# GAN-KAN: Kolmogorov Arnold Networks for Preventing Mode Collapse in Generative Adversarial Networks

**Evan Smith** [1] **Kai Davidson** [1]

## Abstract

In this paper, we compare the use of KAN (Liu et al., 2025), MLP, and GR-KAN (Yang and Wang, 2024) as the first and final layer in a convolutional GAN architecture on the MNIST (Deng, 2012) and CIFAR-10 (Krizhevsky, 2009) datasets. The motivation behind this is previous research on KANs that suggests that they do not suffer from "catastrophic forgetting" as much as MLP networks (Liu et al., 2025). Theoretically, catastrophic forgetting has been linked to mode collapse in GANs (Thanh-Tung and Tran, 2020). For this reason, they may present a viable alternative to MLP in GANs to prevent mode collapse. Our results showed that the KAN and GR-KAN networks were comparable, are suffered frequently from mode collapse in both datasets. With the modification of the WGAN-GP training method (Gulrajani et al., 2017), the CIFAR-10 MLP performance slightly improved, though there were no drastic improvements for KAN or GR-KAN. From our experiments, we concluded that because of the difficulty in training GANs, KAN and GR-KAN do not improve network performance for GANs drastically without further fine-tuning. Here is the Github for the project.

## 1. Introduction

Kolmogorov Arnold Networks (KANs) are an alternative to Multi Layer Perceptrons (MLPs) and can feasibly replace a linear layer in many different architectures (Liu et al., 2025). Earlier in this course, we presented on Kolmogorov Arnold Transformers (Yang and Wang, 2024) which introduces the concept of GR-KAN which is a mix of KAN and MLP networks by having a smaller number of tunable rational activation functions that group MLP edges together. KANs have been shown to mitigate catastrophic forgetting (Liu et al., 2025), which is theorized to cause issues in GAN training (Thanh-Tung and Tran, 2020). in certain cases and have intrinsic benefits over MLP in some tabular datasets (Poeta et al., 2024). More evidence for KANs in image related tasks has been done in Image-Image Translation with promising results for the KAN-CUT model (Mahara et al., 2024). GANs can suffer from the problem of mode collapse where the model only generates a few specific classes or a singular example of a given class, which is related to catastrophic forgetting (Thanh-Tung and Tran, 2020). Our question is whether KANs can mitigate mode collapse through their intrinsic learning patterns. The GitHub for this project can be found at github.com/EvanGibsonSmith/KAN-Mode-Collapse, and the google drive of our models and figures can be found here.

### 1.1. Research contributions

We looked into KAN and GR-KAN as alternatives to MLP in GAN architectures. Our goal is to see if they are effective towards mitigating mode collapse in GANs. Through comparing between MLP, KAN, and GR-KAN techniques we hope to see which is the most effective at augmenting or mitigating the performance of GANs. GANs are known to be difficult to train (Thanh-Tung and Tran, 2020) so by integrating a specific network that combats this problem, we find if this meaningfully affects the results of GAN generation to mitigate mode collapse.

## 2. Proposed Method

We used a convolutional GAN architecture that consisted of a generator and discriminator with comparable architectures, each swapping a linear layer with MLP, KAN or GR-KAN layer (See Figure 1). The generator had the first layer replaced and the discriminator the last layer. Each of these architectures was implemented in PyTorch with slight modifications to the input and outputs for each dataset due to the image dimension differences. The reason for the simplified model is due to training time constraints on each model, KAN layers take more computational power to train so reducing the changes drastically reduces training time. Addi-

[1]Worcester Polytechnic Institute. Correspondence to: Evan Smith <egsmith@wpi.edu>, Kai Davidson <kmdavidson1@wpi.edu>.

*CS/DS 552 Generative Artificial Intelligence Final Project*, Worcester Polytechnic Institute. 

tional information about training parameters of the model is in the experiments section.

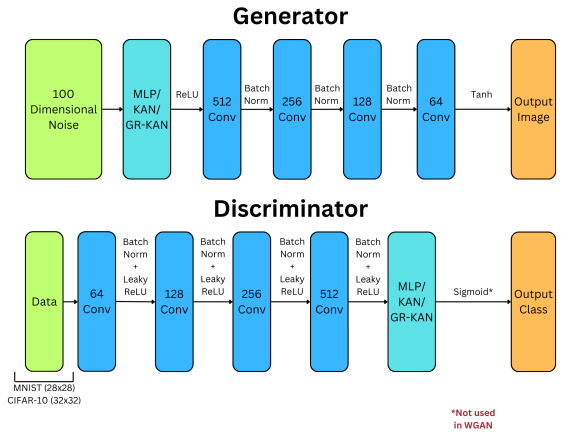


*Figure 1.* GAN Generator and Discriminator architecture with swapped linear layers.

### 2.1. Datasets

For this project we used both the MNIST (Deng, 2012) and CIFAR-10 (Krizhevsky, 2009) datasets. The MNIST dataset served as a simpler task for the generator to learn, yet can still suffer from mode collapse. The CIFAR-10 dataset has multiple color channels which makes it a more difficult generation task, as well as having varied classes of different objects rather than numbers. Choosing a regular GAN to generate data with classes, as opposed to a network like CGAN, is to more easily identify mode collapse in or across classes. This was done using well-trained models on each dataset to detect class representation, t-SNE, and PCA. These metrics allowed for quantifiable mode collapse for each model type.

## 3. Experiments

### 3.1. Preliminary Tests

To get an initial idea of the power of MLP versus KAN's for GANs, we used the GR-KAN network as seen in (Yang and Wang, 2024) and compared it to an MLP network both as discriminators. We wanted to plot the number of parameters versus the accuracy which is what developed these heatmaps that show accuracy compared to number of layers and hidden layer sizes (Figures 2 and 3). The results showed that GR-KANs are unstable when training and had seemingly random training collapse especially in smaller parameter models. This lead us to look at original KANs for our methodology as well as further methods to reduce the instability when training GANs, due to already dealing with models that could have unstable training. Due to the original nature of GR-KAN layer, which was in a transformer architecture, it could be a uniquely worse task to use it for image related tasks, which could have contributed to our later results.

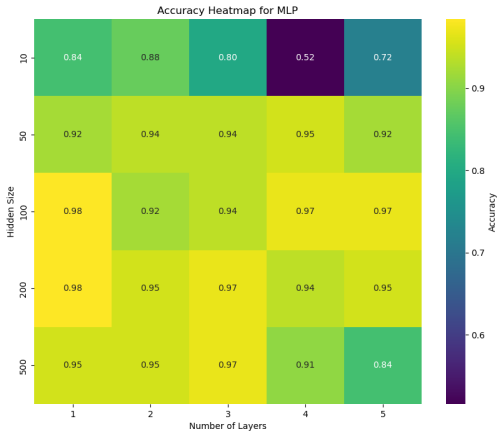


*Figure 2.* MLP heatmap of number of parameters and layers versus accuracy

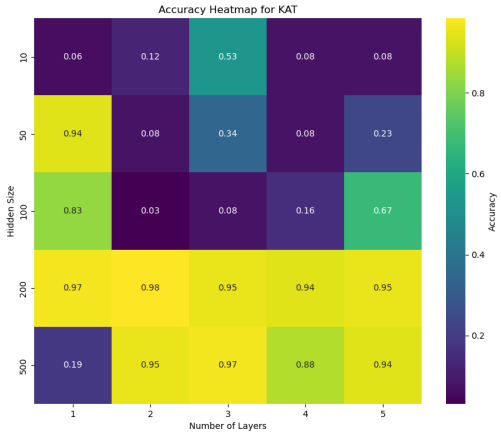


*Figure 3.* GR-KAN heatmap of number of parameters and layers versus accuracy

### 3.2. Methodology

The experiments consisted of running the GAN architecture and varying the linear layer between MLP, KAN, or GR-KAN. An additional variation added after finding our initial results was the WGAN-GP modification to the GAN training. It uses Wasserstein distance, or earth mover's distance, to maintain good gradient flow when the discriminator is very strong. This resulted in a few different training

| Architecture | GAN | WGAN-GP |
|---|---|---|
| Discriminator LR | $2 \times 10^{-5}$ | $1 \times 10^{-4}$ |
| Generator LR | $2 \times 10^{-4}$ | $1 \times 10^{-4}$ |
| Epochs | 100 | 100 |
| Batch Size | 64 | 64 |
| Validation Size | 100 | 100 |
| Optimizer | ADAM | ADAM |

*Table 1.* Parameters for training GAN and WGAN-GP models

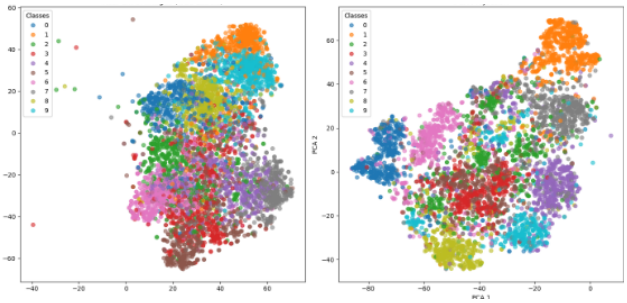

*Figure 4.* Real Data t-SNE vs. MLP GAN MNIST t-SNE

parameters between the two architectures, as seen in Table 1.

## 4. Results

### 4.1. MNIST

On the simpler MNIST dataset with a vanilla GAN, both generators with initial MLP and GR-KAN layers yielded convincing results with the outlier being the KAN network which displayed evidence of mode collapse. To quantitatively compare these models, the Fréchet Inception Score (FID) (Heusel et al., 2018) can be used to determine the similarity between the distribution of the real images and generated images to understand overall quality of the dataset. With the highest FID score of 91.8 (see Table 3) and a latent t-SNE space that shows the mode collapse visually (see Figure 5).

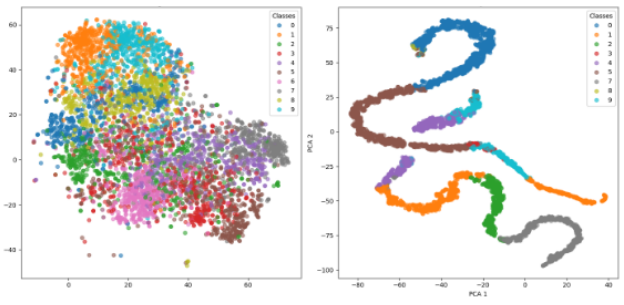

*Figure 5.* Real Data t-SNE vs. KAN GAN MNIST t-SNE

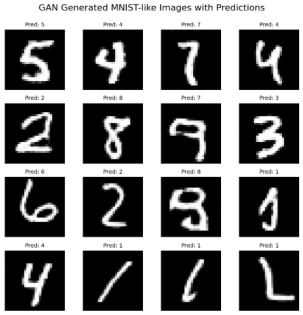


*Figure 6.* Collage of Generated MNIST Results from MLP initial layer

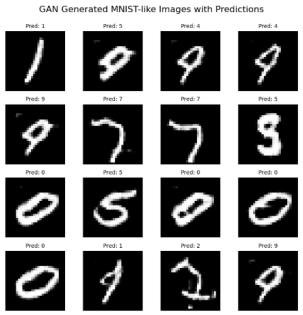


*Figure 7.* Collage of Generated MNIST Results from KAN initial layer.

As seen in the generated collages (Figures 6 and 7) the MLP GAN layer on MNIST creates fairly realistic looking generated digits with the lowest FID score among all model, acting as the baseline which was not surpassed throughout further testing of architectures. The KAN GAN MNIST shows mode collapse more obviously in the generated images (Figure 7) with similarities among each generated number for the classes 0, 9, and 7.

Based on these 16 images, more mode collapse can be seen qualitatively with the KAN final layer, both intra-class within a single class (i.e each of the 7s and 0s generated look very similar) and inter-class between different classes (i.e. We see mostly 7s and 0s).

The MLP final layer does a marginally better job representing the overall distribution between classes than KAN. Because these models were trained to generate hand written digits without knowledge of the specific classes 0-9, our knowledge of these classes from the dataset can be used to examine intra-class and inter-class mode collapse. Inter-mode collapse can be seen in the percentage of the outputs that are classified as each distribution in the t-SNE plots, where often a single class is generated most of the time, with some classes often not being generated at all. The intra-mode collapse can be seen in the tight clustering in the t-SNE reduction. This indicates that the generations for

that class are very similar to each other.

Table 2 shows the FID scores for the MLP, KAN, and GR-KAN over each of the classes. With a trained MNIST classifier, each of the generated results were categorized and then the FID of this category was compared to the total distribution. Naturally, these FIDs will always be worse than the overall FID. In comparing them to each other, a higher FID for an individual class indicates that the generated images of that class are more or less like to total distribution.

| Class | Architecture FID | | |
|---|---|---|---|
| | MLP Conv | KAN Conv | GR-KAN Conv |
| Overall | 5.92 | 91.80 | 7.53 |
| 0 | 106.15 | 198.75 | 105.77 |
| 1 | 115.16 | 140.31 | 114.04 |
| 2 | 42.36 | 102.47 | 42.13 |
| 3 | 67.32 | 129.58 | 66.81 |
| 4 | 90.68 | 120.90 | 89.03 |
| 5 | 61.91 | 112.17 | 60.69 |
| 6 | 53.72 | 120.75 | 54.21 |
| 7 | 65.18 | 99.00 | 64.80 |
| 8 | 75.16 | 156.18 | 75.18 |
| 9 | 49.30 | 121.09 | 49.53 |

*Table 2.* Comparison of FID on MLP, KAN, and GR-KAN between class of generated result and real images across MNIST classes 0-9 for Vanilla GAN implementation.

The CIFAR-10 dataset was then used as more of a challenge for the generator network, and KAN initially produced slightly better FID results than the MLP layer. Unfortunately the GR-KAN had mode collapse as seen in Figure 8. Once again, the t-SNE plot shows specifically the intra-mode collapse that occurs compared to the real dataset.

Due to these results we looked for GAN training strategies

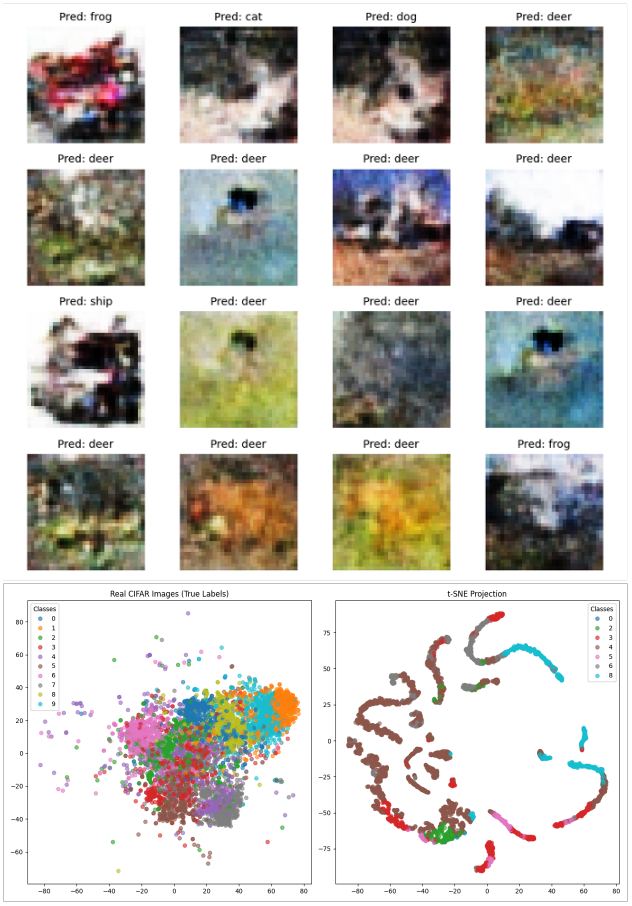


*Figure 8.* GR-KAN CIFAR-10 mode collapse, shown in the generations and t-SNE embedding space compared to real image t-SNE embedding space.

| Dataset | GAN Type | NN Type | FID |
|---|---|---|---|
| CIFAR | Vanilla | MLP | 90.87 |
| CIFAR | Vanilla | KAN | 89.30 |
| CIFAR | Vanilla | GR-KAN | 203.71* |
| **MNIST** | Vanilla | MLP | **5.92** |
| MNIST | Vanilla | KAN | 91.80* |
| MNIST | Vanilla | GR-KAN | 7.53 |
| **CIFAR** | WGAN-GP | MLP | ***79.21*** |
| CIFAR | WGAN-GP | KAN | 265.30* |
| CIFAR | WGAN-GP | GR-KAN | 192.66* |
| MNIST | WGAN-GP | MLP | 7.13 |
| MNIST | WGAN-GP | KAN | 6.97 |
| MNIST | WGAN-GP | GR-KAN | 339.61* |

*Table 3.* Model FID scores for each dataset, GAN Type, and neural network type, * = Suffered from Mode Collapse

that could improve our performance and hopefully boost the effectiveness of GR-KAN and KAN layers so that they can properly train even when faced with a powerful discriminator. For this reason we modified our training to use WGAN method instead (Arjovsky et al., 2017). We used the improved gradient penalty version from Gulrajani et al.. The results from all of these modifications can be seen in Table 3. This improved our performance with MLP on CIFAR-10 using WGAN-GP training leading it to be the best performing once again. Unfortunately for both GR-KAN and KAN on CIFAR-10 with WGAN-GP mode collapse was caused during training as seen by the abnormally high FID scores. Finally for MNIST with WGAN-GP we didn't find any improvements from the previous vanilla GAN approach, though only GR-KAN had mode collapse for this method. The mode collapses for GR-KAN can be seen in Figures 9 and 10.

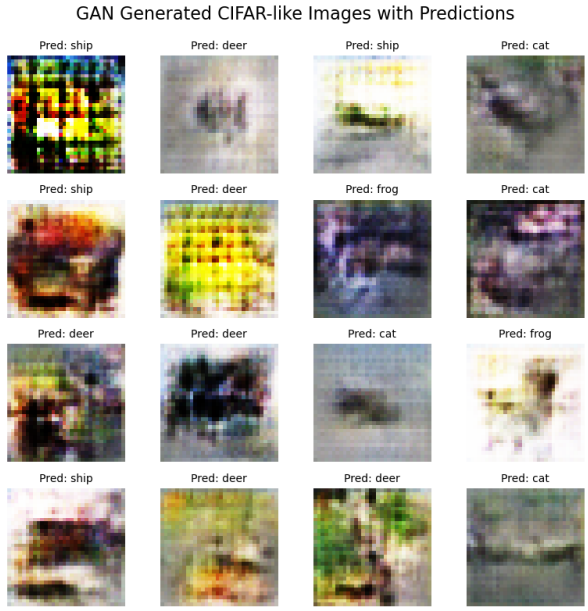


*Figure 9.* GR-KAN Layer with WGAN architecture generating CIFAR images, resulting in mode collapse.

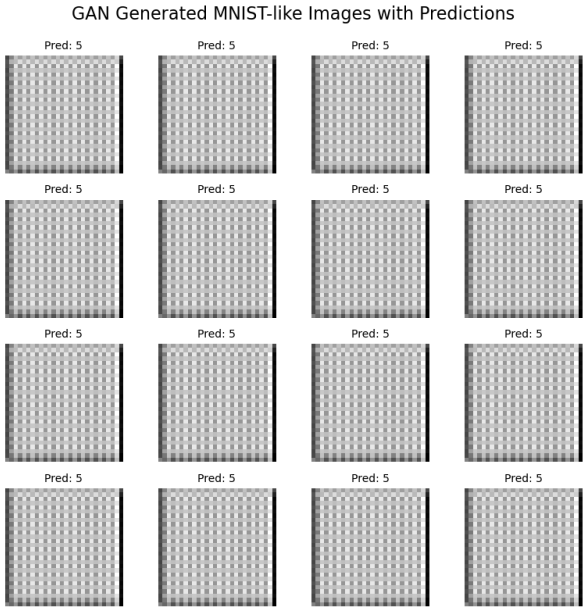


*Figure 10.* GR-KAN Layer with WGAN architecture generating MNIST images, extreme mode collapse caused most likely by unstable training as seen in preliminary experiments.

## 5. Discussion

Based on these results, there is limited evidence for KAN and GR-KANs for GANs. Based on the results the CIFAR and MNIST over many hyperparameter and with both the Vanilla GAN and WGAN-GP there is no evidence they perform better *overall* or perform better in balancing class generation. While these results are preliminary and inconclusive, KANs and GR-KANs seem somewhat more likely to mode collapse and somewhat more unstable than their MLP counterpart. There was no improvement in the context of WGAN-GP with KANs and GR-KANs, as theoretically desired because of the improved gradient flow in the WGAN-GP paired with the resilience of KANs to mode collapse.

Due to the constraints of this project, our training also had some limitations. The models could have been trained more than 100 epochs and none of them were trained until loss convergence, just the same amount of epochs. Most of these limitations were due to time so some simplification was done to allow for comparison and training of all models.

## 6. Conclusions and Future Work

In the context of GANs, through the limited research of this project we found limited success from KANs or the GR-KAN variant when compared to MLP. We did not find that the KANs performed significantly worse, but that they were prone to not training in the case the discriminator learned too quickly. To alleviate this issue we compared KANs in the context of a WGAN-GP and found the same general patterns emerged, despite the possibility for the WGAN-GP to mitigate gradient flow issues from an over fitted or very powerful discriminator.

In the future, KANs could be applied to other promising GAN architectures, particularly the R3GAN (Huang et al., 2025). Choosing a more KAN friendly dataset such as WaveBench (Liu et al., 2024) could have improved the results for a KAN-GAN model compared to MLP-GAN as well.

## References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, December 2017. URL http://arxiv.org/abs/1701.07875. arXiv:1701.07875 [stat].

Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6):141–142, November 2012. ISSN 1558-0792. doi: 10.1109/MSP.2012.2211477. URL https://ieeexplore.ieee.org/document/6296535.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs, December 2017. URL http://

arxiv.org/abs/1704.00028. arXiv:1704.00028 [cs].

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs].

Yiwen Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The GAN is dead; long live the GAN! A Modern GAN Baseline, January 2025. URL http://arxiv.org/abs/2501.05441. arXiv:2501.05441 [cs].

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. April 2009.

Tianlin Liu, Jose Antonio Lara Benitez, Florian Faucher, AmirEhsan Khorashadizadeh, Maarten V. de Hoop, and Ivan Dokmanić. Wavebench: Benchmarking data-driven solvers for linear wave propagation PDEs. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=6wpInwnzs8.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-Arnold Networks, February 2025. URL http://arxiv.org/abs/2404.19756. arXiv:2404.19756 [cs].

Arpan Mahara, Naphtali D. Rishe, and Liangdong Deng. The Dawn of KAN in Image-to-Image (I2I) Translation: Integrating Kolmogorov-Arnold Networks with GANs for Unpaired I2I Translation, August 2024. URL http://arxiv.org/abs/2408.08216. arXiv:2408.08216 [cs].

Eleonora Poeta, Flavio Giobergia, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. A Benchmarking Study of Kolmogorov-Arnold Networks on Tabular Data. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6, September 2024. doi: 10.1109/AICT61888.2024.10740444. URL http://arxiv.org/abs/2406.14529. arXiv:2406.14529 [cs].

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in GANs. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, July 2020. doi: 10.1109/IJCNN48605.2020.9207181. URL https://ieeexplore.ieee.org/document/9207181/. ISSN: 2161-4407.

Xingyi Yang and Xinchao Wang. Kolmogorov-Arnold Transformer, September 2024. URL http://arxiv.org/abs/2409.10594. arXiv:2409.10594 [cs].