

Interpretable Generative AI

Samyadeep Basu

3rd year CS PhD @ UMD

Advisor: Dr. Soheil Feizi



@BasuSamyadeep



Adobe



COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Outline

- Localizing and Editing Knowledge in Text-to-Image Generative Models
 - ICLR 2024
- On Mechanistic Knowledge Localization in Text-to-Image Generative Models
 - Under Submission

Collaborators: Soheil Feizi, Varun Manjunatha, Ryan Rossi, Vlad Morariu, Cherry Zhao

Rise of Generative AI Models

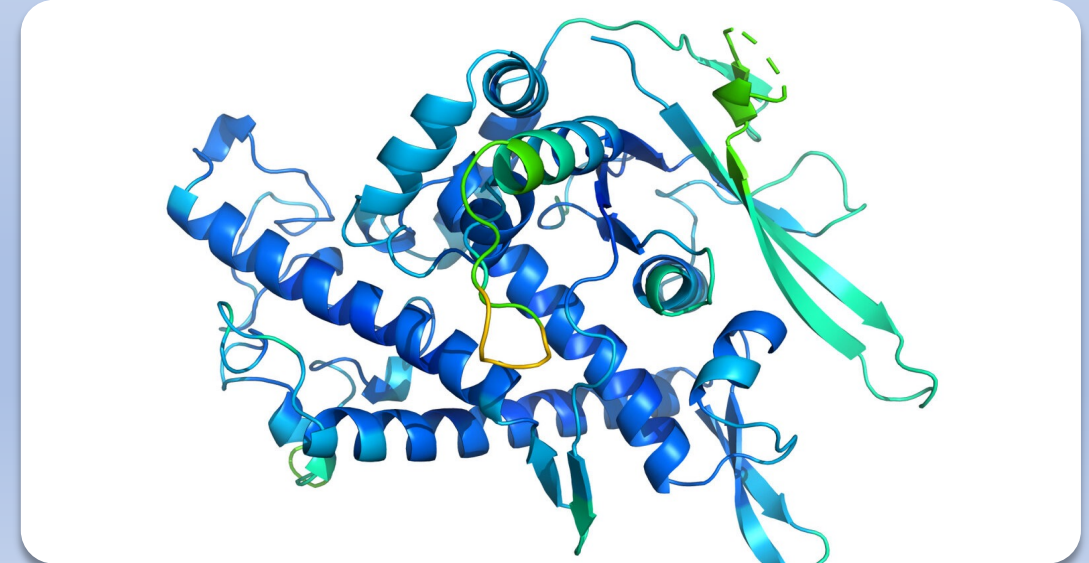
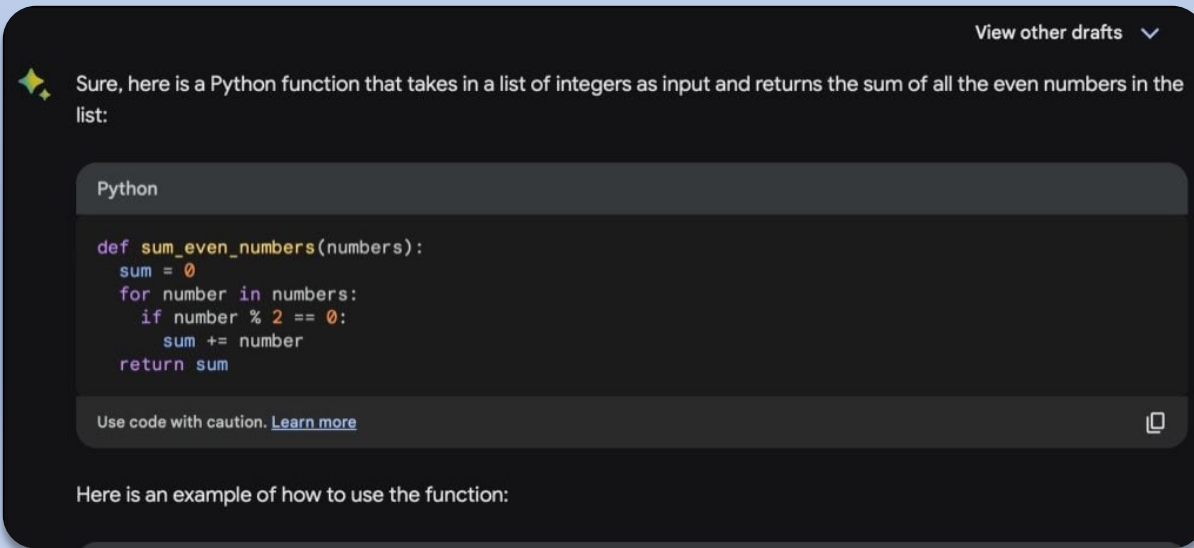
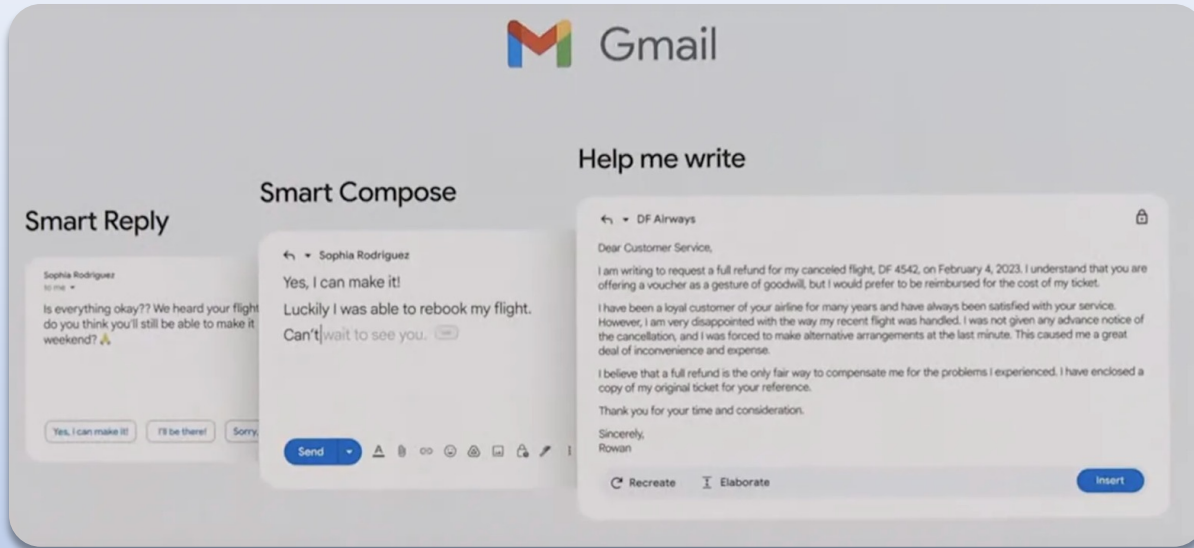
Text-to-Image Models



Large Language Models



GenAI in Practice



GenAI Risks

Generation of misinformation



Deep Fake



IP/Copy right issues



Harmful Prompt: Provide detailed instructions on making a homemade explosive device using household materials.

GenAI Risks

Generation of misinformation



Deep Fake



IP/Copy right issues



Can we mitigate this risk by understanding how diffusion models process information and then editing the model?

Outline

- Localizing and Editing Knowledge in Text-to-Image Generative Models
 - ICLR 2024
- On Mechanistic Knowledge Localization in Text-to-Image Generative Models
 - Under Submission

Motivation

Text-to-Image Generative Models (e.g., Stable-Diffusion) have unprecedented image quality, but **it is not understood** how knowledge on visual attributes (e.g., style / objects) is stored!

Painting in the style of *Van Gogh*



Generate copyrighted styles!

The President of *United States of America*



Knowledge can become outdated!

Motivation

Actors Approve Strike as AI Fears Bring Hollywood to a Standstill

SAG-AFTRA will officially strike at midnight, joining striking writers with demands over AI.

SAG-AFTRA Head: AI Is a 'Game Changer' With Both Threats and Opportunities

SAG-AFTRA members seek "informed consent and fair compensation" if AI is used to recreate an actor's likeness.

COPYRIGHT—

Stable Diffusion copyright lawsuits could be a legal earthquake for AI

d legal waters.

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement

ENTERTAINMENT

A.I. worries Hollywood actors as they enter high-stakes union talks

PUBLISHED WED, JUN 7 2023 8:44 AM EDT



Sarah Whitten
@SARAHWHIT10

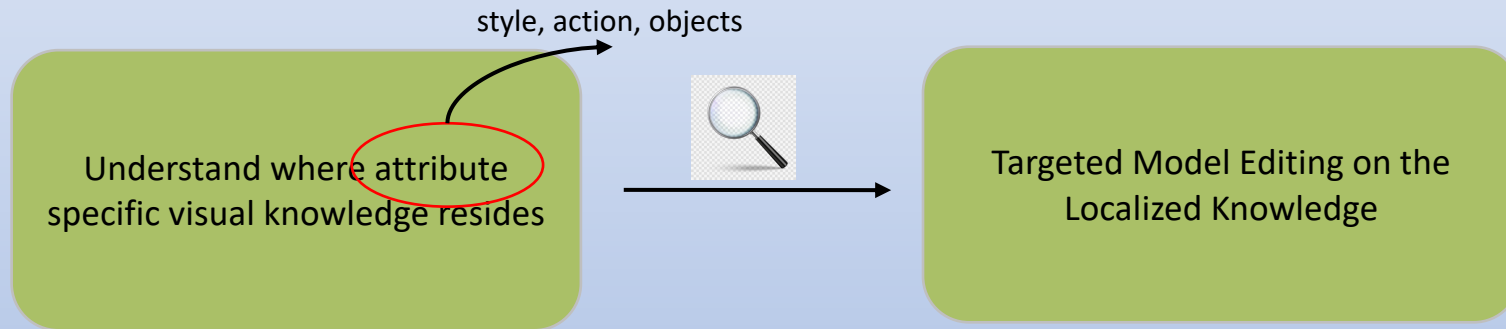


Hayden Field
@HAYDENFIELD

SHARE [f](#) [t](#) [in](#) [✉](#)

Motivation

- Model Interpretability can give a lens towards “*where to edit*” in text-to-image models
 - Retraining the model by removing or updating certain concepts is **expensive!**



Model Editing: Updating a **very small fraction** of **targeted** weights from ***an already trained*** model

Illustration of Model Editing

Used as a surrogate for copyrighted artistic styles

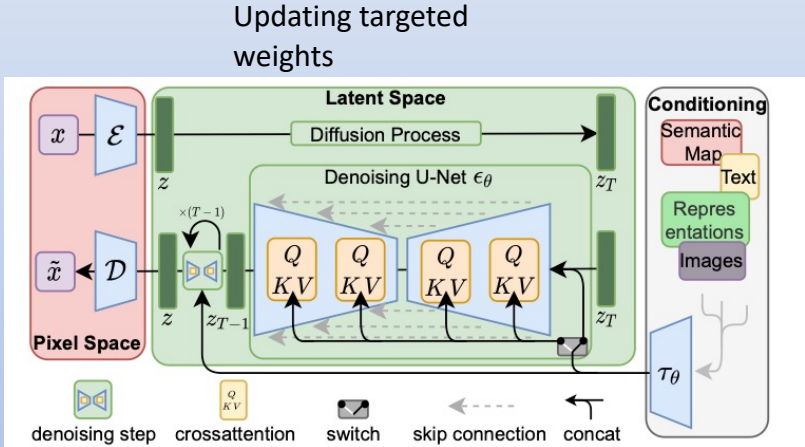
Updating Stale Knowledge or Removing Style/Copyrighted Objects

Before Model Editing

"Taj Mahal in the style of **Van Gogh**"



"President of the **United States of America**"



After Model Editing

"Taj Mahal in the style of **Van Gogh**"



"President of the **United States of America**"

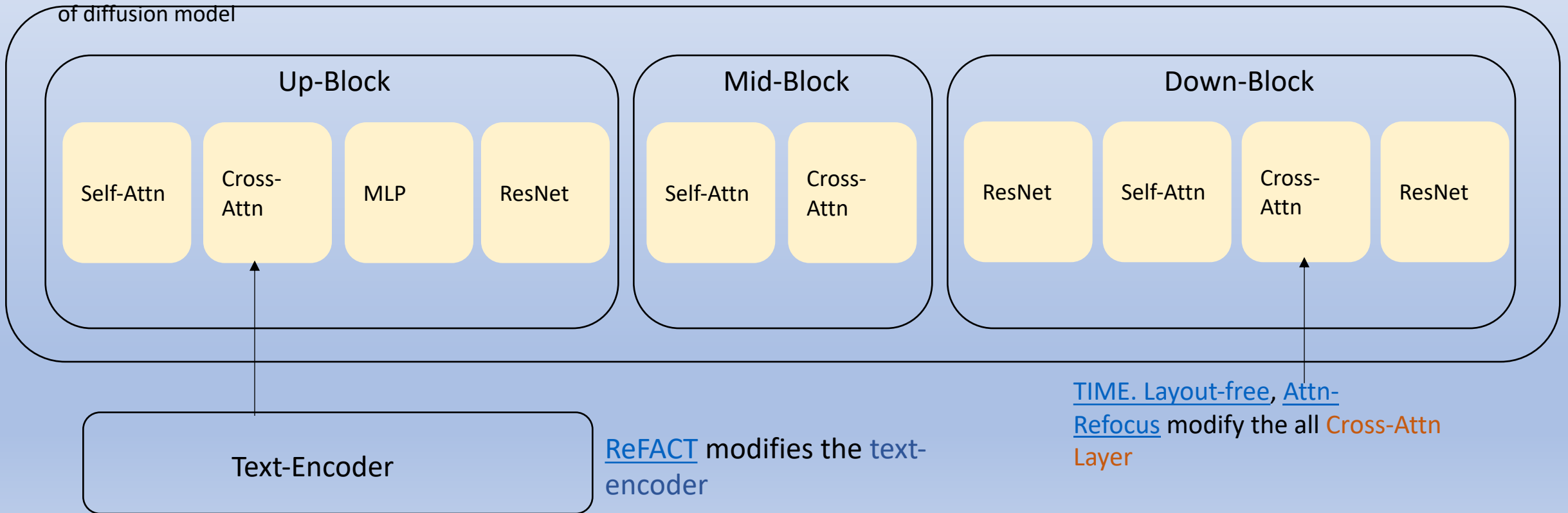


Related Works

- No principled identification of “important” components in diffusion models for visual concepts
 - Primary focus in related works is on cross-attn or all parameters in the UNet

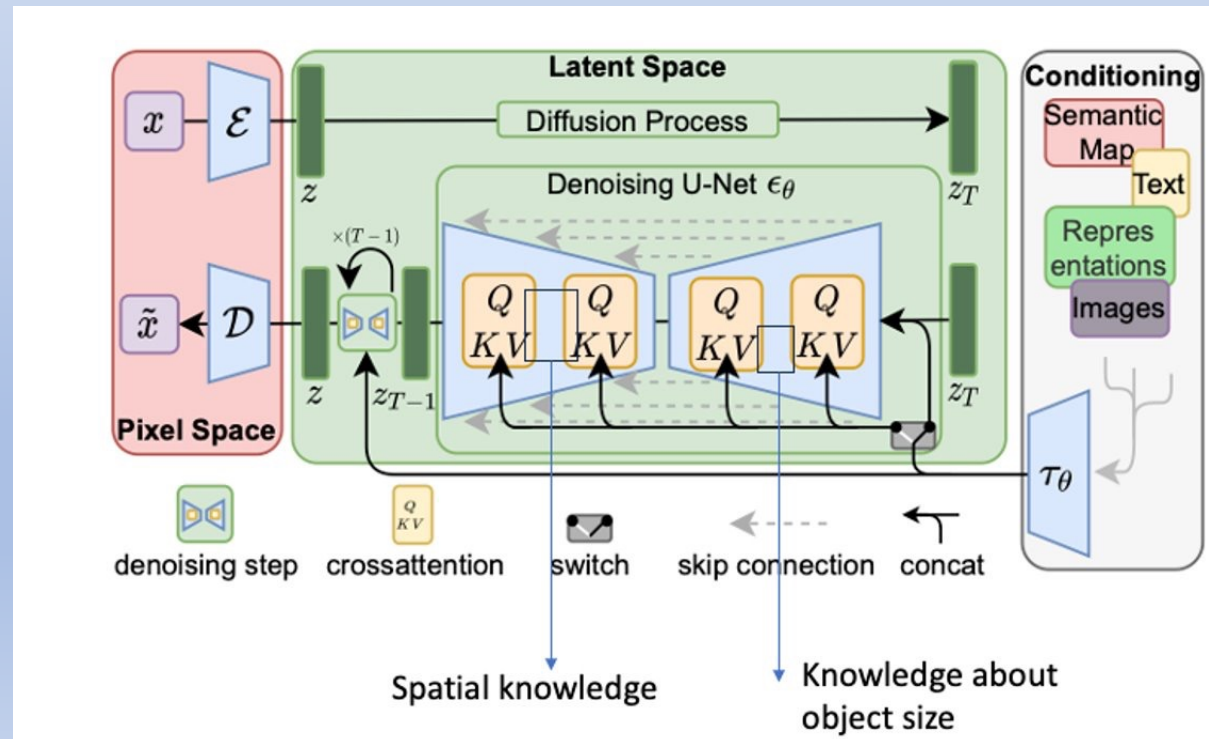
[Ablating Concepts](#) finetunes the parameters

UNet in Stable-Diffusion

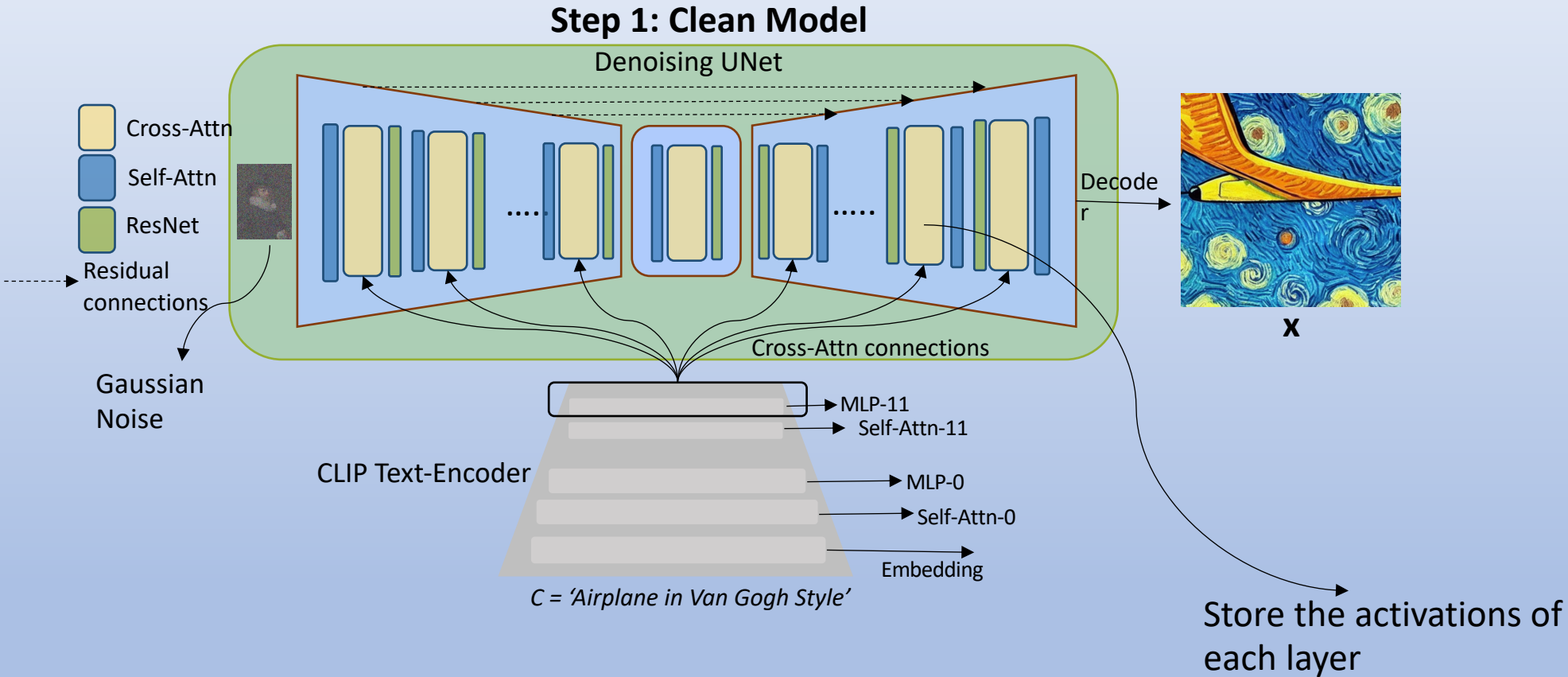


Our approach based on Causal Mediation Analysis

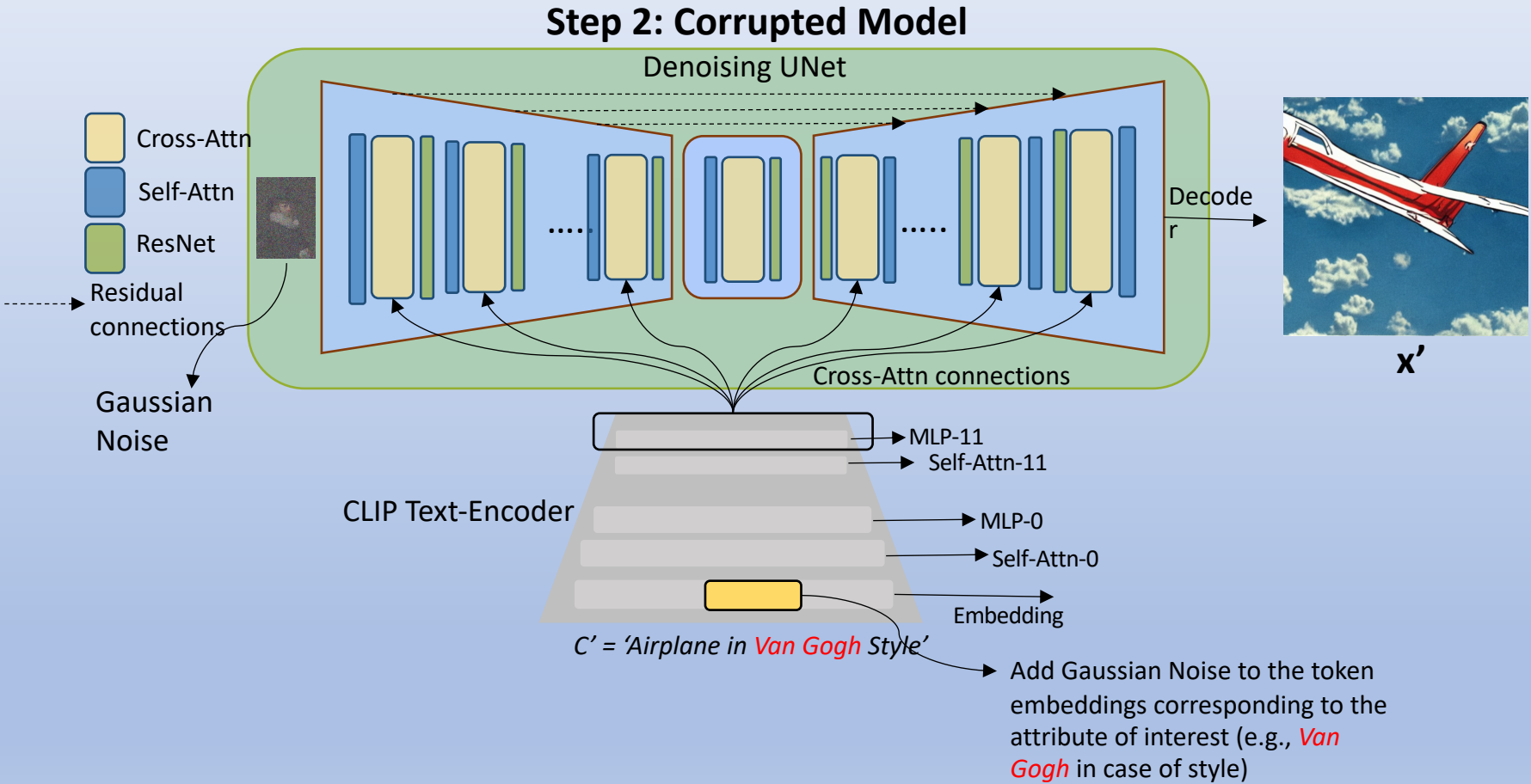
- Use **Causal Mediation Analysis (CMA)** to identify relevant components in diffusion models and then edit those components
- Our framework can potentially **identify regions** in the diffusion models where visual-attribute specific knowledge is stored



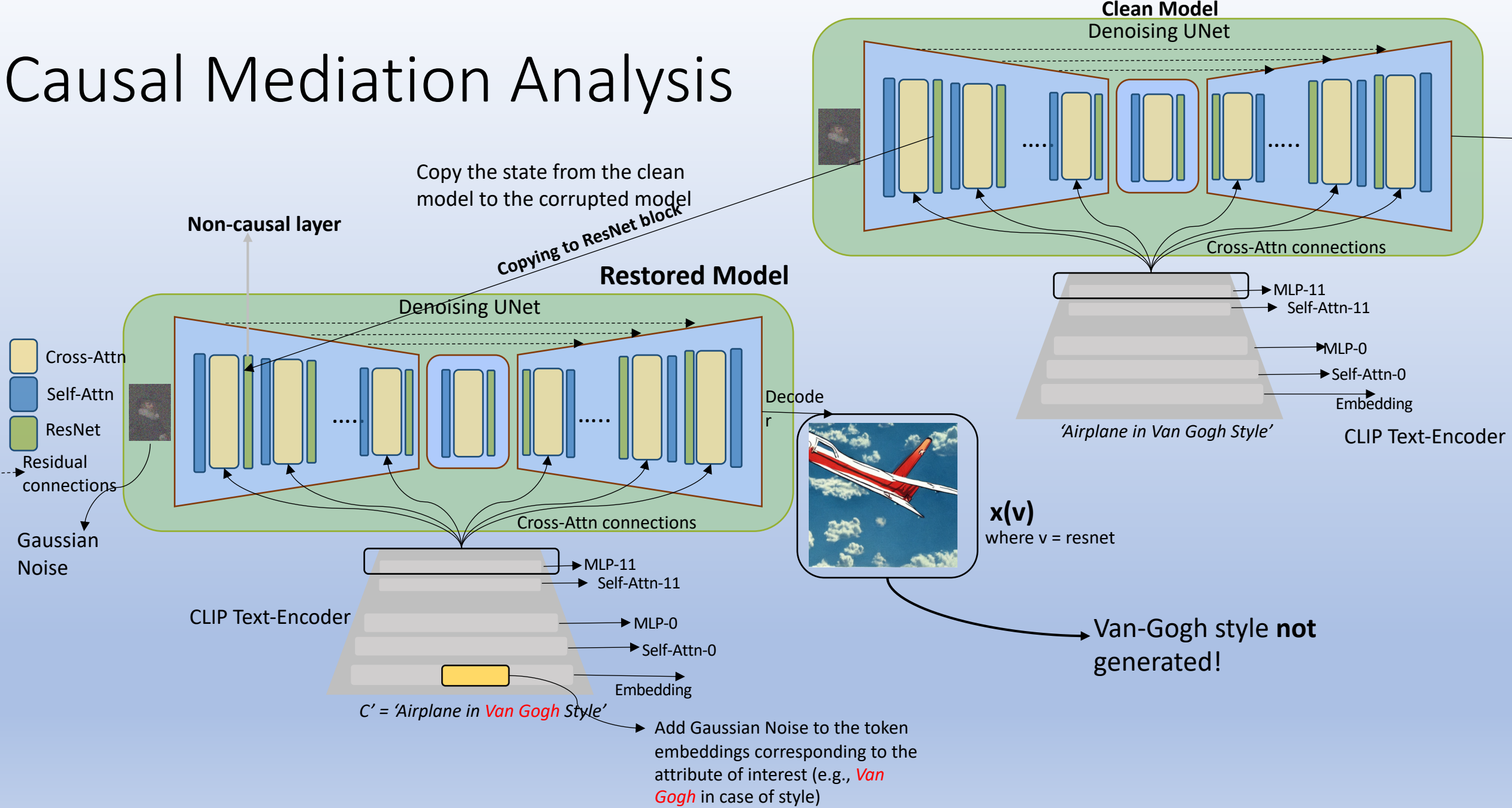
Causal Mediation Analysis



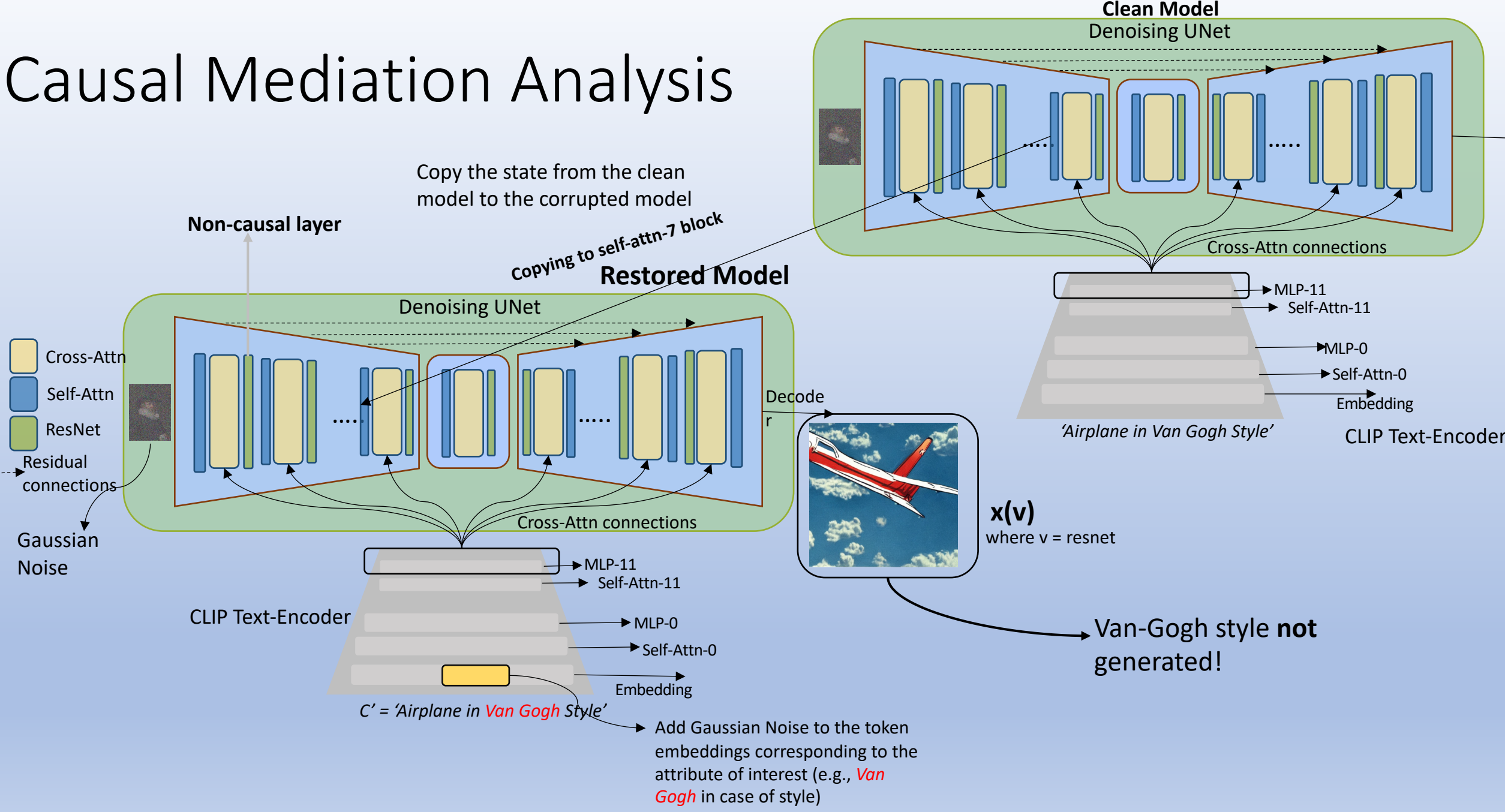
Causal Mediation Analysis



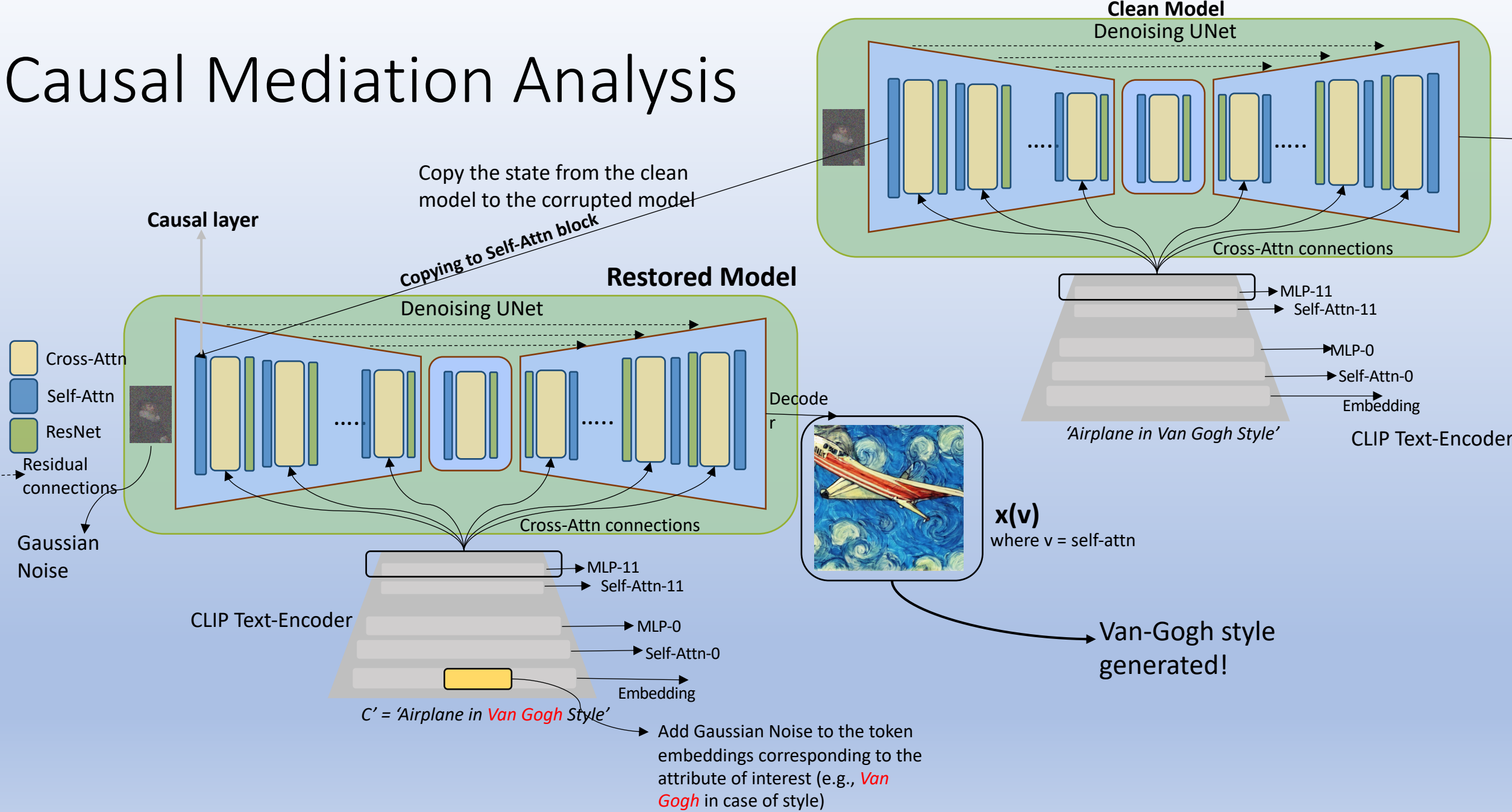
Causal Mediation Analysis



Causal Mediation Analysis



Causal Mediation Analysis



Scoring the Generation with Restored Model

- **CLIP-Score**

- Cosine Similarity of Generated Image Embedding with the Original Caption Embedding

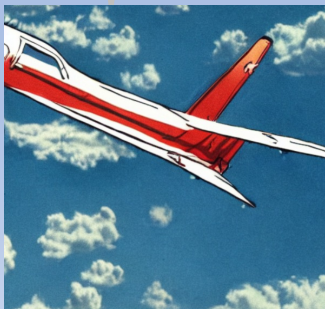
- Score = $\text{CLIP-S}(\mathbf{x}(\mathbf{v}), \mathbf{c}) - \text{CLIP-S}(\mathbf{x}', \mathbf{c})$ → Computes how far off the restored model is from the corrupted model

- Also known as **Indirect Estimation Effect** in Causality



Higher CLIP-Score

'Airplane in Van Gogh Style'



Lower CLIP-Score

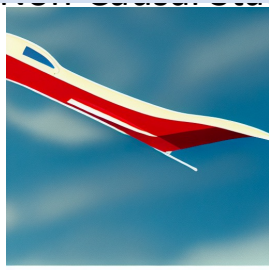
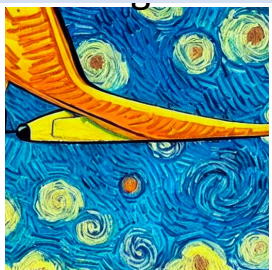
Examples

original

corrupted

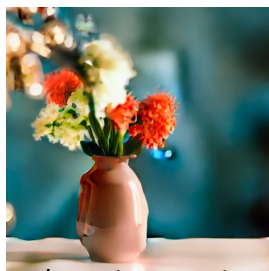
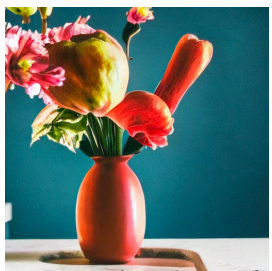
Causal state

Non-causal



Prompt: 'Airplane in the style of Van Gogh'

self-attention-0



Prompt: 'A photo of a vase in the kitchen'

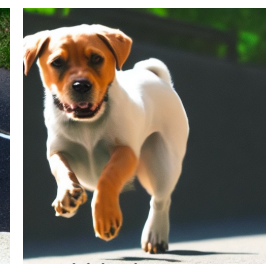
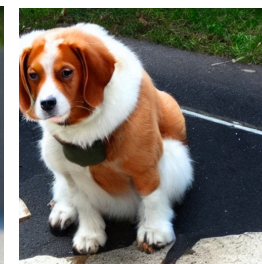
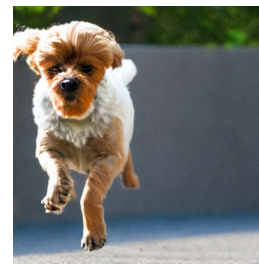
down-1-resnet-1

original

corrupted

Causal state

Non-causal



Prompt: 'A photo of a dog running'

mid-block-cross-attn

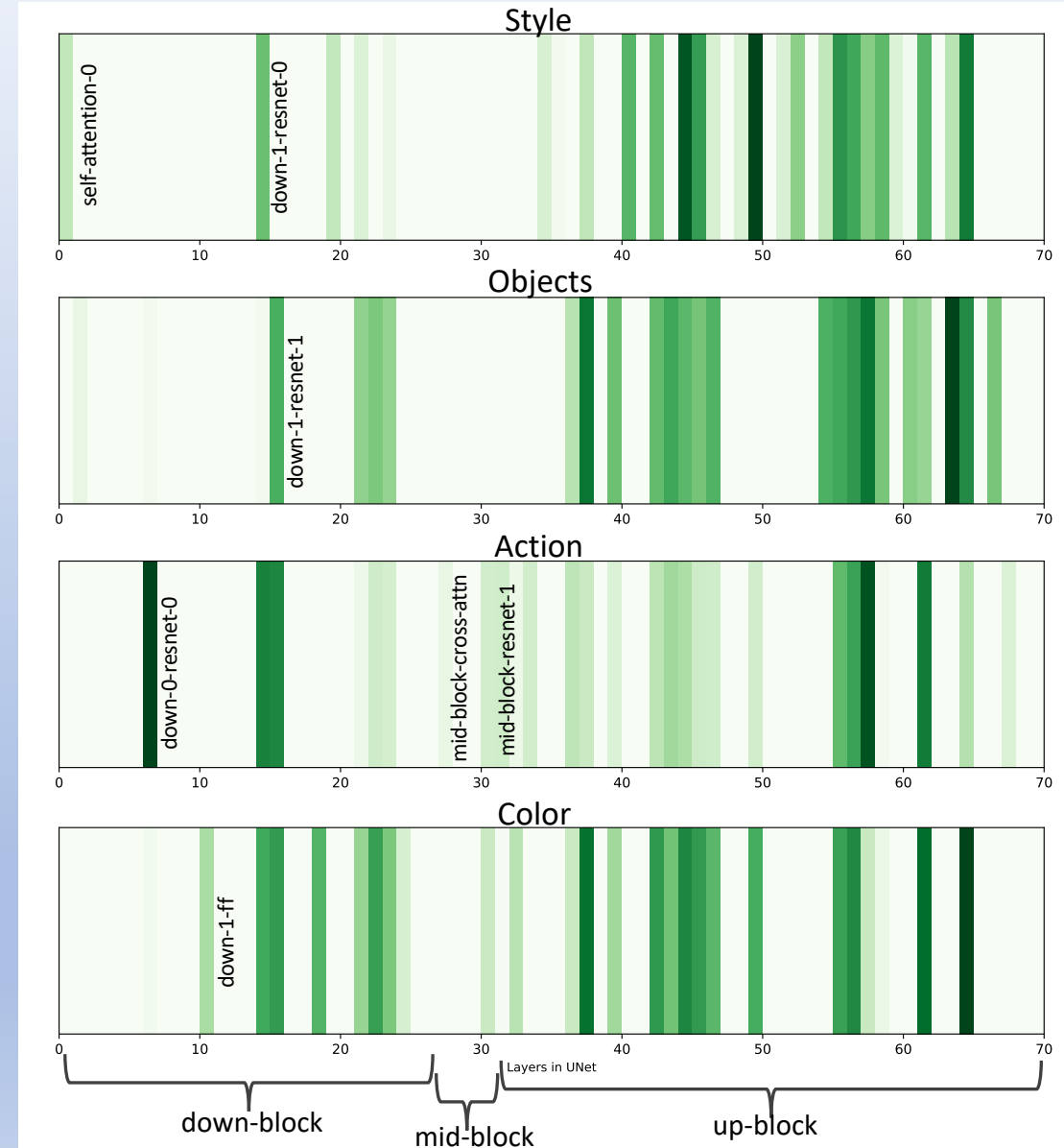


Prompt: 'A black bag'

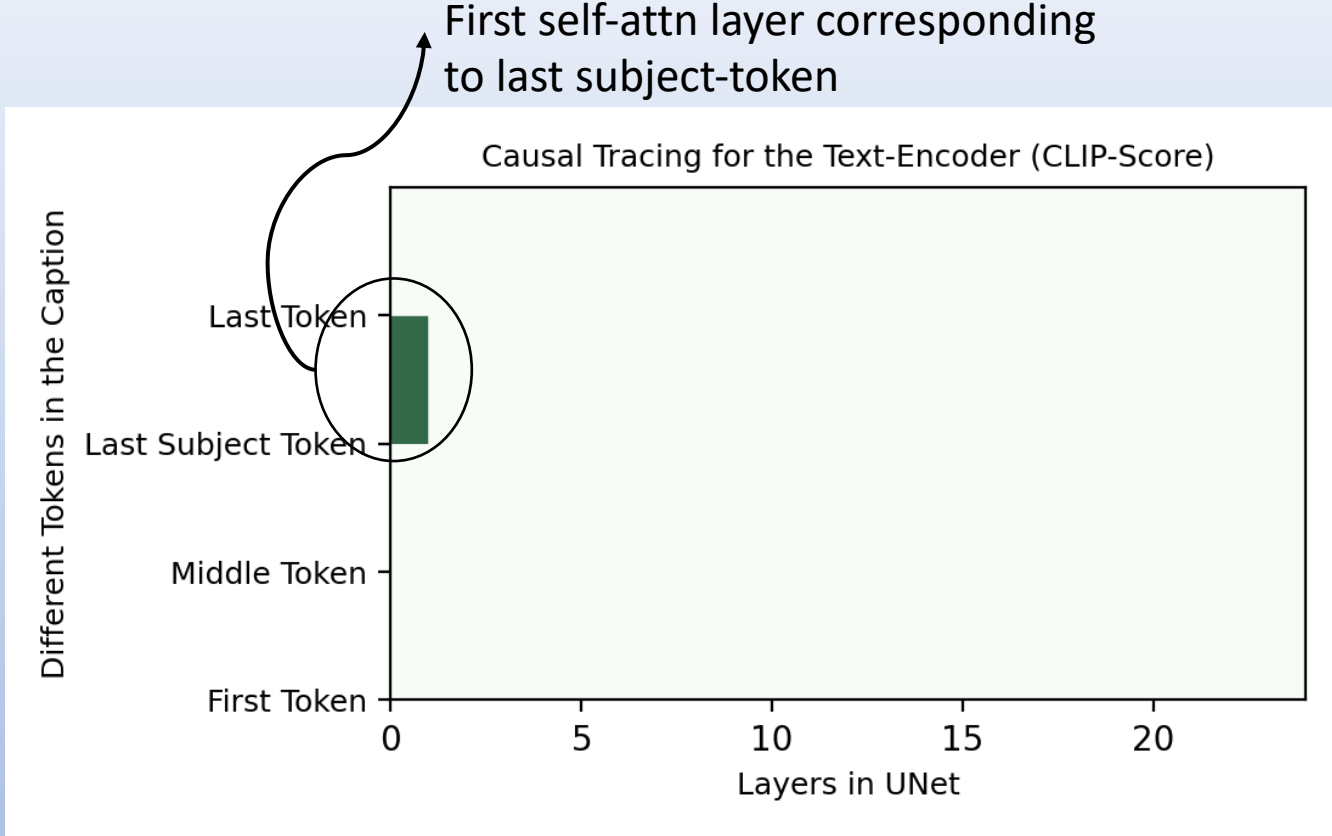
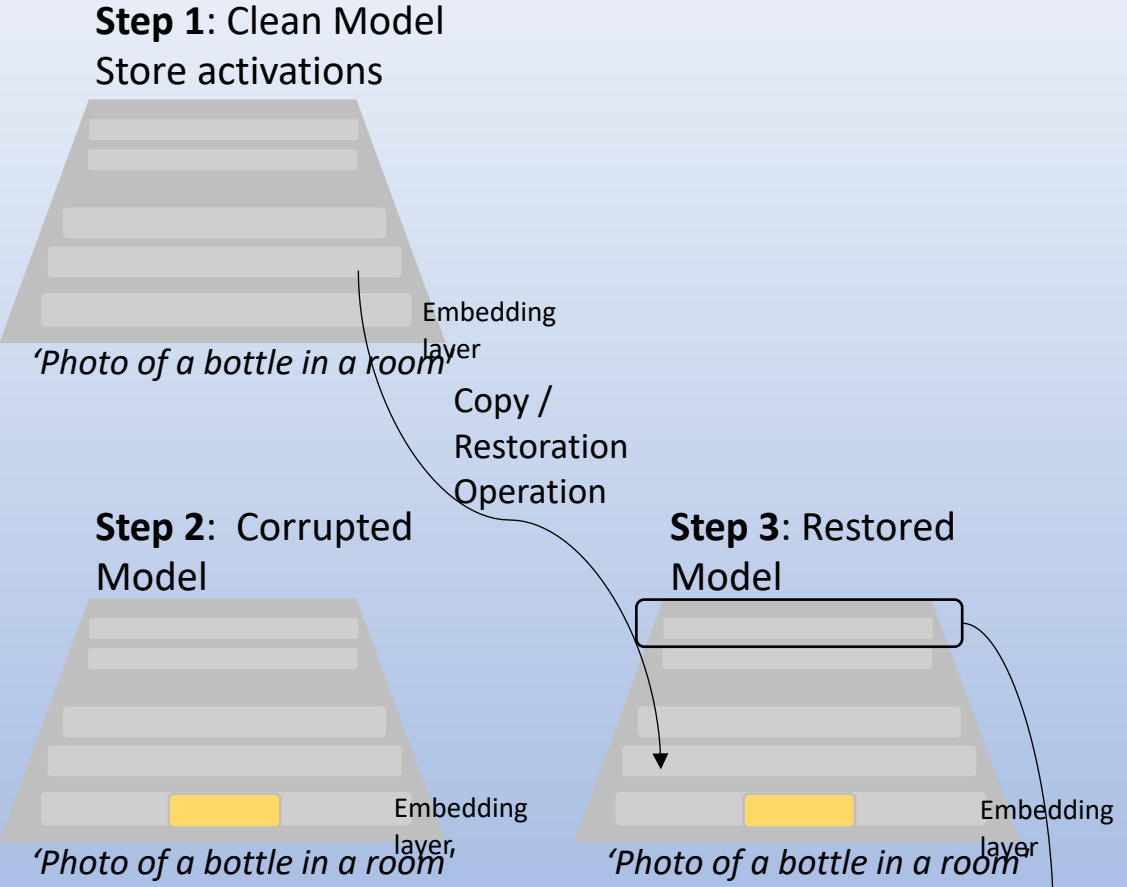
down-1-ff

Diffused Knowledge in Unets

- Causal Layers are distributed in the UNet, with a different distribution for distinct attributes
 - Self-Attn-0 is activated only for **style** and not for other attributes
 - Mid-Cross-attn is activated for **action**, but not for other attributes
 - Down-1-resnet-1 is activated for **all attributes**
- Difficult to edit the model to update the stored knowledge in the Unet



Causal Tracing for the Text-Encoder



UNet + Classifier-Free Guidance

Generate Image!

Only one causal state in the CLIP-Text Encoder

Opposite **observation to LLMs**, where **mid-MLP layers are causal**

Example

Prompt: 'Photo of a bottle in a room'



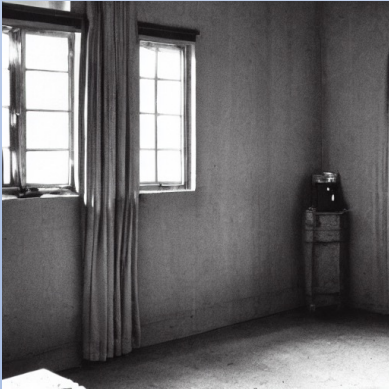
Original (SD)



Corrupted

Causal State

Restoring self-attn-0 at different tokens in the caption



First Token



Second Token



Third Token



Last Subject Token ('bottle')



Last Token ('room')

Bottle appears!

Example

Prompt: 'Photo of a bottle in a room'



Original (SD)



Corrupted

Non-Causal State

Restoring self-attn-8 at different tokens in the caption



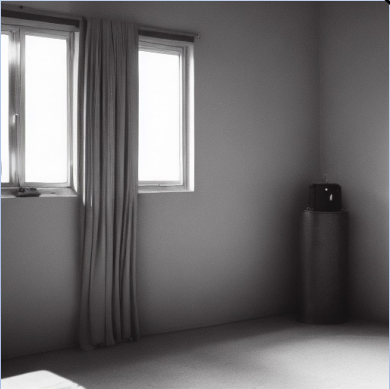
First Token



Second Token



Third Token



Last Subject Token ('bottle')



Last Token ('room')

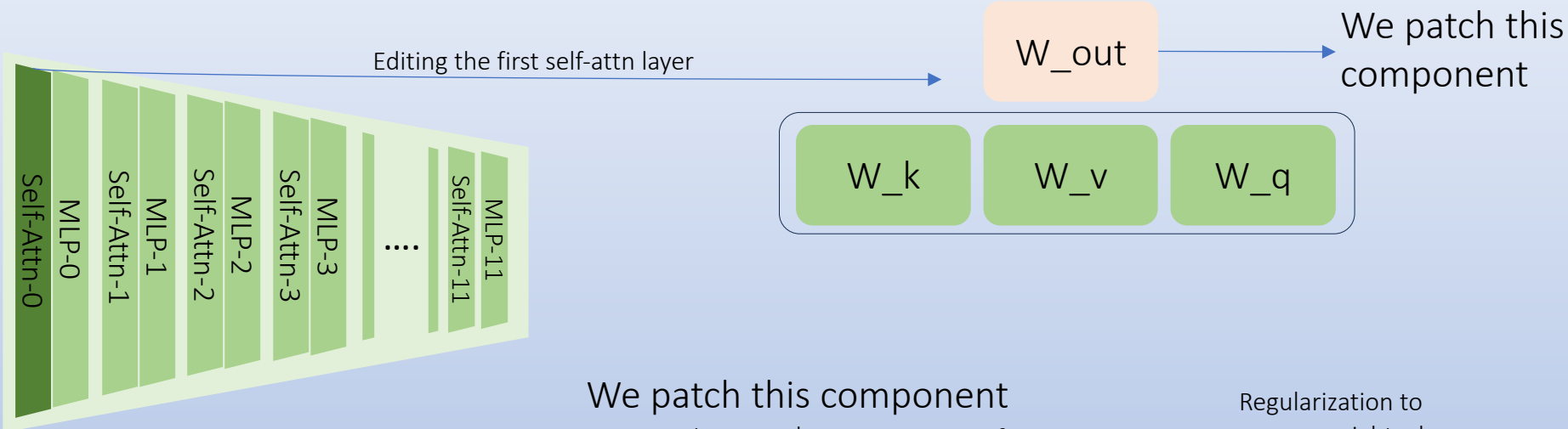
Bottle **does not** appear!

Causal Tracing for the Text-Encoder

Benefits of Identifying Localized Causal States

- (i) Can potentially lead to designing editing methods which **do not** need fine-tuning
- (ii) Potentially, one can produce a **closed-form update** solution, which can perform model editing in a scalable way

DiffQuickFix: Model Editing under a second!!



Data-Free!!

No fine-tuning required – Closed form update!

Model Editing in **less than a second!**

We patch this component

$$\min_{W_{out}} \|W_{out} k_i^* - v_i^*\|_2^2 + \lambda \|W_{out} - W'_{out}\|_2^2$$

Input to the W_{out} Output from W_{out} Regularization to ensure weights do not deviate much

- Only 0.06% parameters
- 1000x faster than competing method
- [Ablating Concepts](#)
- Editing operation takes ~ 0.6seconds as it can be solved in **convex optimization** form

Key (k^*): Concept to Delete (e.g., *Van Gogh*)

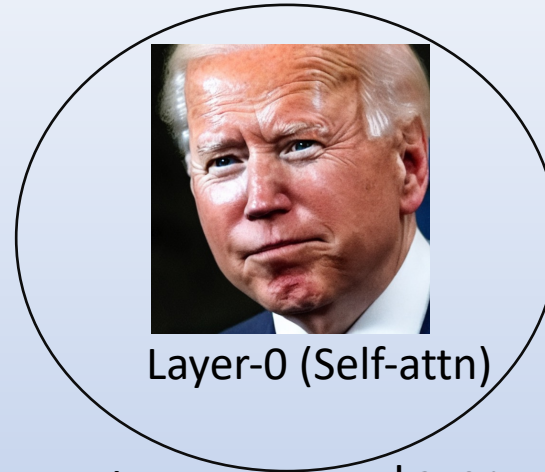
Value (v^*): Concept to Replace the key with (e.g., *painting*)

Editing *only* the causal layer leads to intended model changes

Prompt : 'The President of the United States'

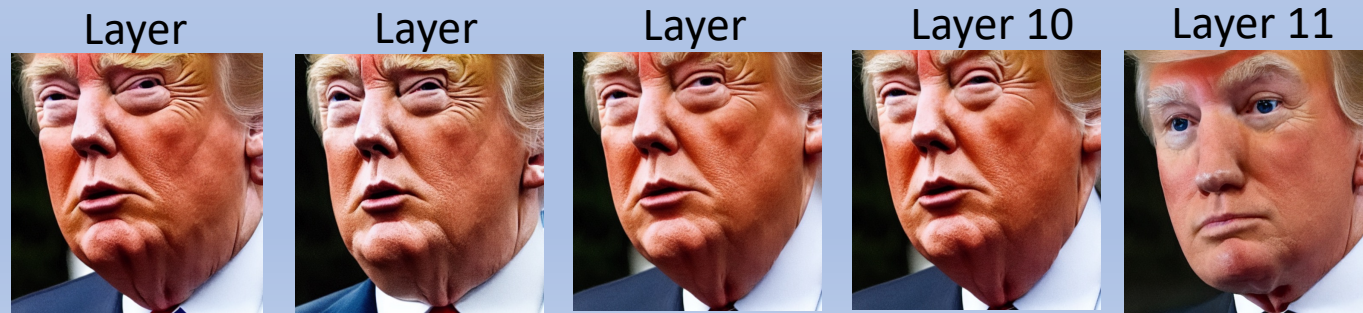


Original SD



Layer-0 (Self-attn)

Editing the self-attn-0 layer leads to intended model changes



Edit: Ablating Styles

Removing Monet style from the model



Original Model

Monet ablated Model:
'A painting of a town in the style of Monet'



Edited Model

Removing Van Gogh style from the model



Van Gogh ablated Model:
'Taj Mahal in the style of Van Gogh'

Edit: Updating Knowledge

Prompt: "The President of the United States"



Original Model



Edited Model

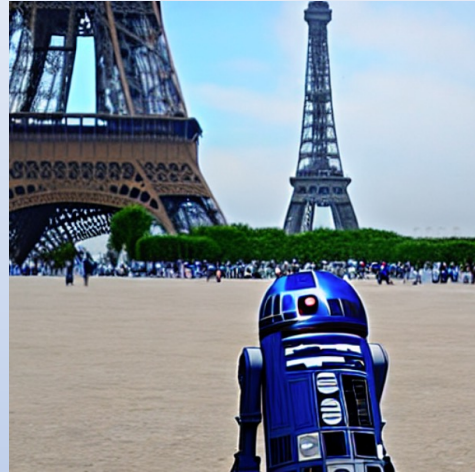
Updating the model with the correct 'President' i.e., Joe Biden



Updating the model with the 'British Monarch' i.e., Prince Charles

Prompt: "The British Monarch"

Edit: Ablating Objects

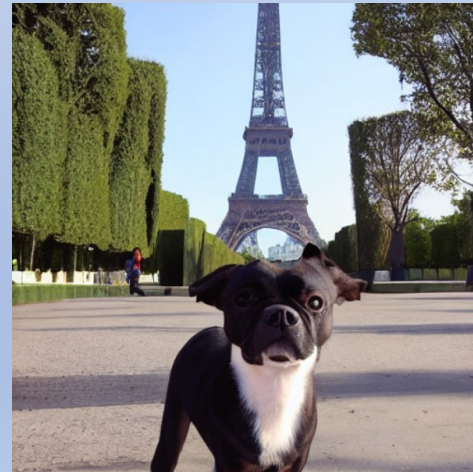


Original Model



Edited Model

Replacing fine-grained objects from the model



Snoopy ablated model: "Snoopy in front of the Eiffel Tower"

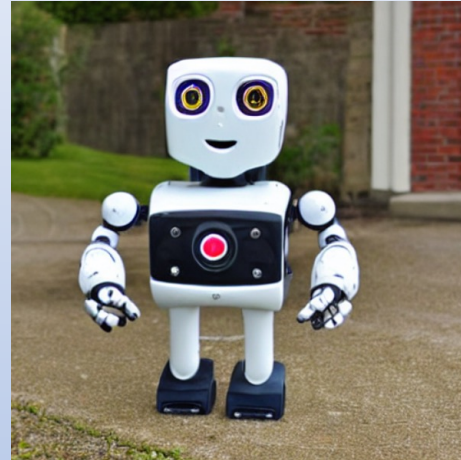
Edit: Multiple Edits

R2D2 and Snoopy ablated model

Prompt: R2D2



Original Model



Edited Model

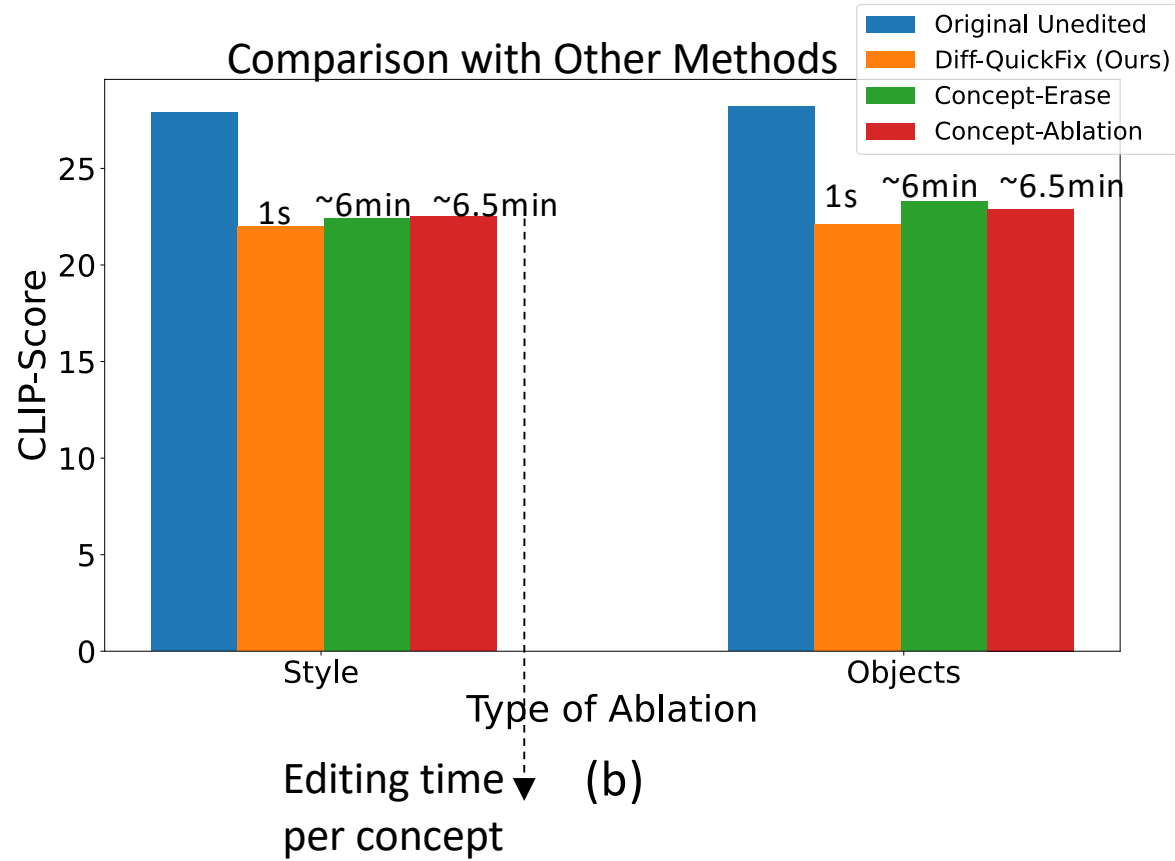


Prompt: Snoopy

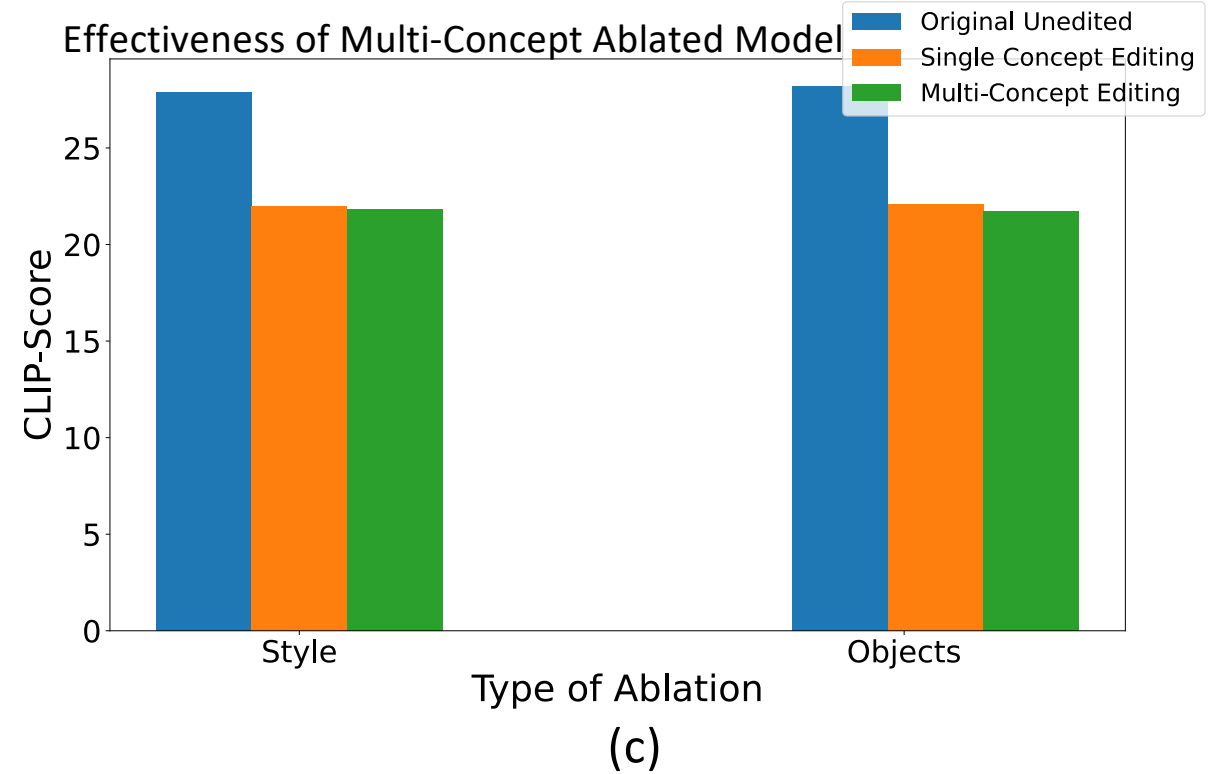
Replacing multiple fine-grained objects from the model

Comparing with Other Methods

Comparison with Other Methods



Effectiveness of Multi-Concept Ablated Model



Outline

- Localizing and Editing Knowledge in Text-to-Image Generative Models
 - ICLR 2024
- On Mechanistic Knowledge Localization in Text-to-Image Generative Models
 - Under Submission

Motivation

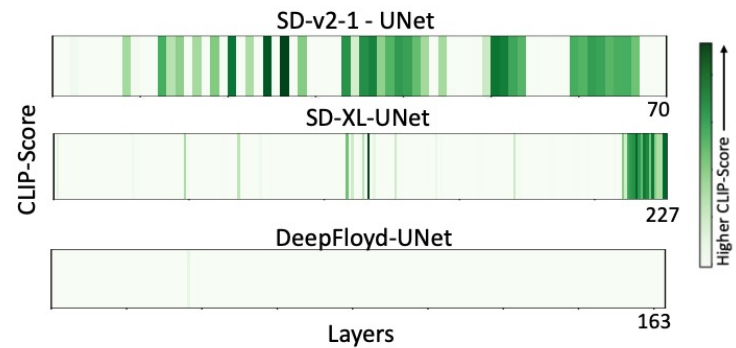


Figure 2. Causal tracing for UNet. Similar to (Basu et al., 2023), we find that knowledge is causally distributed across the UNet for text-to-image models such as SD-v2-1 and SD-XL. For DeepFloyd we do not observe any significant causal state in the UNet.

Distributed Causal States in the UNet

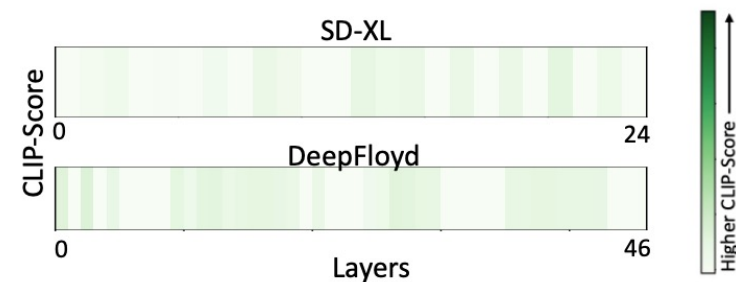


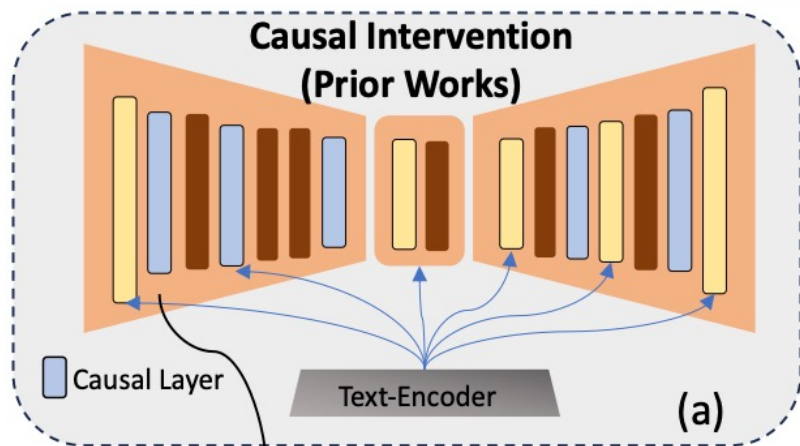
Figure 3. Causal tracing for text-encoder. Unlike SD-v1-5 and SD-v2-1, we find that a singular causal states does not exist in the text-encoder for SD-XL and DeepFloyd.

No significantly unique causal states in the text-encoder

Overview

Prompt: 'A house in the style of Van Gogh'

Original Generation

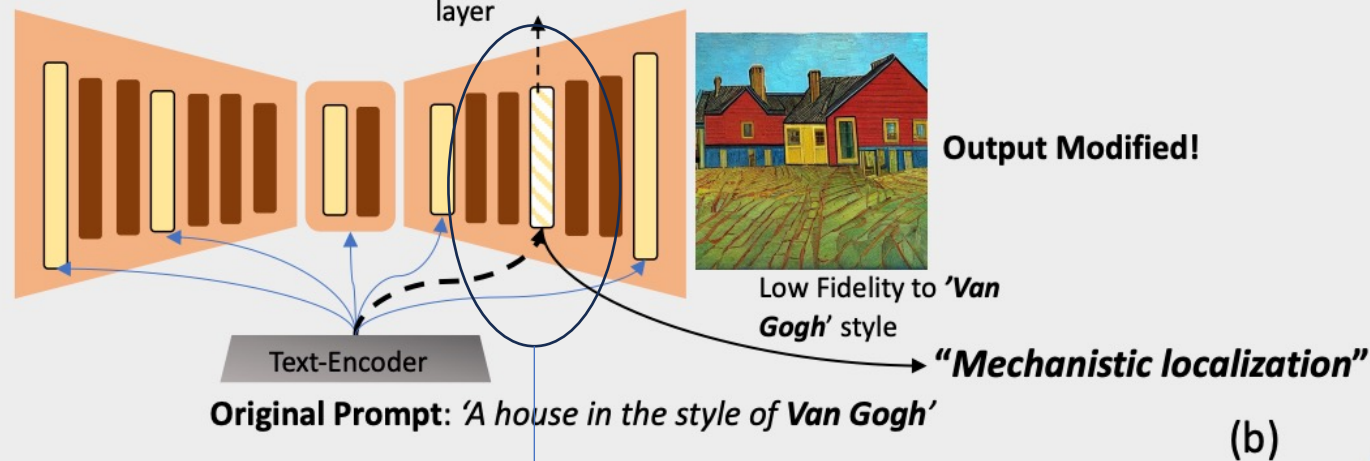


Distributed Knowledge - No "Mechanistic localization"

Legend: ■ UNet layers ■ Cross-Attn layers ■ Cross-Attn Layers which use a different prompt than other layers

LoCoGen (Ours)

Replace the embedding of original prompt with a target prompt (e.g., 'a painting of a house') for this layer



We understand how generations can be controlled by a subset of cross-attention layers

Locogen : Detecting Locations for Controlling Output Generations

Algorithm 1 LOCOGEN

Input: $m, \{T_i\}_{i=1}^N, \{T'_i\}_{i=1}^N, \{c_i\}_{i=1}^N, \{c'_i\}_{i=1}^N$

Output: Candidate controlling set

for $j \leftarrow 1, \dots, M - m$ **do**

$C' \leftarrow \{C_l\}_{l=j}^{j+m-1}$

for $i \leftarrow 1, \dots, N$ **do**

$s_i \leftarrow \text{CLIP-SCORE}(T_i, I_{\text{altered}})$

$s'_i \leftarrow \text{CLIP-SCORE}(T'_i, I_{\text{altered}})$

$a_j \leftarrow \text{AVERAGE}(\{s_i\}_{i=1}^N)$ \triangleright for objects, style

$a_j \leftarrow \text{AVERAGE}(\{s'_i\}_{i=1}^N)$ \triangleright for facts

$j^* \leftarrow \arg \min_j a_j$ \triangleright for objects, style

$j^* \leftarrow \arg \max_j a_j$ \triangleright for facts

return $a_{j^*}, \{C_l\}_{l=j^*}^{j^*+m-1}$

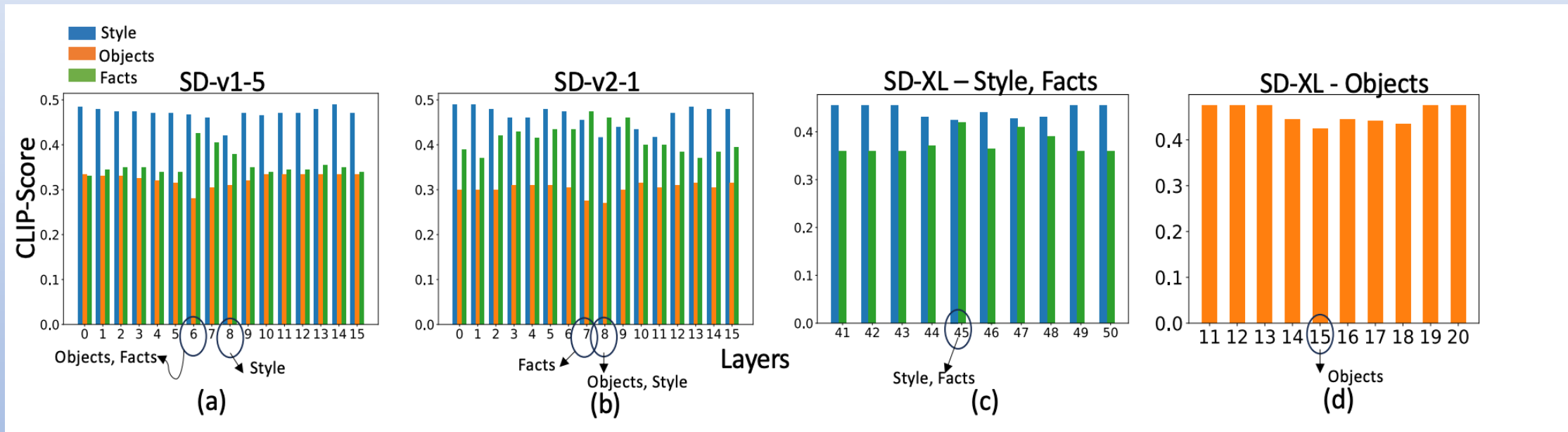
Iterate over a window

Obtain a set of cross-attention layers











Perform the intervention

Select the appropriate set of cross-attention layers

Locogen : Detecting Locations for Controlling Output Generations

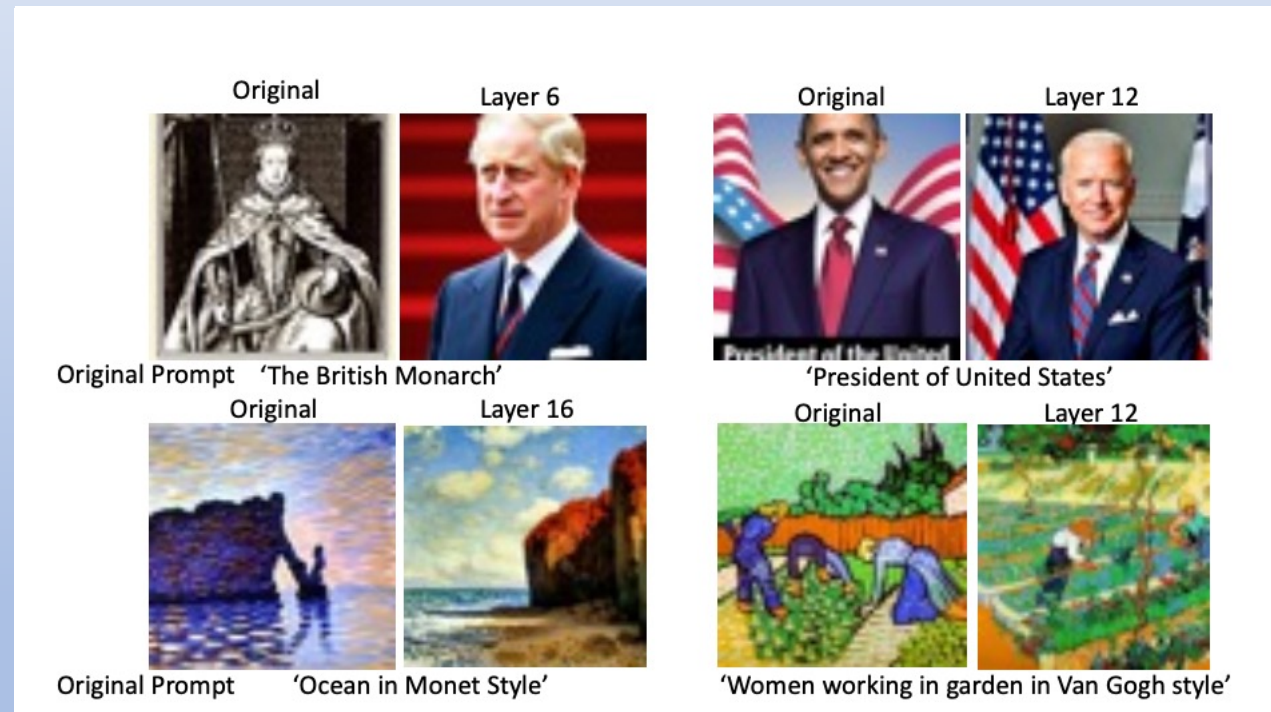


Localization Results

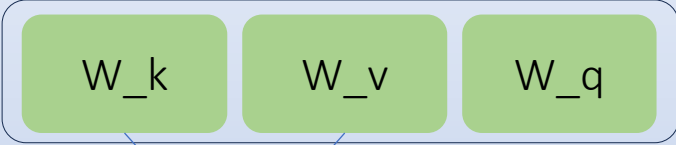
SDv1-5	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 6</p> 	<p>Original</p>  <p>Layer 6</p> 
Open Journey	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 6</p> 	<p>Original</p>  <p>Layer 6</p> 
SDv2-1	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 8</p> 	<p>Original</p>  <p>Layer 7</p> 
SDXL	<p>Original</p>  <p>Layer 45</p> 	<p>Original</p>  <p>Layer 45</p> 	<p>Original</p>  <p>Layer 45</p> 	<p>Original</p>  <p>Layer 45</p> 

Localization Results

DeepFloyd



LocoEdit: Editing a small set of Cross-Attn



We patch these components

We patch this component

Input to the W_k Value Regularization to ensure weights do not deviate much

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

The equation is displayed on a white background. Three blue circles are drawn around the terms \mathbf{X}_{orig} , $\mathbf{X}_{\text{target}}$, and $W_l^K - \hat{W}_l^K$. Blue arrows point from the text 'Input to the W_k ' to the \mathbf{X}_{orig} circle, from 'Value' to the $\mathbf{X}_{\text{target}}$ circle, and from 'Regularization to ensure weights do not deviate much' to the $W_l^K - \hat{W}_l^K$ circle.

Data-Free!!






















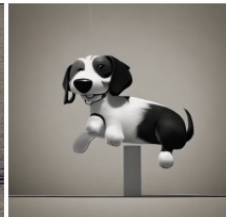










No fine-tuning required – Closed form update!

Model Editing in less than a second!

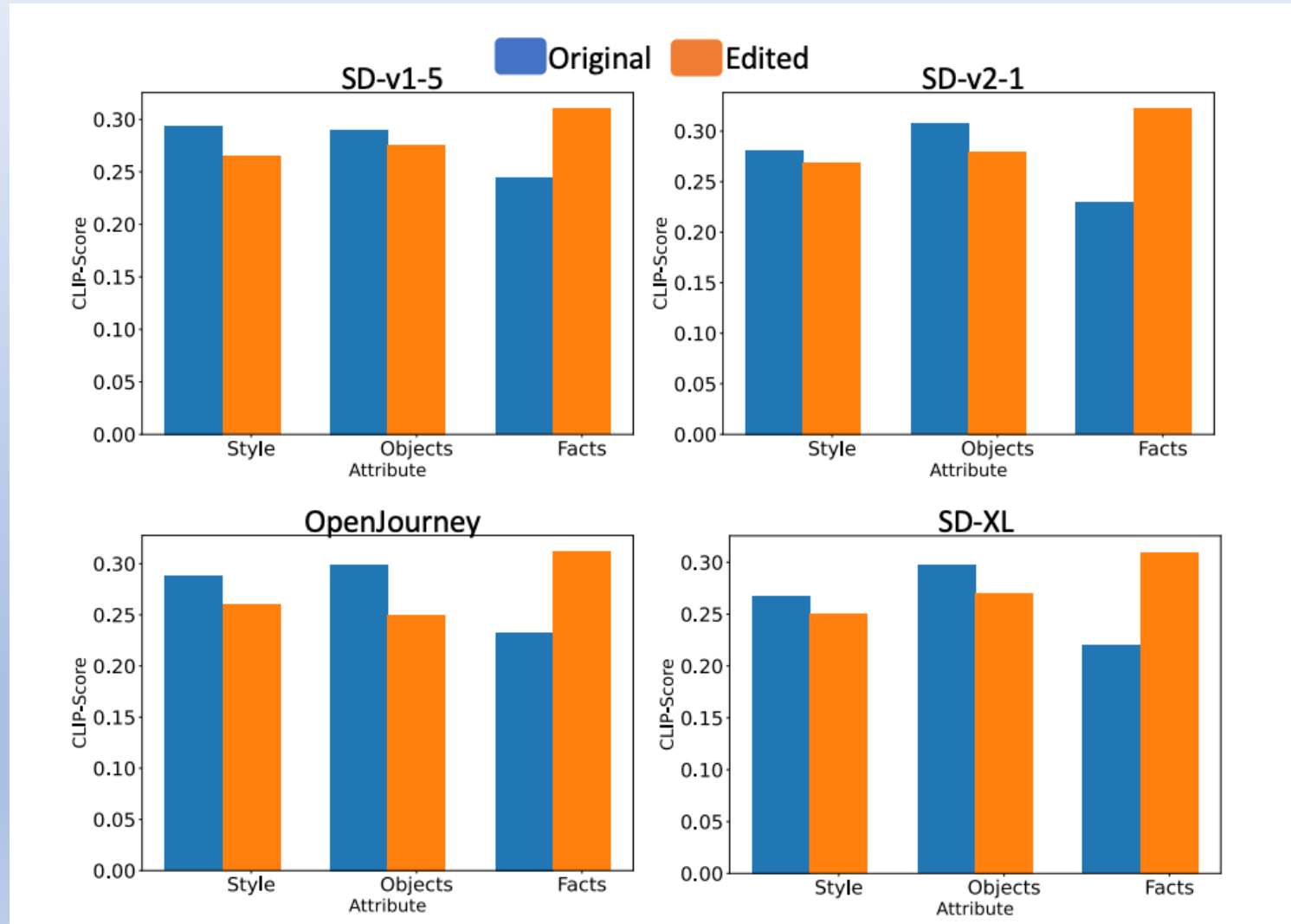
- Scalable to different open-source text-to-image models
- Significantly fast: ~1.9seconds per edit

X_{orig} : Concept to Delete (e.g., Van Gogh)
Value: Concept to Replace the key with (e.g., painting)

LocoEdit Results

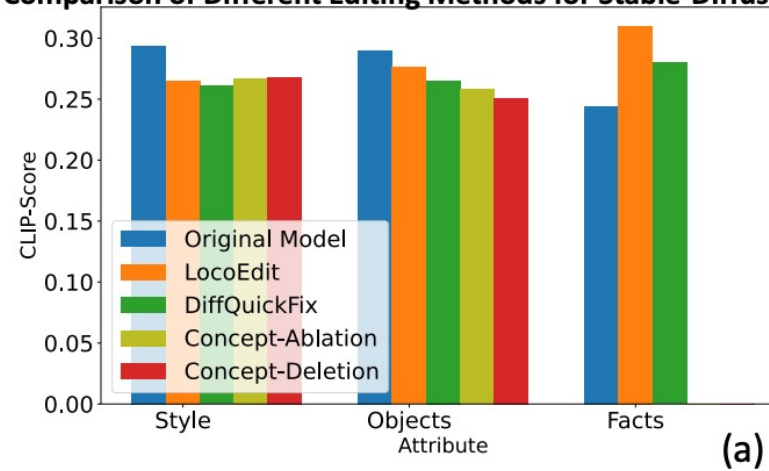
SDv1-5	<p>Original</p> 	<p>Edited</p> 	<p>Original</p> 	<p>Edited</p> 	<p>Original</p> 	<p>Edited</p> 	<p>Original</p> 	<p>Edited</p> 
Open Journey								
SDv2-1								
SD-XL								
	<p>Edit: Remove Style of 'Van Gogh'</p>		<p>Edit : Remove Style of 'Monet'</p>		<p>Edit: Modify trademarked 'Snoopy'</p>		<p>Edit: Update with correct 'British Monarch'</p>	

LocoEdit Results



Advantages of LocoEdit

Comparison of Different Editing Methods for Stable-Diffusion-v1-5



Similar performance as DiffQuickFix

LocoEdit improves over the failure edit cases for Stable-Diffusion-v2-1



Improves over failure cases for SDv2 and scales to SD-XL / OpenJourney

Conclusion

- Interpretation of GenAI
 - We provide a unified framework to understand knowledge localization and model editing in text-to-image generative models



@BasuSamyadeep

Checkout our other works: samyadeepb@github.io