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

# Finite Scalar Quantization as Facial Tokenizer for Dyadic Reaction Generation

Quang Tien Dam[1], Tri Tung Nguyen Nguyen[1], Dinh Tuan Tran[2], and Joo-Ho Lee[2]
[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan
[2] College of Information Science and Engineering, Ritsumeikan University, Japan

*Abstract*— **Creating a human-like interface in human-robot interaction is a formidable challenge. Many efforts have been made to mimic the human ability of attentive listening and synchronous participation in conversations, especially in terms of facial expressions and head movements. By taking advantage of transformer-based sequence generation models and quantization techniques, this advantage is further enhanced in the areas of text, video, and audio generation. Using Finite Scalar Quantization, we develop a facial expression tokenization module that is able to encode facial expressions in a finite, semantically meaningful vocabulary. Using this module, we establish a more powerful cross-modality transformer-based, non-deterministic model that is able to learn multiple appropriate facial responses in a dyadic conversational context.** [1]

## I. Introduction

In human face-to-face conversations, two distinct roles emerge: speakers who articulate their thoughts, and listeners who respond with facial expressions and synchronized head movements, influenced by individual characteristics. The REACT competition [15] has been proposed to address this problem, called the Multiple Appropriate Facial Reaction Generation problem (MARG), and provides a set of evaluation metrics that this work takes advantage of. Specifically, the challenge requires generating multiple outputs in response to the same stimuli from the other agent, and then evaluating the appropriateness, diversity, synchrony, and realism of the generated output. This work won the online track of REACT2024 [14].

Our work localizes into the online MARG as in REACT [15]. Noteworthy recent approaches [10], [14] decompose the main task into two sub-tasks: Facial Motion Representation Learning and Appropriate Facial Reaction Prediction. In Facial Motion Representation Learning, the VAE family plays the pivotal role in grouping similar motions by employing either non-deterministic methods (e.g., VAE [13], Normalizing Flow [6]) or deterministic methods (e.g., VQ-VAE [10], Codebook tokenization [8], etc.). While VAE can model encoded features as normal distributions, search-based vector quantization methods such as VQ-VAE are gaining popularity by providing stable training, better control with discrete latent space, and training efficiency. For generative predictors, the mainstream approach relies on either Transformer [10], [8], [7], Diffusion [13], [3], or Normalizing Flow-Recurrent Neural Net [6]. The generative models with a multinomial sample step can introduce a non-deterministic aspect to the generated motion aligning with the vector-quantization-based tokenizer to achieve desirable balance appropriateness, diversity, and context synchrony [18], [10]. However, the trickiness in VQ-VAE's codebook size definition remains a technical issue varying case by case: larger sizes result in highly sparse codebooks, while smaller ones yield fuzzy and overlapping embeddings between distinct motions, and the problem of codebook collapse when only a small proportion of the codebook is used. Furthermore, generated motions guided by facial features (e.g., 3DMM, AUs) though temporally reasonable, lack clarity between facial components and convey little meaningful intent.

Inspired by the current progress of generative and multi-modality models, this paper aims to contribute to the existing literature in the following aspects: Firstly, we experiment with a method for generating actions based on tokenized facial states in a single-frame manner, utilizing state-of-the-art quantization methods. Secondly, we enhance the attention modules by integrating different modalities, enabling the listener transformer autoregressive predictor to comprehend dyadic conversations over extended contextual spans, thereby generating multiple non-deterministic appropriate facial reactions.

## II. Methodology

We model a sequence-generating problem to generate multiple appropriate facial reactions, in which every state of the face is accordant with a facial token. The generator comprises two parts: (1) face tokenizer and (2) reaction predictor. The face tokenizer is described in II-B and the predictor in II-C.

### A. Problem definitions

To generate meaningful facial reactions, one will digest the content of the conversation, personality, and expressions of the counterpart. Hence, the input for the model as the listener is the speaker's facial status and speech, then output the appropriate facial states. We propose a two-stage learning system. A non-autoregressive VQ-transformer tokenizer $\mathscr{T}$ that learns to discretize facial states into finite states as *tokens*. So that the autoregressive cross-modality transformer model $\mathscr{P}$ can learn the relationship between the speaker's facial expression and voice, and the past listener's expression to predict a multinomial probability of the future face that allows sampling non-deterministically.

This work was not supported by any organization.

[1] Code: https://github.com/ais-lab/FaceAIS_REACT24

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

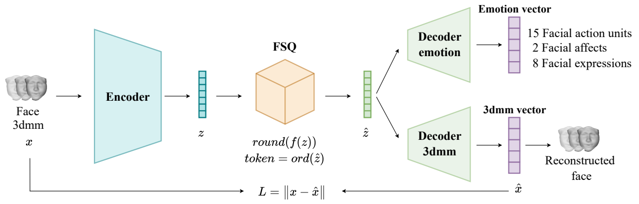

Fig. 1. **Face tokenizer.** To encode a facial image as a token, the proposed tokenizer is designed to learn face-to-face mapping via a self-supervised reconstruction task. Employing Finite Scalar Quantization, this tokenizer eliminates the need for a codebook learning phase and enhances codebook efficiency by utilizing a rounding function $f(z)$ and facilitating token conversion through a bijection function. After training the encoder and decoder with 3DMM, they will be frozen, and the emotion decoder will be trained to generate the corresponding emotion vector.

Let $F_n = \{f_1, f_2, \dots, f_n\}$ denote the sequence of facial states, where $f_i \in \mathbb{R}^{d_f}$ represents the 3D Morphable Face Models (3DMM) vector of the face at timestep $i$. Similarly, $E_n = \{e_1, e_2, \dots, e_n\}$ represents the sequence conveying facial descriptors, with $e_i \in \mathbb{R}^{d_e}$ encapsulating 15 facial action units, facial affect, and categorical facial expressions at timestep $i$. Additionally, $S_n = \{s_1, s_2, \dots, s_k\}$ signifies the sequence of sound, where $s_i \in \mathbb{R}^{d_s}$ represents the sound feature vector, noting that this sequence may not align in length with the facial and emotional sequences. The system accepts the facial sequence $F^{sp}_{i-w:i}$ of the speaker and the corresponding sound feature vector sequence $S^{sp}_{i-w:i}$ to generate the facial sequence $F^{li}_{i+1:i+p}$ and the emotion sequence $E^{li}_{i+1:i+p}$ for the listener, where $i$ denotes the current timestamp, $w \in \mathbb{N}^*$ represents the context window of the system, and $p \in \mathbb{N}$ is the number of future frames that the system predicts at once. The predicted facial sequence then will be fed into a face renderer model to generate a human-like face.

### B. Facial tokenizer

The Vector Quantization method, VQ-VAE [16], in conjunction with the autoregressive model, has been instrumental in the development of various robust image, audio, and video generation models. However, the original VQ-VAE exhibits diminished effectiveness when quantized to codebook sizes exceeding $2^{10}$ [9]. To address this limitation, we propose leveraging a more contemporary quantization approach that is both smaller in scale and more adept at handling larger codebooks. This upgrade is motivated by two primary considerations: Firstly, a broader vocabulary enables the predictor to articulate more nuanced expressions, and secondly, an optimized utilization of quantization space enhances its ability to capture facial features accurately. Consequently, we empirically opt for Finite Scalar Quantization [9] to quantize facial feature.

The tokenizer architecture comprises three primary components as illustrated in Fig. 1: a transformer encoder $E$, a quantizer $Q$, and two transformer decoders $D_{3dmm}$ and $D_{emotion}$. The encoder $E$ transforms the features of the input facial sequence $x$ into a $d$-dimensional latent vector $z \in \mathbb{R}^d$:

$$z = E(x). \tag{1}$$

Subsequently, the quantizer $Q$ maps $z$ to $\hat{z}$, which belongs to a finite set of vectors $C$. The size of $C$ is the vocabulary size of the predictor. This quantization is achieved through a bounding function $f$ that selects $\hat{z}$ from $L$ unique values, with each entry computed as:

$$\hat{z}_i = f(z_i) := \text{round}\left(\left\lfloor \frac{L}{2} \right\rfloor \tanh(z_i)\right) \in \{-1, 0, 1\}. \tag{2}$$

Each $\hat{z}$ vector is represented by a token which is the order of that vector in $C$. The order is computed by a bijection function similar to $token = ord(\hat{z})$.

Finally, the reconstruction is obtained via:

$$\begin{aligned} \hat{x}_{3dmm} &= D_{3dmm}(\hat{z}), \\ \hat{x}_{emotion} &= D_{emotion}(\hat{z}). \end{aligned} \tag{3}$$

As the quantization $Q$ operates heuristically, we solely train $E$ and $D$ using the reconstruction loss through 2 phases of training. To further customize quantization for facial expression representation, we propose the integration of a BlendShapeLoss function during the first phase of tokenizer training. This function leverages the Blend Shape feature utilized by Faceverse for rendering facial states [17]. Specifically, the facial expression in 3DMM vector $\beta_t \in \mathbb{R}^{d_m}$, where $d_m$ denotes the dimension of the coefficient, is segregated into brow coefficients, eye movement coefficients, mouth movement coefficients, face rotation, and translation coefficients. We then compute reconstruction loss separately for each group of coefficients to emphasize these features during tokenizer training:

$$\begin{aligned} L_{Blendshape}(E, D_{3dmm}) = &\left\|x_{eyebrow} - \hat{x}_{eyebrow}\right\| \\ &+ \left\|x_{eyemovement} - \hat{x}_{eyemovement}\right\| \\ &+ \dots \\ &+ \left\|x_{rotation} - \hat{x}_{rotation}\right\| \\ &+ \left\|x_{translation} - \hat{x}_{translation}\right\|. \end{aligned} \tag{4}$$

In the second phase of training the tokenizer, $E$ and $D_{3dmm}$ will be frozen, and only $D_{emotion}$ will be trained to generate the corresponding emotion vector of a face status using an L1 loss:

$$L(D_{emotion}) = \left\|x_{emotion} - \hat{x}_{emotion}\right\|. \tag{5}$$

Additionally, we refine the quantization technique to discretize distinct facial features more efficiently. Through empirical investigation, we have observed that augmenting the vocabulary size of the predictor improves output quality up to a certain threshold. Beyond this threshold, the capacity of the predictor reaches saturation. We formulate the VQ-VAE transformer to represent each facial state as a single token, aiming to provide the facial expression predictor with maximally detailed signals rather than action tokens. To the best of our knowledge, this is the first to explore facial quantization in a single-frame manner.

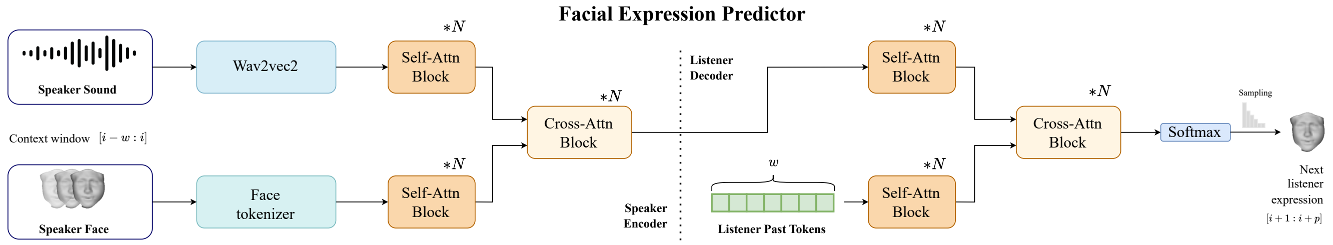

Fig. 2. **Predictor Diagram.** The predictor $\mathscr{P}$ utilizes Wav2Vec2 to extract speaker sound features, employs the face tokenizer $\mathscr{T}$ to quantize facial expressions, incorporates cross-attention to merge sound and facial features, and samples from the distribution to generate the next $p$ facial tokens. The face tokenizer then decodes the tokens according to their facial expressions and emotions.

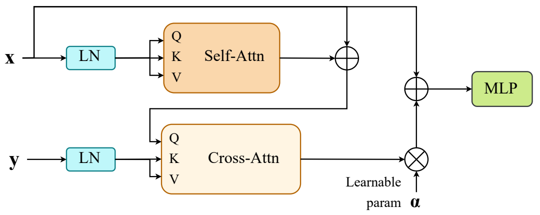

Fig. 3. **Cross-Modality Attention Block** receiving feature vectors from two modalities $x$ and $y$, employing a gated parameter $\alpha$ to regulate feature fusion.

### C. Crossmodal autoregressive facial reaction

The predictor $\mathscr{P}$ takes the tokenized sequence of facial expressions from the speaker, denoted as $F^{sp}$, and the speaker's sound features $S^{sp}$ to predict the subsequent facial expression as illustrated in Fig 2. The predictor is divided into two parts: the speaker encoder and the listener decoder. In the speaker encoder, Wav2Vec2 extracts sound features from raw audio, and the face tokenizer $\mathscr{T}$ turns the speaker facial sequence into face tokens. Then, sound features and face tokens, are processed by $n$ normal self-attention layers and pass through cross-attention blocks to create a *speaker feature* vector. In the listener decoder, the speaker feature and past listener feature will be similarly processed by $n$ self-attention layers and fed to $n$ cross-attention layers to make the output more consistent. In processing the sound signal, we learn that the VQ-Wav2Vec model [1] produces better results than directly using Mel-frequency cepstral coefficients (MFCC) [10], [14] as sound feature. Subsequently, we leverage the enhanced sound feature extraction capabilities of Wav2Vec2 [2] as it outperforms VQ-Wav2Vec, thus selecting it for sound feature extraction. Regarding the attention mechanism, we adopt the architecture of the FIBER attention block [5], which incorporates a learnable gated parameter to modulate the influence of features from different modalities, as illustrated in Fig. 3.

To enhance the model's usability, we adopt a patch-based approach wherein the model predicts the next $p$ tokens instead of a single token, referred to as the model's patch size [10]. Through experimentation, we ascertain that smaller patch sizes yield improved model performance, albeit at a linear increase in computational requirements. Consequently, we empirically set the patch size to 32 frames. To ensure consistent yet dynamic output during test time, we scale logits by a temperature $t$ and cut the top $k$ of the prediction then apply a multinomial sampling on the predicted distribution to create non-deterministic appropriate output. We employ teacher forcing and random masking past tokens during training to make the model more generalized across training data.

## III. Experiments and Results

To evaluate the efficacy of our proposed approach, we utilized the datasets from NoXi [4] and RECOLA [12], which contain internet conference data. We applied the validation techniques outlined in the React Competition [15] and compared our results with the baseline outcomes from the REACT 2024 competition [14].

**Implementation Details.** The tokenizer was implemented with a latent dimension $d_z = 252$, codebook size of 2048, and quantization levels $[8,5,5,5]$. The training process employed a two-phase strategy, optimizing the 3DMMs output and emotion output sequentially. The predictor architecture comprised a block size of 256, an embedding dimension of 256, and 8 layers of attention in each attention block. The tokenizer and predictor models were each trained for 500 epochs on an NVIDIA RTX 4060 Ti GPU.

### A. Metrics

We leverage the metrics proposed by [15], which encompass four categories: appropriateness, diversity, realism, and synchrony. During testing and optimization, the priorities are set as follows: appropriateness, synchrony, diversity, and lastly realism, to guide the design decisions as the resulting video exhibits qualitatively better stability and usability. FRRea is considered a metric that does not directly reflect the model's performance, as it only compares frames after a 30-frame interval and is strongly influenced by the photorealistic renderer, in this case, PIRender [11]. Therefore, this metric is not heavily weighted in comparison with other methods. Moreover, in contrast to the approach proposed in [14], the synchrony metric ought to be evaluated in proximity to the ground truth representation, as a scenario wherein the reaction signal exhibits perfect synchronization with the speaker would be deemed unnatural.

TABLE I

BASELINES. COMPARISON OF OUR APPROACH WITH BASELINE MODELS [14] ON THE TEST SET.

| | **Appropriateness** | | **Diversity** | | | **Realism** | **Synchrony** |
|---|---|---|---|---|---|---|---|
| | **FRCorr** (↑) | **FRDist** (↓) | **FRDiv** (↑) | **FRVar** (↑) | **FRDvs** (↑) | **FRRea** (↓) | **FRSyn** $(\cdot)$ |
| Ground truth | 8.73 | 0.00 | 0.0000 | 0.0724 | 0.2483 | - | 47.69 |
| Random | 0.05 | 237.23 | 0.1667 | 0.0833 | 0.1667 | - | 44.10 |
| Mime | 0.38 | 92.94 | 0.0000 | 0.0724 | 0.2483 | - | 38.54 |
| MeanSeq | 0.01 | 97.13 | 0.0000 | 0.0000 | 0.0000 | - | 45.28 |
| MeanFr | 0.00 | 97.86 | 0.0000 | 0.0000 | 0.0000 | - | 49.00 |
| Trans-VAE | 0.07 | 90.31 | 0.0064 | 0.0012 | 0.0009 | 69.19 | 44.65 |
| BeLFusion(k=10)+BinarizedAUs | 0.12 | 94.09 | 0.0379 | 0.0248 | 0.0397 | - | 49.00 |
| Ours | **0.31** | **84.93** | **0.1164** | **0.0348** | **0.1166** | **34.66** | **47.42** |

$(\cdot)$ means the closer to the ground truth, the better.

🟥 indicates the best average performance among the heuristic baselines for the groups of metrics.

TABLE II

VECTOR QUANTIZE ABLATION. USING THE SAME PREDICTOR, WE ALTERNATED THE QUANTIZATION MODULE AND COMPARED ITS PERFORMANCE ON THE VALIDATION SET.

| | Appropriateness | | Diversity | | | Realism | Synchrony |
|---|---|---|---|---|---|---|---|
| | FRC | FRD | FRDvs | FRVar | FRDiv | FRRea | FRSyn |
| FSQ-val | **0.2737** | **86.6145** | 0.1162 | 0.0345 | 0.1163 | 81.2801 | 45.7206 |
| LFQ-val | 0.2625 | 99.8672 | 0.1213 | 0.0434 | 0.1213 | 73.2092 | 45.8896 |
| VQ-val | 0.2693 | 91.5249 | 0.0943 | 0.0370 | 0.0943 | 96.1280 | 46.2099 |

### B. Baseline comparison

Table I demonstrates that our proposed methods outperform all model-based baselines in the test set. The random baseline produces high diversity but lacks appropriateness. The mime baseline mimics the speaker's reaction as the listener's reaction. Our method achieves similar diversity to the random baseline while maintaining an appropriate level comparable to the mime baseline. Regarding the model baselines, Trans-VAE generates results that lack dynamics, exhibiting low diversity and suboptimal appropriateness. BeLFusion [3] with $k = 10$ and Binarized AU perform slightly better than Trans-VAE but still fall short in terms of diversity and appropriateness. Furthermore, our model produces results that are highly temporally correlated with the baseline, as evidenced by its FRSyn value being closer to the ground truth. Our method achieves better diversity by leveraging the advantages of a generative model, sampling from a multinomial distribution. However, by applying a cutoff to this distribution, it maintains high appropriateness. The Wav2Vec2 feature extraction and fusion attention mechanism enable our model to synchronize effectively with the counterpart, even for subtle movements such as eye blinking and mouth movements (see supplementary video).

### C. Vector quantize ablation

We conducted an ablation study to identify the appropriate vector quantization method for facial reaction generation, as demonstrated in Table II. We evaluated traditional vector quantization (VQ), Finite Scalar Quantization (FSQ) [9], and Lookup-Free Quantization (LFQ) [19] with the same vocabulary size of 2048 and the same predictor. The results indicate that FSQ leverages a significantly larger vocabulary set compared to traditional vector quantization. Additionally, we applied several optimization techniques, such as k-means initialization and quantization decay for the VQ. In terms of implementation complexity, FSQ is simpler than LFQ, despite sharing the idea of eliminating the lookup step. It is noteworthy that the quality of the tokenizer strongly correlates with the predictor output, as the final output must be decoded by the tokenizer and carry the meaning of each token. Consequently, based on the ablation study, we selected FSQ due to its superior performance in terms of appropriateness.

## IV. CONCLUSION

In facial expression generation, employing a tokenization approach to discretize intricate natural data into a finite vocabulary set enhances the capability of sequence-generative transformer-based models to produce diverse and appropriate facial expressions. In our devised framework, the utilization of a single-frame tokenizer employing Finite Scalar Quantization creates a meaningful vocabulary set for the generative task. Integrating a cross-modality transformer model featuring the FIBER cross-attention block enables the generative model to efficiently incorporate both sound and facial expressions. Even though our method won the online MARG track at the REACT 2024 competition, there is still room for optimization, such as adding a controlling parameter to tune the intrinsic characteristics of the listener, making the model more usable in specific use cases.

# References

[1] A. Baevski, S. Schneider, and M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*, 2020.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[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 (ICCV)*, pages 2317–2327, October 2023.

[4] 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*, ICMI '17, page 350–359, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Z.-Y. Dou, A. Kamath, Z. Gan, P. Zhang, J. Wang, L. Li, Z. Liu, C. Liu, Y. LeCun, N. Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

[6] P. Jonell, T. Kucherenko, G. E. Henter, and J. Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

[7] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[8] 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.

[9] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen. Finite scalar quantization: Vq-vae made simple, 2023.

[10] 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.

[11] 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.

[12] 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, 2013.

[13] 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.

[14] S. Song, M. Spitale, C. Luo, C. Palmero, G. Barquero, H. Zhu, S. Escalera, M. Valstar, T. Baur, F. Ringeval, E. Andre, and H. Gunes. React 2024: the second multiple appropriate facial reaction generation challenge, 2024.

[15] S. Song, M. Spitale, Y. Luo, B. Bal, and H. Gunes. Multiple appropriate facial reaction generation in dyadic interaction settings: What, why and how?, 2023.

[16] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[17] 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 (CVPR)*, pages 20333–20342, June 2022.

[18] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[19] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2023.