ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890674

DiffListener: Discrete Diffusion Model for Listener Generation

Siyeol Jung, Taehwan Kim
Artificial Intelligence Graduate School, UNIST, Republic of Korea
{siyeol, taehwankim}@unist.ac.kr

***Abstract*—The listener head generation (LHG) task aims to generate natural nonverbal listener responses based on the speaker's multimodal cues. While prior work either rely on limited modalities (e.g. audio and facial information) or employ autoregressive approaches which have limitations such as accumulating prediction errors. To address these limitations, we propose DiffListener, a discrete diffusion based approach for non-autoregressive listener head generation. Our model takes the speaker's facial information, audio, and text as inputs, additionally incorporating facial differential information to represent the temporal dynamics of expressions and movements. With this explicit modeling of facial dynamics, DiffListener can generate coherent reaction sequences in a non-autoregressive manner. Through comprehensive experiments, DiffListener demonstrates state-of-the-art performance in both quantitative and qualitative evaluations. The user study shows that DiffListener generates natural context-aware listener reactions that are well synchronized with the speaker. The code and demo videos are available in https://siyeoljung.github.io/DiffListener.**

***Index Terms*—Listener Head Generation, Discrete Diffusion, Dyadic Conversation**

I. Introduction

With the rise of applications such as digital avatar generation [1]–[3] and human-computer interaction [4]–[6], appropriate listening reactions have attracted attentions [7], [8]. In face-to-face conversations, appropriate nonverbal feedback from the listener, rather than content-based replies, is often crucial for maintaining the flow of communication [9]. Therefore appropriate nonverbal feedback is important for virtual agents that communicate with humans [10]. In this context, there is growing interest in generating realistic listener head generation (LHG) [11]–[13]. The listener head generation task aims to generate natural and appropriate nonverbal responses, such as head motions and facial expressions, based on the speaker's verbal utterances and facial expressions. Moreover, listener reactions are influenced not only by the speaker's input but also by the listener's internal state, emotions, and personality, leading to non-deterministic responses. To address this non-deterministic property of the listener's reaction, a previous study [11] proposes to model the listener's reactions using a one-dimensional VQ-VAE [14]. This allows modeling unique reactions for different listener identities while preserving the non-deterministic properties and identity-specific reaction styles.

Most existing LHG approaches rely on autoregressive methodologies to generate listener reactions [11], [12], [15]. These autoregressive methods have inherent structural drawbacks. In particular, during the inference stage, autoregressive models are sensitive to accumulating prediction errors. These accumulated errors can lead the generated sequence to significantly differ from the desired target and produce incoherent or unnatural listener responses. Although a NAR(non-autoregressive) generation approach [13] has been explored, it comes with its own constraints. The reported results are limited to short durations, and extending the response length requires a larger codebook size, which may limit the scalability of the approach.

To address the limitations of existing listener head generation methods, we propose *DiffListener* that generates longer listener responses, which is challenging but more practical, in a non-autoregressive manner and fixed codebook size. First, we train a VQ-VAE [14] model to learn a discrete codebook that encodes listener-specific response patterns. Second, we employ a discrete diffusion model [16] to generate diverse and varied listener responses while preserving the codebook representation. The model performs the denoising diffusion process on the codebook tokens, which represent the listener's responsive reactions. Prior research [11], [13], [15] has primarily focused on the speaker's facial and audio cues to generate listener responses. However, this approach may overlook crucial lexical context. Our method incorporates textual information to consider this aspect. Also, compressing speaker modalities into a condition module might result in a loss of temporal rhythmic information. To overcome it, we incorporated the speaker's facial differential information, which helps maintain temporal rhythmic information, potentially enhancing the naturalness and coherence of generated reactions. In our experiments, DiffListener outperforms the existing baselines in both quantitative and qualitative evaluations. These results demonstrate the effectiveness of our proposed approach.

In summary, the key contributions are as follows:

- We propose DiffListener which is a novel non-autoregressive framework for listener head generation. To the best of our knowledge, it is the first to apply the discrete diffusion to listener generation task.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00608, Artificial intelligence research about multimodal interactions for empathetic conversations with humans & No.RS-2020-II201336, Artificial Intelligence graduate school support(UNIST)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00219959).

- We propose utilizing the facial differential and text information into the non-autoregressive generation framework to provide more context information.
- DiffListener achieves state-of-the-art performance on the listener head generation task, generating longer sequences while preserving high quality and relevance.

II. Related Work

A. Listener Head Generation

VICO [15] proposes the LSTM-based model which generates the listener's responsive reactions based on the speaker's facial and audio information. L2L [11] points out the non-deterministic properties of the listener's reaction. They propose using a codebook to represent the listener's reaction. ELP [13] proposes using multiple codebooks based on the estimated emotion from the speaker. LM-Listener [12] utilizes the large language model to generate the listener's reaction only using the text information of the speaker. Most of the existing listener head generation models [11], [12], [15] utilize the autoregressive approach. However, this approach has drawbacks such as accumulated error problems, so we propose the DiffListener to solve this limitation in non-autoregressive manner.

B. Diffusion Model

The denoising diffusion probabilistic model [17] has shown strong performance not only in the image generation [18]–[20] but also in the other various generation tasks [21]–[23]. The Argmax Flow [24] extends the diffusion model to categorical random variables with transition matrices. The VQ-Diffusion [16] improves the discrete diffusion and applies it to the image generation task, which has also been applied in various tasks [25]–[27]. To the best of our knowledge, this is the first time discrete diffusion is applied to the listener generation task.

III. Methodology

A. Problem Definition

To represent the precise 3D facial expressions and motions from video frames, we utilize the 3D Morphable Face Model(3DMM) [28], [29] for each frame. Through this process, we can get the coefficient that corresponds to the facial expression $\beta_t \in \mathbb{R}^{d_m}$ where d_m is the dimension of expression coefficient, head rotation $R_t \in SO(3)$, and identity-specific shape [30]. We discard the shape coefficient to model the representation that is independent of individual identity [11]. Our facial information at time t can be represented by the concatenation of the facial expression and rotation coefficients $f_t = [\beta_t, R_t]$. Let $F^S = \{f_1^S, f_2^S, \cdots, f_{T-1}^S, f_T^S\}$ represent the sequence of the speaker's facial information across T frames, and let A^S, F_Δ^S, and W^S represent their corresponding audio sequence, differential facial information, and text information, respectively. The differential facial information of the speaker is calculated taking the difference between the facial information in each time step and the previous time step: $F_\Delta^S = \{f_x^S - f_{x-1}^S \mid 1 \le x \le T\}$. When DiffListener is given the speaker's information, the listener's responsive head sequence of the corresponding length F^L can be generated:

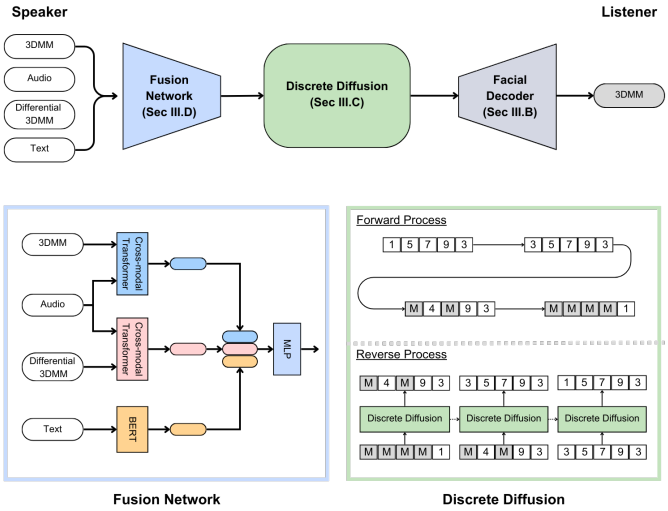


Fig. 1. **The overview of DiffListener.** The discrete diffusion model takes the fused representation and generates the listener's response sequence tokens. These tokens are then passed through the VQ-VAE decoder to obtain the listener's 3DMM coefficients.

$$F^L = DiffListener(F^S, A^S, F_\Delta^S, W^S) \tag{1}$$

B. Quantizing Listener Motion

We use the VQ-VAE [14] to quantize the listener's facial information. The VQ-VAE consists of facial encoder E, facial decoder D and codebook $C = \{c_k\}_{k=1}^K \in \mathbb{R}^{K \times d_z}$ where K is the size of the codebook and d_z is the dimension of each code. Given the sequence of listener facial information $F^L = \{f_1^L, f_2^L, \cdots, f_{T-1}^L, f_T^L\}$ across T frames, we encode F^L into specific representation through the facial encoder $Z = E(F^L) \in \mathbb{R}^{\frac{T}{\tau} \times d}$ where τ represents the ratio of downsampling and d represents the dimension of the representation. After that the vector quantizer Q maps each Z elements to its closest codebook entry. Finally, the facial decoder reconstructs the sequence of the listener's facial information. We utilize the combination of losses to train the VQ-VAE:

$$L_{embed} = \sum_{t=1}^{T/\tau} \parallel z_t - sg[c_i] \parallel^2 \tag{2}$$

$$L_{reconstructed} = \sum_{t=1}^{T} L_1^{smooth}(\hat{f}_t - f_t) \tag{3}$$

$$L_{velocity} = \sum_{t=1}^{T-1} L_1^{smooth}(\hat{f}_{t+1} - \hat{f}_t, f_{t+1} - f_t) \tag{4}$$

where sg is the stop-gradient operation, and L_1^{smooth} is the L1 smooth loss function. The total loss is a weighted sum of these losses.

TABLE I
COMPARISION OF OUR MODEL WITH BASELINES IN TREVOR AND STEPHEN DATASET. BOLD REPRESENTS THE BEST. ↓ INDICATES LOWER IS BETTER, NO ARROW INDICATES THAT CLOSER TO GT IS BETTER. GT DENOTES THE GROUND TRUTH.

	Trevor					Stephen				
Models	L2↓	FD↓	P-FD↓	Diversity	Variation	L2↓	FD↓	P-FD↓	Diversity	Variation
GT				2.59	0.11				3.81	0.18
VICO [15]	0.41	17.94	19.22	2.31	0.08	0.71	31.43	33.64	2.98	0.11
L2L [11]	0.62	25.28	27.11	4.66	0.28	**0.65**	28.69	30.33	2.79	0.09
LM-Listener [12]	0.69	28.67	30.56	4.77	0.29	0.79	37.57	39.21	2.34	0.09
Ours	**0.40**	**15.75**	**17.22**	2.96	0.10	**0.65**	**26.47**	**28.47**	3.56	0.14

C. Discrete Diffsuion Model

We use VQ-Diffusion [16] to generate the listener response conditioned on the speaker's representation. To get the speaker's representation, we utilized the speaker's information I^S which consists of $F^S, A^S, F^S_\Delta, W^S$. This information is then processed through the fusion network described in Section III-D to generate the corresponding speaker's representation R^S. Let x be a token sequence of length T/τ representing the listener's response, where each token is from the codebook. During the forward diffusion process, the data x is progressively corrupted through a fixed Markov chain $q(x_t|x_{t-1})$, which randomly replaces some tokens of x_{t-1} over the diffusion time steps T_d. The reverse diffusion process then restores the data from x_{T_d} based on the architecture. The VQ-Diffusion [16] operates based on a transition probability matrix. The transition probability matrix $[Q_t]_{mn} = q(x_t = m \mid x_{t-1} = n) \in R^{K\times K}$ that represents the probabilities of x_{t-1} transit to x_t. To get better reverse estimation, a [Mask] token is introduced, expanding the transition matrix to $Q_t \in \mathbb{R}^{(K+1)\times(K+1)}$ [16]. To estimate the posterior transition distribution $q(x_{t-1} \mid x_t, x_0)$, we train the denoising network $p_\theta(x_{t-1} \mid x_t, R^s)$. The network is trained to minimize the variational lower bound [17]:

$$L_{vlb} = \sum_{t=1}^{T_d-1} [D_{KL}[q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t, R^s)]] + D_{KL}(q(x_{T_d} \mid x_0) \parallel p(x_{T_d})) \quad (5)$$

where $p(x_{T_d})$ refers to the prior distribution of timestep T_d. To get better results we utilize the reparameterization trick [16]. This leads to a more noiseless distribution at each step.

$$L_{x_0} = -logp_\theta(x_0 \mid x_t, R^s) \quad (6)$$

where L_{x_0} refers the auxiliary denoising loss, which can be further improved when combined with the L_{vlb}.

D. Fusion Network

Initially, we extract the text features by using the pre-trained language model BERT [31] and audio features using Mel-frequency cepstral coefficients (MFCC). The fusion network consists of two components, as illustrated in Figure 1. The first module focuses on fusing audio and 3DMM features. Cross-modal attention is computed using the queries Q_A from A^S, and the keys K_F and values V_F from F^S. The second module handles the fusion of audio with differential 3DMM features.

TABLE II
ABLATION STUDY RESULTS BASED ON CODEBOOK SIZE AT TREVOR DATASET. GT DENOTES THE GROUND TRUTH.

Codebook size	L2↓	FD↓	P-FD↓	Diversity	Variation
GT				2.59	0.11
128	0.47	17.61	19.28	3.63	0.17
256	**0.40**	**15.75**	**17.22**	2.96	0.10
512	0.48	18.28	20.01	3.88	0.18

In this module, we set the queries Q_{F_Δ} from F^S_Δ and the keys K_A and values V_A from A^S. Finally, we concatenate each of the fused representations along with the text representation and feed the combined representation into an MLP layer for further processing.

IV. EXPERIMENT

A. Dataset

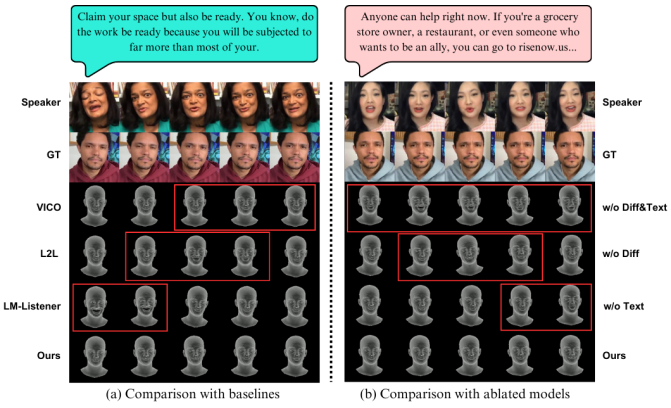
Our research objective is to generate longer identity-specific listener responses. To achieve this goal, we select the dataset [11], [12], which provides sufficiently long sequences of listener behaviors while maintaining individual identity characteristics. We preprocess the dataset by setting the listener's response period to 8 seconds (240 frames), assuming that this duration may be sufficient to understand contextual information through text data [12]. We clip each video to 8 seconds and employ a sliding-window approach to generate more data. Based on this preprocessing, we found sufficient data from Trevor and Stephen identities. To extract text from audio data, we utilize the Whisper [32].

B. Experimental Setup

Following the previous work [11], [12], we use the following evaluation metrics for quantitative evaluation. L2, FD(Frechet Distance), P-FD (Paired FD) which is the distribution distance between listener and speaker, Diversity, and Variation. We use the L2L [11], VICO [15], and LM-Listener [12] as our baselines. The ELP [13], which is based on the non-autoregressive approach, does not release the code, so we are not able to compare ours with it. We use $K = 256$, $T = 240$, $T_d = 100$, $d_z = 512$, $\tau = 8$ with a batch size = 256 when training on the Trevor dataset [12], and batch size 64 when training on the Stephen dataset [12]. The value $K = 256$ is chosen by its superior performance, as shown in Table II. During DiffListener training, we apply early stopping with a patience of 5 epochs. For VQ-VAE training, we set the

TABLE III
ABLATION STUDY RESULTS BASED ON SPEAKER MODALITIES AT TREVOR DATASET.

	L2↓	FD↓	P-FD↓	Diversity	Variation
GT				2.59	0.1127
w/o Diff & Text	0.50	20.77	22.27	3.18	0.1223
w/o Diff	0.44	17.17	18.71	3.21	0.1247
w/o Text	0.41	16.17	17.66	3.05	0.1095
Ours	**0.40**	**15.75**	**17.22**	2.96	0.1027



(a) Comparison with baselines (b) Comparison with ablated models

Fig. 2. The visualization comparison with baselines and ablated models at Trevor dataset. The blue and pink boxes include the speaker's utterance.

weights of each loss term as follows: 0.02 for L_{embed}, 1 for $L_{reconstructed}$, and 0.05 for $L_{velocity}$.

C. Quantitative Comparison

Table I presents the overall performance of our model and the baselines. Our model achieves the lowest L2, FD, and P-FD scores for both Trevor and Stephen's datasets, indicating the highest realism in generating listener motions and the greatest synchrony with the speaker. Our model achieves superior performance in terms of diversity and variation scores in most cases. For the Trevor dataset, the diversity metric is slightly less close to the ground truth compared to the VICO model. However, the difference is minimal, and our model achieves a higher diversity score and lower L2, FD, and P-FD scores than VICO. This indicates that our model can generate more realistic and diverse results.

Table III presents the results of our ablation study on different modality conditions. By incorporating text into a model without differential and text information (w/o Diff & Text), we observed approximately 17.33% decrease in FD score (w/o Diff). Only by incorporating differential 3DMM information into the model to provide rhythmic temporal information, we observed an approximately 22.16% decrease in the FD score (w/o Text). Furthermore, integrating the differential 3DMM value and text, we observed a 24.1% decrease in the FD score and 22.6% decrease in the P-FD score (Ours). This indicates that the combination of differential information with audio, 3DMM, and text data enhances the model's ability to understand the context information, thereby enabling it to generate more appropriate responses.

TABLE IV
USER STUDY RESULTS BASED ON THE COMPARISON OF OUR MODEL WITH BASELINES.

	Trevor			Stephen		
	Win	Tie	Lose	Win	Tie	Lose
Ours vs. Vico [15]	**51%**	23%	26%	**65%**	11%	24%
Ours vs. L2L [11]	**50%**	11%	39%	**41%**	25%	34%
Ours vs. LM-Listener [12]	**67%**	16%	16%	**47%**	12%	41%

D. Qualitative Comparision

In the Listener Generation task, it is important to generate results that closely match the ground truth. Figure 2 (a) presents a visual comparison with the baseline models. The baselines sometimes fail to generate appropriate responses, as indicated by the red boxes. While the ground truth sample is in a neutral state, other baselines sometimes show a smile. In contrast, our model demonstrates more appropriate responses in both head pose and facial expression. This result may come from the advantages of DiffListener. First, NAR approach can avoid the problem of accumulated errors. It makes our model's results more robust during the inference. Second, using various modalities provides sufficient contextual information. Figure 2(b) presents a visual comparison with ablated models. When sufficient contextual information is provided, the model generates more appropriate listener responses. However, when some modalities are excluded, it produces inappropriate responses, as indicated by the red boxes. In addition, we conduct a user study to evaluate human preferences on Amazon Mechanical Turk. We randomly select 25 videos from each of the identity datasets (total 50 videos) and visualize each result as a grayscale mesh video using EMOCA [33] because mesh videos provide a more intuitive way to evaluate facial and head movements compared to photorealistic videos [13]. If photorealistic video is required, it can be generated using a renderer model [34]. Given the (speaker, ours, baseline) tuple of samples, 20 people were asked to choose the video that appeared to be listening and paying more attention to the speaker. The results are presented in Table IV. Our model is more preferred than the baselines in both datasets. More samples can be found on our demo page[1].

V. CONCLUSION

In this work, we propose DiffListener, a novel approach that generates realistic and diverse listener responses in a non-autoregressive manner using a discrete diffusion model. Unlike previous work, our method can generate longer responses while maintaining a fixed codebook size. To better synchronize with the speaker, we introduce a novel speaker modality (the speaker's facial motion differential). Through experiments, we demonstrated that our approach outperforms existing baselines in terms of realism, diversity, and synchrony with the speaker's motions. This work represents a significant step forward in achieving more natural and context-aware listener generation.

[1]https://siyeoljung.github.io/DiffListener/

REFERENCES

[1] L. Chen, C. Cao, F. De la Torre, J. Saragih, C. Xu, and Y. Sheikh, "High-fidelity face tracking for ar/vr via deep lighting adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13059–13069, 2021.

[2] 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*, pp. 13401–13412, 2021.

[3] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv preprint arXiv:2208.15001*, 2022.

[4] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

[5] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5784–5794, 2021.

[6] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan, "Emotalk: Speech-driven emotional disentanglement for 3d face animation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20687–20697, 2023.

[7] A. Tolzin and A. Janson, "Mechanisms of common ground in human-agent interaction: A systematic review of conversational agent research.," in *HICSS*, pp. 342–351, 2023.

[8] E. Cho, N. Motalebi, S. S. Sundar, and S. Abdullah, "Alexa as an active listener: how backchanneling can elicit self-disclosure and promote user experience," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–23, 2022.

[9] J. Cassell and K. R. Thorisson, "The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents," *Applied Artificial Intelligence*, vol. 13, no. 4-5, pp. 519–538, 1999.

[10] E. Tronick, H. Als, and T. B. Brazelton, "Monadic phases: A structural descriptive analysis of infant-mother face to face interaction," *Merrill-Palmer Quarterly of Behavior and Development*, vol. 26, no. 1, pp. 3–24, 1980.

[11] 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*, pp. 20395–20405, 2022.

[12] E. Ng, S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar, "Can language models learn to listen?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10083–10093, 2023.

[13] L. Song, G. Yin, Z. Jin, X. Dong, and C. Xu, "Emotional listener portrait: Neural listener head generation with emotion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20839–20849, 2023.

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

[15] 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*, pp. 124–142, Springer, 2022.

[16] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

[17] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, pp. 2256–2265, PMLR, 2015.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[19] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.

[20] D. Epstein, A. Jabri, B. Poole, A. Efros, and A. Holynski, "Diffusion self-guidance for controllable image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16222–16239, 2023.

[21] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[22] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *arXiv preprint arXiv:2210.08933*, 2022.

[23] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[24] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, "Argmax flows and multinomial diffusion: Learning categorical distributions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12454–12465, 2021.

[25] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[26] Y. Li, Y. Dou, X. Chen, B. Ni, Y. Sun, Y. Liu, and F. Wang, "Generalized deep 3d shape prior via part-discretized diffusion process," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16784–16794, 2023.

[27] S. Han and J. Kim, "Clip-vqdiffusion: Langauge free training of text to image generation using clip and vector quantized diffusion model," *arXiv preprint arXiv:2403.14944*, 2024.

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

[29] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, "3d morphable face models and their applications," in *Articulated Motion and Deformable Objects: 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings 9*, pp. 185–206, Springer, 2016.

[30] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt, "State of the art on monocular 3d face reconstruction, tracking, and applications," in *Computer graphics forum*, vol. 37, pp. 523–550, Wiley Online Library, 2018.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[33] R. Daněček, M. J. Black, and T. Bolkart, "Emoca: Emotion driven monocular face capture and animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20311–20322, June 2022.

[34] 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*, pp. 13759–13768, 2021.