# **Successfully Applying Lottery Ticket Hypothesis to Diffusion Model**

# Chao Jiang<sup>1</sup> Bo Hui<sup>2</sup>\* Bohan Liu<sup>3</sup> Da Yan<sup>4</sup>

Visa<sup>1</sup>, University of Tulsa<sup>2</sup>, Carnegie Mellon University<sup>3</sup>, Indiana University Bloomington<sup>4</sup> chajiang@visa.com, bo-hui@utulsa.edu bohanli2@andrew.cmu.edu, yanda@iu.edu

#### **Abstract**

Despite the success of diffusion models, the training and inference of diffusion models are notoriously expensive due to the long chain of the reverse process. In parallel, the Lottery Ticket Hypothesis (LTH) claims that there exists winning tickets (i.e., a properly pruned sub-network together with original weight initialization) that can achieve performance competitive to the original dense neural network when trained in isolation. In this work, we for the first time apply LTH to diffusion models. We empirically find subnetworks at sparsity 90% - 99%without compromising performance for denoising diffusion probabilistic models on benchmarks (CIFAR-10, CIFAR-100, MNIST). Moreover, existing LTH works identify the subnetworks with a unified sparsity along different layers. We observe that the similarity between two winning tickets of a model varies from block to block. Specifically, the upstream layers from two winning tickets for a model tend to be more similar than the downstream layers. Therefore, we propose to find the winning ticket with varying sparsity along different layers in the model. Experimental results demonstrate that our method can find sparser sub-models that require less memory for storage and reduce the necessary number of FLOPs. Codes are available at https://github.com/osier0524/Lottery-Ticket-to-DDPM.

## 1 Introduction

Diffusion models [37, 19, 39] have achieved state-of-the-art results in a wide range of applications such as image generation [38, 29, 23, 34], text-to-image [35, 30, 33], video generation [43, 20], audio generation [24, 31], and protein generation [44, 41]. These generative models are powerful to produce high-quality data by corrupting the data with slowly increasing noise and then learning to reverse this corruption. For example, Denoising diffusion probabilistic modeling (DDPM) [19] trains a sequence of probabilistic models to reverse each step of the noise corruption.

Although diffusion models have shown impressive performance in capturing distributions and sample quality, they are notoriously slow to generate data due to the long chain of reversing the diffusion process. The noisy data will go through the same U-Net-based generator network thousands of times or even more [19, 42]. At the same time, diffusion models are also notoriously hungry to train. They require many iterations and large size of data to capture the complex data distributions. For example, it takes over two weeks to train DDPM [19] on eight V100 GPUs for  $256 \times 256$  resolution datasets. The reported training time of a state-of-the-art diffusion model in [11] is over 100 days on V100 GPU days to generate high-quality image samples. Moreover, as the image resolution and the size of the training data increases, the training and inference costs grow exponentially. To improve the training efficiency and inference speed, many efficient sampling methods have been proposed, such

<sup>\*</sup>Corresponding author

as DDIM [38], DPM-Solver [26], EDM-Sampling [23]. Different from these fast solvers, we propose to mitigate the computational cost by pruning the reverse model.

The Lottery Ticket Hypothesis (LTH) [14] states that a dense neural network model contains a highly sparse subnetwork (i.e., winning tickets) that can achieve even better performance than the original model. The winning tickets can be identified by training a network and pruning its parameters with the smallest magnitude in an iterative way or one-shot way. Before it is trained in each iteration, the weights will be reset to the original initialization. The identified winning tickets are retrainable to reduce the high memory cost and long inference time of the original neural networks, which has been proved by many works [15, 12, 28, 47, 4]. The existence of winning tickets has been verified in both experiments [27, 40] and theory [45, 36, 3, 13, 28, 9]. LTH has been extended to find the winning tickets for different kinds of neural networks such as GANs [8, 22], Transformers [2, 32, 6, 1] and GNNs [7, 21, 18]. It has also been applied in various domains including computer vision and natural language processing [16, 46, 5]. We for the first time propose to apply LTH to diffusion models. Different from existing works based on efficient sampling, we aim to reduce the number of parameters for efficient training and inference. Specifically, we perform the empirical study to investigate whether there exists a trainable subnetwork of the diffusion model with original initialization that can achieve competitive performance than the original diffusion model. The answer is affirmative. We conclude that the winning tickets achieve the same performance with 90% floating-point operations (FLOPs) saving on the original diffusion model.

We remark that the existing works in LTH identify the winning ticket with a unified sparsity along different layers. That is, we use the same pruning ratio to mask the parameters in the model. In this paper, we empirically found that the similarity between two winning tickets of a given model varies from module to module. Specifically, we introduce centered kernel alignment (CKA) as an index to measure the similarity between the sparsified modules from two winning tickets. We observe that the similarity at the upstream modules is higher than that at the downstream modules. This motivates us to configure the pruning ratio to be different at different modules. Based on the observation, we configure the pruning ratio to be lower at the upstream layer so that the sparsity will be lower for these layers. Intuitively, we need to make sure there are enough parameters to be trained so that meaningful full hidden states can be learned from noisy input data. In the experiment, we verify that our configuration can result in sparser winning tickets without performance compromise.

Since the combination of sparse architectures and initializations in a winning ticket can reveal the potential implications for theoretical study of optimization and generalization in diffusion models, we can take inspiration from winning tickets to design new architectures for the diffusion process, we hope to stimulate the research progress of improving the inference speed of diffusion models.

The contribution of this work can be summarised as:

- We for the first time apply the lottery ticket hypothesis to the diffusion model. Using a pruning method based on magnitude, we identify subnetworks at 99% sparsity in DDMP without performance compromise.
- We propose to identify a winning ticket with a varied sparsity along different layers, which
  is different from existing pruning algorithms in LTH. The proposed method can result in
  winning tickets with higher sparsity.
- The empirical result verifies the quality of pictures generated by a winning ticket is even higher than that generated by the original DDPM.

## 2 Preliminary

We focus on the DDPM in this paper. Given an input  $\mathbf{x_0}$ , the diffusion process gradually adds Gaussian noise based on a variance schedule  $\beta_1,\cdot,\beta_T$ . Denote  $\boldsymbol{\theta}$  as the parameters to learn the distribution  $p_{\boldsymbol{\theta}}(\mathbf{x_{t-1}}|\mathbf{x_t}) = \mathcal{N}(\mathbf{x_{t-1}};\boldsymbol{\mu_{\boldsymbol{\theta}}}(\mathbf{x_t},\mathbf{t}),\sum_{\boldsymbol{\theta}}(\mathbf{x_t},\mathbf{t}))$  in the reverse process.

Given the neural network parameterized by  $\theta$ , a subnetwork is parameterized by  $\theta \odot m$ , where  $\mathbf{m} \in \{\mathbf{0},\mathbf{1}\}^{||\theta||_0}$  is a pruning mask for  $\theta$  and  $\odot$  indicates the element-wise product. We use  $||\cdot||_0$  to represent the  $L_0$  norm counting the number of non-zero elements. The value 0 in  $\mathbf{m}$  means the corresponding parameter  $\theta$  will be masked. The sparsity of a subnetwork is measured as  $1 - \frac{||\mathbf{m}||_0}{||\theta||_0}$ .

Modern pruning methods can be classified into structured pruning and unstructured pruning. In general, structured pruning removes entire groups of neurons, filters, or channels of neural networks

while unstructured pruning results in unstructured sparse matrices. The pruning method based on the magnitude in LTH belongs to the unstructured pruning category. We use an iterative way to find a subnetwork  $\theta_{\tau} \odot m$ , where  $\theta_{\tau}$  is the rewound initialization, which can reach the comparable performance to the full network within a similar training iteration when trained in isolation. After each iteration of training and pruning, we rewind the model with the parameters at  $\tau$  epoch. The combination of  $\theta_{\tau}$  and m with comparable performance is defined as a winning ticket.

# 3 The Existence of Winning Tickets in DDPM

Existing works find the winning ticket by pruning the smallest magnitude in an iterative way. Given a pruning ratio p%, we will sort the magnitude of weights after training and prune p% of parameters with the lowest magnitudes. In practice, existing work prunes the model layer by layer. It will result in a subnetwork where all layers have the same sparsity (p%).

We remark that it is not necessary to find a subnetwork with a unified sparsity along different layers. Intuitively, the input data is highly noisy in the denoising process and we need more parameters to learn a meaningful full hidden state. In this paper, we measure the similarity of two winning tickets based on canonical correlation analysis. Let  $W_i^1$  and  $W_i^2$  be the sparsified weight matrix of ith layer in the first and second winning ticket. We introduce Hilbert-Schmidt Independence Criterion (HSIC) to measure the similarity [17]:  $\mathrm{HSIC}(K,L) = \frac{1}{(n-1)^2} tr(KHLH)$  where  $K_{i,j,k} = k(W_{i,j}^1, W_{i,k}^1)$  and  $L_{i,j,k} = l(W_{i,j}^2, W_{i,k}^2)$ , and H is the centering matrix. Both  $k(\cdot)$  and  $l(\cdot)$  are the RBF kernels. HSIC can be considered as the maximum mean discrepancy between the joint distribution and the product of the marginal distributions [25]. The normalized similarity index is defined as:

$$\mathrm{CKA}(K,L) = \frac{\mathrm{HSIC}(K,L)}{\mathrm{HSIC}(K,K)\mathrm{HSIC}(L,L)}.$$

The effectiveness of this index has been verified by [10]. In Figure 1, we show the CKA between two "Conv2d" modules with the same order in the sequence from two different winning tickets of U-Net. We observe that the similarity is higher at upstream modules. It means there could be a potential implication at these upstream layers.

Motivated by this observation, we propose to configure the pruning ratio to be lower at the upstream modules so that the sparsity will be lower for these modules. Algorithm 1 describes our pruning method. Take the U-Net model as an example. We first train the model parameter as  $\theta_i$  after i iterations. Then for each of J modules (conventional blocks in U-Net) in the sequential list, we prune the parameters with the lowest magnitude. We gradually increase the

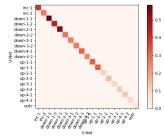


Figure 1: Similarity between two winning tickets

pruning ratio by q as the index of a module increases in the U-Net model. After pruning, we will rewind the parameter to an early stage  $(\tau=5\%*i)$  for the next iteration of pruning. We repeat this process until the desired sparsity  $\delta$  is reached.

#### Algorithm 1 Finding winning tickets for DDPM

**Input**: Initial parameter  $\theta_0$ , initial mask  $\mathbf{m} = \mathbf{1} \in \mathbb{R}^{||\boldsymbol{\theta}||}$ , pruning ratio p, incremental rate q **Output**: Sparsified masks  $\mathbf{m}$ 

```
1: while 1 - \frac{\|\mathbf{m}\|_0}{\|\boldsymbol{\theta}\|_0} < \delta do

2: Train the diffusion model based on gradient \nabla_{\boldsymbol{\theta}} for i iterations

3: Arrive at parameters \boldsymbol{\theta}_i

4: for module j = 0, 1, 2, \cdots, J - 1 do

5: Pruning (p + j * q)\% of the lowest-scored values in jth module \boldsymbol{\theta}^{(j)}

6: creating mask \mathbf{m}^{(j)} for jth module

7: end for

8: Rewinding parameters to \boldsymbol{\theta}_{\tau}

9: \boldsymbol{m} = \{\mathbf{m}^{(0)}, \mathbf{m}^{(1)}, \cdot, \mathbf{m}^{(J-1)}\}

10: end while
```

**Open discussion.** We raise a new question regarding improving the efficiency of reversing: can we use sub-networks with different sparsity in the reverse process? Since a winning ticket can be

considered an equivalent version of the original model, we can leverage a sparser sub-network in the late stage of denoising to further improve efficiency. Intuitively, the noise in the later reverse process has been reduced and it will be easier to optimize. The challenge lies in how to optimize a dense model and a winning ticket while guaranteeing the convergence of training. Another challenge is to decide at which step to use the winning ticket. We leave this open question for future investigation which we hope to stimulate the research on improving the efficiency of the diffusion model.

## 4 Experiment

We conducted experiments to find the winning tickets in DDPM. We use CIFAR-10, CIFAR-100, and MNIST as our benchmark datasets. See the Appendix for more details of the experiment setting. Figure 4 shows the FID score with respect to the sparsity of the pruned U-Net. We observe that a sparsified model can even outperform the original model in terms of FID score. Moreover, by varying the sparsity, we can further reduce the sparsity of a winning ticket. The results verify the existence of winning tickets in DDPM, showing that we can find a winning ticket at sparsity 90% - 99% on the three benchmark datasets.

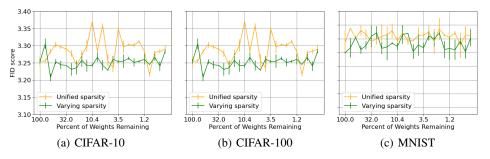


Figure 2: Performance of DDPM w.r.t. sparsity of Unet

We also visualize the quality of pictures generated by the winning ticket. Figure 3 depicts the denoising process of both the winning ticket and the original model on the CIFAR-10 dataset. We can see that the quality of the generated picture is still high when the sparsity is 99.4%. Figure 4 shows samples generated by the winning tickets. It further verifies that a winning ticket can generate a picture with the same quality as the original model. We remark that we reduce 90% of FLOPs with the winning ticket compared with the original model.

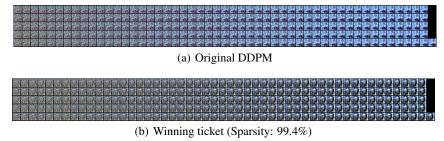


Figure 3: Performance of DDPM w.r.t. sparsity of Unet

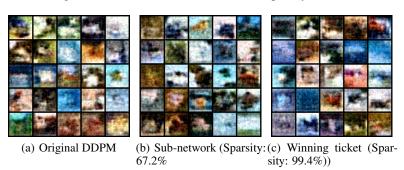


Figure 4: Samples generated on CIFAR-10

#### References

- [1] M. Behnke and K. Heafield. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2664–2674. Association for Computational Linguistics, 2020.
- [2] C. Brix, P. Bahar, and H. Ney. Successfully applying the stabilized lottery ticket hypothesis to the transformer architecture. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 3909–3915. Association for Computational Linguistics, 2020.
- [3] R. Burkholz, N. Laha, R. Mukherjee, and A. Gotovos. On the existence of universal lottery tickets. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.* OpenReview.net, 2022.
- [4] T. Chen, X. Chen, X. Ma, Y. Wang, and Z. Wang. Coarsening the granularity: Towards structurally sparse lottery tickets. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 3025–3039. PMLR, 2022.
- [5] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, M. Carbin, and Z. Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2021, virtual, June 19-25, 2021, pages 16306–16316. Computer Vision Foundation / IEEE, 2021.
- [6] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin. The lottery ticket hypothesis for pre-trained BERT networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual,* 2020.
- [7] T. Chen, Y. Sui, X. Chen, A. Zhang, and Z. Wang. A unified lottery ticket hypothesis for graph neural networks. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1695–1706. PMLR, 2021.
- [8] X. Chen, Z. Zhang, Y. Sui, and T. Chen. Gans can play lottery tickets too. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [9] A. C. W. da Cunha, E. Natale, and L. Viennot. Proving the lottery ticket hypothesis for convolutional neural networks. In *The Tenth International Conference on Learning Representations, ICLR* 2022, *Virtual Event, April* 25-29, 2022. OpenReview.net, 2022.
- [10] M. Davari, S. Horoi, A. Natik, G. Lajoie, G. Wolf, and E. Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.* OpenReview.net, 2023.
- [11] P. Dhariwal and A. Q. Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021.
- [12] J. Diffenderfer and B. Kailkhura. Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning A randomly weighted network. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [13] D. Ferbach, C. Tsirigotis, G. Gidel, and A. J. Bose. A general framework for proving the equivariant strong lottery ticket hypothesis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.* OpenReview.net, 2023.

- [14] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.
- [15] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. The lottery ticket hypothesis at scale. CoRR, abs/1903.01611, 2019.
- [16] Z. Gan, Y. Chen, L. Li, T. Chen, Y. Cheng, S. Wang, J. Liu, L. Wang, and Z. Liu. Playing lottery tickets with vision and language. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 March 1, 2022*, pages 652–660. AAAI Press, 2022.
- [17] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, 16th International Conference, ALT 2005, Singapore, October 8-11, 2005, Proceedings, volume 3734 of Lecture Notes in Computer Science, pages 63–77. Springer, 2005.
- [18] P. Harn, S. D. Yeddula, B. Hui, J. Zhang, L. Sun, M. Sun, and W. Ku. IGRP: iterative gradient rank pruning for finding graph lottery ticket. In S. Tsumoto, Y. Ohsawa, L. Chen, D. V. den Poel, X. Hu, Y. Motomura, T. Takagi, L. Wu, Y. Xie, A. Abe, and V. Raghavan, editors, *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 931–941. IEEE, 2022.
- [19] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.
- [20] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, 2022.
- [21] B. Hui, D. Yan, X. Ma, and W. Ku. Rethinking graph lottery tickets: Graph sparsity matters. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [22] N. M. Kalibhat, Y. Balaji, and S. Feizi. Winning lottery tickets in deep generative models. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 8038–8046. AAAI Press, 2021.
- [23] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.
- [24] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [25] S. Kornblith, M. Norouzi, H. Lee, and G. E. Hinton. Similarity of neural network representations revisited. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.
- [26] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022.
- [27] X. Ma, G. Yuan, X. Shen, T. Chen, X. Chen, X. Chen, N. Liu, M. Qin, S. Liu, Z. Wang, and Y. Wang. Sanity checks for lottery tickets: Does your winning ticket really win the jackpot? In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 12749–12760, 2021.

- [28] E. Malach, G. Yehudai, S. Shalev-Shwartz, and O. Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6682–6691. PMLR, 2020.
- [29] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021.
- [30] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research, pages 16784–16804. PMLR, 2022.
- [31] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR, 2021.
- [32] S. Prasanna, A. Rogers, and A. Rumshisky. When BERT plays the lottery, all tickets are winning. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3208–3229. Association for Computational Linguistics, 2020.
- [33] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.
- [34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.
- [35] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.
- [36] K. Sakamoto and I. Sato. Analyzing lottery ticket hypothesis from pac-bayesian theory perspective. In *NeurIPS*, 2022.
- [37] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015.
- [38] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [39] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [40] J. Su, Y. Chen, T. Cai, T. Wu, R. Gao, L. Wang, and J. D. Lee. Sanity-checking pruning methods: Random tickets can win the jackpot. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.*

- [41] B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, and T. S. Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [42] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou. Diffusion-gan: Training gans with diffusion. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023.
- [43] R. Yang, P. Srivastava, and S. Mandt. Diffusion probabilistic modeling for video generation. CoRR, abs/2203.09481, 2022.
- [44] J. Yim, B. L. Trippe, V. D. Bortoli, E. Mathieu, A. Doucet, R. Barzilay, and T. S. Jaakkola. SE(3) diffusion model with application to protein backbone generation. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 40001–40039. PMLR, 2023.
- [45] Z. Zhang, J. Jin, Z. Zhang, Y. Zhou, X. Zhao, J. Ren, J. Liu, L. Wu, R. Jin, and D. Dou. Validating the lottery ticket hypothesis with inertial manifold theory. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 30196–30210, 2021.
- [46] R. Zheng, B. Rong, Y. Zhou, D. Liang, S. Wang, W. Wu, T. Gui, Q. Zhang, and X. Huang. Robust lottery tickets for pre-trained language models. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 2211–2224. Association for Computational Linguistics, 2022.
- [47] H. Zhou, J. Lan, R. Liu, and J. Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3592–3602, 2019.

## A Appendix

We have introduced three benchmark datasets in the experiment: CIFAR-10, CIFAR-100, and MNIST. A U-Net model is used in the DDPM model. We ran the experiment on a machine with 8 NVIDIA Tesla A100 GPUs. All the training parameters (e.g., training epochs, time steps, and learning rate) are configured as the default of the original DDPM model. We Prune the model for 25 iterations. The default pruning ratio is 20% and the incremental ratio is 1% by default.