# MIDSTEER: OPTIMAL AFFINE FRAMEWORK FOR STEERING GENERATIVE MODELS

**Anonymous authors**Paper under double-blind review

# **ABSTRACT**

The idea of steering intermediate representations of generative models has recently emerged as a simple yet powerful approach for controlling aspects of generated texts and images. However, despite the simplicity of the approach, no theoretical framework has yet been built around steering. In this paper, we aim to bridge this gap, building theory around concept steering. First, we provide a theoretical link between steering and affine concept erasure, showing that the widely used steering approach for erasing unwanted behaviours or concepts from generative models is a special case of LEACE, a closed-form method for affine concept erasure in neural networks. Next, we consider the task of concept switching, the aim of which is to change information about an unwanted concept or behaviour in the model's representations into another, more desired concept or behaviour. Here our contribution is two-fold: first, we formulate a theoretical framework for this task, adapting the existing affine concept erasure framework used for concept erasure. Then, we identify weaknesses of the resulting framework, and propose a new, improved one, that we call **MiDSteer** (Minimal Disturbance concept **Steering**). Our results show that MidSteer performs favourably on a variety of tasks, modalities, and models, including image generative diffusion models and LLMs.

# 1 Introduction

Generative models such as Large Language Models (LLMs) and image diffusion models have achieved remarkable progress in recent years Yang et al. (2024b) Naveed et al. (2023). However, controlling model outputs to enforce desirable behaviors or suppress harmful ones remains challenging Bartoszcze et al. (2025). Yet, this capability is necessary for improving model safety, reliability, alignment, and usefulness in downstream applications.

Concept steering of intermediate representations is an increasingly popular technique that has already proven to be simple yet powerful for controlling behaviour in LLMs. Recently it was also shown to be applicable to image diffusion models Gaintseva et al. (2025). The underlying idea is to change the intermediate representations of a generative model during generation by adding or subtracting a "steering vector" that encodes a target concept. This approach has proven effective for tasks such as erasing unwanted behaviors (toxicity, nudity) or amplifying desirable features (helpfulness, truthfulness). However, despite the simplicity of the approach, its theoretical foundations remain underdeveloped with most of the work around it being highly empirical. Existing methods largely rely on heuristic vector manipulations, which can introduce unintended side effects and lack solid theoretical basis and guarantees. Moreover, naive steering often perturbs unrelated features, undermining the minimal disturbance principle that is critical to maintaining model quality and coherence.

Recently, affine concept erasure frameworks have been developed Ravfogel et al. (2023a) Belrose et al. (2025) for removing undesired concepts from models representations. Yet, these frameworks do not extend naturally to other forms of steering, such as concept addition or concept switching, where the goal is to add information about a desired concept or replace information about one concept with another.

In this work, we address these gaps by **developing a unified theoretical framework** for affine steering of generative models (both language models and diffusion models). We first show that the

widely used steering method for concept deletion in LLMs is a special case of LEACE, a closed-form affine erasure approach. We then extend this perspective to formulate an optimal affine framework for concept switching, establishing its connection to Householder reflections. Finally, we **introduce** MidSteer (Minimal Disturbance Concept Steering), a new framework designed to achieve precise concept switching while minimizing interference with other aspects of the representation.

Through experiments with LLMs and diffusion models, we demonstrate that **MidSteer achieves more reliable concept switching than prior methods**, allowing controllable generation with minimal side effects. Our results highlight the value of grounding steering methods in theory and provide practical tools for aligning generative models with desired behaviors.

# 2 RELATED WORK

Steering generative models. Early work in LLMs demonstrated that adding or subtracting steering vectors derived from contrasting prompts can provoke or suppress targeted concepts such as sentiment, bias, or style from appearing in the generated text Turner et al. (2023). Later, Rimsky et al. (2024) proposed to construct steering vectors of concepts based on the mean activation difference, which was then established as an effective approach to steering Bartoszcze et al. (2025), Zou et al. (2023). Multiple approaches based on steering have since been developed to effectively steer LLMs representations from undesired behaviours or towards desired ones for different tasks. Wang et al. (2025) adaptively adjust steering intensity to improve truthfulness, Stickland et al. (2024) proposed KL-then-steer, a fine-tuning technique that decreases the side effects of steering applied to an LLM, Lu & Rimsky (2024) used steering for bias mitigation, Scialanga et al. (2025) showed that activation steering can be applied to knowledge editing, Rahn et al. (2024) applied steering to LLM agents.

In vision generative models, similar approaches to LLM steering have been developed to control image generations. Kwon et al. (2023) Park et al. (2023), Si et al. (2024), Tumanyan et al. (2023) focus on finding interpretable directions in various intermediate spaces of diffusion models, which can then be used to control the semantics of generated images. SDID Li et al. (2024) constructs learnable concept vectors, which are then added to intermediate activation of a bottleneck layer of the diffusion model during inference to control the level of this concept in generated images. However, these methods rely on different methods of constructing steering vectors than what is used in LLM, in most cases their steering vectors are learnable. Recently, Gaintseva et al. (2025) proposed to apply the steering of cross-attention outputs of the diffusion models to control generation, in a similar manner to the way it was done in LLMs, with steering vectors constructed using mean differences of intermediate activations of the diffusion model.

Affine concept erasure. One related line of research seeks to use affine transformations to remove concepts from representations altogether. Early methods such as INLP Ravfogel et al. (2020) iteratively trained linear classifiers to project out protected attributes like gender or sentiment. More recently, Ravfogel et al. (2023a) introduced log-linear guardedness as a theoretical basis for concept erasure, formalising the conditions under which a representation can be considered free of a concept. Building on this, Belrose et al. (2025) proposed LEACE, a closed-form solution for affine concept erasure that minimizes disturbance to the representation space. Later, Holstege et al. (2025), Singh et al. (2024) extended this framework for preserving task-relevant information under concept removal. These advances provide rigorous guarantees for concept deletion, but they do not directly address the broader goal of concept switching or targeted steering. In our work, we bridge this gap by extending LEACE to the framework of optimal concept flipping, and build a new theory-based framework of optimal concept switching.

# 3 METHODOLOGY

The concept of affine steering of internal representations h of the model to alter its behavior is widely used in Large Language Models and recently was also introduced in Vision Diffusion Models. It works by adding or subtracting a fixed steering vector s carrying information about a desired concept c from intermediate activations of a model during generation. The quantity of addition or subtraction is usually determined as proportionate to the dot product  $\langle h,s \rangle$  or left as a hyperparameter.

The steering vector s is usually formed by collecting the neural activity from a specific part of a neural network based on pairs of positive and negative input stimuli. Formally, for a model M, layer number l and a set of stimuli S consider a function Rep that accepts a model, input, layer number and stimuli and returns a set of neural activity M corresponding to the layer l and stimuli P. Then, if we have two sets of stimuli  $P^{pos} = \{p_i^{pos}\}_1^n$  and  $P^{neg} = \{p_i^{neg}\}_1^n$  containing n samples that differ only in the presence of a concept c, we construct steering vector as follows:

 $s^{c} = \frac{\sum_{i=1}^{n} \left( Rep(p_i^{pos}) - Rep(p_i^{neg}) \right)}{n} \tag{1}$ 

 $s^c$  can be optionally post-processed, e.g. normalized to have  $L^2$ -norm equal to 1. Specifics of Rep function can differ for different models M and different neural activity extraction approaches.

Eq.1 can also be re-written in terms of difference of concept conditional means of the internal representation h:

$$s^{c} = \mathbb{E}[h|C=1] - \mathbb{E}[h|C=0], \tag{2}$$

where  $C \in \{0,1\}$  is a binary variable representing the presence or absence of a concept c. Sometimes, instead of subtracting the negatively conditioned mean, the population mean is used instead:

$$s^{c} = \mathbb{E}[h|C=1] - \mathbb{E}[h] \tag{3}$$

For concepts which are seldom present (e.g. nudity, violence) and have P(C=1) close to 0, this is approximately equivalent.

With calculated steering vectors  $s^c$ , we can define the steering intervention that controls the expressiveness of the concept C in generation result of the model. Let h represent the original internal representation,  $s = s^c$  be a steering vector for a concept c calculated on a sample of data (we omit the subscript of  $s^c$  for clarity). Then the steering intervention is formulated as follows:

$$f_{\text{add}}(h,s) = h + \alpha s \tag{4}$$

 $\alpha$  is a hyperparameter that control the expressiveness of the concept C in model generation. Note that  $\alpha < 0$  leads to suppression of the concept from model generation.

Let us now highlight two special cases of steering setup, varying in choice of  $\alpha$ .

**Concept Deletion** (also referred as concept termination): In this case our aim is to prevent any information of the concept c to be present in current activation vector h.

$$f_{\text{delete}}(h,s) = h - \langle h, s \rangle s \tag{5}$$

The dot product is used as an estimation of the amount of concept c in current activation vector h, and the whole transformation is a projection onto subspace orthogonal to the steering vector s. We can also reformulate  $f_{\rm delete}$  in a matrix form:

$$f_{\text{delete}}(h, s) = (I - ss^T)h \tag{6}$$

**Switch/flip:** In the case of switching/flipping we aim to substitute information about a concept  $c_1$  with information about another concept  $c_2$  in current activation vector h. The steering vector s in this case is formed using pairs of prompts where positive prompts  $p_i^{pos}$  contain  $c_1$  and the negative ones  $p_i^{neg}$  contain  $c_2$ . The formula of intervention here is similar to that of termination up to a scale of dot product value:

$$f_{\text{switch}}(h, s) = h - 2\langle h, s \rangle s$$
 (7)

Note that this formula differs from Eq. 5 by the choice of multiplier of the dot product. In this case the formula becomes a Householder operator, and can be described as a reflection of the vector h around the hyperplane defined by the steering vector s. In this case, we also can reformulate  $f_{\rm switch}$  in a matrix form:

$$f_{\text{switch}}(h, s) = (I - 2ss^{T})h \tag{8}$$

In both cases, if steering is applied to the outputs of self-attention layers in LLM or cross-attention layers of diffusion model, it is possible to incorporate Eq. 6 and Eq. 8 into weight matrices of the model, thus achieving zero inference overhead when steering applied. Refer to Sec.6.1 in supplementary for more details.

The rest of the paper is organized as follows: First, we consider the task of concept erasure and derive a connection between steering setup for erasure and LEACE. In particular, we show that

Eq. 5 is a special case of LEACE. Next, we formulate similar theoretical framework for the task of optimal affine concept switching and show that Eq. 7 is a special case of the proposed framework. Finally, we improve the proposed framework by restricting it to only affecting one concept and derive MiDSteer, an affine optimal concept steering framework. In experimental section, we show that MidSteer outperforms steering in both LLMs and Vision Generative Diffusion Models, enabling precise concept switch while leaving other features of the images or texts intact.

#### 3.1 AFFINE GUARDEDNESS FRAMEWORK

For concept erasure, Ravfogel et al. (2023b) introduced the concept of guardedness. Belrose et al. (2025) expand it to a general formulation of affine guardedness. They prove that it is equivalent to having a zero covariance between concept vector  $Z \in \{0,1\}^k$  and internal representation h. It is also reasonable to assume that among all the transformations that modify h we seek the one with minimal changes to the representation space. Guided by these two assumptions, they prove the following:

**Theorem 1** (Belrose et al.). Let X, Z be random vectors taking values in  $\mathbb{R}^d$  and  $\mathbb{R}^k$  respectively, each of finite second moments. Let  $M \in \mathbb{R}^{d \times d}$  be a symmetric, positive semi-definite matrix defining an inner product in  $\mathbb{R}^d$ :  $\langle x, y \rangle_M = x^T M y$ . Define  $\Sigma_{XX} = \operatorname{Cov}(X, X) \in \mathbb{R}^{d \times d}$  and  $\Sigma_{XZ} = \operatorname{Cov}(X, Z) \in \mathbb{R}^{d \times k}$ . The following optimization problem:

$$E \in \mathbb{R}^{d \times d}$$
. The following optimization problem: 
$$\min_{\substack{A \in \mathbb{R}^{d \times d} \\ b \in \mathbb{R}^d}} \mathbb{E} \Big[ \|AX + b - X\|_M^2 \Big] \quad \textit{s.t.} \quad \operatorname{Cov}(AX + b, Z) = 0$$

has the following solution (almost surely):

$$\widehat{A} = I - W^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}W, \qquad (9)$$

$$\widehat{b} = \mathbb{E}[X] - \widehat{A} \cdot \mathbb{E}[X], \tag{10}$$

where  $W = (\Sigma_{XX}^{1/2})^+$  is whitening transformation.

Here and later we use  $A^+$  to denote the pseudo-inverse of (any) matrix A and  $A^{1/2}$  to denote the square root of a positive semi-definite symmetric matrix A, i.e. for the singular value decomposition  $A = VSV^\top$  with orthonormal matrix V and diagonal matrix S with non-negative singular values on the diagonal, the matrix  $A^{1/2}$  is defined as  $A^{1/2} := VS^{1/2}V^\top$ , where the square root of the diagonal entries of S is computed.

# 3.2 Affine erasure and LEACE

We will now show that steering in deletion mode (Eq. 5) is a special case of LEACE.

**Theorem 2.** Let X be a standardized random vector in  $\mathbb{R}^d$ , i.e. it has zero mean E[x] = 0 and unit covariance matrix  $\Sigma_{XX} = I$ . Let  $C \in \{0,1\}$  be a concept indicator variable. Let  $f_{delete}$  be defined as in Eq. 5 and s be defined as in Eq. 3. Then  $f_{delete}$  as a function of h minimizes

$$\min_{f \in \text{Aff}(\mathbb{R}^d \mapsto \mathbb{R}^d)} \text{E}[\|f(X) - X\|^2] \quad \text{s.t.} \quad \text{Cov}(f(X), C) = 0$$
(11)

This theorem states that steering in deletion mode can be seen as LEACE under the assumptions that the whitening matrix is identity and mean of all vectors is zero. Proof is found in the Appendix.

# 3.3 Concept switching

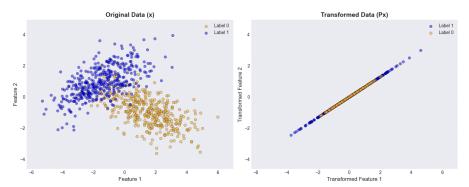
Now we aim to formulate a similar theoretical framework for the optimal affine concept switching. We also show that steering in flipping mode Eq. 7 can be seen as a special case of this framework.

Recall that  $Z \in \{0,1\}^k$  denotes a concept vector. For now, let us say that we want to manipulate the internal representation X so that every concept's label is flipped, i.e.

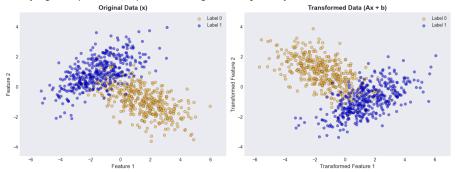
$$Cov(AX + b, \mathbf{1}^k - Z) = Cov(X, Z), \qquad (12)$$

where  $1^k$  denotes a k-dimensional vector where every entry is one. Due to the linearity of the covariance this is equivalent to

$$-Cov(AX + b, Z) = Cov(X, Z).$$



(a) Illustrative example of affine concept erasure. In this case the affine transformation is satisfying Cov(AX+b,Z)=0. This figure is inspired by Belrose et al. (2025)



(b) Illustrative example of affine concept flipping. In this case the affine transformation is satisfying -Cov(AX+b,Z)=Cov(X,Z).

Figure 1: Illustrative example of affine concept erasure and affine concept flipping frameworks

In analogy to Theorem 1, let us characterise the optimal affine-linear transformation that satisfies this constraint.

**Theorem 3** (Optimal concept switching). Let  $X, Z, M, \Sigma_{XX}, \Sigma_{XZ}$  be defined as in Theorem 1. Then the optimization problem

$$\min_{\substack{A \in \mathbb{R}^{d \times d}, \\ b \in \mathbb{R}^d}} \operatorname{E}\left[\|AX + b - X\|_M^2\right] s.t. \operatorname{Cov}(AX + b, Z) = -\operatorname{Cov}(X, Z)$$
(13)

has the following solution (almost surely):

$$\widehat{A} = I - 2W^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}W, \qquad (14)$$

$$\widehat{b} = \mathbb{E}[X] - \widehat{A}\mathbb{E}[X], \tag{15}$$

where  $W = \sum_{XX}^{-1/2}$  is whitening transformation.

Next we show the connection between steering in flipping mode Eq. 7 and the optimal concept switching framework. More specifically, we show that steering in flipping mode Eq. 7 can be seen as a special case of this framework.

**Theorem 4.** Let X be a standardized random vector in  $\mathbb{R}^d$ , i.e. it has zero mean  $\mathbb{E}[x] = 0$  and unit covariance matrix  $\Sigma_{XX} = I$ . Let  $C \in \{0,1\}$  be a concept indicator variable and P(C=1) + P(C=0) = 1. Let  $f_{switch}$  be defined as in Eq. 7 and s be defined as in Eq. 2. Then  $f_{switch}$  as a function of h minimizes

$$\min_{f \in \text{Aff}(\mathbb{R}^d \to \mathbb{R}^d)} \text{E}[\|f(X) - X\|^2] \quad \text{s.t.} \quad \text{Cov}(f(X), C) = -\text{Cov}(X, C)$$
 (16)

Here, C can be seen as labels of a binary classifier which has concepts  $c_1$  and  $c_2$  as classes. Basically, this theorem states that steering in switching mode can be seen as optimal concept switching under the assumptions that the whitening matrix is the identity and the mean of all vectors is zero.

Now note, that this theorem and matrix form reformulation refers to Eq. 7, and in practice both these formulations result in an issue that both concepts  $c_1$  and  $c_2$  are flipped, and in many cases this is not the desired behaviour. Imagine we are solving a problem of truthful content generation for a LLM. In this definition, we can have C=1 for truthful texts, and C=0 if they are untruthful.

Then we proceed to find a transform f(X) = AX + b that minimizes displacement that also satisfies Cov(AX + b, Z) = -Cov(X, Z). This transform will aim to flip the truthfulness, making truthful texts untruthful, which is not what we might want.

In the next section we aim to correct this and present a framework that allows for switching concept  $c_1$  to another concept  $c_2$  while not the opposite.

#### 3.4 Concept steering

Assume that  $Z = (Z_1, Z_2)$ , where  $Z_1, Z_2 \in \{0, 1\}^l$ .  $Z_1$  and  $Z_2$  represent indicators of concepts  $c_1$  and  $c_2$ . Our goal would be to have  $Cov(f(X), Z_1) = Cov(X, Z_2)$ . For reasons that are clarified in the proof, we will require  $Cov(X, Z_1)$  to be full rank. Now we formulate the following theorem.

**Theorem 5** (Affine optimal concept steering). Let X, M, Z be defined as in Theorem 3 and assume k = 2l for l > 0. Let  $Z = (Z_1, Z_2)$ , where  $Z_1, Z_2 \in \{0, 1\}^l$ . Let  $\Sigma_{XZ_i} = Cov(X, Z_i)$ ,  $i \in \{1, 2\}$  be the cross-covariance matrices between X and  $Z_i$ , with  $\Sigma_{XZ_1}$  having full column rank:  $\operatorname{rk}\left(\Sigma_{XZ_1}\right) = l$ . Let W be the ZCA whitening transform  $\Sigma_{XX}^{-1/2}$ . Let  $\Sigma_{WX,Z_i} = W\Sigma_{XZ_i}$ .

Then we have the following optimization problem:

$$\min_{\substack{A \in \mathbb{R}^{d \times d} \\ b \in \mathbb{R}^d}} \mathbb{E}[\|AX + b - X\|_M^2] \quad \text{s.t.} \quad Cov(AX + b, Z_1) = Cov(X, Z_2)$$
 (17)

which has the following solution (almost surely):

$$\widehat{A} = I + W^{+}(\Sigma_{WX,Z_{2}} - \Sigma_{WX,Z_{1}})\Sigma_{WX,Z_{2}}^{+}W$$
(18)

$$\widehat{b} = \mathbb{E}[X] - \widehat{A}\mathbb{E}[X] \tag{19}$$

Steering strength Let us now introduce the steering strength  $\beta$  for erasure and switching. This parameter is present in the initial formulas for CASteer. For Affine erasure (LEACE) and affine flipping, we can re-write the optimal solution obtained in Theorems 1 and 3 as follows:

$$A^* = I - \beta \cdot W^+(\Sigma_{WX,Z})(\Sigma_{WX,Z})^+ W \tag{20}$$

$$b^* = \mu - A^* \mu \tag{21}$$

When  $\beta=1$ , we have optimal affine concept erasure. If  $\beta=2$  (and appropriate covariance matrix  $\Sigma_{XZ}$ ), we have affine optimal concept switching. Note that in both cases the covariance matrix participates in the form  $\Sigma\Sigma^+$ , so any multiplicate constant (such as class probabilities) gets factored out and annihilated.

For MidSteer (concept steering, Theorem 5), the class labels need not be mutually exclusive and cover all the space of X, like what was required in the proof of Theorem 4. Let us now obtain the  $\beta$ -parametrized solution for steering between two classes (k = 1). Recall that:

$$\Sigma_{X,Z_i} = \text{Cov}(X, Z_i) = E[XZ_i] - E[X]E[Z_i] = P(Z_i = 1) \Big( E[X|Z_i = 1] - E[X] \Big) = P(Z_i = 1) \Big( \mu^{(i)} - \mu \Big) = P(Z_i = 1)\tilde{\mu}^{(i)}$$
(22)

Then, equation Eq. 18 becomes

$$A^* = I + W^+ W(\beta \tilde{\mu}^{(2)} - \tilde{\mu}^{(1)})(W \tilde{\mu}^{(1)})^+ W$$
(23)

$$b^* = \mu - A^* \mu \tag{24}$$

where  $\beta = \frac{P(Z_2=1)}{P(Z_1=1)}$ . While it is practically infeasible to estimate  $\frac{P(Z_2=1)}{P(Z_1=1)}$ ,  $\beta$  here can be seem as a hyperparameter defining strength of the intervention, analogous to the parameter in Eq. 21. We will use this equation throughout our experiments.

# 4 EXPERIMENTS

4.1 Experimental setup

For conducting steering experiments for Large Language Models (LLMs), we utilize instruction-tuned Llama 2 Touvron et al. (2023) and Qwen 2.5 family of models Yang et al. (2024a). As for experiments on diffusion models, we use SDXL Podell et al. (2024) and SANA Xie et al. (2025) models.

Following recent work, we apply steering on activations of self-attention (SA) layers in the case of LLMs, and cross-attention (CA) layers in the case of image diffusion models. Activations for estimating class-conditional means (for MidSteer) and steering vectors (for vanilla steering and LEACE) for all the concepts involved in erasure/switching are generated by averaging SA activations corresponding to the last token in prompt in the case of LLMs, or averaging CA activations corresponding to all the images patches in the case of diffusion models. We apply steering to all the SA/CA layers in the models.

In the LLM case, the dataset used to generate the concept prompts was obtained by prompting GPT o4-mini to generate various questions about each concept. To estimate  $\Sigma_{XX}$ , we use prompts from Alpaca dataset Taori et al. (2023). In the case of image diffusion models, both concept prompts and neutral prompts for  $\Sigma_{XX}$  estimation were extracted from RELAION dataset. Examples of prompts used in LLM can be found in the supplementary materials, Sec. 6.3. In all our experiments we used 1000 concept prompts for estimating class-conditional means, and 50000 neutral prompts to estimate  $\Sigma_{XX}$ . We ablate the number of prompts needed for  $\Sigma_{XX}$  in Sec. 6.6.

# 4.2 EXPERIMENTAL RESULTS

#### 4.2.1 Concept erasure

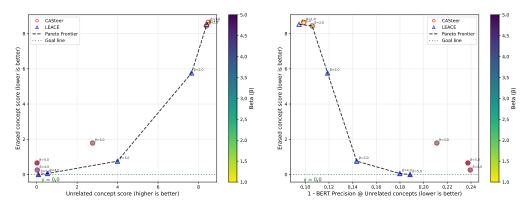
In this section, we compare performance of vanilla steering and LEACE to the task of concept erasure in both LLMs and diffusion models.

In the case of erasure, we aim to completely erase one concept in the model's responses (e.g. prevent the model from generating text about dogs when prompted to do so), while minimally affecting anything else in the generated outputs.

To test erasure of the source concept  $c_s$ , we use 80 template prompts prompting the model to generate output related to  $c_s$ . For each prompt we run 10 such generations varying the random seed. Templates for LLMs and diffusion models can be found in 6.5. We also use five additional concepts  $c_i$ ,  $i=1\ldots 5$  to test the capability of the method to preserve content not related to the concept being removed. We then run the generation on these prompts with and without steering. We use concept score (CS) to estimate the amount of concept present in the model's output. In the case of diffusion models, we use CLIP score Hessel et al. (2021) as CS. In the case of LLMs, we use the Llama-3.1-8B-Instruct Dubey et al. (2024) model as the judge, asking it to score the amount of concept present in the model's response. Prompts used with the judge model and specifics of the scoring is given in 6.4. We also test how much generations based on additional concepts  $c_i$  change after steering is applied to the model. To measure this difference, we calculate FID Heusel et al. (2017) (for diffusion models) and BERT score Zhang et al. (2020) (for LLMs) between the original generations and generations of the steered model. Higher FID value and lower BERT precision scores correspond to more severe generation alteration. For LLMs we also calculate the BERT scores on steered MMLU Hendrycks et al. (2021) generations, which are present in the supplementary.

In the case of ideal erasure, the method should have a low amount of concept  $c_s$  present in the generation (low CS scores on  $c_s$  prompts), and keep all the generations on prompts unrelated to  $c_s$  intact (high CS and better FID/BERT scores on  $c_i$  prompts).

We steer each model with vanilla steering and LEACE with different values of  $\beta$ . Results for Llama-2-7b are presented in Fig. 2. Results for all tested LLMs can be found in the Supplementary Sec. 6.7 and Sec. 6.11 and diffusion models can be found in Sec. 6.9 and Sec. 6.13. It can be clearly seen, that LEACE achieves much better balance between erasure of unwanted concept and preservation of other concepts across different values of  $\beta$ .



(a) CS score on the concept being erased vs CS scores (b) CS score of the concept being erased vs BERT of additional concepts

scores of additional concepts

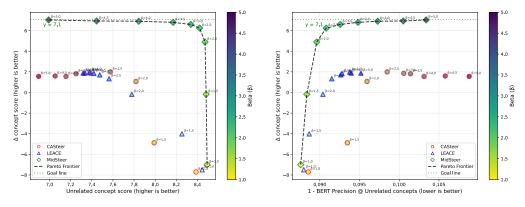
Figure 2: Pareto efficiency frontiers for LLM *erasure* experiments with vanilla steering and LEACE with different values of  $\beta$ . The model used is Llama2 7B.

#### 4.2.2 Concept flipping

In this section, we compare the performances of vanilla steering, LEACE and MidSteer on the task of concept flipping in both LLMs and diffusion models.

In this scenario, we aim to switch one concept in the model's responses to another (e.g. make model output text about cats when prompted to generate text about dogs), while minimally affecting anything else in the generated output.

We adapt the testing procedure described for concept erasure (Sec. 4.2.1) to the concept flipping task. In the case of flipping a concept  $c_1$  to  $c_2$ , our aim is not only to erase concept  $c_1$  from generations, but also add  $c_2$  to them. So we define  $\Delta CS$  as a difference between levels of  $c_2$  and  $c_1$  in the generation. Note that the definition of CS for unrelated concepts remains unchanged.



(a) Difference in CS scores of the source and target (b) Difference in CS scores of the source and target concepts vs CS scores of additional concepts

concepts vs BERT scores of additional concepts

Figure 3: Pareto efficiency frontiers for LLM *flipping* experiments with vanilla steering and LEACE with different values of  $\beta$ . The model used is Llama2 7B.

We steer each model with vanilla steering, LEACE and MidSteer with different values of  $\beta$ . Results for Llama-2-7b are presented in Fig. 3. Results for all tested LLMs can be found in Sec. 6.8 and Sec. 6.12 and diffusion models can be found in Sec. 6.10 and Sec. 6.14. It can be clearly seen, that MidSteer achieves much better balance between level of flipping between desired concepts and preservation of other concepts across different values of  $\beta$ .

To better illustrate differences between vanilla steering, LEACE and MidSteer for concept flipping, in Tab. 1 we also present results on switching a concept of  $c_1$  = "horse" to  $c_2$  = "motorcycle" on the

Table 1: Results on SDXL when flipping from "horse" to "motorcycle". Reported are CLIP-scores (cs) and FID for target and non-target concepts.

		ho	rse n	notorcyc	ele co	ow p	ig d	og legis	slator
method	strength	src-cs	tgt-cs src-cs	tgt-cs	fid cs	fid   cs	fid   cs	fid cs	fid
orig	-	71.0	49.1   51.8	70.7	-   72.7	-   71.8	-   66.3	-   60.8	-
CASteer	2.0	52.1	69.5   68.3	52.9	212.4   70.9	42.7   71.9	18.9   66.1	28.6   60.9	24.6
LEACE	2.0	51.2	68.8   67.6	53.3	207.6   72.2	25.2   71.7	12.6   66.1	20.8   60.6	28.2
MiDSteer (ours)	1.0	51.2	68.7   51.9	70.7	12.7   72.2	23.9   71.8	12.4   66.1	20.7   60.7	27.2



Figure 4: Qualitative results on switching to steer "horses" source concept into target "motorcycles". While all methods similarly successfully performed switching from "horse" to "motorcycle", CASteer and LEACE failed when presented with prompt for the target concept ("motorcycle"), since they do not distinguish between forward and reverse steering. CASteer also additionally failed on "cow" concept, and more significantly altered images of concept "dog"

SDXL model. We compare CASteer and LEACE with  $\beta=2$  and MidSteer with  $\beta=1$ , as these are default parameters for these methods as suggested by Eqs. 8, 29, and 18. The full table with results on all values of  $\beta$  can be found in the supplementary. First note, that all the methods successfully flip "horse" to "motorcycle", having similar CS scores on source ("horse") and target ("motorcycle") concepts. Second, it can be seen that as suggested by definitions Eqs. 8, 29, vanilla steering (CASteer) and LEACE fail to keep the "motorcycle" concept intact when flipping "horse" to "motorcycle", as target CS score goes down. In contrast, MidSteer keeps "motorcycles" intact. This is also illustrated in 4. Next, CS score of "cow" and FID scores of "cow", "pig" and "dog" are worse for CASteer than for other models, showing superiority of LEACE and MidSteer over vanilla steering in ability to keep unrelated concepts intact. Results on other concepts flipping on both LLMs and diffusion models show similar patterns.

#### 5 CONCLUSION

In this work, we bridge the gap between previous empirical research in steering generative models and the theory of affine concept steering. We extend this theoretical framework to concept switching. We define the corresponding optimisation problem and solve it in closed form.

We present MiDSteer, a universal steering method, that is theoretically optimal under certain conditions. It outperforms other methods on concept switching for both LLMs and image diffusion models, while having the advantage of clear matrix form representation. To our knowledge, this is the first theoretical treatment of steering beyond deletion, connecting empirical heuristics and principled affine methods.

# REFERENCES

486

487

488

489

490

491 492

493

494 495

496

497

498

499

500

501

504

505

506

507

509

510

511

512 513

514

515

516

517

518

519

520

521

522 523

524

525

526

527

528

529

530 531

532

534

536

- Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *CoRR*, abs/2502.17601, 2025. doi: 10.48550/ARXIV.2502.17601. URL https://doi.org/10.48550/arXiv.2502.17601.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form, 2025. URL https://arxiv.org/abs/2306.03819.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. CoRR, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https: //doi.org/10.48550/arXiv.2407.21783.
- Tatiana Gaintseva, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Steering diffusion models for controllable generation, 2025. URL https://arxiv.org/abs/2503.09630.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.
- 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.
- Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. *CoRR*, abs/2506.10703, 2025. doi: 10.48550/ARXIV.2506.10703. URL https://doi.org/10.48550/arXiv.2506.10703.
- Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have A semantic latent space. In *ICLR*, 2023.
- 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.
- Dawn Lu and Nina Rimsky. Investigating bias representations in llama 2 chat via activation steering. *CoRR*, abs/2402.00402, 2024. doi: 10.48550/ARXIV.2402.00402. URL https://doi.org/10.48550/arXiv.2402.00402.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *CoRR*, abs/2307.06435, 2023. doi: 10.48550/ARXIV.2307.06435. URL https://doi.org/10.48550/arXiv.2307.06435.

- 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.
- 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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.* OpenReview.net, 2024. URL https://openreview.net/forum?id=di52zR8xgf.
- Nate Rahn, Pierluca D'Oro, and Marc G. Bellemare. Controlling large language model agents with entropic activation steering. *CoRR*, abs/2406.00244, 2024. doi: 10.48550/ARXIV.2406.00244. URL https://doi.org/10.48550/arXiv.2406.00244.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *ACL 2020*, *Online*, *July 5-10*, 2020, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.647. URL https://doi.org/10.18653/v1/2020.acl-main.647.
- Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9413–9431, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023. acl-long.523. URL https://aclanthology.org/2023.acl-long.523/.
- Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9413–9431, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023. acl-long.523. URL https://aclanthology.org/2023.acl-long.523/.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.828. URL https://doi.org/10.18653/v1/2024.acl-long.828.
- Marco Scialanga, Thibault Laugel, Vincent Grari, and Marcin Detyniecki. SAKE: steering activations for knowledge editing. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *ACL 2025*, *Vienna, Austria, July 27 August 1, 2025*, pp. 15966–15978. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.777/.
- Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024.
- Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024.*OpenReview.net, 2024. URL https://openreview.net/forum?id=GwA4go0Mw4.
- Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. Steering without side effects: Improving post-deployment control of language models. *CoRR*, abs/2406.15518, 2024. doi: 10.48550/ARXIV.2406.15518. URL https://doi.org/10.48550/arXiv.2406.15518.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford\_alpaca, 2023.

 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

- Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.
- Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023. doi: 10.48550/ARXIV.2308.10248. URL https://doi.org/10.48550/arXiv.2308.10248.
- Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. ACM, 2025. doi: 10.1145/3696410.3714640. URL https://doi.org/10.1145/3696410.3714640.
- 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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025.* OpenReview.net, 2025. URL https://openreview.net/forum?id=N80j1XhtYZ.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.1515.
- Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4):105:1–105:39, 2024b. doi: 10.1145/3626235. URL https://doi.org/10.1145/3626235.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.
- Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL https://doi.org/10.48550/arXiv.2310.01405.

# 6 APPENDIX

#### 6.1 Incorporating steering into model weights

Recall that the last layer of self-attention block in LLMs or cross-attention block in SDXL/SANA is a 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 SA/CA block of LLM/SDXL/SANA,  $h_{in}$  and  $h_{out}$  are input and output to that layer,  $h_{out}$  being the final output of SA/CA layer.

In this case, by combining last layer of SA/CA block with matrix formulation of steering/LEACE/MidSteer68929, we can incorporate the transformation directly into weights of the model, by multiplying weight matrix of the last layer of SA/CA block with  $I-ss^T$  matrix from of steering/LEACE/MidSteer:

$$h_{out} = (I - ss^{T})W_{proj\_out}h_{in} = W_{proj\_out}^{s}h_{in}$$
(25)

 $W^s_{proj\_out}$  is a matrix of the same size as  $W_{proj\_out}$ . This results in having zero inference overhead compares to original LLM/SDXL/SANA models.

#### 6.2 THEOREM PROOFS

**Theorem 2.** Let X be a standardized random vector in  $\mathbb{R}^d$ , i.e. it has zero mean E[x] = 0 and unit covariance matrix  $\Sigma_{XX} = I$ . Let  $C \in \{0,1\}$  be a concept indicator variable. Let  $f_{delete}$  be defined as in equation 5 and s be defined as in equation 3. Then  $f_{delete}$  as a function of h minimizes

$$\min_{f \in \text{Aff}(R^d \mapsto R^d)} E[\|f(X) - X\|^2] \text{ s.t. } Cov(f(X), C) = 0$$
 (26)

*Proof.* We have k = 1, Z = C, M = I. According to **Theorem 1**,  $f(X) = A^*X + b^*$ , where  $A^*$  is defined as in equation 9 and  $b^*$  is defined as in equation 10, minimizes equation 1.

We conclude b=0 since  $\mathrm{E}[X]=0$ . Further, it can be shown that  $W=\Sigma_{XX}^{-1/2}=I^{-1/2}=I$ ; hence, the transform f (that is optimal according to **Theorem 1**) simplifies to

$$f(X) = \left(I - \Sigma_{XZ} \Sigma_{XZ}^{+}\right) X. \tag{27}$$

Recall that we are working in k=1, so  $\Sigma_{XZ} \in \mathbb{R}^{d \times 1}$  is a column-vector. By definition of the Moore-Penrose inverse for column-vectors,

$$\Sigma_{XZ}^+ = \frac{\Sigma_{XZ}^T}{\|\Sigma_{XZ}\|^2} \,,$$

hence

$$f(X) = X - s s^T X = X - \langle X, s \rangle s$$

for  $s = \Sigma_{XZ}/\|\Sigma_{XZ}\|$ , which is equivalent to  $f_{delete}$  with  $\beta = 1$  and this particular choice of s. Hence,  $f_{delete}$  is the transformation that minimises equation 26 for  $\beta = 1$  and  $s = \Sigma_{XZ}/\|\Sigma_{XZ}\|$ .

**Theorem 3** (Optimal concept switching). Let  $X, Z, M, \Sigma_{XX}, \Sigma_{XZ}$  be defined as in Theorem 1. Let  $\mu = E[X]$ . Then the optimization problem

$$\min_{\substack{A \in \mathbb{R}^{d \times d}, \\ b \in \mathbb{R}^d}} \mathbb{E}\Big[ \|AX + b - X\|_M^2 \Big] \text{ s.t. } \operatorname{Cov}(AX + b, Z) = -\operatorname{Cov}(X, Z)$$
 (28)

has the following solution (almost surely):

$$A^* = I - 2 \cdot W^+(W\Sigma_{XZ})(W\Sigma_{XZ})^+ W \tag{29}$$

(30)

$$b^* = \mu - A^* \mu$$

, where  $W = \Sigma_{XX}^{-1/2}$  is ZCA whitening transform.

*Proof.* The next paragraph of the proof is adapted from Belrose et al. (2025).

 Consider first an orthonormal basis diagonalizing the inner product  $\mathbf{M}$ , so that  $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \sum_{i=1}^{d} \alpha_i x_i y_i$  for fixed  $\alpha_1, \ldots, \alpha_d \geq 0$ . This allows us to treat each row  $\mathbf{A_i} \in \mathbb{R}^d$  of  $\mathbf{A}$  as a separate optimization problem,

$$\underset{\substack{\mathbf{A_i} \in \mathbb{R}^d, \\ b_i \in \mathbb{R}}}{\operatorname{argmin}} \mathbb{E} \left[ \alpha_i \left( \mathbf{A_i}^T \mathbf{v} X + b_i - \mathbf{v} X_i \right)^2 \right] \quad \text{subject to } \operatorname{Cov}(\mathbf{A_i}^T \mathbf{v} X + b_i, \mathbf{v} Z) = -\operatorname{Cov}(\mathbf{v} X, \mathbf{v} Z),$$

at which point the weights  $\alpha_i$  of each subproblem become irrelevant, and our objective may as well be Euclidean. In the rest of the proof we will assume  $\mathbf{M} = I$ .

 The sketch for the rest of the proof will look like this:

1. Find necessary conditions for optimality using Lagrange multipliers method

2. Show that  $A^*, b^*$  satisfy the necessary conditions

 3. Show that optimisation problem is convex over linear constraints, and such, if a local solution exists, it is globally optimal and unique.

Let us formulate the Lagrangian. Here  $\Lambda \in \mathbb{R}^{d \times k}$ , because we have  $d \cdot k$  constraints on covariance matrix.

$$\mathcal{L}(A,b,\Lambda) = \frac{1}{2} \mathbb{E} \Big[ \|AX + b - X\|_2^2 \Big] + \langle \Lambda, \operatorname{Cov}(AX + b, Z) + \operatorname{Cov}(X, Z) \rangle_F =$$

$$\frac{1}{2} \mathbb{E} \Big[ (AX + b - X)^T (AX + b - X) \Big] + \operatorname{Tr} \Big( \Lambda^T (A + I) \Sigma_{XZ} \Big) =$$

$$\mathbb{E} \Big[ \frac{1}{2} X^T A^T A X + b^T A X - X^T A X - X^T b + \frac{1}{2} b^T b + \frac{1}{2} X^T X \Big] + \operatorname{Tr} \Big( \Lambda^T (A + I) \Sigma_{XZ} \Big)$$
(31)

The partial derivatives of the Lagrangian with respect to  $A, b, \Lambda$  are

$$\begin{split} \frac{\partial \mathcal{L}}{\partial A} &= \mathrm{E}[AXX^T + bX^T - XX^T] + \Lambda \Sigma_{XZ}^T \,, \\ &= A\mathrm{E}[XX^T] + b\mathrm{E}[X]^T - \mathrm{E}[XX^T] + \Lambda \Sigma_{XZ}^T \\ \frac{\partial \mathcal{L}}{\partial b} &= \mathrm{E}[AX + b - X] \,, \\ &= A\mathrm{E}[X] + b - \mathrm{E}[X] \,, \\ \frac{\partial \mathcal{L}}{\partial \Lambda} &= (A + I)\Sigma_{XZ} \,. \end{split}$$

Next, we use  $\mu = E(X)$  and  $E[XX^T] = \Sigma_{XX} + \mu \mu^T$  to formulate the necessary conditions

$$0 = \frac{\partial \mathcal{L}}{\partial A} = (A - I) \left( \Sigma_{XX} + \mu \mu^T \right) + b \mu^T + \Lambda \Sigma_{XZ}^T,$$
 (32)

$$0 = \frac{\partial \mathcal{L}}{\partial b} = A\mu + b - \mu \,, \tag{33}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \Lambda} = (A + I)\Sigma_{XZ}. \tag{34}$$

We note that the optimal  $b^*$  as defined in equation 30 satisfies 33. Plugging equation 33 in equation 32 leads to

$$(A - I) \left( \Sigma_{XX} + \mu \mu^{T} \right) + (\mu - A\mu)\mu^{T} + \Lambda \Sigma_{XZ}^{T},$$

$$= A\Sigma_{XX} - \Sigma_{XX} + A\mu\mu^{T} - \mu\mu^{T} + \mu\mu^{T} - A\mu\mu^{T} + \Lambda \Sigma_{XZ}^{T},$$

$$= (A - I)\Sigma_{XX} + \Lambda \Sigma_{XZ}^{T} = 0.$$
(35)

Now let us check that  $A^*$  satisfies 34 and 35. By plugging  $A^*$  into 34 we get

$$0 = (2I - 2W^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}W)\Sigma_{XZ},$$

$$= 2\Sigma_{XZ} - 2W^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}(W\Sigma_{XZ}),$$

$$= 2(\Sigma_{XZ} - W^{+}W\Sigma_{XZ}),$$

$$= 2(\Sigma_{XZ} - (I - P_{\mathcal{N}(W)})\Sigma_{XZ}),$$

$$= 2P_{\mathcal{N}(W)}\Sigma_{XZ},$$
(36)

because Moore-Penrose inverses  $B^+$  of B satisfy  $BB^+B=B$  and  $B^+B=I-P_{\mathcal{B}}$ . Here  $P_{\mathcal{B}}$  denotes the orthogonal projection onto the nullspace  $\mathcal{N}(B)$  of B. Since the columns of  $\Sigma_{XZ}$  always lie within the image of  $\Sigma_{XX}$  (which is the orthogonal complement of the kernel of  $\Sigma_{XX}$ , which is also the kernel of W), we can conclude that equation 36 is always satisfied.

Plugging  $A^*$  into 35 we observe

$$-2 \cdot (W^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}W)\Sigma_{XX} + \Lambda\Sigma_{XZ}^{T} = -2 \cdot W^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}W^{+} + \Lambda\Sigma_{XZ}^{T} = 0$$
 (37)

The identity  $W\Sigma_{XX}=W^+$  holds because  $\Sigma_{XX}$  is symmetric p.s.d., so  $\Sigma_{XX}=UDU^T$  and  $\Sigma_{XX}^{-1/2}\Sigma_{XX}=UD^{-1/2}U^TUDU^T=UD^{1/2}U^T=\Sigma_{XX}^{1/2}$  for some orthogonal U and non-negative diagonal D, and because  $D^{-1/2}$  ignores zero diagonal values.

Next, multiplying equation 37 by W from both sides leads to

$$\begin{split} &-2WW^{+}(W\Sigma_{XZ})(W\Sigma_{XZ})^{+}W^{+}W + W\Lambda\Sigma_{XZ}^{T}W = \\ &-2(W\Sigma_{XZ})(W\Sigma_{XZ})^{+} + W\Lambda(W\Sigma_{XZ})^{T} = -2\Sigma_{WX,Z}\Sigma_{WX,Z}^{+} + \Lambda_{W}\Sigma_{WX,Z}^{T} = 0 \\ &\qquad \qquad \text{(almost surely)} \end{split}$$

, where again  $WW^+ = W^+W = I$  on a subspace covered by X, and thus, almost surely.

Recall that any real matrix has a singular value decomposition, and thus, for orthogonal U, V and diagonal D:

$$\Sigma_{WX,Z} = UDV^T \tag{38}$$

$$\Sigma_{WX,Z}^{+} = VD^{+}U^{T} \tag{39}$$

Plugging this into the equation, it becomes:

$$-2UDV^TVD^+U^T + \Lambda_WVDU^T = -2UD^+DU^T + \Lambda_WVDU^T = 0$$

, which is satisfied for  $\Lambda_W=2UD^+V^T=2(\Sigma_{WX,Z}^+)^T$ . This also makes sense, because  $\Sigma_{WX,Z}\Sigma_{WX,Z}^+$  is an orthogonal projection matrix, and is thus symmetric, so

$$(\Sigma_{WX,Z}^+)^T \Sigma_{WX,Z}^T = (\Sigma_{WX,Z} \Sigma_{WX,Z}^+)^T = \Sigma_{WX,Z} \Sigma_{WX,Z}^+$$

, which again proves the Lagrange conditions for partial derivative w.r.t. A.

 Thus we have shown that the said optimisation problem has a local solution. But because the constraint is linear in A, and it follows from the triangle inequality that  $\|\cdot\|$  is convex, the local optimum is actually the global minimum.

**Theorem 4.** Let X be a standardized random vector in  $\mathbb{R}^d$ , i.e. it has zero mean E[x]=0 and unit covariance matrix  $\Sigma_{XX}=I$ . Let  $C\in\{0,1\}$  be a concept indicator variable and P(C=1)+P(C=0)=1. Let  $f_{switch}$  be defined as in equation 7 and s be defined as in equation 2. Then  $f_{switch}$  as a function of h minimizes

$$\min_{f \in \text{Aff}(R^d \mapsto R^d)} \text{E}[\|f(X) - X\|^2] \text{ s.t. } \text{Cov}(f(X), C) = -\text{Cov}(X, C)$$

$$\tag{40}$$

*Proof.* Let k = 1, Z = C, M = I. According to **Theorem 3**,  $f(X) = A^*X + b^*$ , where  $A^*$  is defined in equation 29 and  $b^*$  is defined in equation 30 minimizes equation 28.

b=0 since  $\mathrm{E}[X]=0$ . Also it can be shown  $W=\Sigma_{XX}^{-1/2}=I^{-1/2}=I$ , so the transform becomes:

$$f(X) = \left(I - 2 \cdot \Sigma_{XZ} \Sigma_{XZ}^{+}\right) X \tag{41}$$

Recall that we are working in k = 1, so  $\Sigma_{XZ} \in \mathbb{R}^{d \times 1}$  is a column-vector. So

$$\Sigma_{XZ} = \text{Cov}(X, Z) = E[XZ] - E[X] \cdot E[Z] = E[X \cdot 1 | Z = 1] \cdot P(Z = 1) + E[X \cdot 0 | Z = 0] \cdot P(Z = 0) - E[X] \cdot P(Z = 1) = P(Z = 1) \cdot \left(E[X | Z = 1] - E[X]\right)$$

Now recall that P(Z=1) + P(Z=0) = 1, so

$$\Sigma_{XZ} = P(Z=1) \cdot \left( \mathbb{E}[X|Z=1] - \mathbb{E}[X] \right) =$$

$$P(Z=1) \cdot \left( \mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=1] P(Z=1) - \mathbb{E}[X|Z=0] P(Z=0) \right) =$$

$$P(Z=1) \cdot P(Z=0) \cdot \left( \mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0] \right)$$

, which is equal to s up to normalization constant. By definition of Moore-Penrose inverse for column-vectors,

$$\Sigma_{XZ}^+ = \frac{\Sigma_{XZ}^T}{\|\Sigma_{XZ}\|^2}$$

, so

$$f(X) = X - 2 \cdot ss^{T}X = X - 2 \cdot s(s^{T}X) = X - 2 \cdot (s^{T}X)s = X - 2 \cdot \langle X, s \rangle s$$

, which is  $f_{switch}$  with  $\beta = 2$ .

**Theorem 5** (Affine optimal concept steering). Let X, M, Z be defined as in theorem 3 and also k=2l. Let  $Z=(Z_1,Z_2)$ , where  $Z_1,Z_2\in\{0,1\}^l$ . Let  $\Sigma_{XZ_i}=Cov(X,Z_i), i\in\{1,2\}$  be the cross-covariance matrices between X and  $Z_i$ , with  $\Sigma_{XZ_1}$  having full column rank:  $\operatorname{rk}\left(\Sigma_{XZ_1}\right)=l$ .

Let W be the ZCA whitening transform  $\Sigma_{XX}^{-1/2}$ . Let  $\Sigma_{WX,Z_i} = W\Sigma_{XZ_i}$ , let  $\mu = E[X]$ .

Then we have the following optimization problem:

$$\min_{\substack{A \in \mathbb{R}^{d \times d} \\ b \in \mathbb{P}^d}} \mathrm{E}[\|AX + b - X\|_M^2] \text{ s.t. } Cov(AX + b, Z_1) = Cov(X, Z_2)$$

$$(42)$$

which has the following solution:

$$A^* = I + W^+ (\Sigma_{WX, Z_2} - \Sigma_{WX, Z_1}) \Sigma_{WX, Z_2}^+ W$$
(43)

$$b^* = \mu - A^* \mu \tag{44}$$

(45)

*Proof.* We will use the same method as previous theorem to prove this. Indeed, the objective is same, and thus convex. The constraint is still linear:

$$Cov(Ax + b, Z_1) = A\Sigma_{XZ_1} = \Sigma_{XZ_2} = Cov(X, Z_2)$$

$$\tag{46}$$

So let us define the Lagrangian, where  $\Lambda \in \mathbb{R}^{d \times l}$ :

$$\mathcal{L}(A,b,\Lambda) = \frac{1}{2} \mathbf{E} \Big[ (AX + b - X)^T (AX + b - X) \Big] + Tr \Big( \Lambda^T (A\Sigma_{XZ_1} - \Sigma_{XZ_2}) \Big)$$
(47)

The derivatives w.r.t. parameters are the following:

$$\frac{\partial \mathcal{L}}{\partial A} = (A - I)(\Sigma_{XX} + \mu \mu^T) + b\mu^T + \Lambda \Sigma_{XZ_1}^T$$
 = 0 (48)

$$\frac{\partial \mathcal{L}}{\partial b} = A\mu - \mu + b \tag{49}$$

$$\frac{\partial \mathcal{L}}{\partial \Lambda} = A \Sigma_{XZ_1} - \Sigma_{XZ_2}$$
 = 0 (50)

Trivially  $b^*$  satisfies equation 49 for suitable  $A^*$ .

Let us see that equation 50 is satisfied. We can plug  $A^*$  and then multiply by W on the left, to get:

$$\begin{split} W\Sigma_{XZ_1} + WW^+ & (\Sigma_{WX,Z_2} - \Sigma_{WX,Z_1})\Sigma_{WX,Z_1}^+ W\Sigma_{XZ_1} - W\Sigma_{XZ_2} = \\ & \Sigma_{WX,Z_1} - \Sigma_{WX,Z_1}\Sigma_{WX,Z_1}^+ \Sigma_{WX,Z_1} + \Sigma_{WX,Z_2}\Sigma_{WX,Z_1}^+ \Sigma_{WX,Z_1} - \Sigma_{WX,Z_2} = \\ & \Sigma_{WX,Z_2} & (\Sigma_{WX,Z_1}^T \Sigma_{WX,Z_1})^{-1}\Sigma_{WX,Z_1}^T \Sigma_{WX,Z_1} - \Sigma_{WX,Z_2} = 0 \quad \text{(almost surely)} \end{split}$$

Here we used  $YY^+Y = Y$  for any Y and  $Y^+ = (Y^TY)^{-1}Y^T$  for Y which columns are linearly independent.

Next, plugging equation 49 into equation 48 we get:

$$(A - I)\Sigma_{XX} + \Lambda \Sigma_{XZ_1}^T = 0 (51)$$

Let us now proceed to show that for  $A^*$  there exists  $\Lambda \in \mathbb{R}^{d \times l}$  so this equality holds. After plugging in  $A^*$  and using previously shown fact  $W\Sigma_{XX} = W^+$ :

$$W^{+}(\Sigma_{WX,Z_{2}} - \Sigma_{WX,Z_{1}})\Sigma_{WX,Z_{1}}^{+}W^{+} + \Lambda\Sigma_{XZ_{1}}^{T} = 0$$
(52)

Again, multiplying by W on both sides and recalling that W is symmetric we get:

923 924

927

928 929

930 931

932 933

934 935 936

937 938

939

940 941

942 943 944

945 946

947 948

949 950

951 952

953 954

955

957 958 959

960 961

962 963

964 965

966 967

968 969 970

971

$$(\Sigma_{WX,Z_2} - \Sigma_{WX,Z_1})\Sigma_{WX,Z_1}^+ + \Lambda_W \Sigma_{WX,Z_1}^T = 0 \qquad \qquad \text{(almost surely)}$$

Now,  $\Sigma_{WX,Z_1}$  is also full column rank, so  $\Sigma_{WX,Z_1}^+ = \left(\Sigma_{WX,Z_1}^T \Sigma_{WX,Z_1}\right)^{-1} \Sigma_{WX,Z_1}^T$ . Thus,  $\Lambda_W = -\left(\Sigma_{WX,Z_2} - \Sigma_{WX,Z_1}\right)\left(\Sigma_{WX,Z_1}^T \Sigma_{WX,Z_1}\right)^{-1}$  satisfies the equation.

# EXAMPLE PROMPTS FOR CLASS-CONDITIONAL STEERING VECTORS

To obtain class-conditional means for LLMs, we constructed 1000 prompts across several categories. In each case, last token activation was used to compute steering vectors, before first token of model output and after corresponding chat template tokens. Below we provide illustrative subsets of the prompts used:

# Horse-related prompts

- How did horses evolve from their early ancestors?
- What factors contributed to the domestication of horses?
- What anatomical features enable a horse to run at high speeds?

# Motorcycle-related prompts

- What was the significance of the earliest gasoline powered motorcycle in the evolution of personal transportation?
- How does the engine cycle mechanism influence power delivery and efficiency in a motorbike?
- What are the primary considerations when selecting the appropriate tire for varying riding conditions?

# Dog-related prompts

- What is believed to be the origin of dog domestication?
- How do dogs communicate through body language?
- Which genetic factors influence coat color in dogs?

#### Cat-related prompts

- What is the origin of domestic cats and how did they become companion animals?
- How did ancient Egyptians incorporate cats into their society and religion?
- Why did cats spread globally alongside human trade routes?

# 6.4 PROMPT FOR CONCEPT SCORING

To assess the quality of steering, we compute a *concept score* that gauges the presence of a specific concept in generated text. The score ranges from 0 to 10:

- A score of 0 indicates the text is completely irrelevant to the concept.
- A score of 10 indicates strong and explicit relevance to the concept.
- Scores from 1 through 9 represent gradations of partial relevance.

The following prompt was used to query the model for concept scoring:

Score the text on a scale of 0 to 10, where the score reflects the extent to which it demonstrates relevance to the concept of {concept}. A score of 0 indicates the text is completely irrelevant and shows no connection to {concept}, whereas a score of 10 signifies strong and explicit relevance to {concept}. For texts that suggest varying degrees of relevance, use the intermediate numbers (1 through 9) to represent the gradation of connection to {concept}. Output just a single number.

Text to evaluate: ''' {text}

982 983 984

972

973

975

976

977

978

979

980

981

The exact score is obtained by doing a forward pass of Llama3.1-7B model and taking argmax across token probabilities corresponding to score tokens (from 0 to 10).

985986987

# 6.5 TEMPLATE PROMPTS FOR STEERING QUALITY ASSESSMENT

988 989

990

991

992

To evaluate the effectiveness of steering, we constructed a set of template prompts. In each case, the { } placeholder was substituted with the concept under consideration (as defined elsewhere). Each template was used to generate multiple model responses. These responses were then evaluated either by a judge model (for text/LLM outputs) or by CLIP score (for images), to compute the *concept score* as well as related metrics such as FID and BERTScore.

993 994 995

996

**LLM Assessment Prompts.** The following template prompts were used for large language model (LLM) evaluation. The placeholder { } was substituted with the target concept.

```
997
       1.
           Write a short story about a {}
998
           Write a poem about a {}
999
       3.
           What is the history of {}
1000
       4.
           What is the most famous {}?
1001
       5.
           What is the most expensive {}?
       6.
          How to make a {}?
1002
       7.
          How to maintain a {}?
1003
       8.
           How to dispose of a {}?
1004
       9.
           How to store a {}?
1005
      10.
          How to transport a {}?
1006
      11.
          How to use a {}?
1007
      12.
          What is important to know about {}?
1008
      13.
          How to tell age of a {}?
1009
      14.
          What is a {}?
1010
      15.
          What types of {} are there?
1011
      16.
           What are the most common {}?
1012
      17.
           Describe an appearance of {} in detail.
      18.
           How does {} look like?
1013
      19.
          How does {} sound like?
1014
           How does {} feel like?
      20.
1015
      21.
           How does {} behave like?
1016
      22.
           What is the purpose of {}?
1017
      23.
           What are the main components of a {}?
1018
      24.
           How to identify a {}?
1019
      25.
          Where can you find a {}?
1020
      26.
           What are the dangers of a {}?
1021
      27.
          What tools do you need for a {}?
1022
      28.
          How much does a {} typically cost?
1023
      29.
           What are alternatives to a {}?
      30.
           How to choose a good {}?
1024
      31.
           What are common problems with a {}?
1025
      32.
           How long does a {} typically last?
```

```
33.
           What size is a typical {}?
1027
      34.
           How to clean a {}?
1028
      35.
           What skills are needed to handle a {}?
1029
      36.
           What are the benefits of having a {}?
           How has {} changed over time?
1030
      37.
1031
      38.
          What cultures use {} the most?
      39.
           How to test if a {} is working properly?
1032
      40.
          What safety precautions are needed for a {}?
1033
           How to upgrade or improve a {}?
      41.
1034
      42.
           How does weather affect a {}?
1035
      43.
           What are the environmental impacts of a {}?
1036
      44.
           How to measure the quality of a {}?
1037
      45.
           What accessories go with a {}?
1038
      46.
           How to protect a {} from damage?
1039
      47.
           What are myths about {}?
1040
      48.
           How to teach someone about a {}?
1041
      49.
           What industries use {}?
1042
      50.
          How is a {} different from similar things?
      51.
          What are the legal considerations for owning a {}?
1043
      52.
           How to pack a {} for moving?
1044
      53.
          What are seasonal considerations for a {}?
1045
      54. How to customize a {}?
1046
      55. What are expert tips for using a {}?
1047
      56.
           How to troubleshoot issues with a {}?
1048
      57.
           What is the lifecycle of a {}?
1049
      58.
           How to estimate the value of a {}?
1050
      59.
           What are cultural significances of a {}?
1051
      60.
           How to take a picture of a {}?
           How to make a sculpture of a {}?
1052
      61.
1053
      62.
           What is the future of {}?
      63.
           How to draw a {}?
1054
      64.
          When was {} first mentioned in human history?
1055
      65.
          Can one ride a {}?
1056
      66. Write a song about {}
1057
      67.
          Define a {}
1058
          Write a positive review on a book about {}
      68.
1059
      69.
           Write a negative review on a book about {}
1060
      70.
           Do people make toys of {}?
1061
      71.
           How is {} used in the economy?
1062
      72.
           Write an abstract for a science paper about {}
1063
      73.
           How does temperature affect a {}?
1064
      74.
           What are the origins of the word {}?
      75.
           What are superstitions about {}?
1065
      76.
           How to simulate a {} digitally?
1066
      77.
           What are the physics of a {}?
1067
      78.
           How to teach children about {}?
1068
      79.
           What are famous artworks featuring {}?
1069
      80.
           What are the nutritional aspects of a {}?
1070
           Describe the most famous {} competitions.
1071
1072
      Image Assessment Prompts. The following template prompts were used for image model evalua-
1073
      tion. The placeholder { } was substituted with the target concept.
1074
1075
           a bad photo of a {}.
       1.
1076
       2.
           a photo of many {}.
1077
          a sculpture of a {}.
       4.
          a photo of the hard to see {}.
1078
       5. a low resolution photo of the {}.
1079
           a rendering of a {}.
```

```
graffiti of a {}.
1081
           a bad photo of the {}.
       9.
           a cropped photo of the {}.
1083
      10.
           a tattoo of a {}.
1084
      11.
           the embroidered {}.
      12.
           a photo of a hard to see {}.
1085
      13.
           a bright photo of a {}.
1086
      14.
           a photo of a clean {}.
1087
           a photo of a dirty {}.
      15.
1088
      16.
           a dark photo of the {}.
1089
      17.
           a drawing of a {}.
1090
      18.
           a photo of my {}.
1091
      19.
           the plastic {}.
1092
      20.
           a photo of the cool {}.
1093
      21.
           a close-up photo of a {}.
1094
      22.
           a black and white photo of the {}.
1095
      23.
           a painting of the {}.
1096
      24.
           a painting of a {}.
      25.
           a pixelated photo of the {}.
1097
      26.
           a sculpture of the {}.
1098
      27.
           a bright photo of the {}.
1099
      28.
           a cropped photo of a {}.
1100
      29.
           a plastic {}.
1101
      30.
           a photo of the dirty {}.
1102
      31.
           a jpeg corrupted photo of a {}.
1103
      32.
           a blurry photo of the {}.
1104
      33.
           a photo of the {}.
1105
      34.
           a good photo of the {}.
1106
      35.
           a rendering of the {}.
1107
      36.
           a {} in a video game.
      37.
           a photo of one {}.
1108
      38.
           a doodle of a {}.
1109
      39.
           a close-up photo of the {}.
1110
      40.
           a photo of a {}.
1111
      41.
           the origami {}.
1112
      42.
           the {} in a video game.
1113
      43.
           a sketch of a {}.
1114
      44.
           a doodle of the {}.
1115
      45.
           a origami {}.
1116
      46.
           a low resolution photo of a {}.
1117
      47.
           the toy {}.
1118
      48.
           a rendition of the {}.
1119
      49.
           a photo of the clean {}.
      50.
           a photo of a large {}.
      51.
          a rendition of a {}.
1121
      52.
           a photo of a nice {}.
1122
      53.
           a photo of a weird {}.
1123
      54.
           a blurry photo of a {}.
1124
      55.
          a cartoon {}.
1125
      56.
           art of a {}.
1126
      57.
           a sketch of the {}.
1127
      58.
           a embroidered {}.
1128
      59.
           a pixelated photo of a {}.
1129
      60.
           itap of the {}.
      61.
           a jpeg corrupted photo of the {}.
1130
      62.
           a good photo of a {}.
1131
      63.
           a plushie {}.
1132
      64.
           a photo of the nice {}.
1133
      65.
           a photo of the small {}.
```

```
1134
      66.
            a photo of the weird {}.
1135
      67.
            the cartoon {}.
1136
      68.
            art of the {}.
            a drawing of the {}.
1137
      69.
1138
      70.
            a photo of the large {}.
      71.
            a black and white photo of a {}.
1139
      72.
            the plushie {}.
1140
      73.
            a dark photo of a {}.
1141
      74.
            itap of a {}.
1142
      75.
            graffiti of the {}.
1143
      76.
            a toy {}.
1144
            itap of my {}.
      77.
1145
      78.
            a photo of a cool {}.
1146
      79.
            a photo of a small {}.
1147
      80.
            a tattoo of the {}.
1148
```

1150 1151

1152

1153

1154

1155

1156

1157

1158

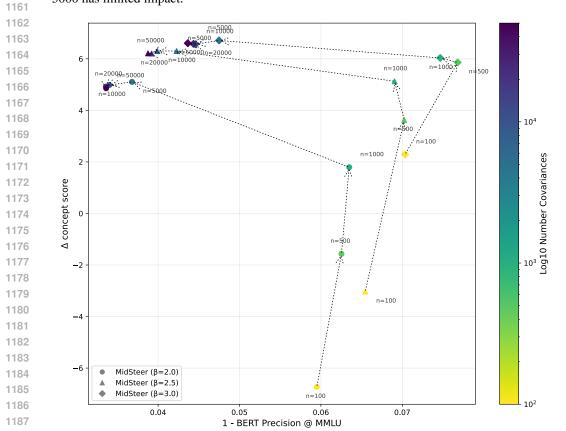
1159

1160

#### 6.6 Number of prompts for covariance calculation

To find the optimal number of prompts used to calculate unconditional covariances  $\Sigma_{XX}$  for concept switching, we perform the following ablation study. For each number of prompts used for covariances generation from the set  $\{100, 500, 1000, 5000, 10000, 20000\}$  we run the same base experiment as outlined in 4.2.2. We then compute  $\Delta CS$  and 1 - BERT Precision @ MMLU metrics on a small set of MiDSteer steering strengths (to show if the steering strength can affect the optimal number of prompts). For LLM experiment we used LLama2-7B model.

We then plot these values of a 2D plane similar to Pareto charts, but this time varying the number of prompts instead. This in essence forms a curve that, after a certain threshold, settles in a small region of metric space. As can be seen from the chart below, increasing the number of prompts used beyond 5000 has limited impact.



# 6.7 RESULTS FOR LLM CONCEPT ERASURE

We tested concept erasure on LLama2-7b, Qwen2.5-7b and Qwen2.5-14b models. In case of smaller models, both clipped and unclipped variant (matrix compliant form) of the steering manipulation were tested.

Table 2: Model llama2-7b (noclip), removal of horses

		horses	moto	orcycles	cow	'S	pigs	S	dog	s	legi	slators
		cs	cs	bertp	cs	bertp	cs	bertp	cs	bertp	cs	bertp
method	strength			_								
No Steering	-	8.6	8.5	-	8.4	-	8.5	-	8.7	-	8.4	-
CASteer	1.0	8.6	8.5	0.91	8.4	0.90	8.4	0.90	8.6	0.90	8.5	0.89
	2.0	8.4	8.5	0.90	8.4	0.89	8.3	0.89	8.5	0.90	8.3	0.89
	3.0	1.5	3.4	0.80	2.4	0.78	1.4	0.78	2.9	0.79	2.6	0.79
	4.0	0.4	0.0	0.76	0.1	0.76	0.0	0.77	0.0	0.77	0.0	0.77
	5.0	1.0	0.0	0.76	0.1	0.77	0.0	0.77	0.0	0.77	0.0	0.77
LEACE	1.0	8.5	8.5	0.91	8.3	0.91	8.4	0.90	8.6	0.91	8.3	0.90
	2.0	5.0	8.3	0.89	7.3	0.87	7.4	0.88	8.0	0.89	7.8	0.88
	3.0	0.2	4.6	0.86	3.3	0.85	3.7	0.85	4.0	0.86	4.6	0.86
	4.0	0.0	1.0	0.82	0.2	0.81	0.2	0.81	0.4	0.82	1.1	0.82
	5.0	0.0	0.1	0.81	0.0	0.81	0.0	0.81	0.1	0.81	0.2	0.81

Table 3: Model llama2-7b (noclip), removal of dogs

		dogs	cats		wol	ves	cow	'S	pigs	3	legi	slators
		cs	cs	bertp	cs	bertp	cs	bertp	cs	bertp	cs	bertp
method	strength											
No Steering	-	8.6	8.6	-	8.5	-	8.4	-	8.4	-	8.5	-
CASteer	1.0	8.6	8.5	0.90	8.4	0.90	8.4	0.90	8.4	0.90	8.5	0.90
	2.0	8.5	8.5	0.90	8.3	0.89	8.3	0.89	8.3	0.89	8.4	0.89
	3.0	2.1	3.1	0.79	3.6	0.80	3.0	0.79	1.9	0.78	3.4	0.79
	4.0	0.1	0.0	0.76	0.1	0.76	0.0	0.75	0.0	0.75	0.0	0.76
	5.0	0.3	0.0	0.76	0.0	0.75	0.0	0.76	0.0	0.76	0.0	0.75
LEACE	1.0	8.5	8.5	0.91	8.4	0.91	8.3	0.91	8.4	0.91	8.4	0.90
	2.0	6.5	7.8	0.88	7.6	0.88	7.2	0.88	7.5	0.88	7.7	0.88
	3.0	1.3	3.8	0.86	4.2	0.86	3.4	0.85	3.7	0.85	4.6	0.86
	4.0	0.1	0.4	0.82	0.2	0.83	0.5	0.82	0.3	0.82	1.0	0.82
	5.0	0.0	0.1	0.81	0.1	0.82	0.1	0.82	0.1	0.81	0.1	0.81

Table 4: Model qwen2.5-7b (noclip), removal of horses

		horses	mote	orcycles	cow	'S	pigs	S	dog	s	legi	slators
		cs	cs	bertp	cs	bertp	cs	bertp	cs	bertp	cs	bertp
method	strength			_								
No Steering	-	8.7	8.6	-	8.4	-	8.4	-	8.7	-	8.5	-
CASteer	1.0	8.6	8.5	0.89	8.4	0.88	8.4	0.88	8.7	0.88	8.6	0.88
	2.0	8.0	7.8	0.84	8.1	0.84	7.8	0.83	8.1	0.83	8.1	0.83
	3.0	3.6	4.0	0.72	5.1	0.73	4.5	0.73	4.7	0.73	3.9	0.72
	4.0	1.2	2.8	0.69	2.6	0.70	1.9	0.70	2.6	0.70	2.0	0.69
	5.0	0.6	1.0	0.70	0.7	0.70	0.3	0.70	0.7	0.70	0.9	0.69
LEACE	1.0	8.5	8.5	0.89	8.4	0.89	8.4	0.89	8.7	0.89	8.5	0.89
	2.0	6.9	8.5	0.89	8.4	0.88	8.4	0.88	8.7	0.88	8.5	0.88
	3.0	0.6	8.5	0.87	8.3	0.85	8.3	0.85	8.6	0.86	8.4	0.86
	4.0	0.2	8.3	0.85	8.1	0.83	8.1	0.82	8.4	0.83	8.2	0.83
	5.0	0.0	7.3	0.76	6.7	0.73	6.2	0.73	6.9	0.73	7.2	0.74

Table 5: Model qwen2.5-7b (noclip), removal of dogs

		dogs	cats		wol	ves	cow		pigs		legi	slators
		cs	cs	bertp	cs	bertp	cs	bertp	cs	bertp	cs	bertp
method	strength											
No Steering	-	8.7	8.5	-	8.3	-	8.4	-	8.4	-	8.5	-
CASteer	1.0	8.6	8.5	0.88	8.3	0.88	8.4	0.88	8.4	0.88	8.5	0.88
	2.0	8.0	8.1	0.84	8.0	0.84	8.1	0.84	7.9	0.84	8.3	0.84
	3.0	3.9	5.0	0.74	4.6	0.74	5.4	0.75	4.8	0.75	4.6	0.75
	4.0	1.4	2.2	0.71	1.7	0.71	2.4	0.71	1.6	0.71	2.1	0.71
	5.0	0.3	0.7	0.71	0.4	0.71	0.7	0.71	0.4	0.72	0.8	0.71
LEACE	1.0	8.7	8.5	0.89	8.4	0.88	8.4	0.89	8.4	0.88	8.5	0.89
	2.0	8.7	8.5	0.88	8.3	0.87	8.4	0.88	8.4	0.87	8.5	0.88
	3.0	7.8	8.5	0.85	8.2	0.84	8.3	0.85	8.3	0.85	8.4	0.86
	4.0	3.2	8.2	0.83	7.7	0.79	8.1	0.83	7.9	0.81	8.2	0.82
	5.0	0.6	6.5	0.72	5.6	0.71	6.6	0.72	5.9	0.72	6.9	0.72

Table 6: Model qwen2.5-14b (noclip), removal of horses

		horses	moto	orcycles	cow	'S	pigs	S	dog	s	legi	slators
		cs	cs	bertp	cs	bertp	cs	bertp	cs	bertp	cs	bertp
method	strength											
No Steering	-	8.7	8.6	-	8.4	-	8.4	-	8.7	-	8.5	-
CASteer	1.0	8.7	8.6	0.90	8.4	0.90	8.5	0.89	8.7	0.90	8.5	0.89
	2.0	8.7	8.6	0.89	8.4	0.89	8.5	0.89	8.7	0.89	8.5	0.89
	3.0	8.5	8.5	0.86	8.4	0.86	8.4	0.86	8.6	0.86	8.4	0.87
	4.0	2.9	4.2	0.78	5.0	0.78	5.0	0.78	4.3	0.79	4.3	0.79
	5.0	0.8	1.5	0.73	1.2	0.73	2.0	0.74	1.4	0.74	0.6	0.73
LEACE	1.0	8.7	8.6	0.90	8.4	0.89	8.4	0.88	8.7	0.89	8.5	0.88
	2.0	8.0	8.5	0.88	8.3	0.87	8.3	0.86	8.7	0.87	8.4	0.86
	3.0	3.0	8.5	0.85	8.3	0.84	8.3	0.83	8.6	0.84	8.3	0.83
	4.0	1.2	8.3	0.82	8.1	0.81	8.0	0.80	8.4	0.81	8.1	0.80
	5.0	0.3	7.8	0.79	7.5	0.79	7.5	0.78	7.9	0.79	7.7	0.79

Table 7: Model qwen2.5-14b (noclip), removal of dogs

		dogs	cats		wol	ves	cow	'S	pigs	S	legi	slators
		cs	cs	bertp	cs	bertp	cs	bertp	cs	bertp	cs	bertp
method	strength											
No Steering	-	8.7	8.5	-	8.4	-	8.4	-	8.5	-	8.5	-
CASteer	1.0	8.7	8.5	0.90	8.4	0.89	8.5	0.89	8.5	0.89	8.6	0.89
	2.0	8.7	8.5	0.89	8.4	0.88	8.4	0.89	8.5	0.89	8.5	0.89
	3.0	8.4	8.5	0.86	8.3	0.86	8.4	0.86	8.4	0.86	8.4	0.87
	4.0	3.5	3.7	0.79	3.1	0.77	4.2	0.78	4.2	0.78	3.4	0.79
	5.0	1.9	0.9	0.73	1.6	0.73	0.9	0.73	1.2	0.73	0.4	0.73
LEACE	1.0	8.7	8.5	0.89	8.4	0.88	8.4	0.89	8.4	0.89	8.5	0.89
	2.0	8.7	8.4	0.87	8.4	0.86	8.3	0.87	8.4	0.86	8.4	0.86
	3.0	7.7	8.4	0.84	8.2	0.83	8.3	0.84	8.2	0.83	8.3	0.83
	4.0	4.7	8.3	0.82	7.9	0.81	8.2	0.82	8.0	0.80	8.1	0.80
	5.0	2.3	7.9	0.79	7.5	0.79	7.7	0.80	7.5	0.78	7.7	0.79

# 6.8 RESULTS FOR LLM CONCEPT SWITCHING

We tested concept switching on LLama2-7b, Qwen2.5-7b and Qwen2.5-14b models. In case of smaller models, both clipped and unclipped variant (matrix compliant form) of the steering manipulation were tested.

Table 8: Model llama2-7b (noclip), flipping from dogs to cats

		dogs		cats			wol	ves			pigs				cow	'S			legi	slators		
		src-cs	tgt-cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp
method	strength		-		-	_			-				-	_							-	_
CASteer	1.0	8.6	1.1	2.9	7.5	0.91		7.8	1.5	0.91	8.5	1.5	0.4	0.91	8.4	0.7	0.3	0.91	8.5	0.3	0.2	0.90
	1.5	8.1	1.9	6.4	3.8	0.90	8.5	7.8	1.5	0.90	8.5	1.5	0.4	0.91	8.4	0.6	0.3		8.5	0.4	0.2	0.90
	2.0	5.9	3.9	7.1	2.8	0.89	8.4	7.8	1.4	0.90	8.5	1.5	0.4		8.4	0.7	0.3	0.90		0.3	0.2	0.90
	2.5	5.0	4.4	7.0	3.0	0.89	8.5	7.7	1.4	0.90	8.4	1.4	0.4	0.90	8.4	0.7	0.3		8.4	0.4	0.2	0.89
	3.0	4.8	4.6	6.6	2.8	0.89	8.5	7.7	1.4	0.90	8.4	1.4	0.4	0.90	8.4	0.7	0.3	0.90	8.5	0.3	0.2	0.89
	3.5	4.6	4.7	6.5	3.0	0.89	8.4	7.7	1.4	0.90	8.5	1.4	0.3	0.90	8.3	0.9	0.3	0.90	8.5	0.4	0.2	0.90
	4.0	4.6	4.7	6.5	2.9	0.89	8.4	7.7	1.5	0.90	8.4	1.4	0.4	0.90	8.2	0.9	0.3	0.90	8.4	0.4	0.2	0.89
	4.5	4.5	4.9	6.5	2.9	0.89	8.3	7.6	1.5	0.90	8.3	1.5	0.5		8.2	1.0	0.4		8.4	0.4	0.2	0.89
	5.0	4.4	4.9	6.5	2.7	0.88	8.3	7.6	1.5	0.89	8.2	1.5	0.6	0.89	7.9	1.2	0.5	0.89	8.2	0.4	0.2	0.89
LEACE	1.0	8.5	1.4	2.1	8.2	0.91	8.4	7.8	1.5	0.91	8.5	1.5	0.5	0.91	8.4	0.7	0.3	0.91		0.3	0.2	0.90
	1.5	5.4	4.5	3.4	6.6	0.91	8.4	7.8	1.5	0.91	8.5	1.4	0.4	0.91	8.4	0.7	0.3	0.91	8.5	0.3	0.1	0.90
	2.0	4.0	5.5	4.9	4.8	0.91	8.4	7.7	1.7	0.91	8.5	1.4	0.4	0.91	8.4	0.6	0.3	0.91	8.5	0.3	0.2	0.90
	2.5	3.8	5.6	5.3	4.1	0.90	8.4	7.7	1.5	0.91	8.5	1.5	0.5	0.91	8.4	0.7	0.3	0.91	8.5	0.3	0.2	0.90
	3.0	3.7	5.6	5.6	4.1	0.90	8.4	7.7	1.6	0.91	8.5	1.5	0.4	0.91	8.4	0.7	0.3	0.91	8.5	0.3	0.2	0.90
	3.5	3.7	5.7	5.5	3.9	0.90	8.4	7.7	1.5	0.91	8.4	1.5	0.5	0.91	8.4	0.7	0.3		8.5	0.3	0.2	0.90
	4.0	3.6	5.6	5.5	3.9	0.90	8.4	7.6	1.6	0.91	8.5	1.5	0.4	0.91	8.4	0.7	0.3	0.91	8.5	0.3	0.2	0.90
	4.5	3.6	5.6	5.5	3.9	0.90	8.4	7.7	1.6	0.91	8.5	1.4	0.4	0.91	8.4	0.7	0.3		8.5	0.3	0.2	0.90
	5.0	3.6	5.5	5.4	3.9	0.90	8.4	7.7	1.6	0.91	8.4	1.5	0.5	0.91	8.3	0.7	0.3	0.91	8.5	0.3	0.1	0.90
mean_matching	1.0	8.0	1.8	1.6	8.6	0.92	8.4	7.8	1.6	0.91	8.5	1.4	0.4	0.91	8.4	0.7	0.3	0.91	8.5	0.3	0.1	0.91
-	1.5	4.1	6.1	1.5	8.6	0.91	8.5	7.8	1.6	0.91	8.4	1.5	0.5	0.91	8.4	0.6	0.3	0.91	8.5	0.4	0.2	0.90
	2.0	2.6	7.7	1.5	8.6	0.91	8.4	7.7	1.7	0.91	8.5	1.5	0.5	0.91	8.4	0.6	0.3	0.91	8.5	0.3	0.2	0.90
	2.5	2.1	8.2	1.5	8.6	0.91	8.3	7.5	1.8	0.91	8.4	1.4	0.5	0.91	8.2	0.7	0.4	0.91	8.4	0.3	0.2	0.90
	3.0	1.9	8.4	1.4	8.7	0.91	7.9	7.0	2.2	0.91	8.3	1.5	0.8		8.0	0.7	0.8	0.91	8.5	0.3	0.2	0.90
	3.5	1.7	8.5	1.3	8.7	0.91	7.4	6.5	3.2	0.90	7.8	1.6	1.5	0.90	7.7	0.9	1.4	0.91	8.4	0.3	0.2	0.90
	4.0	1.6	8.5	1.4	8.6	0.90	6.3	5.6	4.2	0.90	7.3	1.7	2.3	0.90	6.8	1.1	2.3		8.4	0.4	0.4	0.90
	4.5	1.6	8.6	1.3	8.7	0.90	5.4	4.8	5.1	0.89	6.5	1.8	3.2	0.90	5.8	1.2	3.8	0.90	8.3	0.5	0.9	0.90
	5.0	1.6	8.6	1.4	8.7	0.90	4.6	4.1	5.8	0.89	5.5	2.0	4.4	0.89	4.6	1.2	4.8	0.89	8.1	0.6	1.4	0.90

Table 9: Model llama2-7b (noclip), flipping from horses to motorcycles

		horses		motor	cycles		cow	/S			pigs	;			dog	s			legi	slators		
		src-cs	tgt-cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp
method	strength										ļ											
CASteer	1.0	8.6	0.7	1.3	8.5	0.92	8.4	1.7	0.1	0.91	8.5	1.2	0.1	0.91	8.6	1.5	0.2	0.92	8.5	0.3	0.3	0.90
	1.5	6.5	2.9	1.4	8.5	0.92	8.4	1.7	0.1	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.2	0.91	8.5	0.3	0.3	0.90
	2.0	2.7	6.9	2.1	7.9	0.91	8.3	1.6	0.1	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.2	0.91	8.5	0.3	0.3	0.90
	2.5	2.6	7.2	3.7	5.5	0.91	8.1	1.4	0.2	0.90	8.3	1.2	0.1	0.91	8.5	1.5	0.2	0.91	8.5	0.3	0.3	0.90
	3.0	2.9	6.8	4.4	4.2	0.90	7.8	1.3	0.3	0.90	8.2	1.2	0.1	0.90	8.6	1.4	0.2	0.91	8.5	0.3	0.2	0.90
	3.5	3.0	6.5	4.7	3.6	0.90	7.2	1.3	0.3	0.90	8.1	1.1	0.1	0.90	8.5	1.3	0.2	0.91	8.4	0.3	0.3	0.90
	4.0	3.0	6.1	4.6	3.4	0.89	6.9	1.1	0.4	0.90	8.1	1.1	0.2	0.90	8.3	1.4	0.3	0.90	8.5	0.3	0.3	0.90
	4.5	3.1	5.8	5.0	3.2	0.89	6.5	1.1	0.4	0.89	7.9	1.1	0.2	0.90	8.2	1.4	0.3	0.90	8.5	0.4	0.4	0.90
	5.0	3.0	5.7	4.8	3.1	0.89	6.4	1.1	0.5	0.89	7.7	1.0	0.3	0.89	8.0	1.3	0.3	0.90	8.4	0.4	0.4	0.90
LEACE	1.0	8.5	0.7	1.2	8.5	0.92	8.4	1.7	0.1	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.1	0.92	8.5	0.3	0.3	0.90
	1.5	8.2	1.0	1.7	8.2		8.4	1.6	0.2	0.91	8.4	1.3	0.1	0.91	8.6	1.5	0.2	0.92	8.5	0.4	0.3	0.90
	2.0	5.4	3.5	4.0	5.4		8.3	1.7	0.2	0.91	8.4	1.3	0.1	0.91	8.6	1.5	0.2	0.92	8.5	0.3	0.2	0.90
	2.5	4.0	4.9	4.9	3.9	0.90	8.3	1.6	0.2	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.2	0.91	8.5	0.3	0.2	0.91
	3.0	3.7	5.2	5.5	3.2	0.90	8.2	1.5	0.2	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.2	0.91	8.5	0.3	0.3	0.90
	3.5	3.6	5.3	5.4	3.0	0.90	8.1	1.4	0.2	0.91	8.3	1.2	0.1	0.91	8.6	1.4	0.2	0.91	8.5	0.3	0.2	0.90
	4.0	3.5	5.5	5.5	2.9	0.90	8.0	1.3	0.2	0.90	8.3	1.1	0.1	0.91	8.6	1.5	0.2	0.91	8.5	0.4	0.3	0.90
	4.5	3.4	5.2	5.5	2.8	0.90	7.7	1.2	0.2	0.90	8.2	1.2	0.1	0.91	8.5	1.5	0.2	0.91	8.5	0.3	0.2	0.90
	5.0	3.4	5.3	5.6	2.9	0.90	7.6	1.3	0.3	0.90	8.2	1.1	0.1	0.91	8.5	1.5	0.2		8.5	0.4	0.3	0.90
mean_matching	1.0	8.6	0.7	1.2	8.5	0.92		1.6	0.1	0.91	8.5	1.2	0.1	0.91	8.6	1.4	0.2	0.92	8.5	0.3	0.3	0.90
	1.5	5.9	3.5	1.2	8.5		8.4	1.6	0.1	0.91	8.4	1.2	0.1	0.91	8.6	1.6	0.2	0.92	8.4	0.3	0.3	0.90
	2.0	2.6	7.2	1.2	8.5		8.3	1.5	0.1	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.2	0.91	8.4	0.3	0.3	0.90
	2.5	1.8	8.2	1.2	8.5		8.3	1.6	0.2	0.91	8.4	1.2	0.1	0.91	8.6	1.5	0.2	0.91	8.5	0.3	0.3	0.90
	3.0	1.7	8.4	1.2	8.5	0.92	8.3	1.6	0.2	0.91	8.2	1.1	0.1	0.91	8.6	1.6	0.2	0.91	8.5	0.3	0.3	0.90
	3.5	1.6	8.4	1.2	8.5		8.0	1.4	0.4	0.90		1.0	0.2	0.90	8.5	1.5	0.2	0.91	8.4	0.4	0.3	0.90
	4.0	1.6	8.4	1.2	8.5	0.91	7.6	1.3	0.7	0.90		1.1	0.4		8.4	1.4	0.2	0.91	8.4	0.3	0.3	0.90
	4.5	1.6	8.5	1.2	8.5	0.91	7.0	1.3	1.4	0.90		1.0	0.6		8.3	1.4	0.4		8.5	0.3	0.3	0.90
	5.0	1.4	8.5	1.2	8.5	0.91	6.1	1.2	2.2	0.89	7.3	0.9	0.9	0.90	8.2	1.4	0.5	0.90	8.4	0.3	0.3	0.90

Table 10: Model qwen2.5-7b (noclip), flipping from dogs to cats

		dogs		cats			wol	ves			pigs				cow	'S			legi	slators		
		src-cs	tgt-cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp			tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp
method	strength																					
CASteer	1.0	5.6	6.9	1.9	8.4	0.89	8.3	7.6	2.0	0.89	8.4	1.6	0.6	0.89	8.4	0.8	0.4	0.89	8.5	0.3	0.2	0.89
	2.0	3.2	7.5	2.2	7.8	0.88	7.6	6.4	3.7	0.88	8.1	1.6	1.9	0.88	7.9	1.1	1.8	0.89	8.4	0.5	0.9	0.89
	3.0	3.4	7.5	2.3	7.6	0.87	7.3	6.2	3.9	0.87	7.2	1.7	3.0	0.88	7.0	1.4	3.0	0.88	8.4	0.7	2.0	0.88
	4.0	3.0	7.1	2.4	7.2	0.86	7.2	6.2	3.7	0.87	7.4	1.6	2.2	0.87	7.1	1.3	2.8	0.87	8.3	0.6	1.6	0.88
	5.0	2.9	6.8	2.4	6.6	0.85	5.9	5.2	4.3	0.85	7.3	1.6	1.7	0.86	7.5	1.2	1.8	0.86	8.3	0.5	1.0	0.87
LEACE	1.0	8.7	1.0	1.8	8.5	0.89	8.3	7.7	1.6	0.89	8.4	1.6	0.4	0.89	8.4	0.9	0.3	0.89	8.5	0.3	0.2	0.89
	2.0	7.4	4.2	6.0	6.7	0.88	8.3	7.6	1.6	0.89	8.4	1.4	0.4	0.89	8.4	0.9	0.3	0.89	8.5	0.3	0.2	0.89
	3.0	6.3	6.7	7.1	5.5	0.88	8.3	7.6	1.7	0.89	8.4	1.6	0.4	0.89	8.4	0.9	0.3	0.89	8.5	0.3	0.2	0.89
	4.0	5.8	6.6	6.9	4.7	0.87	8.3	7.5	1.7	0.89	8.4	1.5	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.3	0.2	0.89
	5.0	5.3	6.5	6.7	4.3	0.87	8.3	7.6	1.8	0.89	8.4	1.5	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.4	0.2	0.89
mean_matching	1.0	8.7	1.0	1.7	8.5	0.89	8.3	7.6	1.6	0.89	8.5	1.5	0.5	0.89	8.4	0.8	0.3	0.89	8.5	0.4	0.2	0.89
	2.0	7.3	5.0	1.7	8.5	0.89	8.4	7.6	1.7	0.89	8.5	1.5	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.3	0.2	0.89
	3.0	5.7	7.7	1.7	8.5	0.89	8.3	7.5	2.0	0.89	8.4	1.4	0.6	0.89	8.4	0.8	0.5	0.89	8.5	0.3	0.2	0.89
	4.0	5.0	8.2	1.7	8.5	0.89	8.1	7.4	2.5	0.89	8.3	1.6	1.0	0.89	8.3	1.0	0.7	0.89	8.5	0.4	0.3	0.89
	5.0	4.3	8.3	1.7	8.5	0.89	7.8	6.9	3.2	0.89	8.3	1.7	1.1	0.89	8.2	1.1	1.0	0.89	8.4	0.4	0.3	0.89

Table 11: Model qwen2.5-7b (noclip), flipping from horses to motorcycles

		horses		motore	cycles		cow	'S			pigs				dog	s			legi	slators		
		src-cs	tgt-cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp
method	strength				-				-					_			-					_
CASteer	1.0	8.6	0.8	1.2	8.4	0.90	8.4	1.8	0.2	0.89	8.4	1.1	0.1	0.89	8.6	1.4	0.3	0.89	8.5	0.4	0.4	0.89
	2.0	4.3	5.7	5.7	4.4	0.87	8.2	1.6	0.3	0.89	8.3	0.9	0.1	0.88	8.6	1.3	0.2	0.89	8.5	0.4	0.3	0.89
	3.0	4.1	5.5	6.1	3.2	0.86	7.4	1.3	0.8	0.88	8.0	0.8	0.3	0.87	8.5	1.3	0.3	0.88	8.5	0.3	0.3	0.89
	4.0	4.2	5.2	6.2	2.8	0.86	6.6	1.5	1.3	0.86	7.5	0.9	0.4	0.87	8.1	1.3	0.4	0.88	8.4	0.3	0.3	0.89
	5.0	3.8	5.3	6.4	2.9	0.85	6.3	1.4	1.2	0.86	7.1	0.9	0.6	0.86	7.8	1.3	0.5	0.87	8.4	0.3	0.3	0.88
LEACE	1.0	8.6	0.7	1.0	8.4	0.90	8.4	1.9	0.2	0.89	8.4	1.2	0.1	0.89	8.7	1.5	0.2	0.89	8.5	0.4	0.3	0.89
	2.0	4.4	6.2	2.8	7.7	0.89	8.4	1.8	0.2	0.89	8.4	1.1	0.1	0.89	8.7	1.5	0.2	0.89	8.6	0.4	0.3	0.89
	3.0	3.1	6.7	5.3	5.0	0.88	8.3	1.7	0.2	0.89	8.4	1.0	0.1	0.89	8.7	1.4	0.2	0.89	8.5	0.4	0.3	0.89
	4.0	3.2	6.3	5.5	4.1	0.87	8.2	1.6	0.4	0.89	8.3	1.0	0.1	0.89	8.6	1.4	0.2	0.89	8.5	0.4	0.4	0.89
	5.0	3.2	6.1	5.6	3.8	0.87	7.9	1.5	0.7	0.88	8.3	0.9	0.2	0.88	8.6	1.3	0.2	0.89	8.5	0.4	0.4	0.89
mean_matching	1.0	8.6	0.8	0.9	8.5	0.90	8.3	1.9	0.2	0.89	8.4	1.1	0.1	0.89	8.7	1.5	0.2	0.89	8.5	0.3	0.3	0.89
_	2.0	2.9	7.1	0.8	8.5	0.90	8.4	1.8	0.2	0.89	8.4	1.1	0.1	0.89	8.7	1.4	0.2	0.89	8.5	0.3	0.3	0.89
	3.0	1.4	8.2	0.9	8.5	0.90	8.2	1.7	0.3	0.89	8.4	1.0	0.2	0.89	8.6	1.4	0.3	0.89	8.5	0.3	0.3	0.89
	4.0	1.3	8.4	0.8	8.5	0.89	7.7	1.7	1.0	0.89	8.2	0.9	0.3	0.88	8.6	1.2	0.3	0.89	8.4	0.3	0.4	0.89
	5.0	1.2	8.4	0.9	8.6	0.90	6.3	1.5	2.8	0.88	7.8	1.0	1.0	0.88	8.2	1.4	0.8	0.89	8.5	0.3	0.4	0.89

Table 12: Model qwen2.5-14b (noclip), flipping from dogs to cats

		dogs		cats			wol	ves			pigs				cow	'S			legi	slators		
		src-cs	tgt-cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp
method	strength		-		-	_			-	_			-	_			-				-	_
CASteer	1.0	8.7	0.9	2.7	8.2	0.89	8.4	7.7	1.3	0.89	8.5	1.4	0.4	0.89	8.4	0.8	0.3	0.90	8.5	0.3	0.2	0.89
	2.0	7.8	4.5	7.3	4.9	0.86	8.4	7.7	1.4	0.89	8.5	1.4	0.4	0.89	8.4	0.8	0.2	0.89	8.5	0.3	0.2	0.89
	3.0	7.0	5.9	7.3	4.3	0.86	8.4	7.6	1.3	0.89	8.5	1.3	0.3	0.89	8.4	0.8	0.3	0.89	8.5	0.4	0.2	0.89
	4.0	6.8	6.2	7.2	4.6	0.86	8.4	7.6	1.3	0.89	8.5	1.3	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.4	0.2	0.89
	5.0	6.8	6.1	7.3	5.1	0.86	8.3	7.5	1.3	0.89	8.5	1.3	0.4	0.89	8.3	0.8	0.2	0.89	8.5	0.4	0.2	0.89
LEACE	1.0	8.7	1.0	1.9	8.4	0.89	8.4	7.7	1.3	0.89	8.5	1.5	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.3	0.2	0.89
	2.0	7.9	4.2	6.0	7.3	0.88	8.4	7.7	1.3	0.89	8.5	1.4	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.3	0.2	0.89
	3.0	6.9	6.5	7.0	5.4	0.86	8.4	7.6	1.4	0.89	8.5	1.3	0.3	0.89	8.4	0.8	0.3	0.90	8.5	0.3	0.1	0.89
	4.0	6.2	6.7	6.9	4.2	0.86	8.4	7.7	1.4	0.89	8.5	1.4	0.4	0.89	8.4	0.8	0.3	0.90	8.5	0.3	0.2	0.89
	5.0	6.0	6.6	6.9	3.7	0.85	8.4	7.7	1.4	0.89	8.4	1.4	0.4	0.89	8.4	0.8	0.2	0.89	8.5	0.3	0.2	0.89
mean_matching	1.0	8.7	1.2	1.8	8.5	0.89	8.4	7.7	1.4	0.89	8.5	1.5	0.4	0.89	8.5	0.8	0.3	0.89	8.5	0.3	0.1	0.89
	2.0	7.1	6.4	1.6	8.5	0.89	8.4	7.7	1.5	0.89	8.5	1.4	0.4	0.89	8.4	0.8	0.3	0.89	8.5	0.4	0.2	0.89
	3.0	4.8	7.8	1.6	8.6	0.89	8.4	7.6	1.8	0.89	8.4	1.5	0.5	0.89	8.4	0.7	0.3	0.89	8.5	0.3	0.2	0.89
	4.0	3.3	8.2	1.6	8.5	0.89	8.3	7.5	2.3	0.89	8.4	1.6	0.8	0.89	8.4	0.9	0.5	0.89	8.5	0.4	0.2	0.89
	5.0	2.1	8.3	1.6	8.6	0.89	8.1	7.0	2.9	0.89	8.3	1.8	1.2	0.89	8.3	0.9	0.7	0.89	8.5	0.4	0.3	0.89

Table 13: Model qwen2.5-14b (noclip), flipping from horses to motorcycles

		horses		motoro	cycles		cow	'S			pigs				dog	s			legi	slators		
		src-cs	tgt-cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp	cs	src-cs	tgt-cs	bertp
method	strength										j											
CASteer	1.0	8.7	0.7	1.2	8.6	0.90	8.4	1.7	0.2	0.89	8.4	1.1	0.1	0.89	8.7	1.4	0.2	0.90	8.5	0.3	0.3	0.89
	2.0	6.4	4.4	4.9	6.0	0.87	8.3	1.5	0.3	0.89	8.4	1.0	0.1	0.89	8.6	1.4	0.2	0.89	8.4	0.3	0.3	0.89
	3.0	5.9	5.0	6.9	2.9	0.84	7.9	1.4	0.5	0.88	8.4	0.9	0.2	0.88	8.6	1.3	0.2	0.89	8.5	0.3	0.3	0.89
	4.0	6.0	4.4	7.3	1.8	0.84	7.6	1.4	0.6	0.87	8.1	0.9	0.3	0.87	8.5	1.4	0.3	0.88	8.5	0.3	0.3	0.89
	5.0	5.5	4.3	7.4	1.6	0.83	7.5	1.3	0.5	0.86	7.9	0.9	0.3	0.86	8.3	1.4	0.3	0.88	8.5	0.3	0.3	0.89
LEACE	1.0	8.6	0.7	1.1	8.6	0.90	8.4	1.8	0.2	0.90	8.5	1.1	0.1	0.89	8.7	1.4	0.2	0.90	8.5	0.4	0.3	0.89
	2.0	6.4	4.6	3.8	7.5	0.88	8.4	1.6	0.2	0.89	8.4	1.0	0.1	0.89	8.7	1.3	0.2	0.90	8.5	0.4	0.4	0.89
	3.0	4.2	6.0	6.1	4.6	0.85	8.2	1.5	0.4	0.89	8.4	1.0	0.1	0.89	8.7	1.4	0.2	0.89	8.5	0.3	0.3	0.89
	4.0	4.2	6.1	6.6	3.5	0.84	7.9	1.5	0.6	0.89	8.3	1.0	0.2	0.88	8.7	1.4	0.3	0.89	8.5	0.4	0.4	0.89
	5.0	4.1	5.9	6.8	3.1	0.83	7.7	1.4	0.7	0.88	8.3	1.0	0.2	0.88	8.6	1.3	0.4	0.89	8.5	0.4	0.3	0.89
mean_matching	1.0	8.7	0.7	0.9	8.6	0.90	8.4	1.8	0.1	0.89	8.5	1.1	0.1	0.89	8.7	1.4	0.2	0.90	8.5	0.4	0.3	0.89
-	2.0	3.9	6.5	0.9	8.6	0.90	8.3	1.6	0.2	0.89	8.4	1.1	0.1	0.89	8.7	1.3	0.2	0.90	8.5	0.3	0.3	0.89
	3.0	1.4	8.2	0.9	8.6	0.90	8.3	1.5	0.2	0.89	8.4	1.0	0.1	0.89	8.7	1.4	0.3	0.89	8.5	0.3	0.3	0.89
	4.0	1.2	8.2	0.9	8.6	0.90	8.3	1.5	0.2	0.89	8.4	1.0	0.2	0.89	8.7	1.3	0.3	0.89	8.4	0.4	0.4	0.89
-	5.0	1.2	8.2	0.8	8.6	0.90	8.1	1.6	0.7	0.89	8.3	0.9	0.4	0.89	8.6	1.3	0.3	0.89	8.5	0.3	0.3	0.89

# 6.9 RESULTS FOR IMAGE DIFFUSION CONCEPT ERASURE

Concept erasure was tested on SDXL and SANA 1.6B models. Both clipped and unclipped variants of the steering function were tested.

Table 14: Model sdxl (noclip), removal of snoopy

		snoopy	mick	ev	pikac	hu	spons	gebob	dog		legis	lator
		cs	cs	fid	cs	fid	cs	fid	cs	fid	cs	fid
method	strength											
No Steering	-	74.3	73.1	-	72.6	-	75.1	-	66.3	-	60.8	-
CASteer	1.0	55.8	70.1	54.9	72.5	30.3	73.9	50.9	66.2	30.6	60.9	22.6
	1.5	49.9	67.9	71.8	72.5	39.9	72.8	66.0	66.2	39.4	60.9	27.5
	2.0	47.0	65.2	90.5	72.6	51.2	71.0	85.2	66.2	48.1	60.8	31.7
	2.5	45.6	62.2	111.0	72.5	65.7	68.6	109.4	66.1	58.1	60.9	35.3
	3.0	45.3	58.8	132.1	72.2	83.7	65.3	138.2	66.2	68.0	60.8	38.7
	4.0	45.3	53.5	169.0	71.6	123.4	59.0	189.5	66.1	83.3	60.9	45.8
	5.0	45.9	50.7	195.3	69.3	153.0	55.7	218.2	65.6	99.9	61.0	52.9
LEACE	1.0	56.7	72.2	35.7	72.9	21.3	74.1	42.2	66.3	20.7	60.6	26.9
	1.5	51.2	71.7	42.3	73.0	25.8	73.7	50.1	66.3	26.5	60.5	32.5
	2.0	48.3	71.0	48.2	73.2	29.6	73.3	57.8	66.3	31.6	60.4	36.4
	2.5	46.5	70.3	53.9	73.3	33.0	72.8	66.5	66.4	36.5	60.2	41.3
	3.0	45.8	69.6	59.8	73.5	36.7	72.1	75.3	66.4	40.8	60.0	46.5
	4.0	45.8	67.9	72.2	73.7	44.8	70.8	91.8	66.5	49.5	59.4	56.1
	5.0	47.0	66.1	85.9	73.7	53.8	69.4	114.1	66.4	57.1	58.6	69.2

Table 15: Model sdxl (noclip), removal of chihuahua

		chihuahua	muff	in	dog		wolf		cat		legisl	lator
		cs	cs	fid	cs	fid	cs	fid	cs	fid	cs	fid
method	strength											
No Steering	-	75.9	68.2	-	66.3	-	71.8	-	67.5	-	60.8	-
CASteer	1.0	54.6	68.1	19.7	65.0	58.2	72.5	25.9	67.0	35.2	60.9	22.7
	1.5	48.5	68.2	24.0	61.2	99.9	72.6	34.0	66.5	48.9	60.8	27.8
	2.0	47.6	68.0	27.3	54.1	155.5	72.6	44.1	64.6	69.0	60.9	31.9
	2.5	47.2	67.9	31.1	50.7	177.8	72.2	61.2	60.5	102.6	60.8	35.8
	3.0	46.9	67.9	34.6	49.7	187.7	70.0	96.2	55.8	141.5	60.8	39.4
	4.0	47.8	67.7	42.2	49.0	198.2	62.2	191.6	50.7	186.3	60.7	45.7
	5.0	49.7	67.6	49.5	48.9	209.3	57.8	228.4	49.4	201.5	60.7	52.1
LEACE	1.0	55.0	68.2	20.0	65.8	35.2	72.3	17.1	67.4	22.1	60.9	21.8
	1.5	48.5	68.1	25.0	65.5	47.6	72.5	21.3	67.3	27.1	60.8	27.0
	2.0	47.4	68.1	29.0	65.0	61.1	72.6	25.0	67.3	31.4	60.9	31.0
	2.5	47.0	68.2	32.6	64.1	74.8	72.8	28.6	67.2	35.1	60.8	34.2
	3.0	47.2	68.1	36.0	62.7	92.4	72.9	32.4	67.1	38.5	60.8	36.5
	4.0	48.6	68.0	42.3	57.6	131.4	73.1	39.4	66.9	45.1	60.7	41.9
	5.0	50.2	68.0	49.4	53.7	162.6	73.1	48.3	66.5	52.5	60.5	48.3

Table 16: Model sdxl (noclip), removal of horse

		horse	moto	rcycle	cow		pig		dog		legisl	lator
		cs	cs	fid	cs	fid	cs	fid	cs	fid	cs	fid
method	strength											
No Steering	-	71.0	70.7	-	72.7	-	71.8	-	66.3	-	60.8	-
CASteer	1.0	59.3	70.7	12.9	71.9	30.1	71.8	20.8	65.9	29.9	61.0	21.3
	1.5	49.8	70.7	15.6	71.2	46.6	71.8	27.6	65.8	36.5	61.1	26.2
	2.0	48.3	70.6	17.4	69.2	79.8	71.9	36.5	65.7	42.2	61.0	30.3
	2.5	47.9	70.7	19.5	62.1	152.5	72.0	45.9	65.4	48.4	60.9	33.6
	3.0	47.8	70.7	21.7	54.7	211.1	72.0	60.1	65.0	54.7	60.9	37.0
	4.0	48.0	70.7	26.9	51.1	227.9	71.8	92.3	63.9	69.4	60.8	43.3
	5.0	49.3	70.8	35.1	50.4	238.4	69.8	138.4	62.2	89.4	60.6	49.4
LEACE	1.0	57.1	70.6	11.5	72.3	20.5	71.8	11.6	66.1	19.7	60.7	25.1
	1.5	49.6	70.6	14.0	72.0	26.0	71.9	14.1	66.1	23.7	60.5	29.6
	2.0	48.4	70.6	15.9	71.8	33.1	71.9	16.2	66.0	28.0	60.4	34.1
	2.5	48.0	70.6	17.5	71.4	39.9	72.0	17.3	66.1	30.9	60.3	38.2
	3.0	48.1	70.6	19.3	70.7	53.5	72.0	19.1	66.1	33.6	60.3	41.6
	4.0	48.4	70.5	23.0	64.5	115.1	72.1	23.1	66.1	38.6	59.9	49.2
	5.0	49.7	70.3	27.4	56.4	197.4	72.2	27.8	66.1	44.6	59.4	58.8

Table 17: Model sana (noclip), removal of horse

		horse	moto	rcycle	cow		pig		dog		legisl	ator
		cs	cs	fid	cs	fid	cs	fid	cs	fid	cs	fid
method	strength											
No Steering	-	72.1	70.5	-	73.8	-	73.5	-	68.1	-	60.4	-
CASteer	1.0	70.8	70.1	21.3	74.1	36.4	73.7	28.2	67.8	29.2	60.2	21.2
	2.0	52.2	70.9	45.3	72.0	93.0	74.1	49.3	67.4	44.2	59.8	36.3
	3.0	51.3	69.3	105.0	61.5	216.9	65.9	261.3	65.5	71.9	59.3	58.4
	4.0	56.7	62.3	186.2	59.0	242.8	67.3	175.3	62.5	118.4	58.9	90.1
	5.0	52.5	60.4	221.0	58.7	249.9	65.0	205.0	59.3	161.3	58.4	130.3
LEACE	1.0	71.1	70.5	7.5	73.8	14.4	73.4	11.4	68.0	9.2	60.4	11.1
	2.0	52.0	70.4	10.0	73.9	20.9	73.5	16.7	68.0	14.2	60.4	15.6
	3.0	49.7	70.3	12.2	74.0	25.8	73.5	21.4	67.9	18.4	60.3	19.1
	4.0	48.8	70.4	13.8	74.2	30.9	73.5	25.6	67.8	21.3	60.2	21.8
	5.0	53.7	70.3	15.2	74.1	36.2	73.6	29.7	67.8	24.4	60.1	24.2

Table 18: Model sana (noclip), removal of snoopy

		snoopy	mick	ey	pikac	hu	spon	gebob	dog		legis	lator
		cs	cs	fid	cs	fid	cs	fid	cs	fid	cs	fid
method	strength											
No Steering	-	79.7	76.1	-	74.0	-	79.0	-	68.1	-	60.4	-
CASteer	1.0	60.6	75.3	64.0	74.1	41.3	79.0	43.9	68.0	42.1	60.8	23.3
	2.0	46.0	70.5	168.3	74.3	103.6	74.7	146.6	68.0	74.3	61.1	38.2
	3.0	42.4	64.2	189.7	72.0	164.1	63.4	222.2	67.7	100.9	60.8	55.3
	4.0	40.9	58.5	202.0	62.6	204.2	55.7	258.7	66.8	116.9	60.4	74.8
	5.0	40.9	55.4	208.1	55.5	231.6	52.9	276.5	65.1	127.6	60.0	94.8
LEACE	1.0	57.0	76.1	18.2	74.1	6.7	79.0	13.9	68.1	17.3	60.3	9.1
	2.0	44.8	76.2	30.5	74.1	11.7	78.9	19.3	68.1	25.3	60.2	13.6
	3.0	41.6	76.1	49.0	74.2	16.4	75.2	200.5	68.0	32.2	60.2	16.8
	4.0	40.9	75.6	73.4	74.2	21.3	78.7	29.0	68.0	38.0	60.1	19.5
	5.0	41.4	74.4	109.4	74.2	26.1	78.7	34.1	68.1	44.1	60.0	22.0

Table 19: Model sana (noclip), removal of chihuahua

		chihuahua	muffi	in	dog		wolf		cat		legisl	lator
		cs	cs	fid	cs	fid	cs	fid	cs	fid	cs	fid
method	strength											
No Steering	-	76.4	66.3	-	68.1	-	73.2	-	68.5	-	60.4	-
CASteer	1.0	75.6	66.6	19.8	67.4	49.4	73.6	25.6	68.3	32.5	55.3	268.2
	2.0	49.5	66.8	30.8	59.9	143.6	73.4	53.4	66.4	65.2	60.5	33.7
	3.0	48.9	67.1	44.1	52.6	214.4	64.6	265.3	58.6	151.2	60.4	48.6
	4.0	49.5	67.0	58.0	52.3	223.8	62.0	263.8	54.6	205.0	60.2	71.9
	5.0	50.4	66.3	76.7	53.2	233.5	59.8	282.5	53.6	220.9	59.8	102.3
LEACE	1.0	73.0	66.3	5.8	68.0	28.1	73.2	6.4	68.5	10.4	60.4	9.9
	2.0	49.3	66.2	9.7	67.8	47.1	73.3	9.7	68.5	16.0	60.4	14.7
	3.0	47.3	66.2	12.6	67.6	68.1	73.3	12.3	68.6	20.6	60.4	17.9
	4.0	47.2	66.1	15.2	67.1	88.7	73.3	14.7	68.6	24.6	60.3	20.9
	5.0	48.5	66.1	17.7	66.2	113.6	73.3	16.8	68.7	27.5	60.3	23.0

# 6.10 RESULTS FOR IMAGE DIFFUSION CONCEPT SWITCHING

Concept switching was tested on SDXL and SANA 1.6B models. Both clipped and unclipped variants of the steering function were tested.

Table 20: Model sdxl (noclip), flipping from horse to motorcycle

		horse		motoro	cycle		cow				pig				dog				legis	lator		
		src-cs	tgt-cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid
method	strength																				-	
No Steering	-	71.0	49.1	51.8	70.7	-	72.7	54.6	41.5	-	71.8	49.5	43.6	-	66.3	52.4	44.9	-	60.8	44.8	42.4	-
CASteer	1.0	70.0	50.8	52.6	71.3	27.8	72.3	54.4	42.4	21.9	71.8	49.0	43.8	13.3	66.2	52.0	45.0	20.0	60.9	44.8	42.5	16.3
	1.5	53.4	68.3	60.1	63.7	121.3	72.0	54.4	43.0	28.7	71.9	48.9	44.1	16.2	66.1	51.9	45.1	24.7	60.9	44.8	42.5	21.0
	2.0	52.1	69.5	68.3	52.9	212.4		54.4	44.9		71.9	48.9	44.4	18.9	66.1	51.8	45.4	28.6		44.8	42.6	
	2.5	51.7	69.4	69.4	51.7	213.0	62.5	54.2	55.8	105.2	72.0	48.7	44.6	22.0	66.1	51.7	45.6	33.0	60.9	44.9	42.7	27.0
	3.0	51.4	69.1	69.9	50.9	210.0	52.4	53.2	66.2	186.9	72.0	48.5	44.9	25.8	66.0	51.6	45.7	37.2	60.9	44.9	42.7	29.7
	4.0	51.0	68.7	70.6	49.9	207.9	48.6	52.3	69.5	222.6	72.0	48.5	46.5	37.2	65.7	51.4	46.7	46.5	60.9	44.9	42.8	35.0
	5.0	50.7	68.5	70.9	49.5	207.4		51.8	69.6	231.6		49.1		63.7				61.6	61.0	45.1		39.0
LEACE	1.0	65.0	56.5	52.7	71.2	28.6	72.5	54.4	42.0	17.0	71.7	49.3	43.6	8.9	66.2		44.8	14.1		44.8		21.1
	1.5	52.1	68.6	57.1	67.0	84.6	72.2	54.4	42.3	21.3	71.7	49.3	43.7	10.9	66.2		44.8	17.7	60.7	44.8	42.5	
	2.0	51.2	68.8	67.6	53.3	207.6		54.5	42.7	25.2	71.7	49.3	43.7		66.1	52.3	44.8		60.6			28.2
	2.5	50.8	68.5	69.0	51.5	213.3		54.5	43.1	30.0	71.7	49.4	43.7		66.1	52.3	44.8		60.6	44.9		31.0
	3.0	50.5	68.2	69.6	50.5	210.6		54.4	43.8	37.0	71.7	49.4	43.8		66.0		44.8	24.5		44.8		33.2
	4.0	50.2	68.0	70.4	49.6	206.7	64.6		52.6	85.7	71.8	49.4	43.9	17.2	66.0		44.7	28.1	60.5	44.8	42.6	37.3
	5.0	49.9	67.7	70.8	49.1	204.4		53.2	63.6	166.2		49.4	44.0		65.9		44.7		60.4	44.8	42.7	42.0
mean_matching		51.2	68.7	51.9	70.7	12.7	72.2		42.5	23.9		49.2	43.9		66.1	52.3	44.8	20.7		45.0	42.3	27.2
	1.5	50.4	68.1	51.9	70.7	15.1	71.6	54.6	43.5	34.5		49.1	44.1		66.1	52.2	44.8		60.5	45.1	42.3	32.3
	2.0	50.0	67.8	52.0	70.8	17.2	65.9	54.3	50.9	77.3		49.0	44.3	17.6		52.1	44.8		60.6	45.2	42.3	35.9
	2.5	49.5	67.4	52.0	70.7		55.1	53.3	62.6	162.3		49.0	44.5		66.0		44.8		60.5	45.5	42.4	39.8
	3.0	49.2	67.2	52.0	70.6		51.0	52.7	66.7	199.3		49.0	44.8		66.0		44.8		60.2	45.5	42.4	43.1
	4.0	48.8	67.2	52.0	70.5	22.9	48.4	51.9	68.7	222.5		49.0		29.5				41.4		46.0	42.7	49.9
	5.0	48.3	66.9	52.0	70.4	28.0	47.9	51.5	68.7	229.8	71.9	49.2	46.8	40.0	65.5	51.9	45.2	48.7	59.1	46.7	43.2	58.2

Table 21: Model sdxl (noclip), flipping from chihuahua to muffin

		chihua	ıhua	muffin	ı		dog				wolf				cat				legis	lator		
		src-cs	tgt-cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid												
method	strength																					
No Steering	-	75.9	54.6	42.6	68.2	-	66.3	57.9	52.5	-	71.8	52.7	45.6	-	67.5	53.4	54.2	-	60.8	42.6	40.1	-
CASteer	1.0	71.7	54.7	59.3	61.5	145.9		54.4	52.4	37.4	71.9	51.7	44.7	19.1	67.2		53.9	26.2	60.8	42.4	39.9	19.2
	1.5	47.0	61.0	66.5	57.3	211.5	65.2	53.1	52.4	53.3	71.8	51.2	44.3	24.4	67.1	52.4	53.8	33.8	60.8	42.3	39.6	23.2
	2.0	43.7	63.2	69.5	56.7	226.4	63.2	51.7	53.0	79.7	71.8	51.0	43.9	30.4	66.8	52.0	53.6	41.7	60.8	42.3	39.6	26.7
	2.5	42.3	63.8	71.5	56.3	241.1	55.2	48.9	55.5	140.5	71.5	50.6	43.6	36.8	66.3	51.7	53.6	53.8	60.9	42.3	39.6	29.8
	3.0	41.4	64.0	72.9	56.3	253.8	48.1	45.8	58.2	192.4	70.8	50.3	43.6	45.7	63.6	50.7	54.3	82.4	60.9	42.4	39.5	32.7
	4.0	40.1	64.3	74.8	56.2	276.0	44.2	43.5	60.3	226.9	65.6	49.0	45.6	96.6	52.5	47.2	57.0	168.7	60.9	42.4	39.4	38.0
	5.0	39.3	64.1	75.6	56.1	291.6	42.9	42.5	60.8	243.3	53.5	47.0	51.5	206.1	46.9	45.2	58.0	213.0	60.8	42.5	39.3	43.9
LEACE	1.0	67.1	55.7	60.4	60.6	160.6	66.0	55.8	52.4	25.4	72.0	52.1	45.3	11.9	67.5	53.2	54.1	17.5	60.8	42.5	40.2	20.0
	1.5	46.2	61.5	66.4	57.3	212.1	65.9	54.9	52.3	33.1	72.1	52.0	45.1	14.7	67.5	53.1	54.0	21.1	60.9	42.4	40.3	24.6
	2.0	42.8	63.3	69.3	56.9	227.7	65.8	54.2	52.3	40.3	72.1	51.8	44.9	16.9	67.6	53.0	53.9	24.6	60.9	42.3	40.3	28.1
	2.5	41.4	63.6	71.0	56.7	240.7	65.5	53.4	52.2	48.6	72.1	51.5	44.7	19.0	67.6	52.9	53.9	27.8	61.0	42.3	40.4	31.3
	3.0	40.6	63.9	72.3	56.7	250.8	65.1	52.5	52.3	57.1	72.2	51.4	44.6	21.5	67.6	52.7	53.8	30.4	61.0	42.5	40.5	34.0
	4.0	39.3	63.8	73.7	56.6	269.4	63.8	51.3	52.6	76.7	72.2	51.0	44.2	25.7	67.5	52.6	53.8	35.4	61.0	42.7	40.6	39.4
	5.0	38.5	63.7	74.6	56.5	282.5	60.3	49.7	53.7	104.1	72.2	50.7	44.0	29.8	67.4	52.4	53.7	41.1	61.0	43.0	41.1	43.4
mean_matching	1.0	44.3	62.5	42.0	68.3	23.9	65.9	54.7	52.4	35.1	72.2	51.7	44.9	16.9	67.6	53.0	54.1	22.7	60.8	42.5	40.1	23.9
	1.5	41.5	63.9	41.9	68.2	28.3	65.5	53.5	52.4	48.8	72.2	51.2	44.6	21.2	67.7	52.8	54.1	27.7	60.8	42.4	40.2	29.0
	2.0	39.8	64.0	41.7	68.2	32.2	64.5	52.2	52.7	65.6	72.2	50.8	44.1	25.4	67.7	52.6	54.1	31.6	60.9	42.5	40.4	32.5
	2.5	38.9	64.3	41.6	68.1	35.7	63.2	51.3	53.1	83.0	72.2	50.6	44.0	29.0	67.6	52.4	54.0	35.7	61.0	42.4	40.5	36.3
	3.0	38.0	64.2	41.5	68.0	39.5	59.4	49.7	54.4	109.8	72.2	50.3	43.8	33.4	67.5	52.2	54.0	40.1	60.9	42.4	40.7	39.4
	4.0	37.5	64.0	41.3	67.9	48.2	48.3	44.8	58.7	178.3	71.6	50.3	44.2	43.9	67.0	51.8	54.3	50.7	60.6	42.4	41.1	45.1
	5.0	37.9	63.6	41.3	67.5	58.9	43.4	42.2	61.1	209.8	70.2	50.4	44.8	63.1	64.5	50.7	54.5	69.6	60.4	42.5	41.5	51.4

Table 22: Model sdxl (noclip), flipping from snoopy to mickey

		snoop					pikac	hu			spon	gebob			dog				legisl	ator		
		src-cs	tgt-cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid
method	strength																					
No Steering	-	74.3	58.7	56.0	73.1	-	72.6	41.3	51.2	-	75.1	49.0	52.5		66.3	56.1	52.0	-	60.8	41.6	45.0	-
CASteer	1.0	65.6	68.4	58.9	70.8	47.5	72.7	41.4	51.2	13.8	75.1	48.9	52.5	25.0	66.4	55.0	52.1	25.4	60.8	41.6	45.0	14.9
	1.5	59.8	71.2	61.7	68.6	59.6	72.7	41.4	51.2	17.1		48.9	52.6	29.7	66.3	54.4	52.1	33.9		41.6	45.0	18.4
	2.0	57.0	72.4	66.8	65.3	72.3	72.7	41.5	51.1	19.4		48.9	52.6	33.5		53.9	52.2		60.8	41.5		
	2.5	55.4	72.7	69.6	61.3	83.9	72.7	41.6		21.2		48.8	52.5	36.5		53.3		48.5	60.9	41.6	45.0	23.3
	3.0	54.5	72.7	70.3	59.4	91.8	72.7	41.6	51.2	22.9	75.1	48.9	52.6	39.1	66.4	52.8	52.3	54.8	60.8	41.6	45.0	25.6
	4.0	53.0	71.9	70.8	57.4	106.3	72.7	41.7	51.2	25.9	75.2	48.7	52.6	43.3	66.4	52.1	52.6	66.5	60.7	41.6	45.0	28.9
	5.0	52.0	71.3	71.2	56.4	117.5	72.7	41.7	51.2	28.9	75.0	48.7	52.7	47.7	66.3	51.3	52.8	79.9	60.7	41.5	44.9	31.5
LEACE	1.0	65.9	67.9	58.8	70.6	50.0	72.8	41.4	51.1	18.5	75.0	48.9	52.5	33.5	66.3	55.8	52.0	18.6	60.9	41.7	45.1	22.2
	1.5	60.1	71.0	61.0	68.1	62.6	72.8	41.5	51.1	22.0	75.0	48.8	52.5	39.0	66.3	55.8	52.1	23.9	60.9	41.7	45.0	26.4
	2.0	57.5	72.2	64.3	64.6	77.0	72.9	41.5	51.1	24.5	75.1	48.9	52.5	43.6	66.3	55.6	52.1	28.4	60.9	41.9	45.2	29.7
	2.5	56.0	72.7	67.3	60.7	89.9	72.9	41.6	51.1	26.7	75.2	48.9	52.5	47.7	66.3	55.5	52.2	32.0	61.0	41.9	45.1	31.9
	3.0	55.1	72.7	68.2	58.2	101.5	73.0	41.7	51.2	29.1	75.2	48.8	52.5	51.3	66.4	55.4	52.3	35.7	61.0	41.9	45.1	34.8
	4.0	53.7	72.3	68.7	55.9	118.6	73.2	41.9	51.2	33.4	75.0	49.0	52.6	57.7	66.5	55.3	52.5	43.4	60.9	42.1	45.2	38.9
	5.0	53.0	72.0	69.5	55.2	134.6	73.4	42.4	51.4	37.6	74.8	49.1	52.7	64.0	66.5	55.2	52.7	50.0	60.9	42.3	45.3	42.2
mean_matching	1.0	59.7	71.0	55.5	72.8	30.1	72.8	41.3	51.3	19.0	74.7	48.9	52.6	36.5	66.2	55.9	52.1	20.4	60.9	41.7	45.0	24.5
	1.5	56.0	72.5	55.3	72.7	34.8	72.9	41.2	51.2	22.3	74.6	48.6	52.6	42.7	66.2	55.8	52.1	25.9	60.8	41.8	45.0	29.1
	2.0	54.5	72.7	55.0	72.4	38.9	72.9	41.3	51.3	25.3	74.5	48.6	52.6	46.8	66.3	55.6	52.1	30.6	60.7	42.0	45.1	32.6
	2.5	53.4	72.4	54.8	72.4	42.2	73.0	41.2	51.3	27.3	74.3	48.4	52.6	50.8	66.3	55.6	52.3	34.4	60.7	42.0	45.1	35.2
	3.0	52.5	71.8	54.7	72.2	44.7	73.1	41.3	51.3	30.0	74.2	48.5	52.7	55.2	66.4	55.5	52.5	38.3	60.7	42.1	45.3	37.6
	4.0	51.5	71.2	54.6	71.9	49.6	73.2	41.4	51.4	34.5	73.4	48.4	53.0	64.7	66.4	55.3	52.7	47.2	60.4	42.3	45.4	43.3
	5.0	51.0	70.4	54.4	71.5	55.1	73.3	41.3	51.4	39.1	72.7	48.6	53.4	72.9	66.5	55.1	52.9	55.3	60.3	42.5	45.6	48.8

Table 23: Model sana (noclip), flipping from horse to motorcycle

		horse		motor	cycle		cow				pig				dog				legisl	lator		
		src-cs	tgt-cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid
method	strength																					
No Steering	-	72.1	50.6	50.9	70.5	-	73.8	55.3	42.5	-	73.5	48.3	44.5	-	68.1	51.7	46.1	-	60.4	45.8	43.9	-
CASteer	1.0	71.0	52.0	52.2	71.3	48.1	73.6	54.8	43.0	22.1	73.8	48.0	44.8	13.8	68.1	51.6	46.2	15.7	60.2	46.0	44.3	15.9
	2.0	65.5	60.4	67.5	56.0	229.0	73.1	54.2	43.7	37.9	74.0	47.7	45.0	22.6	68.1	51.4	46.3	24.1	60.0	46.0	44.5	21.5
	3.0	54.9	68.7	69.5	51.8	234.6	72.3	54.0	45.0	55.8	74.2	47.4	45.3	30.5	68.1	51.3	46.5	32.2	59.7	46.3	44.9	25.9
	4.0	52.6	70.3	70.2	50.7	228.7	69.3	55.1	53.0	107.1	74.5	47.4	45.9	39.9	68.2	51.2	46.7	40.4	59.5	46.5	45.3	31.1
	5.0	51.8	71.0	70.5	50.3	224.9	62.3	54.8	61.7	167.0	74.7	47.5	46.9	48.9	68.2	51.2	47.0	51.3	59.3	46.7	45.8	36.8
LEACE	1.0	68.7	56.2	51.6	71.6	36.3	73.7	55.2	42.6	10.2	73.6	48.2	44.7	8.0	68.1	51.8	46.0	8.1	60.4	45.7	43.9	11.6
	2.0	53.1	70.7	58.7	69.4	125.0	73.7	55.2	42.8	16.2	73.7	48.1	44.7	12.6	68.1	51.8	46.0	13.1	60.5	45.7	43.8	17.0
	3.0	51.9	71.3	65.7	58.2	229.3	73.7	55.1	43.1	21.0	69.8	57.6	52.1	220.6	68.0	51.8	46.0	17.7	60.7	45.8	43.8	21.6
	4.0	51.9	71.1	67.4	53.7	242.8	73.7	55.0	43.3	25.9	73.9	48.0	44.9	20.4	68.0	51.8	46.0	21.5	55.1	52.2	51.9	294.6
	5.0	51.7	70.8	68.4	52.6	236.2	73.7	54.9	43.6	31.8	74.0	48.0	45.0	24.0	68.0	51.8	45.9	25.0	60.8	45.8	43.8	30.1
mean_matching	1.0	55.5	68.9	50.9	70.4	8.7	73.6	55.1	42.7	14.7	73.7	48.2	44.7	10.4	68.0	51.8	46.1	9.3	60.5	45.8	44.0	13.0
	2.0	51.7	71.1	50.9	70.4	11.7	73.6	55.0	43.1	23.2	73.9	48.1	44.8	16.5	68.0	51.7	46.1	15.0	60.6	45.8	43.9	19.1
	3.0	51.4	70.7	50.9	70.4	13.9	73.4	54.7	43.4	32.5	74.1	48.0	45.0	22.2	67.9	51.7	46.1	19.8	60.8	45.9	44.0	23.9
	4.0	51.4	70.9	50.9	70.4	15.9	73.1	54.5	43.9	43.8	74.3	48.1	45.3	27.4	67.9	51.6	46.1	24.3	60.8	45.9	44.1	28.9
	5.0	51.5	71.4	51.0	70.4	18.0	69.8	55.2	50.0	78.2	74.5	48.1	45.7	34.3	67.9	51.6	46.2	28.4	61.0	46.0	44.2	34.3

Table 24: Model sana (noclip), flipping from chihuahua to muffin

		chihua	hua	muffin	ı		dog				wolf				cat				legisl	ator		
		src-cs	tgt-cs	src-cs	tgt-cs	fid	CS	src-cs	tgt-cs	fid												
method	strength																					
No Steering	-	76.4	55.0	43.4	66.3	-	68.1	62.0	52.7	-	73.2	52.8	46.1	-	68.5	53.4	53.0	-	60.4	42.7	40.8	-
CASteer	1.0	75.8	55.8	45.3	65.6	46.6	67.7	58.4	52.9	39.4	73.1	52.7	45.8	16.3	68.3	52.8	53.5	24.5	60.2	42.8	41.0	13.8
	2.0	55.3	60.1	62.0	60.4	200.7	67.1	56.3	54.1	78.8	72.9	52.4	45.6	28.4	68.3	52.1	54.3	39.2	60.1	42.7	41.1	18.9
	3.0	45.4	60.5	60.0	47.6	278.5	50.6	49.0	59.2	198.9	72.4	52.4	45.8	46.2	67.9	51.0	55.0	61.7	59.9	42.8	41.3	23.4
	4.0	44.7	60.7	70.6	54.9	250.7	45.1	46.0	60.8	223.3	68.6	52.0	46.6	113.8	65.5	50.1	55.5	109.2	59.9	42.7	41.4	27.6
	5.0	44.1	60.7	72.1	54.2	260.7	43.4	45.3	61.1	229.3	58.9	51.1	50.3	216.2	56.9	48.0	57.6	179.1	59.9	42.7	41.6	31.3
LEACE	1.0	72.3	58.3	46.1	65.0	57.7	67.9	59.9	52.7	21.0	73.2	52.9	46.1	5.8	68.4	53.2	52.9	8.9	60.4	42.7	40.8	9.1
	2.0	48.9	61.4	59.8	63.0	185.8	67.8	58.3	52.8	35.1	73.2	52.9	46.1	9.4	68.4	53.1	53.0	13.6	60.4	42.5	40.8	13.7
	3.0	45.5	60.7	65.7	57.2	227.5	67.7	57.1	52.8	48.4	73.2	52.9	46.0	12.5	68.5	53.1	53.0	17.5	60.4	42.7	40.7	16.9
	4.0	44.6	60.5	68.1	55.8	240.4	67.7	56.3	53.0	60.8	73.2	52.9	46.0	15.4	68.5	53.0	53.0	21.0	60.4	42.7	40.7	19.4
	5.0	44.2	60.1	69.4	55.1	246.3	67.4	55.9	53.6	76.7	73.1	52.9	46.0	18.1	68.6	52.9	53.0	24.0	60.4	42.9	40.6	21.5
mean_matching	1.0	53.3	60.5	43.4	66.2	7.8	66.2	62.5	48.3	216.4	73.2	52.8	45.9	7.1	68.4	53.2	53.1	12.1	60.4	42.6	40.8	10.7
	2.0	44.4	58.7	43.4	66.0	13.1	67.6	56.8	53.0	57.9	73.1	52.8	45.8	11.2	68.4	52.9	53.1	17.8	60.4	42.5	40.8	15.8
	3.0	44.1	56.4	43.5	65.8	17.7	65.8	55.4	54.4	96.5	73.0	52.8	45.7	14.5	68.5	52.7	53.2	23.0	60.3	42.4	40.8	19.8
	4.0	45.0	53.8	43.6	65.6	22.3	56.4	51.6	56.3	154.4	73.0	52.9	45.6	17.3	68.5	52.5	53.2	27.2	60.2	42.5	40.8	22.9
	5.0	46.1	51.8	43.6	65.5	27.3	49.3	48.5	57.2	193.2	73.0	52.9	45.5	19.9	68.5	52.2	53.4	31.3	60.1	42.4	40.6	25.9

Table 25: Model sana (noclip), flipping from snoopy to mickey

		snoop	У	mickey			pikachu			spongebob				dog			legislator					
		src-cs	tgt-cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	src-cs	tgt-cs	fid	cs	STC-CS	tgt-cs	fid
method	strength																					
No Steering	-	79.7	58.0	56.3	76.1	-	74.0	41.5	50.9	-	79.0	50.7	53.8	-	67.3	57.0	57.1	-	60.4	42.9	46.6	-
CASteer	1.0	77.9	61.7	57.8	76.3	31.8	74.1	41.5	50.8	8.6	79.0	50.7	53.8	14.5	68.1	54.0	52.0	218.8	60.3	43.0	46.8	12.5
	2.0	65.6	71.0	61.0	76.5	53.8	74.1	41.6	50.8	13.9	78.9	50.8	53.8	19.8	68.2	53.5	52.0	218.3	60.2	43.0	46.9	17.7
	3.0	55.8	72.3	71.7	74.8	78.8	74.2	41.7	50.7	18.6	78.9	50.8	53.8	23.9	68.1	53.2	52.1	218.3	60.1	43.0	46.8	21.4
	4.0	52.0	71.4	78.5	67.6	102.9	74.3	41.8	50.7	22.7	78.9	50.8	53.8	27.9	68.2	52.8	52.1	218.5	60.1	43.0	47.0	25.4
	5.0	50.0	70.4	79.8	61.9	121.0	74.5	42.0	50.7	26.7	79.0	50.8	53.8	31.8	68.2	52.4	52.2	218.0	60.1	43.0	47.2	28.5
LEACE	1.0	76.6	63.0	58.2	76.4	36.0	74.1	41.6	51.0	3.2	79.0	50.7	53.7	10.8	68.1	54.2	52.0	219.3	60.2	42.9	46.6	5.9
	2.0	62.6	71.3	57.9	72.9	170.9	74.1	41.7	51.0	5.5	78.9	50.7	53.7	15.1	68.0	54.0	52.0	220.0	60.2	42.9	46.6	9.6
	3.0	55.9	71.7	75.1	73.4	88.9	74.2	41.7	51.0	7.5	78.9	50.7	53.7	18.2	68.0	53.7	52.1	218.7	60.1	42.9	46.7	12.2
	4.0	52.9	68.8	79.2	66.4	111.0	74.2	41.8	51.1	9.5	78.9	50.8	53.8	21.1	68.0	53.6	52.1	218.8	60.1	42.9	46.7	14.6
	5.0	51.0	65.5	79.6	60.8	130.5	74.3	41.9	51.1	11.2	78.9	50.8	53.8	24.1	68.0	53.3	52.2	219.7	59.9	42.9	46.7	16.4
mean_matching	1.0	64.9	70.8	56.0	76.2	16.0	74.1	41.5	51.0	5.8	79.1	50.6	53.8	12.3	68.1	54.0	52.0	219.2	60.2	42.9	46.7	8.3
	2.0	53.5	71.1	55.6	76.2	25.2	74.2	41.5	51.1	10.0	79.2	50.6	53.9	17.4	68.1	53.6	52.1	218.2	60.1	42.9	46.7	12.4
	3.0	50.0	67.4	55.2	76.2	34.8	74.4	41.5	51.2	13.5	79.2	50.6	54.0	21.6	68.0	53.2	52.2	218.6	59.9	42.8	46.7	15.6
	4.0	48.4	62.6	54.9	76.4	45.4	74.5	41.6	51.3	16.8	79.2	50.5	54.0	25.8	67.9	52.6	52.3	216.6	59.8	42.8	46.7	17.8
	5.0	48.1	60.7	54.5	76.6	55.3	74.6	41.5	51.3	20.0	79.2	50.3	54.1	29.6	68.0	52.3	52.4	215.1	55.0	47.5	52.8	271.2

6.11 PARETO CHARTS FOR LLM CONCEPT ERASURE

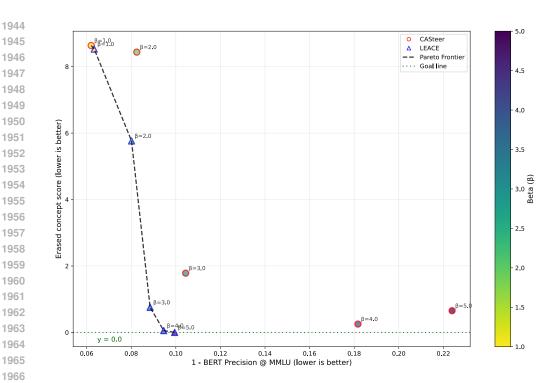


Figure 5: Pareto plot for concept erasure on model llama2-7b (noclip)

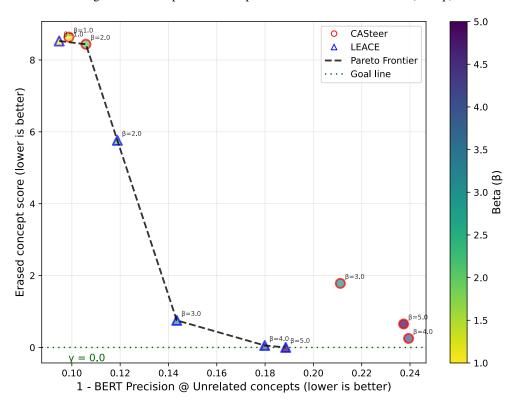


Figure 6: Pareto plot for concept erasure on model llama2-7b (noclip)

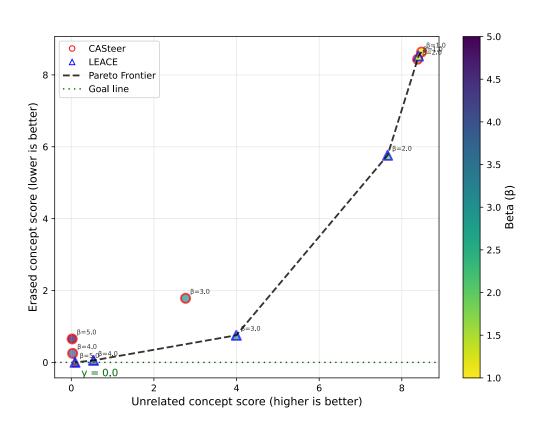


Figure 7: Pareto plot for concept erasure on model llama2-7b (noclip)

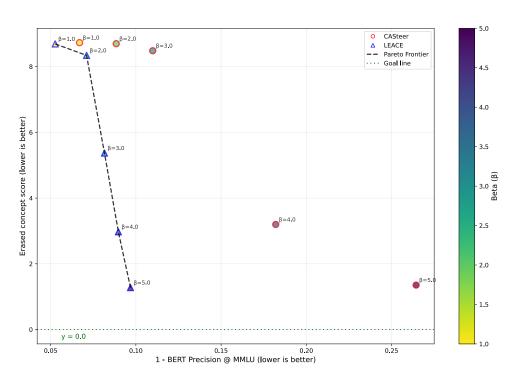


Figure 8: Pareto plot for concept erasure on model qwen-14b (noclip)

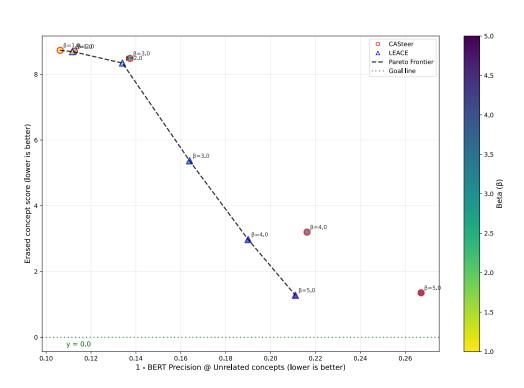


Figure 9: Pareto plot for concept erasure on model qwen-14b (noclip)

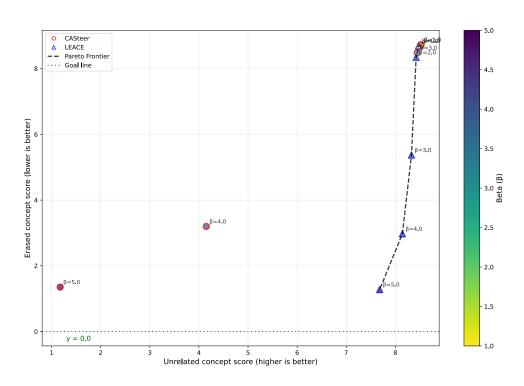


Figure 10: Pareto plot for concept erasure on model qwen-14b (noclip)

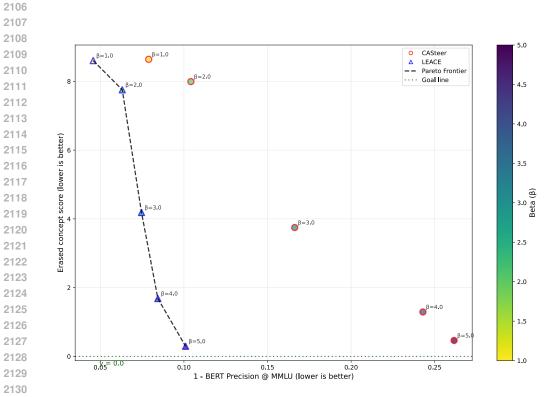


Figure 11: Pareto plot for concept erasure on model qwen-7b (noclip)

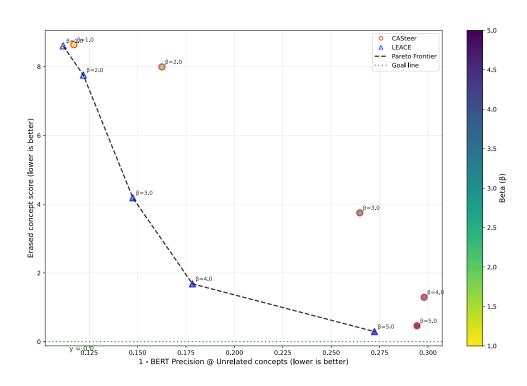


Figure 12: Pareto plot for concept erasure on model qwen-7b (noclip)

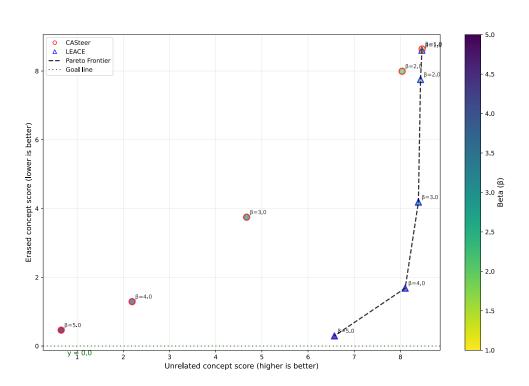


Figure 13: Pareto plot for concept erasure on model qwen-7b (noclip)

6.12 PARETO CHARTS FOR LLM CONCEPT SWITCHING

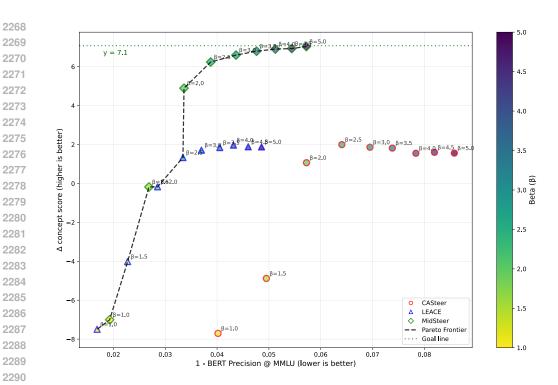


Figure 14: Pareto plot for concept flip on model llama2-7b (noclip)

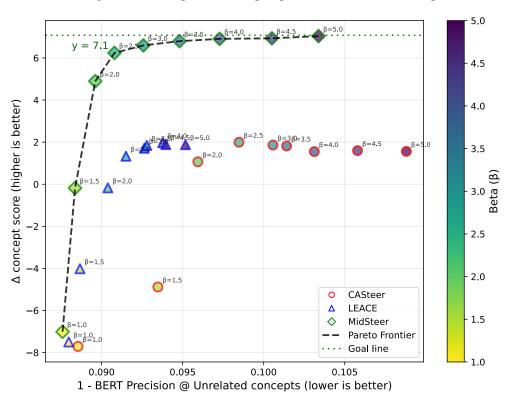


Figure 15: Pareto plot for concept flip on model llama2-7b (noclip)

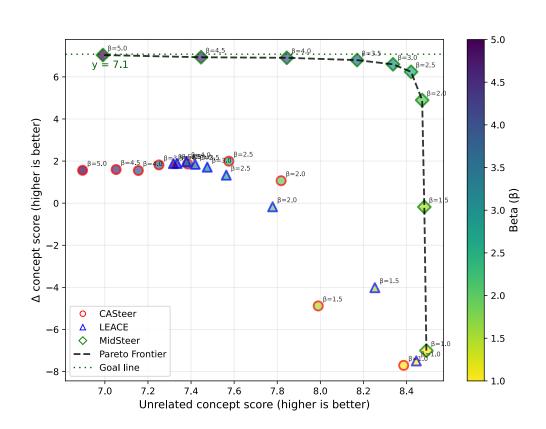


Figure 16: Pareto plot for concept flip on model llama2-7b (noclip)

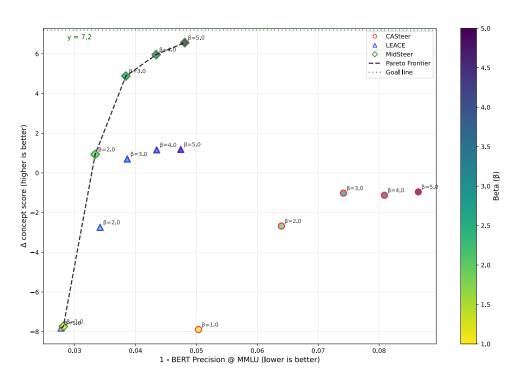


Figure 17: Pareto plot for concept flip on model qwen-14b (noclip)

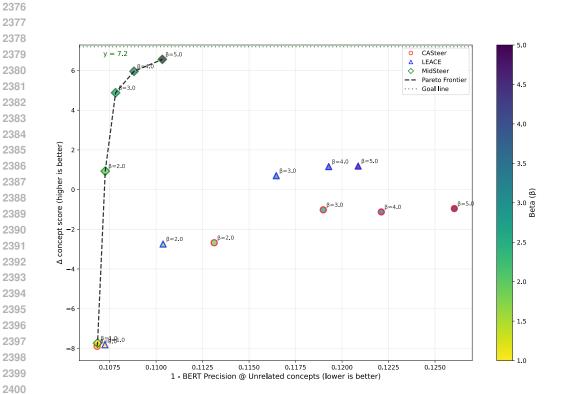


Figure 18: Pareto plot for concept flip on model qwen-14b (noclip)

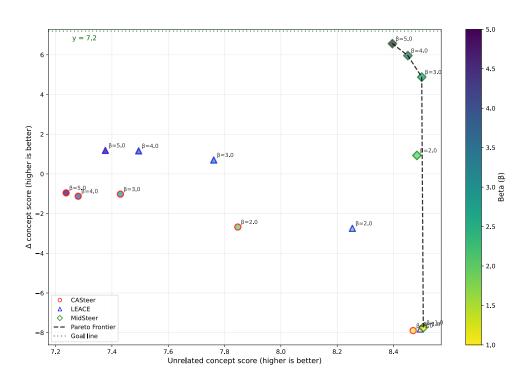


Figure 19: Pareto plot for concept flip on model qwen-14b (noclip)

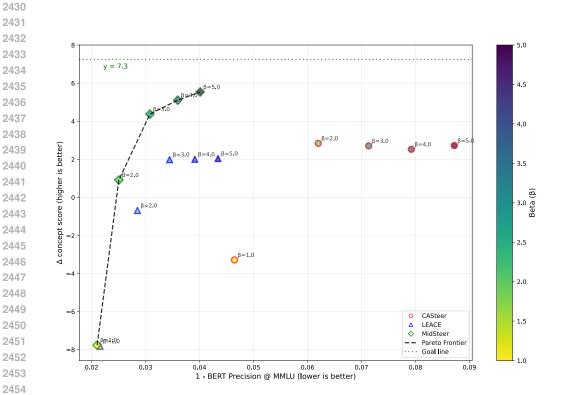


Figure 20: Pareto plot for concept flip on model qwen-7b (noclip)

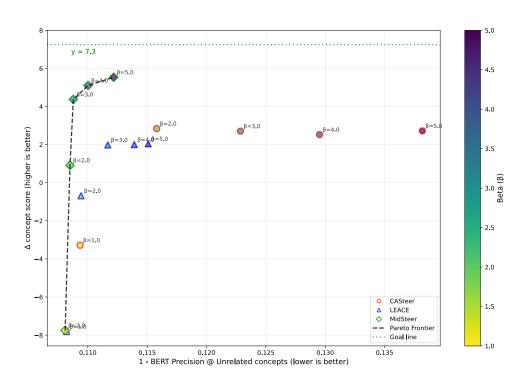


Figure 21: Pareto plot for concept flip on model qwen-7b (noclip)

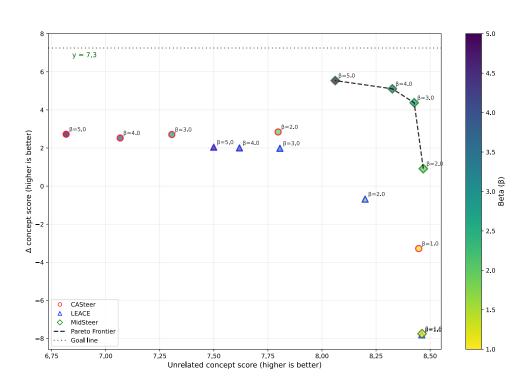


Figure 22: Pareto plot for concept flip on model qwen-7b (noclip)

6.13 PARETO CHARTS FOR IMAGE DIFFUSION CONCEPT ERASURE

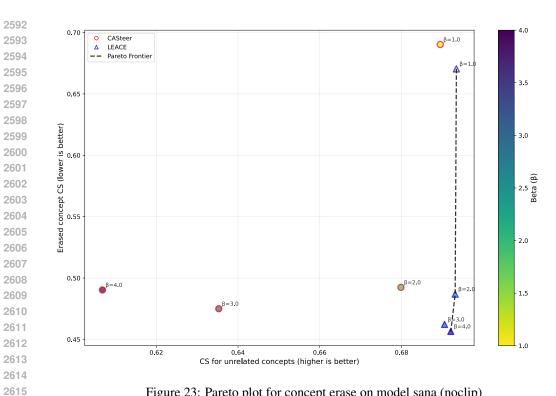


Figure 23: Pareto plot for concept erase on model sana (noclip)

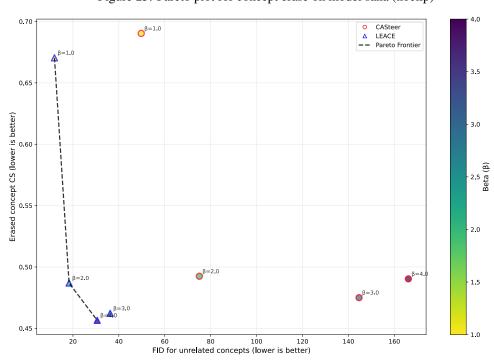


Figure 24: Pareto plot for concept erase on model sana (noclip)

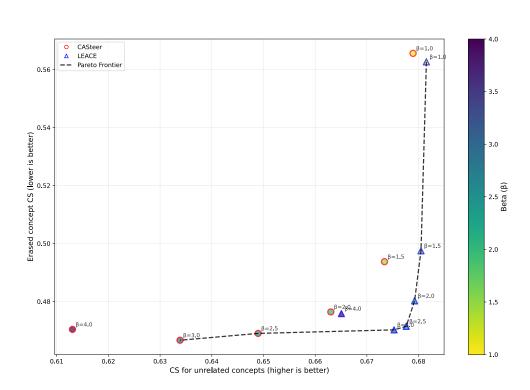


Figure 25: Pareto plot for concept erase on model sdxl (noclip)

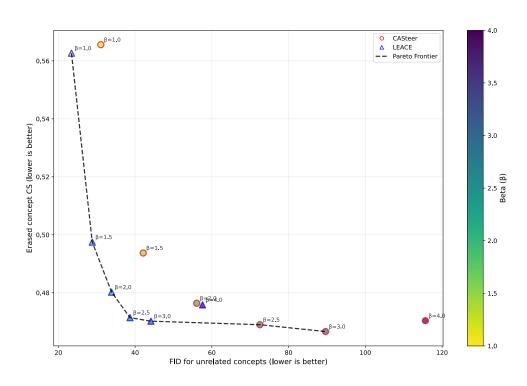


Figure 26: Pareto plot for concept erase on model sdxl (noclip)

6.14 PARETO CHARTS FOR IMAGE DIFFUSION CONCEPT SWITCHING

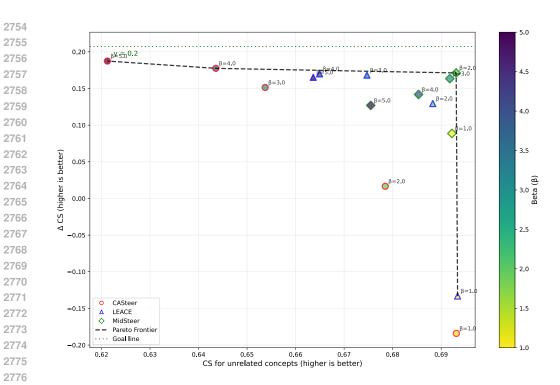


Figure 27: Pareto plot for concept flip on model sana (noclip)

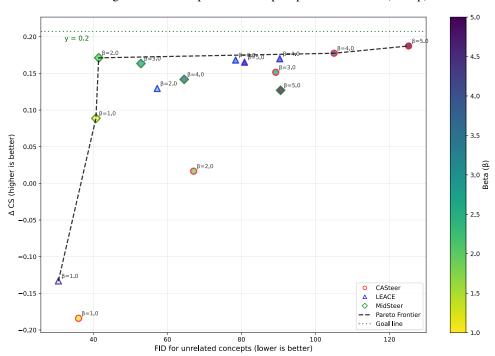


Figure 28: Pareto plot for concept flip on model sana (noclip)

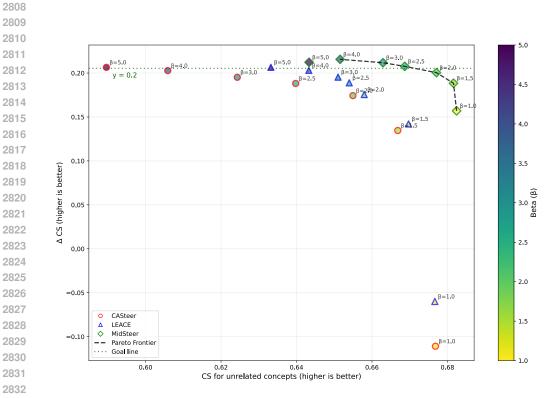


Figure 29: Pareto plot for concept flip on model sdxl (noclip)

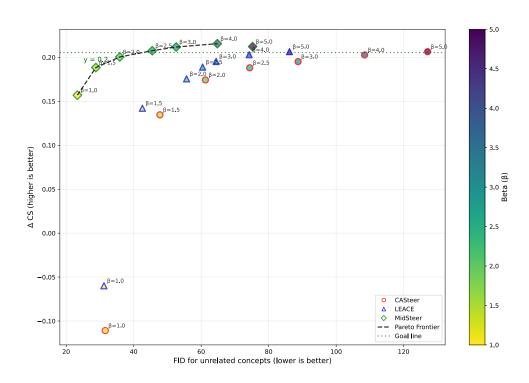


Figure 30: Pareto plot for concept flip on model sdxl (noclip)