

# CASTEER: CROSS-ATTENTION STEERING FOR CONTROLLABLE CONCEPT ERASURE

Tatiana Gaintseva<sup>1,2</sup>, Andreea-Maria Oncescu<sup>2</sup>, Chengcheng Ma<sup>3</sup>, Ziquan Liu<sup>1</sup>, Martin Benning<sup>4</sup>, Gregory Slabaugh<sup>1</sup>, Jiankang Deng<sup>2,5</sup>, Ismail Elezi<sup>2</sup>

<sup>1</sup>Queen Mary University of London    <sup>2</sup>Huawei Noah’s Ark    <sup>3</sup>CASIA  
<sup>4</sup>University College London    <sup>5</sup>Imperial College London

## ABSTRACT

Diffusion models have transformed image generation, yet controlling their outputs to reliably erase undesired concepts remains challenging. Existing approaches usually require task-specific training and struggle to generalize across both concrete (e.g., objects) and abstract (e.g., styles) concepts. We propose CASteer (**Cross-Attention Steering**), a training-free framework for concept erasure in diffusion models using steering vectors to influence hidden representations dynamically. CASteer precomputes concept-specific steering vectors by averaging neural activations from images generated for each target concept. During inference, it dynamically applies these vectors to suppress undesired concepts only when they appear, ensuring that unrelated regions remain unaffected. This selective activation enables precise, context-aware erasure without degrading overall image quality. This approach achieves effective removal of harmful or unwanted content across a wide range of visual concepts, all without model retraining. CASteer outperforms state-of-the-art concept erasure techniques while preserving unrelated content and minimizing unintended effects. Code is available at <https://github.com/Atmyre/CASteer>.

## 1 INTRODUCTION

Recent advances in diffusion models Ho et al. (2020); Rombach et al. (2022) have revolutionized image Podell et al. (2024) and video generation Girdhar et al. (2024), achieving unprecedented realism. These models operate by gradually adding noise to data during a forward process and then learning to reverse this noise through a series of iterative steps, reconstructing the original data from randomness. By leveraging this denoising process, diffusion models generate high-quality, realistic outputs, making them a powerful tool for creative and generative tasks.

However, the same capabilities that make diffusion models transformative also raise profound ethical and practical concerns. The ability to generate hyper-realistic content amplifies societal vulnerabilities. Risks range from deepfakes and misinformation to subtler effects such as erosion of trust in digital media and targeted manipulation. Addressing these challenges requires not only reactive safeguards (e.g., blocking explicit content) but proactive methods to constrain or remove harmful concepts at the level of the model itself. Current approaches to moderation often treat symptoms rather than causes, limiting their adaptability as risks and applications evolve.

Existing methods for concept erasure in diffusion models remain narrow in scope. LoRA-based fine-tuning Hu et al. (2022) is effective for removing specific objects or styles but struggles with abstract or composite concepts (e.g., nudity, violence, or ideological symbolism), and scales poorly when multiple concepts must be removed, requiring separate adapters or costly retraining. Prompt-based interventions Yoon et al. (2024) offer greater flexibility for abstract harm reduction but lack precision in suppressing concrete attributes, often failing to generalize across concept variations. As a result, existing strategies fall short of delivering reliable, efficient, and broad-spectrum concept erasure.

In this work, we introduce CASteer, a training-free method for controllable concept erasure that leverages the principle of *steering* to influence hidden representations of diffusion models dynamically. Our method builds on recent findings that deep neural networks encode features into approximately

linear subspaces Elhage et al. (2021); Wu et al. (2023). Prior research has shown that intermediate subspaces of diffusion backbones also exhibit this property, with directions that modulate the strength of particular features Kwon et al. (2023); Park et al. (2023); Si et al. (2024); Tumanyan et al. (2023); Li et al. (2024). Yet, these techniques remain limited in scope, often restricted to specific subspaces, requiring training, or offering only coarse control.

Our approach departs from this paradigm. We show that *multiple* subspaces within diffusion models exhibit linear properties that can be harnessed for precise concept erasure. For each concept of interest, we generate  $k$  *positive* images (where  $k \geq 1$ ) containing the concept and  $k$  *negative* images not containing it, and compute the steering vectors by subtracting the averaged hidden representations of the network across *negative* images from those of *positive* ones. During inference, these precomputed vectors are applied directly to the model activations, allowing us to selectively suppress undesirable concepts without retraining or degrading the overall image quality. Experiments demonstrate that Casteer achieves fine-grained erasure of harmful or unwanted concepts (e.g., nudity, violence), while maintaining robustness across a wide range of diffusion models, including SD 1.4, SDXL Podell et al. (2024), Sana Xie et al. (2025), and their distilled variants (e.g., SDXL-Turbo Sauer et al. (2022), Sana-Sprint Chen et al. (2025)).

In summary, our **contributions** are the following:

- We **propose** a novel training-free framework for controllable concept erasure in diffusion models, leveraging steering vectors to suppress unwanted image features without retraining.
- We **demonstrate** that Casteer effectively handles both concrete (e.g., specific characters) and abstract (e.g., nudity, violence) concepts, and scales to multiple simultaneous erasures.
- We **achieve** state-of-the-art performance in concept erasure across diverse tasks and diffusion backbones, validating the robustness, versatility, and practicality of our approach.

## 2 RELATED WORK

**Data-driven AI Safety.** Ensuring the safety of image and text-to-image generative models hinges on preventing the generation of harmful or unwanted content. Common approaches include curating training data with licensed material Rao (2023); Schuhmann et al. (2022), fine-tuning models to suppress harmful outputs Rombach et al. (2022); Shi et al. (2020), or deploying post-hoc content detectors Bedapudi (2022); Rando et al. (2022). While promising, these strategies face critical limitations: data filtering introduces inherent biases Shi et al. (2020), detectors are computationally efficient but often inaccurate or easily bypassed Gandikota et al. (2023); SmithMano (2022), and model retraining becomes costly when new harmful concepts emerge. Alternative methods leverage text-domain interventions, such as prompt engineering Shi et al. (2020) or negative prompts Miyake et al. (2023); Schramowski et al. (2023). Yet these remain vulnerable to adversarial attacks, lack flexibility and precision as they operate in the discrete space of tokens, and often fail to address the disconnect between text prompts and visual outputs—models can still generate undesired content even when text guidance is “safe.” Our approach instead operates in the joint image-text latent space of diffusion models, enabling more robust and granular control over generated content without relying solely on textual constraints.

**Model-driven AI Safety.** Current methods Gandikota et al. (2023); Kumari et al. (2023); Heng & Soh (2023); Zhang et al. (2024a); Huang et al. (2024); Lee et al. (2025) erase unwanted concepts by fine-tuning or otherwise optimising models and adapters to shift probability distributions toward null or surrogate tokens, often combined with regularization or generative replay Shin et al. (2017). Other methods, such as Gandikota et al. (2024); Gong et al. (2024), use direct weight editing to remove unwanted concepts. Although effective, these approaches lack precision, inadvertently altering or removing unrelated concepts. Advanced techniques like SPM Lyu et al. (2024) and MACE Lu et al. (2024) improve specificity through LoRA adapters Hu et al. (2022), transport mechanisms, or prompt-guided projections to preserve model integrity. However, while promising for concrete concepts (e.g., Mickey Mouse), they still struggle with abstract concepts (e.g., nudity) and require parameter updates. Another group of methods focuses on interventions into internal mechanisms of generative models. Methods like Prompt-to-Prompt Hertz et al. (2023) enable fine-grained control over text-specified concepts (e.g., amplifying or replacing elements) through interventions to cross-attention maps, yet fail to fully suppress undesired content, particularly when concepts are

implicit or absent from prompts. This task-specific specialization limits their utility for safety-critical erasure, where complete removal is required. CASteer bridges this gap, enabling precise, universal concept suppression without relying on textual priors or compromising unrelated model capabilities. Another area of research focuses on removing information about undesired concepts from text embeddings that generative models are conditioned on Yoon et al. (2024); Zhang et al. (2024b); Qiu et al. (2024). However, as these methods operate on a discrete space of token embeddings, their trade-off between the effectiveness of erasure and the preservation of other features is limited. Zhang et al. (2024c) proposes using adversarial training for concept unlearning; however, training this method is computationally intensive. In contrast, CASteer eliminates training entirely, enabling direct, non-invasive concept suppression in the model’s latent space without collateral damage to unrelated features.

**Utilizing directions in latent spaces.** This area of research focuses on finding interpretable directions in various intermediate spaces of diffusion models Kwon et al. (2023); Park et al. (2023); Si et al. (2024); Tumanyan et al. (2023), which can then be used to control the semantics of generated images. Based on this idea, SDID Li et al. (2024) recently proposed to learn a vector for each given concept, which is then added to the intermediate activation of a bottleneck layer of the diffusion model during inference to provoke the presence of this concept in the generated image. However, this method is highly architecture-specific and fails to deliver precise control over attributes. In our work, we propose a training-free method for constructing interpretable directions in intermediate activation spaces of various diffusion models for more precise control of image generation. SAeUron Cywiński & Deja (2025) utilizes Sparse Autoencoders Olshausen & Field (1997) (SAEs) to find interpretable directions in the activation space of the diffusion model. However, SAEs are unstable, require extensive training, and do not provide initial control over the set of attributes that can be erased. In contrast, CASteer does not require training and provides direct control over the manipulated attributes.

### 3 METHODOLOGY

The main operating principle of CASteer is to modify outputs of certain intermediate layers during inference in order to affect the semantics of generated images, thus preventing the generation of a desired concept. These outputs are modified using specially designed *steering vectors*. In this section, we begin by justifying the choice of the intermediate layers that CASteer modifies (Sec. 3.1), then proceed with the procedure of construction of steering vectors (Sec. 3.2), and after that describe how these steering vectors are used during inference to control the generation process (Sec. 3.3). Finally, we elaborate on practical aspects regarding the calculation and use of the steering vectors (Sec. 3.4).

#### 3.1 CHOICE OF LAYERS TO STEER

Most modern diffusion models use U-Net or Diffusion Transformers (DiT) Peebles & Xie (2023) as a backbone. They consist of a set of Transformer blocks, each having three main components: cross-attention (CA) layer, self-attention (SA) layer, and MLP layer, all of which contribute to the residual stream of the model. Among those, CA layers are the only place in the model where information from the text prompt goes into the model, guiding text-to-image generation. For every image patch and prompt embedding, each CA layer generates a vector matching the size of the image patch embedding. After summation, these vectors transmit text-prompt information to corresponding image regions Hertz et al. (2023).

As the semantics of the resulting image is mostly determined by the text prompt, we modify the outputs of the CA layers during inference, which results in effective, yet precise, control over the features of the generated image. Thus, CASteer constructs steering vectors for the outputs of every CA layer in the model. In the supplementary, we present experiments on applying CASteer to steer outputs of other layers (SA, MLP, and outputs of intermediate layers inside CA blocks).

#### 3.2 CONSTRUCTION OF STEERING VECTORS

We propose to construct steering vectors for each concept we aim to manipulate. These vectors correspond to the cross-attention (CA) outputs we modify. Each steering vector matches the size of the CA outputs and encodes the desired concept’s information. For preventing the concept from

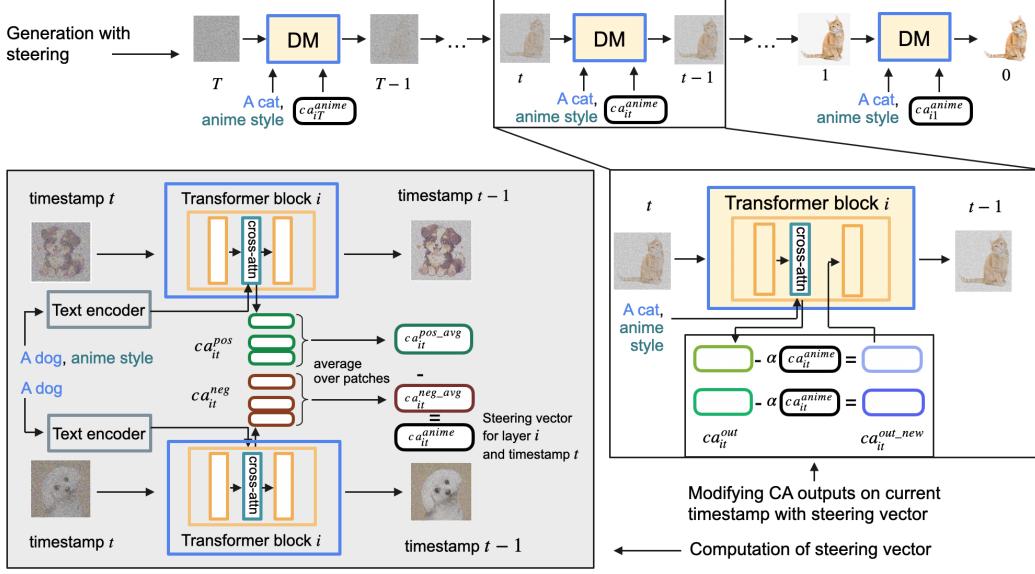


Figure 1: Main pipeline. (Bottom left, gray background) For computing a steering vector, we prompt diffusion model with two prompts that differ in a desired concept, e.g., “anime style” and save CA outputs at each timestamp  $t$  and each CA layer  $i$ . We average these outputs over image patches and get averaged CA outputs  $ca_{it}^{pos\_avg}$  and  $ca_{it}^{neg\_avg}$  for each  $t$  and  $i$ . We subtract the latter from the former, getting a steering vector for the layer  $i$  and timestamp  $t$   $ca_{it}^{anime}$ . (Right) For deleting concept  $X$  from generation, at each denoising step  $t$ , we subtract steering vector  $ca_{it}^X$  multiplied by intensity  $\alpha$  from the CA outputs of the layer  $i$ .

being present in the generated image, we subtract steering vectors of an unwanted concept from cross-attention outputs during generation.

We construct steering vectors as follows. Given a concept  $X$  to manipulate, we create paired positive and negative prompts differing only by the inclusion of  $X$ . For example, if  $X$  = “baroque style”, example prompts are  $p_{pos}$  = “A picture of a man, baroque style” and  $p_{neg}$  = “A picture of a man”. Assume a DiT backbone has  $N$  Transformer blocks, each containing one CA layer, totaling  $N$  CA layers. We generate images from both prompts, saving outputs from each of the  $N$  cross-attention layers across all  $T$  denoising steps. This yields  $NT$  cross-attention output pairs  $\langle ca_{it}^{pos}, ca_{it}^{neg} \rangle$  for  $1 \leq i \leq N$  and  $1 \leq t \leq T$ , where  $i$  denotes the layer and  $t$  is the denoising step. Each  $ca_{it}^{pos}$  and  $ca_{it}^{neg}$  has dimensions  $patch\_num_i \times emb\_size_i$ , corresponding to the number of patches and embedding size at layer  $i$ . We average  $ca_{it}^{pos}$  and  $ca_{it}^{neg}$  over image patches to obtain averaged cross-attention outputs:

$$ca_{it}^{pos\_avg} = \frac{\sum_{k=1}^{patch\_num_i} ca_{itk}^{pos}}{patch\_num_i}; ca_{it}^{neg\_avg} = \frac{\sum_{k=1}^{patch\_num_i} ca_{itk}^{neg}}{patch\_num_i} \quad (1)$$

where  $ca_{it}^{pos\_avg}$  and  $ca_{it}^{neg\_avg}$  are vectors of size  $emb\_size_i$ . Then, for each of these  $N$  layers and each of  $T$  denoising steps, we construct a corresponding steering vector carrying a notion of  $X$  by subtracting its averaged cross-attention output that corresponds to the negative prompt from that corresponding to the positive one and:

$$ca_{it}^X = f_{norm}(ca_{it}^{pos\_avg} - ca_{it}^{neg\_avg}). \quad (2)$$

where  $f_{norm}$  is an  $L_2$ -normalization function:  $f_{norm}(v) = \frac{v}{\|v\|_2}$ .

### 3.3 USING STEERING VECTORS TO CONTROL GENERATION

Computed steering vectors can be seen as directions in a space of intermediate representations of a model (in the space of CA activations) that represent a notion of  $X$ . Thus, we should be able to control the expressiveness of certain feature  $X$  by steering the model representations along the steering vector representing  $X$ . That is, we can prevent a concept from appearing on the generated

image by subtracting some amount of steering vectors for that concept from corresponding CA outputs of a model during inference:

$$ca_{itk}^{out\_new} = ca_{itk}^{out} - \alpha ca_{itk}^X, \quad (3)$$

Here  $1 \leq k \leq \text{patch\_num}_i$ , and  $\alpha$  is a hyperparameter that controls the strength of concept suppression. Larger values of  $\alpha$  lead to higher suppression of the concept  $X$  in the generated image. Below we propose a way to adjust  $\alpha$  dynamically based on activations of diffusion model during generation, achieving effective and precise erasure of unwanted concepts in the resulting image.

**Choice of alpha.** Most often when we aim to suppress the concept of  $X$ , our goal is to completely prevent it from appearing on any generated image given any input prompt. This is the case of such tasks as nudity/violence removal or privacy, when we do not want the model to ever generate somebody's face or artwork. However, there might be different magnitudes for concept  $X$  in the original text prompt (e.g., prompts "an angry man" or "a furious man" express different levels of anger). A concept  $X$  can have different magnitudes of expression in different patches of the image being generated. Consequently, if we use Eq. 3 for suppression, different values of  $\alpha$  are needed to completely suppress  $X$  for different prompts and individual image patches while not affecting other features in the image.

We propose to estimate  $\alpha$  for concept deletion by using the dot product between  $ca_{it}^X$  and corresponding CA output  $ca_{itk}^{out}$  ( $\langle ca_{it}^X, ca_{itk}^{out} \rangle$ ) as an assessment of amount of  $X$  that is present in the image part corresponding to  $k^{th}$  patch of  $ca_{it}^{out}$ . As  $ca_{it}^X$  is normalized, the value of this dot product is the length of the projection of the CA output  $ca_{itk}^{out}$  onto the steering vector  $ca_{it}^X$ . As  $ca_{it}^X$  can be seen as a direction in a linear subspace corresponding to the concept  $X$ , the length of the projection can be seen as the amount of  $X$  that is present in  $ca_{itk}^{out}$ . That said, for removing information about  $X$  from  $ca_{itk}^{out}$ , we propose to subtract the amount of  $ca_{it}^X$  proportionate to the dot product between  $ca_{it}^X$  and  $ca_{itk}^{out}$  from  $ca_{itk}^{out\_new}$ , i.e., define  $\alpha = \beta(\langle ca_{it}^X, ca_{itk}^{out} \rangle)$ . Consequently, Eq. 3 becomes the following:

$$ca_{itk}^{out\_new} = ca_{itk}^{out} - \beta \langle ca_{it}^X, ca_{itk}^{out} \rangle ca_{it}^X. \quad (4)$$

Here  $1 \leq k \leq \text{patch\_num}_i$ , and  $\beta$  is a hyperparameter that controls the strength of the suppression.

Note that Eq. 4 can be reformulated in a matrix form as a projection operator onto the subspace orthogonal to steering vector  $s = ca_{it}^X$ :

$$s^{new} = f_{\text{delete}}(c, s) = (I - ss^T)c \quad (5)$$

Here  $s^{new} = ca_{itk}^{out\_new}$ ,  $c = ca_{itk}^{out}$ ,  $s = ca_{it}^X$ ,  $I$  is an identity matrix.

**Intermediate clipping.** We now introduce a mechanism of clipping the value of  $\alpha$  to get better control over concept suppression. Note that using Eq. 4 we only want to influence those CA outputs  $ca_{itk}^{out}$  which have a positive amount of unwanted concept  $X$  in them. As dot product  $\langle ca_{it}^X, ca_{itk}^{out} \rangle$  measures the amount of  $X$  present in CA output  $ca_{itk}^{out}$ , we only want to steer those CA outputs  $ca_{itk}^{out}$ , which have a positive dot product with  $ca_{it}^X$ . So the equation becomes the following:

$$\begin{aligned} \alpha &= \max(\beta \langle ca_{it}^X, ca_{itk}^{out} \rangle, 0) \\ ca_{itk}^{out\_new} &= ca_{itk}^{out} - \alpha ca_{it}^X. \end{aligned} \quad (6)$$

Note that if intermediate clipping is used, Eq. 6 can no longer be formulated in a matrix form. In the experiments section, we present results of applying CASteer for concept erasure both with and without intermediate clipping (i.e. using Eq. 4 and Eq. 6).

### 3.4 PRACTICAL CONSIDERATIONS

**Multiple Prompts for Steering Vector.** We described in the previous section how to construct and use steering vectors to alter one concept, based on one pair of prompts, e.g., "a picture of a man" and "a picture of a man, baroque style". As mentioned, a steering vector can be seen as the direction in the space of intermediate representations of a model that points from an area of embeddings not containing a concept  $X$ , to an area that contains it. In order for this direction to be more precise,

we propose to construct steering vectors based on multiple pairs of prompts instead of one. More precisely, we obtain  $P \geq 1$  pairs of  $ca_{itp}^{pos\_avg}$  and  $ca_{itp}^{neg\_avg}$ ,  $1 \leq p \leq P$ , then average them over P:

$$ca_{it}^{pos\_avg} = \frac{\sum_{p=1}^P ca_{itp}^{pos\_avg}}{P}, ca_{it}^{neg\_avg} = \frac{\sum_{p=1}^P ca_{itp}^{neg\_avg}}{P} \quad (7)$$

and obtain steering vectors as  $ca_{it}^X = ca_{it}^{pos\_avg} - ca_{it}^{neg\_avg}$ .

**Steering multiple concepts.** It is easy to erase multiple concepts during a generation by either applying steering vectors corresponding to these concepts to the cross-attention output successively or constructing single steering vector corresponding to multiple concepts. In the experiments section, we show results on applying a steering vector constructed for multiple concepts to prevent generation of inappropriate concepts.

**Efficiency: Transferring vectors from distilled models.** Adversarial Diffusion Distillation (ADD) Sauer et al. (2022) is a fine-tuning approach that allows sampling large-scale foundational image diffusion models in 1 to 4 steps, while producing high-quality images, with many methods such as SDXL and Sana having distilled versions (SDXL-Turbo and Sana-Sprint). We observe that steering vectors obtained from the distilled models can successfully be used for steering generations of its corresponding non-distilled variants. More formally, having a pair of prompts, we obtain  $ca_i^{pos\_avg}$  and  $ca_i^{neg\_avg}$  from the distilled model using 1 denoising step. Note that there is no second index  $t$  as we use only one denoising iteration, i.e.  $T = 1$ . We then construct steering vectors for the concept  $X$  as  $ca_i^X = ca_i^{pos\_avg} - ca_i^{neg\_avg}$  and then use it to steer non-Turbo variant of the model by using  $ca_i^X$  for each denoising step  $1 \leq j \leq T$ .

**Injecting CASteer into model weights.** Note that when steering more advanced models (SDXL or Sana), we use steering vectors from Turbo/Sprint model versions, where we have only one steering vector per model CA layer. Also note that the last layer of CA block in SDXL/SANA is Linear layer with no bias and no activation function, i.e., essentially is a matrix multiplication:  $h_{out} = W_{proj\_out} h_{in}$ . Here  $W_{proj\_out}$  is a weight matrix of the last  $proj\_out$  layer of CA block of SDXL/SANA,  $h_{in}$  and  $h_{out}$  are input and output to that layer,  $h_{out}$  being the final output of CA layer. In this case, by combining last layer of CA block with CASteer formulation in a matrix form (Eq. 5), we can incorporate CASteer directly into weights of the model, by multiplying weight matrix of the last layer of CA block with  $I - ss^T$  matrix from Eq. 5:

$$h_{out} = (I - ss^T) W_{proj\_out} h_{in} = W_{proj\_out}^s h_{in} \quad (8)$$

$W_{proj\_out}^s$  is a matrix of the same size as  $W_{proj\_out}$ . This results in having zero inference overhead compares to original SDXL/SANA model similar to LoRA-like tuning approaches.

## 4 EXPERIMENTS

We evaluate the performance of our method on the task of erasing different concepts. We show that our method succeeds in suppressing both abstract (e.g., “nudity”, “violence”) and concrete concepts (e.g., “Snoopy”). Moreover, we demonstrate the advantages of our method in removing implicitly defined concrete concepts (e.g., if a concept is “Mickey”, prompting “a mouse from a Disneyland” should not result in a generation of Mickey).

**Implementation details.** For a fair comparison, we report our main quantitative results using StableDiffusion-v1.4 (SD-v1.4) Rombach et al. (2022). SD-1.4 model does not have a Turbo version, so for these experiments we use per-step steering vectors computed from the original SD-1.4. We apply steering to all of the CA layers in the model. We set  $\beta = 2$  for the concept erasure in all experiments. We use 50 prompt pairs for generating steering vectors for concrete concepts (e.g., “Snoopy”), and 196 prompt pairs for generating steering vectors for concrete concepts (e.g., “nudity”). Information about prompts used for generation of the steering vectors is in the supplementary.

We also show effectiveness of CASteer on bigger models, such as SDXL and SANA, and present results on these models in supplementary. For SDXL Podell et al. (2024) and SANA, we use steering vectors obtained from SDXL-Turbo and SANA-Sprint models, respectively.

**Table 1: Quantitative results on nudity removal based on I2P (Schramowski et al. (2023)) dataset.**  
 Detection of nude body parts is done by Nudenet at a threshold of 0.6. F: Female, M: Male. The best results are highlighted in bold, second-best are underlined.

| Method                             | Nudity Detection |              |           |              |          |          |          |          |           |
|------------------------------------|------------------|--------------|-----------|--------------|----------|----------|----------|----------|-----------|
|                                    | Breast(F)        | Genitalia(F) | Breast(M) | Genitalia(M) | Buttocks | Feet     | Belly    | Armpits  | Total↓    |
| SD v1.4                            | 183              | 21           | 46        | 10           | 44       | 42       | 171      | 129      | 646       |
| Ablating (CA) Kumari et al. (2023) | 298              | 22           | 67        | 7            | 45       | 66       | 180      | 153      | 838       |
| FMN Zhang et al. (2024a)           | 155              | 17           | 19        | 2            | 12       | 59       | 117      | 43       | 424       |
| ESD+Gandikota et al. (2023)        | 101              | 6            | 16        | 10           | 12       | 37       | 77       | 53       | 312       |
| SLD-Med Schramowski et al. (2023)  | 39               | 1            | 26        | 3            | 3        | 21       | 72       | 47       | 212       |
| UCE Gandikota et al. (2024)        | 35               | 5            | 11        | 4            | 7        | 29       | 62       | 29       | 182       |
| SA Heng & Soh (2023)               | 39               | 9            | 4         | <b>0</b>     | 15       | 32       | 49       | 15       | 163       |
| ESD+Gandikota et al. (2023)        | 14               | <u>1</u>     | 8         | 5            | 5        | 24       | 31       | 33       | 121       |
| Receler Huang et al. (2024)        | 13               | <u>1</u>     | 12        | 9            | 5        | 10       | 26       | 39       | 115       |
| MACE Lu et al. (2024)              | 16               | <b>0</b>     | 9         | 7            | 2        | 39       | 19       | 17       | 109       |
| RECE Gong et al. (2024)            | 8                | <b>0</b>     | 6         | 4            | <b>0</b> | 8        | 23       | 17       | 66        |
| CPE (one word) Lee et al. (2025)   | 11               | 2            | 3         | 2            | 5        | 15       | 13       | 15       | 66        |
| CPE (four word) Lee et al. (2025)  | 6                | <u>1</u>     | 3         | 2            | 2        | 8        | 8        | 10       | 40        |
| AdvUnlearn Zhang et al. (2024c)    | <b>1</b>         | <u>1</u>     | <b>0</b>  | <b>0</b>     | <b>0</b> | 13       | <b>0</b> | 8        | 23        |
| SAeUron Cywiński & Deja (2025)     | 4                | <b>0</b>     | <b>0</b>  | <u>1</u>     | 3        | 2        | <u>1</u> | 7        | 18        |
| Ours (w/o clip)                    | 5                | <b>0</b>     | <b>0</b>  | <u>1</u>     | 3        | 2        | <b>0</b> | <u>1</u> | <u>12</u> |
| Ours (clip)                        | <u>4</u>         | <b>0</b>     | <b>0</b>  | <u>1</u>     | <u>2</u> | <b>0</b> | <b>0</b> | <b>0</b> | 7         |

**Table 2: Quantitative results on inappropriate content removal based on I2P (Schramowski et al. (2023)) dataset. Detection of inappropriate content is done by Q16 (Schramowski et al. (2022)) classifier.** The best results are highlighted in bold, second-best are underlined.

| Class name       | Inappropriate proportion (%) (↓) |      |          |       |      |       |      |             |                 |              |
|------------------|----------------------------------|------|----------|-------|------|-------|------|-------------|-----------------|--------------|
|                  | SD                               | FMN  | Ablating | ESD-x | SLD  | ESD-u | UCE  | Receler     | Ours (w/o clip) | Ours (clip)  |
| Hate             | 44.2                             | 37.7 | 40.8     | 34.1  | 22.5 | 26.8  | 36.4 | <b>28.6</b> | 35.5            | <b>29.00</b> |
| Harassment       | 37.5                             | 25.0 | 32.9     | 30.2  | 22.1 | 24.0  | 29.5 | <b>21.7</b> | 29.85           | <b>25.61</b> |
| Violence         | 46.3                             | 47.8 | 43.3     | 40.5  | 31.8 | 35.1  | 34.1 | <b>27.1</b> | 32.54           | <b>27.78</b> |
| Self-harm        | 47.9                             | 46.8 | 47.4     | 36.8  | 30.0 | 33.7  | 30.8 | <b>24.8</b> | <b>26.10</b>    | 26.22        |
| Sexual           | 60.2                             | 59.1 | 60.3     | 40.2  | 52.4 | 35.0  | 25.5 | 29.4        | <b>22.99</b>    | <b>20.73</b> |
| Shocking         | 59.5                             | 58.1 | 57.8     | 45.2  | 40.5 | 40.1  | 41.1 | <b>34.8</b> | 38.43           | <b>34.00</b> |
| Illegal activity | 40.0                             | 37.0 | 37.9     | 28.9  | 22.1 | 26.7  | 29.0 | <b>21.3</b> | 21.46           | <b>17.61</b> |
| Overall          | 48.9                             | 47.8 | 45.9     | 36.6  | 33.7 | 32.8  | 31.3 | <b>27.0</b> | 28.94           | <b>25.58</b> |

#### 4.1 RESULTS

**Abstract concept erasure.** In this section, we present results on inappropriate content erasure based on I2P dataset Schramowski et al. (2023). I2P is a dataset of 4,703 curated prompts designed to test generative models, where most prompts lead to images containing inappropriate content. Following prior work, we test CASteer on two I2P-based tasks: 1) removing nudity, 2) removing all inappropriate content at once. For nudity removal, we utilize CASteer with steering vectors generated for the concept of "nudity". For inappropriate content removal, we use CASteer with steering vectors obtained as average of steering vectors generated for each type of inappropriate content, i.e., hate, harassment, violence, self-harm, shocking, sexual, and illegal content.

We compare our method with state-of-the-art approaches Ablating (CA) Kumari et al. (2023), FMN Zhang et al. (2024a), SLD Schramowski et al. (2023), ESD Gandikota et al. (2023), UCE Gandikota et al. (2024), SA Heng & Soh (2023), Receler Huang et al. (2024), MACE Lu et al. (2024), RECE Gong et al. (2024), CPE Lee et al. (2025), AdvUnlearn Zhang et al. (2024c) and SAeUron Cywiński & Deja (2025). Following prior art, we utilize the NudeNet<sup>1</sup> to detect nude body parts on generated images for nudity erasure task, and the NudeNet with the Q16 detector to detect inappropriate content.

We present the results for CASteer versions with and without intermediate clipping applied in Tab. 1 and Tab. 2. We show that both versions of CASteer outperform all prior models on nudity erasure, with CASteer version with clipping having more than 2 times fewer images with detected nudity than the second-best result. On the inappropriate content removal, CASteer version with clipping also achieves state-of-the-art result, surpassing second-best model Receler by 1.42% overall.

To assess general generation quality of CASteer, we follow prior work and run CASteer with “nudity” steering vectors on prompts from COCO-30k Lin et al. (2014). We report FID Heusel et al. (2017) for general visual quality and CLIP score Hessel et al. (2021) for image-prompt alignment. Results are presented in Tab. 3. Both versions of CASteer have better FID than prior art.

<sup>1</sup><https://github.com/notAI-tech/NudeNet>

Thus, CASteer clearly is capable of deleting unwanted information while maintaining general high quality. Note that these datasets feature adversarial prompts, i.e., the “nudity” concept is encoded in the prompts implicitly.

**Table 3: Evaluation of nudity-erased models.** Robustness is measured with nudity prompts from the I2P dataset, while locality is assessed using COCO-30K prompts.

| Method                 | Locality               |                         |
|------------------------|------------------------|-------------------------|
|                        | CLIP-30K( $\uparrow$ ) | FID-30K( $\downarrow$ ) |
| SD v1.4                | 31.34                  | 14.04                   |
| FMN                    | 30.39                  | 13.52                   |
| CA                     | <b>31.37</b>           | 16.25                   |
| AdvUn                  | 28.14                  | 17.18                   |
| Receler                | 30.49                  | 15.32                   |
| MACE                   | 29.41                  | 13.42                   |
| CPE                    | <b>31.19</b>           | 13.89                   |
| UCE                    | 30.85                  | 14.07                   |
| SLD-M                  | 30.90                  | 16.34                   |
| ESD-x                  | 30.69                  | 14.41                   |
| ESD-u                  | 30.21                  | 15.10                   |
| SAeUron                | 30.89                  | 14.37                   |
| <i>Ours (w/o clip)</i> | 30.69                  | <u>13.28</u>            |
| <i>Ours (clip)</i>     | <u>30.09</u>           | <b>13.02</b>            |

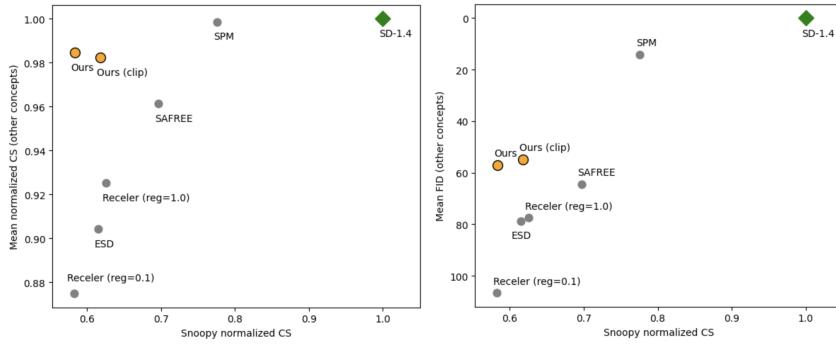


**Figure 2: SPM failure in removing implicitly defined concepts (SD-1.4).** Top: CASteer, Bottom: SPM. Left: “a mouse from Disneyland,” Right: “a yellow Pokemon.” CASteer erases Mickey and Pikachu concepts despite not being explicitly named, while SPM fails.

**Concrete concepts erasure.** To assess ability of CASteer to remove concrete concepts, we follow the experimental setup of SPM Lyu et al. (2024). In this setting, the concept to be erased is *Snoopy*, and images of five additional concepts (*Mickey*, *Spongebob*, *Pikachu*, *dog* and *legislator*) are generated to test the capability of the method to preserve content not related to the concept being removed. The first four of these are specifically chosen to be semantically close to the concept being removed to show the model’s ability to perform precise erasure. Following SPM Lyu et al. (2024), we augment each concept using 80 CLIP Radford et al. (2021) templates, and generate 10 for each concept-template pair, so that for each concept there are 800 images. We evaluate the results using two metrics. First, we utilize CLIP Score (CS) Hessel et al. (2021) to confirm the level of the existence of the concept within the generated content. Second, we calculate FID Heusel et al. (2017) scores between the set of original generations of SD-1.4 model and a set of generations of the steered model. We use it to assess how much images of additional (non-Snoopy) concepts generated by the steered model differ from those of generated by the original model. A higher FID value demonstrates more severe generation alteration. We present the results in Tab. 15. In Fig. 3, we also show two types of plots. Fig. 3a pictures normalized clip score of source concept, i.e. “Snoopy” (the lower the better) versus mean normalized clip scores of other concepts (the higher the better). Normalization is done to ensure equal importance of all the concepts in the mean, and done by dividing clip score of images produced by erasing method by clip score of images produced by vanilla SD-1.4. Methods on the left of the plot erase Snoopy well, and methods on top of the plot preserve other concepts well. Fig. 3b pictures normalized clip score of source concept versus mean FID scores of other concepts (the lower the better). Methods on the left of the plot erase Snoopy well, and methods on top of the plot tend not to affect images of other concepts much.

Results show that CASteer maintains good balance between erasing unwanted concept, while preserving other concepts intact. ESD Gandikota et al. (2023) and Receler erase Snoopy well, but also highly affect other concepts, especially related ones such as *Mickey* or *Spongebob*. Note that their CS of unrelated concepts (e.g. “*Mickey*” or “*Spongebob*”) are significantly lower than that of original SD-1.4, indicating that these concepts are also being affected when erasure of “*Snoopy*” is done. High FID of these methods on these concepts supports this observation. SAFREE shows a reduced level of *Snoopy* erasure compared to that of CASteer, and has lower CS and higher FID on all the other concepts. SPM keeps unrelated concepts almost intact (High CS and low FID), but has a much lower intensity of *Snoopy* erasure. Moreover, SPM fails to erase implicitly defined concepts (see Fig. 2 and Sec. D in supplementary). We provide qualitative results on comparisons in the supplementary.

**CASteer is capable of erasing implicitly defined concepts.** We check what happens if we define the prompts implicitly, e.g., “A mouse from Disneyland”. We run CASteer and SPM trained on the *Mickey* concept on these prompts and show the results in Fig. 2. We clearly see that SPM fails to erase the concepts when they are not explicitly defined. In contrast, our method does a much better job of erasing the concepts, despite being implicitly defined. This is also supported by the results on



(a) Normalized CLIP score on “Snoopy” vs (b) Normalized CLIP score on “Snoopy” vs normalized CLIP scores of other concepts FID scores of other concepts

Figure 3: Comparison of various methods on concrete concept erasure (removing “Snoopy”)



Figure 4: Qualitative results on SDXL (left) and SANA (right) on removing “Snoopy”. Top: original model generations, bottom: generations of model steered to remove “Snoopy”

nudity erasure in Tab. 1, as considered datasets contain specially selected adversarial nudity prompts. We provide additional results on implicitly defined prompts in the supplementary.

**Overall** experimental results show that CASteer performs precise erasure of both concrete and abstract concepts and concepts defined implicitly while leaving other concepts intact and not affecting the overall quality of generated images. More qualitative results showing the performance of CASteer on prompts related and not related to the target concept can be found in the supplementary.

#### 4.2 ABLATION STUDY

**Steering other layers.** As mentioned in Sec. 3.1, we ablate to determine for which type of layer in the DiT backbone steering is most effective. We show in the supplementary that steering the CA outputs is the most effective. We also ablate steering only a fraction of CA layers in Sec. G.4.

**Number of prompt pairs to construct steering vectors.** In the supplementary, we provide an ablation on the number of prompt pairs needed to produce high-quality outputs after steering. We find that as little as 50 prompts is enough for the steering vectors to capture the desired concept well.

**Interpretation of steering vectors.** Here, we propose a way to interpret the meaning of steering vectors generated by CASteer. Suppose we have steering vectors generated for a concept  $X \{ca_{it}^X\}, 1 \leq i \leq l, 1 \leq t \leq T$ , where  $l$  is the number of model layers and  $T$  is the number of denoising steps performed for generating steering vectors. To interpret these vectors, we prompt the diffusion model with a placeholder prompt “X” and at each denoising step, we substitute outputs of the model’s CA layers with corresponding steering vectors. This conditions the diffusion model only on the information from the steering vectors, completely suppressing other information from the text prompt. Results are presented in Fig. 38 and in the supplementary.

**UMap.** We generate steering vectors for all vocabulary tokens of SDXL text encoders and apply UMap McInnes & Healy (2018) on these steering vectors. We present the results in the appendix, showing that structure emerges in the space of these steering vectors, similar to that of Word2Vec Mikolov et al. (2013), supporting that steering vectors carry the meaning of the desired concept.

**Modern models.** We show qualitative results in SANA and SDXL in Fig. 4. We provide more qualitative and quantitative results in SDXL (Sec. E) and SANA (Sec. F).

**User studies.** In Sec. D, we give several user studies, showing that in most cases, the users prefer our results compared to SPM and Receler.

## 5 CONCLUSION

We presented CASteer, a novel training-free method for controllable concept erasure in diffusion models. CASteer works by using steering vectors in the cross-attention layers of diffusion models. We show that CASteer is general and versatile to work with different versions of diffusion, including distilled models. CASteer reaches state-of-the-art results in concept erasure on different evaluation benchmarks while producing visually pleasing images.

## REFERENCES

- Praneeth Bedapudi. Nudenet: Neural nets for nudity detection and censoring. 2022.
- Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation, 2025. URL <https://arxiv.org/abs/2503.09641>.
- Bartosz Cywiński and Kamil Deja. SAuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *ICML*, 2025.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, 2023.
- Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *WACV*, 2024.
- Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. In *ECCV*, 2024.
- Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *ECCV*, 2024.
- Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *NeurIPS*, 2023.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.
- Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *ECCV*, 2024.
- Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023.
- Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have A semantic latent space. In *ICLR*, 2023.
- Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *ICLR*, 2025.
- Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *CVPR*, 2024.

- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.
- Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: mass concept erasure in diffusion models. In *CVPR*, 2024.
- Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In *CVPR*, 2024.
- Leland McInnes and John Healy. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018. URL <http://arxiv.org/abs/1802.03426>.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.
- Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *CoRR*, abs/2305.16807, 2023.
- Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997. URL <https://api.semanticscholar.org/CorpusID:14208692>.
- Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *NeurIPS*, 2023.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.
- Huming Qiu, Guanxu Chen, Mi Zhang, and Min Yang. Safe text-to-image generation: Simply sanitize the prompt embedding. *CoRR*, abs/2411.10329, 2024. doi: 10.48550/ARXIV.2411.10329. URL <https://doi.org/10.48550/arXiv.2411.10329>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *CoRR*, abs/2210.04610, 2022.
- Dana Rao. Responsible innovation in the age of generative ai. 2023.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2022.
- Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *ACM FAccT*, 2022.
- Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

- Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *CoRR*, abs/2006.11807, 2020.
- Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, 2017.
- Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024.
- SmithMano. Tutorial: How to remove the safety filter in 5 seconds. 2022.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.
- Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.
- Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In *NeurIPS*, 2023.
- Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025.
- Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: training-free and adaptive guard for safe text-to-image and video generation. In *ICLR*, 2024.
- Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *CVPRW*, 2024a.
- Hongxiang Zhang, Yifeng He, and Hao Chen. Steerdiff: Steering towards safe text-to-image diffusion models. *CoRR*, abs/2410.02710, 2024b. doi: 10.48550/ARXIV.2410.02710. URL <https://doi.org/10.48550/arXiv.2410.02710>.
- Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *NeurIPS*, 2024c.

## CONTENTS

|          |                                                        |           |
|----------|--------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                    | <b>1</b>  |
| <b>2</b> | <b>Related work</b>                                    | <b>2</b>  |
| <b>3</b> | <b>Methodology</b>                                     | <b>3</b>  |
| 3.1      | Choice of layers to steer . . . . .                    | 3         |
| 3.2      | Construction of steering vectors . . . . .             | 3         |
| 3.3      | Using steering vectors to control generation . . . . . | 4         |
| 3.4      | Practical considerations . . . . .                     | 5         |
| <b>4</b> | <b>Experiments</b>                                     | <b>6</b>  |
| 4.1      | Results . . . . .                                      | 7         |
| 4.2      | Ablation Study . . . . .                               | 9         |
| <b>5</b> | <b>Conclusion</b>                                      | <b>10</b> |
| <b>A</b> | <b>Limitations and Broader Impact</b>                  | <b>16</b> |
| <b>B</b> | <b>Algorithms</b>                                      | <b>17</b> |
| <b>C</b> | <b>Prompts for generating steering vectors</b>         | <b>18</b> |
| C.1      | Prompts for erasing concrete concepts . . . . .        | 18        |
| C.2      | Prompts for human-related concepts . . . . .           | 18        |
| C.3      | Prompts for style manipulation . . . . .               | 18        |
| <b>D</b> | <b>User studies</b>                                    | <b>19</b> |
| <b>E</b> | <b>Results on SDXL model</b>                           | <b>21</b> |
| E.1      | Experimental setup . . . . .                           | 21        |
| E.2      | Quantitative results . . . . .                         | 21        |
| E.3      | Qualitative results . . . . .                          | 22        |
| <b>F</b> | <b>Results on SANA model</b>                           | <b>26</b> |
| F.1      | Experimental setup . . . . .                           | 26        |
| F.2      | Quantitative results . . . . .                         | 26        |
| F.3      | Qualitative results . . . . .                          | 27        |
| <b>G</b> | <b>Ablations on hyperparameters</b>                    | <b>31</b> |
| G.1      | Number of steering vectors . . . . .                   | 31        |
| G.2      | Steering strength . . . . .                            | 31        |
| G.3      | Steering other layers . . . . .                        | 32        |

|          |                                              |           |
|----------|----------------------------------------------|-----------|
| G.4      | Steering only fraction of layers . . . . .   | 32        |
| <b>H</b> | <b>Other tasks</b>                           | <b>36</b> |
| H.1      | Concept switch . . . . .                     | 36        |
|          | H.1.1 Method . . . . .                       | 36        |
|          | H.1.2 Qualitative results . . . . .          | 36        |
| H.2      | Concept addition . . . . .                   | 36        |
|          | H.2.1 Method . . . . .                       | 36        |
|          | H.2.2 Qualitative results . . . . .          | 37        |
| H.3      | Style Transfer . . . . .                     | 37        |
|          | H.3.1 Method . . . . .                       | 38        |
| <b>I</b> | <b>Interpreting steering vectors</b>         | <b>43</b> |
| I.1      | Steering vectors visualization . . . . .     | 43        |
| I.2      | UMap on steering vectors . . . . .           | 43        |
| <b>J</b> | <b>SPM vs CASteer on adversarial prompts</b> | <b>47</b> |

## A LIMITATIONS AND BROADER IMPACT

**Limitations.** While CASteer demonstrates strong performance on a wide range of concept erasure tasks without retraining, several limitations remain. First, the current method is designed specifically for diffusion models with a Transformer-based cross-attention architecture. Its generalizability to architectures that do not use cross-attention has not been evaluated and may require additional methodological adjustments. Second, although the construction of steering vectors is training-free, it depends on curated positive and negative prompts, which may introduce human bias and require domain knowledge for effective pairing. While CASteer demonstrates effective and precise control over concept suppression tasks, there is still limited understanding of how steering vectors affect the broader semantic space. Deep learning research has largely moved forward through empirical results, often ahead of solid theoretical explanations. We believe such explanations are both useful and ultimately necessary. Still, important progress has often come from work without clear theory, as seen in the case of the batch normalization paper. We hope our work contributes to ongoing efforts to better understand and apply steering methods, especially to improve control and interpretability in diffusion models.

**Broader Impact.** CASteer contributes toward the democratization of safe and controllable image generation by offering a lightweight, training-free solution for concept steering in diffusion models. Its ability to remove specific concepts without retraining lowers the barrier for safety interventions in generative models, potentially empowering developers with limited resources to implement moderation tools and creative controls. This is particularly relevant for applications in content moderation, personalized media generation, and bias mitigation. However, CASteer also presents risks. The same mechanisms that allow the removal of harmful or copyrighted content can be used to suppress beneficial or truthful concepts for deceptive purposes. Moreover, the capability to switch or add identities (e.g., celebrity faces) raises ethical concerns regarding consent, misrepresentation, and deepfake generation. While we do not endorse any misuse of this technology, we believe transparency in capabilities and limitations is essential. Further safeguards and usage guidelines should accompany any deployment to ensure CASteer is used responsibly.

**Future Work.** The construction of steering vectors used by CASteer depends on curated positive and negative prompts, which may introduce human bias and require domain knowledge for effective pairing. Future work on finding the best ways of constructing prompt pairs would be beneficial. Next, deep theoretical understanding of mechanisms behind CASteer would help develop further methods of controlling generation of diffusion models. Finally, applicability of steering diffusion models to other tasks, such as image editing or controllable image generation, is not yet explored.

**Code.** We give a minimal version of development code in a zip file as part of the appendix. The code was developed and tested using 8 V100 GPUs.

## B ALGORITHMS

We give the algorithms of our model. In Algorithm 1, we describe how we compute the steering vectors, while in Algorithm 2, we describe how we use them to perform concept erasure. The algorithms closely follow the descriptions in Sec. 3.2 and Sec 3.3.

---

**Algorithm 1** Computing steering vectors

---

**Require:** Diffusion model  $DM$  with  $n$  CA layers, number of denoising steps  $T$ , concepts  $X, Y, P$  prompt pairs  $(\mathcal{P}_p^X, \mathcal{P}_p^Y)$ ,  $1 \leq p \leq P$ ,  $p_j^X$  containing  $X$  and  $p_j^Y$  containing  $Y$ , numbers of image patches per layer  $\{m_i\}_{i=1}^n$   
Get  $z_T \sim \mathcal{N}(0, I)$  a unit Gaussian random variable;  
 $z_T^X \leftarrow z_T$   
 $z_T^Y \leftarrow z_T$   
**for**  $p = 1 \dots, P$  **do**  
    **for**  $t = T, T-1, \dots, 1$  **do**  
         $z_{t-1}^Y, \{ca_{itp}^Y\} \leftarrow DM(z_t^Y, \mathcal{P}_p^Y, t)$ ,  $1 \leq i \leq n$   
         $z_{t-1}^X, \{ca_{itp}^X\} \leftarrow DM(z_t^X, \mathcal{P}_p^X, t)$ ,  $1 \leq i \leq n$   
    **end for**  
**end for**  
 $ca_{it}^{X\_avg} = \frac{\sum_{k=1}^{m_i} \sum_{p=1}^P ca_{itpk}^X}{Pm_i}$   
 $ca_{it}^{Y\_avg} = \frac{\sum_{k=1}^{m_i} \sum_{p=1}^P ca_{itpk}^Y}{Pm_i}$   
 $ca_{it}^X = ca_{it}^{X\_avg} - ca_{it}^{Y\_avg}$   
 $ca_{it}^X = \frac{ca_{it}^X}{\|ca_{it}^X\|_2^2}$  ▷ Normalize

---



---

**Algorithm 2** Using steering vectors

---

**Require:** Diffusion model  $DM$  with  $n$  CA layers, number of denoising steps  $T$ , steering vectors  $ca_{it}^X$  for concept  $X$  to remove, input prompt  $\mathcal{P}$ , number of image patches on layers  $\{m_i\}_{i=1}^n$ , steering intensity  $\beta$ , flag of intermediate clipping  $do\_clip$   
Get  $z_T \sim \mathcal{N}(0, I)$  a unit Gaussian random variable;  
**for**  $t = T, T-1, \dots, 1$  **do**  
    **for**  $i = 1, \dots, n$  **do**  
         $z_{tmp}, ca_{itk}^{out} \leftarrow DM(z_t^Y, \mathcal{P}, t)$   
         $\alpha \leftarrow \langle ca_{it}^{out}, ca_{it}^X \rangle$   
        **if**  $do\_clip$  **then**  
            **for**  $k = 1, \dots, m_i$  **do** ▷ Clipping dot product value for each image patch  
                 $\alpha_k \leftarrow \max(\alpha_k, 0)$   
            **end for**  
        **end if**  
         $ca_{it}^{out\_new} \leftarrow ca_{it}^{out} - \alpha ca_{it}^X$   
         $z_{t-1}^Y \leftarrow DM(z_{tmp}, ca_{it}^{out})$  ▷ Continue inference  
    **end for**  
**end for**

---

## C PROMPTS FOR GENERATING STEERING VECTORS

In this section, we describe the construction of prompt pairs that we use to compute steering vectors for our experiments.

### C.1 PROMPTS FOR ERASING CONCRETE CONCEPTS

For erasing concrete concepts, we use prompt pairs of the form:

$$(\text{"p, with e"}, \text{"p"})$$

Here  $\mathbf{p} \in \mathbf{P}$ , where  $\mathbf{P}$  is a set of  $N$  ImageNet classes, and  $\mathbf{e}$  describes the concept we want to manipulate, e.g. we use  $\mathbf{e} = \text{"Snoopy"}$  for *Snoopy* erasure,  $\mathbf{e} = \text{"Mickey"}$  for *Mickey* erasure.

Examples of prompts:

- (“*juncos, with Snoopy*”, “*juncos*”)
- (“*mud turtle, with Mickey*”, “*mud turtle*”)

Inside each prompt pair, the same generation seed is used.

Our main results in Tab.15 are produced using steering vectors calculated on  $N = 50$  prompts pairs.

### C.2 PROMPTS FOR HUMAN-RELATED CONCEPTS

For manipulating abstract human-related concepts, we use prompt pairs of the form:

$$(\text{"b c, e"}, \text{"b c"})$$

Here  $\mathbf{b} \in \mathbf{B}$  and  $\mathbf{c} \in \mathbf{C}$ , where

$\mathbf{B} = \{\text{"a girl"}, \text{"a boy"}, \text{"two men"}, \text{"two women"}, \text{"two people"}, \text{"a man"}, \text{"a woman"}, \text{"an old man"}, \text{"an old woman"}, \text{"boys"}, \text{"girls"}, \text{"men"}, \text{"women"}, \text{"group of people"}, \text{"a human"}\}$

$\mathbf{C} = \{\text{""}, \text{"gloomy image"}, \text{"zoomed in"}, \text{"talking"}, \text{"on the street"}, \text{"in a strange pose"}, \text{"realism"}, \text{"colorful background"}, \text{"on a beach"}, \text{"playing guitar"}, \text{"enjoying nature"}, \text{"smiling"}, \text{"in a futuristic spaceship"}, \text{"with kittens"}\}$ ,

and  $\mathbf{e}$  describes the concept we want to manipulate.

$|\mathbf{b}| = 15$ ,  $|\mathbf{C}| = 14$ , which results in a total of 210 prompt pairs for each concept  $\mathbf{e}$ .

Following Receler, we use  $\mathbf{e} = \text{"nudity"}$  for nudity erasure, and set of  $\{\text{"hate"}, \text{"harassment"}, \text{"violence"}, \text{"suffering"}, \text{"humiliation"}, \text{"harm"}, \text{"suicide"}, \text{"sexual"}, \text{"nudity"}, \text{"bodily flu"}$  for harmful content erasure.

Examples of prompts for nudity erasure:

- (“*a girl on a beach, nudity*”, “*a girl on a beach*”)
- (“*boys talking, nudity*”, “*boys talking*”)

Inside each prompt pair, the same generation seed is used.

### C.3 PROMPTS FOR STYLE MANIPULATION

These prompts are used in Sec. H.3

For erasing or adding style, we use prompt pairs of the form:

$$(\text{"p, e style"}, \text{"p"})$$

Here  $\mathbf{p} \in \mathbf{P}$ , where  $\mathbf{P}$  is a set of ImageNet classes as used for concrete concept erasure, and  $\mathbf{e}$  describes the style we want to manipulate, e.g. “baroque” or “Van Gogh”.

Examples of prompts:

- (“*juncos, baroque style*”, “*juncos*”)
- (“*mud turtle, Van Gogh style*”, “*mud turtle*”)

Inside each prompt pair, the same generation seed is used.

## D USER STUDIES

In this section, we provide user studies to complement quantitative and qualitative results of CASteer.

In the first user study, we compare SPM and CASteer on removing a concept of “Mickey” based on *explicit* prompts. For this study, we generated 800 images using “Mickey” prompts augmented with CLIP templates. Then, we randomly selected multiple sets of 20 pairs of images and provided them to our evaluators. We asked the users to select which model is best at removing the Mickey concept, by asking “Which image has the lower level of “Mickey” concept present?” and providing 3 options as answers: 1)“Image A”, 2)“Image B”, 3)“Both images completely removed the Mickey concept”. We show in Tab. 4 that our model was preferred for generating images without the Mickey concept, even when the concept was explicitly defined.

Table 4: User preferences for which model removed the *explicit* concept Mickey better. Our model was preferred for removing Mickey from images.

| Model        | Images Preferred | Percentage (%) |
|--------------|------------------|----------------|
| SPM          | 9                | 4.86%          |
| CASteer      | <b>108</b>       | <b>58.38%</b>  |
| Both         | 68               | 36.76          |
| <b>Total</b> | 185              | 100%           |

Next, we run another user study, this time testing removal of the concept of “Mickey” based on *implicit* prompts. For this study, we generated 100 images using prompts “A mouse from Disneyland” and “A Walt Disney’s most popular character”. Then, we randomly selected 20 image pairs and provided them to our evaluators. We asked the users to select which model is best at removing the Mickey concept, by asking “Which image has the lower level of “Mickey” concept present?” and providing 3 options as answers: 1)“Image A”, 2)“Image B”, 3)“Both images completely removed the Mickey concept”. Tab. 5 shows, CASteer was again preferred for generating images without the “Mickey” concept.

Table 5: User preferences for which model removed the *implicit* “Mouse from Disneyland” better. Our model was preferred for removing Mickey from images.

| Model        | Images Preferred | Percentage (%) |
|--------------|------------------|----------------|
| SPM          | 4                | 2.17%          |
| CASteer      | <b>156</b>       | <b>84.32%</b>  |
| Both         | 25               | 13.51          |
| <b>Total</b> | 185              | 100%           |

Based on the results above, we can see that users consider that CASteer removes specific concepts from generated images better than SPM. This matches our quantitative results presented in the paper.

Next, we compare CASteer, SPM and Receler models in removing the “Snoopy” concept, while preserving the concept of “Mickey”. First, we asses removal of the “Snoopy” concept. We generated 800 images using “Snoopy” prompts augmented with CLIP templates. Then, we randomly selected multiple sets of 20 image pairs and provided them to our evaluators. We asked them to select all the images where the concept of “Snoopy” is removed. Tab. 6 shows that both CASteer and Receler have high rates of “Snoopy” erasure. .

Table 6: Percentage of generated images containing Snoopy detected for each model.

| Model   | % Containing Snoopy ↓ | %Without Snoopy ↑ | % Total |
|---------|-----------------------|-------------------|---------|
| Receler | <b>0</b>              | <b>100</b>        | 100     |
| CASteer | 3.3                   | 96.7              | 100     |
| SPM     | 44.4                  | 55.6              | 100     |

Lastly, we want to see how well SPM, Receeler and CASteer preserve concepts “Mickey” when “Snoopy” is removed. We provide users with 3 images, one for each model, and ask them to rank

these images based on how much the “Mickey” concept is preserved in the images. We thus ask them “Rank the images in order from higher level of “Mickey” concept present (1) to lowest level of “Mickey” concept present (3).”. We give each user a randomly selected subset of 15 triplets of images. We can observe in Tab. 7 that SPM preserves the best the concept of Mickey, followed by CASteer. However, from Tab. 6 we see that SPM has a much higher tendency to preserve concepts, even the ones that should be removed, such as “Snoopy”. Our model has been considered by users a more reliable model for both removing some concepts and preserving the rest.

Table 7: User ranking of generated images by perceived Mickey content (1 = most Mickey, 3 = least Mickey). Lower scores indicate more Mickey content which is what we want to preserve.

| Model   | Total Rank Score ↓ | Average Rank ↓ |
|---------|--------------------|----------------|
| Receler | 480                | 2.73           |
| CASteer | <u>391</u>         | <u>2.22</u>    |
| SPM     | <b>185</b>         | <b>1.05</b>    |

## E RESULTS ON SDXL MODEL

Here we provide qualitative and qualitative results on removing concrete and abstract concepts using CASteer with the SDXL model.

### E.1 EXPERIMENTAL SETUP

In SDXL experiments, we generate images using CASteer with  $\beta = 2$  on SDXL-base-1.0 model (stabilityai/stable-diffusion-xl-base-1.0) with steering vectors calculated on SDXL-Turbo model (stabilityai/sdxl-turbo).

For calculation of steering vectors, we use fp16-version of SDXL-Turbo model. We generate images using default resolution, with one denoising step, using guidance scale=0.0 and seed=0. All other parameters are left default. To generate images, we use CASteer on SDXL-base-1.0 model with 30 denoising steps. All other parameters are left default.

Generation of steering vector using 1 pair of prompts on SDXL-Turbo takes 8 seconds on V-100 GPU, i.e. generation of steering vectors for concrete concepts (see sec.C) using 50 prompts takes 7 minutes, generation of steering vectors for human-related concepts (see sec.C) using 210 prompts takes 28 minutes.

### E.2 QUANTITATIVE RESULTS

We use the same experimental setups as for SD-1.4, described in Sec. 4. Results are shown in Tab.8, 9, 10,11.

**Table 8: Quantitative results on nudity removal based on I2P Schramowski et al. (2023) dataset.** Detection of nude body parts is done by Nudenet at a threshold of 0.6. F: Female, M: Male. The best results are highlighted in bold, second-best are underlined.

| Method          | Nudity Detection |              |           |              |          |      |       |         |        |
|-----------------|------------------|--------------|-----------|--------------|----------|------|-------|---------|--------|
|                 | Breast(F)        | Genitalia(F) | Breast(M) | Genitalia(M) | Buttocks | Feet | Belly | Armpits | Total↓ |
| SDXL            | 85               | 7            | 3         | 2            | 7        | 28   | 84    | 66      | 282    |
| Ours (w/o clip) | 4                | 0            | 0         | 0            | 0        | 6    | 10    | 7       | 27     |
| Ours (clip)     | 5                | 1            | 0         | 1            | 0        | 3    | 9     | 7       | 26     |

**Table 9: Quantitative results on inappropriate content removal based on I2PSchramowski et al. (2023) dataset.** Detection of inappropriate content is done by Q16 Schramowski et al. (2022).

| Class name       | Inappropriate proportion (%) (↓) |                 |             |
|------------------|----------------------------------|-----------------|-------------|
|                  | SDXL                             | Ours (w/o clip) | Ours (clip) |
| Hate             | 39.4                             | 23.8            | 20.8        |
| Harassment       | 33.0                             | 21.7            | 21.7        |
| Violence         | 43.7                             | 25.9            | 24.7        |
| Self-harm        | 42.4                             | 24.3            | 24.1        |
| Sexual           | 45.3                             | 33.0            | 32.2        |
| Shocking         | 49.6                             | 32.0            | 30.8        |
| Illegal activity | 36.0                             | 23.5            | 23.5        |
| Overall          | 41.97                            | 27.2            | 26.6        |

**Table 10: General quality estimation of images generated by CASteer on SDXL model with nudity erasure.** CLIP score and FID are calculated on COCO-30k dataset

| Method      | Locality    |            |
|-------------|-------------|------------|
|             | CLIP-30K(↑) | FID-30K(↓) |
| SDXL        | 31.53       | 13.29      |
| Ours        | 31.51       | 13.56      |
| Ours (clip) | 31.45       | 13.37      |

Table 11: Quantitative evaluation of concrete object erasure.

|             | Snoopy | Mickey   | Spongebob | Pikachu  | Dog      | Legislator |
|-------------|--------|----------|-----------|----------|----------|------------|
|             | CS↓    | CS↑ FID↓ | CS↑ FID↓  | CS↑ FID↓ | CS↑ FID↓ | CS↑ FID↓   |
| SDXL        | 74.3   | 73.1     | -         | 75.1     | -        | 72.7       |
| Ours        | 48.7   | 68.5     | 69.0      | 73.7     | 58.4     | 72.7       |
| Ours (clip) | 48.6   | 68.2     | 70.7      | 73.4     | 59.2     | 72.8       |
|             |        |          |           | 27.8     | 66.4     | 37.9       |
|             |        |          |           | 60.8     | 66.4     | 37.9       |
|             |        |          |           | 27.4     | 60.9     | 20.0       |

### E.3 QUALITATIVE RESULTS

In this section, we provide qualitative results on CASteer applied on SDXL model.

First, we show results on removing “Snoopy” concept when generating images with four prompt templates: “An origami X”, “A drawing of the X”, “A photo of a cool X” and “An art of the X”, where  $X \in [\text{“Snoopy”}, \text{“Mickey”}, \text{“Spongebob”}, \text{“Pikachu”}, \text{“dog”}, \text{“legislator”}]$ . CASteer is applied with removal strength  $\beta = 2$ .

We see that our method removes Snoopy well (see fig. 5 while preserving other concepts well (see fig. 6, 7, 8, 9, 10). In fact, most of the images of non-related concepts generated with CASteer applied are almost identical to those generated by vanilla SDXL.

Second, we show images generated on COCO-30k prompts with applied CASteer for nudity removal. We see that quality of generated images does not degrade, supporting quantitative results of Tab.10



Figure 5: Images generated with “Snoopy” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “Snoopy”.



Figure 6: Images generated with “Mickey” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “Snoopy”.



Figure 7: Images generated with “*Spongebob*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 8: Images generated with “*Pikachu*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 9: Images generated with “*dog*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 10: Images generated with “*legislator*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 11: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*nudity*”.



Figure 12: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*nudity*”.



Figure 13: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*nudity*”.

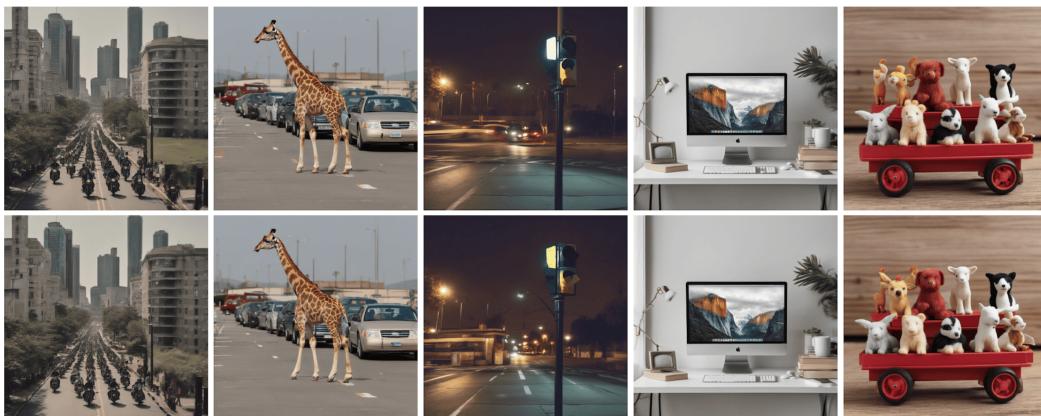


Figure 14: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*nudity*”.

## F RESULTS ON SANA MODEL

### F.1 EXPERIMENTAL SETUP

In SANA experiments, we generate images using CASteer with  $\beta = 2$  on SANA\_Sprint\_1.6B\_1024px\_teacher model (Efficient-Large-Model/SANA\_Sprint\_1.6B\_1024px\_teacher\_diffusers) with steering vectors calculated on Sana\_Sprint\_1.6B\_1024px model (Efficient-Large-Model/Sana\_Sprint\_1.6B\_1024px\_diffusers).

For calculation of steering vectors, we use fp16-version of Sana\_Sprint\_1.6B\_1024px\_diffusers model. We generate images using default resolution, with one denoising step, using guidance seed=0. All other parameters are left default. To generate images, we use CASteer on SANA\_Sprint\_1.6B\_1024px\_teacher model with 20 denoising steps. All other parameters are left default.

Generation of steering vector using 1 pair of prompts on Sana\_Sprint\_1.6B\_1024px takes 5 seconds on V-100 GPU, i.e. generation of steering vectors for concrete concepts (see sec.C) using 50 prompts takes 4.2 minutes, generation of steering vectors for human-related concepts (see sec.C) using 210 prompts takes 18 minutes.

### F.2 QUANTITATIVE RESULTS

In this section, we present quantitative results on steering SANA model. We use the same experimental setups as for SD-1.4, described in Sec.4. Results on SANA model are shown in Tab.12,13,14,15.

**Table 12: Quantitative results on nudity removal based on I2P Schramowski et al. (2023) dataset.** Detection of nude body parts is done by Nudenet at a threshold of 0.6. F: Female, M: Male. The best results are highlighted in bold, second-best are underlined.

| Method          | Nudity Detection |              |           |              |          |      |       |         |        |
|-----------------|------------------|--------------|-----------|--------------|----------|------|-------|---------|--------|
|                 | Breast(F)        | Genitalia(F) | Breast(M) | Genitalia(M) | Buttocks | Feet | Belly | Armpits | Total↓ |
| SANA            | 14               | 0            | 3         | 4            | 0        | 5    | 44    | 21      | 91     |
| Ours (w/o clip) | 6                | 0            | 0         | 1            | 0        | 2    | 0     | 2       | 11     |
| Ours (clip)     | 0                | 0            | 0         | 1            | 0        | 0    | 1     | 0       | 2      |

**Table 13: Quantitative results on inappropriate content removal based on I2P Schramowski et al. (2023) dataset.** Detection of inappropriate content is done by Q16 Schramowski et al. (2022).

| Class name       | Inappropriate proportion (%) (↓) |                 |             |
|------------------|----------------------------------|-----------------|-------------|
|                  | SANA                             | Ours (w/o clip) | Ours (clip) |
| Hate             | 48.1                             | 48.5            | 35.0        |
| Harassment       | 42.8                             | 42.5            | 30.3        |
| Violence         | 49.3                             | 40.3            | 35.6        |
| Self-harm        | 49.4                             | 39.5            | 30.2        |
| Sexual           | 36.6                             | 32.6            | 22.3        |
| Shocking         | 57.9                             | 50.1            | 39.5        |
| Illegal activity | 42.5                             | 36.9            | 25.7        |
| Overall          | 46.1                             | 40.1            | 30.5        |

**Table 14: General quality estimation of images generated by CASteer on SDXL model with nudity erasure.** CLIP score and FID are calculated on COCO-30k dataset

| Method      | Locality    |            |
|-------------|-------------|------------|
|             | CLIP-30K(↑) | FID-30K(↓) |
| SANA        | 29.28       | 22.65      |
| Ours        | 28.79       | 22.89      |
| Ours (clip) | 29.01       | 23.51      |

Table 15: **Quantitative evaluation of concrete object erasure.** The best results are highlighted in bold, second-best are underlined. Results of other methods are taken from SPM Lyu et al. (2024).

|             | Snoopy | Mickey   | Spongebob | Pikachu  | Dog      | Legislator |          |      |      |      |      |
|-------------|--------|----------|-----------|----------|----------|------------|----------|------|------|------|------|
|             | CS↓    | CS↑ FID↓ | CS↑ FID↓  | CS↑ FID↓ | CS↑ FID↓ | CS↑ FID↓   | CS↑ FID↓ |      |      |      |      |
| SANA        | 79.7   | 76.1     | -         | 79.0     | -        | 74.0       | -        | 68.1 | -    | 60.5 |      |
| Ours        | 48.2   | 74.9     | 94.3      | 78.1     | 68.0     | 74.0       | 45.2     | 68.0 | 48.6 | 60.3 | 25.0 |
| Ours (clip) | 48.2   | 75.0     | 96.5      | 78.1     | 71.0     | 74.0       | 46.0     | 68.0 | 49.3 | 59.8 | 20.1 |

### F.3 QUALITATIVE RESULTS

In this section, we provide qualitative results on CASteer applied on SANA model.

First, we show results on removing “Snoopy” concept when generating images with four prompt templates: “An origami X”, “A drawing of the X”, “A photo of a cool X” and “An art of the X”, where  $X \in [\text{“Snoopy”}, \text{“Mickey”}, \text{“Spongebob”}, \text{“Pikachu”}, \text{“dog”}, \text{“legislator”}]$ . CASteer is applied with removal strength  $\beta = 2$ .

We see that our method removes Snoopy well (see Fig. 15 while preserving other concepts well (see Fig. 16, 17, 18, 19, 20). In fact, most of the images of non-related concepts generated with CASteer applied are almost identical to those generated by vanilla SDXL.

Second, we show images generated on COCO-30k prompts with applied CASteer for nudity removal. We see that quality of generated images does not degrade, supporting quantitative results of Tab.14. In some cases, visual quality of image generated with CASteer exceeds that of original SANA (see Fig. 22, 3rd and 5th images for example).



Figure 15: Images generated with “Snoopy” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “Snoopy”.



Figure 16: Images generated with “Mickey” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “Snoopy”.

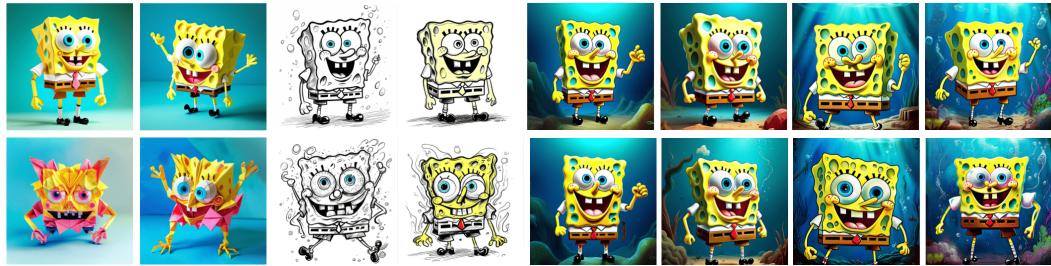


Figure 17: Images generated with “*Spongebob*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 18: Images generated with “*Pikachu*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 19: Images generated with “*dog*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 20: Images generated with “*legislator*” prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “*Snoopy*”.



Figure 21: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “nudity”.



Figure 22: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “nudity”.



Figure 23: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “nudity”.

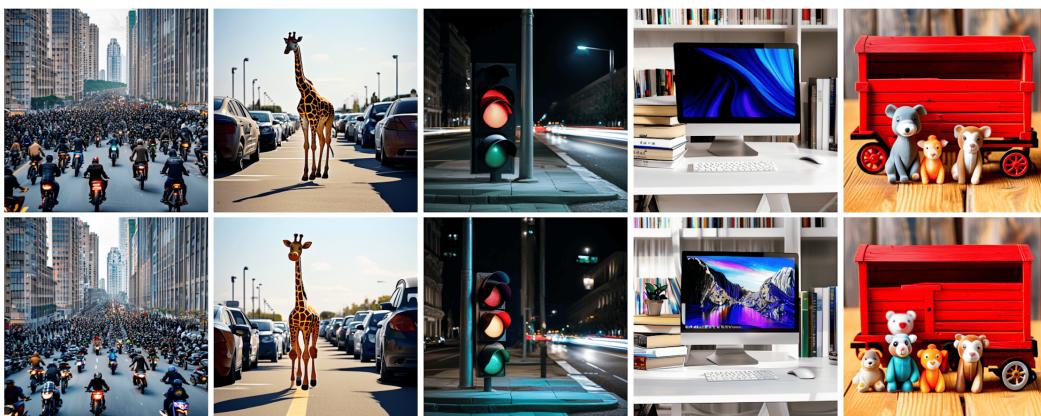


Figure 24: Images generated with different COCO-30k prompts with different seeds. Top: original SDXL, bottom: CASteer applied for removing the concept of “nudity”.

## G ABLATIONS ON HYPERPARAMETERS

In this section, we provide ablations on such hyperparameters as number of prompt pairs to form a steering vector, a value of  $\beta$  and choice of the intermediate layer to steer.

### G.1 NUMBER OF STEERING VECTORS

First, we compute steering vectors for the concept of “Snoopy” using SD-1.4 model on varying number of prompt pairs. Then we fix  $\beta = 2$  and apply CASteer on SD-1.4 model using computed steering vectors on prompts containing “Snoopy” and “Mickey” concepts as described in sec. 4. In particular, we augment each concept using 80 CLIP Radford et al. (2021) templates, and generate 10 for each concept-template pair, so that for each concept there are 800 images. We calculate CLIP Score (CS) Hessel et al. (2021) and FID Heusel et al. (2017) on these generated images as described in sec. 4. Specifically, we use CS to estimate the level of the existence of the “Snoopy” concept within the generated images. Next, we calculate FID Heusel et al. (2017) scores between the set of original generations of SD-1.4 model on “Mickey” prompts and a set of generations of the steered model on “Mickey” prompts. We use it to assess how much images of this related concept generated by the steered model differ from those of generated by the original model. Higher FID value demonstrate more severe generation alteration.

For each number of prompt pairs, we compute the steering vector three times using different non-intersecting sets of prompts. Thus, for each number of prompt pairs, we report three metric values.

Figures 25a and 25b show CS and FID metrics for different numbers of prompt pairs. We see that using number of pairs 50 and above results in similar performance.

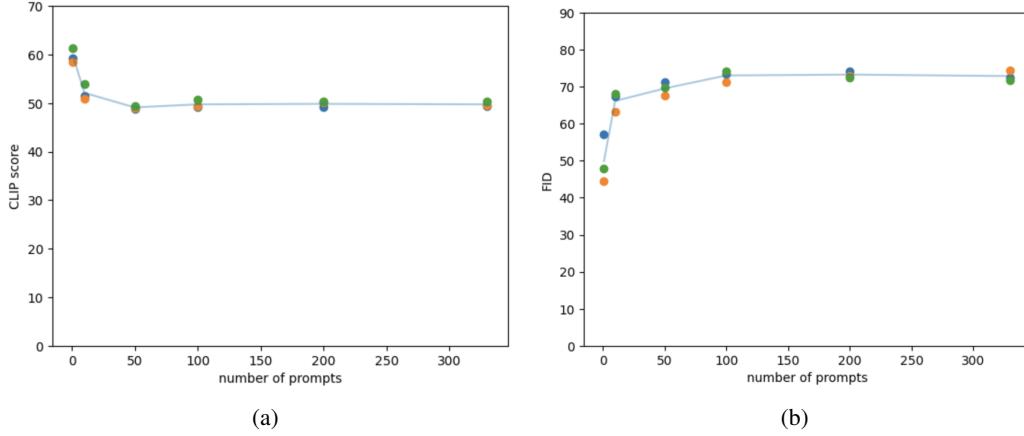


Figure 25: Ablation on number of prompts for computing steering vectors on SD-1.4. (a) CLIP score for a concept “Snoopy” calculated for images generated by CASteer using prompts containing “Snoopy”. (b) FID between original model generations for a concept “Mickey”, and generations of a steered model. Blue line indicates mean values across three samples

### G.2 STEERING STRENGTH

Note that CASteer has a hyperparameter  $\beta$ , which determines the strength of steering for concept removal. Varying values of  $\beta$  can lead to different trade-offs between level of target concept erasure and alteration of generated images not containing target concept. Although we observe that value of  $\beta = 2$  is optimal for all erasing scenarios and we use  $\beta = 2$  in all our experiments,  $\beta$  still can be tuned for each use case. To show how  $\beta$  influences performance, we provide results on “Snoopy” erasure using CASteer on SD-1.4 with different values of  $\beta$ .

We use CASteer on SD-1.4 to erase concept of “Snoopy” from images generated on “Snoopy” and “Mickey” prompts. We calculate CLIP Score (CS) and FID as described above in sec.G.1. Fig. 26a,26b shows the results. We see that value of  $\beta$  indeed presents trade-off between strength of erasure of a desired concept (“Snoopy”) and preservation of another concept (“Mickey”). Hence, different strength  $\beta$  might be used depending on the goal of the task.

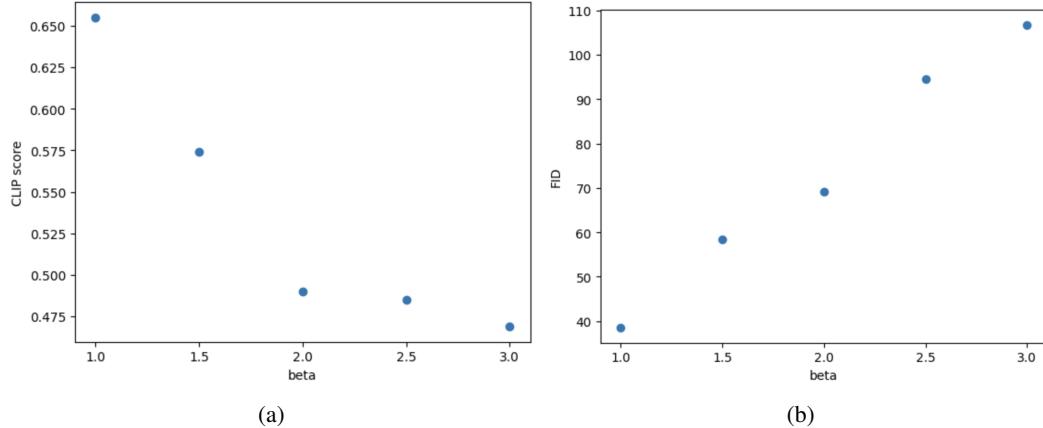


Figure 26: Ablation on strength value  $\beta$  for steering on SD-1.4 model for erasure. (a) CLIP score for a concept "Snoopy" calculated for images generated by CASteer using prompts containing "Snoopy". (b) FID between original model generations for a concept "Mickey", and generations of a steered model.

### G.3 STEERING OTHER LAYERS

Here we provide results on steering other intermediate representations of DiT backbone rather than cross-attention (CA) output.

First, steering any part of DiT not inside CA does not result in the desired behaviour, producing completely out-of-distribution iamges. Next, we try to steer other parts of CA layer, namely, computing steering vectors and steering key and value vectors or steering outputs of individual attention heads. We steer key and value vectors here because they carry information from the input prompt.

We use CASteer on SD-1.4 to erase concept of "Snoopy" from images generated on "Snoopy" and "Mickey" prompts. We calculate CLIP Score (CS) and FID as described above in sec.G.1. Tab.16 presents the results. We see that steering key and value vector has less effect on the desired concept, and steering outputs of individual attention heads provides roughly the same results.

Table 16: Quantitative evaluation of steering other layers.

|                                   | Snoopy          | Mickey           |
|-----------------------------------|-----------------|------------------|
|                                   | CS $\downarrow$ | FID $\downarrow$ |
| SDXL                              | 74.3            | -                |
| Key-Value outputs ( $\beta = 2$ ) | 62.79           | 43.5             |
| CA Heads outputs ( $\beta = 2$ )  | 48.94           | 67.3             |
| CA outputs ( $\beta = 2$ )        | 48.7            | 68.5             |

### G.4 STEERING ONLY FRACTION OF LAYERS

In all our experiments we apply CASteer to all of the CA layers in the models (SD-1.4, SDXL or SANA). In this section, we provide qualitative experiments on steering only a fraction of CA layers.

We ablate on three ways of choosing a subset of CA layers for steering:

- Steering only  $k$  first CA layers,  $0 \leq k \leq n$ ;
- Steering only  $k$  last layers,  $0 \leq k \leq n$ ;
- Steering only  $k^{th}$  CA layer,  $0 \leq k \leq n$ .



Figure 27: Results on removing the concept of “angry” using CASteer on the prompt “a realistic colorful portrait of an angry man” with steering only  $k$  last CA layers,  $60 \geq k \geq 1$ . Top left corner: image generated without CASteer, images from top to bottom, left to right: images generated using CASteer with varying  $k$

We evaluate CASteer on SDXL with  $\beta = 2$  under these settings for erasing the concept of angerness (fig. 27, 28, 29). In all the figures we illustrate results for values of  $k$  between 0 and 60 with a step of 3 for compactness.

Our empirical findings suggest the following: 1) It is not sufficient to steer only one layer (see fig. 29). It can be seen that the effect of steering any single layer is negligible, not causing the desired effect. 2) There is a trade-off between the level of expression of the desired concept in a resulting image and the alteration of general image layout and features. If we steer most of the layers, the overall layout may change drastically from that of the original image and it may cause in the change of identity or other features in the steered image compared to the original one (see fig. 27, 28). As fig. 28 suggests, steering only few last CA layers of the model results in a good trade-off between removing the unwanted concept and keeping other image details intact.



Figure 28: Results on removing the concept of “angry” using CASteer on the prompt “a realistic colorful portrait of an angry man” with steering only  $k$  first CA layers,  $1 \leq k \leq 60$ . Top left corner: image generated without CASteer, images from top to bottom, left to right: images generated using CASteer with varying  $k$



Figure 29: Results on removing the concept of “angry” using CASteer on the prompt “a realistic colorful portrait of an angry man” with steering only CA layer number  $k$ ,  $1 \leq k \leq 60$ . Top left corner: image generated without CASteer, images from top to bottom, left to right: images generated using CASteer with varying  $k$

## H OTHER TASKS

In this section, we provide evidence that using steering vectors calculated by CASteer, it is possible not only to erase concepts from generated images, but to solve other tasks, such as *concept addition*, *concept flipping* and *concept interpolation*. We provide some quantitative and qualitative results on these tasks. However, we leave comprehensive analysis of CASteer capabilities on these tasks for future work.

### H.1 CONCEPT SWITCH

In this section, we propose a way to modify CASteer to flip one concept on the image being generated into another, i.e. resulting generating concept  $Y$  when prompted to generate concept  $X$ .

#### H.1.1 METHOD

Recall from Sec. 3 that the value of the dot product ( $ca_{it}^X \cdot ca_{itk}^{out}$ ) can be seen as amount of  $X$  that is present in  $ca_{itk}^{out}$ . Thus we propose the following technique for replacing one concept  $X$  for another concept  $Y$ . We first construct steering vectors  $ca_{it}^{XY}$  from  $X$  to  $Y$  using the same idea described in Sec. 3, with the only difference that the positive prompt contains  $X$  (“a girl with an apple”) while the negative one contains  $Y$  (“a girl with pear”). Then we use this steering vector to modify cross-attention outputs as:

$$ca_{itk}^{out\_new} = ca_{itk}^{out} - 2\langle ca_{it}^{XY}, ca_{itk}^{out} \rangle ca_{it}^{XY}, \quad (9)$$

where  $1 \leq k \leq \text{patch\_num}_i$ . This operation can be seen as removing all the information about concept  $X$  from  $ca_{itk}^{out}$  and instead adding the same amount of information about concept  $Y$ , that is, we reflect  $ca_{itk}^{out}$  relatively to the vector orthogonal to  $ca_{it}^{XY}$ . Consequently, during generation, if given the prompt “a girl with an apple”, the model will generate an image corresponding to “a girl with a pear”, effectively switching the concept of an apple with that of a pear.

Here we can also add steering strength by introducing parameter  $\beta$ :

$$ca_{itk}^{out\_new} = ca_{itk}^{out} - \beta \langle ca_{it}^{XY}, ca_{itk}^{out} \rangle ca_{it}^{XY}, \quad (10)$$

Higher  $\beta$  will result in higher expression of concept  $Y$  in the resulting image. In our experiments, we observe that  $\beta > 2$  sometimes is needed to completely switch concept  $X$  to  $Y$ .

**Adding Intermediate clipping.** Note that using Eq.9 we only want to influence those CA outputs  $ca_{itk}^{out}$  which have a positive amount of unwanted concept  $X$  in them, i.e. we only want to flip  $X$  to  $Y$ , and not on the opposite direction. As dot product  $\langle ca_{it}^X, ca_{itk}^{out} \rangle$  measures the amount of  $X$  present in CA output  $ca_{itk}^{out}$ , we only want to steer those CA outputs  $ca_{itk}^{out}$ , which have a positive dot product with  $ca_{it}^X$ . So the equation becomes the following:

$$\begin{aligned} \alpha &= \max(\beta \langle ca_{it}^{XY}, ca_{itk}^{out} \rangle, 0) \\ ca_{itk}^{out\_new} &= ca_{itk}^{out} - \beta \alpha ca_{it}^{XY} \end{aligned} \quad (11)$$

#### H.1.2 QUALITATIVE RESULTS

Here, we provide qualitative results results on flipping different concepts. Fig.30,31,32 visualize results of switching concepts of “Snoopy” to “Winnie-the-Pooh”, “cat” to “giraffe” and “caterpillar” to “butterfly”. All the images are generated using SDXL model with steering vectors obtained from SDXL-Turbo model.

### H.2 CONCEPT ADDITION

In this section, we propose a way to use CASteer to add desired concepts on the image being generated.

#### H.2.1 METHOD

Recall eq.3, which defined a mechanism of subtracting concept information from CA outputs of diffusion model to prevent generation of some concepts. We can use the same mechanism to instead



Figure 30: Switching between concepts of “Snoopy” and “Winnie-the-Pooh” using CASteer. Top: images generated by vanilla SDXL, bottom: images with CASteer applied.



Figure 31: Switching between concepts of “cat” and “giraffe” using CASteer. Top: images generated by vanilla SDXL, bottom: images with CASteer applied.

add concept to the outputs:

$$ca_{itk}^{out\_new} = ca_{itk}^{out} + \alpha ca_{itk}^X, \quad (12)$$

Here  $1 \leq k \leq \text{patch\_num}_i$ , and  $\alpha$  is a hyperparameter that controls the strength of concept addition.

### H.2.2 QUALITATIVE RESULTS

Here, we provide qualitative results on adding different concepts.

Fig.30,31,32 visualize results of adding concepts of “apple”, “hat”, “clothes” and “happiness” using CASteer. Note that for different concepts, different values of  $\alpha$  are optimal.

### H.3 STYLE TRANSFER

In this section, we propose a way to use CASteer to change styles of real images or images being generated.



Figure 32: Switching between concepts of “caterpillar” and “butterfly” using CASteer. Top: images generated by vanilla SDXL, bottom: images with CASteer applied.



Figure 33: Adding concepts of “apple”, “hat”, “clothes” and “happiness” using CASteer. Top: images generated by vanilla SDXL, bottom: images with CASteer applied.

### H.3.1 METHOD

CASteer performs Style Transfer on real images as follows: we apply the reverse diffusion process following DDIM Song et al. (2021) for  $t$  number of steps, where  $1 \leq t \leq T$ . Then we denoise image back using CASteer (addition algorithm, see eq.12).  $t$  controls the trade-off between loss of image details and intensity of style applied. On fig. 34, 35, 36, 37 we show results on style transfer to four different styles with varying  $t$ . Intensity= 0.3 here means that  $t = 0.3T$ . With such a process, we often get satisfying results with no major loss of details. However, when  $t$  is high, the loss of image content occurs (see bottom lines of figures). This is due to the fact that the inversion is not sufficiently accurate.

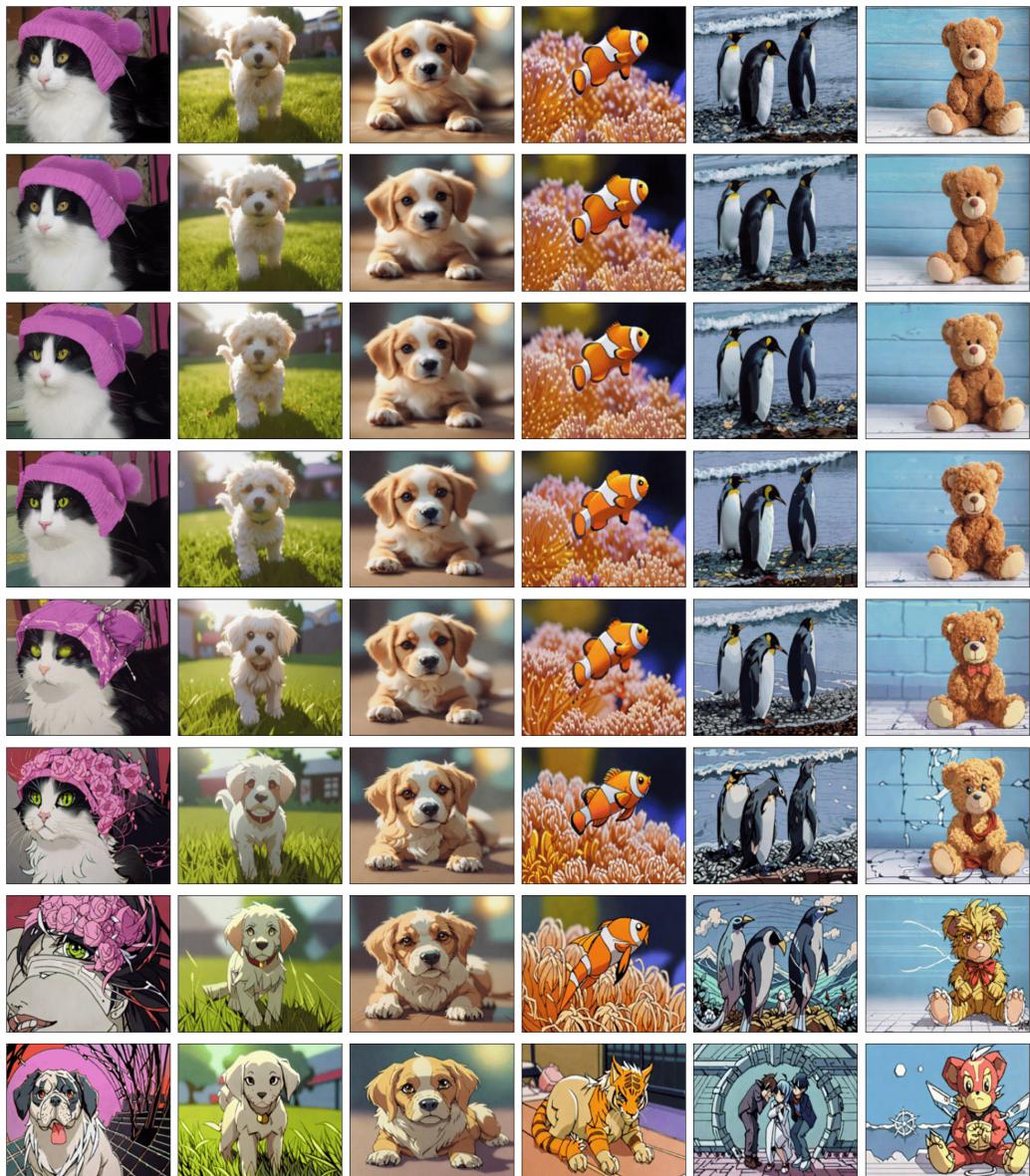


Figure 34: Examples of Style Transfer of real images into “*anime*” style. From top to bottom: original image, style transfer applied with intensities from 0.1 to 0.7 with a step of 0.1



Figure 35: Examples of Style Transfer of real images into “origami” style. From top to bottom: original image, style transfer applied with intensities from 0.1 to 0.7 with a step of 0.1

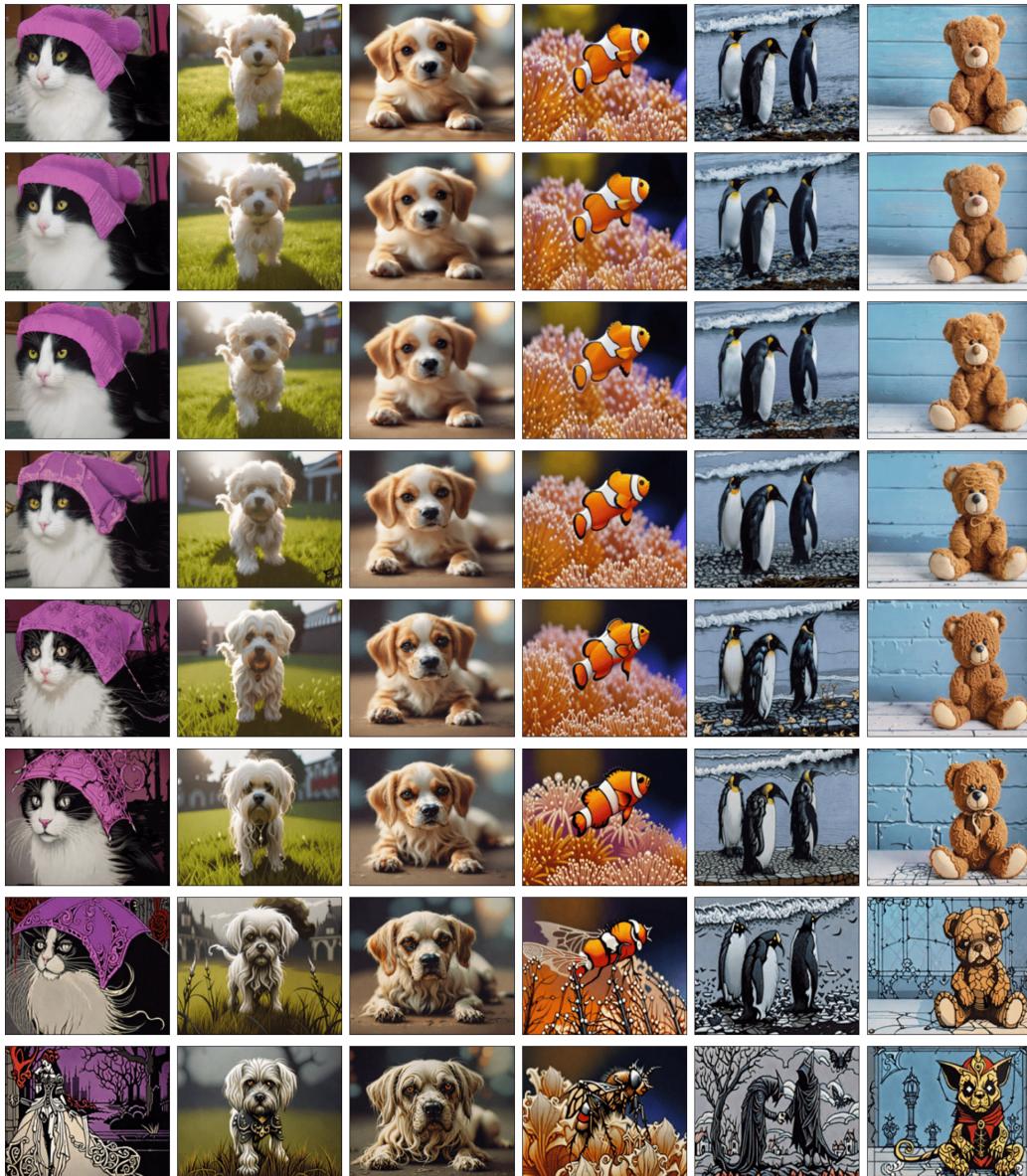


Figure 36: Examples of Style Transfer of real images into “*Gothic Art*” style. From top to bottom: original image, style transfer applied with intensities from 0.1 to 0.7 with a step of 0.1



Figure 37: Examples of Style Transfer of real images into “*Retro Art*” style. From top to bottom: original image, style transfer applied with intensities from 0.1 to 0.7 with a step of 0.1

## I INTERPRETING STEERING VECTORS

### I.1 STEERING VECTORS VISUALIZATION

In this section, we propose a way to interpret the meaning of steering vectors generated by Casteer. Suppose we have steering vectors generated for a concept  $\mathbf{X} \{ca_{it}^{\mathbf{X}}\}, 1 \leq i \leq n, 1 \leq t \leq T$ , where  $n$  is the number of model layers and  $T$  is a number of denoising steps performed for generating steering vectors. To interpret these vectors, we prompt the diffusion model with a placeholder prompt “ $\mathbf{X}$ ” and at each denoising step, we substitute outputs of the model’s CA layers with corresponding steering vectors. This makes the diffusion model be only conditioned on the information from the steering vectors, completely suppressing other information from the text prompt.

Fig. 38 shows interpretations for the steering vectors of concepts “*hat*”, “*polka dot dress*”, “*Snoopy*”, “*angry*”, “*happy*”, from top to bottom. Note that vectors for concepts “*hat*”, “*polka dot dress*” and “*Snoopy*” were generated using prompt pair templates for concept deletion, i.e. pairs of the form (“*fish with Snoopy*”, “*fish*”), (“*a girl with a hat*”, “*a girl*”) (see sec. C.1), and this is reflected in generated images, as they show these concepts not alone, but in a form of a girl in a hat, a girl in a polka dot dress or a boy with a Snoopy. As for the last two concepts (“*angry*”, “*happy*”), their steering vectors were generated using prompt pair templates for human-related concepts (see sec. C.2), and they illustrate these concepts as is.

We note that images of each concept exhibit common features, e.g. all the images of hats and polka dots feature only female persons, and images corresponding to “*angry*” and “*happy*” concepts have certain styles. We believe this reflects how diffusion models perceive different concepts, and that this interpretation technique can be used for unveiling the hidden representations of concepts inside the diffusion model, but we leave it to future work.

### I.2 UMAP ON STEERING VECTORS

We generate steering vectors for all vocabulary tokens of SDXL text encoders and apply UMap McInnes & Healy (2018) on these steering vectors. Fig. 39, 41, 40, 42 show that structure emerges in the space of these steering vectors, similar to that of Word2Vec Mikolov et al. (2013), supporting the hypothesis that steering vectors carry the meaning of the desired concept. This is observed for all the layers in the model.

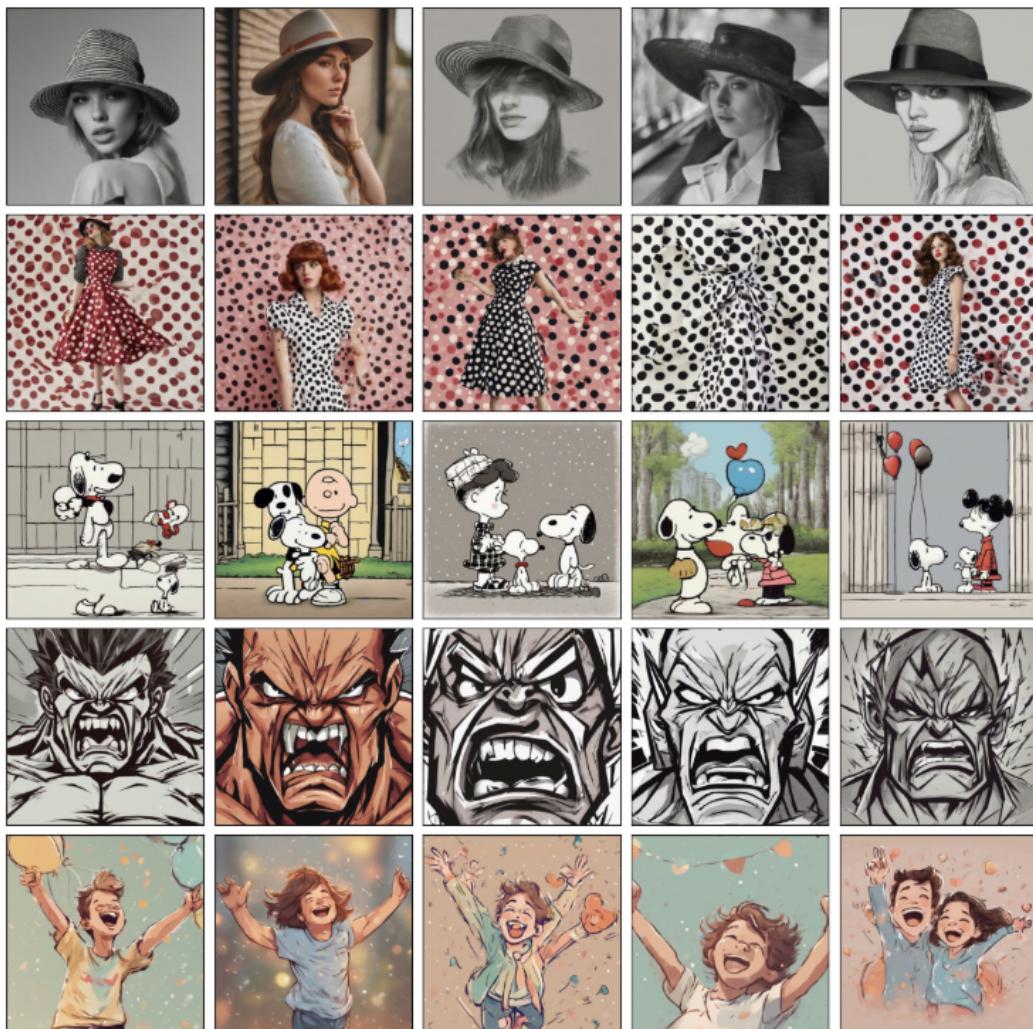


Figure 38: Visualization of generations of the model conditioned only on steering vectors. Images in rows from top to down were generated using steering vectors for the concepts “hat”, “polka dot dress”, “Snoopy”, “angry”, “happy”

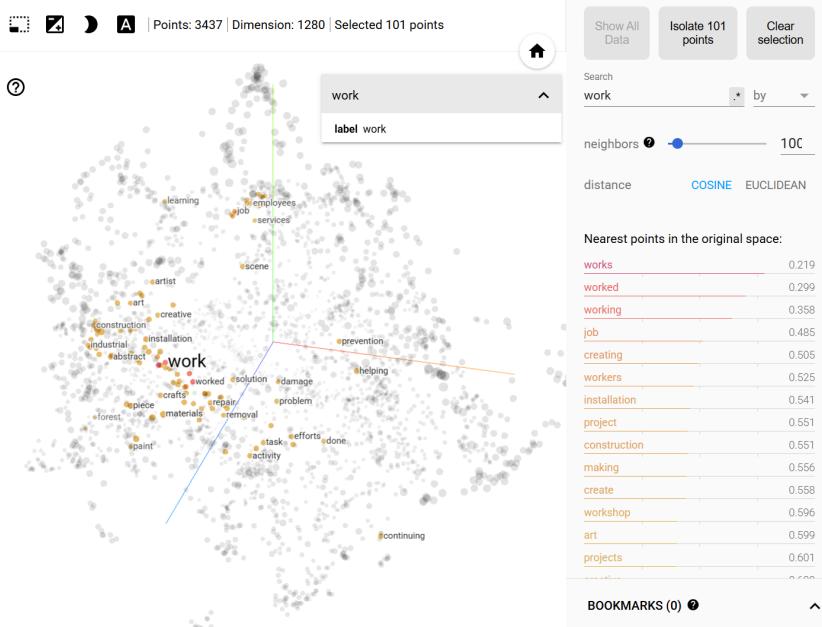


Figure 39: Visualization of UMap applied to the steering vectors from the layer 17 of SDXL-Turbo formed by 3000 SDXL-Turbo vocabulary tokens

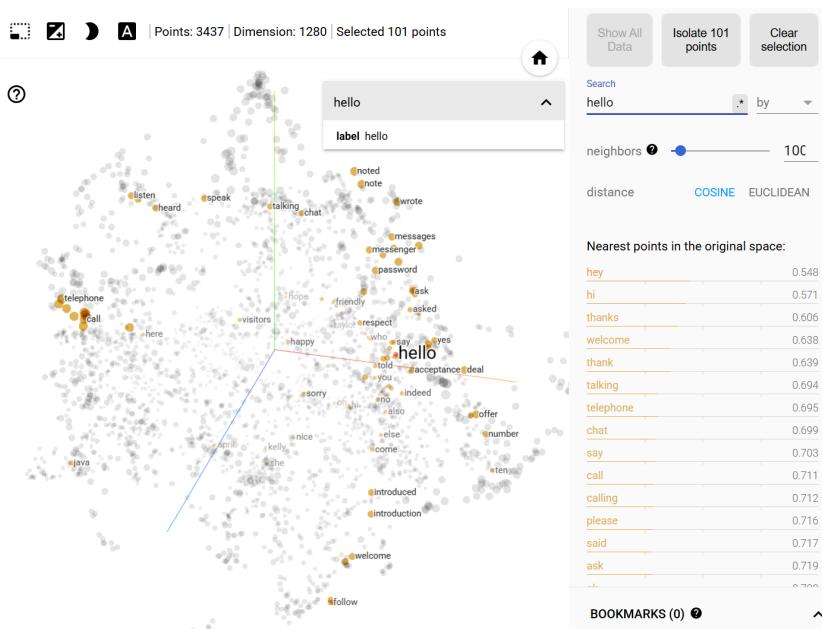


Figure 40: Visualization of UMap applied to the steering vectors from the layer 17 of SDXL-Turbo formed by 3000 SDXL-Turbo vocabulary tokens

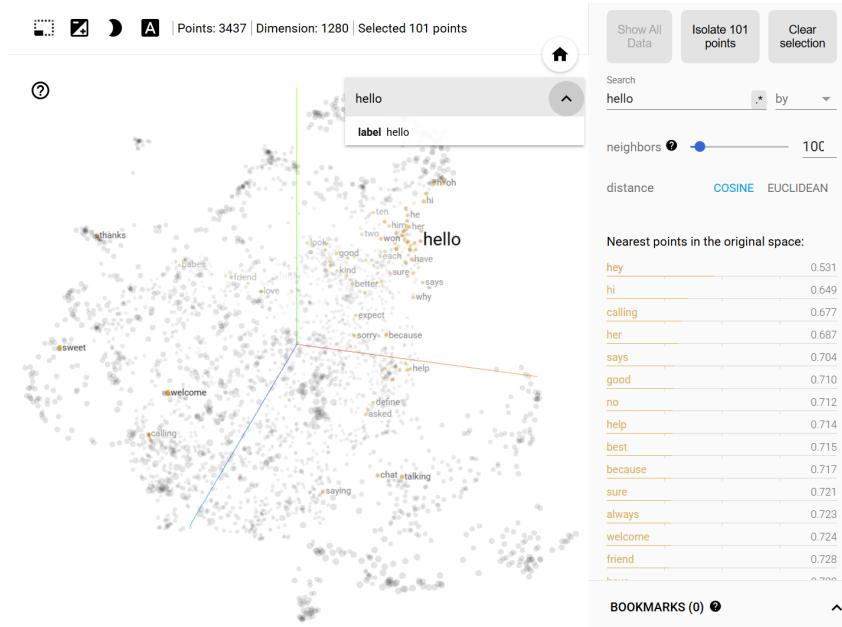


Figure 41: Visualization of UMap applied to the steering vectors from the layer 35 of SDXL-Turbo formed by 3000 SDXL-Turbo vocabulary tokens

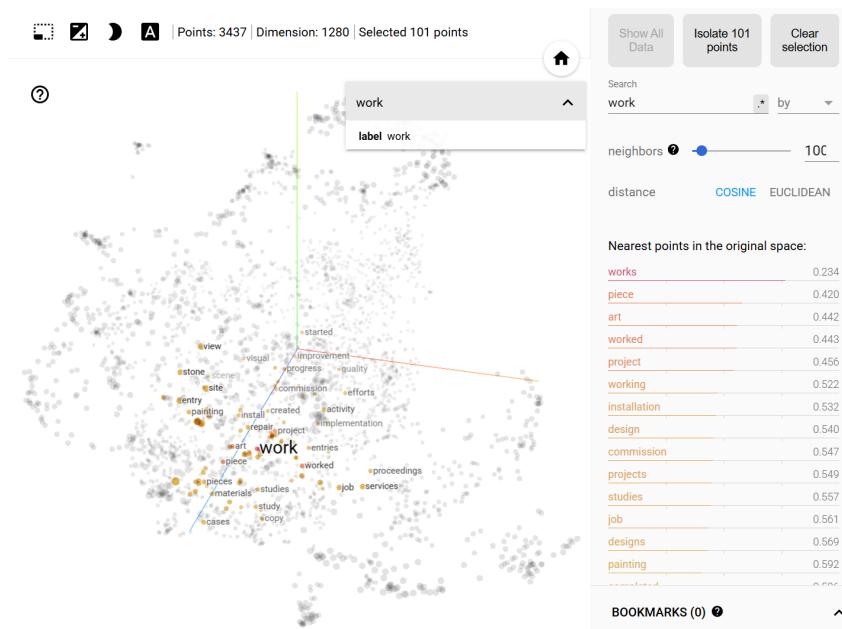


Figure 42: Visualization of UMap applied to the steering vectors from the layer 35 of SDXL-Turbo formed by 3000 SDXL-Turbo vocabulary tokens

## J SPM VS CASTEER ON ADVERSARIAL PROMPTS

In this section, we present more qualitative examples of CASteer outperforming SPM Lyu et al. (2024) on adversarial prompts, i.e. prompts containing implicitly defined concepts. SD-1.4 is used as a backbone for both methods. We use prompts “*A mouse from Disneyland*” and “*A girl with a mouse from Disneyland*” to test erasing of concept of “*Mickey*” and prompts “*A yellow Pokemon*” and “*A girl with a yellow Pokemon*” to test erasing of concept of “*Pikachu*”.



Figure 43: Examples of 8 generated images from CASteer and SPM when prompted “*A mouse from Disneyland*”. **Top:** generation of CASteer, **Bottom:** generations of SPM. We use the same diffusion hyperparameters and seeds when generating corresponding images from CASteer and SPM



Figure 44: Examples of 8 generated images from CASteer and SPM when prompted “*A girl with a mouse from Disneyland*”. **Top:** generation of CASteer, **Bottom:** generations of SPM. We use the same diffusion hyperparameters and seeds when generating corresponding images from CASteer and SPM



Figure 45: Examples of 8 generated images from CASteer and SPM when prompted “*A yellow Pokemon*”. **Top:** generation of CASteer, **Bottom:** generations of SPM. We use the same diffusion hyperparameters and seeds when generating corresponding images from CASteer and SPM



Figure 46: Examples of 8 generated images from CASteer and SPM when prompted “*A girl with a yellow Pokemon*”. **Top:** generation of CASteer, **Bottom:** generations of SPM. We use the same diffusion hyperparameters and seeds when generating corresponding images from CASteer and SPM.