Factorized Diffusion

Everett Lee, Sahib Manjal, Ryan Noonan, Andrew Richmond, Chad Yu

`https://github.com/chadyu9/factorized-diffusion`

1 Introduction

In this project, we reimplemented "Factorized Diffusion," a technique developed by Geng, Park, and Owens (2025) that enables the creation of perceptual illusions through controlled image generation. We focused specifically on reproducing their hybrid image results—images that change appearance depending on viewing distance, revealing different content when viewed up close versus from afar. Our implementation utilized the DeepFloyd IF diffusion model to factorize noise estimates during the denoising process. By decomposing images into frequency components and conditioning these components on different text prompts, we successfully generated hybrid images without requiring specialized fine-tuning or additional networks. This reimplementation demonstrates how existing generative models can be leveraged to create compelling visual illusions through careful manipulation of the diffusion process.

2 Chosen Result

In this re-implementation of Factorized Diffusion, our objective was to reproduce the hybrid image generation results utilizing publicly available diffusion models.

3 Methodology

We used DeepFloyd IF[1], a pixel space diffusion model integrated through Hugging Face's Diffusers library.

At each denoising step of the diffusion process, we did two parallel passes of the noisy pixel representation through the DeepFloyd model, one pass conditioned on each prompt. With the resulting noise estimates, we then factorized those estimates separately, recombined the results, and used that as our final noise estimate to complete the denoising step. To perform the factorization of the noise estimates we relied on different factorization processes. For the Frequency Hybrids, the factorization process involved applying a Fast Fourier Transform to each individual noise estimate to separate high and low frequency components, extracting only the component that we wanted to take from each prompt, and then using those components to recombine. Mathematically, this means for a decomposition of an image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ into N components in pixel space

$$\mathbf{x} = \sum_{i=1}^{N} f_i(\mathbf{x})$$

where each component corresponds to a different prompt y_i we have that at the update step in the denoising process where we are trying to reconstruct $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$, we replace the noise estimate $\epsilon_t\epsilon_\theta(\mathbf{x}_t, y, t)$ with N individual noise estimates $\epsilon_{i,t}\epsilon_\theta$

We used a combination of original prompts and prompts taken from the paper. We focused on qualitative evaluation as the nature of the task was inherently perceptual.

4 Results & Analysis

We re-implemented three of the four factorization techniques presented in the paper, opting not to re-implement inverse hybrids and instead faithfully recreating the paper's frequency, color, and motion blur factorizations.

Our output quality is comparable to the original paper (Figure 1) for frequency factorization, and less so for the color and motion images. One challenge was the models being too large to run on our

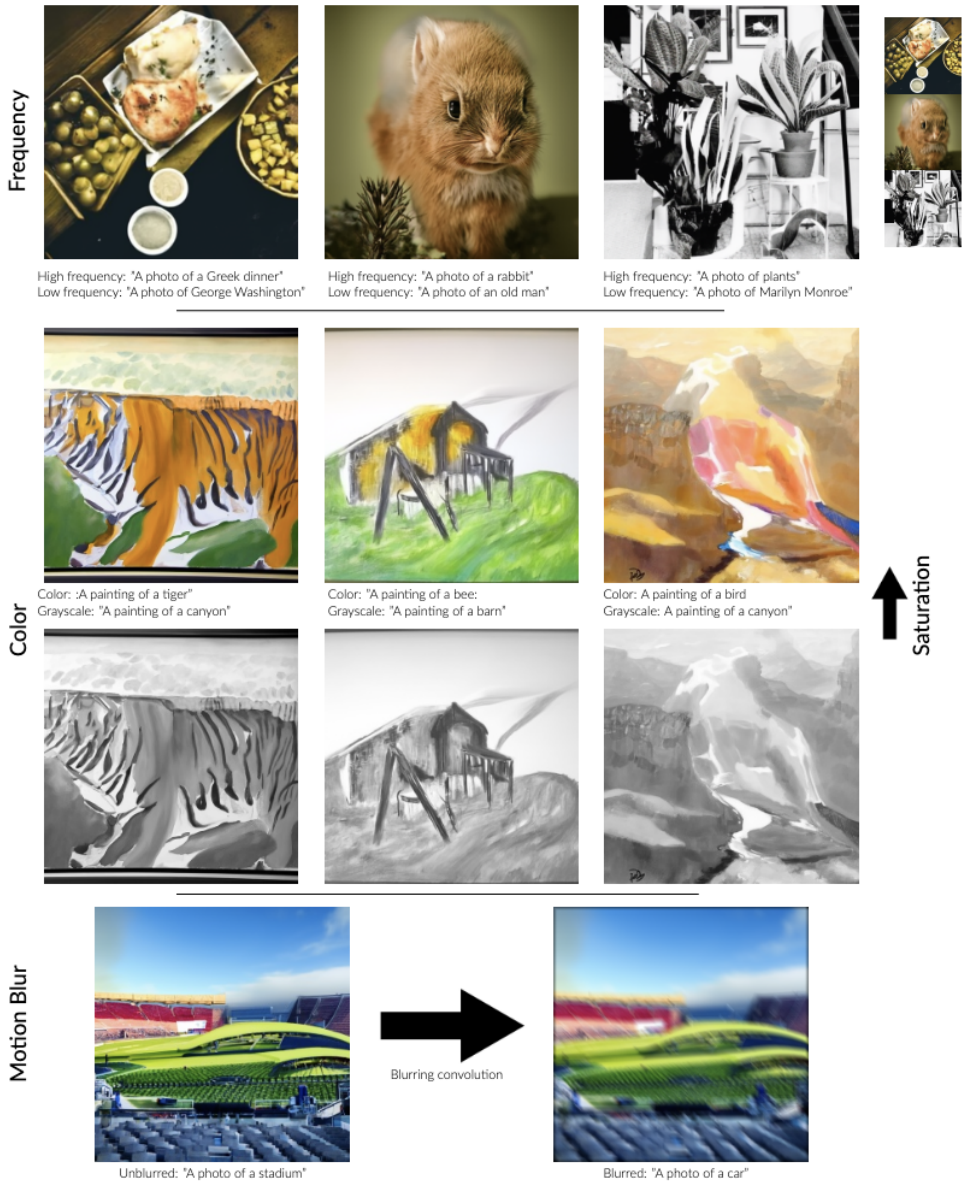

Figure 1: Hybrid images generated using various factor decompositions and prompts

GPUs. In turn, we had to switch to CPU, slowing down inference and testing. In addition, choice and order of prompts heavily impacted quality. Finally latent space diffusion models did not work since the factorizations were not meaningful in latent space.

Overall, our results support the paper's main contribution, that perceptual factors like frequency and color can be explicitly manipulated in diffusion models via noise decomposition. This approach opens

promising directions in the broader generative modeling space, enabling more controllable and interpretable image synthesis.

5 Reflections

In the usual ML paradigm, 'training' is an essential part of any model, so much so that we often conceive it as a necessary part of any machine learning implementations. Using existing diffusion models in a unique way was great exposure to how interesting results can obtained using pre-existing models, showing that ML and training models can be distinct.

We faced various challenges during this project, which led to many lessons learned. We learned that Computer architecture (notably GPU strength) is highly important for fast inference in Diffusion models.

Future directions we could take are implementing inverse hybrids - A technique used to create a hybrid image of an existing image and a prompt, as opposed to a hybrid image from 2 prompts.

6 References

[1] Daniel Geng, Inbum Park, and Andrew Owens. "Factorized Diffusion: Perceptual Illusions by Noise Decomposition". In: European Conference on Computer Vision (ECCV). 2024. URL: https://arxiv.org/abs/2404.11615.

[2] IF by DeepFloyd Lab at StabilityAI. https://github.com/deep-floyd/IF.

[3] Visual Anagrams | Factorized Diffusion. https://github.com/dangeng/visual_anagrams.