Early-Bird Diffusion: Investigating and Leveraging Timestep-Aware Early-Bird Tickets in Diffusion Models for Efficient Training

Supplementary Material

We provide additional supporting experiments and visualizations in this supplementary material. Specifically, Section A presents further empirical evidence for the existence of Early-Bird (EB) tickets in diffusion models across four additional dataset-model pairs. Section B offers additional empirical evidence of Timestep-Aware Early-Bird (TA-EB) tickets using four additional dataset-model pairs. Section C supplies an ablation study to analyze the impact of varying the number of iterations within a pseudo-epoch. Section D showcases additional image generations from the CelebA [4] dataset, further illustrating the qualitative performance of our proposed methods. Section E presents additional experimental results using the Diffusion Transformer (DiT [5]) architecture, demonstrating the versatility of our approach when applied to this widely-used architecture. Section F provides experimental results for different number of timestep regions, showcasing that our method is robust to number or regions. Section G presents an ablation study on the number of iterations chosen for the TA-EB tickets, explaining how we determined the appropriate iteration count for our models.

A. More Results for Finding EB Tickets

To provide further empirical evidence for identifying EB tickets in diffusion models, in Figure 2 (a) we present visualizations of pairwise mask distances across four dataset-model pairs: CelebA [4], LSUN Church [8], and LSUN Bedroom [8], utilizing Denoising Diffusion Probabilistic Models (DDPMs) [3], as well as ImageNet-1K [1], using the Latent Diffusion Model (LDM) [6]. For all settings, we apply magnitude-based structural pruning with a pruning rate of 50%.

Due to the large number of training samples, we adopt a "pseudo-epoch" of 1K steps/iterations to expedite the identification of EB tickets in the case of LSUN Church, LSUN Bedroom, and ImageNet-1K. The (i, j)-th element represents the Hamming distances between the pruned subnetworks extracted at the i-th and j-th pseudo-epochs.

Lighter colors correspond to lower inter-mask Hamming distances, darker colors indicate higher distances. The epoch/pseudo-epoch where the EB ticket is identified is marked in red font. Unless otherwise stated, we use a convergence threshold of $\eta = 0.1$ and a FIFO queue of length 5 for these visualizations. The results demonstrate that EB tickets are consistently identified during the early stages of training. This set of visualizations further empirically confirms the existence of EB tickets in diffusion model training.

Figure 1. More generations from the CelebA [4] dataset. a) Generations from the unpruned model. b) Generations from the "Scratch" model with 50% pruning rate, following the procedure in Section 5.1 of the manuscript. c) Generations from our EB-Diff-Train (EB) method with 50% pruning rate. d) Generations from our EB-Diff-Train (TA-EB) method with 64% average pruning rate.

B. More Results for Finding TA-EB Tickets

To provide further empirical evidence for identifying TA-EB tickets, we present visualizations of pairwise mask distances across four dataset-model pairs in Figure 2 (b): CelebA [4], LSUN Church [8], and LSUN Bedroom [8], all utilizing the DDPM [3], as well as ImageNet-1K [1], using the LDM [6]. For these experiments, we use magnitude-based structural pruning with pruning rates of 30%, 60%, and 80%, respectively. Similar to the case of EB tickets (Section A), we adopt a "pseudo-epoch" approach for LSUN Church, LSUN Bedroom, and ImageNet-1K datasets to identify TA-EB tickets more efficiently. Unless otherwise stated, we use a pseudo-epoch of 1000 steps/iterations.

The (i, j)-th element of each matrix represents the Hamming distance between subnetworks pruned at the i-th and j-th pseudo-epochs across the designated timestep regions. Lighter colors correspond to lower inter-mask Hamming distances, darker colors indicate higher distances. The epoch/pseudo-epoch where the TA-EB ticket is identified is marked in red font. Unless otherwise stated, we use a convergence threshold of $\eta = 0.1$ and a FIFO queue of length 5 for these visualizations. The results show that TA-EB tickets are consistently identified during the early stages of training, further confirming their existence in diffusion model training.

C. Ablation Study on Pseudo-Epoch Choices

We conduct ablation studies to evaluate the impact of the number of iterations per pseudo-epoch on drawn tickets' performance. We employ magnitude-based structural pruning to prune DDPMs [3], using the LSUN Church dataset [8] as a representative case. All training hyperparameters are

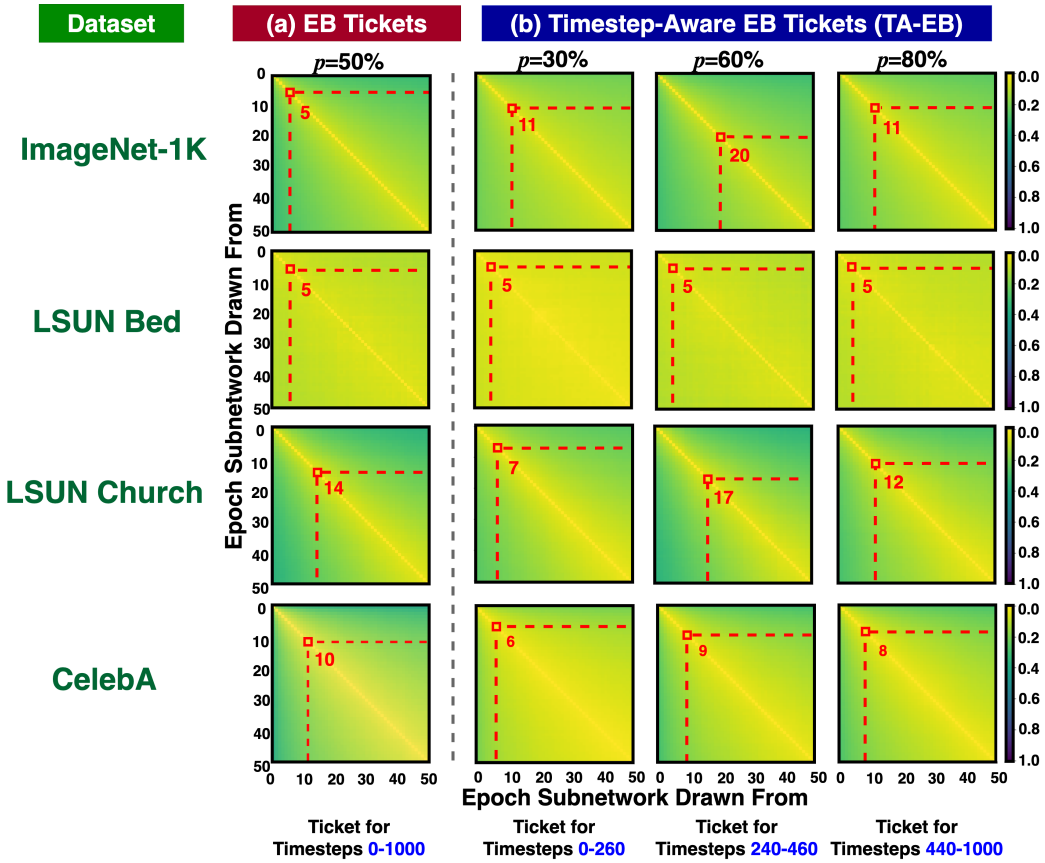

Figure 2. (a) Visualizations of pairwise Hamming distance matrices of EB tickets for the CelebA [4], LSUN Church [8], LSUN Bedroom [8], and ImageNet-1K [1] datasets when using structural magnitude pruning at pruning rate of 50%. (b) Visualizations of Hamming distance matrices of TA-EB tickets when using structural magnitude pruning with pruning rates of 30%, 60%, and 80% across timestep periods of 0-260, 240-460, and 440-1000, respectively.

consistent with those in [2], and we use 100 DDIM [7] timesteps to generate images. Table 1 presents results examining the effect of varying the number of iterations within a single pseudo-epoch, ranging from 100 to 5K. To provide a robust comparison, we also include results for a pruned network trained from scratch, which involves pruning a fully trained network to establish its connectivity and subsequently reinitializing it with random weights for training. As prior work [2] indicates, training pruned networks from scratch is highly competitive, making this a critical baseline for analysis. We observe that the EB tickets identified in our experiments are not sensitive to the choice of pseudo-epoch, as the pseudo-epoch choice marginally impacts the final image generation quality. In all cases, we achieve a 13.88∼19.78 lower FID as well as 3.09× speedups compared to the "scratch" baseline.

Table 1. Ablation study on the choice of iterations in one pseudo-epoch (PE) for the LSUN Church 256×256 dataset using the DDPM [3] model with 100 DDIM timesteps.

LSUN Church 256×256					
Method	**Iters per PE**	**#Params ↓**	**MACs ↓**	**FID ↓**	**Speed-Up ↑**
Unpruned	-	113.7M	248.7G	26.58	1.00×
Scratch	-	28.5M	62.4G	54.18	0.68×
EB-Diff-Train (EB)	100			37.89	2.10×
EB-Diff-Train (EB)	300			40.30	2.10×
EB-Diff-Train (EB)	500			38.05	2.10×
EB-Diff-Train (EB)	1000			37.60	2.10×
EB-Diff-Train (EB)	3000			34.40	2.10×
EB-Diff-Train (EB)	5000			36.90	2.10×

D. Additional Generations

To further qualitatively show the efficacy of our EB-Diff-Train (EB) and EB-Diff-Train (TA-EB) methods, we show additional generations from the CelebA [4] dataset in Fig. 1. We compare our EB and TA-EB methods against two baselines: (1) the original unpruned network ("Unpruned") and (2) a 50% magnitude-based structurally pruned network trained from scratch following the methodology outlined in Section B ("Scratch"). The generations show that our

DiT @ ImageNet-1K 256×256					
Method	**#Params ↓**	**MACs ↓**	**FID ↓**	**Iters**	**Speed-Up ↑**
Unpruned	675.1M	118.7G	28.12	100K	1.00×
Scratch	337.5M	57.0G	58.23	100K	0.72×
EB	337.5M	57.0G	**33.16**	100K	2.63×
TA-EB	242.1M	40.7G	52.10	40K	**3.11**×

EB and TA-EB tickets can generate images of high quality while being up to 6.74× faster than the "Scratch" baseline.

E. Experiments Using the DiT

To highlight that our method can be applied to a variety of model architectures, we include results from the popular Diffusion Transformer (DiT [5]) architecture in Table 2 using the ImageNet-1K [1] dataset. Following the methodology of Section B, we compare against an unpruned network ("Unpruned") and a magnitude-based structurally pruned network trained from scratch ("Scratch"). The results show that our EB-Diff-Train (EB) and EB-Diff-Train (TA-EB) methods can be successfully applied to the DiT architecture, yielding reductions in FID scores of $0.13 \sim 19.07$ while also improving training speed by $3.65\times \sim 4.32\times$ respectively, as compared to the "Scratch" baseline.

F. Ablation Study on Number of Regions

To showcase that our method is robust to the number of regions selected, we test 2, 3, and 4 regions, following the same settings as the submitted manuscript, in Table 2. For the 2 region case, we merged the first two timestep regions; for the 4 region case, we split the last region in half. The pruning rate was adjusted to maintain a timestep-weighted average of $\sim 64\%$. From this ablation study, we see that our method is robust to the number of regions selected.

G. Ablation Study on Training Iterations

To determine the most suitable training iteration under the same pruning rate, we summarize how generation quality (e.g., FID scores) changes with increased training in Table 3. Specifically, the generation quality improves significantly, with FID scores decreasing by 1.06 when training iterations increase from 100K to 200K. Beyond 200K iterations, however, the generation quality gradually declines, suggesting an optimal training window between 100K and 200K iterations. The decline in generation quality after 200K likely reflects overfitting to specific timestep regions.

Table 2. Results for CIFAR-10 32×32 using the DDPM [3] model with 100 DDIM steps under different region numbers. The number of iterations for the TA-EB methods are recorded as the total number of iterations for the subnetwork of longest training time

DDPM @ CIFAR10 32×32						
#Regions	**Avg. Pruning Rate**	**#Params ↓**	**MACs ↓**	**FID ↓**	**Iters**	**Speed-Up ↑**
1 Region	0%	35.7M	6.1G	5.15	800K	1.00×
2 Regions	64.2%	7.4M	1.4G	7.37	200K	6.30×
3 Regions	64.0%	7.2M	1.3G	7.29	200K	5.78×
4 Regions	64.0%	7.2M	1.3G	7.55	200K	5.78×

Table 3. Effect of extending training iterations and time on an A100 GPU under a 30% pruning rate for DDPM on CIFAR-10.

Training Time (hr)	**Training Iterations**	**FID↓**
3.1	100K	7.95
6.2	**200K**	**6.89**
9.3	300K	7.02
12.4	400K	7.19
15.5	500K	7.44

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2, 3

[2] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems*, pages 16716–16728, 2023. 2

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1, 2, 3

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, pages 3730–3738, 2015. 1, 2

[5] William Peebles et al. Scalable diffusion models with transformers. *arXiv*, 2022. 1, 3

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 1

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[8] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 2