized variational autoencoder (VQ-VAE) to synthesize facial reactions that are contextually apt. We rigorously evaluate our model on the IEEE FG REACT2024 dataset, where it demonstrates superior performance, outshining baseline methods in terms of effectiveness. The results underscore the potential of our framework to enhance the fidelity of digital human interactions, paving the way for more nuanced and emotionally intelligent systems.**

I. INTRODUCTION

Dyadic interactions, the fundamental units of human communication, involve two individuals engaged in a reciprocal exchange. These interactions can be richly informative, encompassing both verbal dialogue and a myriad of non-verbal cues such as facial expressions, gestures, and body postures. The study of dyadic interactions has garnered increasing attention, with research aiming to decode and replicate the nuanced dynamics of these exchanges. Previous studies [5, 6, 12, 14, 21] focused on generating reactions that mirror a ground-truth response, often through deterministic models that replicate exact reactions.

The challenge lies in capturing the spontaneity and diversity of human responses, which are not always predictable or uniform. As such, the field is moving towards models that not only reproduce but also anticipate the complex array of potential reactions, contributing to more realistic and empathetic human-computer interaction. The REACT2023 challenge [17] delved into this evolving landscape, focused on generating human-compatible facial responses across a range of dyadic interaction situations, ensuring that all participants are evaluated under identical conditions. The challenge proposed two sub-challenges: Offline Multiple Appropriate Facial Reaction Generation and Online Multiple Appropriate Facial Reaction Generation, respectively.

As highlighted in the referenced study [15], participants utilize a combination of verbal and non-verbal cues in dyadic communications. This interaction allows for a broad spectrum of potential responses that are considered appropriate for a listener to produce, contingent upon their psychological disposition. For investigating the one-to-many appropriate reaction mappings, Evonne Ng et al. [9] have been conducted to synthesize multiple listener facial motions with the given speaker's behavior with a novel encoding VQ-VAE. UniFaRN [7] utilizes the Transformer architecture to address challenges, capitalizing on its versatility in processing multimodal data and its capacity to steer the generation process. BEAMER [4] worked on a Transformer-VAE architecture, they proposed a behavioral encoder that takes as input a speaker's behavior and texture and later encodes it into a latent space of an appropriate listener. A recent approach by Jun et al. [20] worked on the latent diffusion model with modification to enhance the competency of modeling the context. The inherent property of stochasticity in the diffusion model enables their model to generate multiple reactions.

This paper presents a vector quantized diffusion model (VQ-Diff) to tackle the REACT2024 challenge [16], which is based on recent work by Barquero et al. [1]. We employed a Vector Quantized Variational autoencoder model (VQ-VAE) [18] that learns listener reactions from a discrete latent space and leveraged Latent Diffusion Model [11] with DDIM sampler to predict the lower-dimensional representation of the listener's appropriate facial reaction from the speaker's reaction as input.

To summarize, our contribution includes:

- We utilize the Latent Diffusion Model (LDM) to generate appropriate facial reaction of the listener. Compared to REACT2024 baseline methods, our approach archives significant results in terms of diversity metrics while maintaining decent effectiveness in terms of appropriateness evaluation.
- We employed a VQ-VAE as an auxiliary model to learn listener reaction features (e.g., AUs, facial affects, and expressions) from a discrete latent space instead of continuous distribution by the vanilla VAE model.

II. METHOD

We proposed the vector quantized diffusion (VQ-Diff) model for Multiple Appropriate Facial reaction generation. Our approach utilizes a vector quantized variation autoencoder (VQ-VAE), whose latent space is modeled by a Latent Diffusion Model (LDM). In this part, we outline the problem statement and related concepts followed by a concise overview of our model.

979-8-3503-9494-8/24/$31.00

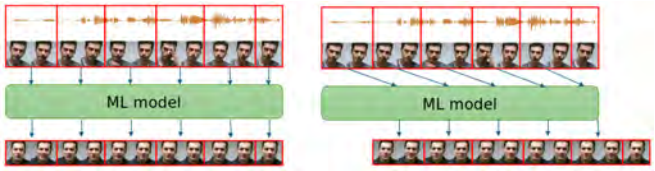

Fig. 1. Offline MAFRG (left) aims to gradually generate facial reactions and Online MAFRG (right) is the same as offline, but speaker behavior must be shifted by the window size to only use past information.

A. Problem Statement

The REACT2024 challenge introduced two separate sub-challenges designed for online and offline appropriate facial reaction generation. Assume T is the time size and w is the size of the segmented window. The task Offline Multiple Appropriate Facial Reaction Generation (MAFRG) by the Belfusion baseline [1] aims to gradually generate facial reaction frames by a window-based approach where the T/w reactions are predicted independently. Subsequently, the reactions spanning w-frames, denoted as T/w, are sequenced to form the entire reaction sequence. Following the methodology outlined in the referenced study [3] for the online sub-challenge (online MAFRG), the visual features of the listener within the interval $[t, t+w]$ are predicted on the preceding features of the speaker from the interval $[t-w, t]$. For the initial segment $[0, w]$, all features are initialized to zero. In the context of the offline sub-challenge (offline MAFRG), the reaction generation is contingent upon the speaker's features during the concurrent interval $[t, t+w]$. As shown in Fig. 1, our approach aligns with the Belfusion baseline, offline MAFRG tends to predict the listener's reaction to the current speaker's behavior while online MAFRG takes past speaker's behavior to generate future listener's reaction.

B. Model architechture

Our approach method includes two stages which are trained separately. In the first stage, we train a VQ-VAE with two LSTM [3] based layers at the encoder and encoder to learn a codebook with discrete listener's reaction embeddings space, as shown in Fig. 2. The model aims to learn a lower representation of the listener's visual features (e.g., AUs, facial affects, and expressions) of w frames. For the prior distribution $p(z)$, we define a vector quantized layer following[18], we choose dimensions D=128 and K=200 for the size of discrete latent space. The encoder takes the input and gives the latent embedded variables z_e which are calculated as the nearest neighbor look-up based on L_2 distance to select z_q from the codebook. Afterward, the decoder consumes z_q and recreates the lower representation of the listener's visual features. We also incorporate a regressor after the decoder to convert the decoded reaction to a sequence of 3D Morphable Model (3DMM) parameters. VQ-VAE loss is composed of three components: reconstruction loss which optimizes the encoder and decoder; codebook loss to bypass the embedding as the codebook learning by L_2 error; and commitment loss to make sure the encoder commits to an embedding. The total loss is defined as follows:

$$L = \log p\left(x \mid z_q(x)\right) + \|\mathrm{sg}\left[z_e(x)\right] - e\|_2^2 + \beta \left\|z_e(x) - \mathrm{sg}[e]\right\|_2^2 \quad (1)$$

Where sg represents the stop gradient operator meaning no gradient, β denotes the commitment loss hyper-parameter equal to 0.25 as mentioned in the paper.

In the second stage, a latent diffusion model is addressed to predict the corresponding embedded variables z_0 from given the lower representation of the speaker's visual features. We first reuse the encoder from the VQ-VAE model to transform the listener's reaction to z_0 and the speaker's reactions to condition c for LDM. The process involves the initial latent variable z_0 undergoing a forward Markov chain progression over M steps, resulting in z_M, which approximates a normal distribution $\mathcal{N}(0, 1)$. During the model's training, the network is tasked with estimating z_{t-1} from z_t, the timestep t, and the given context c, effectively reversing the Markov chain using the DDIM sampler across the designated M=10 steps (Fig. 3). Once we have predicted latent variables z_0 at the inference process, listener's reaction and predicted sequence of 3DMM coefficients can be reconstructed by using the decoder from the first stage (Fig. 4) which can first retrieve the closest neighbor for variable z_q. The loss function is calculated as the mean of Mean Square Error (MSE) in the latent space and MSE in the reconstructed space.

III. EXPERIMENTS

A. Datasets

To evaluate our proposed model, we conduct experiments on the REACT2024 dataset, which employs two video conference corpora: NoXi [2] and RECOLA [10]. It consists of 5910 clips of 30 seconds each, about 71,8 hours of dyadic videos in total. The dataset also provided a comprehensive set of 25 facial attributes for each frame. These include the occurrences of 15 Action Units (AUs) - specifically AU1, AU2, AU4, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU23, AU24, AU25, and AU26. These AUs are forecasted using the advanced GraphAU model [13] and [8]. Additionally, two measures of facial affect valence and arousal intensities are supplied, along with the probabilities of eight distinct facial expressions: Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, and Contempt.

B. Evaluation Metrics

We adopt the evaluation methodology provided by [15]. There are four dimensions: appropriateness, diversity, synchrony, and realism with several metrics. Appropriateness is gauged by comparing the similarity between the generated facial reactions and the actual reactions observed. This involves two metrics: FRDist (Dynamic Time Warping) and FRCorr (Concordance Correlation Coefficient). Diversity measures the variation in reactions both within a single frame and across multiple frames. To quantify diversity, FRVar,

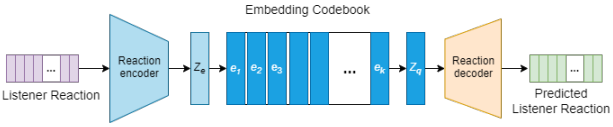

Fig. 2. We leverage a VQ-VAE model to reconstruct the lower representation of listener facial features. The model employs vector quantization to produce discrete latent representations by a codebook.

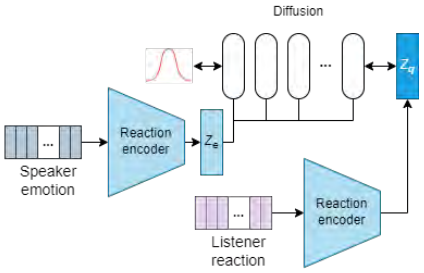

Fig. 3. Latent diffusion model training process.

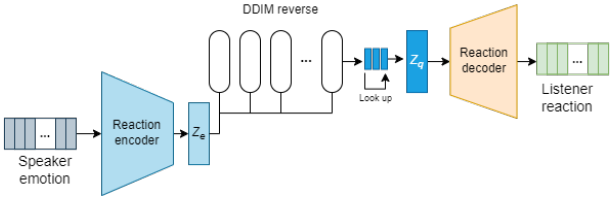

Fig. 4. Latent diffusion inference process.

FRDiv, and FRDvs are computed. Moreover, FRSyn (Time Lagged Cross Correlation) for synchrony examines how well the generated facial expressions align with the speaker's behavior. Additionally, the authenticity of the created facial reaction videos is evaluated using the Frechet Inception Distance (FID), referred to as FRRea, which gauges the realism of the videos by comparing them to genuine facial reactions. These metrics collectively ensure that the generated reactions are not only contextually appropriate but also exhibit a natural variety and are in sync with the speaker's behavior, contributing to the overall realism of the interaction.

C. Implementation Details

All the experiments are conducted in PyTorch using a single RTX 8000 GPU. The training process for our model involves two distinct phases. Firstly, the VQ-VAE is trained by 1000 epochs with batch size 32, the window size is set to 50, the learning rate is 1e-3 and the weight decay by 5e-4. We adjust the same optimizer parameters in the second stage and train the LDM for 200 epochs. The LDM finally leverages the learnt decoder from pre-trained VQ-VAE which converts predicted listener latent variables to the final listener reaction features and a sequence of listener 3DMM coefficients. We trained two LDM models which are different in the inference process, we omit the vector quantized layer in the decoder for the second model for experiments.

D. Results

Our method is evaluated against generative approaches that are Transformer-based Variational autoencoder (Trans-VAE), Belfusion, and Reversible Graph Neural Network (REGNN) [19]. The experimental results are conducted on the REACT2024 test set as shown in Table I and Table II. Additionally, we consider both non-binarized and binarized Action unit features (AUs) from the generated output in our evaluations.

The results obtained by both baseline models show the ability to generate facial reactions that align well with the actual listener's reaction observed. Especially in the offline task evaluation, REGNN indicates the lowest FRDist and is more synchronized with speaker behavior. In terms of appropriateness evaluation, our VQ-Diff may be less effective at quantifying the nearest appropriate facial reaction by FRDist metric, but it demonstrates a better resemblance to the ground truth listener's reaction by getting a much higher FRC than all other baselines. In sharp contrast, our approach surpasses all previous methods in three key diversity measures: FRDiv, FRVar, and FRDvs. When it comes to synchrony, our VQ-Diff non-binarize AUs version archives a decent good TLCC value among other models.

Our ablation study lies in the comparison between with and without the codebook layer in the listener's reaction decoder. The model without a vector quantized layer can lead to more arbitrary reactions that break down the FRDist score trading off with incredibly elevated diversity evaluations. However, the model beats by a fair margin most baselines in terms of FRCorr. This is an interesting insight that would benefit from further exploration in the future.

IV. CONCLUSION

In this work, we present the VQ-Diff model, which leverages both discrete learning and the denoise diffusion model to address the Multiple Appropriate Facial Reaction Generation challenges. Particularly, the model consists of a VQ-VAE model which focuses on reconstructing the listener's reactions, and a Latent Diffusion model to learn the lower representation of the listener's reaction to a given speaker's behaviors. The listener's outcome can be predicted by the auxiliary VQ-VAE decoder afterward. Our approach enables significantly the generation of diverse responses

TABLE I
OFFLINE FACIAL REACTION GENERATION RESULTS ACHIEVED ON THE TEST SET

Method	Appropriateness		Diversity			Synchrony
	FRCorr (↑)	FRDist (↓)	FRDiv (↑)	FRVar (↑)	FRDvs (↑)	FRSyn (↓)
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	**0.30**	91.65	0.0743	0.0348	0.0745	45.33
Ours + Binarized AUs	0.29	96.53	0.1455	0.0682	0.1461	49.00
Ours (w/o codebook)	0.30	115.02	0.3712	0.1738	0.3730	45.58
Ours (w/o codebook) + Binarized AUs	0.29	119.30	**0.4907**	**0.2298**	**0.4922**	47.19

Note: k is the number of denoise steps.

TABLE II
ONLINE FACIAL REACTION GENERATION RESULTS ACHIEVED ON THE TEST SET

Method	Appropriateness		Diversity			Synchrony
	FRCorr (↑)	FRDist (↓)	FRDiv (↑)	FRVar (↑)	FRDvs (↑)	FRSyn (↓)
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	**0.30**	91.86	0.0737	0.03463	0.07434	45.18
Ours + Binarized AUs	0.29	96.83	0.1486	0.0699	0.1499	49.00
Ours (w/o codebook)	0.30	115.72	0.3800	0.1778	0.3809	45.37
Ours (w/o codebook)+ Binarized AUs	0.29	118.62	**0.4719**	**0.2208**	**0.4728**	48.66

while preserving the appropriateness and synchrony of generated reactions within dyadic exchanges for the REACT2024 MAFRG challenge.

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

[2] Angelo Cafaro et al. "The NoXi database: multimodal recordings of mediated novice-expert interactions". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, pp. 350–359.

[3] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[4] Ximi Hoque et al. "BEAMER: Behavioral Encoder to Generate Multiple Appropriate Facial Reactions". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 9536–9540.

[5] Yuchi Huang and Saad Khan. "A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions". In: *Proceedings of*

the 20th ACM international conference on multimodal interaction. 2018, pp. 437–445.

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

[7] Cong Liang et al. "Unifarn: Unified transformer for facial reaction generation". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 9506–9510.

[8] Cheng Luo et al. "Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition". In: *arXiv preprint arXiv:2205.01782* (2022).

[9] Evonne Ng et al. "Learning to listen: Modeling non-deterministic dyadic facial motion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20395–20405.

[10] Fabien Ringeval et al. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, pp. 1–8.

[11] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[12] Zilong Shao et al. "Personality recognition by modelling person-specific cognitive processes using graph representation". In: *proceedings of the 29th ACM international conference on multimedia*. 2021, pp. 357–366.

[13] Siyang Song et al. "Gratis: Deep learning graph representation with task-specific topology and multi-dimensional edge features". In: *arXiv preprint arXiv:2211.12482* (2022).

[14] Siyang Song et al. "Learning person-specific cognition from facial reactions for automatic personality recognition". In: *IEEE Transactions on Affective Computing* (2022).

[15] Siyang Song et al. "Multiple Appropriate Facial Reaction Generation in Dyadic Interaction Settings: What, Why and How?" In: *arXiv preprint arXiv:2302.06514* (2023).

[16] Siyang Song et al. "REACT 2024: the Second Multiple Appropriate Facial Reaction Generation Challenge". In: *arXiv preprint arXiv:2401.05166* (2024).

[17] Siyang Song et al. "React2023: The first multiple appropriate facial reaction generation challenge". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 9620–9624.

[18] Aaron Van Den Oord, Oriol Vinyals, et al. "Neural discrete representation learning". In: *Advances in neural information processing systems* 30 (2017).

[19] Tong Xu et al. "Reversible graph neural network-based reaction distribution learning for multiple appropriate facial reactions generation". In: *arXiv preprint arXiv:2305.15270* (2023).

[20] Jun Yu et al. "Leveraging the latent diffusion models for offline facial multiple appropriate reactions generation". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 9561–9565.

[21] Mohan Zhou et al. "Responsive listening head generation: a benchmark dataset and baseline". In: *European Conference on Computer Vision*. Springer. 2022, pp. 124–142.