AN AUDIO-TEXTUAL DIFFUSION MODEL FOR CONVERTING SPEECH SIGNALS INTO ULTRASOUND TONGUE IMAGING DATA

Yudong Yang[1], Rongfeng Su[1,], Xiaokang Liu[1,2], Nan Yan[1,3], Lan Wang[1,3,*]*

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
[2]University of Chinese Academy of Sciences, Beijing, China.
[3]Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen, China

ABSTRACT

Acoustic-to-articulatory inversion (AAI) is to convert audio into articulator movements, such as ultrasound tongue imaging (UTI) data. An issue of existing AAI methods is only using the personalized acoustic information to derive the general patterns of tongue motions, and thus the quality of generated UTI data is limited. To address this issue, this paper proposes an audio-textual diffusion model for the UTI data generation task. In this model, the inherent acoustic characteristics of individuals related to the tongue motion details are encoded by using wav2vec 2.0, while the ASR transcriptions related to the universality of tongue motions are encoded by using BERT. UTI data are then generated by using a diffusion module. Experimental results showed that the proposed diffusion model could generate high-quality UTI data with clear tongue contour that is crucial for the linguistic analysis and clinical assessment. The codes and examples can be found on the website[1].

Index Terms— Diffusion Model, Acoustic-to-Articulatory Inversion, Ultrasound Tongue Imaging, Generation

1. INTRODUCTION

The problem of estimating the movement of articulators from speech signals is known as acoustic-to-articulatory inversion (AAI). In recent years, AAI has garnered increasing attention due to its potential applications in speech technology, such as automatic speech recognition [1, 2], speech therapy [3, 4], and speech assessment [5, 6]. The articulatory movements can be measured through various techniques, for instance real-time magnetic resonance imaging (rt-MRI) [7], X-ray microbeam [1], electromagnetic articulography (EMA) [2], and ultrasound tongue imaging (UTI) [8, 9]. Compared to the more invasive EMA and the operationally complex rt-MRI, UTI has the characteristics of simple operation, non-invasive,

---

*Corresponding authors: Rongfeng Su(email: rf.su@siat.ac.cn); Lan Wang(email: lan.wang@siat.ac.cn).
[1]https://yangyudong2020.github.io/wav2uti/

and clear visualization. The tongue contour in UTI data is the most important part to display the tongue motor function [10]. Suitable AAI methods to generate UTI data with clear tongue contour is of great importance in clinical applications.

The current AAI methods can be divided into two categories: discriminative and generative methods. Discriminative methods generate the sequence data of the tongue by identifying the mapping relationship between the speech signal of each pronunciation and the corresponding tongue motions [1, 8, 11]. Due to the lack of modelling ability regarding the temporal dependencies across multiple consecutive pronunciations, the quality of generated tongue motions is limited. For example, the state-of-the-art DNN based AAI system generated UTI data with blurry tongue contour [8], which is difficult to obtain tongue motion trajectories for further applications. In contrast, generative methods learn the joint probability distribution of speech and articulatory motions, which can generate results with better temporal coherence [7, 12, 13]. Researchers hoped to directly derive the general patterns of tongue movements from large amounts of personalized acoustic inputs by using various neural network structure [8, 9]. For the UTI data generation task in this paper, it is difficult to acquire high-quality UTI data from limited speech-ultrasound parallel training data in many real applications. In fact, the tongue motions related to the tongue contour in UTI data can be divided into the individual and universal parts. The individual part means the personal details in tongue movements, such as the height of the tongue position. The universal part means the general patterns of tongue movements, which should be consistent for the same consecutive pronunciations of different speakers. Obviously, additional text inputs contain rich universal information with respect to the tongue movements, which may reduce the requirement of parallel training data volume and thus improve the quality of generated UTI data.

In this study, inspired by the successful use of diffusion model in medical image generation tasks with richer details [14, 12, 15, 16], a multimodal diffusion model that integrates speech and textual information is proposed for the

UTI data generation task. To best of our knowledge, this study is the first to apply the diffusion model for converting audio into UTI data. The proposed model consists of two stages: conditional encoding and UTI data generation. In the first stage, the inherent acoustic characteristics of individuals related to the details of tongue movements are encoded by using wav2vec2.0 [17], while the ASR transcriptions in the textual space related to the universality of tongue motions are encoded by using BERT [18]. The personalized acoustic information and universal textual information are then fused by using cross-attention mechanism. In the second stage, the fused features are used as the conditional inputs of the diffusion module [19]. In this module, the long-term dependencies between consecutive pronunciations and UTI data are captured by using cyclic denoising sampling strategy. Experimental results on a Mandarin speech-ultrasound dataset showed that the proposed audio-textual diffusion AAI system outperformed the state-of-the-art DNN-based AAI system by a LPIPS [20] improvement of 67.95% relative, with a FID [21] decreased from 256.80 to 22.02. This demonstrates that the generated results exhibit data distribution and diversity closer to real UTI data. Besides, the proposed diffusion AAI system generated high-quality UTI data with clear tongue contour, which is crucial for clinical assessment tasks, such as tongue function assessment.

2. METHODS

In this section, audio-only diffusion model and the audio-textual diffusion model will be presented for generating ultrasound tongue imaging (UTI) data from speech signals.

2.1. Audio-only Diffusion Model for UTI Generation

The audio-only diffusion model for UTI data generation is illustrated in Fig.1(a). The original speech inputs were first pre-processed by the Wav2vec2.0 [17] to obtain the acoustic embeddings for describing different pronunciations. The acquired acoustic embeddings c and the corresponding time steps t were then fed into the UTI data generation module.

The core of the UTI data generation module is the diffusion model, which is shown in Fig.I. The generic pipeline of diffusion models involves a forward process and a reverse process to learn a data distribution, as well as a sampling procedure to generate novel data. For the forward process, supposed that x is the UTI data, $q_{utl}(x)$ represents the distribution of real UTI data and the corresponding standard deviation is σ_{utl}. By adding a i.i.d Gaussian noise with the standard deviation σ to the UTI data, a family of distributions $p(x;\sigma)$ can be obtained. When $\sigma_{max} \gg \sigma_{utl}$, $p(x;\sigma_{max})$ is approximated to the Gaussian noise. The key point of the diffusion model is to randomly sample a noisy start point $x_0 \sim ((0, \sigma_{max}^2, \mathbf{I})$, and then denoise it into the UTI data x_i with noise level $\sigma_{max} = \sigma_0 > \sigma_1 > \cdots > \sigma_N = 0$, where $x_i \sim p(x;\sigma_i)$. The finally generated UTI data x_N will have

the approximated distribution of real data $q_{utl}(x)$.

In order to reverse the forward process of the diffusion model and gradually restore to the true data distribution from a purely random state. A denoising function $D(x,\sigma)$ is introduced to minimize the $L2$ denoising error of all extracted samples from q_{utl}, the loss function is defined as follow:

$$\mathcal{L} = \mathbb{E}_{y\sim q_{utl}} \mathbb{E}_{n\sim(0,\sigma_{max}^2,\mathbf{I})}||D(y+n,\sigma)-y||_2^2 \quad (1)$$

Where y is a training data point and n is noise. By using the ordinary differential equations (ODE) [22], the noise level of each data point can be increased or decreased by moving it forward or backward in the diffusion process, respectively. The requirement of ODE is satisfied by:

$$dx = -\sigma(t)\,\sigma(t)\,\nabla_x \log p(x;\sigma(t))dt \quad (2)$$

where the dot denotes a time derivative, $\sigma(t) = t$ is a schedule, $\nabla_x \log p(x;\sigma)$ is the score function [23], a vector field that points towards higher density of data at a given noise level. In this paper, the tongue contour in UTI data is supposed to be a higher density vector field. According to the equation (2) used in the denoising process, the network optimization will pay more attention to the field related to the tongue contour in UTI data, and thus the UTI data with clearer tongue contour can be acquired.

The aim of this study is to generate coherent and natural UTI data under given speech conditions. First, the low-quality UTI data sequence v_0 is generated from a pretrained speech encoder, such as Wav2vec2.0 used in this paper. Then, v_0 is used as a conditional input for the next diffusion model, which follows the concept of a Cascade Diffusion Model to generate a high-quality and clearer v_1. As shown in Fig.1(c), compared to traditional image diffusion models, the architecture based on Unet [19] incorporates changes in different levels and adds time-aware layers through 3D convolutions. We use the raw speech signal V_{audio} as a condition with the Elucidated Diffusion Model (EDM) [22] setup, and apply it to UTI data generation. The denoising models in the cascade are defined as D_{θ_s}, where s defines the rank of the cascaded models, and D_{θ_0} is defined as:

$$\begin{aligned} D_{\theta_0}(x;\sigma,\lambda_c) = c_{skip}(\sigma)\,x \\ + c_{out}(\sigma)\,F_{\theta_s}(c_{in}(\sigma)\,x;c_{noise}(\sigma),\lambda_c) \end{aligned} \quad (3)$$

Where F_{θ_s} is the neural network transformation of D_{θ_0} and D_{θ_0} outputs v_0. Supposed that σ_{utl} is the standard deviation of the real UTI data distribution, c_{skip}, c_{out}, c_{in} and c_{noise} are defined as:
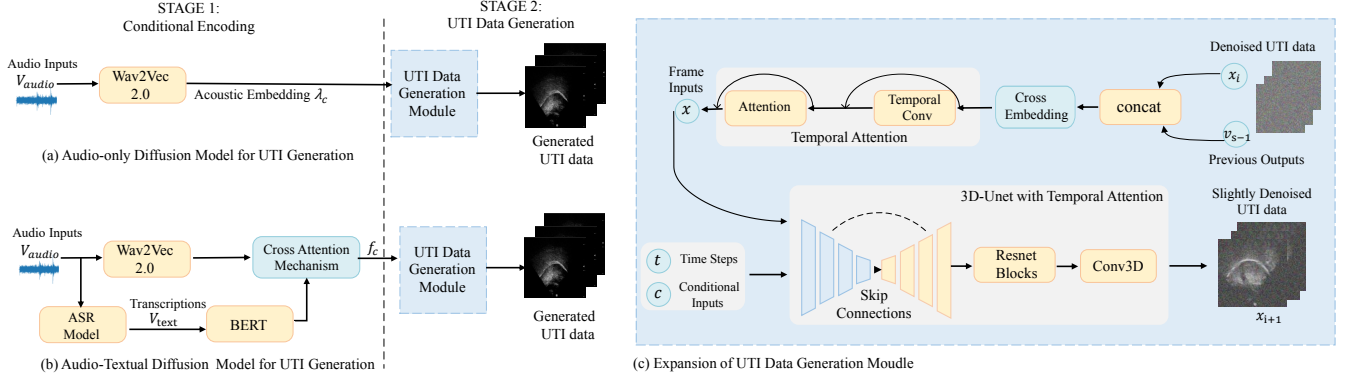
$$c_{skip}(\sigma) = (\sigma_{q_{utl}}^2)/(\sigma^2 + \sigma_{q_{utl}}^2) \quad (4)$$

$$c_{out}(\sigma) = \sigma * \sigma_{q_{utl}} * 1/(\sigma^2 * \sigma_{q_{utl}}^2)^{0.5} \quad (5)$$

$$c_{in}(\sigma) = 1/(\sigma^2 * \sigma_{q_{utl}}^2)^{0.5} \quad (6)$$

$$c_{noise}(\sigma) = \log(\sigma_t)/4 \quad (7)$$

The sampling procedure is accomplished through a random sampling method. The sampling begins with a noise sample $x_0 \sim ((0, t_0^2, \mathbf{I})$, and the noise level is given by $\sigma(t_i) = t_i$, Then, we calculate the variable factor γ_i, which is

2171

Fig. 1: The overall structure of the audio-only and audio-textual diffusion models for ultrasound tongue imaging (UTI) data generation, where f_c is the fusion conditions from the acoustic and textual encoding.

used to control the magnitude of noise level adjustment $\sigma(t_i)$ over the time step t_i. The variable factor γ_i is defined as:

$$\gamma_i(t_i) = min(s_{churn}/N, \sqrt{2} - 1)\forall t_i \in (s_{tmin}, s_{tmax}) \quad (8)$$

Where N represents the number of sampling steps. For $\forall i \in \{0, \ldots, N-1\}$, we sample the $\epsilon_i \sim ((0, S_{noise}, \mathbf{I})$ and calculate the slightly increased noise level $\hat{t}_i = (\gamma_i(t_i)) + 1)t_i$. The previous sample is $\hat{x}_i = x_i + (\hat{t}_i^2 - t_i^2)^{0.5}\epsilon_i$. Then, this sample is applied in the denoising model D_θ and the local slope is $d_i = (\hat{x}_i - D_\theta(\hat{x}_i; \hat{t}_i))/\hat{t}_i$, which is used to predict the next sample $x_{i+1} = \hat{x}_i + (\hat{t}_{i+1} - \hat{t}_i)d_i$. Except for the last step, a correction to x_{i+1} is applied: $d_i' = (x_{i+1} - D_\theta(x_{i+1}; t_{i+1}))/t_{i+1}$ and $x_{i+1} = \hat{x}_i + (t_{i+1} - \hat{t}_i)(d_i + d_i')/2$. This process is then iterated sequentially in the cascade model, using the previous cascade output v_{s-1} as a condition, and having the speech encoding λ_c.

2.2. Audio-Textual Diffusion Model for UTI Generation

As discussed before, for the given consecutive pronunciations, the tongue motions can be divided into individual detailed information and universal movement patterns. In fact, the personalized speech signals contain rich individual detailed information related to the tongue motions, such as the tongue position. The text inputs, as the speaker-independent semantic representation, inherently contain the generality information about tongue movements. In this paper, we hope to improve the model ability to generate clear and coherent UTI data by integrating the personalized detailed acoustic information and the universal textual information.

The proposed audio-textual diffusion model is shown in Fig. 1(b). In the proposed model, the audio inputs were fed into the large Wav2vec 2.0 to obtain the acoustic embeddings. To obtain reliable text inputs from the speech signals, a well-trained large Whisper ASR model is used in this paper. This ASR model achieved a character error rate (CER) of 0.02% on the whole Mandarin speech-ultrasound dataset. Inspired by the success of BERT [18] applied in language understanding tasks, the ASR transcriptions was preprocessed by a well-trained BERT to derive the word embeddings. Such word

embeddings are supposed to contain the high-level semantic information that can reflex the universality of the tongue motions. The acquired acoustic and textual embeddings were then fused by using the cross-attention mechanism [24]. The fused features f_c were used as the conditional inputs c of the UTI data generation module, as shown in Fig. 1(c). The training process of the audio-textual diffusion model is similar to the one described in Sec. 2.1. The only difference is D_{θ_0} of equation (2). Supposed that the audio inputs are V_{audio} and the ASR transcriptions are V_{text}, D_{θ_0} is defined as:

$$D_{\theta_0}(x; \sigma, f_c) = c_{skip}(\sigma)x + c_{out}(\sigma)F_{\theta_s}(c_{in}(\sigma)x; c_{noise}(\sigma), f_c) \quad (9)$$

3. EXPERIMENT

3.1. Datasets

Experiments were conducted on Mandarin speech-ultrasound dataset. This dataset was collected from 44 healthy persons with three different speech tasks (vowel, word and sentence), totally 6.85 hours. The training set consists of 40 speakers, while the test set consists of 4 speakers. No overlap exists between the training and test sets. The UTI data were recorded in mid sagittal orientation using a Focus&Fusion Finus 55 ultrasound system with a sampling rate of 60 fps and a resolution of 920×700. The P5-2 phased array probe was fixed by an ultrasound stabilization headset. The speech data were recorded by a BOYA BY-WM4 PRO microphone with a sampling frequency of 16khz and single channel. The speech signals and UTI data were synchronized by using an external sound card. The downsampled UTI data under different resolutions were used for training different models.

3.2. Implement Details

Following the state-of-the-art AAI system based on DNN for the UTI data generation task [8], a similar DNN-based AAI baseline system was reimplemented in this paper. All neural networks were trained on four NVIDIA A6000 GPUs with a batch size of 4. The experimental setup included hyperpa-

rameters with a minimum noise level of 0.002, a maximum noise level of 160, a data distribution standard deviation of 0.25, and a learning rate of 5×10^{-4}. All neural networks were trained with 5×10^6 iterations, and the parameters were updated every 1000 iterations. For the audio part, the large Wav2Vec2.0 were obtained from a 56k-hour corpus. For the text part, the well-trained large BERT has 110M parameters. During the training process of all neural networks, the parameters of both large Wav2Vec2.0 and large BERT were frozen.

3.3. Evaluation Metrics

Various metrics are used to comprehensively assess the quality of generated UTI data. The root Mean Square Error (RMSE) and Peak Signal-to-Noise Ratio (PSNR) metrics are used to evaluate the pixel-level similarity and quality differences. Besides, Learned Perceptual Image Patch Similarity(LPIPS) and Fréchet Inception Distance (FID) are used to evaluate the perception quality of UTI data and the similarity of statistical distributions, which are computed using VGG [25] and I3D [26] networks, respectively.

3.4. Experimental Result

The performance of the DNN-based AAI system and the proposed diffusion AAI system is shown in Table 1. It could be observed that the proposed AAI system consistently outperformed the DNN-based AAI baseline in all metrics (RMSE, LPIPS, PSNR and FID). For example, a LPIPS improvement of 67.95% relative and a FID improvement of 91.43% relative were obtained from the diffusion system using audio-textual information over the DNN-based AAI baseline. The generated results are displayed in Fig. 2.

Table 1: Performance comparison of the DNN-based AAI baseline system and the proposed diffusion AAI system under a UTI output size of 112×112. "A" represents the audio-only inputs, and "A+T" represents the audio-textual inputs.

| Models	Inputs	RMSE↓	PSNR↑	LPIPS↓	FID↓
DNN [8]	A	5.8489	32.8234	0.3323	256.80
Ours	A	5.6635	33.1081	0.1206	27.81
	A+T	**5.6341**	**33.1482**	**0.1065**	**22.02**

We also compared the performance of the proposed diffusion AAI system with or without additional textual information under different UTI output resolutions. As shown in Table 2, the diffusion AAI system using textual inputs consistently outperformed the counterpart without textual inputs. For example, even if under the largest UTI output size of 112*112, the audio-textual AAI system outperformed the audio-only AAI system by a LPIPS reduction of 11.69% relative and a FID reduction of 20.82% relative, respectively. The results in Table 2 have given the strong evidences that the introduction of additional textual information can significantly enhance the quality of the generated UTI data, although the used speech-ultrasound training data is only 6.39 hours.

In order to intuitively illustrate the performance of different AAI systems, some examples of the UTI data generated

Table 2: Performance comparison of the diffusion AAI systems with or without additional textual information under various UTI output resolutions.

#Resolution	Inputs	RMSE↓	PSNR↑	LPIPS↓	FID↓
64×64	A	5.8114	32.8734	0.0972	80.51
	A+T	5.7512	32.9820	0.0946	78.65
96×96	A	5.7681	32.9355	0.1547	50.30
	A+T	5.7025	33.0411	0.1195	37.01
112×112	A	5.6635	33.1081	0.1206	27.81
	A+T	**5.6341**	**33.1482**	**0.1065**	**22.02**

from the DNN-based AAI baseline and the proposed diffusion systems are shown in Fig. 2. As shown in Fig. 2, compared to the other AAI systems, the diffusion AAI system using additional textual information can generated the best UTI data with the most clear tongue contour.

(a) Real   (b) DNN   (c)Ours(A)   (d)Ours(A+T)
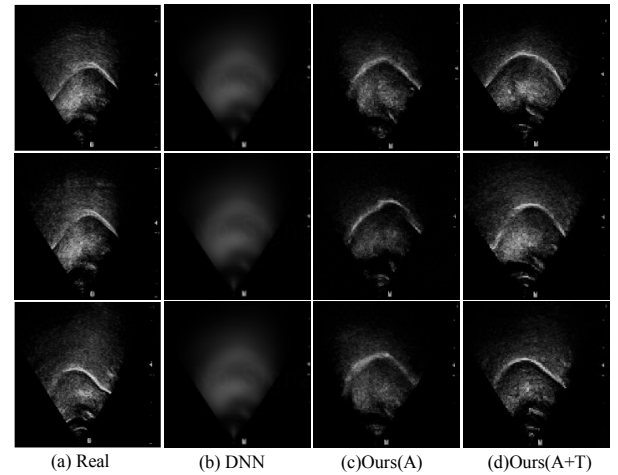
Fig. 2: Examples of real and generated UTI data.

4. CONCLUSION

In this paper, the diffusion model is first applied for converting audio into ultrasound tongue imaging (UTI) data. Using additional textual inputs, the proposed diffusion AAI system can generate high quality UTI data with clear tongue contour, which is crucial for clinical assessment tasks, such as tongue function assessment. Future works will focus on the new applications of diffusion model related to speech orders.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] Tanja Schultz and Michael Wand et al, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[2] Abdolreza Sabzi Shahrebabaki and Giampiero Salvi et al, "Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 135–147, 2022.

[3] Philipp Arens and Thomas Fischer et al, "Ultrasound shear wave elastography of the tongue during selective hypoglossal nerve stimulation in patients with obstructive sleep apnea syndrome," *Ultrasound in Medicine Biology*, vol. 47, no. 10, pp. 2869–2879, 2021.

[4] Manuel Sam Ribeiro and Joanne Cleland et al, "Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors," *Speech Communication*, vol. 128, pp. 24–34, 2021.

[5] Manuel Sam Ribeiro and Jennifer Sanger et al, "Tal: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *SLT 2021*, 2021, pp. 1109–1116.

[6] Rui-Chen Zheng and Yang Ai et al, "Speech reconstruction from silent tongue and lip articulation by pseudo target generation and domain adversarial training," in *ICASSP 2023*, 2023, pp. 1–5.

[7] Sathvik Udupa and Prasanta Kumar Ghosh, "Real-time mri video synthesis from time aligned phonemes with sequence-to-sequence networks," in *ICASSP 2023*, 2023, pp. 1–5.

[8] Dagoberto Porras and Alexander Sepúlveda-Sepúlveda et al, "Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging," in *IJCNN 2019*, 2019, pp. 1–8.

[9] Jianrong Wang and Yalong Yang et al, "Continuous ultrasound based tongue movement video synthesis from speech," in *ICASSP 2016*, 2016, pp. 1716–1720.

[10] Jonathan L Preston and Tara McAllister Byun et al, "Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors," *Journal of Visualized Experiments*, , no. 119, pp. e55123, 2017.

[11] Sathvik Udupa and Siddarth C et al, "Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models," in *ICASSP 2023*, 2023, pp. 1–5.

[12] Soowon Kim and Young-Eun Lee et al, "Diff-E: Diffusion-based Learning for Decoding Imagined Speech EEG," in *Proc. INTERSPEECH 2023*, 2023, pp. 1159–1163.

[13] Kevin Scheck and Tanja Schultz, "STE-GAN: Speech-to-Electromyography Signal Conversion using Generative Adversarial Networks," in *Proc. INTERSPEECH 2023*, 2023, pp. 1174–1178.

[14] Hadrien Reynaud and Mengyun Qiao et al, "Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis," *arXiv preprint arXiv:2303.12644*, 2023.

[15] Ruihan Yang and Prakhar Srivastava et al, "Diffusion probabilistic modeling for video generation," *arXiv preprint arXiv:2203.09481*, 2022.

[16] Chitwan Saharia and William Chan et al, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *NeurIPS 2022*, vol. 35, pp. 36479–36494, Dec. 2022.

[17] Alexei Baevski and Yuhao Zhou et al, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS 2020*, vol. 33, pp. 12449–12460, 2020.

[18] Jacob Devlin and Ming-Wei Chang et al, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Jonathan Ho and Tim . Salimans et al, "Video diffusion models," *arXiv preprint arXiv:2204.03458*, 2022.

[20] Richard Zhang and Phillip Isola et al, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR 2018*, 2018, pp. 586–595.

[21] Martin Heusel and Hubert Ramsauer et al, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS 2017*, vol. 30, 2017.

[22] Tero Karras and Miika Aittala et al, "Elucidating the design space of diffusion-based generative models," *NeurIPS 2022*, vol. 35, pp. 26565–26577, 2022.

[23] Yang Song and Jascha Sohl-Dickstein et al, "Score-based generative modeling through stochastic differential equations," in *ICLR 2021*, 2021.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "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.

[25] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] Ivan Skorokhodov and Sergey Tulyakov et al, "Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2," in *CVPR 2022*, 2022, pp. 3626–3636.