Diffusion Models and Rendering of Text

By Paul Thompson, June 12, 2025

Why are Diffusion Models so bad at rendering text (and how could you fix them?).

Have you ever puzzled out how to get GPT-4o (or any diffusion model) to render images with correct nice-looking text ?

Every time text appears, I get random spelling glitches (e.g. *"cloccks"* in this example). Retry leads to new errors.

This is a known failure mode for diffusion models, and although they do train on some images with text in them (LAION [1], [2], etc.), the perceptual losses (used to decide if the output is good) don’t penalize tiny deformations that mess up legibility of text for us

Are DeepFloyd ( [3], [4] ) or Imagen ([5], [6]) better ? Any automatic solutions (other than me typing on images like I would in PPT) ?

\*\*How the Fisher spectrum reveals effective complexity in (a subclass of) natural images:

Training a VAE deep neural network to be able to tell the time on one brand of digital clocks produces a sharply peaked Fisher spectrum: only a few latent directions capture meaningful variation (digit identity, minor distortions) but most parameters are unused (flat directions).

When trained on multiple 7-segment to n-segment digital clock brands, the model needs to account for font style, segment shapes, brightness+other factors, gradually filling out more directions in latent space.

The Fisher spectrum becomes heavier-tailed as true data complexity grows, but still leaves many flat directions unused.

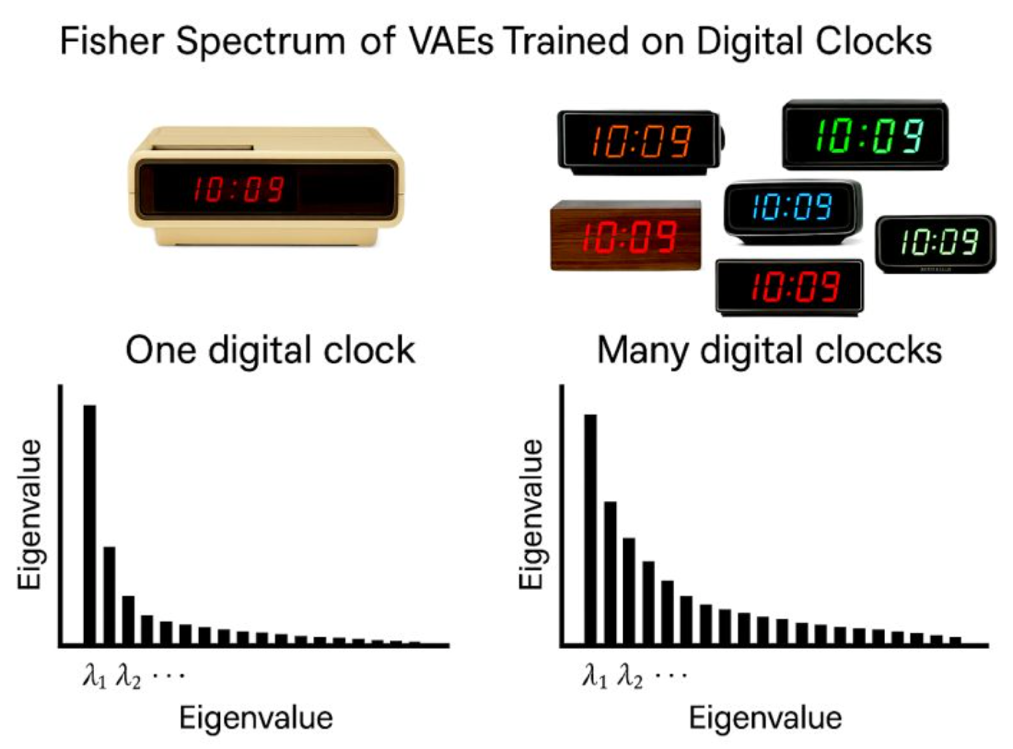
These flat directions compress the posterior volume, reducing effective complexity despite overparameterization. (FIM spectrum analysis [7] inspired by last week's cool paper by Frank Nielsen [8])

1 ) The Fisher spectrum for an encoder-decoder VAE that can draw text correctly is super sharp - you need a very precise latent direction for each letter.

if diffusion models wobble even a tiny bit, we get *"cloccks"* instead of *"clocks"* + sometimes unreadable text ... Good text-rendering DiTs would have extremely sharp Fisher spectrum directions because tiny parameter changes lead to massive perceptual failure

2 } Bonus problem: can you design a diffusion model(or just its encoder-decoder) to render text nicely? You'd want to add an auxiliary OCR model to identify any rendered text and decide what it says using an OCR loss. And then in those text regions you'd want an trace penalty to encourage sharp spectrum, . This Fisher loss pushes the model to allocate capacity to sharp text submanifolds.

We need more BIG models trained for typography.



[1] LAION: <https://laion.ai/blog/>

[2] LAION LeoLM: Lingustically Enhanced Open Language Model on HuggingFace: <https://huggingface.co/LeoLM>

[3] [Stability AI releases DeepFloyd IF, a powerful text-to-image model that can smartly integrate text into images, April, 2023](https://stability.ai/news/deepfloyd-if-text-to-image-model)

[4] <https://github.com/deep-floyd/IF>

[5] Imagen: <https://deepmind.google/models/imagen/>

[6] [Instruct-Imagen: Image Generation with Multi-modal Instruction, H. Hu et al, Google DeepMind, 2024](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/articles/generative_models/Instruct-Imagen-Image_Generation_with_Multi-modal_Instruction_Hu_2024.pdf)

[7] [Musings on the Latest Generative AI Developments by Paul Thompson, 6/7/2025](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/docs/Musings_on_the_Latest_Generative_AI_developments_by_Paul_Thompson.docx)

[8] [Geometric Modeling of Occam’s Razor in Deep Learning, Ke Sun, Frank Nielsen, 2025](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/geometric_deep_learning/A_Geometric_Modeling_of_Occams_Razor_in_Deep_Learning_Sun_2025.pdf)

[9] [Hints, Yaser-Abu Mostafa, 1995](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/inductive_bias/Hints_Yaser_Abu_Moustafa_1995.pdf)

[10] [The Lack of Apriori Distinctions Between Learning Algorithms, David H. Wolpert, 1996](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/inductive_bias/the-lack-of-a-priori-distinctions-between-learning-Wolpert_1996.pdf)

[11] [The Role of Occam's Razor in Knowledge Discovery, Pedro Domingos, 1999](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/inductive_bias/The_Role_of_Occams_Razor_in_Knowledge_Discovery_Domingos_1999.pdf)

[12] [Ockham’s Razor, Truth, and Information, Kevin T. Kelly, CMU, 2008](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/inductive_bias/Ockhams_Razor_Truth_and_Information_Kelly_2008.pdf)