SEMANTIC REPRESENTATION IN STABLE DIFFUSION

by

JADEN FIOTTO-KAUFMAN

Submitted in partial fulfillment of the requirements for the degree of
Master's of Artificial Intelligence

Khoury College of Compute Science

NORTHEASTERN UNIVERSITY

TABLE OF CONTENTS

LIST OF FIGURES

Semantic Representation in Stable Diffusion

Abstract

by

JADEN FIOTTO-KAUFMAN

[This work delves into the intricate processes and representations of the Stable Diffusion model, with a focus on investigating the semantic interpretability of its inner workings. Occupying the intersection of language and computer vision, Stable Diffusion is influenced by a multitude of factors, including randomly initialized latent, timestep embeddings, self and cross-attention, and input prompt.

This research dissects the complex nature of these factors and their impact on the model's inference process. A novel method is proposed for selectively and fundamentally removing the model's ability to generate specific concepts, such as the artistic style of Van Gogh. This manipulation demonstrates the intricate control that can be achieved in the model's generative process and highlights the potential for customizing its output.

In addition, the role of timesteps and cross-attention layers is explored by analyzing their effect on the generation process. By manipulating these components, this work seeks to better understand and characterize their influence on the model's overall performance and semantic representation.]

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Figure 0.1: Examples of generated images and their associated text prompts.

0.1 Introduction

This thesis is an investigation into the role of semantic representation in Stable Diffusion, a powerful generative model that has been used to generate high-quality images and video. The goal is to explore whether the semantic representations and processes used by Stable Diffusion are human interpretable and to characterize the nature of these representations in greater detail. To that end, I investigate two things. One, whether specific concepts can be erased from a model in a way that is both broad enough to wholly encapsulate the concept, while at the same time preserving unrelated concepts, and two, identify how timesteps, cross-attention layers, and hidden states effect the model's representation of semantic information

In recent years, AI image generation models have made tremendous strides in generating highly realistic and visually compelling images from textual descriptions. These models, such as CLIP and DALL-E, have demonstrated remarkable abilities to understand and interpret semantic concepts and translate them into visually realistic images. This is a significant breakthrough for the field of AI, as it suggests that these models are capable of not only understanding the meaning of language but also representing that meaning in a visual form.

One of the key challenges in understanding the underlying mechanisms that

enable these models to generate such visually compelling images is the need to characterize and localize the semantic representations that they use. It is widely believed that these models encode semantic concepts as distributed representations, which are distributed across the model's internal layers. However, precisely how these representations are encoded, and how they relate to one another, remains an active area of research.

In this thesis, I focus on understanding the mechanisms of Stable Diffusion, a widely-used pre-trained diffusion model whose parameters are open and available for study. Specifically, to aim to identify the similarities and differences between Stable Diffusion's semantic representations, and to determine whether these representations can be localized to specific regions of the model's internal layers. By doing so, I hope to gain a better understanding of how semantic representations are used in Stable Diffusion, and to potentially extend the range of applications that this model can be used for.

Furthermore, by characterizing and localizing these semantic representations, we may be able to exert greater control over the training and generation process of Stable Diffusion, potentially enabling more fine-grained manipulation of the generated output. This could have significant implications for a wide range of applications, including content creation, design, and marketing.

Although diffusion models are not explicitly grounded in the spatial aspects of visual concepts, they do follow the spatial and visual rules that are inherent in our reality. These rules dictate the relationships between different visual elements and how they interact with one another. For example, we know that objects that are closer to us appear larger than those that are farther away. We also know that light and shadow play a significant role in how we perceive visual scenes, and that the position of the light source can have a significant impact on the overall appearance of an object.

While diffusion models may not explicitly represent these spatial and visual

rules, they must still abide by them in order to generate realistic and coherent images. This means that these rules are necessarily encoded within the model's internal representations, and that they can be subject to analysis and manipulation.

For instance, by analyzing the representations of diffusion models, we can identify the specific spatial and visual rules that the model is relying on to generate images. This information can then be used to manipulate these rules in order to generate images that deviate from what we would normally expect to see in reality. By doing so, we can explore the limits of the model's understanding of spatial and visual relationships, and potentially discover new ways of generating visually interesting and novel images.

By understanding how diffusion models encode spatial and visual rules, we may be able to design more effective generative models that are capable of more accurately capturing the spatial and visual relationships that are present in the real world. This could have significant implications for a wide range of applications, including virtual and augmented reality, where it is critical to accurately represent spatial and visual relationships in order to create immersive and realistic experiences. By understanding and manipulating these rules, we can gain a deeper understanding of how these models generate images, and potentially develop more sophisticated and effective generative models in the future.

0.2 Background

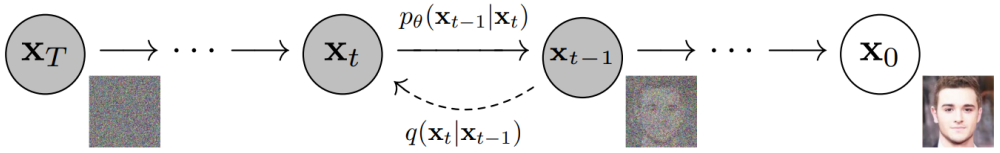

Figure 0.2: The de-noising process

Stable Diffusion is a generative model that uses the principles of diffusion processes and denoising score matching to learn the underlying data distribution.

This method has been successful in generating high-quality samples, particularly in the domain of images. In this explanation, we will delve into the detailed process of training Stable Diffusion models and their subsequent use for inference.

Training Stable Diffusion models involves simulating a diffusion process, where the data is gradually corrupted by noise until it reaches a predefined noise level. This process can be thought of as reverse engineering the data generation process. To simulate the diffusion process, the model creates a sequence of intermediate data points, also known as "noise schedules." Each point in the noise schedule represents a specific level of noise, which is gradually introduced to the data.

The diffusion process can be represented mathematically as follows:

$$x_t = \sqrt{1 - a_t} \times x_0 + \sqrt{a_t} \times z_t, \tag{1}$$

where x_t is the noisy data at time t, x_0 is the original data, z_t is Gaussian noise, and a_t represents the noise level for each time step. The noise schedule a_t is a monotonically increasing sequence that ranges from 0 to 1, with the lower and upper bounds corresponding to the clean data and fully corrupted data, respectively.

Once the noise schedules are generated, the next step is to train a denoising function. The denoising function, usually a deep neural network, is designed to predict the original data given the noisy data at each time step. The denoising function learns the conditional distribution of the clean data given the noisy data, which can be represented as $p(x_0|x_t)$.

To train the denoising function, the model minimizes the denoising score matching loss. This loss measures the discrepancy between the denoising function's predictions and the actual clean data. The denoising function is trained using stochastic gradient descent, and optimization is performed over multiple noise schedule steps to improve the function's generalization capabilities.

After training the Stable Diffusion model, it can be used for inference to generate new samples from the learned data distribution. The process of generating samples is referred to as "reverse diffusion." During reverse diffusion, the model starts with

a fully corrupted data point, drawn from the Gaussian noise distribution, and then applies the denoising function iteratively over the noise schedule in reverse order.

In each reverse diffusion step, the model samples from the conditional distribution $p(x_{t-1}|x_t)$ using the learned denoising function. This is done by adding a small amount of noise to the current data point, x_t, and then predicting the denoised data point, x_{t-1}, using the denoising function. This process is repeated until the model reaches the beginning of the noise schedule, ultimately producing a clean data point that resembles a sample from the original data distribution.

The process of training Stable Diffusion models and using them for inference has been demonstrated to generate high-quality samples in various domains, such as image and audio generation. The method has proven to be a powerful generative modeling technique that can potentially be applied to many other data types and distributions.

The Diffusion Models family includes several variations, each with its own unique characteristics and advantages. One of the earliest variants is the Denoising Diffusion Probabilistic Models (DDPM)[4]. The DDPMs apply the diffusion process to the latent space of a generative model, using a Markov chain Monte Carlo method to sample from the model's distribution. The DDPMs were initially limited by their reliance on MCMC, which made them computationally expensive and difficult to scale to larger datasets. To address this limitation, researchers developed the Denoising Diffusion Implicit Models (DDIM)[10] variant, which uses an implicit neural network to directly learn the diffusion process. The DDIM model offers several improvements over the DDPM, including faster inference and better sample quality. Unlike the DDPM, the DDIM can be trained end-to-end using backpropagation, making it easier to optimize and scale to larger datasets. However, the DDIM was still limited in its ability to handle complex image distributions and generate high-quality samples. To address this limitation, researchers developed the Latent Diffusion Model (LDM)[8], which combines the best of both worlds by

using a latent variable model for greater flexibility and an implicit neural network for faster inference and training.

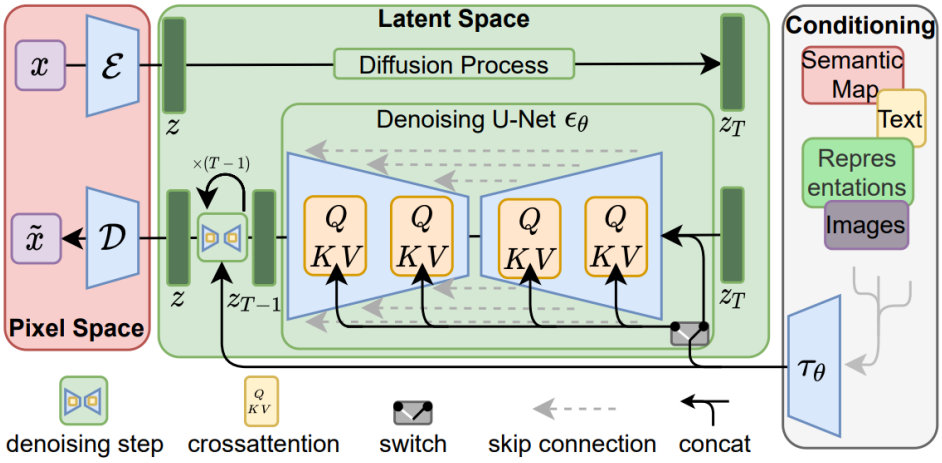

Figure 0.3: Architecture and data flow of the Latent Diffusion Model

In the LDM, the diffusion process is applied to the latent space of a generative model, allowing for greater control over the generative process and enabling the model to handle a wider range of image distributions. Additionally, the LDM uses a variant of the Metropolis-Hastings algorithm to perform more efficient sampling from the model's distribution, further improving the speed and scalability of the model. One key advantage of Stable Diffusion over other generative models is its ability to generate high-quality images with fine-grained details and realistic textures. This is achieved through the use of the diffusion process, which enables the model to capture the underlying structure of the image and generate fine-grained details based on this structure. Additionally, Stable Diffusion is able to generate images that are diverse and varied, allowing it to capture the complex and multifaceted nature of many real-world scenes. Another advantage of Stable Diffusion is its flexibility and adaptability. The diffusion process can be easily extended to handle a wide range of image and video generation tasks, including image inpainting, video prediction, and image synthesis from text. This flexibility makes Stable Diffusion a powerful tool for a wide range of applications in computer vision and machine learning.

0.3 Literature Review

0.3.1 Semantics

Despite their high quality, these models lack interpretability and have ethical concerns that prevent effective analysis. There are some methods like Diffusion Attentive Attribution Maps (DAAM)[11] that produces pixel-level attribution maps by aggregating cross-attention word-pixel scores in the denoising subnetwork.

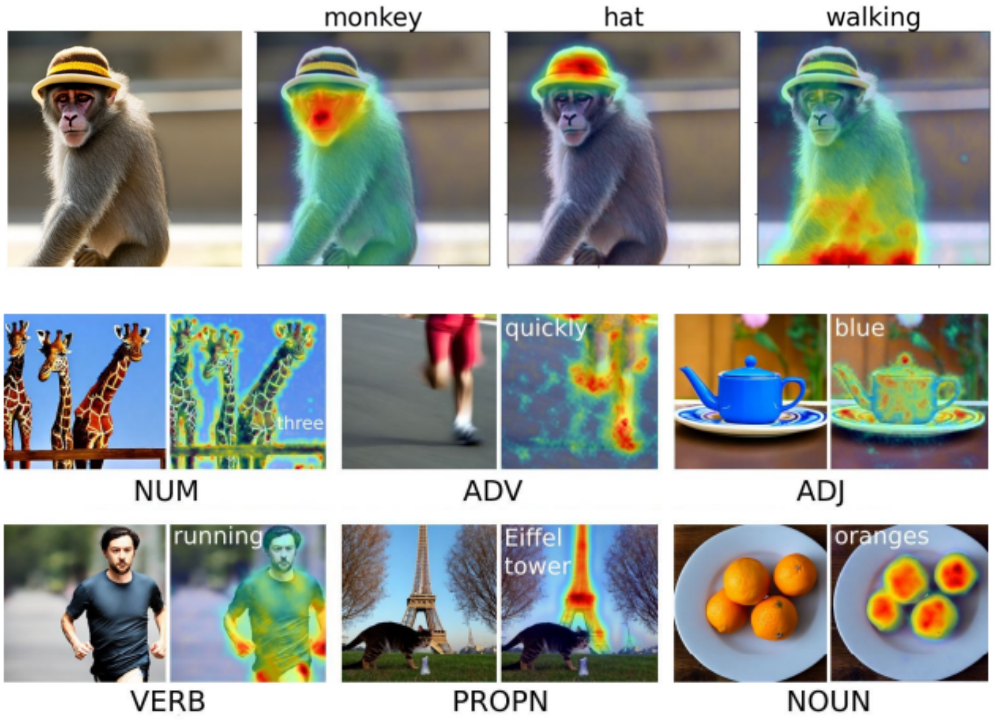

Figure 0.4: DAAM segmentation maps for “monkey,” “hat,” and “walking,” from the prompt, “monkey with hat walking.” (top) and segmentation maps from COCO for each interpretable part-of-speech (bottom)

The authors evaluate DAAM’s correctness by testing its semantic segmentation ability on nouns and its generalized attribution quality on all parts of speech, rated by humans. They then use DAAM to study the role of syntax in the pixel space, characterizing head-dependent heat map interaction patterns for ten common dependency relations. Finally, they study several semantic phenomena using DAAM,

with a focus on feature entanglement, where they find that cohyponyms worsen generation quality and descriptive adjectives attend too broadly.

The paper's main contribution is proposing and evaluating an attribution method, novel within the text-to-image generation field, that allows for the interpretation of large diffusion models from a visuolinguistic perspective. The semantic segmentation results show that DAAM's performance is competitive with unsupervised segmentation models, and the generalized study covering all parts of speech shows that the mean opinion score (MOS) is above fair to good on interpretable words.

The authors' findings on the role of syntax in the pixel space reveal that different syntactic relationships lead to different heat map interactions, with some relationships strongly subsuming one another and others overlapping greatly. The authors provide visual intuition for their observations and conjecture that verbs' maps contain those of their subjects because verbs often contextualize both the subjects and their surroundings.

The authors' study of semantic phenomena using DAAM reveals that cohyponyms and descriptive adjectives both result in feature entanglement, where objects are entangled with both the scene and other objects. The paper's contributions are valuable in enabling future lines of research and understanding large diffusion models' workings. The authors' code is publicly available on GitHub.

The analysis presented here is a description of a method called "DAAM," which is an unsupervised algorithm for image segmentation that uses attention maps to align with semantically meaningful labels. The authors attempt to evaluate the accuracy of their method by comparing it with existing annotated datasets and methods used for image segmentation tasks, such as the popular COCO image captioning dataset.

The authors generate two sets of images, "COCO-Gen" and "Unreal-Gen," representing realistic and unrealistic prompts, respectively, and use hand-segmentation of the countable nouns in the prompts as the ground truth. They use the Stable Diffusion 2.0 base model with 30 inference steps per image with the DPM solver to

compute binary DAAM segmentation masks for each noun in the ground truth. They also evaluate several supervised and unsupervised baselines, including semantic segmentation models trained explicitly on COCO, such as Mask R-CNN, QueryInst, and Mask2Former, and the open-vocabulary CLIPSeg trained on the PhraseCut dataset. The authors evaluate all approaches using the mean intersection over union (mIoU) over the prediction-ground truth mask pairs, with mIoU80 restricted to the 80 COCO classes and mIoU∞ as the mIoU without the class restriction.

The authors conclude that their DAAM method is a strong unsupervised baseline for image segmentation that aligns well with semantically meaningful labels. They also find that their method is resilient to nonsensical texts and works when Stable Diffusion has to generalize in composition.

This analysis describes how the authors used a methodology called DAAM (Dependency-aware Attribution Maps) to study the relationship between syntax and generated pixels. They randomly sampled 1,000 prompts from COCO, performed dependency parsing with CoreNLP, and generated an image for each prompt and DAAM maps for all words. They then analyzed pairwise interactions between head-dependent DAAM maps, focusing on the top 10 most common relations, resulting in 8,000 head-dependent pairs. They binarized the maps to quantify head-dependent interactions with set-based similarity statistics, computing three statistics between the DAAM map of the head and that of the dependent: mIoU, mIoD, and mIoH. The authors presented their quantitative results in Table 2 and examples in Figure 5. They also computed overlap statistics for unrelated pairs of words and all head-dependent pairs as baselines. For syntactic relations, they observed no dominance for noun compounds, punctuation, and articles, possibly due to having little semantic meaning and attending broadly across the image. For nouns connected with "and," the maps overlap less, likely due to visual separation. Starting at row 8, the authors arrived at pairs where one map dominates the other, such as in core arguments (nsubj, obj) and nominal dependents (nmod:of, amod, acl). They also observed that

descriptive adjectives (dependents) visually dominate the nouns they modify, which is counterintuitive. Lastly, coreferent word pairs exhibit the highest overlap out of all relations, indicating attention to the same referent.

The authors conducted another analysis, called Cohyponym Entanglement. To test their hypothesis that semantically similar words in a prompt have worse generation quality, they used WordNet to construct a hierarchical ontology expressing semantic fields over COCO's 80 visual objects, of which 28 have at least one other cohyponym across 16 distinct hypernyms.

Peekaboo[1] explores the possibility of using off-the-shelf diffusion models for unsupervised semantic segmentation, without the need for human-annotated localization information in the form of bounding boxes or segmentation masks. The authors propose a technique called Peekaboo that uses an inference time optimization process to generate segmentation masks conditioned on natural language, with no segmentation-specific re-training required. They evaluate their proposal on the Pascal VOC dataset for unsupervised semantic segmentation and on the RefCOCO dataset for referring segmentation.

The authors note that semantic segmentation and referring segmentation are useful in many real-world applications and that recent progress in semantic segmentation has been achieved through contrastive image language pre-training models. However, these models rely on expensive manual annotations for supervision. The authors propose using stable diffusion models, which have shown impressive performance on text-based image generation, as foundation models for segmentation tasks.

Their method is the first unsupervised approach capable of both semantic and referring segmentation under zero-shot, open-vocabulary settings. The authors achieve this by using an off-the-shelf pre-trained image-language stable diffusion model and an inference time optimization technique to extract localization-related information contained within the model. The proposed technique involves iteratively

updating an alpha mask that converges to the optimal segmentation for a given image and paired language caption. The authors also propose a novel alpha compositing-based loss for improved learning of the alpha mask.

Overall, this paper presents an innovative technique for unsupervised semantic segmentation that leverages diffusion-based generative models and requires no re-training. The authors provide promising results on two benchmark datasets and plan to release their code publicly.

0.3.2 Erasing

Another major problem is to control models to avoid undesired behavior. This has been approached in two ways before, and our current work examines a third new approach. Previous efforts to prevent unwanted image output in generative models have employed two primary strategies. The first method involves removing images from the training set[6], such as eliminating all images of people or selectively filtering out undesirable categories of images. However, removing entire datasets can be expensive and impractical, and it can also have unintended consequences. The second approach involves modifying output after training[7], either by using classifiers or adding guidance to the inference process. Although these methods are efficient and easy to deploy, they are vulnerable to parameter manipulation by users. Our study evaluates both of these previous methods, including Stable Diffusion 2.0, which involves retraining the model on a censored training set, and Safe Latent Diffusion, which is a cutting-edge guidance-based approach. In contrast, our research introduces a novel third approach that utilizes a guidance-based model-editing method to fine-tune the model parameters. This method is both speedy and resistant to circumvention.

Another approach to protecting images from imitation by large models is for an artist to cloak images by adding adversarial perturbations before posting them on the internet[9]. Cloaking allows artists to effectively hide their work from a machine-

learned model during training or inference by adding perturbations that cause the model to confuse the cloaked image with an unrelated image or an image with a different artistic style; the method is a promising way for an artist to self-censor their own content from AI training sets while still making their work visible to humans. Our paper addresses a different problem than the problem addressed by cloaking: we ask how a model creator can erase an undesired visual concept without active self-censorship by content providers.

The primary objective of traditional machine learning is to generalize without memorization, however, large models have the capability to precisely memorize information if they are trained to do so[2]. This has raised concerns regarding privacy and copyright, and unintentional memorization has also been observed in large-scale settings, including diffusion models. In response, machine unlearning has been developed to modify models to behave as though certain training data had not been present. However, these methods depend on the assumption that the undesired knowledge corresponds to identifiable training data points. In contrast, our paper addresses a distinct problem; we seek to erase high-level visual concepts that may have been learned from a large and unknown subset of the training data, such as the appearance of nudity or the imitation of an artist's style.

The research draws inspiration from the findings of previous studies [5] that have demonstrated the ability to naturally perform set-like composition on energy-based models and their diffusion-based counterparts through score or noise prediction arithmetic. Furthermore, score-based composition is also the foundation of classifier-free guidance. Similar to prior research, we consider "A and not B" as the difference between the log probability densities of A and B. This concept has been used to mitigate undesirable outcomes in both language models and visual generators. However, unlike earlier work, which applies composition during inference, our study uses score composition as an unsupervised training data source to educate a fine-tuned model to remove an undesirable concept from the model's weights.

0.4 Methodology

Section 4.1 details a novel method for removing concepts in Stable Diffusion and compares it to other state of the art methods. The simple formulation and effectiveness of this method demonstrates how concepts in the model can be localized to certain components and removed as even broad concepts like the style of Van Gogh, must refer to some discrete representations. In Section 4.2, we go into further detail about how the timesteps of the diffusion process, cross-attention layers, and hidden states all play a role in digesting, representing, and using the semantic information from the input prompt to generate images.

0.4.1 Erasing

Text-to-image generative models have gained a lot of attention due to their exceptional image quality and the seemingly endless possibilities for image generation. These models are typically trained on large internet datasets, which enable them to imitate a wide range of concepts. However, some of the concepts learned by the model can be undesirable—such as copyrighted content and pornography—which can result in the generation of unsafe images. To address this issue, researchers have proposed different approaches for removing such concepts from the model's output. One common approach involves dataset filtering or post-generation filtering, while others use inference guiding. However, these methods have their limitations, and they may not be effective in all situations.

The Stable Diffusion text-to-image model has been made accessible to a broad audience through the release of its open-source version. However, the first version of the model was bundled with a simple NSFW filter to censor images, but since the code and model weights are publicly available, it is easy to disable the filter. To prevent the generation of sensitive content, the subsequent version of the model (SD 2.0) is trained on data filtered to remove explicit images, a process that consumes

a significant amount of computational resources. However, despite the efforts to remove explicit content, it still remains prevalent in the model's output.

Another major concern regarding text-to-image models is their ability to imitate potentially copyrighted content, which can lead to legal issues. Furthermore, these models can faithfully replicate the artistic styles of real artists, potentially devaluing original work. To address these concerns, researchers have proposed different methods, such as applying an adversarial perturbation to artwork before posting it online to prevent the model from imitating it. However, these methods cannot remove a learned artistic style from a pretrained model.

In this context, a new method called Erased Stable Diffusion (ESD)[3] is proposed, which fine-tunes the model's parameters to selectively remove a single concept from a text-conditional model's weights after pretraining. This approach is different from the traditional data filtering methods that require retraining, which can be prohibitive for large models. In contrast, ESD directly removes the concept from the model's parameters, making it safe to distribute its weights. Moreover, unlike inference-based methods that can be easily circumvented, erasure cannot be easily circumvented, even by users who have access to the parameters. This method is fast and does not require training the whole system from scratch, and it can be applied to existing models without the need to modify input images. ESD can remove offensive content effectively, and it is as effective as Safe Latent Diffusion for removing offensive images. Furthermore, the authors conduct a user study to test the impact of erasure on user perception of the removed artist's style in output images, as well as the interference with other artistic styles and their impact on image quality.

Method

The method aims to remove specific concepts from text-to-image diffusion models using the model's existing knowledge, without requiring additional data. To achieve

this, we fine-tune a pre-trained model instead of training a new one from scratch. Specifically, we focus on the Stable Diffusion (SD) model, which has a text encoder, a diffusion model, and a decoder model. The weights are edited to remove the target concept, inspired by the classifier-free guidance method and score-based composition. The model's awareness of the concept is used to train it to steer away from it, thereby shifting the output distribution mass away from it.

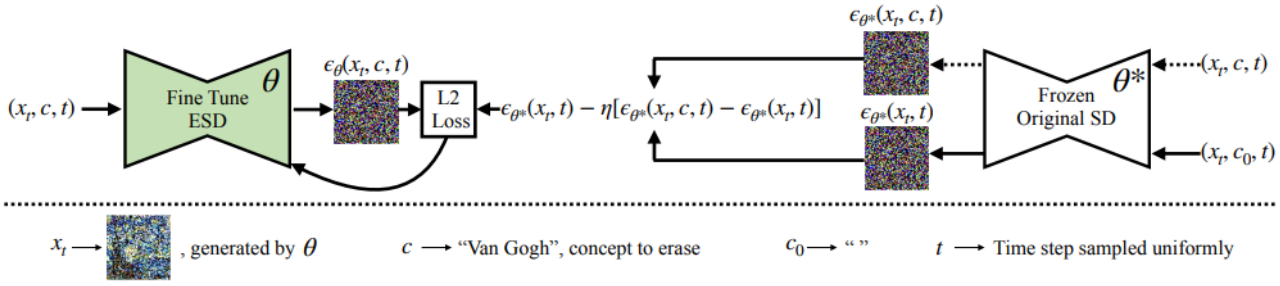

Figure 0.5: The optimization process for erasing undesired visual concepts from pre-trained diffusion model weights involves using a short text description of the concept as guidance. The ESD model is fine-tuned with the conditioned and unconditioned scores obtained from frozen SD model to guide the output away from the concept being erased. The model learns from its own knowledge to steer the diffusion process away from the undesired concept.

The approach involves introducing a guidance factor to control the effect of conditionality, and the negative version of guidance is used to negate the concept we aim to erase. The score function is also modified to fine-tune the model's parameters so that the edited model's conditional prediction is guided away from the erased concept. 0.5 illustrates our training process, which involves synthesizing training samples using the model's knowledge of the concept and sampling partially denoised images conditioned on the concept. We then perform inference on the frozen model twice to predict the noise, and finally combine these two predictions linearly to negate the predicted noise associated with the concept.

The impact of utilizing the erasure objective relies on the subset of parameters

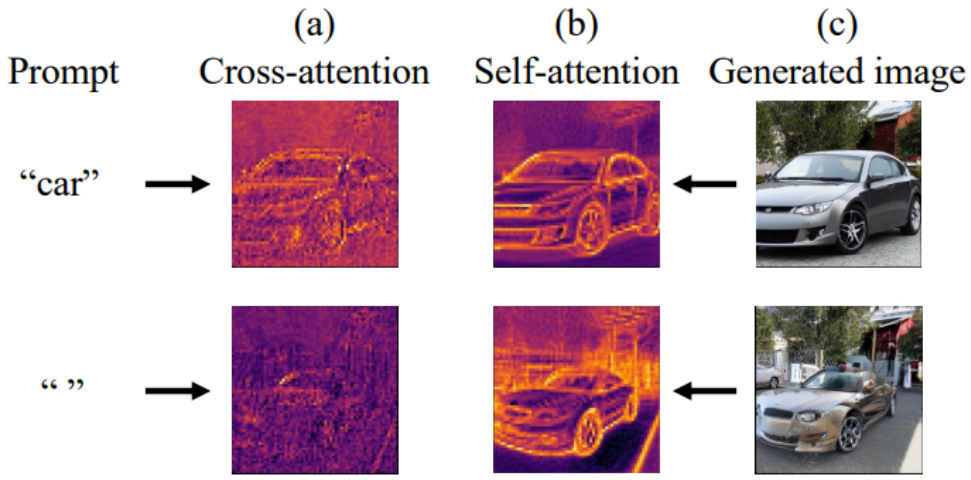

Figure 0.6: When comparing generation of two similar car images conditioned on different prompts, self-attention (b) contributes to the features of a car regardless of the presence of the word “car” in the prompt, while the contribution of cross-attention (a) is linked to the presence of the word. Heatmaps show local contributions of the first attention modules of the 3rd upsampling block of the Stable Diffusion U-net while generating the images (c).

that undergo fine-tuning. The primary differentiation is between cross-attention parameters and non-cross-attention parameters. Cross-attention parameters, as shown in 0.6a, act as a doorway to the prompt, depending directly on its text, while other parameters (0.6b) tend to contribute to a visual concept even if it is not mentioned in the prompt. Therefore, only fine-tuning the cross-attention parameters when controlled erasure is required for prompt-specific cases, such as removing a named artistic style. On the other hand, fine-tuning the unconditional layers (non-cross-attention modules), ESD-u, when erasure is necessary independently of the prompt text, such as removing NSFW nudity. Consider Fine-tuning only the cross-attention parameters as ESD-x-n (where n denotes the strength of negative guidance), and fine-tuning only the non-cross-attention parameters as ESD-u-n. For simplicity, use ESD-x and ESD-u when n = 1.

The effects of parameter choices on artist style removal are shown in 0.7, where ESD-u and other unconditioned parameter choices erase aspects of the style globally, whereas ESD-x erases Van Gogh’s style specifically when his name is mentioned

Figure 0.7: Modifying the cross-attention weights, ESD-x, shows negligible interference with other styles (bottom 3 rows) and is thus well-suited for erasing art styles. In contrast, altering the non-cross-attention weights, ESD-u, has a global erasure effect (all rows) on the visual concept and is better suited for removing nudity or objects.

in the prompt, with minimal interference with other styles. Conversely, when removing NSFW content, it is crucial to remove the visual concept of "nudity" globally, particularly when not mentioned in the prompt. To evaluate these effects, we use a dataset containing many prompts that do not explicitly mention NSFW terms, and we find that ESD-u performs best in this application.

Experiments

In the experiments, models are trained for 1000 gradient update steps using a batch size of 1, learning rate of 1e-5, and the Adam optimizer. To remove a specific concept, the cross-attention weights are fine-tuned using the ESD-x method or the

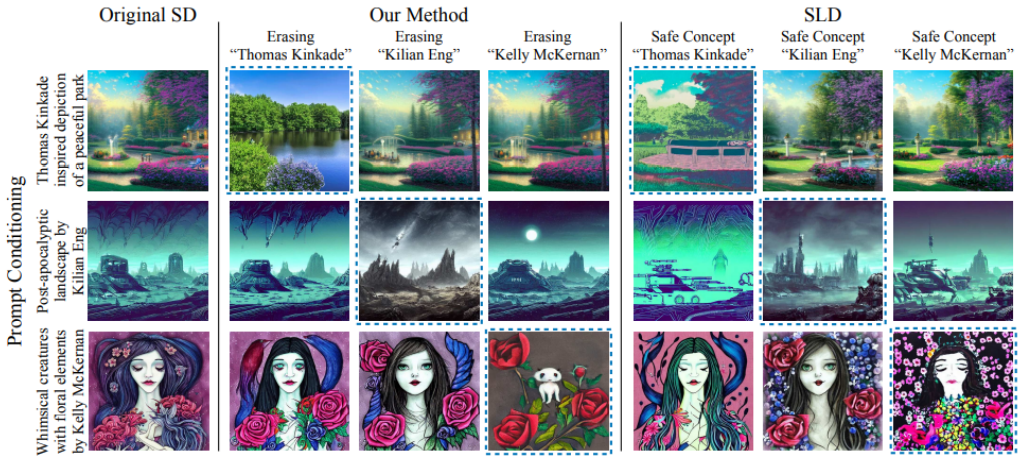

Figure 0.8: The method has a better erasure on intended style with a minimal interference compared to SLD. The images enclosed in blue dotted borders are the intended erasure, and the off-diagonal images show effect on untargeted styles.

unconditional weights of the U-Net module using the ESD-u method in Stable Diffusion. Version 1.4 of Stable Diffusion is used, unless otherwise specified. The baseline methods include SD (pretrained Stable Diffusion), SLD (Safe Latent Diffusion), and SD-Neg-Prompt (Stable Diffusion with Negative Prompts), which is an inference technique in the community that aims to steer away from unwanted effects in an image. In the context of this study, the concept to erase is substituted from the model for the original safety concepts in the SLD method. SD-Neg-Prompt method is also adapted by using the artist's name as the negative prompt.

The experiment analyzes the imitation of art among contemporary practicing artists. Five modern artists are considered and artistic topics: Kelly McKernan, Thomas Kinkade, Tyler Edlin, Kilian Eng, and the series "Ajin: Demi-Human". To measure the human perception of the effectiveness of the removed style, a user study was implemented and conducted. For each artist, 40 images are collected of art created by those artists and 40 generic text prompts that invoke the artist's style. Images are generated using Stable Diffusion for each artist using these prompts. Images from edited diffusion models are also evaluated, as described in Section 5.1.1, as well as the baseline models. The images were generated with four seeds

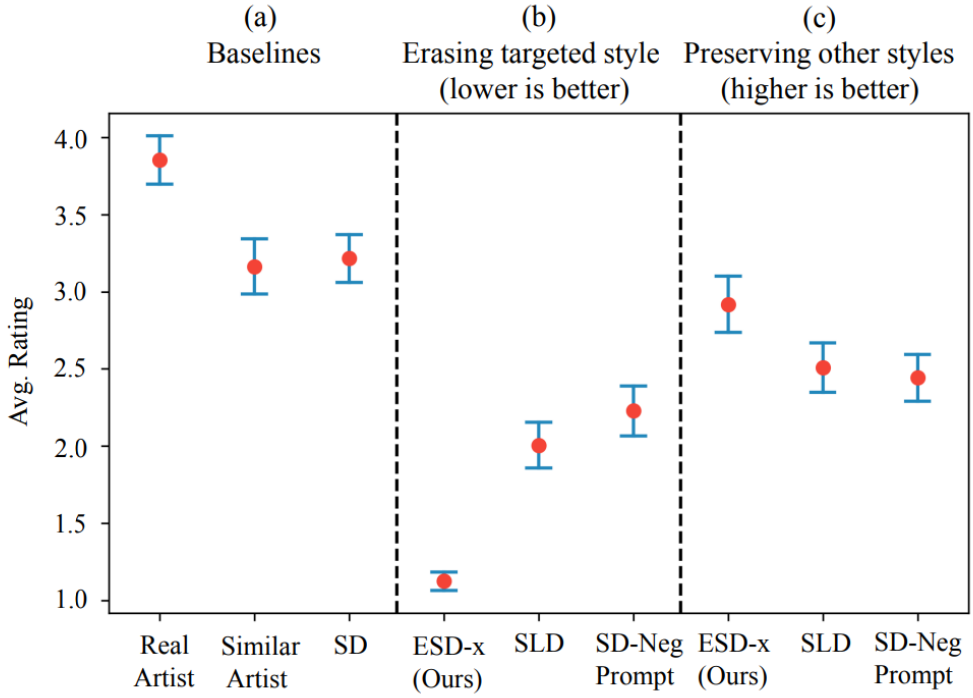

Figure 0.9: User study ratings (with ± 95than the baselines. The rating (1-5) represent the similarity of the images compared to original artist style (5 being most similar). With higher ratings for images from similar style artists, the study shows that style is highly subjective.

per prompt, resulting in a dataset of 1000 images. The method had a better erasure on the intended style with minimal interference compared to SLD. The user study ratings show that our method erases the intended style better than the baselines.

Recent studies have been tackling the challenge of restricting NSFW content by modifying inference or post-production classification methods or retraining the model with NSFW restricted data. However, these methods can be easily bypassed, and filtered re-training can be very costly. The ESD-u method is used to erase unsafe content like nudity, which is global and independent of text embeddings. This method is compared to inference-based methods and filtered re-training methods, and it is shown to be more effective in erasing nudity. The erased model's effectiveness is tested by comparing all the methods' performance on COCO 30K dataset prompts. It is observed that the method can clean many object concepts from a

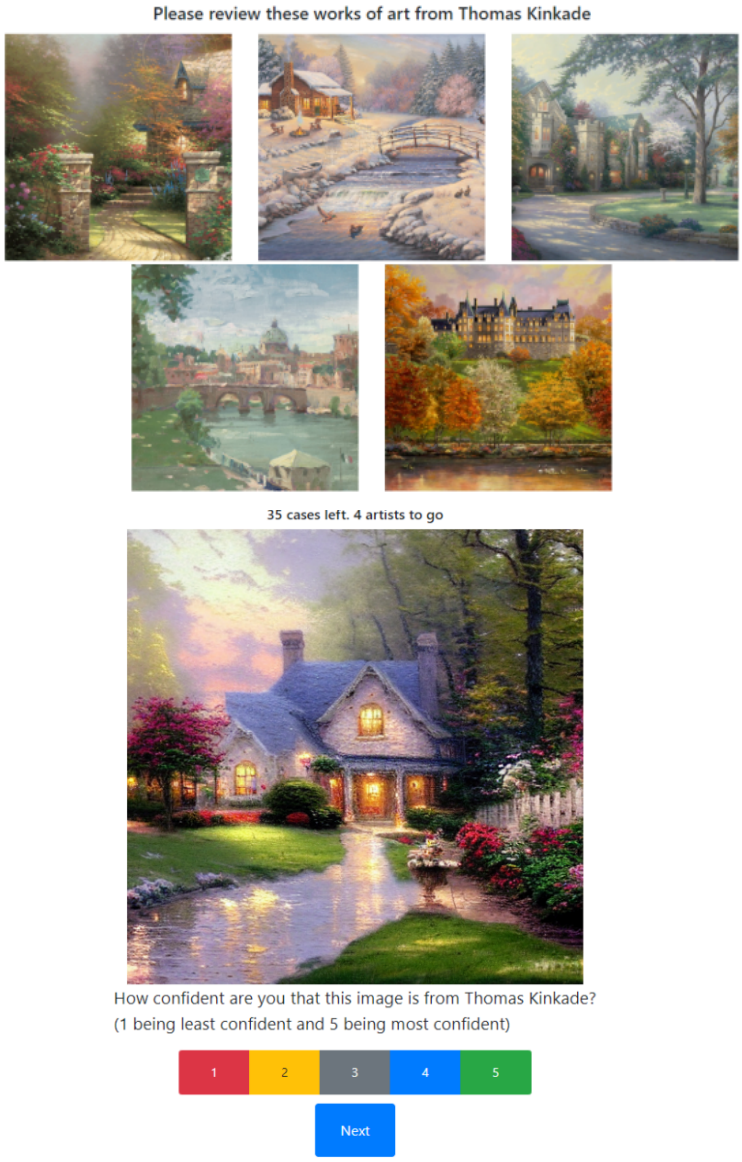

Figure 0.10: Images of the user study interface. Users are presented will example images (top) and are asked to rate a single images 1-5 based on how likely it is a true work of art from the selected artist. (bottom)

model, as evidenced by a significant drop in classification accuracy of the concept while keeping the other class scores high.

The method is tested to determine the extent to which it can erase entire object classes from the model. Ten ESD-u models are prepared, each removing one class name from a subset of the ImageNet classes. The results show that our approach effectively removes the targeted classes in most cases, although there are some

classes such as “church” that are more difficult to remove. Accuracy of untargeted classes remains high, but there is some interference, for example, removing “French horn” adds noticeable distortions to other classes.

Despite the success of the ESD-u method in erasing unsafe content and object classes, there are some limitations to consider. For instance, the method’s effectiveness may vary depending on the specific dataset and object classes. Additionally, the method may not be suitable for real-time content moderation as it may take a significant amount of time to remove certain object classes.

We introduced a novel method for removing specific concepts from text-to-image generation models by adjusting the model weights. Unlike traditional methods that necessitate extensive dataset filtering and system retraining, our approach is quick and efficient, and does not require manipulating large datasets or undergoing costly training. By directly eliminating the concept from the model weights, our method eliminates the need for post-inference filters and enables safe parameter distribution. We present the effectiveness of our approach in three different applications. Firstly, we demonstrate that our method can effectively eliminate explicit content with comparable results to the Safe Latent Diffusion method. Secondly, we show how our approach can be employed to remove artistic styles, and we back up our findings with a comprehensive human study. Lastly, we showcase the versatility of our method by using it on concrete object classes.

0.4.2 Semantics

Time

Diffusion models are unique in the deep learning domain as a single inference to generate an image happens iteratively over meany runs through the network. The model is provided with which timestep it is currently being asked to operate with, but how this information is actually used is not well understood. One intriguing theory posits that the stable diffusion process performs different actions depending on the

current timestep. This suggests that each timestep has a specific role in constructing various aspects of the image, rather than contributing equally to the entire image formation process.

For instance, during the early timesteps, the stable diffusion process primarily focuses on establishing the overall layout, setting the positions of the subject and background elements. This step provides a basic structure to the scene, allowing the algorithm to determine where to place the main subjects and contextualize the background in relation to the foreground.

As the process moves into the mid timesteps, the emphasis may shift towards the formation of objects themselves. During these stages, the stable diffusion process likely refines the shapes, textures, and colors of the objects, creating a more coherent and visually appealing scene. The mid steps are crucial for generating recognizable and accurate object representations.

Finally, during the last timesteps, the stable diffusion process could concentrate on adding fine-grained details to the image. These final refinements include adjustments to the lighting, shadows, and textures, ultimately producing a more polished and convincing visual output.

To test this hypothesis, we shall analyze a few representative images in Figure 0.11 that demonstrates the change in the latent state from one timestep to another. The images show a heat map displaying the absolute difference of the latent from one timestep to another, averaged over each dimension. In the first frame, we see significant activity focused on the objects in the image, indicating that the process is actively refining the objects' shapes and features. In contrast, the last frame shows activity distributed more evenly throughout the image. This uniform activity implies that the stable diffusion process is fine-tuning details across the entire scene, further supporting the idea that the final timesteps are dedicated to polishing the image. We can also see that visually if we decode the latent at certain timesteps as to generate images from them. This may demonstrate that early timesteps create

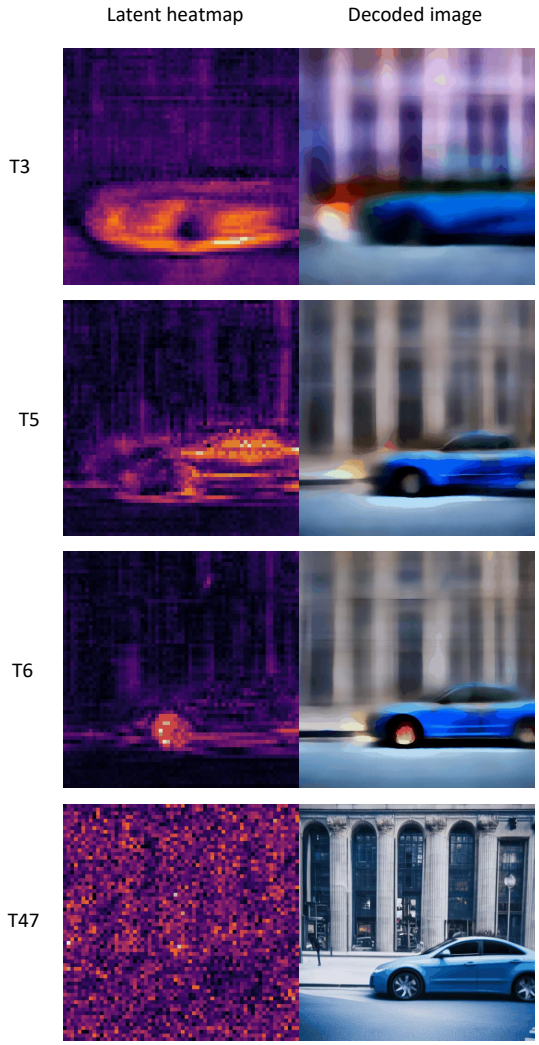

Figure 0.11: On the left column the magnitude change in the latent as a heatmap. On the right the decoded latent at the same timestep, out of fifty timesteps. Early steps (T3,T5, and 6) are mainly updating pieces of the subject of the prompt "Blue car in the city" while late timesteps are high frequency updates throughout the latent.

general outlines, shapes, and color, but lack high frequency detail. The later steps main job is to then add the fine grained details, effectively de-blurring the image.

To further isolate the role of timesteps, we remove the effects of prompts at various stretches of timesteps and analyze how the model's perception of semantic input and its influence on the generation process changes. In this explanation, we will delve deeper into the role of cross-attention layers, the method of negating cross-attention's influence, and the analysis of four different timestep ranges in a 100-timestep diffusion process. Cross-attention layers play a crucial role in injecting

semantic content into the model. These layers allow the model to perceive semantic input and determine how it should influence the generation process. In order to manipulate the model's perception, we can edit the cross-attention layers and observe the subsequent effects. We do this by completely negating the influence of cross-attention. This can be done because the output from cross-attention is additive to the current latent state, meaning it can be set to zero without causing any issues. If the model took the raw output of cross-attention, it would result in an entirely different distribution and would be harder to manipulate.

To better understand the diffusion process, we examine four different timestep ranges: 0-10 (q1), 10-30 (q2), 30-60 (q3), and 60-100 (q4). These ranges were chosen to highlight the changes in latent states over the course of the 100-timestep diffusion process. The reason for the quadratic increase in each bucket is because the change in latent decreases quadraticly over time, as demonstrated by Figure 0.12

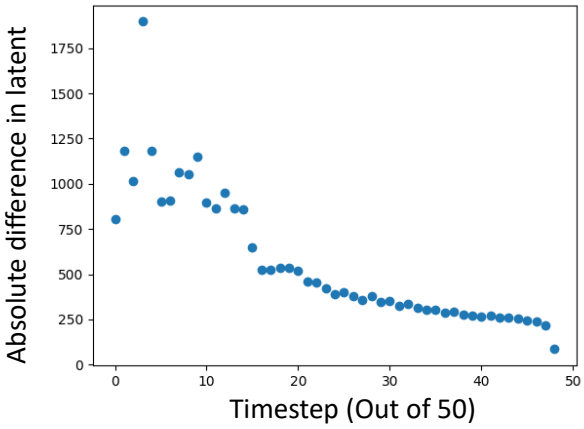

Figure 0.12: Graph showing the magnitude of change in the latent after each timestep. Early steps are much more concerned with the semantic guidance from the prompt as opposed to later timesteps.

A few insights can be gained from doing the removal. Further reinforcing that early timestep's primary role is to decide the overall layout, Figure 0.13 demonstrates that removing using the first q1 set of timesteps, the model still produces a reasonable image for the caption, just with a wildly different layout. Here its probably relying on the random structure provided by the random initial latent, but still creating an image that captures the semantic content of the prompt. Also, you may notice this more randomly chosen

layout looks less "believable" than control image, where the control image has depth and perspective where as the random image does not. Again, if the initial layers mainly perform composition, relying on the random input latent is going to give you a composition not based on anything. Removing the q3, and q4 thus removes the ability for the model to de-blur the image, reinforcing that the later layers sharpen the image and add high frequency details.

Figure 0.13: Here each row represents the final generated images for a single prompt and the same seed, given some intervention. Each column denotes the effects of severing the ability for cross-attention modules to edit the latent at specific buckets of timesteps in the generation process, out of one hundred total steps. The buckets are as follows 0-10 (q1), 10-30 (q2), 30-60 (q3), and 60-100 (q4). The last column showing the normal, un-intervened result.

Figure 0.14: If we increase the size of q1 to 0-26(q1.2) we can see that while it’s a viable image, it less captures the input prompt. Increasing the size of the early removal buckets further highlights the large difference in layout from the control image, especially highlighting it’s inability to match the color adjective with its subject.

We also do the reverse where we remove the effect at all timesteps except those of a certain range, call them o1, o2, o3, and o4. In Figure 0.15(o1), the model has generally decided on a color palette, and uses that palette to outline where some things will go. For example in the ‘woman’ prompt, its told future steps "I want the woman in the red dress here and the the jungle there" which is expressed through red and green. If you squint hard enough at ‘car’ | o1, you can almost make out a vehicle, although at that point it seems the car would be white and the background blue, opposite of how the normal image turns out. In fact, the model certainly "changes it’s mind" at points during the generation process. Look at ‘car’ | q2 and it looks like it’s in a superposition where the car is half coming toward the viewer, and half side to side.

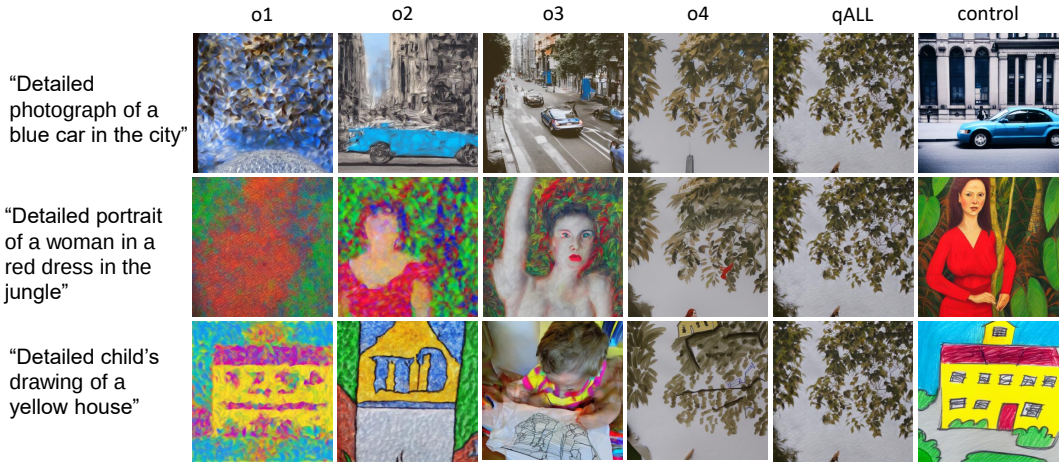

Figure 0.15: Here each row represents the final generated images for a single prompt and the same seed, given some intervention. Each o column denotes the effects of severing the ability for cross-attention modules to edit the latent at all EXCEPT specific buckets of timesteps in the generation process, out of one hundred total steps. The buckets are as follows 0-10 (o1), 10-30 (o2), 30-60 (o3), and 60-100 (o4). The last column showing the normal, un-intervened result and qALL showing the effect of removing all timesteps.

A second way we can remove the influence of semantics at certain timesteps is to feed the cross-attention a blank prompt at the same buckets as before. This way the cross-attention still gets to perform some of it's role just without the semantics of the prompt. This too gives us various insights. Consider the q buckets from before as p and the o buckets as r. Here in Figure 0.17 a side by side figure showing 'car' | q3 vs 'car' | p3. While in q3 the ability for the model to de-blur was blocked, in p3 it still can. Consider that at the state in which the model receives the latent in the later steps is very largely set in stone. If one was asked to de-blur the image in later steps, you would not need to tell them its a blue car in the city.

Figure 0.16: In this setup, buckets are the same as those in 0.13, however instead of severing the cross-attentions connection to the residual stream, the input text embedding is replaced with that of a blank ‘’ prompt. Notice a few things. One, in bucket 3 how the model is still able to perform some amount of de-blurring in contrast to the previous setup where we severed the cross attention all together. Also how different the layout of the image is when doing the intervention at early buckets as opposed to late. The images also have less quality and believability of layout due to them relying on the random beginning latent for more time.

Figure 0.17: Comparing the ablation setup (q3) vs the blank prompt setup (p3) demonstrates the role cross-attention has outside of that which relies on the semantic information from the prompt. In this case, a blurring effect is retained in p3, but absent in q3.

This means the cross-attention does more than inject semantics into the image, and plays roles like de-blur which are agnostic of the prompt. We can also plot the sum of the cross-attention probabilities over all layers for only the tokens in the prompt. Again showing that layer timesteps are less focused on the concept in the prompt.

Like before, we can see when we remove the effects of early timesteps, the concepts from the prompt are not as well understood. The early steps dictate to layer steps where objects should generally go by using blobs of colors, if we have no prompt guidance in the early steps, it relies on the random decisions of the model to attempt to create the objects.

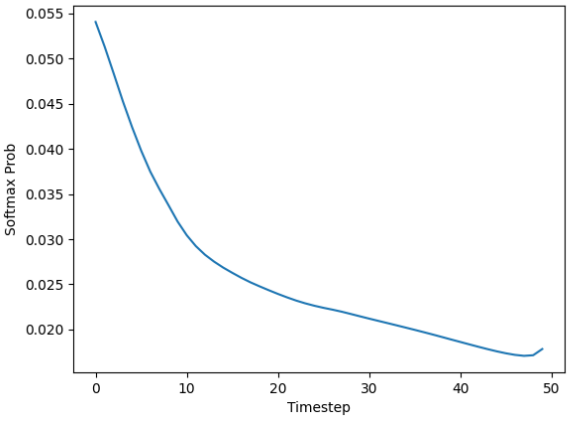

Figure 0.18: Cross-attention probabilities for words in the prompt over timesteps. Averaged over all layers, five prompts and ten seeds. This demonstrates that cross-attention layers pay less attention to, and therefore are less concerned with, the semantic information from the input prompt and more concerned with the current state of the latent.

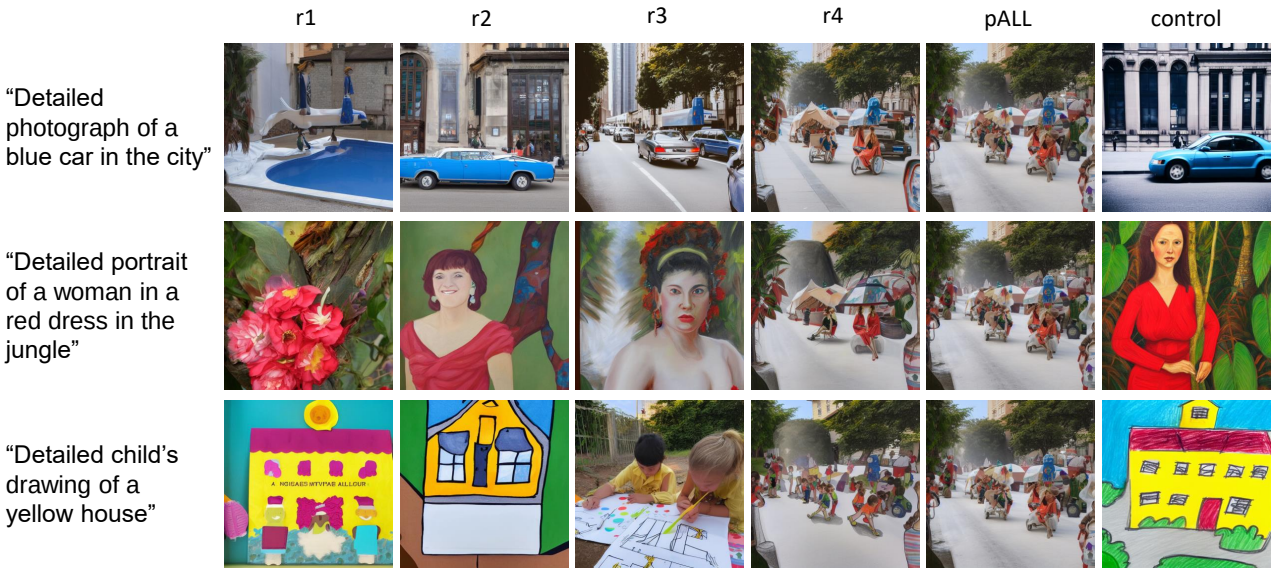

Figure 0.19: Here r designates buckets that are not effected by the prompt swap, just like with the ablation experiment in 0.15. Swapping in the blank prompt as opposed to ablation reinforces that early layers decide composition though use of color. Removing this ability causes the model to generate less believable, less connected to the prompt images just like the ablation experiment.

Layers

In this section we turn to the individual role each cross-attention component has, building upon the theory that the stable diffusion process operates differently over components and processes. In deep learning transformer models , layers and attention heads contribute to different aspects of the generated output, allowing the model to learn and represent various semantic information.

Different layers within the model can be thought of as responsible for capturing different levels of abstraction in the image generation process. As the backbone of Stable Diffusion is a U-net (figure pointing to background) high resolution layers might concentrate on the high frequency aspects like textures, while low resolution layers may delve into more complex relationships and interactions between objects in the scene. Consequently, the model learns to represent and process semantic information at various levels of granularity.

We hypothesize that just like how timesteps play distinctly different roles in the image generation process, separate cross attention layers also perform fundamentally different roles. In order to do this, we perform the same experiments as before, where in one setting we removed the effects of cross-attention by severing it's connection and another where we fed the model the text embeddings of a blank '' prompt, and instead of doing the operation at certain timesteps, we do it at specific layers. There are sixteen separate cross-attention modules, six down, nine up, and one mid. The down layers start at a high resolution and end at a low resolution while up layers start low resolution and end high..

In Figure 0.20, effects of three layers ablated are shown. oUP9 being the highest resolution up layer, oMID being the lowest resolution, and oDOWN5 being a low resolution down layer although higher than oMID. In addition, qHIGHRES is five of the highest resolution layers (DOWN1, DOWN2, UP7, UP8, UP) and qLOWRES is the six lowest resolution layers (DOWN5, DOWN6, MID, UP1, UP2, UP3).

The three singleton layers shown don't have a significant effect on any particular

Figure 0.20: Here each row represents the final generated images for a single prompt and the same seed, given some intervention. Each q column denotes the effects of severing the ability for cross-attention modules to edit the latent at specific cross-attention layers, or buckets of such layers. UP9 denotes the ninth up layer, MID the singular middle layer, and DOWN5 the fifth down layer. HIGHRES and LOWRES are buckets of multiple layers when the latent is at its highest and lowest resolution respectively. Ablating any one layer (qUP9, qMID, and qDOWN5) does not have a drastic effect on the semantics of the resulting image. Layers with the lowest resolution (qLOWRES) generally effect the composition and semantics more than the high resolution layers (qHIGHRES).

interpretable concept. However, the smallest resolution layers (qMID, qDOWN5) do have the largest compositional changes from the control images. This makes sense as being the lowest resolution, each patch of the latent at that point has the largest "view" of the other sections and represents the largest amount of the image, which is important when deciding on the layout of the semantic concepts to generate from the prompt. The inverse is true where when the latent state is in it’s highest resolution (UP9), its responsibilities are much more localized and high frequency. Note this all general observations and it seems having each layer active leads to the most visually pleasing and well constructed images.

When removing whole groups of layers, we can still see that removing the high

resolution ones (qHIGHRES), while certainly changes and degrades the image, doesn't lead to any loss of semantic accuracy in relation to the input prompt. The low resolution group (qLOWRES) however, has completely changed the composition of the image, leaving only vestiges of color and vague connections to semantics.

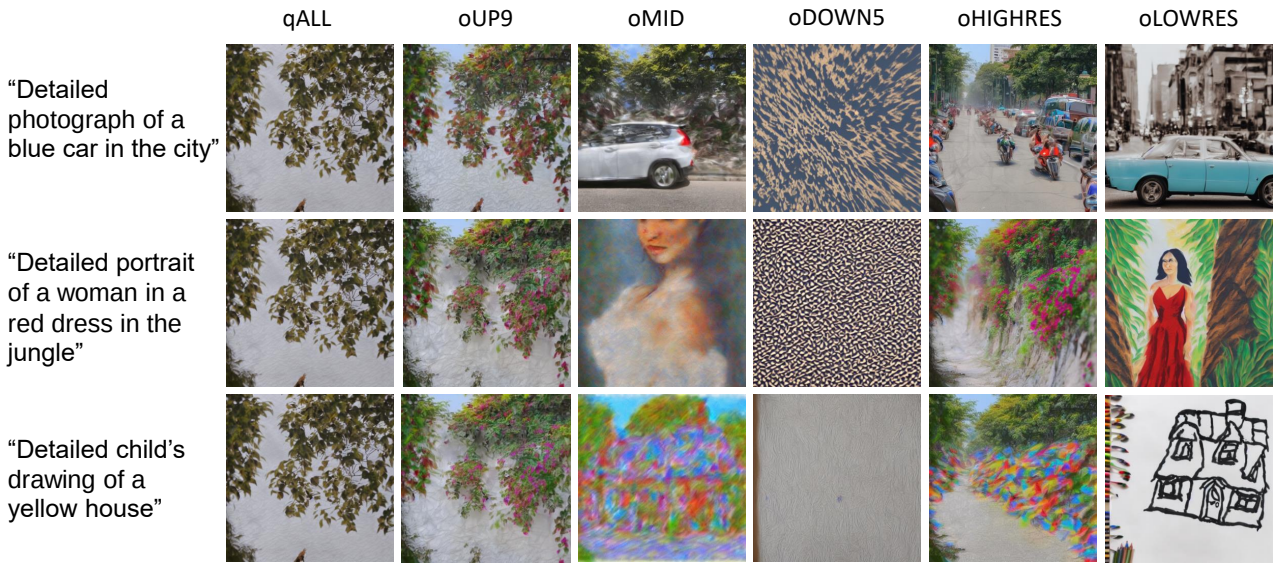

Figure 0.21: Here each row represents the final generated images for a single prompt and the same seed, given some intervention. Each o column denotes the effects of severing the ability for cross-attention modules to edit the latent at all EXCEPT at specific cross-attention layers, or buckets of such layers. qALL shows the effect of removing at all cross-attention layers. Ablating all but the highest resolution layers (oUP9 and oHIGHRES) leaves us with an image closest to the completely ablated image (qALL), while ablating all but the lowest resolution (oMID and oLOWRES) provided images that most match the semantic concept in the prompt. However note for the low resolution layers, the lack of clarity and color specifically requested in the input prompt. A middle resolution layer (oDOWN5) is harder to characterize, but suggests a role in the overall style of the image i.e photograph portrait, or drawing.

Figure 0.21 displays the effects of applying the ablation to all but one layer or group. Here we start to see some jarring and abstract results. qALL shows what the resulting image looks like with all of the layers removed. Like how qUP9 is not much different than the control image, here oUP9 is not very different than qALL demonstrating how it has a relatively small effect on the semantic composition of the final image. Interestingly, restoring the mid layer (oMID) goes quite far in restoring the semantic content of the resulting images. The 'car' prompt looks like a realistic

photo of a car and the ‘woman’ prompt provides a portrait-like image of a woman. If you squint hard enough, the ‘house’ prompt takes the shape of a house as well. Do note that the first two prompts are distinctly without their color. There is a lot of evidence to characterize the mid, lowest resolution layer, is crucial for generating the most important semantic concepts and objects involved in the prompt. Conversely, color does not seem like it plays a major role here relying on the ranom colors from the initial latent, even with the ‘house’ | oMID image as you could characterize the collage of colors as random or indecisive in that aspect. oDOWN5 in contrast, has anything but meaningful semantic content. This layer throughout many prompts, generates a single swath of some repeating pattern. A sensible characterization of this layer might be that of texture. ‘woman’ | oDOWN5 has a blotchy, amorphous pattern while the ‘car’ | oDOWN5 has sharp clear edges you might expect from a photograph. The ‘house’ | oDOWN5 image with its wavy black lines may well be related to the prompt being a child’s drawing. To drive the high vs low resolution point home, if we look at oLOWRES and oHIGHRES, and see how much more the low resolution layers matter when it comes to creating an image congruent with the semantic content of the prompt.

Furthermore, attention heads within each layer play a pivotal role in determining how the model focuses on and generates specific content. Different attention heads can specialize in different aspects of the input prompt, enabling the model to disentangle and process the semantic information effectively. For example, one attention head might be responsible for understanding spatial relationships, while another might focus on object recognition or color patterns.

Like with timesteps, we can look to visualise the role of layers by providing the cross-attention module with a blank prompt at specific layers as shown in Figure 0.22. Similar insights are gained in terms of the roles of low resolution vs high resolution layers. We can again see in rMID that the semantic content is largely restored, though with much higher fidelity and less artifacts. If we look at ‘car’ |

Figure 0.22: Here r designates buckets that are not effected by the blank prompt swap, just like with the ablation experiment in 0.21. Notice the same insight as before where allowing just high resolution layers (rUP9 and rHIGHRES) most closely match the completely blank prompt swapped image (qALL) suggesting they don’t matter as much when dealing with the semantic information of the prompt. Images in the rMID column are of much higher quality than 0.21 and yet still without the colors we specified. rDOWN5 gives a realized version of the texture/palette we saw in qDOWN5.

rDOWN5 and ‘house’ | rDOWN5, instead of textures we get flashed out components of the input prompt. Perhaps the textures from before overall signal ‘city’ and ‘drawing’ respectively.

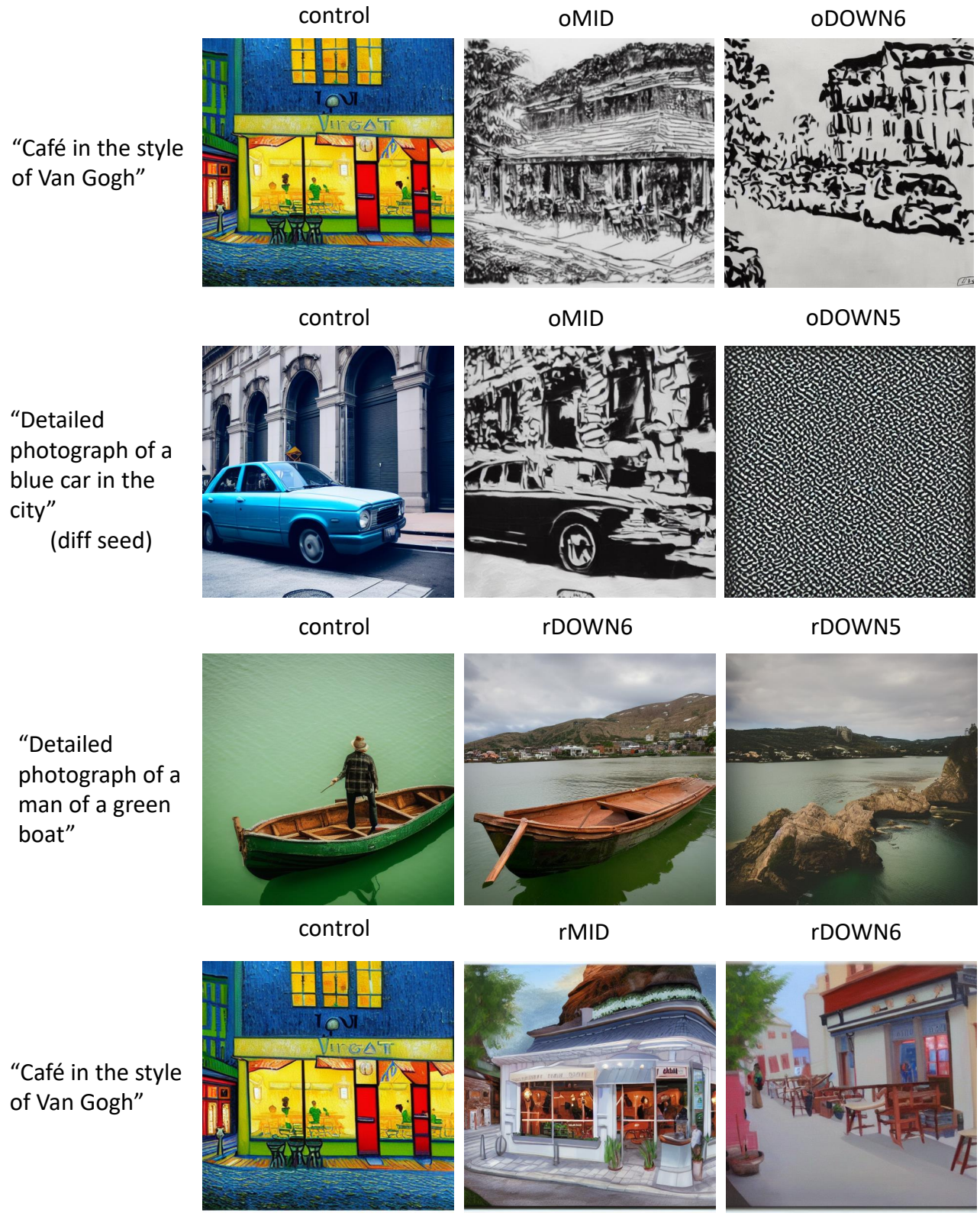

Figure 0.23: Some more interesting examples from a couple of the experiments. Note for the 'cafe' and 'car' prompts (rows 1 and 2), we see when we ablate all but one low resolution layer, we are left with am image containing the semantic content from the prompt, a cafe and a car respectively, but with no color. For the 'boat' prompt (row 3), we perform the prompt swap experiment on a couple layers which result in a nice image, but without the man in one and without the man or the boat in another! Perhaps the semantic information for boat and man are in other layers?

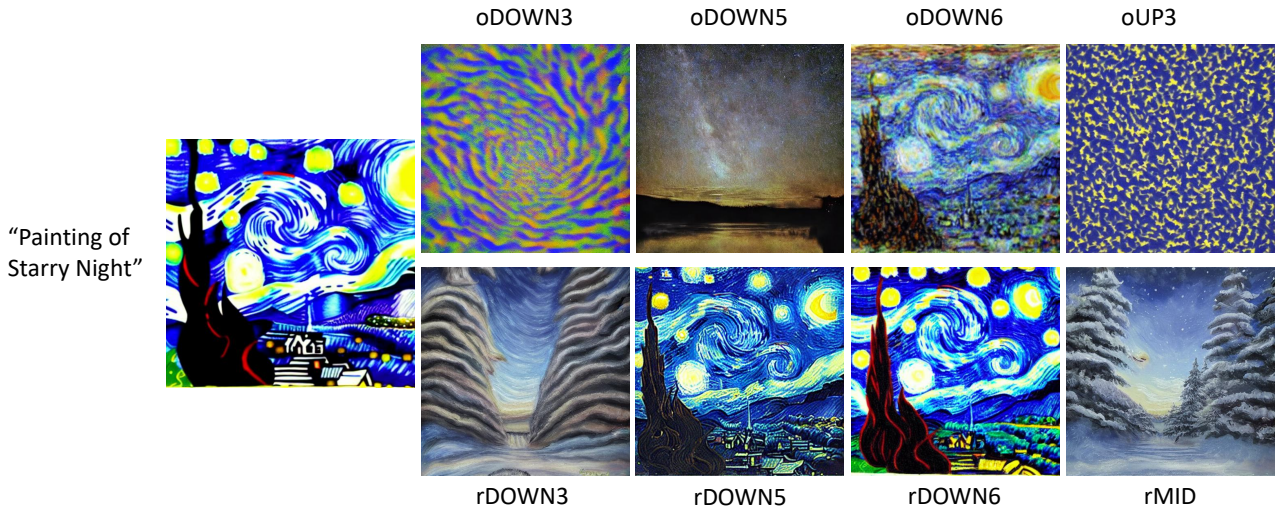

Figure 0.24: Stable Diffusion memorizes some images fairly heavily. in oDOWN5, we get a lovely picture of a starry night, but not the Starry Night we were looking for. It would seem that layer either favors processing just the starry night piece literally, or is not responsible for Van Gogh's Starry Night (or paintings in general for that matter). With oDOWN6, we get a mostly complete albeit blurring version of Starry Night. Is this the layer that has the most of the information for Van Gogh's famous work? We also have oDOWN3 and oUP3 perhaps providing the swirling and color palette this paining is know for respectively.

Another way of measuring what each layer is doing is by creating a heatmap of the magnitude change the cross-attention module is making to the residual stream. In Figure 0.25, we can see these heatmaps averaged over all timesteps (100) for specific layers. For DOWN3, a mid-resolution layer, the heatmaps clearly select for the primary subject in their respective prompts, car, woman, and house. UP6 (another mid-resolution layer) highlights secondary characteristic of the subject, dealing heavily with color. In this case blue for the car (notice how compared to DOWN3, its not highlighting the dark shadow of the car or the dark wheels), the red dress and brown hair of the woman, and the red roof/door, green shrubbery and windowsills of the house. Lastly, UP8 (high-resolution) focuses on edges and fine details like the outlines of the car, woman, and windows. This matches the previous experiments where high-resolution layers were seen to have a large role in high-frequency sharp details.

If you remember from the timestep section, it seemed that in general, early

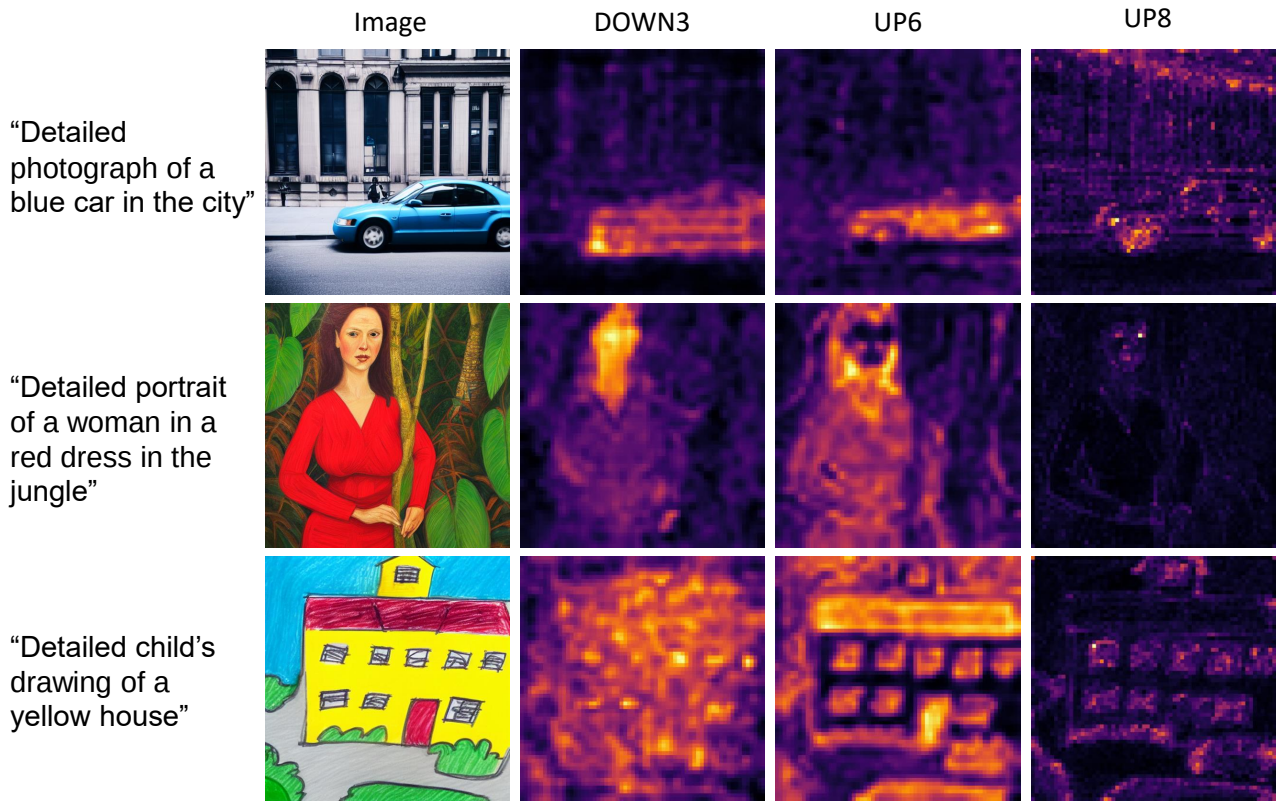

Figure 0.25: The absolute magnitude of change layers make thought the diffusion process.

and mid timesteps were heavily involved in the composition and semantics of the generation process, while late timesteps dealt in fine-details. This parallels the results here where the low resolution layers are responsible for composition and semantics and high resolution ones the fine details. Is it the case then that low resolution layers are most active in the early timesteps and high resolution ones the later ones? It would seem so. In Figure 0.26 , the magnitude of update the cross-attention is adding to the residual stream is plotted against timesteps. Low resolution layers like MID are most active in the first quarter or so of timesteps and heavily drop off. The highest resolution layers like UP9 are by far the most active right at the very end of the process. This is consistent for many prompts and seeds tested. The mid resolution layers like UP6 vary depending on the prompt

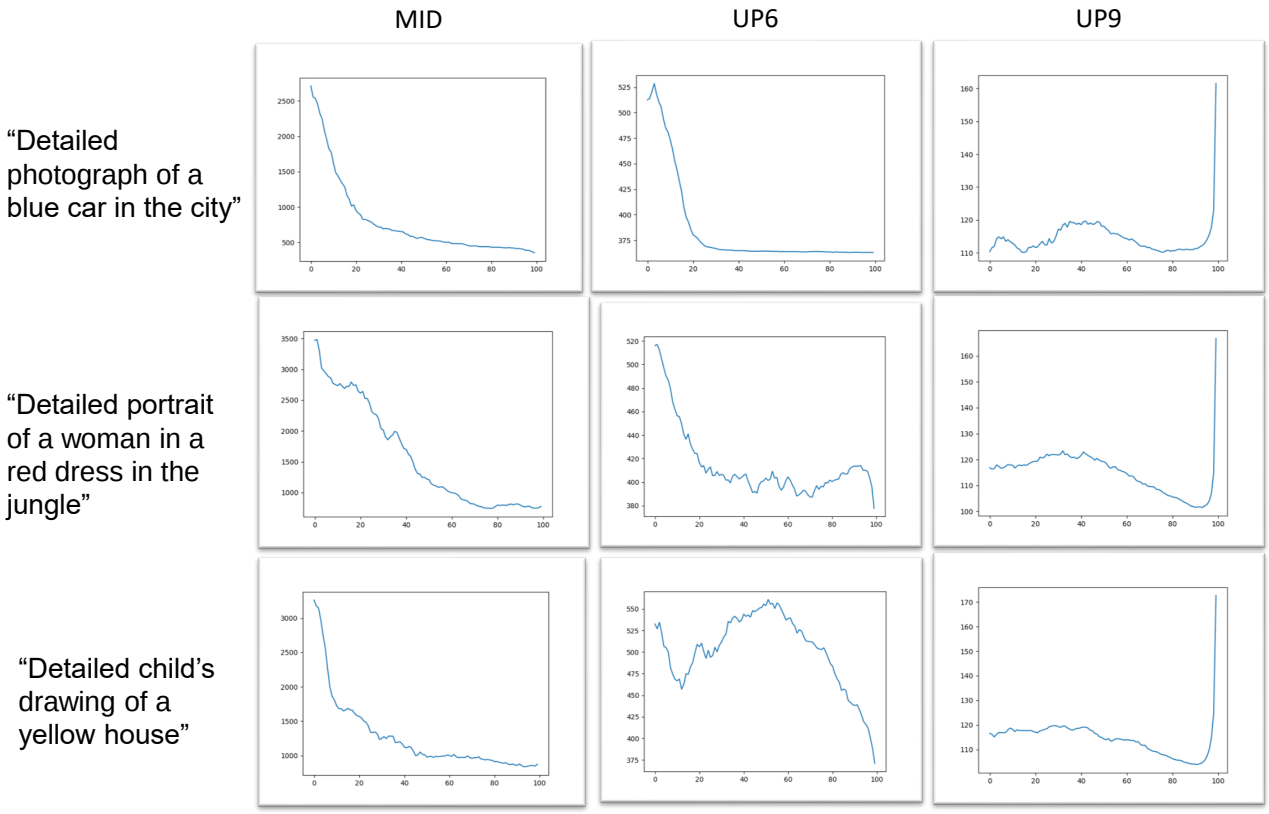

Figure 0.26: The absolute magnitude of change (y-axis) for selected layers over 100 timesteps (x-axis)

Hidden States

The key and value weights in the cross-attention mechanism reveal how the model represents semantic information regardless of the current state of the latent. By examining the hidden states generated by these weights, we can gain insights into the model's internal representation and understanding of the input prompt. These hidden states enable the model to maintain semantic coherence throughout the image generation process, even as it transitions between different timesteps and adjusts the image's content accordingly.

In essence, the dynamic interplay between timesteps, layers, and attention heads in a transformer image generation model highlights the complexity and adaptability of the stable diffusion process. This multi-faceted approach allows the model to generate images with rich semantic content that accurately reflect the input prompt. As a result, understanding the role of each layer and attention head in relation to the timesteps can provide valuable insights into the inner workings of the image generation process and help refine the model's performance even further.

Each cross-attention layer in the model processes and represent the input prompt in a different way. Not only that, but each layer has eight separate heads which in and of themselves have different representations. The dimension size of their representation varies too where the lowest resolution layers have the highest number. To get a window into how some of the heads represent input prompts, we can use PCA to drop the number of dimensions down to two. To make it a fair comparison, thirty thousand prompts from the COCO image captioning dataset are processed, and for each layer and each head we fit PCA on the hidden states of the tokens in the prompt.

Referencing Figure 0.27, the top row shows the PCA components for various categories. The categories are fairly easily separable at many attention heads throughout the network. In the bottom row, the categories are objects of the same color. The head UP4H3 has good separation of the color groupings. MIDH1 on the

other hand doesn't at all which tracks with previous observations that middle, low resolution layers, care less about color than higher resolution layers.

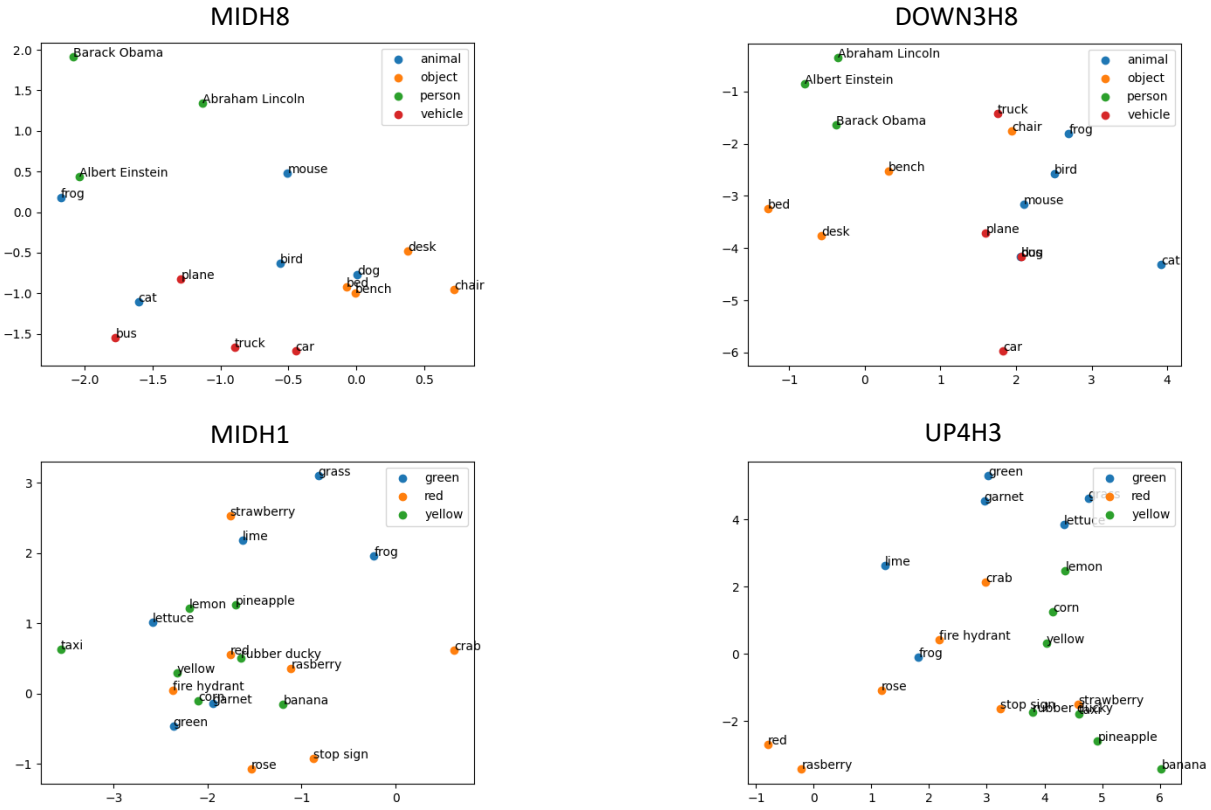

Figure 0.27: This figure shows the hidden states generated by the key matrix for various prompts, reduced to two components by PCA. MIDH8 and MIDH1 being the eighth and first attention head of the middle layer respectively. DOWN3H8 being the eighth head of the third down layer and UP4H3 being the third head of the fourth up layer. Some attention heads like UP4H3, separate objects of the same color very well, while other like MIDH1 do not, matching our previous observation in Figure 0.21 that middle layers are not, or are less, responsible for color from the input prompt. MIDH8 and DOWN3H8 demonstrate how even similar prompts semantically are similar in their hidden states.

0.5 Conclusion

The exploration into erasing concepts from the model has proven informative about how Stable Diffusion represents semantic concepts. For instance, utilizing the negative training objective to remove a concept like "Van Gogh" eliminates not only the generation of Van Gogh himself or his style when prompted with "Van Gogh," but also prevents his style from being generated in any context. This is evident when prompts like "Starry Night" produce a simple image of the night sky instead of Van Gogh's iconic painting, suggesting that both prompts reference the same or similar representation within the model, and it is this representation that is removed.

Achieving successful effects when removing a concept like 'car,' however, necessitates applying the method to the self-attention rather than the cross-attention layers. There may be several reasons for this. One possibility is that the concept of a car is so prevalent in the model's training data that it has multiple disparate representations within the model. As a result, removing the concept when prompted with 'car' does not affect the cars generated with a prompt like 'auto show,' which could reference a different internal representation. Another possibility is that, due to its commonality in the training data, the model often hallucinates cars even without a specific prompt, and these cars might appear in the self-attention or convolutional layers.

Experiments manipulating the cross-attention layers reveal that not all layers and timesteps are created equal. Restricting the method to subsets of these aspects might lead to less interference with other concepts and a greater degree of manipulation. Early timesteps and low-resolution layers, in particular, have the most significant impact on the semantic information injected into the final generated image. Additionally, it is apparent that some concepts have similar internal representations as related concepts.

Insights from this investigation leaves many open doors for future research

directions. Given the nature of visual generation models, quantifying the effects of specific interventions on the model is challenging, often relying on qualitative analysis. Implementing methods such as measuring actual frequency values or using a semantic classification model could enable some of these experiments to be scaled up and measured quantitatively. Components like self-attention and convolutional layers also contribute to the model's performance and warrant further experimentation and analysis. One of the challenges in isolating the roles of these components is the unique time dimension of diffusion models. Effects made to one layer can cascade into other layers as timesteps increase. Additionally, the timestep embedding that the model receives always coincides with the latent being de-noised to a certain degree, complicating the understanding of whether the model's behavior at specific timesteps is due to the timestep embedding or the state of the latent. Further experiments manipulating the timestep embedding itself may help clarify this relationship.

In summary, the erasing work in section 4.1 demonstrated that concepts in Stable Diffusion can be removed due to the reversible nature of the training objective and the similarity of representation for related concepts. Seciton 4.2 also highlights that some timesteps and layers can be characterized by their contributions to the generation process and the absence of certain elements when removed. Through our experiments, we have observed that it is possible to not only characterize processes and representations for semantic content Stable Diffusion, but to edit them in a way to remove certain semantic concepts all together in a robust and minimally interfering way. This leads us to conclude that interpretabily for diffusion model is a promising research direction, as ultimately increased research into the mechanisms behind diffusion models will pave the way for the development of more useful tools that align with our desires and needs.

BIBLIOGRAPHY

[1] Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S. Ryoo. *Peekaboo: Text to Image Diffusion Models are Zero-Shot Segmentors*. 2022. eprint: arXiv:2211.13224.

[2] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. *Quantifying Memorization Across Neural Language Models*. 2022. eprint: arXiv:2202.07646.

[3] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. *Erasing Concepts from Diffusion Models*. 2023. eprint: arXiv:2303.07345.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. eprint: arXiv:2006.11239.

[5] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. *Compositional Visual Generation with Composable Diffusion Models*. 2022. eprint: arXiv:2206.01714.

[6] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. 2021. eprint: arXiv:2112.10741.

[7] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. *Red-Teaming the Stable Diffusion Safety Filter*. 2022. eprint: arXiv:2210.04610.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. eprint: arXiv:2112.10752.

[9] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. *Raising the Cost of Malicious AI-Powered Image Editing*. 2023. eprint: arXiv:2302.06588.

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. 2020. eprint: arXiv:2010.02502.

[11] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. *What the DAAM: Interpreting Stable Diffusion Using Cross Attention*. 2022. eprint: arXiv:2210.04885.