Appendix

A Additional Experiments

Class Conditioning. As both GET and ViT share the same class injection interface, we perform an ablation study on ViT. We consider two types of input injection schemes for class labels: 1) additive injection scheme 2) injection with adaptive layer normalization (AdaLN-Zero) as used in DiT [25]. Despite using almost the same parameters as unconditional ViT-B, the class-conditional ViT-B using additive injection interface has an FID of 12.43 at 200k, while the ViT-B w/ AdaLN-Zero class embedding [25] set up an FID of 17.19 at 200k iterations. Another surprising observation is that ViT-B w/ AdaLN-Zero class embedding performs worse than unconditional ViT in terms of FID score. Therefore, it seems that adaptive layer normalization might not be useful when used only with class embedding.

Table 6: Ablation on class conditioning.

Model	FID↓	IS↑	Params↓
ViT-Uncond	15.20	8.27	85.2M
ViT-AdaLN-Zero	17.19	8.38	128.9M
ViT-Inj-Interface	12.43	8.69	85.2M

B Related Work

Transformers. Transformers were first proposed by Vaswani et al. [36] for machine translation and since then have been widely applied in many domains like natural language processing [10, 19, 28, 33], reinforcement learning [9, 24], self-supervised learning [8], vision [12, 21], and generative modeling [13, 15, 25, 31]. Many design paradigms for transformer architectures have emerged over the years. Notable ones include encoder-only [10, 18, 20], decoder-only [7, 28, 29, 37, 38], and encoder-decoder architectures [17, 30, 36]. We are interested in scalable transformer architectures for generative modeling. Most relevant to this work are two encoder-only transformer architectures: Vision Transformer (ViT) [12] and Diffusion Transformer (DiT) [25]. Vision Transformer (ViT) closely follows the original transformer architecture. It first converts 2D images into patches that are flattened and projected into an embedding space. 2D Positional encoding is added to the patch embedding to retain positional information. This sequence of embedding vectors is fed into the standard transformer architecture. Diffusion Transformers (DiT) are based on ViT architecture and operate on sequences of patches of an image that are projected into a latent space through an image encoder [34]. In addition, DiTs adapt several architectural modifications that enable their use as a backbone for diffusion models and help them scale better with increasing model size, including adaptive Layer Normalization (AdaLN-Zero) [6, 11, 16, 26] for time and class embedding, and zero-initialization for the final convolution layer [14].

Deep equilibrium models. Deep Equilibrium models (DEQs) [2] solve for a fixed point in the forward pass. Specifically, given an input $\mathbf{x}$ and a layer or a block f_θ, DEQ approximates an infinite-depth representation of f_θ by solving for its fixed point $z^\star$: $z^\star = f_\theta(z^\star; \mathbf{x})$. For the backward pass, one can differentiate analytically through $z^\star$ by the implicit function theorem. DEQs do not have any convergence guarantees and can be highly unstable to train [4]. As a result, recent efforts focus on addressing these issues by designing variants of DEQs with provable guarantees [32, 39], or through optimization techniques such as Jacobian regularization [4], and fixed-point correction [5]. DEQs have been successfully applied on a wide range of tasks such as image classification [3], semantic segmentation [3, 40], optical flow estimation [5], landmark detection [23], out-of-distribution generalization [1], language modelling [2], unsupervised learning [35], and generative modelling [22, 27].

C Model Configuration

We set the EMA momentum to 0.9999 for all the models.

The configuration of different GET architectures are listed in Table 7. Here, L_i and L_e denote the number of transformer blocks in the Injection transformer and Equilibrium transformer, respectively. D denotes the width of the network. E corresponds to the expanding factor of the FFN layer in the Equilibrium transformer, which results in the hidden dimension of $E \times D$. For the injection transformer, we always adopt an expanding factor of 4.

Table 7: Details of configuration for GET architectures.

Model	Params	L_i	L_e	D	E
GET-Tiny	8.9M	6	3	256	6
GET-Mini	19.2M	6	3	384	6
GET-Small	37.2M	6	3	512	6
GET-Base	62.2M	1	3	768	12
GET-Base+	83.5M	6	3	768	8

We have listed relevant model configuration details of ViT in Table 8. The model configurations are adopted from DiT [25], whose effectiveness was tested for learning diffusion models. In this table, L denotes the number of transformer blocks in ViT. D stands for the width of the network. We always adopt an expanding factor of 4 following the common practice [12, 25, 36].

Table 8: Details of configuration for ViT architectures.

Model	Params	L	D
ViT-B	85.2M	12	768
ViT-L	302.6M	24	1024

References

[1] Cem Anil, Ashwini Pokle, Kaiqu Liang, Johannes Treutlein, Yuhuai Wu, Shaojie Bai, J Zico Kolter, and Roger B Grosse. Path independent equilibrium models can better exploit test-time computation. *Advances in Neural Information Processing Systems*, 35:7796–7809, 2022. 1

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Neural Information Processing Systems (NeurIPS)*, 2019. 1

[3] Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. *Advances in Neural Information Processing Systems*, 33:5238–5250, 2020. 1

[4] Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Stabilizing Equilibrium Models by Jacobian Regularization. In *International Conference on Machine Learning (ICML)*, 2021. 1

[5] Shaojie Bai, Zhengyang Geng, Yash Savani, and J Zico Kolter. Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 620–630, 2022. 1

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[9] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 1

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems (NeurIPS)*, 2021. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

[15] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *International Conference on Machine Learning (ICML)*, 2021. 1

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[17] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019. 1

[18] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 1

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 1

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[22] Cheng Lu, Jianfei Chen, Chongxuan Li, Qiuhao Wang, and Jun Zhu. Implicit normalizing flows. *arXiv preprint arXiv:2103.09527*, 2021. 1

[23] Paul Micaelli, Arash Vahdat, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Recurrence without recurrence: Stable video landmark detection with deep equilibrium models. *arXiv preprint arXiv:2304.00600*, 2023. 1

[24] Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020. 1

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1, 2

[26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018. 1

[27] Ashwini Pokle, Zhengyang Geng, and J Zico Kolter. Deep equilibrium approaches to diffusion models. *Advances in Neural Information Processing Systems*, 35:37975–37990, 2022. 1

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[32] Max Revay, Ruigang Wang, and Ian R Manchester. Lipschitz bounded equilibrium networks. *arXiv preprint arXiv:2010.01732*, 2020. 1

[33] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019. 1

[34] 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*, pages 10684–10695, 2022. 1

[35] Russell Tsuchida and Cheng Soon Ong. Deep equilibrium models as estimators for continuous latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1671. PMLR, 2023. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2

[37] Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, pages 22964–22984. PMLR, 2022. 1

[38] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 1

[39] Ezra Winston and J Zico Kolter. Monotone operator equilibrium networks. *Advances in neural information processing systems*, 33:10718–10728, 2020. 1

[40] Sai Zhang, Liangjia Zhu, and Yi Gao. An efficient deep equilibrium model for medical image segmentation. *Computers in Biology and Medicine*, 148:105831, 2022. 1