On Mechanistic Knowledge Localization in Text-to-Image Generative Models

tldr; We identify and edit a small set of layers across various diffusion models (e.g., SD-XL) which controls for visual attributes such as *style, objects and facts*

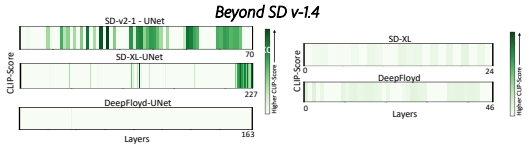
Samyadeep Basu*, Keivan Rezaei*, Priyatham Kattakinda, Vlad Morariu, Cherry Zhao, Ryan Rossi, Varun Manjunatha, Soheil Feizi

Adobe

University of Maryland 18 56

Motivation

Earlier Work using Causal Tracing for early SD variants (v1.4) has shown that knowledge about various visual attributes is (i) Distributed in UNet and (ii) Localized in the CLIP text-encoder



In UNet: Knowledge is distributed in SD-XL sparsely, but no relevant causal layers in DeepFloyd

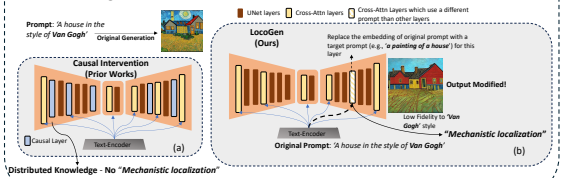
In text-encoder: No localization in the text-encoder

*A **small set of cross-attention layers** control visual attributes such as style, objects, facts*

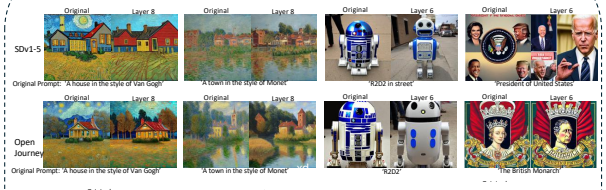
Salient Research Questions

- Can we localize knowledge 'universally' across different text-to-image models including SD-XL and DeepFloyd? - **LocoGen**
- Are these locations editable?
- Can we introduce closed-form edits across these locations to remove copyrighted content from text-to-image models? - **LocoEdit**

***LocoGen:* Identifying a small number** of cross attention layers which **control the generation** of distinct **visual attributes.**



LocoGen: Localization Results



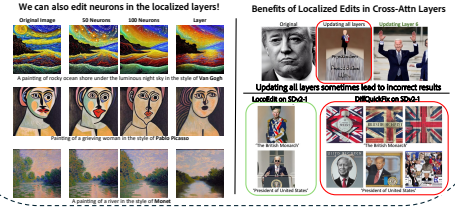
Replacing the text-embeddings of the original target prompt within those layers ***alters generated image***.

LocoEdit: Editing the Localized Layers

Update key and value matrices in cross-attention mechanism of identified layers.

$$\min_{W_l^K} \|\mathbf{X}_{\text{orig}} W_l^K - \mathbf{X}_{\text{target}} \bar{W}_l^K\|_2^2 + \lambda_K \|W_l^K - \bar{W}_l^K\|_2^2$$

Edited models **do not generate images** with **specific style, trademark objects, or old facts.**



ICML International Conference On Machine Learning

Paper Link

Code Link