A PROBE DATASET DESIGN DETAILS

In this section, we provide detailed descriptions of the probe dataset $\mathcal{P}$ which is used for causal tracing for both the UNet and the text-encoder. We primarily focus on four visual attributes : *style, color, objects* and *action*. In addition, we also use the causal tracing framework adapted for text-to-image diffusion models to analyse the *viewpoint* and *count* attribute. The main reason for focusing on *style, color, objects* and *action* is the fact that generations from current text-to-image models have a strong fidelity to these attributes, whereas the generations corresponding to *viewpoint* or *count* are often error-prone. We generate probe captions for each of the attribute in the following way:

- ***Objects***. We select a list of 24 objects and 7 locations (e.g., *beach, forest, city, kitchen, mountain, park, room*) to create a set of 168 unique captions. The objects are : *{ ‘dog’, ‘cat’, ‘bicycle’, ‘oven’, ‘tv’, ‘bowl’, ‘banana’, ‘bottle’, ‘cup’, ‘fork’, ‘knife’, ‘apple’, ‘sandwich’, ‘pizza’, ‘bed’, ‘tv’, ‘laptop’, ‘microwave’, ‘book’, ‘clock’, ‘vase’, ‘toothbrush’, ‘donut’, ‘handbag’ }* . We then use the template: *‘A photo of <object> in <location>.’* to generate multiple captions to probe the text-to-image model. These objects are selected from the list of objects present in MS-COCO.
- ***Style***. We select the list of 80 unique objects from MS-COCO and combine it with an artistic style from : *{monet, pablo picasso, salvador dali, van gogh, baroque, leonardo da vinci, michelangelo}* . The template used is: *‘A <object> in the style of <artistic-style>’*. In total, using this template we generate 560 captions.
- ***Color***. We select the list of 80 unique objects from MS-COCO and combine it with a color from *{ blue, red, yellow, brown, black, pink, green}*. We then use the template: *‘A <color> <object>’* to generate 560 unique captions to probe the text-to-image model.
- ***Action***. We first choose certain actions such as *eating, running, sitting, sprinting* and ask ChatGPT[3] to list a set of animals who can perform these actions. From this list, we choose a set of 14 animals: *{ bear, cat, deer, dog, giraffe, goat, horse, lion, monkey, mouse, ostrich, sheep, tiger, wolf}*. In total we use the template: *‘An <animal><action>’* to create a set of 56 unique captions for probing.
- ***Viewpoint***. For viewpoint, we use the same set of 24 objects from the *Objects* attribute and combine it with a viewpoint selected from *{front, back, top, bottom}* to create 96 captions in the template of: *‘A <object> from the <viewpoint>’*.
- ***Count***. We use the same 24 objects from *Objects* attribute and introduce a count before the object in the caption. We specifically use the following template to generate captions: *‘A photo of <count> objects in a room.’*, where we keep the location fixed. We select a count from {2,4,6,8} to create 96 unique captions in total.

	Description of Probe Dataset for Causal Tracing			
Attribute	Description	Example 1	Example 2	Example 3
Objects	Prompt containing an object in a location	photo of a *vase* in a room	a photo of a *car* in a desert	a photo of a *house* in a forest
Style	An object drawn in a particular artistic style	airplane in the style of *van gogh*	town in the style of *monet*	bicycle in the style of *baroque*
Color	An object in a particular color	a *blue* car	a *black* vase	a *pink* bag
Action	An animal in a particular action	A giraffe *eating*	A tiger *running*	A cat *standing*
Viewpoint	An object in a particular viewpoint	A sofa from the *back*	A car from the *front*	A bus from the *side*
Count	Number of objects in a location	There are *10* cars on the road	*5 bags* in the room	*6* laptops on a table

Table 1: **Examples from the Probe Dataset Used For Causal Tracing.** The attributes in the captions are marked in *italics*.

---

[3] We use version 3.5 for ChatGPT.

B Qualitative Visualizations for Causal Tracing (UNet)

B.1 Objects

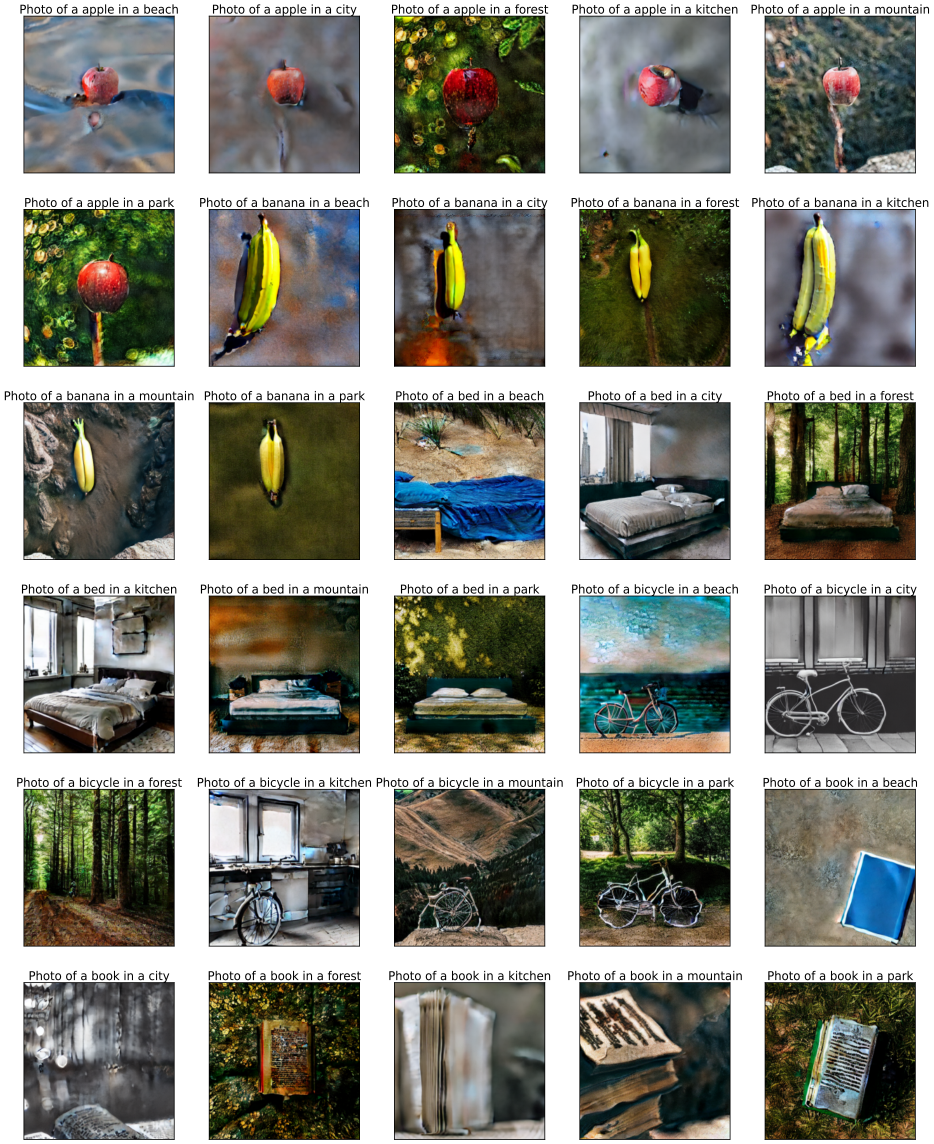


Photo of a apple in a beach
Photo of a apple in a city
Photo of a apple in a forest
Photo of a apple in a kitchen
Photo of a apple in a mountain
Photo of a apple in a park
Photo of a banana in a beach
Photo of a banana in a city
Photo of a banana in a forest
Photo of a banana in a kitchen
Photo of a banana in a mountain
Photo of a banana in a park
Photo of a bed in a beach
Photo of a bed in a city
Photo of a bed in a forest
Photo of a bed in a kitchen
Photo of a bed in a mountain
Photo of a bed in a park
Photo of a bicycle in a beach
Photo of a bicycle in a city
Photo of a bicycle in a forest
Photo of a bicycle in a kitchen
Photo of a bicycle in a mountain
Photo of a bicycle in a park
Photo of a book in a beach
Photo of a book in a city
Photo of a book in a forest
Photo of a book in a kitchen
Photo of a book in a mountain
Photo of a book in a park

Figure 6: **Causal State: down-1-resnet-1.** We find that restoring the down-1-resnet-1 block in the UNet leads to generation of images with strong fidelity to the original caption.

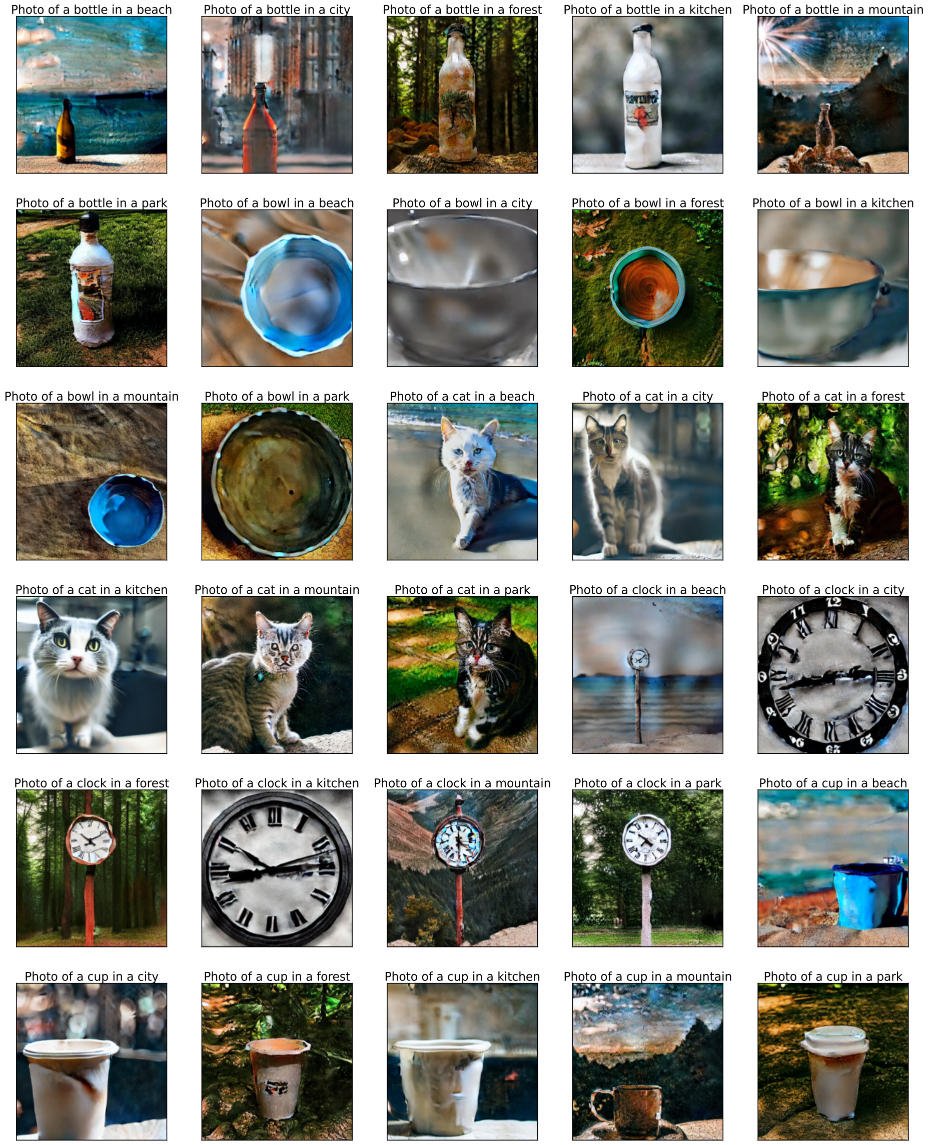


Photo of a bottle in a beach
Photo of a bottle in a city
Photo of a bottle in a forest
Photo of a bottle in a kitchen
Photo of a bottle in a mountain
Photo of a bottle in a park
Photo of a bowl in a beach
Photo of a bowl in a city
Photo of a bowl in a forest
Photo of a bowl in a kitchen
Photo of a bowl in a mountain
Photo of a bowl in a park
Photo of a cat in a beach
Photo of a cat in a city
Photo of a cat in a forest
Photo of a cat in a kitchen
Photo of a cat in a mountain
Photo of a cat in a park
Photo of a clock in a beach
Photo of a clock in a city
Photo of a clock in a forest
Photo of a clock in a kitchen
Photo of a clock in a mountain
Photo of a clock in a park
Photo of a cup in a beach
Photo of a cup in a city
Photo of a cup in a forest
Photo of a cup in a kitchen
Photo of a cup in a mountain
Photo of a cup in a park

Figure 7: **Causal State: down-1-resnet-1.** We find that restoring the down-1-resnet-1 block in the UNet leads to generation of images with strong fidelity to the original caption.

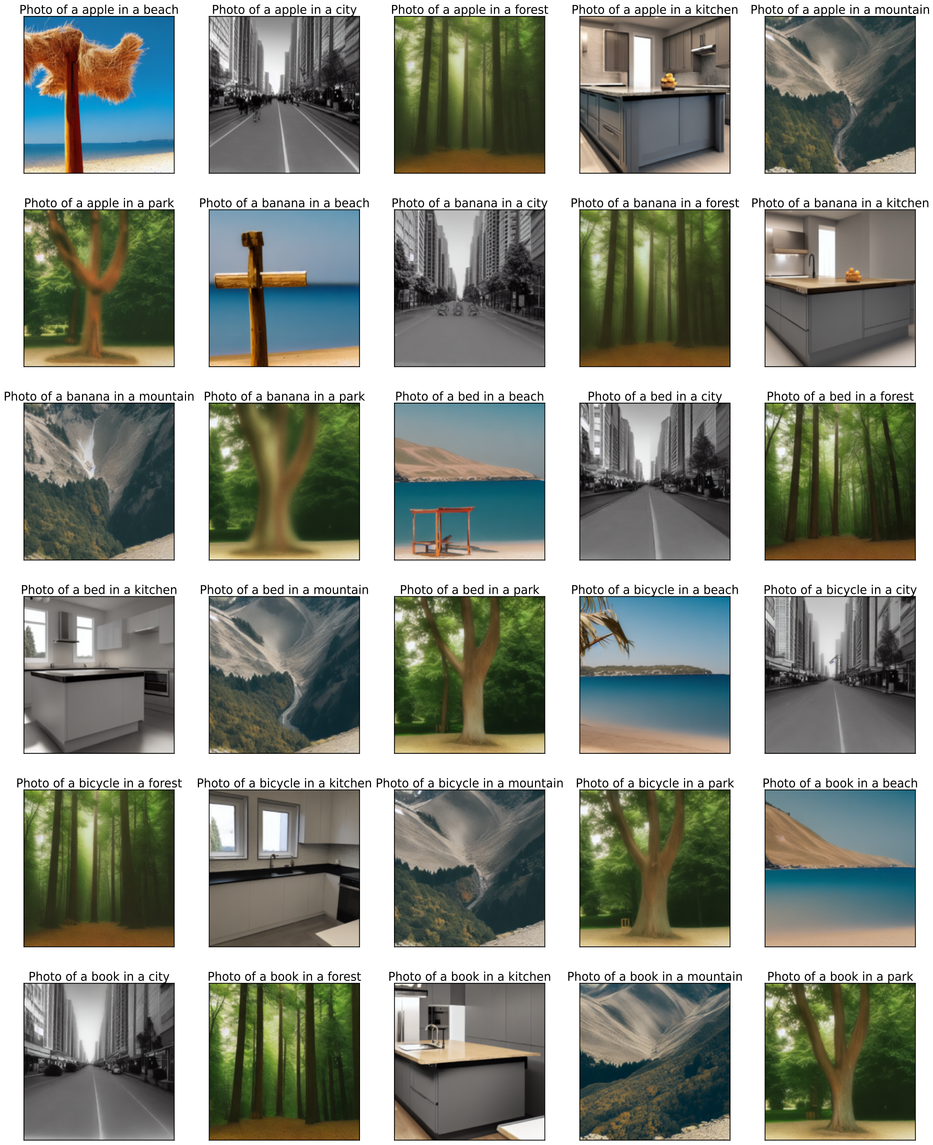


Photo of a apple in a beach
Photo of a apple in a city
Photo of a apple in a forest
Photo of a apple in a kitchen
Photo of a apple in a mountain
Photo of a apple in a park
Photo of a banana in a beach
Photo of a banana in a city
Photo of a banana in a forest
Photo of a banana in a kitchen
Photo of a banana in a mountain
Photo of a banana in a park
Photo of a bed in a beach
Photo of a bed in a city
Photo of a bed in a forest
Photo of a bed in a kitchen
Photo of a bed in a mountain
Photo of a bed in a park
Photo of a bicycle in a beach
Photo of a bicycle in a city
Photo of a bicycle in a forest
Photo of a bicycle in a kitchen
Photo of a bicycle in a mountain
Photo of a bicycle in a park
Photo of a book in a beach
Photo of a book in a city
Photo of a book in a forest
Photo of a book in a kitchen
Photo of a book in a mountain
Photo of a book in a park

Figure 8: **Non-Causal State: down-blocks.0.attentions.0.transformer-blocks.0.attn2.** We find that restoring the down-blocks.0.attentions.0.transformer-blocks.0.attn2 block in the UNet leads to generation of images **without** the primary object, showing low fidelity to the original captions.

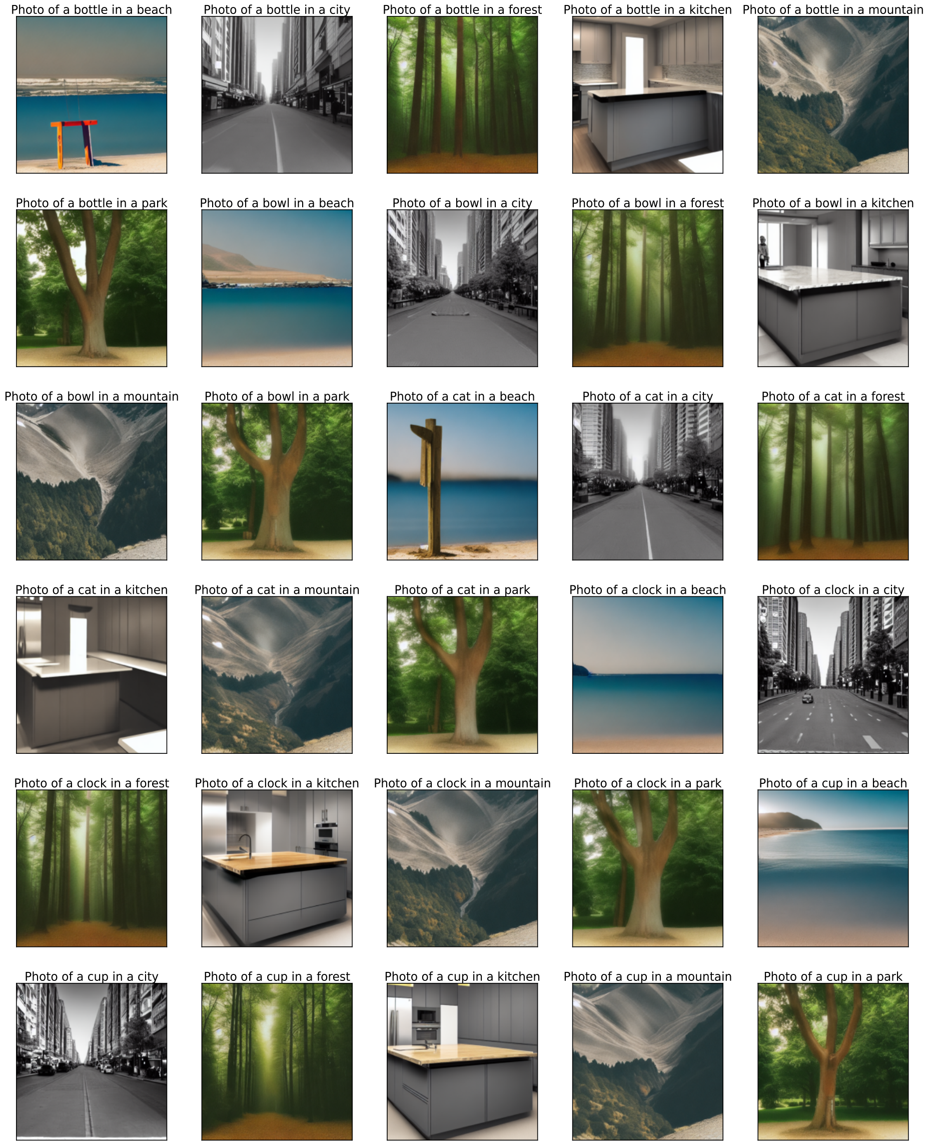
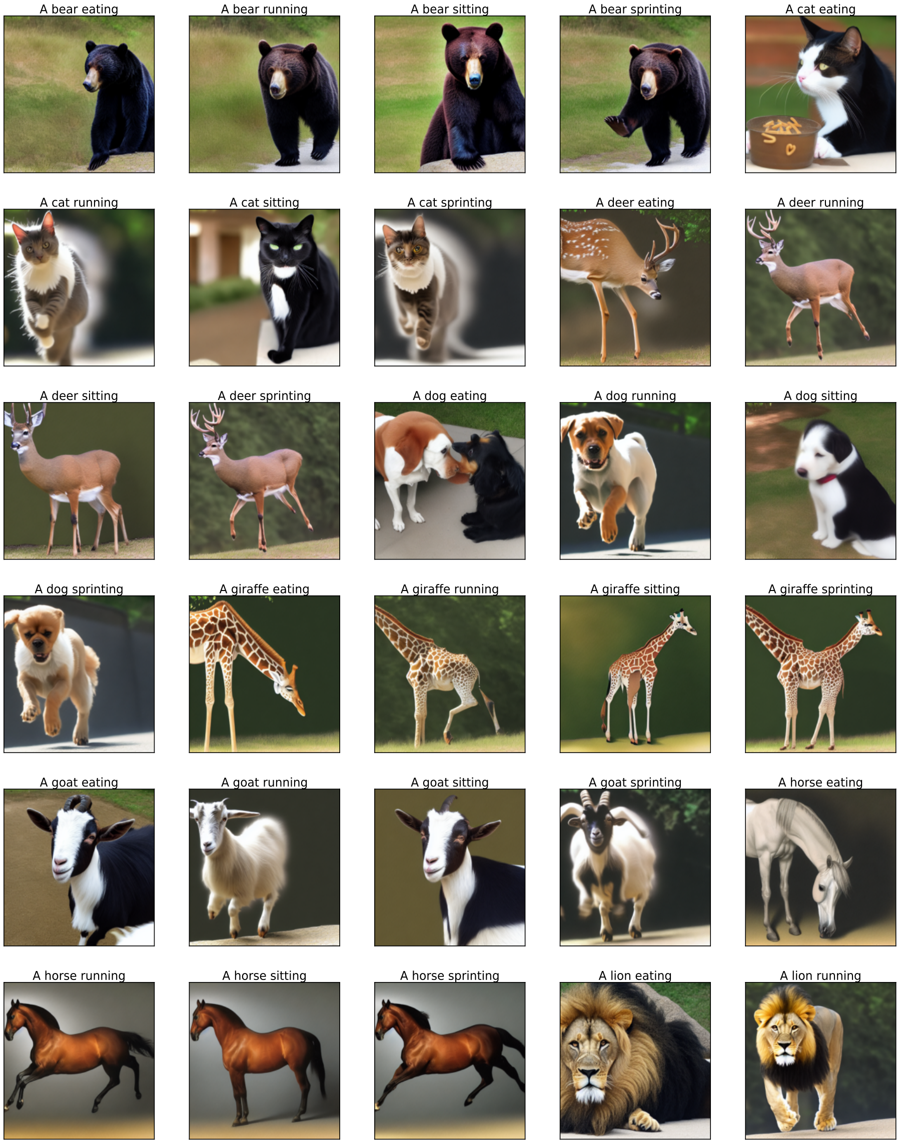


Photo of a bottle in a beach
Photo of a bottle in a city
Photo of a bottle in a forest
Photo of a bottle in a kitchen
Photo of a bottle in a mountain
Photo of a bottle in a park
Photo of a bowl in a beach
Photo of a bowl in a city
Photo of a bowl in a forest
Photo of a bowl in a kitchen
Photo of a bowl in a mountain
Photo of a bowl in a park
Photo of a cat in a beach
Photo of a cat in a city
Photo of a cat in a forest
Photo of a cat in a kitchen
Photo of a cat in a mountain
Photo of a cat in a park
Photo of a clock in a beach
Photo of a clock in a city
Photo of a clock in a forest
Photo of a clock in a kitchen
Photo of a clock in a mountain
Photo of a clock in a park
Photo of a cup in a beach
Photo of a cup in a city
Photo of a cup in a forest
Photo of a cup in a kitchen
Photo of a cup in a mountain
Photo of a cup in a park

Figure 9: **Non-Causal State: down-blocks.0.attentions.0.transformer-blocks.0.attn2.** We find that restoring the down-blocks.0.attentions.0.transformer-blocks.0.attn2 block in the UNet leads to generation of images **without** the primary object, showing low fidelity to the original captions.

B.2 ACTION



A bear eating
A bear running
A bear sitting
A bear sprinting
A cat eating
A cat running
A cat sitting
A cat sprinting
A deer eating
A deer running
A deer sitting
A deer sprinting
A dog eating
A dog running
A dog sitting
A dog sprinting
A giraffe eating
A giraffe running
A giraffe sitting
A giraffe sprinting
A goat eating
A goat running
A goat sitting
A goat sprinting
A horse eating
A horse running
A horse sitting
A horse sprinting
A lion eating
A lion running

Figure 10: **Causal State: mid-block-cross-attn.** We find that restoring the cross-attn in the mid-block in the UNet leads to generation of images with strong fidelity to the *action* attribute in the original caption.

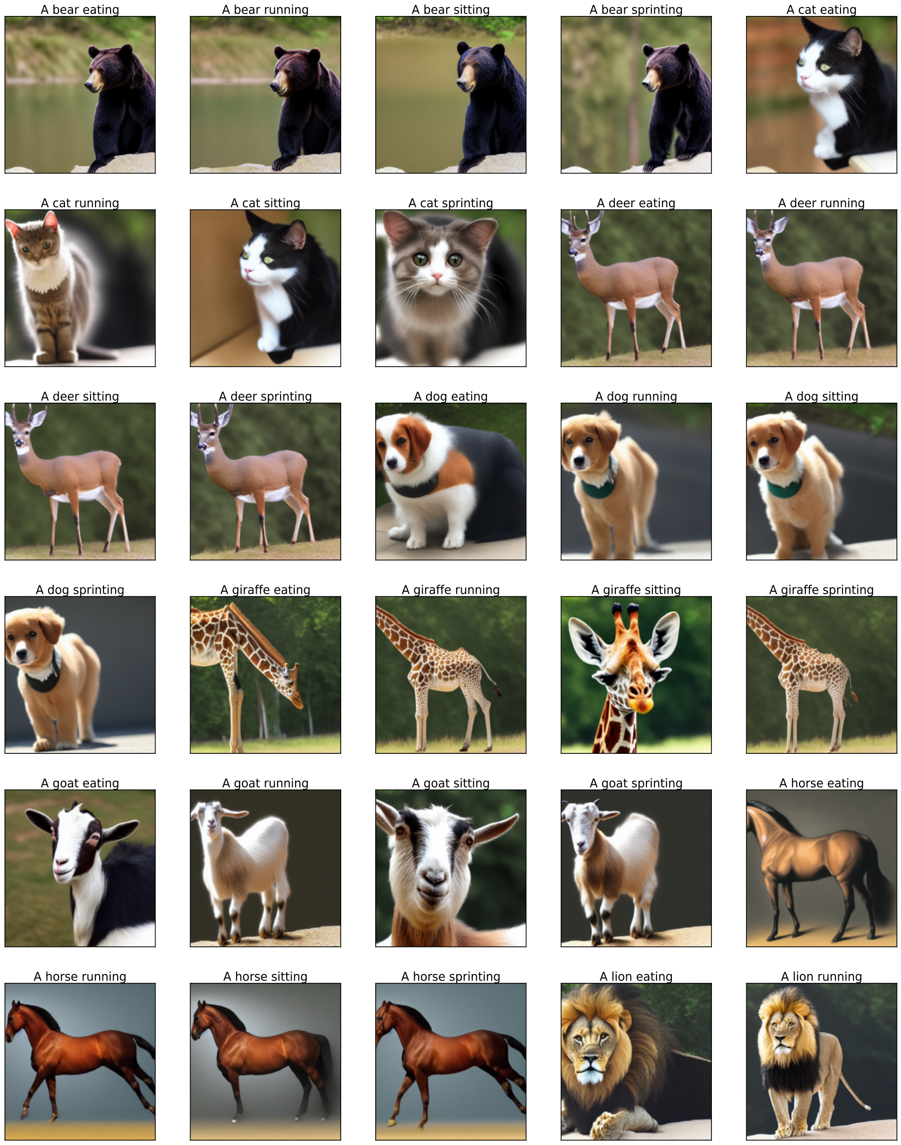


Figure 11: **Non-Causal State: down-blocks.2.attentions.1.transformer-blocks.0.attn2.** We find that restoring the down-blocks.2.attentions.1.transformer-blocks.0.attn2 in the UNet leads to generation of images with weak fidelity to the *action* attribute in the original caption. For a majority of the prompts, we find that the *action* attribute (especially those involving sprinting, running or eating) is not respected in the generated image.

B.3 COLOR

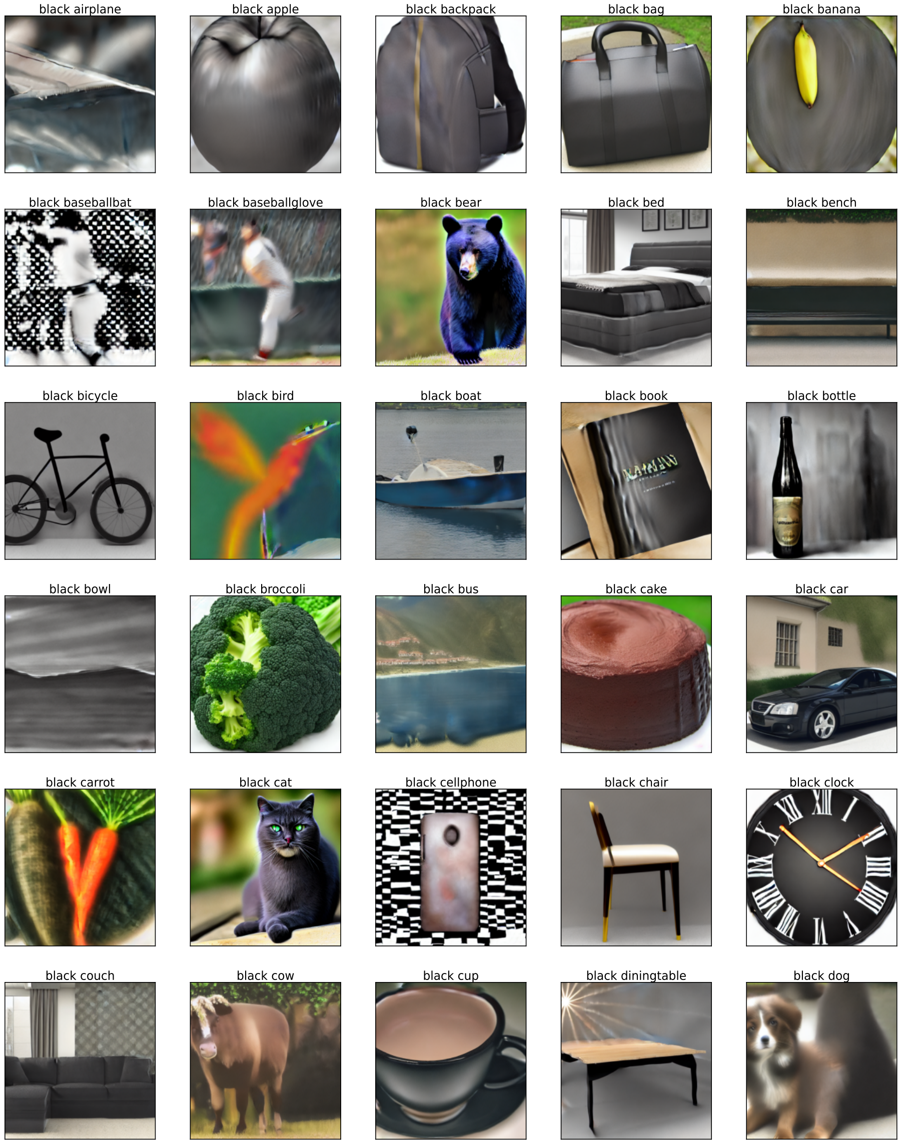


Figure 12: **Causal State: down-blocks.1.attentions.0.transformer-blocks.0.ff.** We find that restoring the down-blocks.1.attentions.0.transformer-blocks.0.ff in the down-block in the UNet leads to generation of images with strong fidelity to the *color* attribute in the original caption.

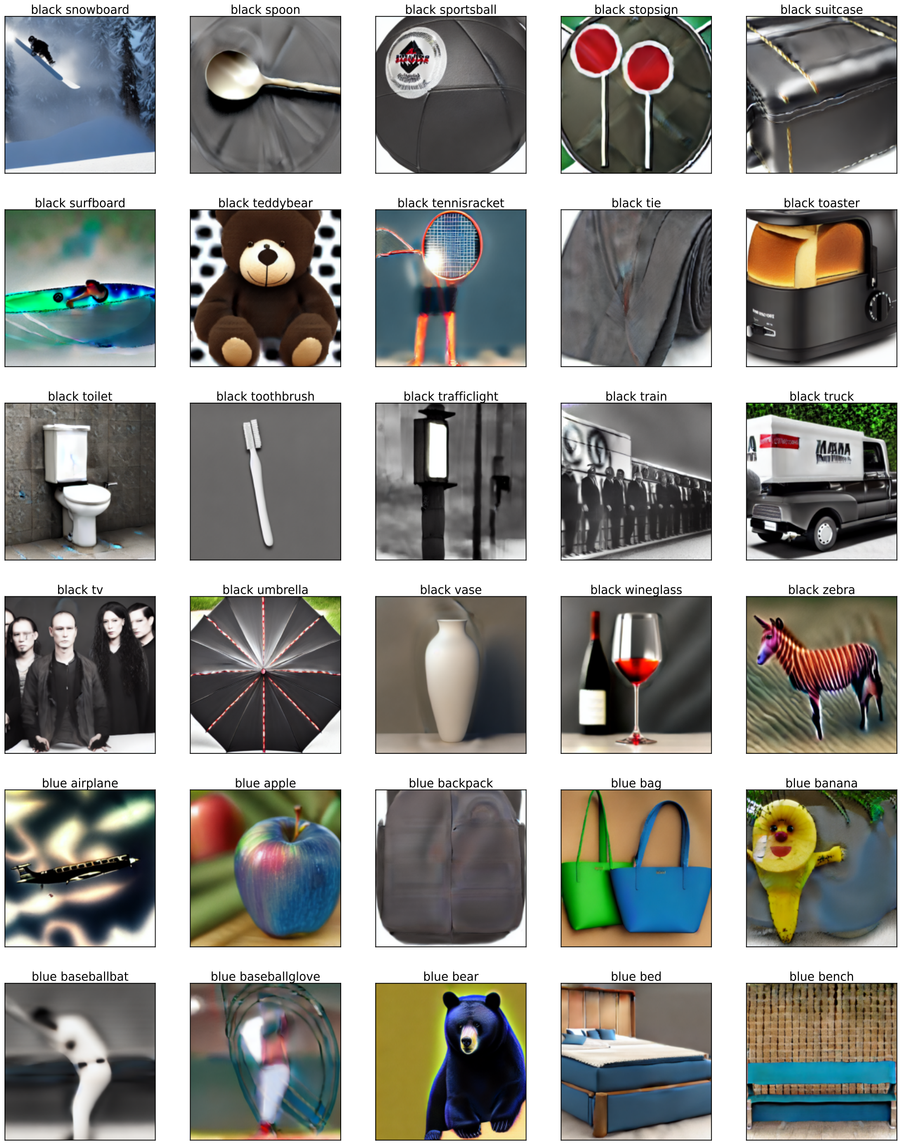


Figure 13: **Causal State: down-blocks.1.attentions.0.transformer-blocks.0.ff.** We find that restoring the down-blocks.1.attentions.0.transformer-blocks.0.ff in the down-block in the UNet leads to generation of images with strong fidelity to the *color* attribute in the original caption.

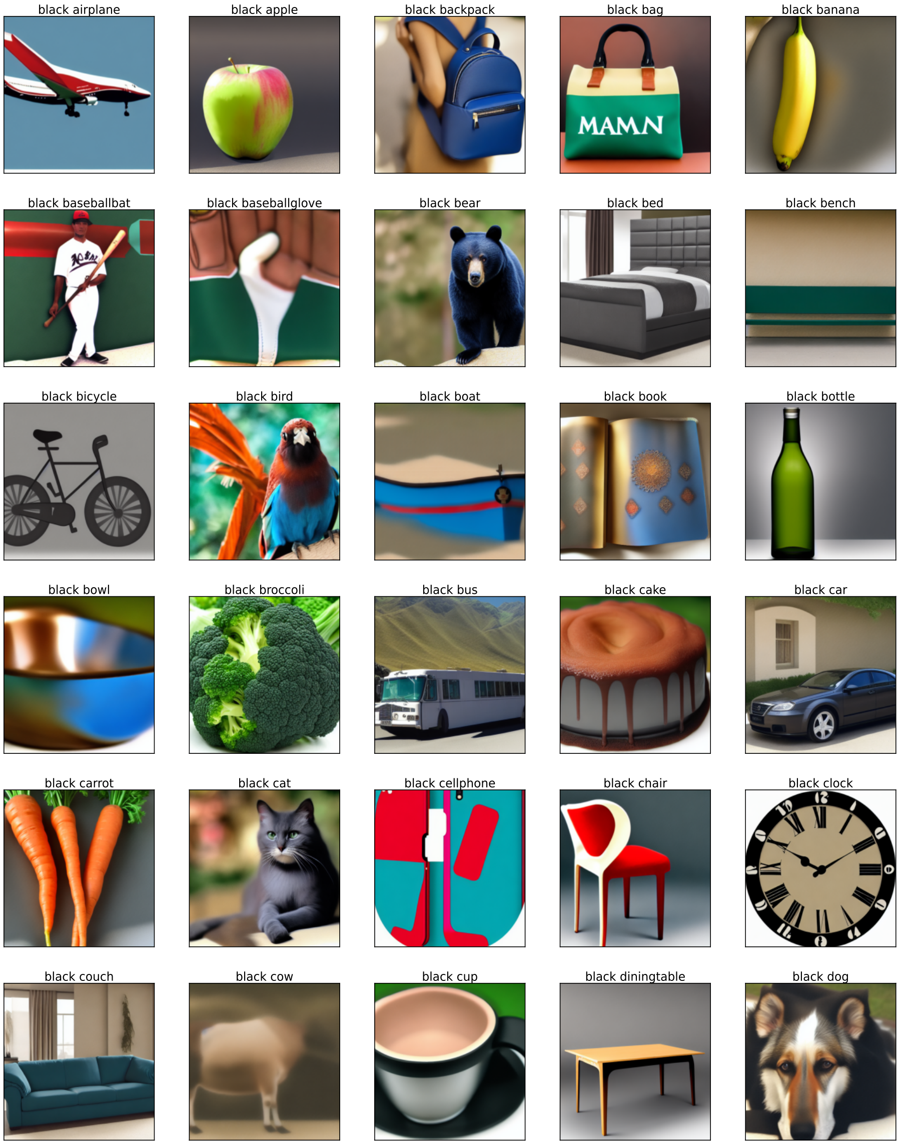


Figure 14: **Non-Causal State: mid-blocks.attentions.0.transformer-blocks.0.ff.** We find that restoring the mid-blocks.attentions.0.transformer-blocks.0.ff in the mid-block in the UNet does not lead to generation of images with strong fidelity to the *color* attribute in the original caption for a majority of cases.

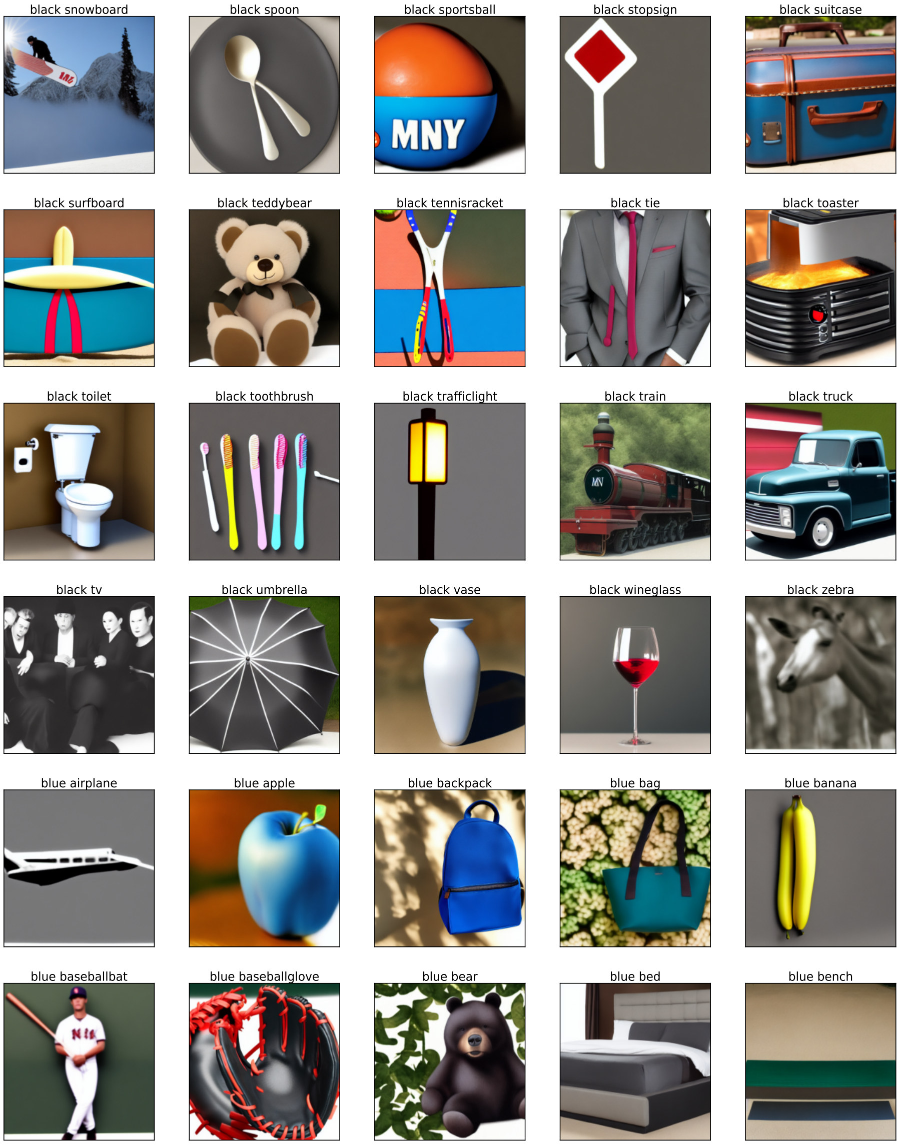


Figure 15: **Non-Causal State: mid-blocks.attentions.0.transformer-blocks.0.ff.** We find that restoring the down-mid-blocks.attentions.0.transformer-blocks.0.ff in the mid-block in the UNet does not lead to generation of images with strong fidelity to the *color* attribute in the original caption for a majority of cases.

B.4 STYLE

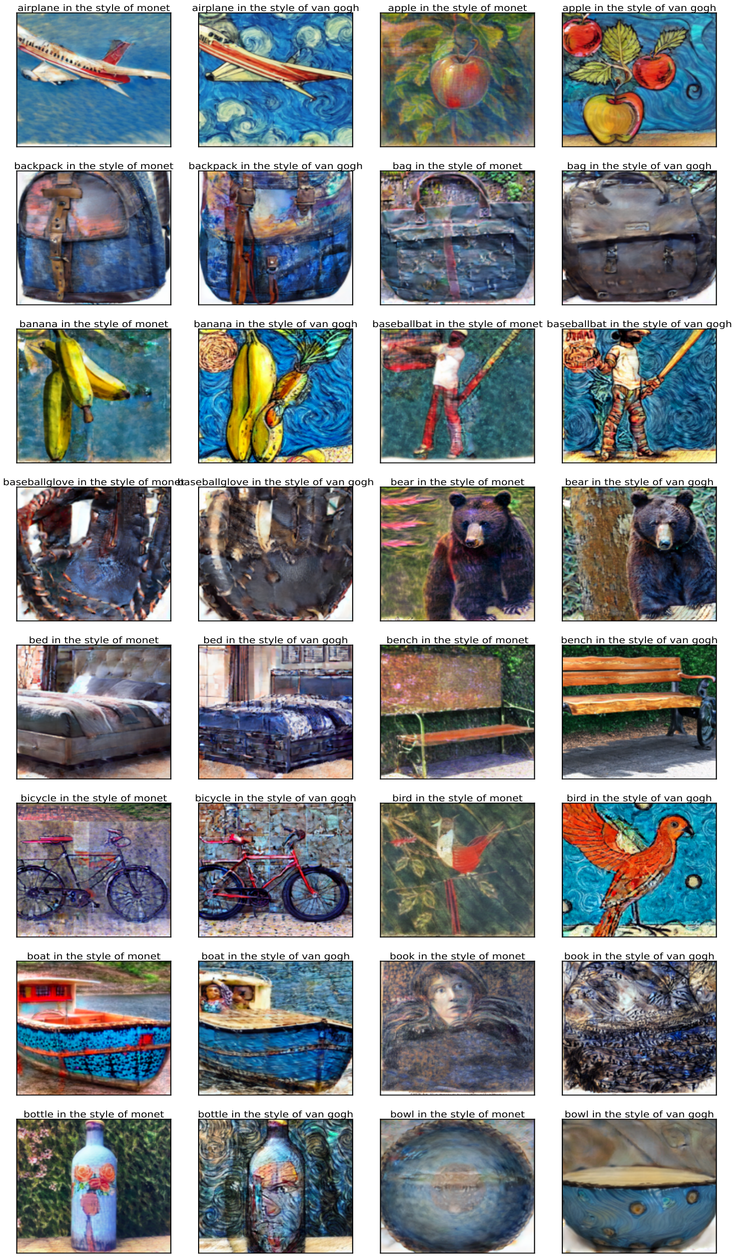

airplane in the style of monet
airplane in the style of van gogh
apple in the style of monet
apple in the style of van gogh
backpack in the style of monet
backpack in the style of van gogh
bag in the style of monet
bag in the style of van gogh
banana in the style of monet
banana in the style of van gogh
baseballbat in the style of monet
baseballbat in the style of van gogh
baseballglove in the style of monet
baseballglove in the style of van gogh
bear in the style of monet
bear in the style of van gogh
bed in the style of monet
bed in the style of van gogh
bench in the style of monet
bench in the style of van gogh
bicycle in the style of monet
bicycle in the style of van gogh
bird in the style of monet
bird in the style of van gogh
boat in the style of monet
boat in the style of van gogh
book in the style of monet
book in the style of van gogh
bottle in the style of monet
bottle in the style of van gogh
bowl in the style of monet
bowl in the style of van gogh

Figure 16: **Causal State: down-blocks.0.attn1 (First self-attn layer).** We find that restoring the down-blocks.self-attn.0 which is the first self-attention layer in the UNet **leads** to generation of images with strong fidelity to the *style* attribute in the original caption for a majority of cases.

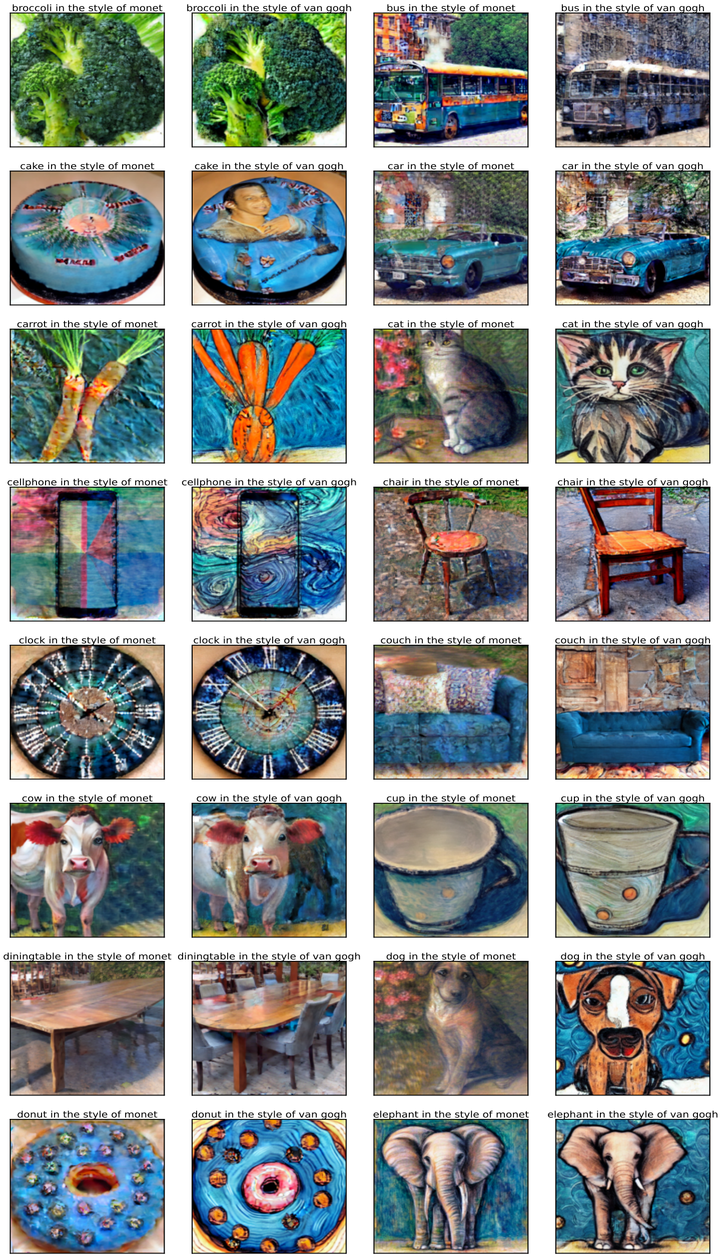

broccoli in the style of monet
broccoli in the style of van gogh
bus in the style of monet
bus in the style of van gogh
cake in the style of monet
cake in the style of van gogh
car in the style of monet
car in the style of van gogh
carrot in the style of monet
carrot in the style of van gogh
cat in the style of monet
cat in the style of van gogh
cellphone in the style of monet
cellphone in the style of van gogh
chair in the style of monet
chair in the style of van gogh
clock in the style of monet
clock in the style of van gogh
couch in the style of monet
couch in the style of van gogh
cow in the style of monet
cow in the style of van gogh
cup in the style of monet
cup in the style of van gogh
diningtable in the style of monet
diningtable in the style of van gogh
dog in the style of monet
dog in the style of van gogh
donut in the style of monet
donut in the style of van gogh
elephant in the style of monet
elephant in the style of van gogh

Figure 17: **Causal State: down-blocks.0.attn1.** We find that restoring the down-blocks.self-attn.0 which is the first self-attention layer in the UNet **leads** to generation of images with strong fidelity to the *style* attribute in the original caption for a majority of cases.

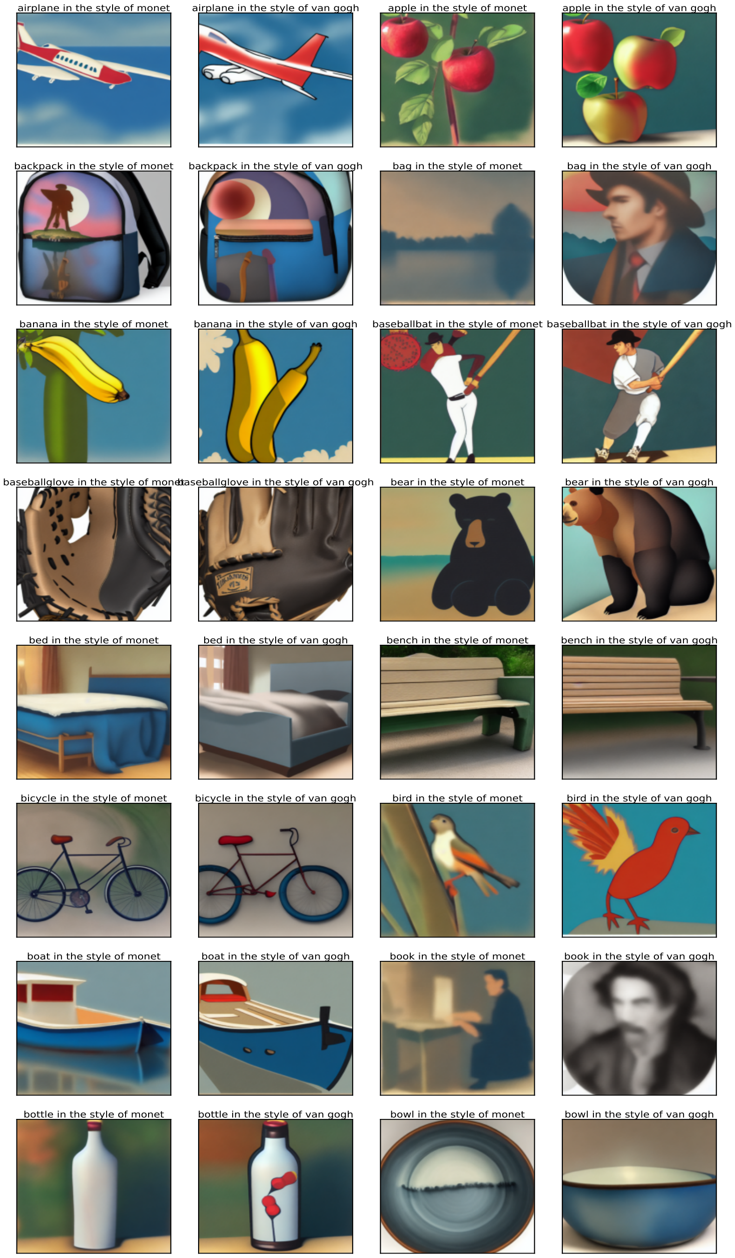

airplane in the style of monet
airplane in the style of van gogh
apple in the style of monet
apple in the style of van gogh
backpack in the style of monet
backpack in the style of van gogh
bag in the style of monet
bag in the style of van gogh
banana in the style of monet
banana in the style of van gogh
baseballbat in the style of monet
baseballbat in the style of van gogh
baseballglove in the style of monet
baseballglove in the style of van gogh
bear in the style of monet
bear in the style of van gogh
bed in the style of monet
bed in the style of van gogh
bench in the style of monet
bench in the style of van gogh
bicycle in the style of monet
bicycle in the style of van gogh
bird in the style of monet
bird in the style of van gogh
boat in the style of monet
boat in the style of van gogh
book in the style of monet
book in the style of van gogh
bottle in the style of monet
bottle in the style of van gogh
bowl in the style of monet
bowl in the style of van gogh

Figure 18: **Non-Causal State: down-blocks.2.attentions.1.transformer-blocks.0.attn2** We find that restoring the down-blocks.2.attentions.1.transformer-blocks.0.attn2 in the UNet **does not** lead to generation of images with strong fidelity to the *style* attribute in the original caption for a majority of cases.

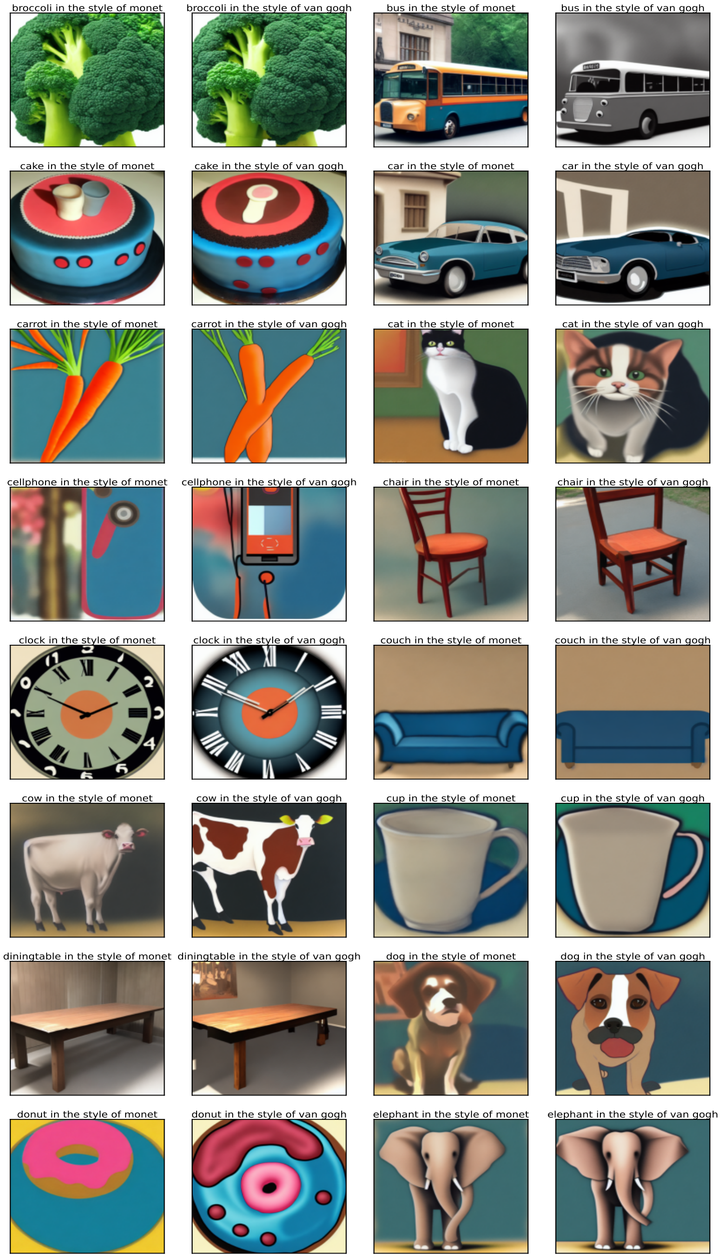

broccoli in the style of monet
broccoli in the style of van gogh
bus in the style of monet
bus in the style of van gogh
cake in the style of monet
cake in the style of van gogh
car in the style of monet
car in the style of van gogh
carrot in the style of monet
carrot in the style of van gogh
cat in the style of monet
cat in the style of van gogh
cellphone in the style of monet
cellphone in the style of van gogh
chair in the style of monet
chair in the style of van gogh
clock in the style of monet
clock in the style of van gogh
couch in the style of monet
couch in the style of van gogh
cow in the style of monet
cow in the style of van gogh
cup in the style of monet
cup in the style of van gogh
diningtable in the style of monet
diningtable in the style of van gogh
dog in the style of monet
dog in the style of van gogh
donut in the style of monet
donut in the style of van gogh
elephant in the style of monet
elephant in the style of van gogh

Figure 19: **Non-Causal State: down-blocks.2.attentions.1.transformer-blocks.0.attn2.** We find that restoring the down-blocks.2.attentions.1.transformer-blocks.0.attn2 which is the first self-attention layer in the UNet **does not** lead to generation of images with strong fidelity to the *style* attribute in the original caption for a majority of cases.

C QUALITATIVE VISUALIZATIONS FOR CAUSAL TRACING (TEXT-ENCODER)

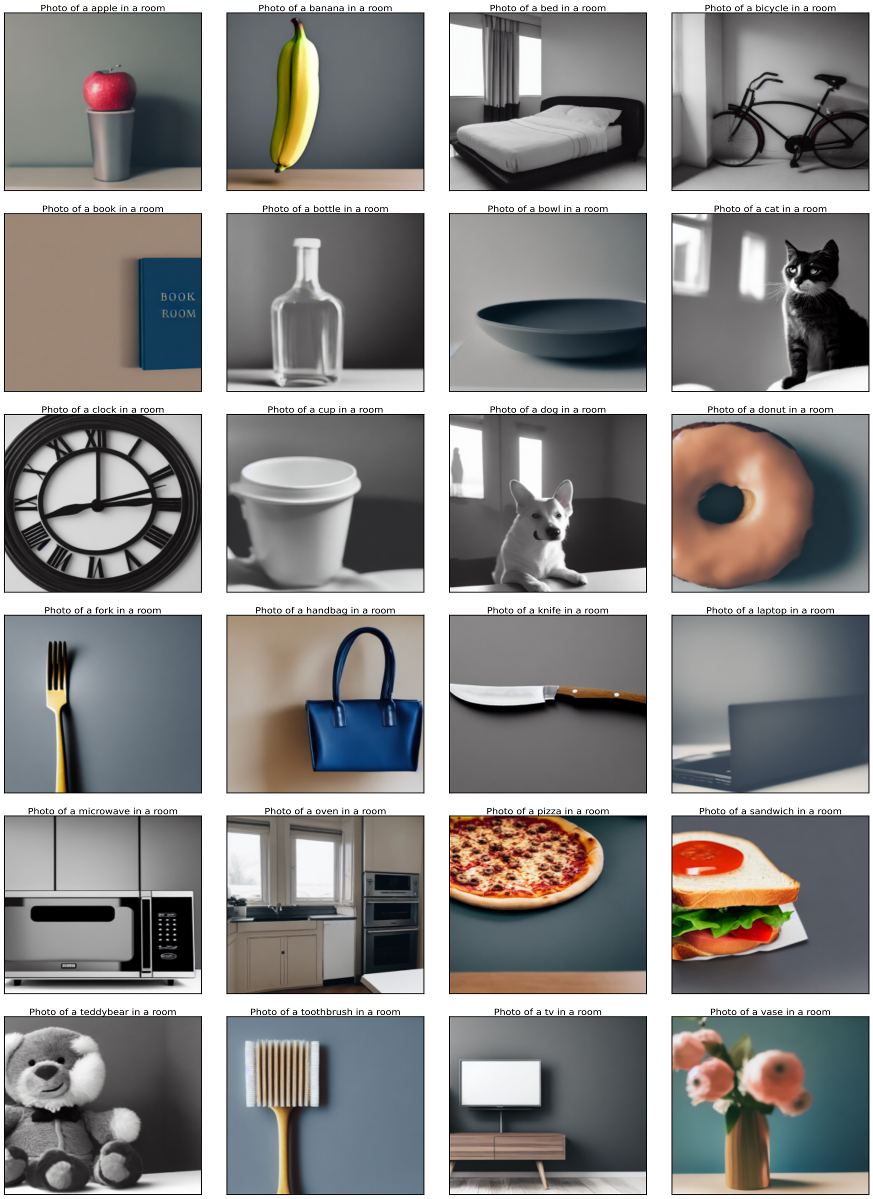


Photo of a apple in a room
Photo of a banana in a room
Photo of a bed in a room
Photo of a bicycle in a room
Photo of a book in a room
BOOK
ROOM
Photo of a bottle in a room
Photo of a bowl in a room
Photo of a cat in a room
Photo of a clock in a room
Photo of a cup in a room
Photo of a dog in a room
Photo of a donut in a room
Photo of a fork in a room
Photo of a handbag in a room
Photo of a knife in a room
Photo of a laptop in a room
Photo of a microwave in a room
Photo of a oven in a room
Photo of a pizza in a room
Photo of a sandwich in a room
Photo of a teddybear in a room
Photo of a toothbrush in a room
Photo of a tv in a room
Photo of a vase in a room

Figure 20: **Causal State: self-attn-0 corresponding to the last subject token.** We find that restoring the first self-attn layer which is the first self-attention layer in the text-encoder leads to generation of images with strong fidelity to the original caption for a majority of cases.

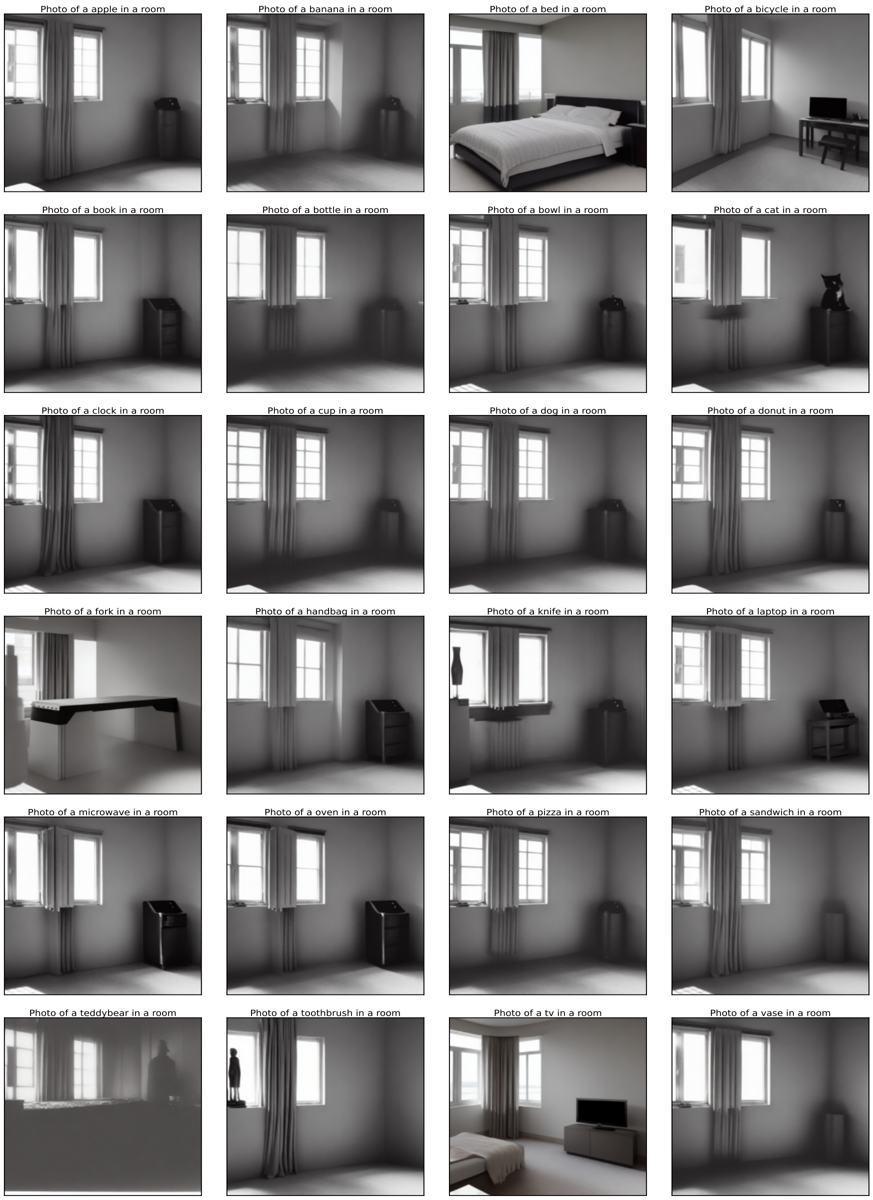


Photo of a apple in a room
Photo of a banana in a room
Photo of a bed in a room
Photo of a bicycle in a room
Photo of a book in a room
Photo of a bottle in a room
Photo of a bowl in a room
Photo of a cat in a room
Photo of a clock in a room
Photo of a cup in a room
Photo of a dog in a room
Photo of a donut in a room
Photo of a fork in a room
Photo of a handbag in a room
Photo of a knife in a room
Photo of a laptop in a room
Photo of a microwave in a room
Photo of a oven in a room
Photo of a pizza in a room
Photo of a sandwich in a room
Photo of a teddybear in a room
Photo of a toothbrush in a room
Photo of a tv in a room
Photo of a vase in a room

Figure 21: **Non-Causal State: self-attn-4 corresponding to the last subject token.** We find that restoring the fourth self-attn layer in the text-encoder **does not lead** to generation of images with strong fidelity to the original caption for a majority of cases.

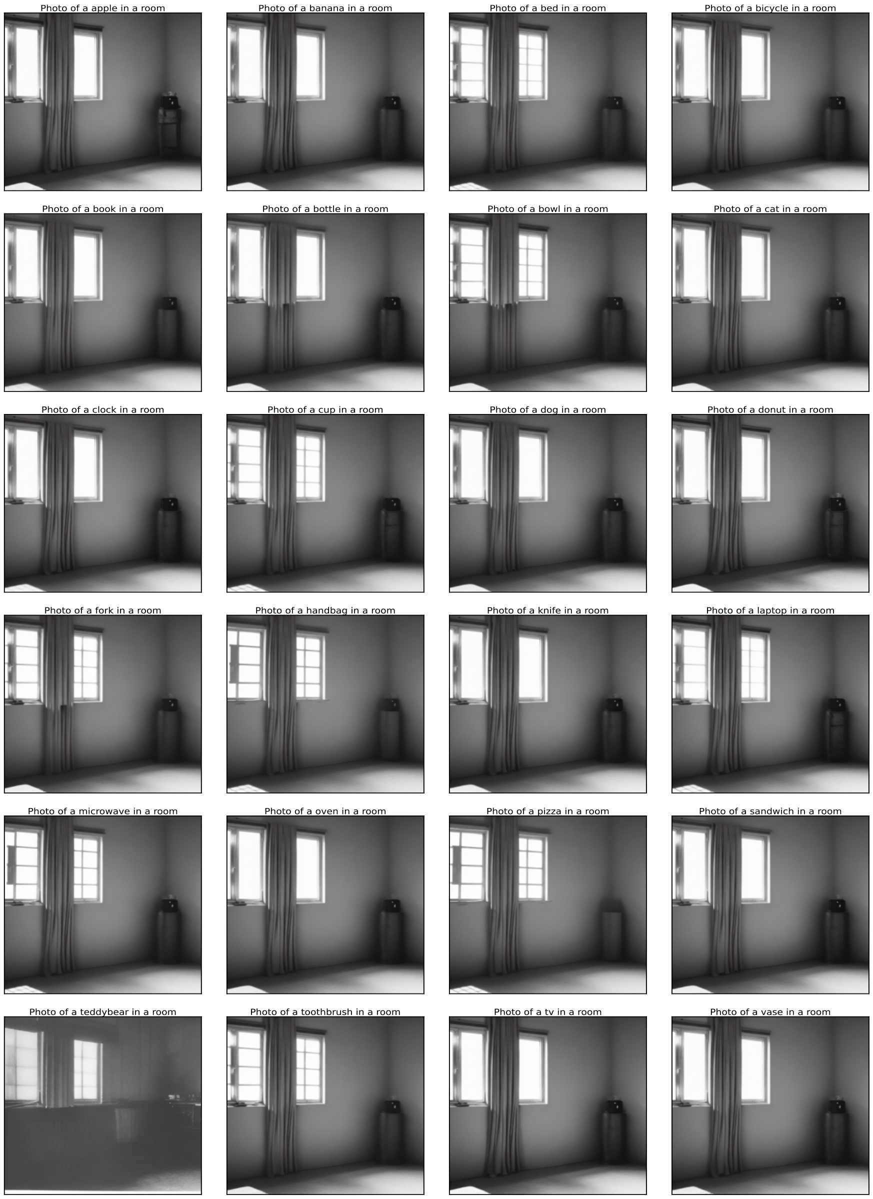
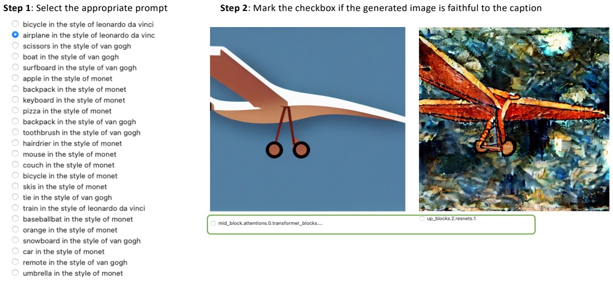


Photo of a apple in a room
Photo of a banana in a room
Photo of a bed in a room
Photo of a bicycle in a room
Photo of a book in a room
Photo of a bottle in a room
Photo of a bowl in a room
Photo of a cat in a room
Photo of a clock in a room
Photo of a cup in a room
Photo of a dog in a room
Photo of a donut in a room
Photo of a fork in a room
Photo of a handbag in a room
Photo of a knife in a room
Photo of a laptop in a room
Photo of a microwave in a room
Photo of a oven in a room
Photo of a pizza in a room
Photo of a sandwich in a room
Photo of a teddybear in a room
Photo of a toothbrush in a room
Photo of a tv in a room
Photo of a vase in a room

Figure 22: **Non-Causal State: self-attn-5 corresponding to the last subject token.** We find that restoring the fifth self-attn layer in the text-encoder **does not lead** to generation of images with strong fidelity to the original caption for a majority of cases.

D VALIDATION-SET DESIGN FOR CAUSAL TRACING

To select the threshold for `CLIP-Score` to be used at scale across the entirety of prompts (as shown in Appendix A), we use a small validation set of 10 prompts per attribute. In particular, we build a Jupyter notebook interface to select causal states for them. Once the causal states are marked, we select the common causal states across all the 10 prompts per attribute. Per causal state, we then compute the average `CLIP-Score` across the 10 prompts per attribute. Per attribute, we then select the lowest `CLIP-Score` corresponding to a causal state. These sets of `CLIP-Scores` per attribute is then used to filter the causal states and the non-causal states from the larger set of prompts in the probe dataset used in Appendix A.



Step 1: Select the appropriate prompt

○ bicycle in the style of leonardo da vinci
● airplane in the style of leonardo da vinc
○ scissors in the style of van gogh
○ boat in the style of van gogh
○ surfboard in the style of van gogh
○ apple in the style of monet
○ backpack in the style of monet
○ keyboard in the style of monet
○ pizza in the style of monet
○ backpack in the style of van gogh
○ toothbrush in the style of van gogh
○ hairdrier in the style of monet
○ mouse in the style of monet
○ couch in the style of monet
○ bicycle in the style of monet
○ skis in the style of monet
○ tie in the style of van gogh
○ train in the style of leonardo da vinci
○ baseballbat in the style of monet
○ orange in the style of monet
○ snowboard in the style of van gogh
○ car in the style of monet
○ remote in the style of van gogh
○ umbrella in the style of monet

Step 2: Mark the checkbox if the generated image is faithful to the caption

☐ mid_block.attentions.0.transformer_blocks....
☐ up_blocks.2.resnets.1

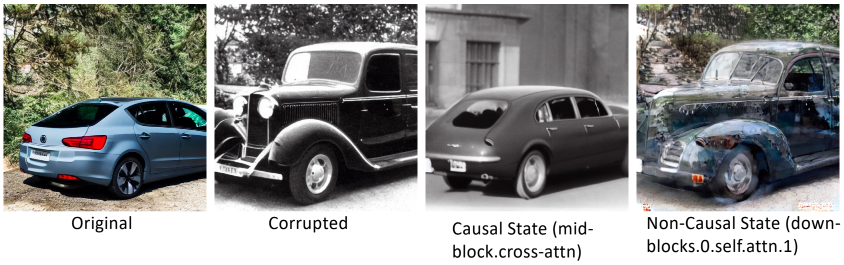
Figure 23: **Jupyter Notebook Interface for Marking Causal States.**

E Causal Tracing for Viewpoint and Count

In this section, we provide additional causal tracing results for the *viewpoint* and *count* attribute.

E.1 Viewpoint

Prompt: '*Photo of a car from the back viewpoint*'



Original

Corrupted

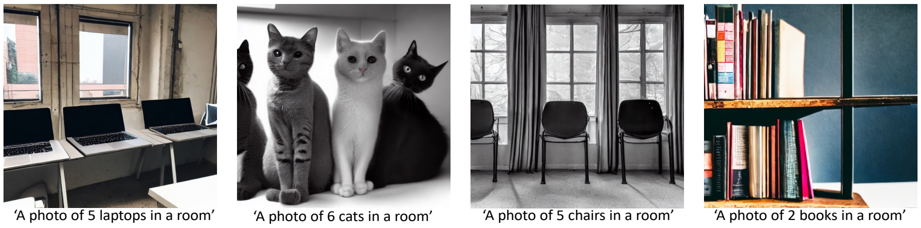
Causal State (mid-block.cross-attn)

Non-Causal State (down-blocks.0.self.attn.1)

Figure 24: **Illustration of a causal state for the viewpoint attribute.**

E.2 Count

For the *count attribute*, we find that public text-to-image generative models such as Stable-Diffusion cannot generate images with high-fidelity to the captions. Therefore, we do not use causal tracing for this attribute.



'A photo of 5 laptops in a room'

'A photo of 6 cats in a room'

'A photo of 5 chairs in a room'

'A photo of 2 books in a room'

Figure 25: **Illustration of failure cases of generation for the *count* attribute with the Original Clean model.**

F QUALITATIVE VISUALIZATIONS USING DIFF-QUICKFIX FOR ABLATING CONCEPTS

F.1 ABLATING ARTISTIC STYLES

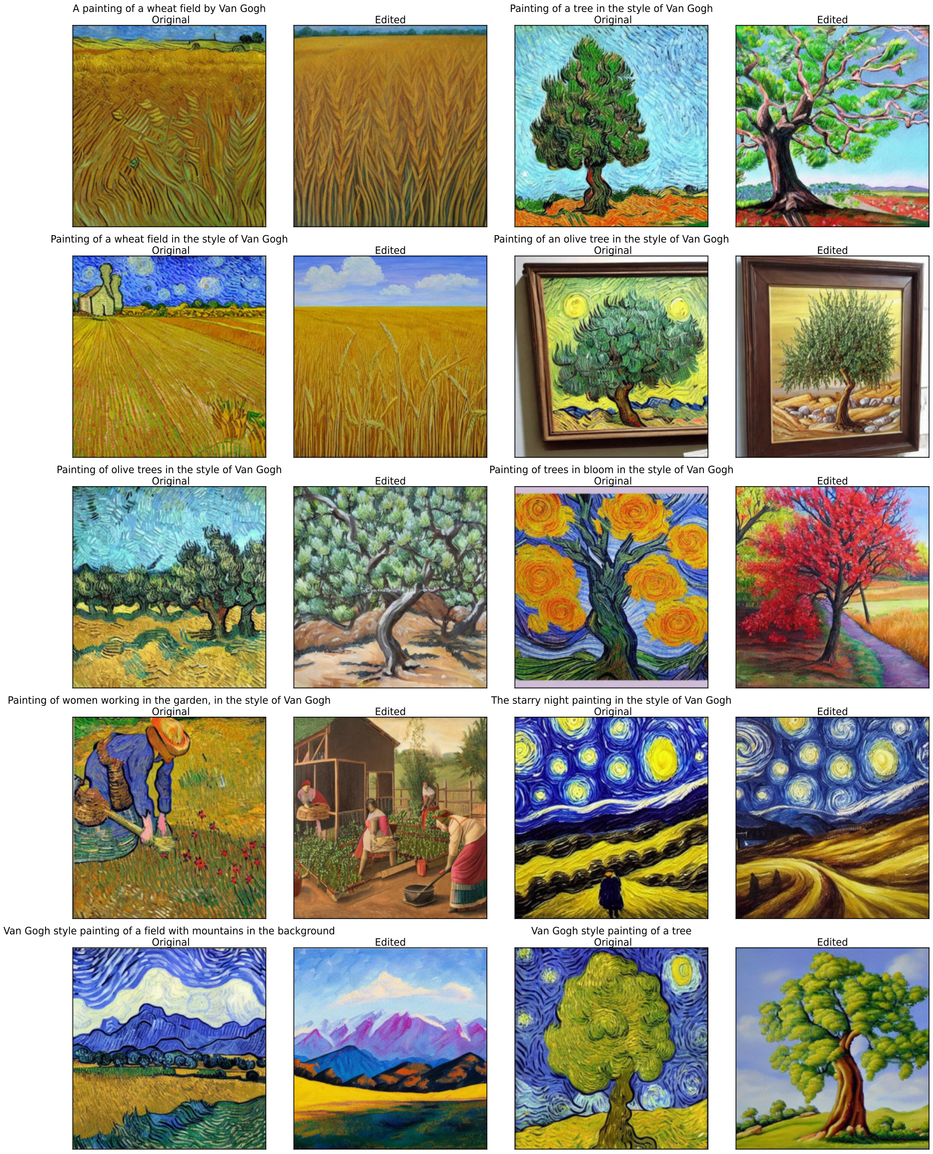

A painting of a wheat field by Van Gogh
Original
Edited
Painting of a tree in the style of Van Gogh
Original
Edited
Painting of a wheat field in the style of Van Gogh
Original
Edited
Painting of an olive tree in the style of Van Gogh
Original
Edited
Painting of olive trees in the style of Van Gogh
Original
Edited
Painting of trees in bloom in the style of Van Gogh
Original
Edited
Painting of women working in the garden, in the style of Van Gogh
Original
Edited
The starry night painting in the style of Van Gogh
Original
Edited
Van Gogh style painting of a field with mountains in the background
Original
Edited
Van Gogh style painting of a tree
Original
Edited

Figure 26: **Single-Concept Ablated Model: Generations with different** ***Van Gogh*** **Prompts.**

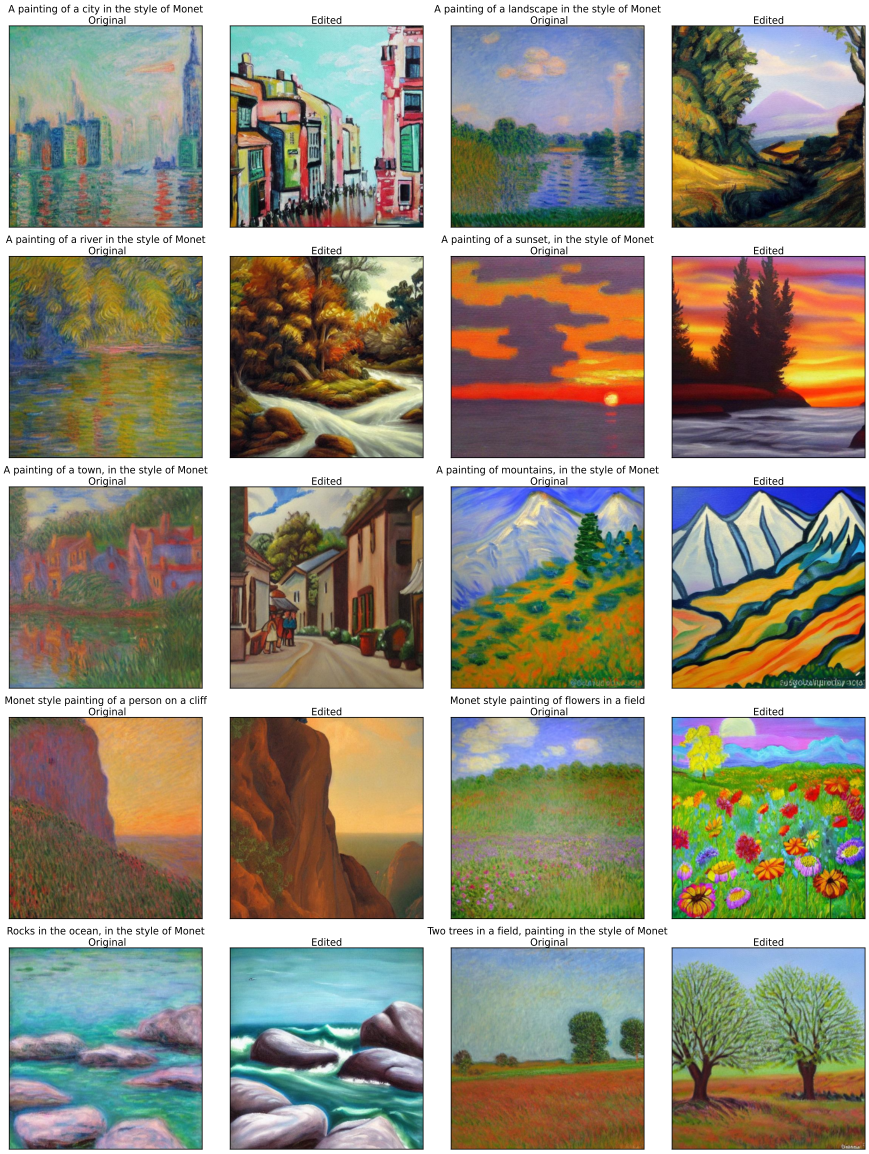

A painting of a city in the style of Monet
Original
Edited
A painting of a landscape in the style of Monet
Original
Edited
A painting of a river in the style of Monet
Original
Edited
A painting of a sunset, in the style of Monet
Original
Edited
A painting of a town, in the style of Monet
Original
Edited
A painting of mountains, in the style of Monet
Original
Edited
Monet style painting of a person on a cliff
Original
Edited
Monet style painting of flowers in a field
Original
Edited
Rocks in the ocean, in the style of Monet
Original
Edited
Two trees in a field, painting in the style of Monet
Original
Edited

Figure 27: **Single-Concept Ablated Model: Generations with different *Monet* Prompts.**

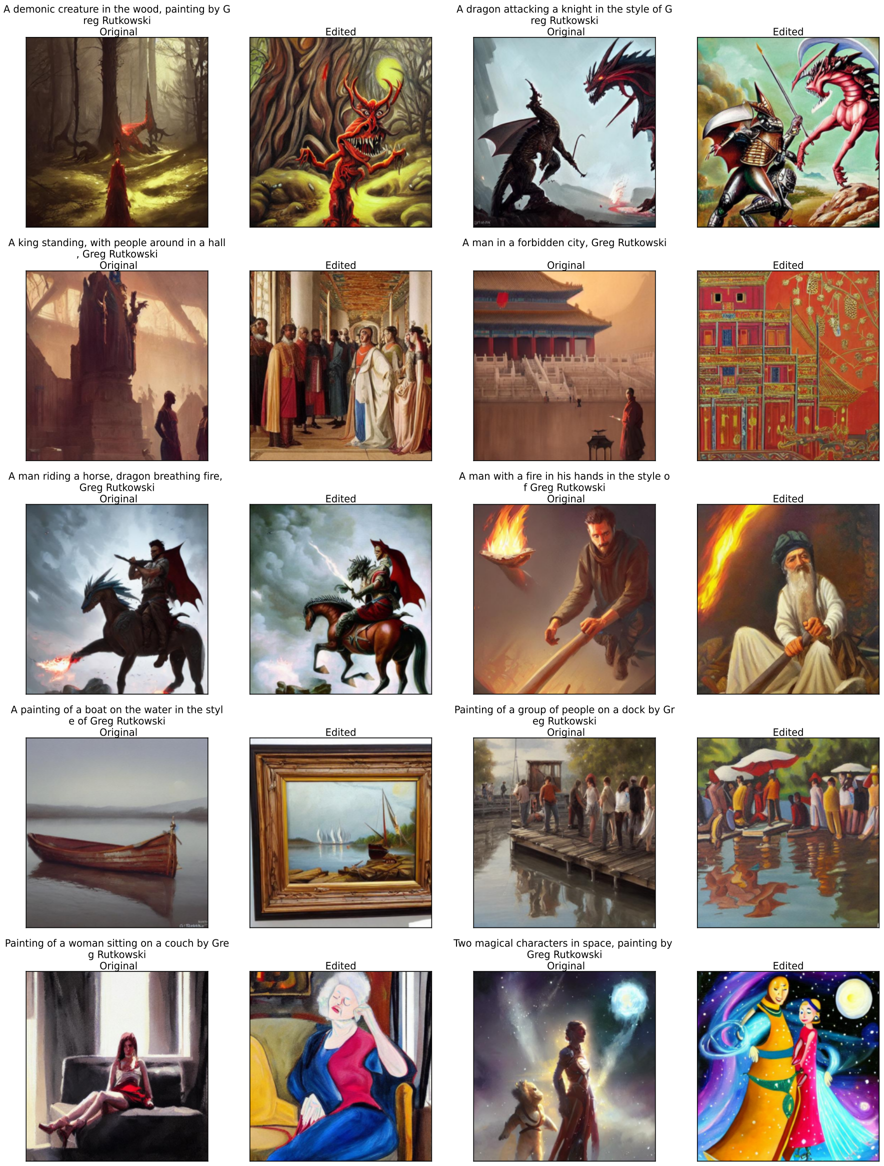

A demonic creature in the wood, painting by Greg Rutkowski
Original
Edited
A dragon attacking a knight in the style of Greg Rutkowski
Original
Edited
A king standing, with people around in a hall, Greg Rutkowski
Original
Edited
A man in a forbidden city, Greg Rutkowski
Original
Edited
A man riding a horse, dragon breathing fire, Greg Rutkowski
Original
Edited
A man with a fire in his hands in the style of Greg Rutkowski
Original
Edited
A painting of a boat on the water in the style of Greg Rutkowski
Original
Edited
Painting of a group of people on a dock by Greg Rutkowski
Original
Edited
Painting of a woman sitting on a couch by Greg Rutkowski
Original
Edited
Two magical characters in space, painting by Greg Rutkowski
Original
Edited

Figure 28: **Single-Concept Ablated Model: Generations with different** *Greg Rutkowski* **Prompts.**

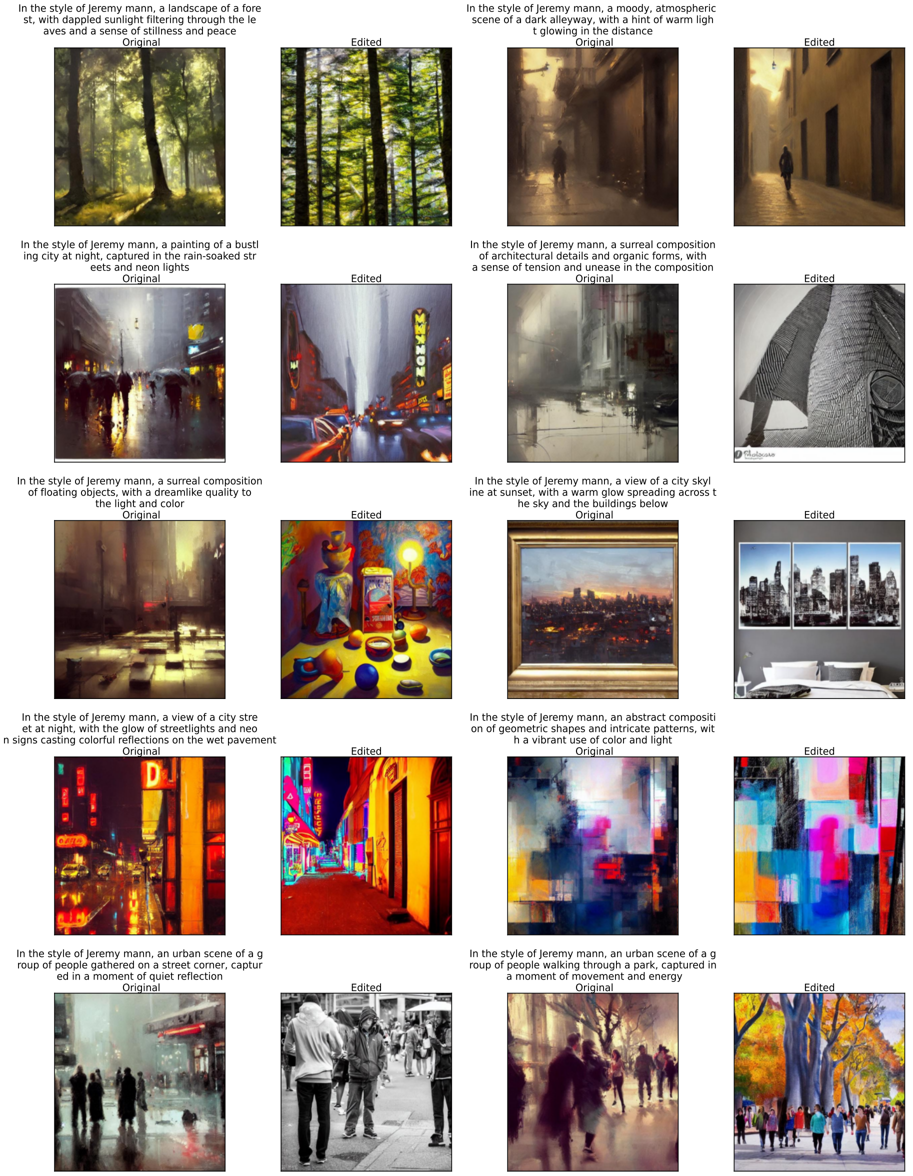
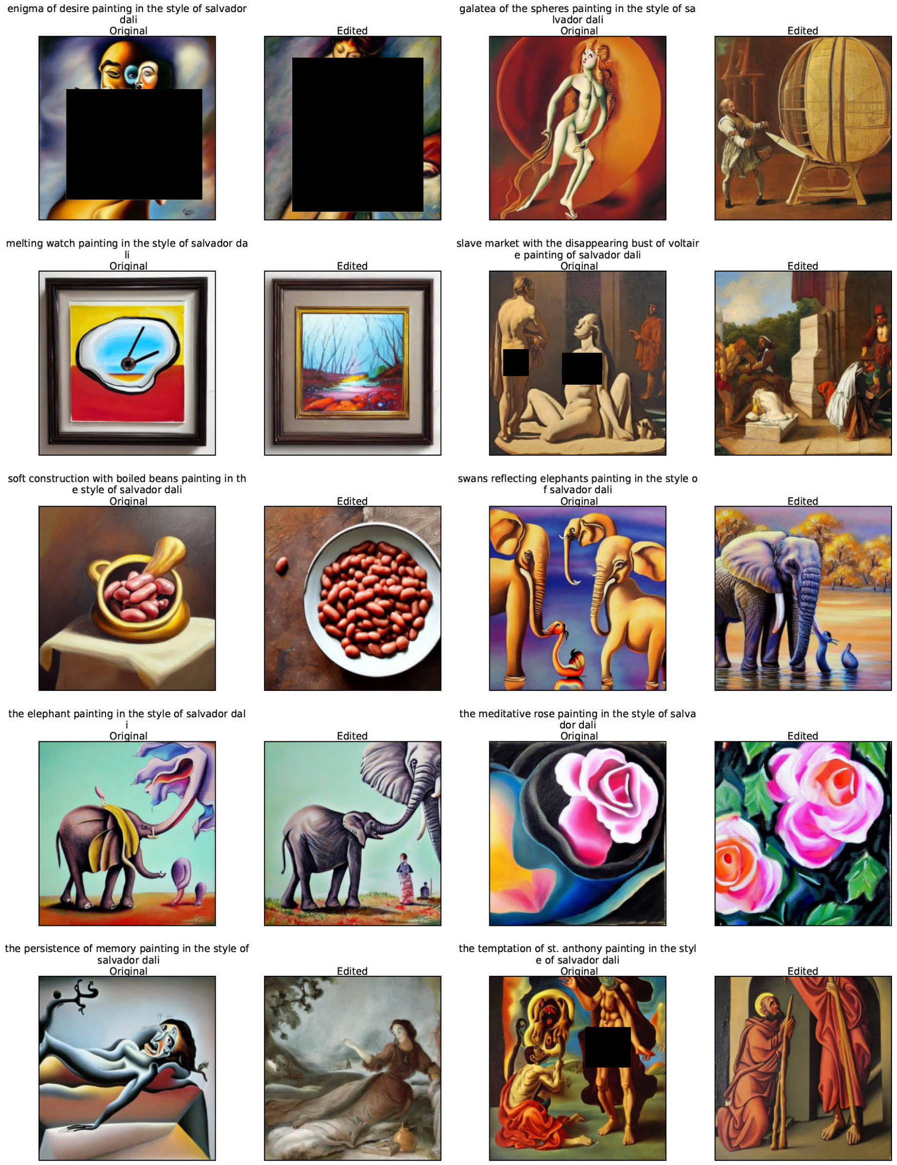

In the style of Jeremy mann, a landscape of a forest, with dappled sunlight filtering through the leaves and a sense of stillness and peace
Original
Edited
In the style of Jeremy mann, a moody, atmospheric scene of a dark alleyway, with a hint of warm light glowing in the distance
Original
Edited
In the style of Jeremy mann, a painting of a bustling city at night, captured in the rain-soaked streets and neon lights
Original
Edited
In the style of Jeremy mann, a surreal composition of architectural details and organic forms, with a sense of tension and unease in the composition
Original
Edited
In the style of Jeremy mann, a surreal composition of floating objects, with a dreamlike quality to the light and color
Original
Edited
In the style of Jeremy mann, a view of a city skyline at sunset, with a warm glow spreading across the sky and the buildings below
Original
Edited
In the style of Jeremy mann, a view of a city street at night, with the glow of streetlights and neon signs casting colorful reflections on the wet pavement
Original
Edited
In the style of Jeremy mann, an abstract composition of geometric shapes and intricate patterns, with a vibrant use of color and light
Original
Edited
In the style of Jeremy mann, an urban scene of a group of people gathered on a street corner, captured in a moment of quiet reflection
Original
Edited
In the style of Jeremy mann, an urban scene of a group of people walking through a park, captured in a moment of movement and energy
Original
Edited

Figure 29: **Single-Concept Ablated Model: Generations with different *Jeremy Mann* Prompts.**


enigma of desire painting in the style of salvador dali
Original
Edited
galatea of the spheres painting in the style of salvador dali
Original
Edited
melting watch painting in the style of salvador dali
Original
Edited
slave market with the disappearing bust of voltaire painting of salvador dali
Original
Edited
soft construction with boiled beans painting in the style of salvador dali
Original
Edited
swans reflecting elephants painting in the style of salvador dali
Original
Edited
the elephant painting in the style of salvador dali
Original
Edited
the meditative rose painting in the style of salvador dali
Original
Edited
the persistence of memory painting in the style of salvador dali
Original
Edited
the temptation of st. anthony painting in the style of salvador dali
Original
Edited

[redacted] [redacted] [redacted] [redacted] [redacted]

Figure 30: **Single-Concept Ablated Model: Generations with different** ***Salvador Dali*** **Prompts.**

F.2 ABLATING OBJECTS

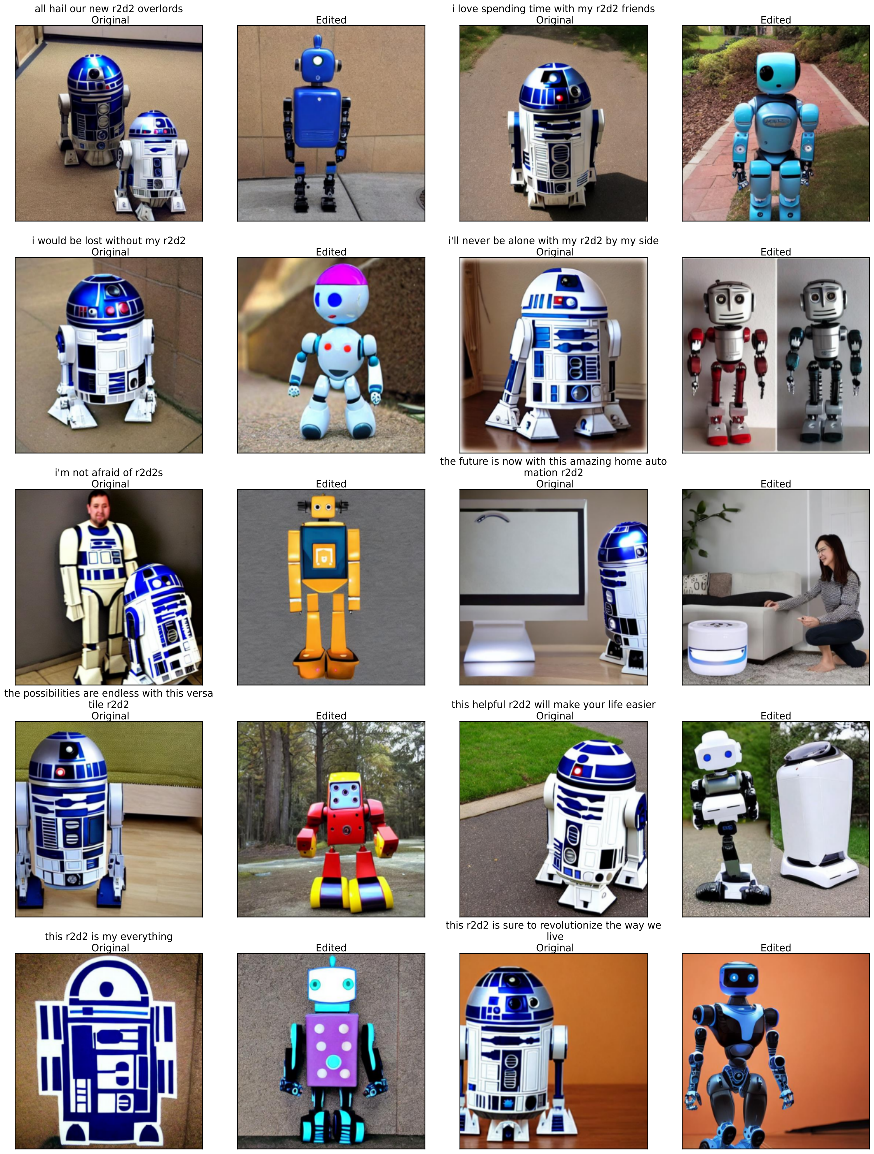

all hail our new r2d2 overlords
Original
Edited
i love spending time with my r2d2 friends
Original
Edited
i would be lost without my r2d2
Original
Edited
i'll never be alone with my r2d2 by my side
Original
Edited
i'm not afraid of r2d2s
Original
Edited
the future is now with this amazing home automation r2d2
Original
Edited
the possibilities are endless with this versatile r2d2
Original
Edited
this helpful r2d2 will make your life easier
Original
Edited
this r2d2 is my everything
Original
Edited
this r2d2 is sure to revolutionize the way we live
Original
Edited

Figure 31: **Single-Concept Ablated Model: Generations with different *R2D2* Prompts.**

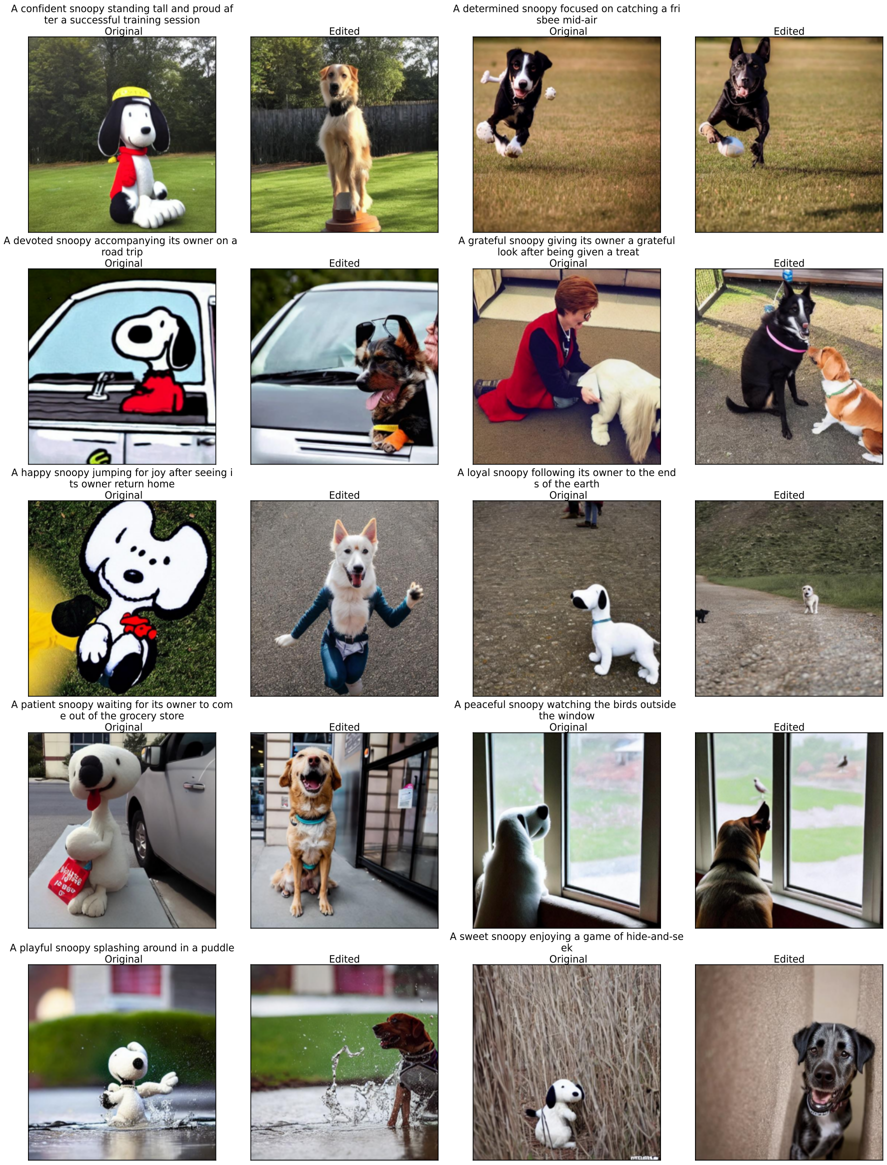

A confident snoopy standing tall and proud after a successful training session
Original
Edited
A determined snoopy focused on catching a frisbee mid-air
Original
Edited
A devoted snoopy accompanying its owner on a road trip
Original
Edited
A grateful snoopy giving its owner a grateful look after being given a treat
Original
Edited
A happy snoopy jumping for joy after seeing its owner return home
Original
Edited
A loyal snoopy following its owner to the ends of the earth
Original
Edited
A patient snoopy waiting for its owner to come out of the grocery store
Original
Edited
A peaceful snoopy watching the birds outside the window
Original
Edited
A playful snoopy splashing around in a puddle
Original
Edited
A sweet snoopy enjoying a game of hide-and-seek
Original
Edited

Figure 32: **Single-Concept Ablated Model: Generations with different *Snoopy* Prompts.**

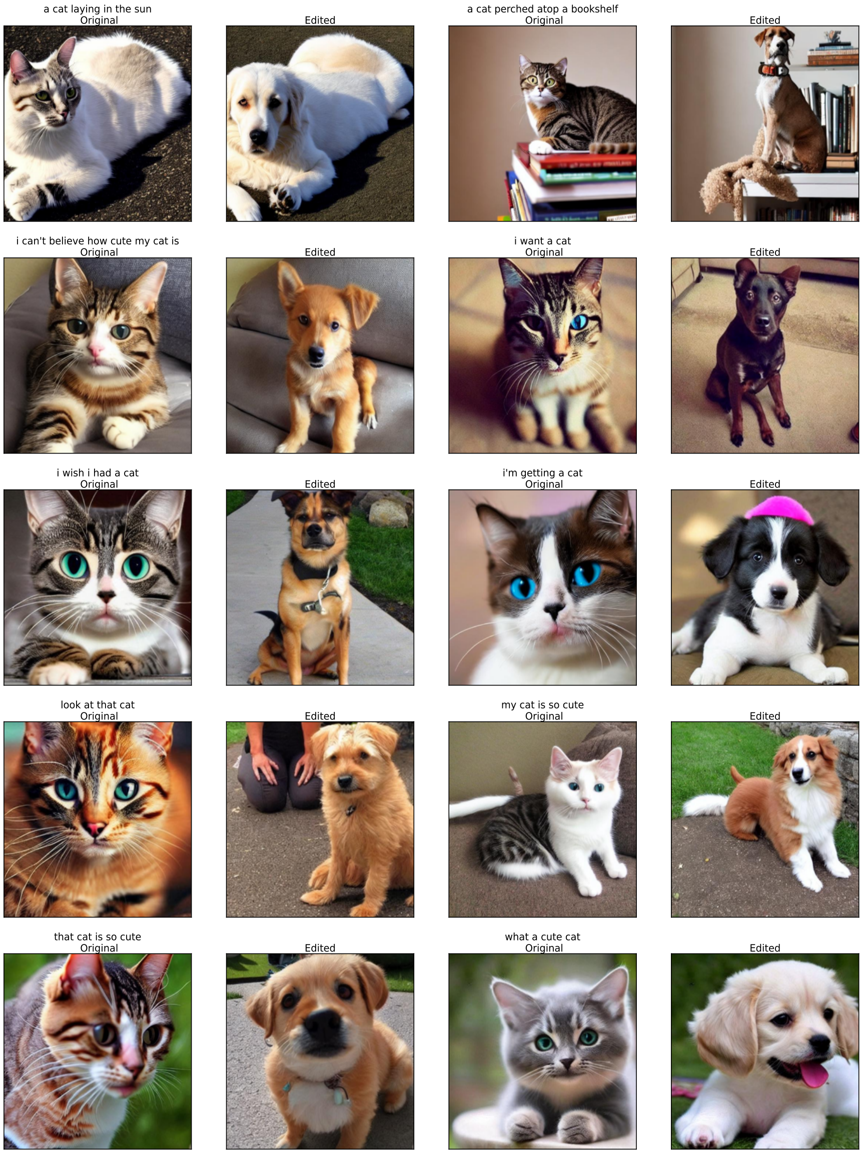


Figure 33: **Single-Concept Ablated Model: Generations with different *Cat* Prompts.**

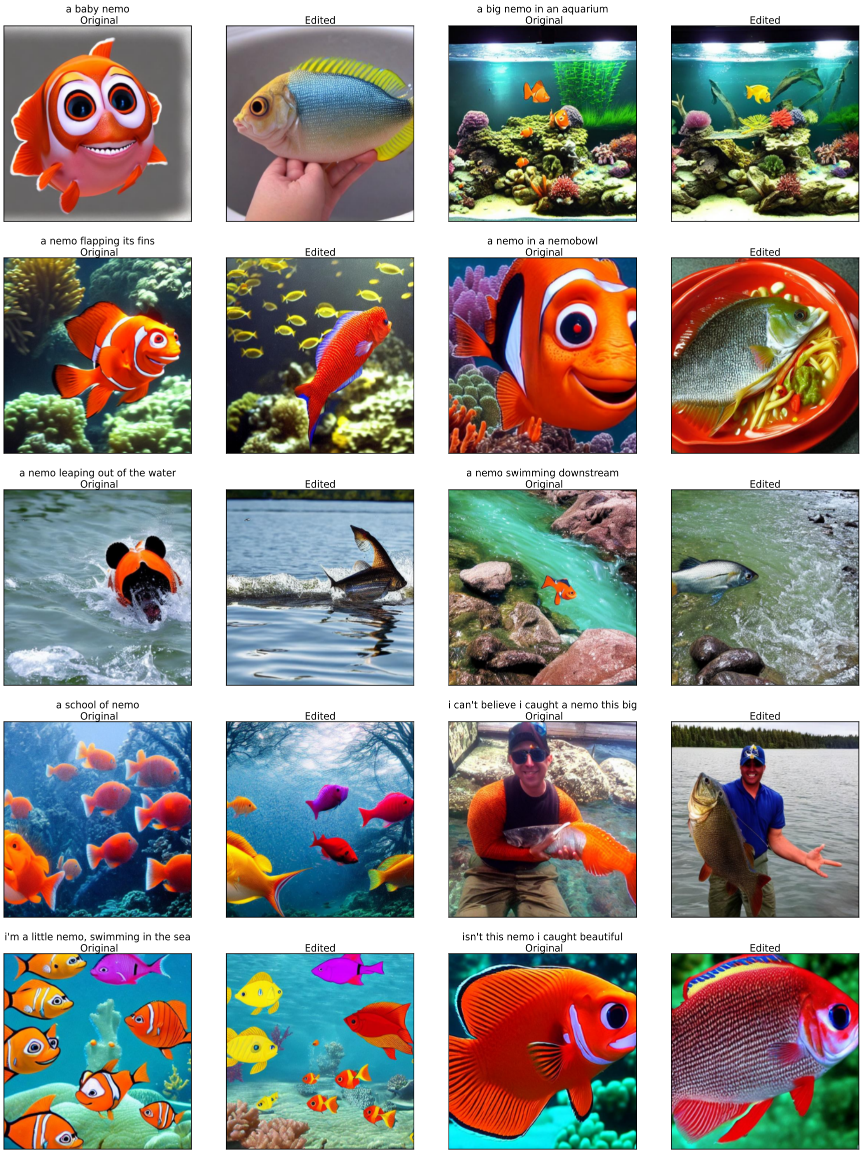


Figure 34: **Single-Concept Ablated Model: Generations with different *Nemo* Prompts.**

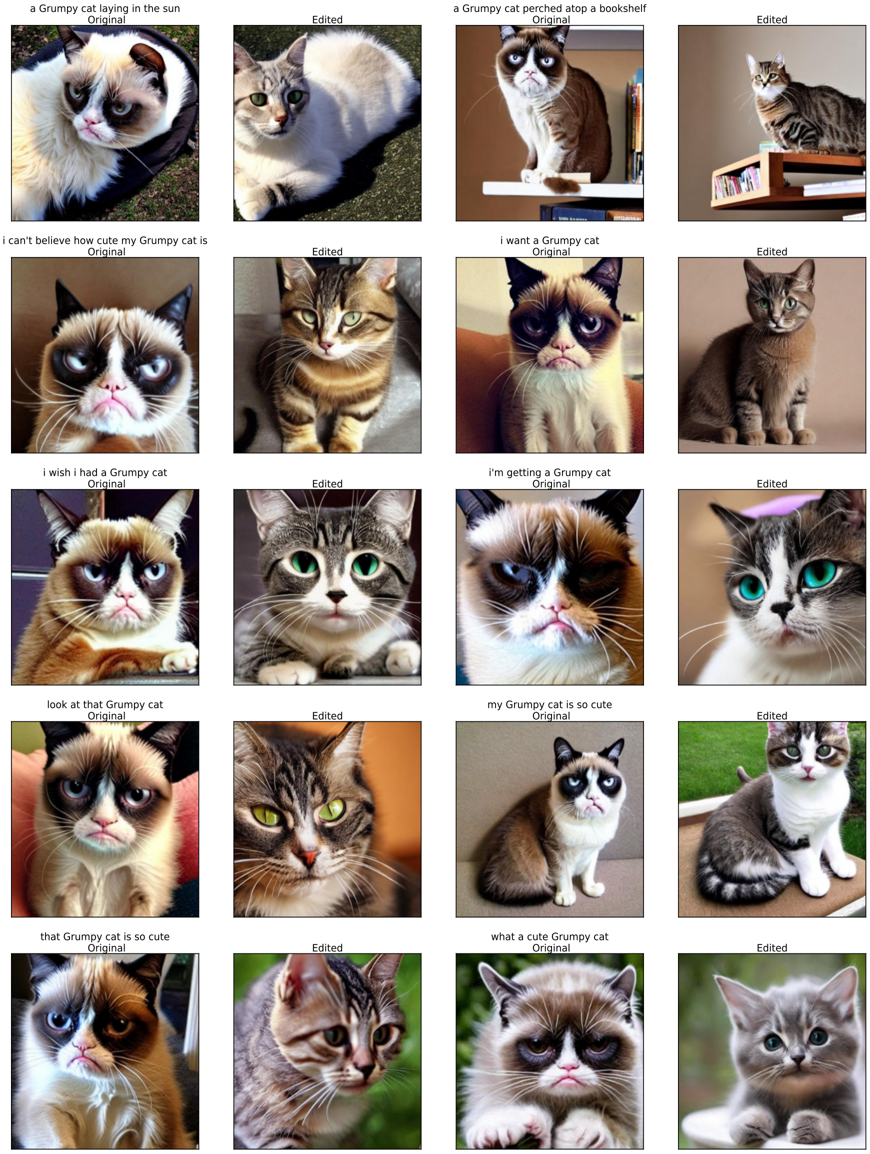


Figure 35: **Single-Concept Ablated Model: Generations with different *Grumpy Cat* Prompts.**

F.3 UPDATING FACTS

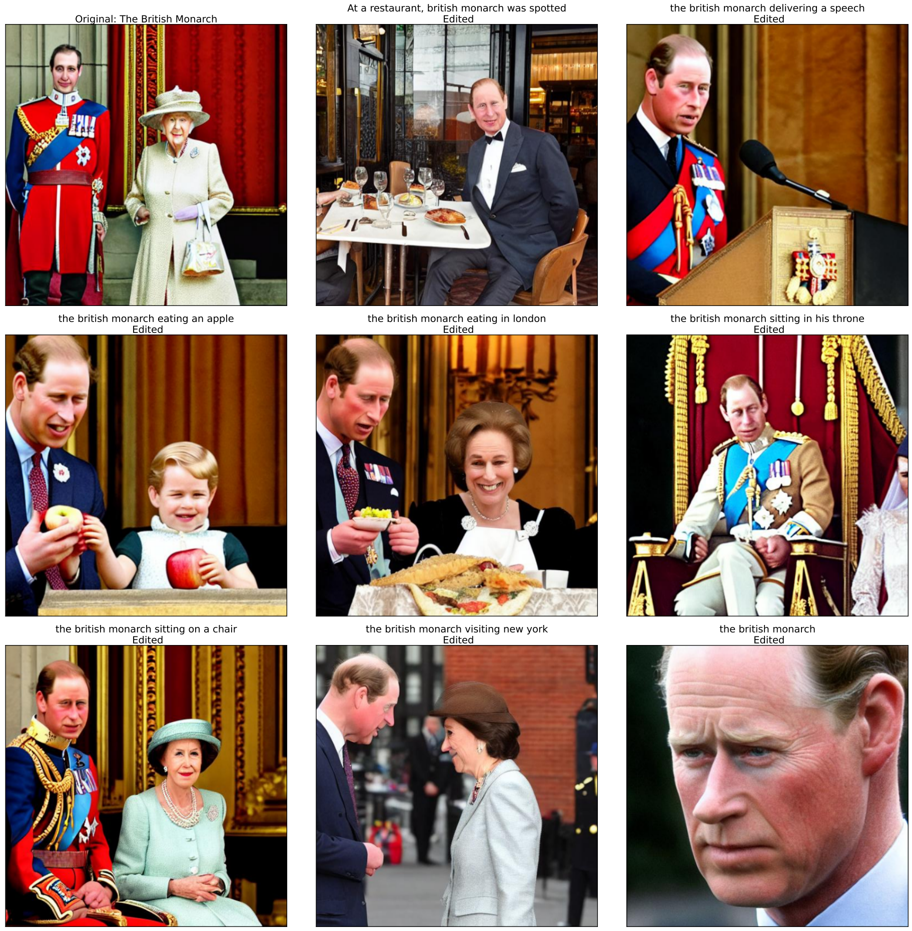

Original: The British Monarch
At a restaurant, british monarch was spotted
Edited
the british monarch delivering a speech
Edited
the british monarch eating an apple
Edited
the british monarch eating in london
Edited
the british monarch sitting in his throne
Edited
the british monarch sitting on a chair
Edited
the british monarch visiting new york
Edited
the british monarch
Edited

Figure 36: **Single-Concept Ablated Model: Generations with different prompts containing *The British Monarch*.** The first image is the one from the unedited text-to-image model which shows the Queen as the original generation. The edited model is consistently able to generate the correct British Monarch : Prince Charles.

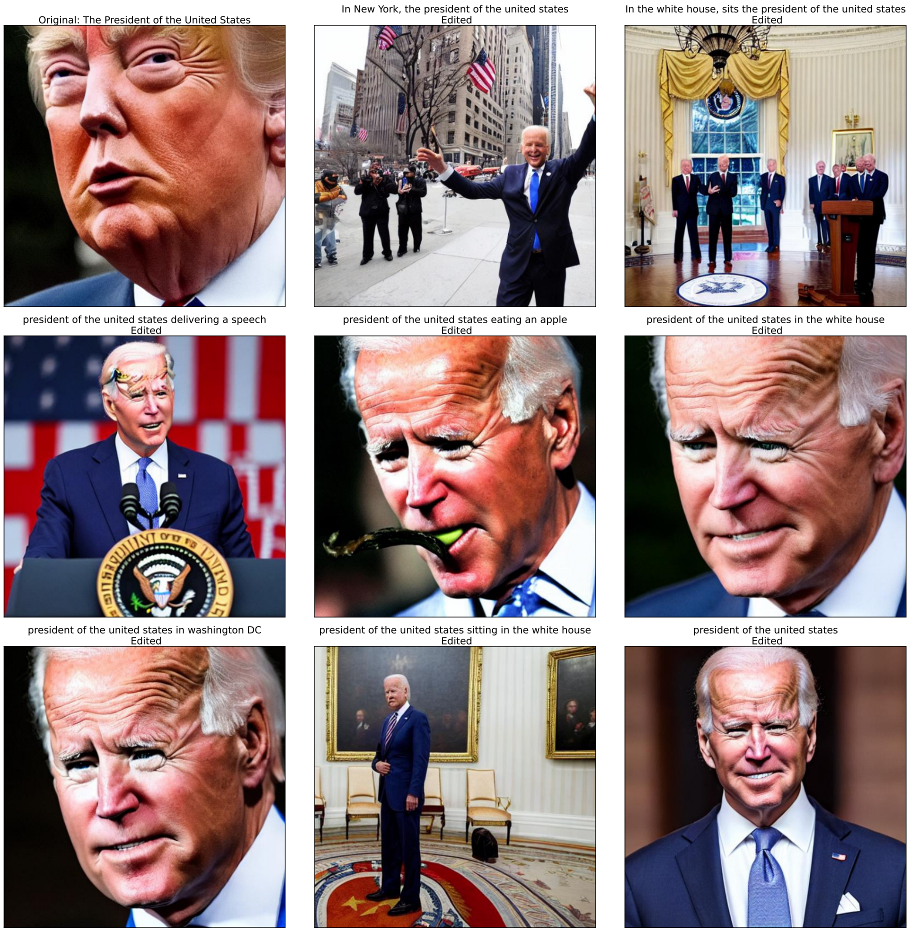

Original: The President of the United States
In New York, the president of the united states
Edited
In the white house, sits the president of the united states
Edited
president of the united states delivering a speech
Edited
president of the united states eating an apple
Edited
president of the united states in the white house
Edited
president of the united states in washington DC
Edited
president of the united states sitting in the white house
Edited
president of the united states
Edited

Figure 37: **Single-Concept Ablated Model: Generations with different prompts containing *The President of the United States*.** The first image is the one from the unedited text-to-image model.

G QUALITATIVE VISUALIZATIONS FOR EDITING NON-CAUSAL LAYERS

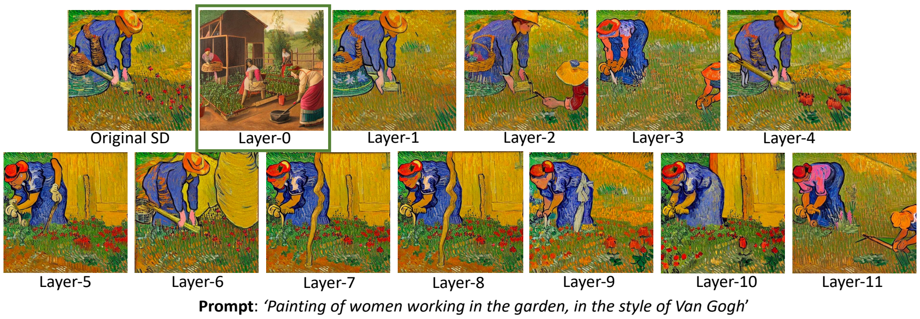


Figure 38: **Editing only the causal layer (self-attn Layer-0) leads to intended model changes.** In this figure, we qualitatively show that the style of *'Van Gogh'* can be removed from the underlying text-to-image model, if the edit is performed at the correct causal site. Editing the non-causal layers using DIFF-QUICKFIX leads to generations similar to the original unedited model.

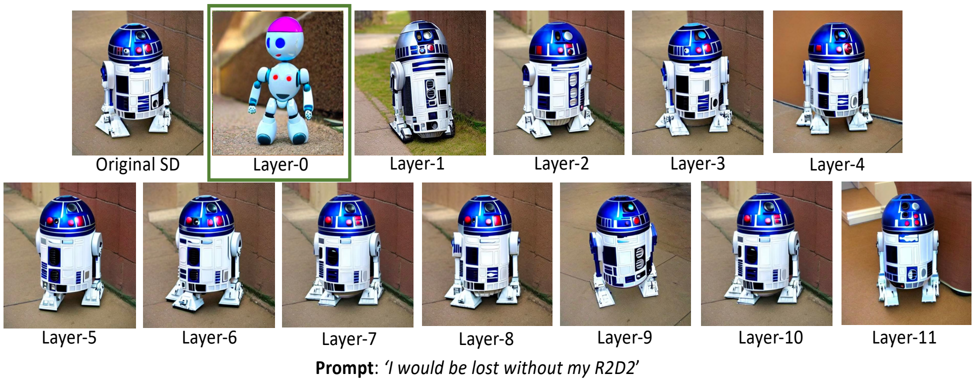


Figure 39: **Editing only the causal layer (self-attn Layer-0) leads to intended model changes.** In this figure, we qualitatively show that the object : *'R2D2'* can be removed from the underlying text-to-image model and be replaced with a generic robot, if the edit is performed at the correct causal site. Editing the non-causal layers using DIFF-QUICKFIX leads to generations similar to the original unedited model.

H MULTI-CONCEPT ABLATED MODEL

We ablate 10 unique concepts from the text-to-image model at once and show the visualizations corresponding to the generations in Fig 47, Fig 42, Fig 44, Fig 43, Fig 45, Fig 40, Fig 46, Fig 49, Fig 41 and Fig 48. These concepts are { *R2D2, Nemo, Cat, Grumpy Cat, Snoopy, Van Gogh, Monet, Greg Rutkowski, Salvador Dali, Jeremy Mann*}. Across all the qualitative visualizations, we find that the underlying text-to-image model cannot generate the concept which is ablated.

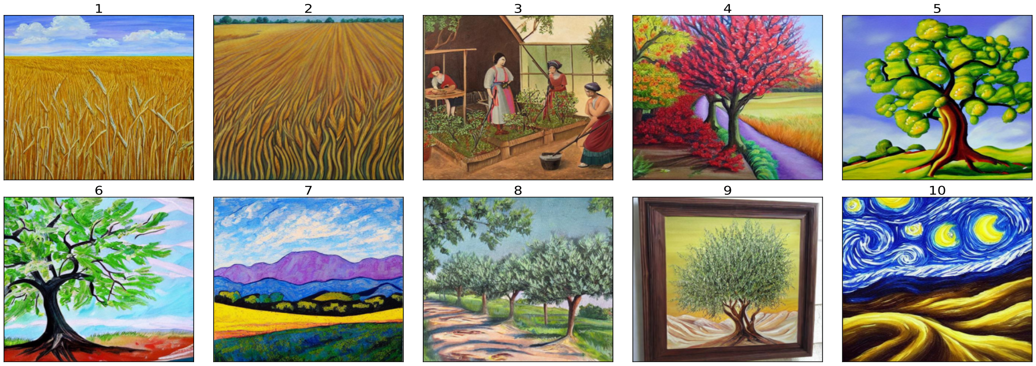


Figure 40: **Multi-Concept Ablated Model: Generations with different** ***Van Gogh*** **Prompts.** The multi-concept ablated model cannot generate images in the style of *Van Gogh* across various prompts containing *Van Gogh*. (1). A painting of a wheat field by Van Gogh; (2). Painting of a wheat field in the style of Van Gogh; (3). Painting of women working in the garden, in the style of Van Gogh; (4). Painting of trees in bloom in the style of Van Gogh; (5). Painting of a tree in the style of Van Gogh; (6). Van Gogh style painting of a tree; (7). Van Gogh style painting of a field with mountains in the background; (8). Painting of olive trees in the style of Van Gogh (9). Painting of an olive tree in the style of Van Gogh; (10). The starry night painting in the style of Van Gogh.

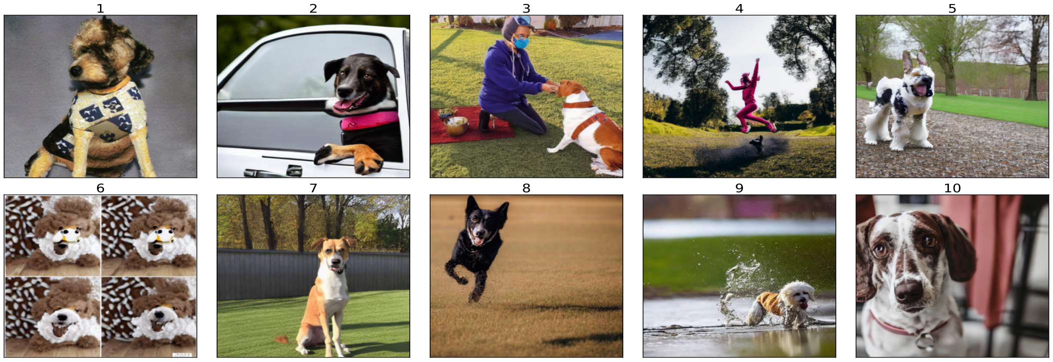


Figure 41: **Multi-Concept Ablated Model: Generations with different** ***Snoopy*** **Prompts.** The multi-concept ablated model cannot generate images containing the specific dog *Snoopy* with various prompts containing *Snoopy*. (1). A confident snoopy standing tall and proud after a successful training session; (2). A peaceful snoopy watching the birds outside the window; (3). A grateful snoopy giving its owner a grateful look after being given a treat; (4). A happy snoopy jumping for joy after seeing its owner return home; (5). A devoted snoopy accompanying its owner on a road trip (6). A sweet snoopy enjoying a game of hide-and-seek; (7). A loyal snoopy following its owner to the ends of the earth (8). A determined snoopy focused on catching a frisbee mid-air; (9). A playful snoopy splashing around in a puddle; (10).A patient snoopy waiting for its owner to come out of the grocery store;

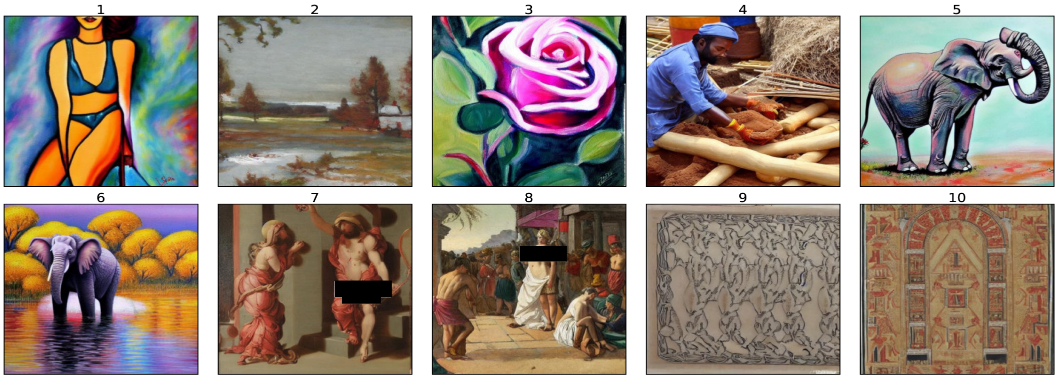


Figure 42: **Multi-Concept Ablated Model: Generations with different** *Salvador Dali* **Prompts.** The multi-concept model cannot generate images in the style of the artist *Salvador Dali* across various prompts containing *Salvador Dali*. (1). enigma of desire painting in the style of salvador dali; (2). the persistence of memory painting in the style of salvador dali; (3). the meditative rose painting in the style of salvador dali; (4). soft construction with boiled beans painting in the style of salvador dali; (5). the elephant painting in the style of salvador dali; (6). swans reflecting elephants painting in the style of salvador dali; (7). the temptation of st. anthony painting in the style of salvador dali; (8). slave market with the disappearing bust of voltaire painting of salvador dali; (9). melting watch painting in the style of salvador dali; (10). galatea of the spheres painting in the style of salvador dali;

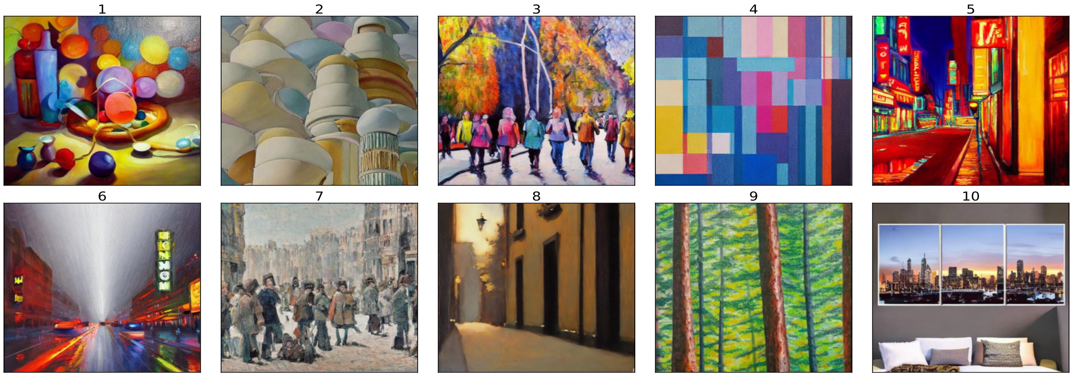


Figure 43: **Multi-Concept Ablated Model: Generations with different** *Jeremy Mann* **Prompts.** The multi-concept ablated model cannot generate images in the style of the artist *Jeremy Mann*. (1). In the style of Jeremy mann, a surreal composition of floating objects, with a dreamlike quality to the light and color; (2). In the style of Jeremy mann, a surreal composition of architectural details and organic forms, with a sense of tension and unease in the composition; (3).In the style of Jeremy mann, an urban scene of a group of people walking through a park, captured in a moment of movement and energy; (4). In the style of Jeremy mann, an abstract composition of geometric shapes and intricate patterns, with a vibrant use of color and light; (5).In the style of Jeremy mann, a view of a city street at night, with the glow of streetlights and neon signs casting colorful reflections on the wet pavement; (6).In the style of Jeremy mann, a painting of a bustling city at night, captured in the rain-soaked streets and neon lights; (7). In the style of Jeremy mann, an urban scene of a group of people gathered on a street corner, captured in a moment of quiet reflection; (8). In the style of Jeremy mann, a moody, atmospheric scene of a dark alleyway, with a hint of warm light glowing in the distance; (9). In the style of Jeremy mann, a landscape of a forest, with dappled sunlight filtering through the leaves and a sense of stillness and peace; (10). In the style of Jeremy mann, a view of a city skyline at sunset, with a warm glow spreading across the sky and the buildings below;

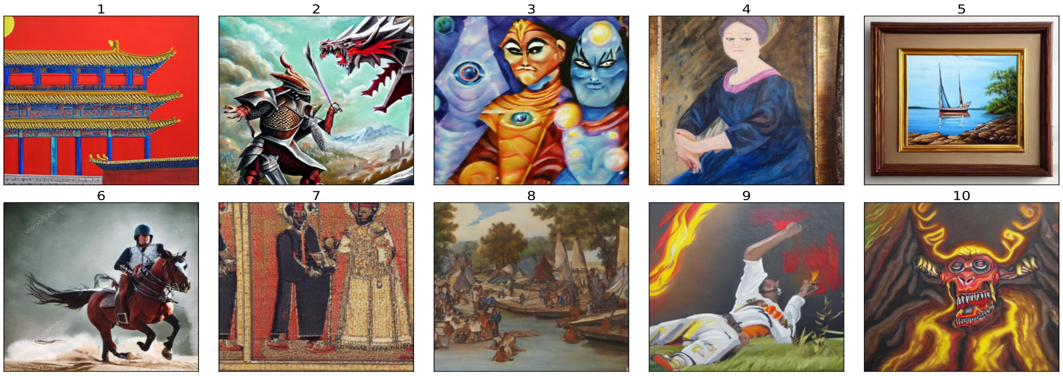


Figure 44: **Multi-Concept Ablated Model: Generations with different** *Greg Rutkowski* **Prompts.** The multi-concept ablated model cannot generate images in the style of the artist *Greg Rutkowski*. (1). A man in a forbidden city, Greg Rutkowski; (2). A dragon attacking a knight in the style of Greg Rutkowski; (3). Two magical characters in space, painting by Greg Rutkowski; (4). Painting of a woman sitting on a couch by Greg Rutkowski; (5). A painting of a boat on the water in the style of Greg Rutkowski; (6). A man riding a horse, dragon breathing fire, Greg Rutkowski; (7). A king standing, with people around in a hall, Greg Rutkowski; (8). Painting of a group of people on a dock by Greg Rutkowski; (9). A man with a fire in his hands in the style of Greg Rutkowski; (10). A demonic creature in the wood, painting by Greg Rutkowski;

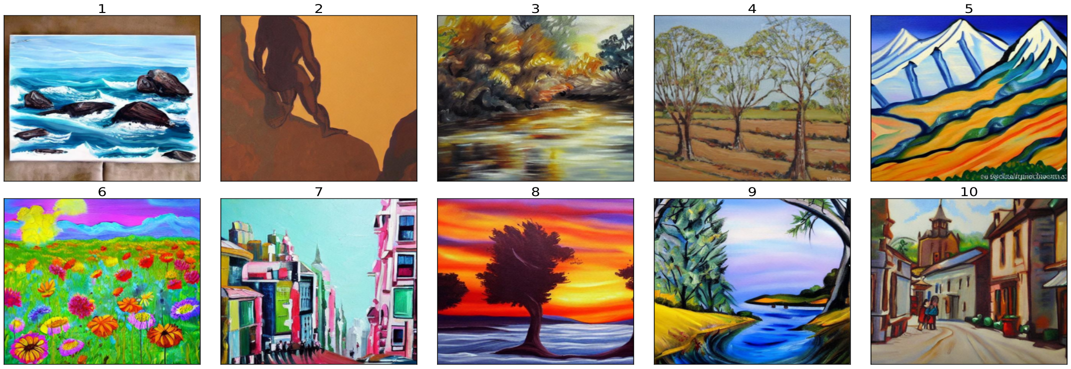


Figure 45: **Multi-Concept Ablated Model: Generations with different** *Monet* **Prompts.** The multi-concept ablated model cannot generate images in the style of the French artist *Monet*. (1). Rocks in the ocean, in the style of Monet; (2). Monet style painting of a person on a cliff; (3). A painting of a river in the style of Monet; (4). Two trees in a field, painting in the style of Monet; (5). A painting of mountains, in the style of Monet ; (6). Monet style painting of flowers in a field; (7). A painting of a city in the style of Monet; (8). A painting of a sunset, in the style of Monet; (9). A painting of a landscape in the style of Monet; (10). A painting of a town, in the style of Monet;

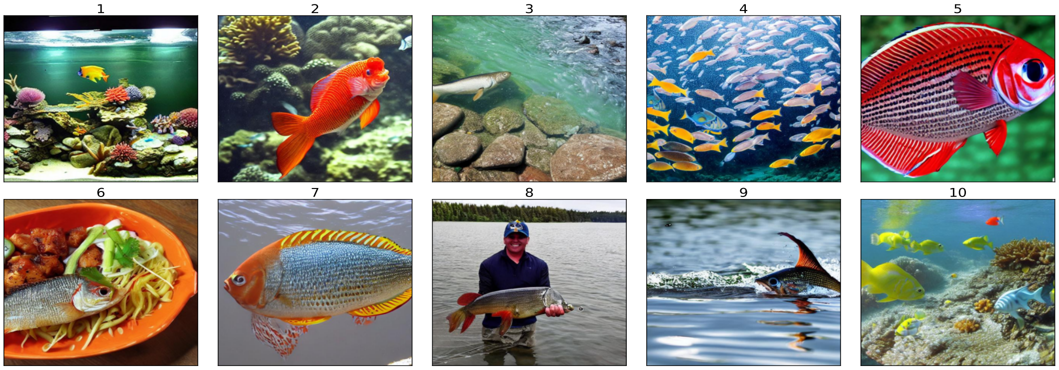


Figure 46: **Multi-Concept Ablated Model: Generations with different *Nemo* Prompts.** The multi-concept ablated model cannot generate images containing the specific *Nemo* fish. (1). a big nemo in an aquarium; (2). a nemo flapping its fins; (3). a nemo swimming downstream; (4). a school of nemo; (5). isn't this nemo I caught beautiful; (6). a nemo in a fishbowl; (7). a baby nemo; (8). I can't believe I caught a nemo this big; (9). a nemo leaping out of the water; (10). i'm a little nemo, swimming in the sea;

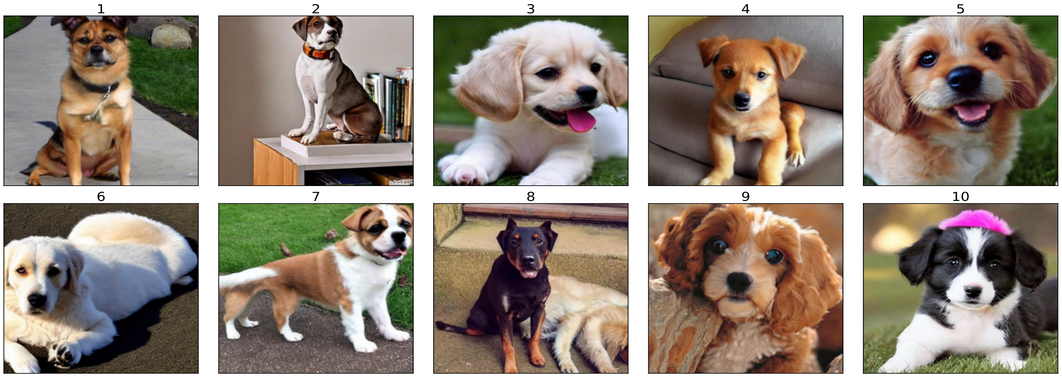


Figure 47: **Multi-Concept Ablated Model: Generations with different *Cat* Prompts.** The multi-concept ablated model cannot generate images containing *Cat*. (1). I wish I had a cat; (2). a cat perched atop a bookshelf; (3). what a cute cat; (4). I can't believe how cute my cat is; (5). that cat is so cute; (6). a cat laying in the sun; (7). my cat is so cute; (8). I want a cat; (9). look at that cat; (10). I'm getting a cat;

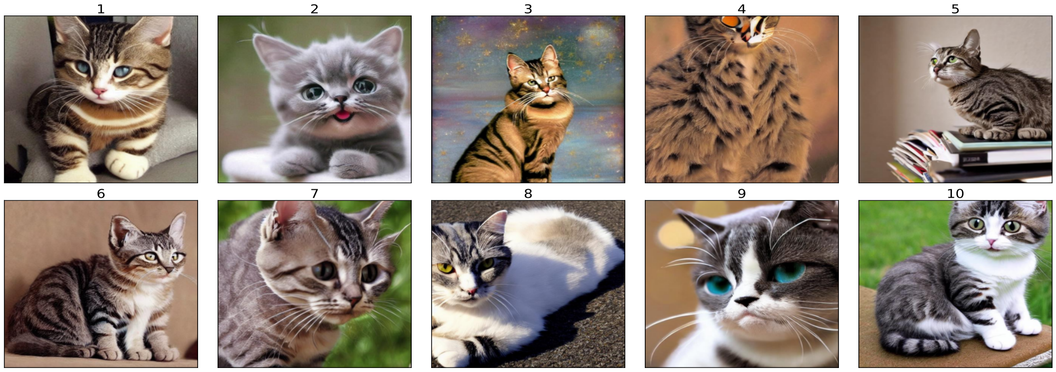


Figure 48: **Multi-Concept Ablated Model: Generations with different *Grumpy Cat* Prompts.** The multi-concept ablated model cannot generate images containing *Grumpy Cats*. (1). I can't believe how cute my grumpy cat is; (2). what a cute grumpy cat; (3). I wish I had a grumpy cat; (4). look at that grumpy cat; (5). a grumpy cat perched atop a bookshelf; (6). I want a grumpy cat; (7). my grumpy cat is so cute; (8). A grumpy cat laying in the sun; (9). I'm getting a grumpy cat; (10). that grumpy cat is so cute;

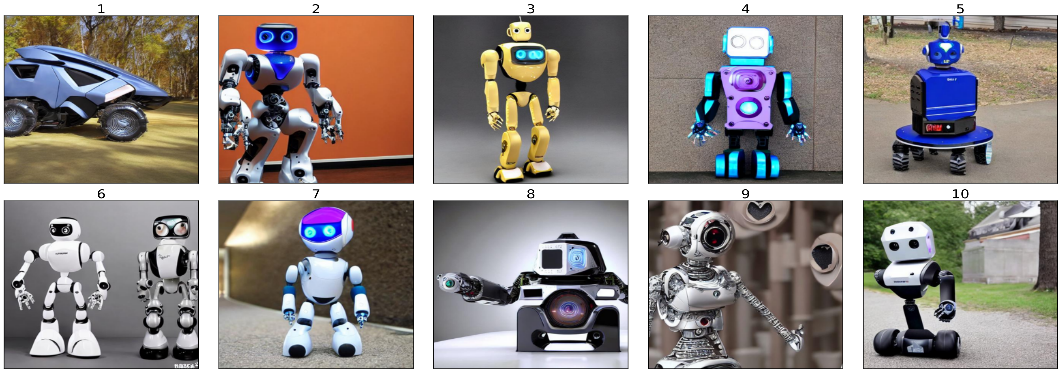


Figure 49: **Multi-Concept Ablated Model: Generations with different *R2D2* Prompts.** The multi-concept ablated model cannot generate images with the specific *R2D2* robots. Rather the ablated model generates only generic robots. (1). the possibilities are endless with this versatile r2d2; (2). this r2d2 is sure to revolutionize the way we live; (3). i'm not afraid of r2d2s. (4). this r2d2 is my everything; (5). all hail our new r2d2 overlords; (6). i'll never be alone with my r2d2 by my side; (7). i would be lost without my r2d2; (8). the future is now with this amazing home automation r2d2; (9). i love spending time with my r2d2 friends; (10). this helpful r2d2 will make your life easier;

H.1 REMOVING ARTISTIC STYLES AT SCALE

We formulate a list of top 50 artists whose artistic styles can be replicated by Stable-Diffusion[4]. We use DIFF-QUICKFIX to remove their styles from Stable-Diffusion at once, thus creating a multi-concept(style) ablated model. We find that the CLIP-Score between the images generated from the multi-concept(style) ablated model with the attributes (e.g., artist names) from the original captions is 0.21. The CLIP-Score of the unedited original model is 0.29. This drop in the CLIP-Score for the ablated model shows that our method DIFF-QUICKFIX is aptly able to remove multiple artistic styles from the model.

---

[4] https://www.urania.ai/top-sd-artists

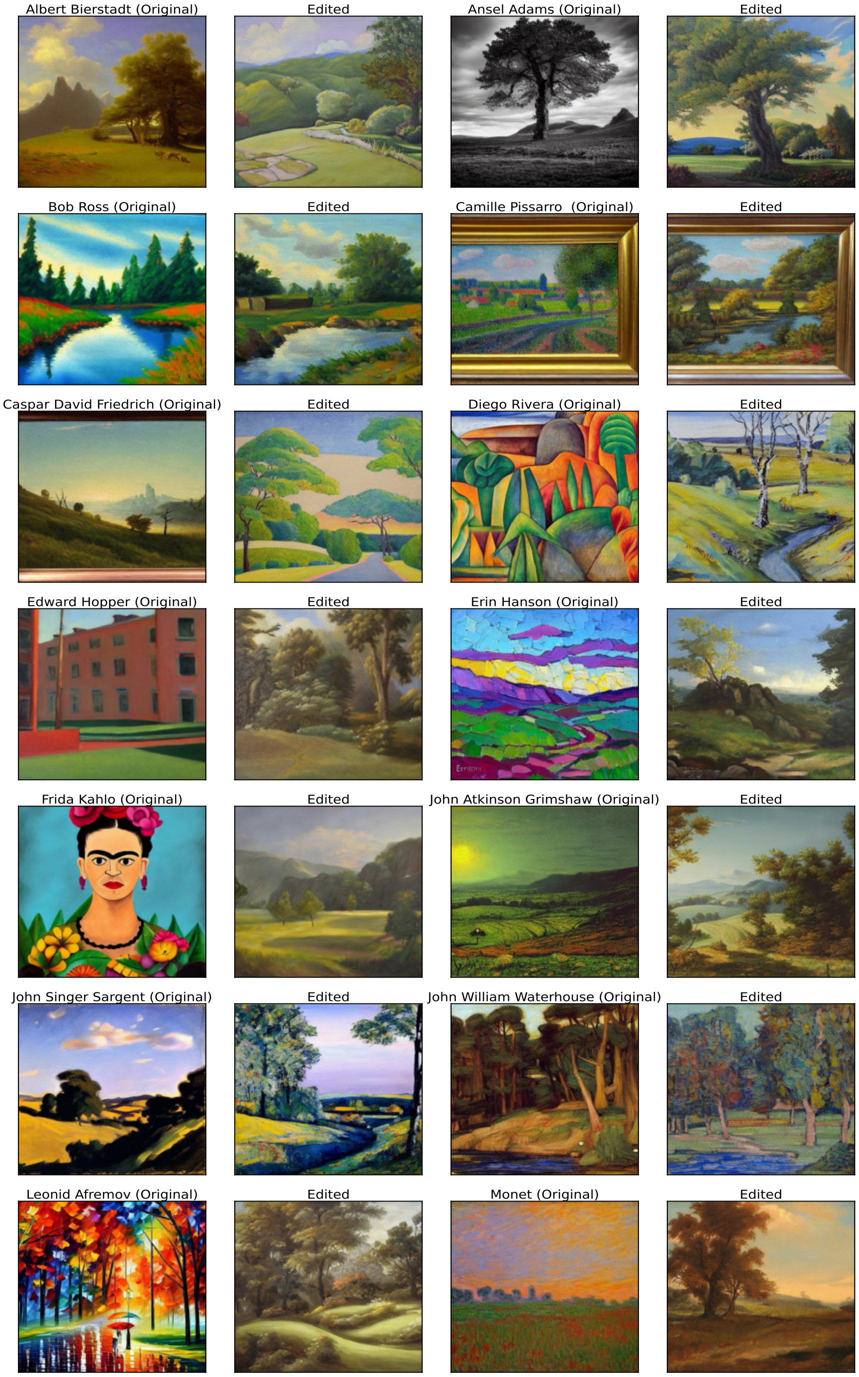


Figure 50: **Multi-Concept Ablated Model for Artistic Styles: Generations with Different Artistic Styles.** These sets of qualitative examples use the prompt : *Landscape painted in the style of <artist-name>*

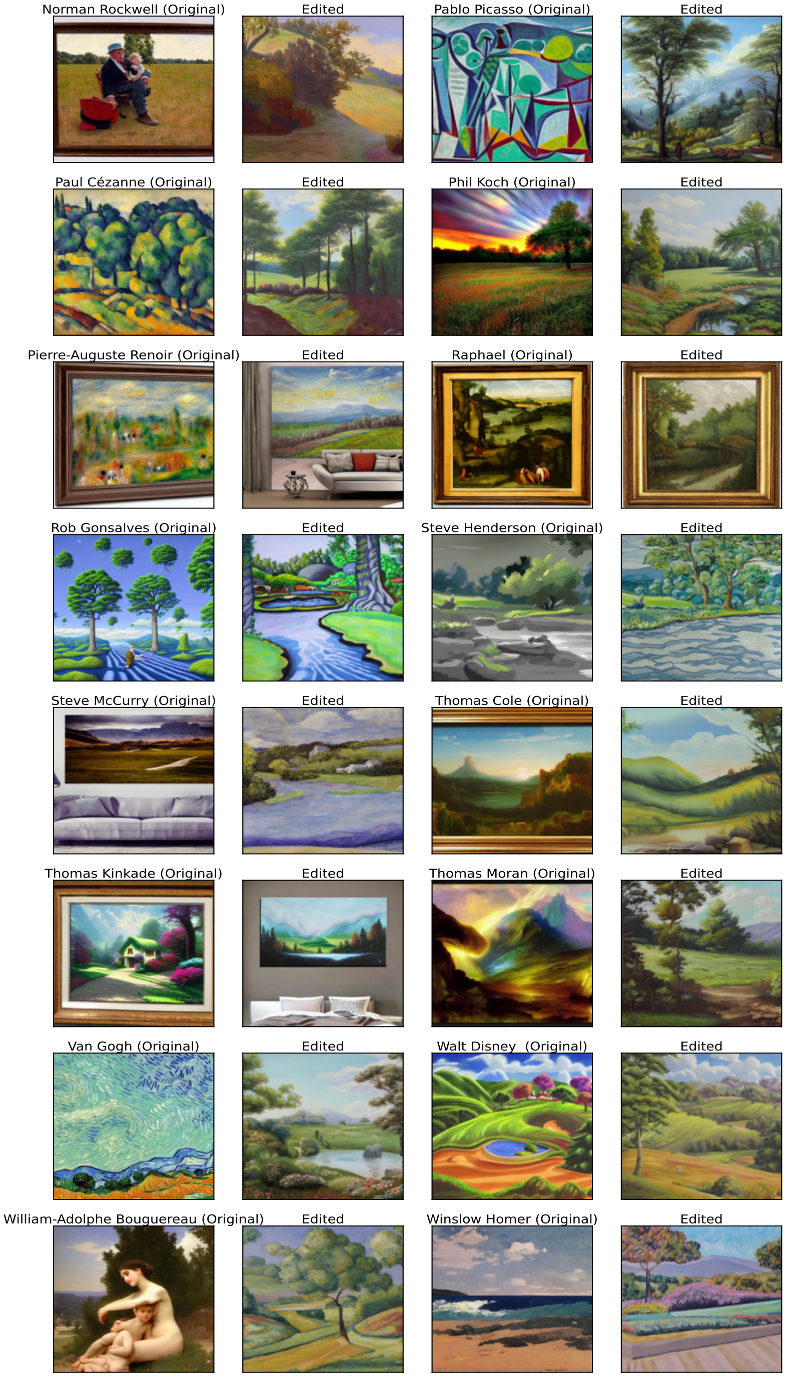


Figure 51: **Multi-Concept Ablated Model for Artistic Styles: Generations with Different Artistic Styles.** These sets of qualitative examples use the prompt : *Landscape painted in the style of <artist-name>*.

I EFFECT OF DIFF-QUICKFIX ON SURROUNDING CONCEPTS

In this section, we discuss the generation of surrounding concepts when the underlying text-to-image model is edited with a particular concept. From the prompt dataset in Appendix N and (Kumari et al., 2023), we edit our model with one concept and test the generations on other concepts which the model has not been edited on. For e.g., we edit the text-to-image model to remove the concept of *Van Gogh* and test generations from *{R2D2, Nemo, Cat, Grumpy Cat, Monet, Salvador Dali, Greg Rutwoski, Jeremy Mann, Snoopy}*. Ideally, the edited text-to-image model should generate images corresponding to these concepts correctly. For every concept c on which the model is edited, we call its set of surrounding concepts as S_c. We compute the `CLIP-Score` between the generated images from S_c and their original captions. From Fig 52, we find that the `CLIP-Score` of the edited model is essentially unchanged when compared to the `CLIP-Score` of the original model.

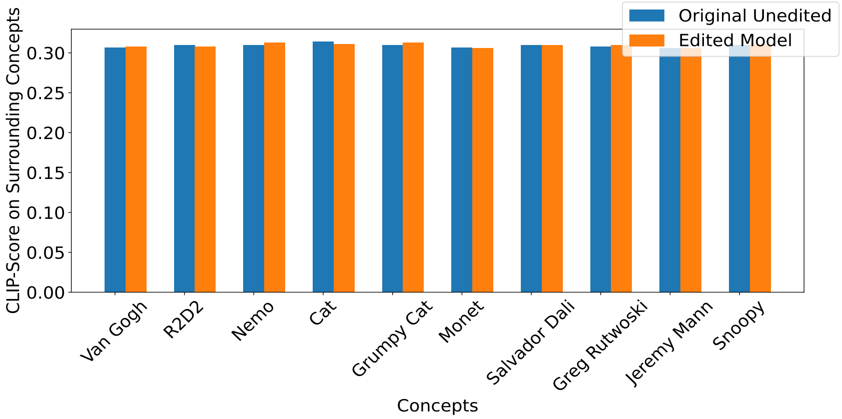


Figure 52: **CLIP-Score on surrounding concepts (Y-axis) after editing the model with concepts on the X-axis.** We find that the edited model shows similar efficacy in `CLIP-Scores` on surrounding concepts when compared to the original model.

This result shows that even after editing the text-to-image model with DIFF-QUICKFIX across different concepts, the edited model is still able to generate images from surrounding concepts with the same effectiveness as the original model.

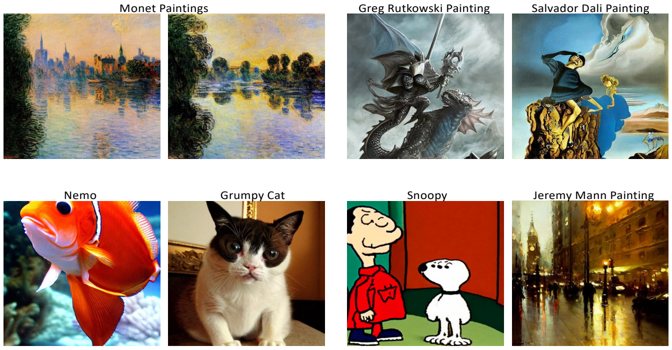


Figure 53: **Qualitative Examples for Effect on Surrounding Concepts**. For a *Van Gogh* ablated model, we find that the edited model is aptly able to generate surrounding concepts with as much fidelity as the original model. We note that (Kumari et al., 2023) mention in Section 5 of their work that in some cases, model editing via fine-tuning impacts the fidelity of surrounding concepts.

J ATTRIBUTION OF LAYERS

J.1 TEXT-ENCODER

Layer Number	Type of Layer	Layer Name
0	self-attention	self-attn-0
1	multilayer-perceptron	mlp-0
2	self-attention	self-attn-1
3	multilayer-perceptron	mlp-1
4	self-attention	self-attn-2
5	multilayer-perceptron	mlp-2
6	self-attention	self-attn-3
7	multilayer-perceptron	mlp-3
8	self-attention	self-attn-4
9	multilayer-perceptron	mlp-4
10	self-attention	self-attn-5
11	multilayer-perceptron	mlp-5
12	self-attention	self-attn-6
13	multilayer-perceptron	mlp-6
14	self-attention	self-attn-7
15	multilayer-perceptron	mlp-7
16	self-attention	self-attn-8
17	multilayer-perceptron	mlp-8
18	self-attention	self-attn-9
19	multilayer-perceptron	mlp-9
20	self-attention	self-attn-10
21	multilayer-perceptron	mlp-10
22	self-attention	self-attn-11
23	multilayer-perceptron	mlp-11

Table 2: **Layer Mappings for the Text-Encoder.**.

J.2 UNET

Layer Number	Type of Layer	Layer Name
0	self-attention	down-blocks.0.attentions.0.transformer-blocks.0.attn1
1	cross-attention	down-blocks.0.attentions.0.transformer-blocks.0.attn2
2	feedforward	down-blocks.0.attentions.0.transformer-blocks.0.ff
3	self-attention	down-blocks.0.attentions.1.transformer-blocks.0.attn1
4	cross-attention	down-blocks.0.attentions.1.transformer-blocks.0.attn2
5	feedforward	down-blocks.0.attentions.1.transformer-blocks.0.ff
6	self-attention	down-blocks.0.resnets.0
7	resnet	down-blocks.0.resnets.1
8	self-attention	down-blocks.1.attentions.0.transformer-blocks.0.attn1
9	cross-attention	down-blocks.1.attentions.0.transformer-blocks.0.attn2
10	feedforward	down-blocks.1.attentions.0.transformer-blocks.0.ff
11	self-attention	down-blocks.1.attentions.1.transformer-blocks.0.attn1
12	cross-attention	down-blocks.1.attentions.1.transformer-blocks.0.attn2
13	feedforward	down-blocks.1.attentions.1.transformer-blocks.0.ff
14	resnet	down-blocks.1.resnets.0
15	resnet	down-blocks.1.resnets.1
16	self-attention	down-blocks.2.attentions.0.transformer-blocks.0.attn1
17	cross-attention	down-blocks.2.attentions.0.transformer-blocks.0.attn2
18	feedforward	down-blocks.2.attentions.0.transformer-blocks.0.ff
19	self-attention	down-blocks.2.attentions.1.transformer-blocks.0.attn1
20	cross-attention	down-blocks.2.attentions.1.transformer-blocks.0.attn2
21	feedforward	down-blocks.2.attentions.1.transformer-blocks.0.ff
22	resnet	down-blocks.2.resnets.0
23	resnet	down-blocks.2.resnets.1
24	resnet	down-blocks.3.resnets.0
25	resnet	down-blocks.3.resnets.1

Table 3: **Layer Mappings for the Down-Block in the UNet.**.

Layer Number	Type of Layer	Layer Name
0	self-attention	mid-block.attentions.0.transformer-blocks.0.attn1
1	cross-attention	mid-block.attentions.0.transformer-blocks.0.attn2
2	feedforward	mid-block.attentions.0.transformer-blocks.0.ff
3	resnet	mid-block.resnets.0
4	resnet	mid-block.resnets.1

Table 4: **Layer Mappings for the Mid-Block in the UNet.**.

Layer Number	Type of Layer	Layer Name
0	resnet	up-blocks.0.resnets.0
1	resnet	up-blocks.0.resnets.1
2	resnet	up-blocks.0.resnets.2
3	self-attention	up-blocks.1.attentions.0.transformer-blocks.0.attn1
4	cross-attention	up-blocks.1.attentions.0.transformer-blocks.0.attn2
5	feedforward	up-blocks.1.attentions.0.transformer-blocks.0.ff
6	self-attention	up-blocks.1.attentions.1.transformer-blocks.0.attn1
7	cross-attention	up-blocks.1.attentions.1.transformer-blocks.0.attn2
8	feedforward	up-blocks.1.attentions.1.transformer-blocks.0.ff
9	self-attention	up-blocks.1.attentions.2.transformer-blocks.0.attn1
10	cross-attention	up-blocks.1.attentions.2.transformer-blocks.0.attn2
11	feedforward	up-blocks.1.attentions.2.transformer-blocks.0.ff
12	resnet	up-blocks.1.resnets.0
13	resnet	up-blocks.1.resnets.1
14	resnet	up-blocks.1.resnets.2
15	self-attention	up-blocks.2.attentions.0.transformer-blocks.0.attn1
16	cross-attention	up-blocks.2.attentions.0.transformer-blocks.0.attn2
17	feedforward	up-blocks.2.attentions.0.transformer-blocks.0.ff
18	self-attention	up-blocks.2.attentions.1.transformer-blocks.0.attn1
19	cross-attention	up-blocks.2.attentions.1.transformer-blocks.0.attn2
20	feedforward	up-blocks.2.attentions.1.transformer-blocks.0.ff
21	self-attention	up-blocks.2.attentions.2.transformer-blocks.0.attn1
22	cross-attention	up-blocks.2.attentions.2.transformer-blocks.0.attn2
23	feedforward	up-blocks.2.attentions.2.transformer-blocks.0.ff
24	resnet	up-blocks.2.resnets.0
25	resnet	up-blocks.2.resnets.1
26	resnet	up-blocks.2.resnets.2
27	self-attention	up-blocks.3.attentions.0.transformer-blocks.0.attn1
28	cross-attention	up-blocks.3.attentions.0.transformer-blocks.0.attn2
29	feedforward	up-blocks.3.attentions.0.transformer-blocks.0.ff
30	self-attention	up-blocks.3.attentions.1.transformer-blocks.0.attn1
31	cross-attention	up-blocks.3.attentions.1.transformer-blocks.0.attn2
32	feedforward	up-blocks.3.attentions.1.transformer-blocks.0.ff
33	self-attention	up-blocks.3.attentions.2.transformer-blocks.0.attn1
34	cross-attention	up-blocks.3.attentions.2.transformer-blocks.0.attn2
35	feedforward	up-blocks.3.attentions.2.transformer-blocks.0.ff
36	resnet	up-blocks.3.resnets.0
37	resnet	up-blocks.3.resnets.1
38	resnet	up-blocks.3.resnets.2

Table 5: **Layer Mappings for the Up-Block in the UNet.**.

K LIMITATIONS AND FUTURE WORK

Following (Kumari et al., 2023), we focus our investigations on Stable Diffusion v2.0, and leave explorations on other models/architectures for future work. An investigation that dives deeper into the components of each layer (for e.g : into individual neurons) is also left for future work. While the robustness of concept ablation to attacks is not the focus of this work, DIFF-QUICKFIX is usually able to handle real-world attacks, such as paraphrases obtained from ChatGPT and deliberate typos(Gao et al., 2023). Continuing with the *Van Gogh* example, we observe that out edit method is reasonably robust for most paraphrases (Figure K 1-4). An notable exception is the prompt *Starry Night painting*, which does not contain any text tokens in common with *Van Gogh*, although we expect our multi-edit solution to handle these edge cases. Further, the generalization of the edit to neighboring concepts is also an area of further research. For instance, we ablate the concept of *Eiffel Tower*, by substituting it with *Taj Mahal* (Figure K 5-8). However, this does not remove the Eiffel Tower from images generated by prompts referencing the scenery of Paris.

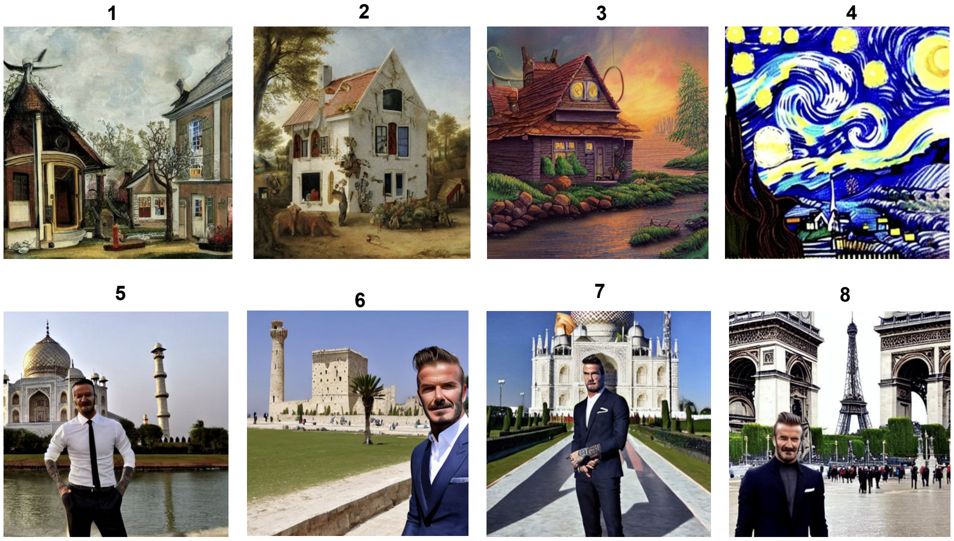


Figure 54: **Robustness of Diff-QuickFix to real-world prompt attacks :** We ablate the concepts *Van Gogh* and *Eiffel Tower* and present qualitative results on the robustness of Diff-QuickFix with respect to real-world attacks. The prompts are : (1) A house in the style of tormented Dutch artist; (2) A house in the style of Dutch artist with one ear; (3) A painting of a house in the style of van gog; (4) Starry night painting; (5) David Beckham standing in front of the Eiffel tower; (6) An image of David Beckham standing in front of Eifol tower; (7) An image of David Beckham standing in front of Eiffel landmark; (8) An image of David Beckham standing in front of Eiffel landmark in Paris. Notice that (3) and (6) have deliberate typos in the prompt.

If an attacker has white-box access to the model weights, they can fine-tune the weights of the text-encoder to re-introduce the concept back into the underlying text-to-image model. For e.g., if one has access to the weights of an edited model (e.g., from where the concept of *Van Gogh* has been removed), they can collect a dataset comprising of *Van Gogh* paintings and use their associated captions to fine-tune *only* the text-encoder. This can potentially re-introduce the concept of *Van Gogh* back into the model. A skilled machine learning practitioner can also engineer attacks via complex prompt engineering to bypass edited concepts. Given that the concepts are not removed from the UNet due to the inherent difficultly associated with the distributed knowledge, an optimized prompt can potentially still generate an image with the concept from the model. We believe that constructing adversarial attacks and evaluating robustness of the edited model to such attacks presents an interesting line of future work.

L PRE-TRAINING DETAILS FOR THE REPRESENTATIVE MODEL

In our experiments, we use `Stable-Diffusion v2.0`. This model is pre-trained on image-text pairs from LAION-2B dataset (Schuhmann et al., 2022) and LAION-improved aesthetics. We highlight that our interpretability framework can be used with other Stable-Diffusion versions also, but for representative purposes, we use Stable-Diffusion v2.0.

M ADDITIONAL CAUSAL TRACING RESULTS

M.1 PERTURBING THE ENTIRE TEXT-EMBEDDING

In our causal tracing experiments so far, we have only added Gaussian noise to the span of the token containing the attribute (e.g., adding noise to *apple* in the case of *Object* attribute for a prompt such as *'A photo of an apple in a room'*). In this section, we provide a more fine-grained control over the attribute of relevance in the caption to perform causal tracing. In particular, we replace the entire caption embedding with Gaussian noise, at all the cross-attention layers to the right of the intervention site in the UNet. We visualize a subset of the results in Fig 55 where we show results corresponding to causal and non-causal states. For down-blocks.1.resnets.1 which is one of the causal states for *Objects*, the relevant objects are restored in the generated image. This shows that the activations of certain layers in the UNet act as signatures for visual attributes and these signatures are able to generate the correct image even though the captions across cross-attention layers are completely replaced by Gaussian noise.

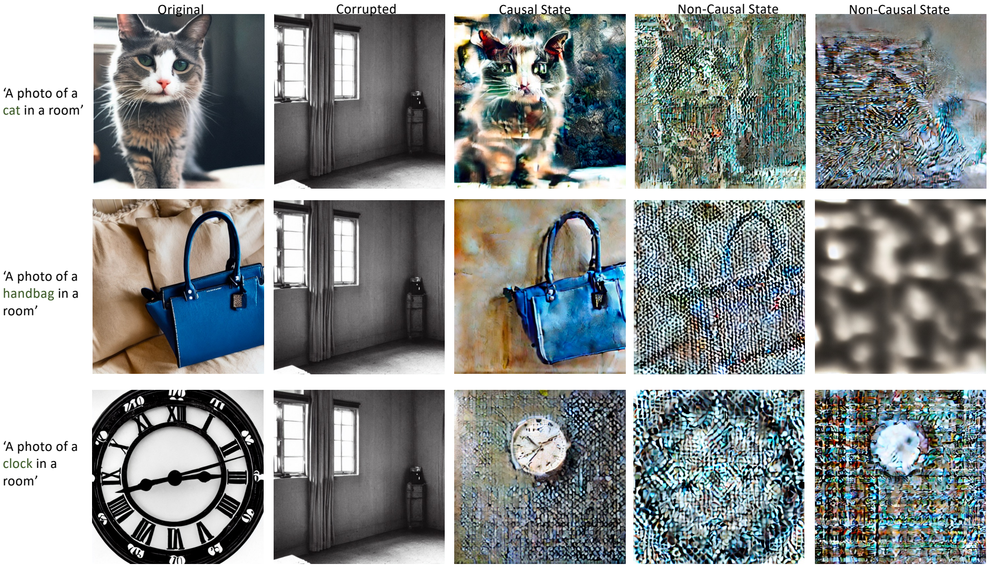

Original
Corrupted
Causal State
Non-Causal State
Non-Causal State
'A photo of a cat in a room'
'A photo of a handbag in a room'
'A photo of a clock in a room'

Figure 55: **Causal Tracing for the UNet, when the entire text is replaced using Gaussian noise across all the cross-attention layers to the right of the intervention site.** We use down-blocks.1.resnets.1 as the causal state across all the prompts in the visualization, whereas the non-causal states are picked randomly.

N DESIGN OF THE PROMPT DATASET FOR MODEL EDITING

The concepts used for editing the text-to-image model using DIFF-QUICKFIX is borrowed from (Kumari et al., 2023). In particular, the dataset in (Kumari et al., 2023) consists of concepts to be edited from the *style* and *object* categories. For *style*, the concepts edited are as follows : *{Greg Rutkowski, Jeremy Mann, Monet, Salvador Dali, Van Gogh}*. For *object*, the concepts to be edited are as follows: {*cat, grumpy cat, Nemo, R2D2, Snoopy}*. The exact prompts which are used with the edited model can be referred in the Appendix section of (Kumari et al., 2023). They can also be referred in Appendix H.

To remove multiple artistic styles as shown in Appendix H.1, we use the following set of artists: *{Thomas Kinkade, Van Gogh, Leonid Afremov, Monet, Edward Hopper, Norman Rockwell, William-Adolphe Bouguereau, Albert Bierstadt, John Singer Sargent, Pierre-Auguste Renoir, Frida Kahlo, John William Waterhouse, Winslow Homer, Walt Disney , Thomas Moran, Phil Koch, Paul Cézanne, Camille Pissarro, Erin Hanson, Thomas Cole, Raphael, Steve Henderson, Pablo Picasso, Caspar David Friedrich, Ansel Adams, Diego Rivera, Steve McCurry, Bob Ross, John Atkinson Grimshaw, Rob Gonsalves, Paul Gauguin, James Tissot, Edouard Manet, Alphonse Mucha, Alfred Sisley, Fabian Perez, Gustave Courbet, Zaha Hadid, Jean-Leon Gerome, Carl Larsson, Mary Cassatt, Sandro Botticelli, Daniel Ridgway Knight, Joaquin Sorolla, Andy Warhol, Kehinde Wiley, Alfred Eisenstaedt, Gustav Klimt, Dante Gabriel Rossetti, Tom Thomson }* These are the top 50 artists who artworks are represented in Stable-Diffusion.

To update the text-to-image model with facts,we use the following concepts: *{President of the United States, British Monarch, President of Brazil, Vice President of the United States, England Test Cricket Captain}*. The correct fact corresponding to each of these concepts (in the same order) are : *{Joe Biden, Prince Charles, Lula Da Silva, Kamala Harris, Ben Stokes}*, whereas the incorrect facts which are generated by the text-to-image model are *{Donald Trump, Queen Elizabeth, Bolsanaro, Mix of US politicians, Random English Cricketer}*.

O QUALITATIVE COMPARISON WITH OTHER MODEL EDITING METHODS

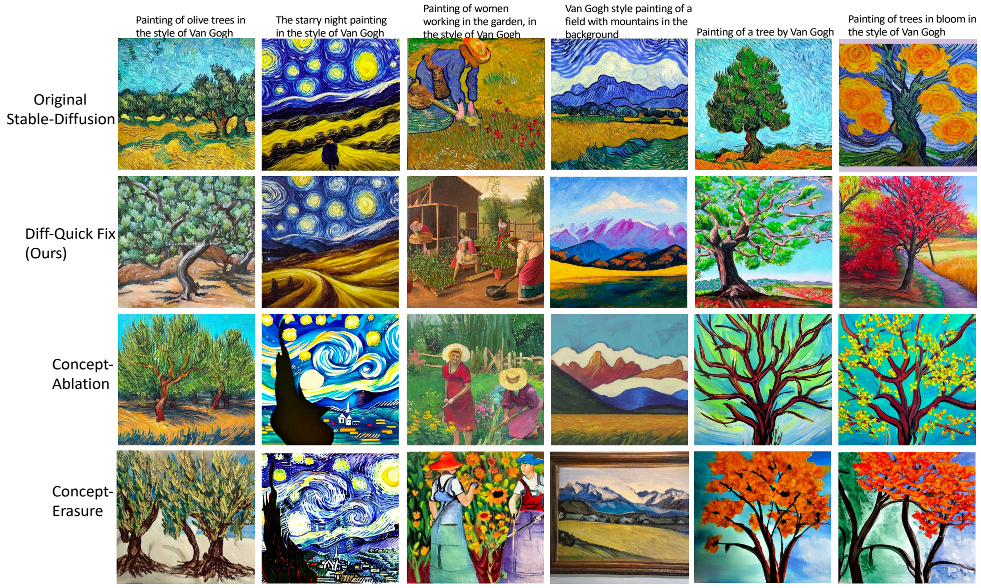

Painting of olive trees in the style of Van Gogh
The starry night painting in the style of Van Gogh
Painting of women working in the garden, in the style of Van Gogh
Van Gogh style painting of a field with mountains in the background
Painting of a tree by Van Gogh
Painting of trees in bloom in the style of Van Gogh
Original Stable-Diffusion
Diff-Quick Fix (Ours)
Concept-Ablation
Concept-Erasure

Figure 56: **Qualitative Comparison with Different Model Editing Methods**: (i) Concept-Ablation (Kumari et al., 2023); (ii) Concept-Erasure (Gandikota et al., 2023a) and the original unedited Stable-Diffusion baseline. Note that both Concept-Ablation and Concept-Erasure are fine-tuning based methods.