

LinEAS: End-to-end Learning of Activation Steering with a Distributional Loss

Pau Rodríguez* Michal Klein Eleonora Gualdoni Valentino Maiorca Arno Blaas
 Apple Apple Apple Sapienza, Apple Apple

Luca Zappella Marco Cuturi Xavier Suau*
 Apple Apple Apple

Abstract

The growing use of generative models in daily life calls for efficient mechanisms to control their generation, to *e.g.*, produce safe content or provide users with tools to explore style changes. Ideally, such mechanisms should require low volume of unpaired data (*i.e.*, without explicit preference), and should be cheap, both at train and inference time, while preserving output quality. Recent research has shown that such mechanisms can be obtained by intervening exclusively on model *activations*, with the goal of correcting *distributional* differences between activations seen when using prompts from a source vs. a target set (*e.g.*, toxic and non-toxic sentences). While cheap, these fast methods are inherently crude: their maps are tuned locally, not accounting for their impact on downstream layers, resulting in interventions that cause unintended shifts when used out-of-sample. We propose in this work linear end-to-end activation steering (LinEAS), an approach trained with a global loss that accounts simultaneously for all layer-wise distributional shifts. In addition to being more robust, the loss used to train LinEAS can be regularized with sparsifying norms, which can automatically carry out neuron selection. LinEAS only requires a handful of unpaired samples to be effective, and beats similar baselines on toxicity mitigation in language models, becoming competitive with oracle-dependent methods that have access to strong supervision. LinEAS is modality-agnostic and we empirically find that it outperforms existing activation steering methods at mitigating and including new concepts at the output of single-step text-to-image generation models.

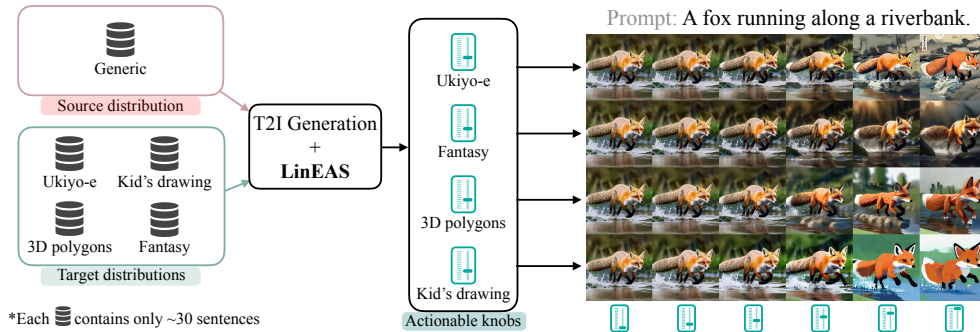


Figure 1: LinEAS learns lightweight maps to steer pretrained model activations. With LinEAS, we gain fine-grained control on text-to-image generation to induce precise styles (in the figure) or remove objects (*e.g.*, Section 4.4). The same procedure also allows controlling LLMs (*e.g.*, Section 4.1).

1 Introduction

Modern generative models are typically trained in two distinct phases. The first phase, known as pre-training, involves learning from a large corpus of data using tasks like next-token prediction or text-to-image generation. This is followed by an alignment phase designed to adjust the model towards a more specific, desired behavior. This alignment can be achieved through instruction fine-tuning [1], reinforcement learning from human feedback (RLHF) [2] for LLMs, or guidance [3] and LoRA adapters [4] in text-to-image diffusion. Many of these approaches propose to modify the model’s internal mechanisms, realigning its parameters by leveraging new data with, ideally, a minimal impact on the utility of the model.

The rapid growth of model sizes, coupled with the potentially infinite combination of alignment goals, calls for alignment mechanisms that readily adapt to new and evolving user needs. Ideally, adaptable alignment methods should adhere to the following desiderata: low training cost (potentially on device), memory-efficiency (small set of parameters), low inference time overhead, data-efficiency (few annotated samples), and fine-grained control. Working in low-data regimes makes data collection simple or even unmanned, and makes training potentially faster while fine-grained control makes alignment methods more customizable [5]. Such advantages together enable an agile control of generative models (see example in Figure 1), giving tools to users to customize their experience, and to administrators to quickly intervene to prevent harmful model behaviors.

A body of work known as *activation steering* [6–11] has proven to be effective at conditioning models while being memory and compute efficient, but still falls short in other desiderata. Some methods require data points to be *paired* with their corresponding counterfactuals [12], while others require a reward model (or a human) to indicate preference among generated outputs [2]. While this is a form of *strong supervision* that provides a robust signal to alignment methods, it is costly or even impossible to obtain in many scenarios. A weaker and more flexible form of supervision consists of dealing with two sets of examples – one for the target and one for the source behavior (*e.g.*, non-toxic and toxic sentences). In this setting, sentence pairs are not required, and no additional supervision is needed. Methods using this weaker signal typically analyze the distribution of activations from each set, thus we refer to them as a *distributional approaches*.

In this work, we propose Linear End-to-end Activation Steering (LinEAS), a **low-data** and **weakly-supervised** (unpaired data, no reward model) method that is trained with a **global distributional cost** grounded in optimal transport theory as signal for steering. Our hypothesis, which we validate empirically, is that our end-to-end learning accounts for the interactions between maps at different layers, leading to improved results over other steering methods. An additional advantage of optimizing a global cost is that it allows us to introduce additional objectives such as a regularization coefficient, which results in more targeted interventions, preserving the utility of the model.

Our main contributions are:

- We propose LinEAS, a novel framework to steer activations based on affine optimal transport maps between activations trained jointly across layers (*e.g.*, T_1 , T_2 in Figure 2) with a global loss that enforces a global distributional alignment across all layers.¹ Our method provides low-budget conditioning, with a continuous and theory-grounded application strength $\in [0, 1]$ (*e.g.*, Figure 1).
- We show how LinEAS can be coupled with a *sparse lasso* regularizer [13] that can detect a small subset of activations that matter for a steering goal, reducing the intervened support by 100×, resulting in improved model utility.
- We show that LinEAS learns effective interventions with as few as 32 source and 32 target unpaired data points, showing state-of-the-art toxicity mitigation among steering methods.
- We validate fine-grained control of text-to-image (T2I) models, showing superior coherence with the prompt semantics while achieving the desired alignment goal.

2 Related Work

We classify steering methods based on whether they require strong (paired data or use of a human supervision / oracle) or weak supervision (unpaired data and no further supervision).

¹<https://github.com/apple/ml-lineas>.

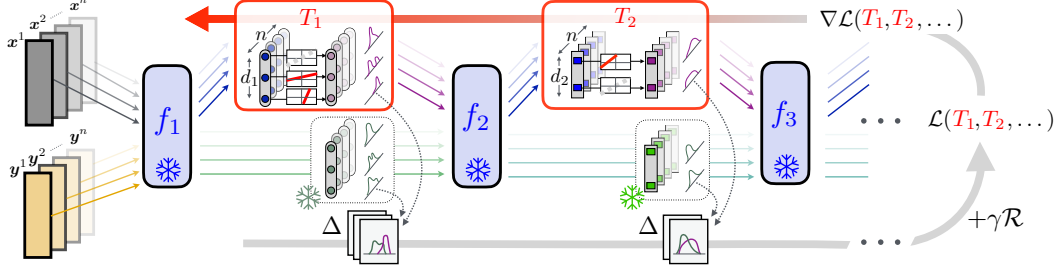


Figure 2: Given a frozen computational graph (blue) of $L + 1$ layers of interest, we interlace it with L transport blocks (red). Each transport is defined as a collection of *coordinate-wise* affine transformations, as displayed in the 3 and 2 boxes for maps T_1 and T_2 respectively. All transport maps are jointly trained to minimize a sum of distributional losses Δ between the neural activation distributions collected from samples $x_1, \dots, x_n \sim p$ (one shade of grey per sample) and $y_1, \dots, y_n \sim q$ (resp. yellow). We learn the parameters of these maps *jointly* by minimizing the penalized sum of Δ terms, where Δ is a 1D Wasserstein distances evaluated on the d_ℓ activations of layer ℓ . Using a global optimization, we can consider sparsifying regularizers (included in $\mathcal{R}$), to, e.g., select a sparse subset of activations that require interventions. For instance, when adding a regularizer that promotes sparsity, both T_1 and T_2 do not intervene on one neuron, the first and the second, respectively.

Strong supervision. CAA [7] calculates the steering vector as the mean of differences between two sets of paired prompts. ReFT [14] optimizes low-rank projections of representations. LoFiT [11] learns a bias added to the representations of pre-selected attention heads. LoFiT training is similar to ReFT (although simpler, since only biases are trained). BitFit [15] directly finetunes the bias terms of pre-selected linear layers. Most of these methods can be viewed as a simplified version of LoRA [16]. Similarly to LoRA they also require either paired data or a human /oracle supervision during training.

Weak supervision. Some methods propose to add a vector to the activations. For example, ACTADD [17] uses the difference between 2 prompts and Mean-AcT the difference in means [10]. ITI-C [9] uses a steering vector orthogonal to the hyperplane learned by a binary linear classifier on the activations from two sets of sentences. With a different approach, AURA by Suau et al. [18] dampens activations proportionally to each neuron’s ability to classify toxic and non-toxic sentences, effectively mitigating toxicity. REPE, by Zou et al. [8], does require paired data, however, we place this algorithm in the weak supervision family since it computes steering vectors based on a single prompt pair. Closest to our work, Lin-ACT [10] uses an affine map to steer activations. In Section 3.1 we discuss in more detail the differences between our proposed LinEAS with Lin-ACT and other methods.

3 End-to-end Learning of Steering Maps

We propose LinEAS, a method to optimize activation-specific interventions in a joint manner. Our hypothesis is that a global estimation that exploits causal interdependencies between activations across layers is needed to maintain the model’s utility while achieving the steering goal.

3.1 Interventions Setup and Distributional Loss

We consider a generative model and target a set of L intermediate activation layers. We describe the model as a composition of $L + 1$ abstract pretrained functions, where the output, given an input prompt $x \in \mathcal{S}$, can be written as $o = f_{L+1} \circ f_L \circ \dots \circ f_1(x)$. For convenience, we denote d_ℓ the dimension of the activations obtained at layer ℓ , i.e., the size of the activation vector $f_\ell \circ \dots \circ f_1(x)$. We interlace this pretrained computational graph of $L + 1$ intermediate frozen layers with L comparatively much simpler vector-to-vector maps: $o = f_{L+1} \circ T_L \circ f_L \circ \dots \circ T_2 \circ f_2 \circ T_1 \circ f_1(x)$, where for $1 \leq \ell \leq L$, the map $T_\ell : \mathbb{R}^{d_\ell} \rightarrow \mathbb{R}^{d_\ell}$ acts on the intermediate activations observed after layer ℓ , by altering its coordinates and outputting a vector of the same size.

Layerwise Distributional Losses. We consider two distinct probability distributions over prompts, the source distribution p and the desired target distribution q . For example, p and q could be such that a sample from the source distribution p corresponds to toxic sentence, while a sample from the target

distribution q corresponds to a non-toxic sentence (see Section 4.1). We then view each sampled prompt through the lens of their sequence of L activations. In practice, this means having access to samples $\mathbf{x}^1, \dots, \mathbf{x}^n \sim p$ and $\mathbf{y}^1, \dots, \mathbf{y}^n \sim q$, tracking their execution trace of their $\ell \leq L + 1$ activations, either modified through interleaved transports for samples from p :

$$\xi_\ell^i := T_\ell \circ f_\ell \circ T_{\ell-1} \circ \dots \circ T_1 \circ f_1(\mathbf{x}^i), \quad \ell \leq L, \quad (1)$$

or ran through the original network for samples of q :

$$\eta_\ell^j := f_\ell \circ \dots \circ f_1(\mathbf{y}^j), \quad \ell \leq L. \quad (2)$$

Our goal is to learn *jointly* all L transport maps so that, for each $\ell \leq L$, the families of vectors $(\xi_\ell^i)_i$ and $(\eta_\ell^j)_j$ are *similar* with respect to a distributional metric Δ between probability measures, making the cost below small:

$$\mathcal{C}(T_1, \dots, T_L; (\mathbf{x}^i)_i, (\mathbf{y}^j)_j) = \sum_{\ell \leq L} \Delta((\xi_\ell^i)_i, (\eta_\ell^j)_j). \quad (3)$$

Sliced Wasserstein Losses. To define Δ at each layer, we adopt the approach of Rodriguez et al. [10] and sum d_ℓ univariate Wasserstein distances between the d_ℓ marginal distributions at each layer ℓ . This choice is motivated by the fact that in the typical setting targeted in this work, we must deal with a high-dimensionality / low sample regime, $d_\ell \gg N$, that would hinder the use of more complex multivariate distributional losses that account more closely for cross-variable effects. We observe that adding univariate quantities yields a more robust loss estimation that translates to better downstream tasks than considering, e.g., Sinkhorn divergences [19].

To define Δ at each layer, we adopt the approach of Rodriguez et al. [10] and use d univariate Wasserstein distances between the d marginal distributions. The activations can be arranged as matrices $U := [\xi_\ell^1, \dots, \xi_\ell^n]$ and $V = [\eta_\ell^1, \dots, \eta_\ell^n]$, both in $\mathbb{R}^{n \times d_\ell}$. To compute their 1D-Wasserstein distance [20, Chap. 2], these activations must be first sorted in increasing order along the feature axis:

$$\tilde{U} = \text{sort}(U, \text{axis} = -1), \tilde{V} = \text{sort}(V, \text{axis} = -1)$$

to define the sliced Wasserstein distance [21] computed only on the *canonical directions*, namely:

$$\Delta(U, V) := \sum_{j=1}^d W_2^2(U_{\cdot j}, V_{\cdot j}) = \frac{1}{n} \sum_{j=1}^d \|\tilde{U}_{\cdot j} - \tilde{V}_{\cdot j}\|^2. \quad (4)$$

Differences with ITI-C, Lin-ACT and ReFT. Both ITI-C [9] and Lin-ACT [10] optimize each T_ℓ *independently* across layers, minimizing a single distributional difference in closed form (Lin-ACT) or learning a linear classifier (ITI-C), and assuming all other layers are frozen. While arguably much faster, this also generates causal inconsistencies, which can be partially resolved as in Lin-ACT by using a sequential approach: when $T_{\ell-1}$ is trained, $T_{\ell-1}$ is reused to recompute all activation distributions used for T_ℓ . We claim that this suboptimality is to blame for poor generalization. This independent approach also precludes trade-offs when choosing which activations to turn on/off across layers, which we can easily be surfaced using sparsity regularizers. One fundamental difference with ReFT [12] is that LinEAS does not need paired data, *i.e.*, sample $\mathbf{y}^i$ does not need to be a counterfactual of $\mathbf{x}^i$, which is the case for ReFT. While counterfactual data is important for applications like translation, there is a vast amount of applications (*e.g.*, toxicity) where paired data is not available.

3.2 Parameterization and Regularization

Building on the approach outlined by Rodriguez et al. [10], we propose to parameterize each map T_ℓ as an affine map for each activation $\ell \leq L$, namely for $z \in \mathbb{R}^{d_\ell}$, one has

$$T_\ell(z) := \omega_\ell \odot z + b_\ell, \quad \omega_\ell, b_\ell \in \mathbb{R}^{d_\ell}, \quad (5)$$

where $\odot$ is the element-wise product. We write $\mathbf{w} := (\omega_1, \dots, \omega_L)$ and $\mathbf{b} := (b_1, \dots, b_L)$ for the collections of all scale and intercept parameters. In what follows since each map T_ℓ is entirely parameterized through its scale/intercept parameters $\mathbf{w}, \mathbf{b}$, we overload notations to define

$$\mathcal{C}(\mathbf{w}, \mathbf{b}; (\mathbf{x}^i)_i, (\mathbf{y}^j)_j) := \mathcal{C}(T_1, \dots, T_{L-1}; (\mathbf{x}^i)_i, (\mathbf{y}^j)_j).$$

Sparse Regularization. We propose to use a sparsity regularizer that will carry out both *layer* and *within-layer* selection of activations. This can be achieved by using structured regularization, using either 1-norms or 2-norms:

$$\mathcal{R}_1(\mathbf{w}, \mathbf{b}) := \sum_{\ell} \|\omega_{\ell} - \mathbf{1}\|_1 + \|\mathbf{b}_{\ell}\|_1 \quad \text{and} \quad \mathcal{R}_G(\mathbf{w}, \mathbf{b}) := \sum_{\ell} \sqrt{d_{\ell}} (\|\omega_{\ell} - \mathbf{1}\|_2 + \|\mathbf{b}_{\ell}\|_2).$$

resulting in a *sparse group lasso* regularizer [13, 22], $\mathcal{R} := \lambda_1 \mathcal{R}_1 + \lambda_G \mathcal{R}_G$, which can be added to the cost to result in:

$$\mathcal{L}(\mathbf{w}, \mathbf{b}) := \mathbb{E}_{\substack{(\mathbf{x}^i)_{i \sim p}, \\ (\mathbf{y}^j)_{j \sim q}}} [\mathcal{C}(\mathbf{w}, \mathbf{b}; (\mathbf{x}^i)_i, (\mathbf{y}^j)_j)] + \gamma \mathcal{R}(\mathbf{w}, \mathbf{b}), \quad (6)$$

where γ controls the amount of sparsity, *i.e.*, larger γ will result in fewer activations and/or layers being intervened on, with solution such that $\omega \approx \mathbf{1}$ and $\mathbf{b} \approx \mathbf{0}$.

On the choice of sparsity hyperparameters We only consider two hyperparameters at the moment, λ_1 and λ_G , which is relatively small since LinEAS intervention models have none. Empirically, we found that tuning $\gamma \in [0, 1]$ with $\lambda_1 = \lambda_G = 1$ already provides interesting trade-offs between conditioning and utility as reported in Figures 4 and 8. This also reduces the number of additional hyper-parameters to tune to just one (γ) and enables automatic layer selection (Figure 10).

3.3 Optimization

Proximal SGD. We optimize $\mathcal{L}$ in (6) with proximal stochastic gradient descent (PSGD) with a learning rate of 0.1 and cosine decay. We assume access to 2 sets of unpaired N prompts $(\mathbf{x}^i)_{i=1}^N$ and $(\mathbf{y}^i)_{i=1}^N$ and run PSGD on minibatches of activations $(\xi_{\ell}^i)_i$ and $(\eta_{\ell}^i)_i$ of size n . Note that activations originating from $\mathbf{y}^i$ use the untouched network, and can be pre-computed beforehand. In Appendix D we provide details on the algorithm and proximal operators.

4 Experimental Results

4.1 Toxicity Mitigation in LLMs

We analyze the effectiveness of LinEAS at the important task of toxicity mitigation. To that end, we compare with prompting, CAA [7], ReFT [12], ITI-c [9] and Lin-ACT [10] on three LLMs ranging from 1.5B to 7B parameters, by aligning the activations of $N = 32$ toxic to 32 non-toxic sentences sampled from the Jigsaw dataset [23].

Toxicity Metrics. We evaluate toxicity mitigation on the *RealToxicityPrompts* (RTP) dataset [24] and the *Thoroughly Engineered Toxicity* (TET) dataset [25]. For RTP, we follow Rodriguez et al. [10] by sampling 1000 prompts from the dataset and let the model (intervened or not) complete them. For TET, we use the 2546 prompts provided. Then, we score the generations using the open-source Roberta toxicity classifier (RTC) [26]. We report $\text{Tox}_{\text{RTC}}^{\text{RTP}}$ and $\text{Tox}_{\text{RTC}}^{\text{TET}}$, the respective percentage of generations flagged as toxic on RTP and TET.

Utility Metrics. To measure whether the utility of the model is affected by these interventions, we report PPL_{WIK} , the perplexity obtained on a fixed set of 20k Wikipedia sentences [27], as well as the overall 5-shot accuracy on the MMLU compendium [28].

Oracle Baselines. In addition, we introduce two baselines that require a strong supervision signal directly from the RTC oracle, *i.e.*, the classifier used to compute *test-time metrics*, giving them a significant advantage. More precisely, we train a LoRA adapter and LoFIT [11] using the RTC labels, yielding the oracle methods LRTC-RL and LoFIT-RL. We report in Table 1 only the results of LoFIT-RL as it performed better than LRTC-RL (available in Table 7). See Appendix I for details on the oracle training protocol.

Setup. All methods have access to only 32 toxic and 32 non-toxic sentences (unpaired). We optimize LinEAS for 1K steps using SGD and a learning rate of 0.1. As we focus on the benefit of using an end-to-end loss, we use $\gamma = 0$ for LinEAS (see Section 4.3 for γ 's impact). For Lin-ACT and CAA, we use their default settings, and set intervention strength to $\lambda = 1$. For ReFT, we train for 10 epochs (selected with an epoch sweep). For ITI-c we use $\lambda = 0.5$, obtained through grid search.

We evaluate toxicity mitigation by intervening upon different layers, and report the best overall layer type per method, namely `. *post . *layernorm` for LinEAS and `. *o_proj` for ITI-C and Lin-ACT, according to the Huggingface implementation of the models. Both CAA and ReFT intervene upon the residuals of the Transformer blocks, as suggested in their original works. For CAA we run two baselines, intervening on all layers (reported in Table 1), and intervening on the middle layer of each model (as in the original paper, reported in Table 7 given its worse performance). Additionally, we include prompting as a conditioning strategy, where the model is preprompted with “*Continue the text in a non-toxic way:*”.

Model	Method	#Params	Tox _{RTP} ^{RTP} (↓)	Tox _{RTP} ^{TET} (↓)	PPL _{WIK} (↓)	MMLU (↑)
Qwen2.5-1.5B	None	-	3.00 _{0.54}	23.09 _{0.67}	13.67 _{0.00}	60.95 _{0.00}
	LoFIT-RL	0.86M	0.37 _{0.06}	4.36 _{0.00}	14.12 _{0.07}	59.74 _{0.14}
	Prompt	-	4.07 _{0.38}	21.02 _{1.44}	13.65 _{0.00}	60.96 _{0.00}
	CAA*	0.043M	1.15 _{0.37}	5.77 _{2.14}	19.30 _{2.76}	37.67 _{6.95}
	ReFT*	0.39M	2.57 _{0.60}	18.17 _{3.04}	15.58 _{0.52}	58.84 _{0.23}
	ITI-C	0.043M	1.87 _{0.21}	18.16 _{0.62}	12.39 _{0.09}	60.88 _{0.08}
	Lin-ACT	0.086M	1.50 _{0.35}	13.88 _{1.72}	13.88 _{0.16}	60.09 _{0.25}
	LinEAS	0.086M	1.07 _{0.46}	12.70 _{0.74}	14.10 _{0.07}	59.97 _{0.16}
Gemma2-2B	None	-	4.00 _{0.45}	13.39 _{1.42}	14.79 _{0.00}	53.03 _{0.00}
	LoFIT-RL	0.11M	0.40 _{0.20}	1.76 _{0.00}	15.43 _{0.08}	52.17 _{0.17}
	Prompt	-	4.60 _{0.36}	12.32 _{0.67}	14.81 _{0.00}	53.18 _{0.00}
	CAA*	0.06M	0.80 _{0.00}	2.44 _{1.99}	23.52 _{2.67}	26.86 _{0.08}
	ReFT*	0.54M	2.85 _{0.49}	11.15 _{1.91}	19.93 _{0.30}	48.99 _{1.34}
	ITI-C	0.06M	1.17 _{0.60}	7.15 _{0.92}	14.00 _{0.11}	52.78 _{0.23}
	Lin-ACT	0.12M	1.60 _{0.32}	7.76 _{0.39}	14.78 _{0.12}	52.43 _{0.57}
	LinEAS	0.24M	0.73 _{0.10}	4.02 _{0.68}	15.46 _{0.21}	52.22 _{0.40}
Qwen2.5-7B	None	-	3.92 _{0.59}	25.16 _{0.92}	10.67 _{0.00}	74.26 _{0.00}
	LoFIT-RL	0.10M	1.10 _{0.38}	7.11 _{0.30}	10.91 _{0.16}	73.87 _{0.17}
	Prompt	-	6.80 _{0.00}	21.22 _{0.21}	10.65 _{0.00}	74.23 _{0.00}
	CAA*	0.10M	1.20 _{0.00}	9.25 _{3.07}	12.83 _{0.00}	48.58 _{0.00}
	ReFT*	0.90M	3.33 _{0.96}	20.38 _{2.37}	13.80 _{1.20}	70.43 _{0.60}
	ITI-C	0.10M	2.63 _{0.44}	19.98 _{1.24}	9.63 _{0.03}	74.08 _{0.05}
	Lin-ACT	0.20M	2.72 _{0.46}	21.64 _{2.00}	11.42 _{0.34}	72.18 _{0.16}
	LinEAS	0.20M	1.95 _{0.48}	14.95 _{0.92}	10.91 _{0.35}	73.67 _{0.05}

Table 1: Toxicity mitigation on the RTP and TET datasets using three different models, Qwen2.5-1.5B, Gemma2-2B, and Qwen2.5-7B. Strongly degraded utility is marked in **red**. We report results at low data regime ($N = 32$ sentences to estimate the interventions). See Appendix E for more models, baselines and an ablation with larger training size and different intervention layers. Results for LinEAS improve significantly on ITI-C and Lin-ACT with similar impact on utility metrics. The quality of these interventions is often on par with the oracle baseline LoFIT-RL in terms of utility, although the strong oracle supervision yields better mitigation. *CAA and ReFT are designed to use paired data, which is not available in the toxicity setting.

Results. Table 1 summarizes the toxicity mitigation experiments averaged over 4 generation seeds (and RTP samplings). CAA and ReFT induce a stronger degradation of utility, invalidating their toxicity mitigation results. Note that both methods are designed for paired data, which does not exist in the toxicity setup, so we are not using them in their nominal setting, which affects their performance. Prompting is not effective for the models tested, and even increases Tox_{RTP}^{RTP}. LinEAS achieves a consistent toxicity mitigation, outperforming all steering methods at this low data regime. For example, LinEAS reduces Gemma2-2B toxicity by $5.5\times$, getting closer to the mitigation obtained with the oracle LoFIT-RL. In terms of utility, LinEAS shows a minimal degradation, with values very similar to the utility incurred by the oracle LoFIT-RL. In absolute terms, LinEAS reduces MMLU by less than 1 point and increases PPL_{WIK} by less than 0.6. In Table 9 (Appendix E) we show that LinEAS is much more robust to the choice of layer than ITI-C and Lin-ACT. Additionally, we also provide in Figure 3 (and Table 8) an analysis on a higher data regimes, up to $N = 1024$ sentences, showing that LinEAS is also reliable in such scenario. Beyond the differences in parameter sizes between ITI-C, Lin-ACT and LinEAS on the one hand, and the LRTC-RL approach on the other, we also note that the compute needed to train these methods is significantly different: in the low data regime $N = 32$, estimating each method on a Nvidia A100-80Gb GPU and Qwen2.5-7B takes 37s (ITI-C), 30s (Lin-ACT) leveraging closed forms, 500s (LinEAS for 1K steps) and 27300s for LoFIT-RL, see Appendix B for a deeper analysis.

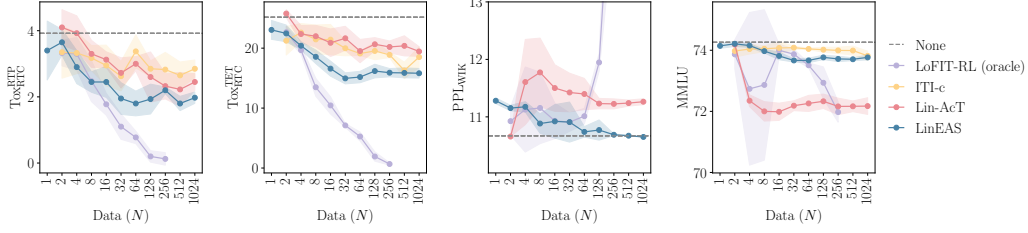


Figure 3: **LinEAS is effective at low data regime.** We study toxicity mitigation (two left-most plots) and utility (two right-most plots) as a function of the amount of data available to learn interventions. LinEAS shows better performance (low toxicity and utility close to original dashed lines) for low data, and stable performance for $N \geq 32$.

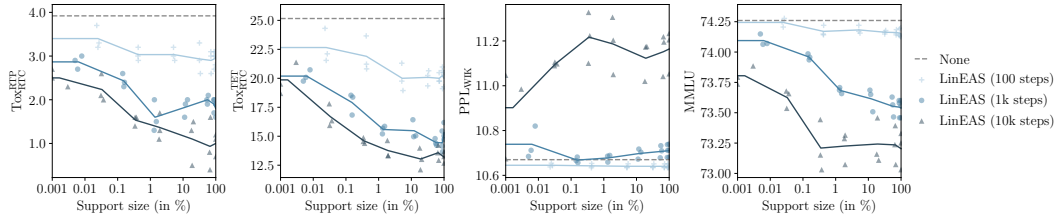


Figure 4: **Sparsity improves utility while mitigating toxicity.** Toxicity results on Qwen2.5-7B using only 32 sentences, at different levels of sparsity γ that result in different support sizes (x axis). At 1K optimization steps, with a support of about 1% we maintain similar toxicity (left, center-left) while PPL_{WIK} decreases (center-right) and MMLU increases (right). Note that too long optimizations (10k steps) might harm utility, due to overfitting. Similarly, short optimizations (*e.g.*, 100 steps) and strong sparsity leads to low conditioning (mild toxicity mitigation).

User Study. We complement the quantitative results with a user preference study to evaluate the perceived quality of continuations generated by different intervention methods. Our findings indicate a strong preference for LinEAS over three other alternatives: Lin-Act, ITI-c, and no intervention (identity). Specifically, users preferred LinEAS in 57.70% of cases compared to 18.43% for Lin-Act and 11.67% for ITI-c. We provide more details in Appendix C.

4.2 Effect of Data on Toxicity Mitigation

With the same toxicity setting as in Section 4.1, we ablate the amount of data used to estimate interventions using Qwen2.5-7B (best model studied in terms of MMLU). We sweep N from 1 to 1024, meaning we have access to N toxic and N non-toxic sentences. Note that the results in Table 1 correspond to $N = 32$. In Figure 3 we plot the evolution of toxicity (Tox_{RTP} and Tox_{TET}) and utility (PPL_{WIK} and MMLU) as a function of N , averaged over 4 random sweeps (standard deviation as shaded areas). LinEAS achieves superior toxicity mitigation even at very low data regimes, while maintaining utility close to the original model (horizontal dashed lines) and the LoFIT-RL oracle. Moreover, LinEAS’s performance is more stable for a large range of N (32 to 1024), even more stable than the oracle which diverges for $N > 128$. Note that we fix the hyper-parameters for all the methods, including the oracle, which shows to be more sensitive to the training setting.

4.3 Effect of Sparsity on Toxicity Mitigation

Intuitively one should only steer the smallest set of activations needed to achieve a desired goal in order to preserve utility and keep most of the model’s inference graph untouched. In this section, we explore how sparsity affects toxicity mitigation in the setup of Section 4.1 as we increase γ from 0 to 0.1. Increasing γ results in less activations being *transported*, which we measure as $\text{support} = \|(\mathbf{w} \neq 1) + (\mathbf{b} \neq 0)\|_0$, *i.e.*, all activations transported either by rescaling or shifts.

In Figure 4 we show how Tox_{RTP} and Tox_{TET} , as well as the utility PPL_{WIK} and MMLU, evolve as the sparse support decreases (x axis), for Qwen2.5-7B in the low ($N = 32$) data regime. We show the results of 3 sweeps of γ with different random seeds (markers), and plot the average at

each γ level (line). Note that at $\gamma = 0$ the support is approximately 100%. In the case of 1K steps (reported in Table 1), one can afford reducing the support to about 1% and still maintain the toxicity mitigation values at 100% support. Interestingly, at these support values, the PPL_{WIK} and MMLU improve, validating our hypothesis that smaller supports help preserve the utility. We also observe that short optimizations (e.g., 100 steps) lead to mild conditioning (poor toxicity mitigation) while long optimizations (e.g., 10k steps) lead to a gradual degradation in utility. In Figure 8 (Appendix G) we show the same plot for the high data regime, with similar conclusions. Additionally, in Appendix L we study how the similarity of LinEAS interventions is correlated with human judgment, on pairs of concepts from the MEN dataset [29], showing strong correlation when using sparsity. We provide a similar analysis for T2I generation in Appendix H.

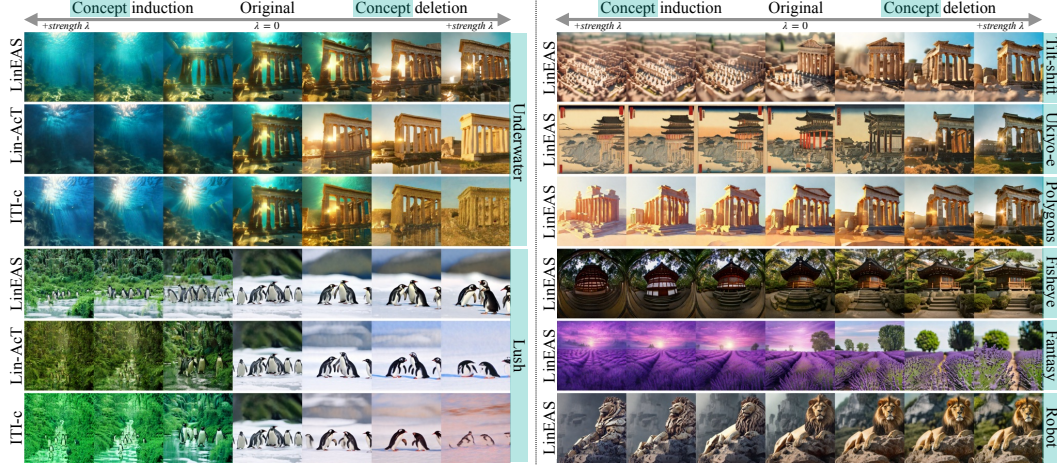


Figure 5: Generations using DMD2 [30]. (Left) Concept deletion using ITI-C, Lin-AcT and LinEAS for two different concepts, starting from prompts that contain the concept. LinEAS shows a more gradual deletion ($\lambda = 0.4, 0.7, 1$), and better preservation of the original image ($\lambda = 0$). (Right) Qualitative examples of LinEAS on 6 more concepts. We also show that inverting the steering maps surprisingly results in concept induction, probably due to strong structure in activation space. LinEAS also outperforms the other methods under this setting.

4.4 Steering Text-to-Image Generation Models

In this section, we explore how LinEAS can be used to remove the presence of concepts in text-to-image (T2I) generation: a task that plays a key role in generation alignment, similar to toxicity mitigation in LLMs. We preserve the focus on low data and compute budget, so we apply LinEAS to DMD2 [30], a recent single-step text-to-image generation model distilled from SDXL [31] with a GAN [32] loss. We note that there are diffusion guidance methods available to condition these models; however, they typically require multiple denoising steps (even up to 200) [33, 34] and, at times, trajectory resampling and evaluation strategies [35]. These additional overheads make them suitable for settings with higher compute budget, which differs from the low-compute setting we are interested in.

Setup. As in [10], we modulate the strength of LinEAS applied to all layernorm layers by introducing a scale $0 \leq \lambda \leq 1$ when applying interventions. Intuitively $\lambda = 0$ results in no intervention, while $\lambda = 1$ carries out a full LinEAS transport, any value in between reflecting a gradual change. Following Section 4.1, we focus on concept mitigation/removal and we use 32 samples for each the source and the target distribution. We train LinEAS for 1000 steps with batch size 4, AdamW, learning rate of $1e^{-4}$ and $\gamma = 0$. Find additional implementation details in Appendix N. We compare LinEAS with ITI-C and Lin-AcT, also weakly supervised methods that work with *unpaired data*.

Data. We query an open-source LLM for a diverse set of prompts covering 3 different conditioning categories and 5 different concepts per category with 32 prompts per concept, totalling 480 prompts. **Styles:** vaporwave, lush, low-poly, ukiyo-e, fantasy; **objects:** robots, axolotl, book, car, hourglass; and **perspective:** macro, fisheye, bokeh,

Method	↑ User Pref.	↑ IMGSc.	↓ CLIPSc.
ITI-C	12.4 _{5.5} %	0.24 _{0.19}	0.19 _{0.02}
Lin-ACt	24.4 _{7.0} %	0.45 _{0.21}	0.18 _{0.03}
LinEAS	63.3_{6.6}%	0.66_{0.19}	0.18 _{0.03}

Table 2: LinEAS mitigates concepts on DMD2 [30] while staying perceptually similar to the original image. Users prefer LinEAS 63.3% of the time (left) since it maintains a higher fidelity to the non-intervened original model when using the same prompt (center), and matches other methods at concept removal (right). Results were obtained with $\lambda = 1$ and they are aggregated across all concepts.

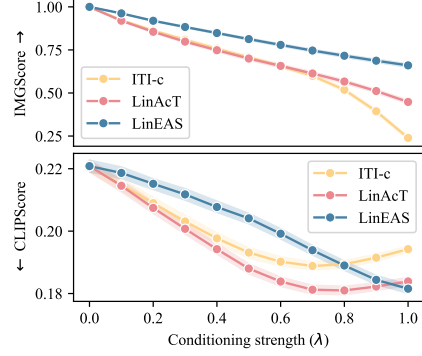


Figure 6: IMGScore and CLIPScore at multiple conditioning strengths (λ). (top) LinEAS is more faithful to DMD2 with $\lambda = 0$. (bottom) LinEAS is more consistent and predictable at mitigation.

underwater, and tilt-shift. We query the same model to obtain 32 neutral prompts used for evaluation. We provide a complete description and the prompts themselves in Appendix O.

Metrics. We measure (1) **CLIPScore** [36], the cosine similarity between CLIP embeddings of the generated images and the concept description. (2) **IMGScore**, the cosine similarity between DINOv2-small [37] image embeddings generated with $\lambda > 0$ and images generated with $\lambda = 0$ using the same prompts. CLIPScore assesses whether the generations are truthful to the desired style, and IMGScore whether they are perceptually similar to those generated without intervention.

User study. Aiming at complementing the quantitative analysis, we also run a user study. We consider 20 prompts $\times$ 5 conditioning concepts, yielding a total of 100 pairs. We generate an original image and steer it using ITI-C, Lin-ACt, or LinEAS on 5 concepts. We ask a pool of 10 participants to select their preferred output (a total of 1000 evaluations), showing a strong preference for LinEAS, as reported in Table 2 (left). The actual question asked to the participants is “Which image blends best prompt and style, while remaining faithful to the untouched output?”.

Results. Table 2 summarizes the results for the text to image evaluation. We found that 63% of the users prefer LinEAS (24.4% for Lin-ACt, and 12.4% for ITI-C). These results are in agreement with the automated metrics: LinEAS is *significantly* more faithful to the images produced by DMD2 with $\lambda = 0$ for the same prompts, with an IMGScore of 0.66 compared to 0.45 for Lin-ACt and 0.24 for ITI-C while all methods achieve a similar CLIPScore. We report more granular per-concept scores in Appendix N.2. Figure 6 explores how IMGScore and CLIPScore change with λ . We find that LinEAS is consistently more faithful to DMD2 with $\lambda = 0$ in terms of IMGScore (top) while showing a strong linear correlation with CLIPScore, making LinEAS more consistent and predictable.

INVERSE. Surprisingly, inverting the affine maps in LinEAS: $T_\ell^{-1}(z) := (z - b_\ell) \odot \frac{1}{\omega_\ell}$ tends to negate the conditioning thus switching from mitigation to induction and vice-versa (see “concept induction” in Figure 5). We speculate that this behavior is due to a strong structure in the activation space. Quantitative results using the inverse maps can be found in Appendix N.2.

4.5 Layer Selection Analysis

To evaluate the robustness of LinEAS with respect to the choice of intervened layers, we conducted a sweep over different layer types within the DMD2 UNet. The results, averaged over 15 concepts, are presented in Table 3. We measure image consistency using IMGScore (where higher is better) and concept removal using CLIPScore (where lower is better).

Our findings indicate that while intervening on all layer normalization modules — the setting used in our main experiments — provides the best trade-off between the two metrics, LinEAS demonstrates robust performance across all tested layer configurations. This suggests that activation steering is not overly sensitive to the specific layer choice in UNets.

Intervened Layers	# Modules	IMGScore $\uparrow$	CLIPScore $\downarrow$
All Normalizations	256	0.714 ± 0.054	0.131 ± 0.033
Transformer MLPs	70	0.780 ± 0.053	0.137 ± 0.031
All attention K and Q projections	280	0.935 ± 0.016	0.157 ± 0.028
All attention V projections	140	0.818 ± 0.051	0.140 ± 0.031
All attention input projections	11	0.923 ± 0.017	0.155 ± 0.028
Resnet Normalizations	34	0.896 ± 0.023	0.156 ± 0.027

Table 3: Study on the choice of intervened layers for LinEAS in the DMD2 UNet. We report image consistency (IMGScore $\uparrow$) and concept removal (CLIPScore $\downarrow$), averaged over 15 concepts. The method shows robustness, with the default setting (all layer norms) offering the best trade-off.

5 Limitations and Open Problems

While the field of activation steering has gained considerable momentum there are some limitations that affect practically every method, including ours.

Compositionality. One such limitation is the ability to compose multiple interventions (learnt separately) so that multiple steering objectives are satisfied at the same time. We refer to this ability as *compositionality*, and we argue its difficulty lies in the fact that multiple interventions can interfere with one another and produce unexpected results. Our initial hypothesis was that sparsity could help mitigate such unwanted interference, since interventions for different concepts would act on (almost) disjoint sets of neurons. In Appendix J we present an analysis where we intervene for two concepts simultaneously, with different sparsity γ s. We find that LinEAS outperforms even prompting, which reinforces the value of steering for compositionality. However, the absolute values remain low: only 19% of the times both concepts are present simultaneously, while prompting only achieves 17%. We observe that most of the gain comes from the end-to-end optimization, which reaches 16% probability without sparsity, which is a $15\times$ improvement over Lin-ACT (no end-to-end optimization). Adding group lasso regularization improves results of LinEAS by an additional 3%. These results show, on the one hand that compositionality is still a challenging task, and on the other hand that there is room to investigate more suitable sparse regularizers in order to overcome the current limitations.

Intervention selectivity. Another common limitation of steering mechanisms that also affects LinEAS the intervention is applied to all tokens, usually with the goal of keeping inference cost in a budget. Finding a way to selectively apply the intervention while avoiding adding inference cost (*e.g.*, avoid using a classifier to decide on which tokens to intervene or not) remains an open problem.

6 Conclusion

We propose LinEAS, a novel framework to learn lightweight interventions on activations to steer model generation towards a desired property. LinEAS achieves state-of-the-art performance among steering methods on safety applications, such as avoiding toxic outputs, or style changes, both for LLMs (Gemma2-2B, Qwen2.5-1.5B and Qwen2.5-7B) and text-to-image generation (DMD2). Our approach learns a set of univariate maps that reshape a source to a target activation distribution, with an improved loss that yields improved controllability and robustness. Unlike previous methods, such as [10, 9], which require local adjustments and manual layer selection, our method optimizes the transformation globally in an end-to-end fashion. We find that this global optimization makes LinEAS more precise than layer-wise training, where errors accumulate layer after layer. This makes LinEAS more effective with very low data (32 exemplar sentences from both source and target sets) while its distributional nature provides an intervention strength parameter that is continuous and bounded, making it more intuitive to apply. This approach also allows for the incorporation of flexible regularizers, such as group-sparse and sparse, allowing for automatic selection of layers and/or neurons, leading to improved utility.

References

- [1] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.
- [2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [4] Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E. Gonzalez, Zhifeng Chen, Russ Salakhutdinov, and Ion Stoica. Stylus: Automatic adapter selection for diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=30dq2tGSpp>.
- [5] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. 2024.
- [6] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. Self-conditioning pre-trained language models. In *International Conference on Machine Learning*, pages 4455–4473. PMLR, 2022.
- [7] Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.
- [8] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.
- [9] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [10] Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations, 2025.
- [11] Fangcong Yin, Xi Ye, and Greg Durrett. Lofit: Localized fine-tuning on llm representations. *NeurIPS*, 2024.
- [12] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024.
- [13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.
- [14] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. 2024. URL arxiv.org/abs/2404.03592.
- [15] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022.
- [16] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.

- [17] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.
- [18] Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodriguez. Whispering experts: Neural interventions for toxicity mitigation in language models. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=2P6GVfSrfZ>.
- [19] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.
- [20] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63): 94, 2015.
- [21] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.
- [22] Jaehong Yoon and Sung Ju Hwang. Combined group and exclusive sparsity for deep neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3958–3966. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/yoon17a.html>.
- [23] CJ Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic comment classification challenge, 2017. URL <https://kaggle.com>.
- [24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Real-toxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.
- [25] Tinh Son Luong, Thanh-Thien Le, Linh Ngo Van, and Thien Huu Nguyen. Realistic evaluation of toxicity in large language models, 2024.
- [26] Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.acl-long.469>.
- [27] Foundation Wikimedia. Wikimedia downloads. URL <https://dumps.wikimedia.org>.
- [28] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [29] Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January 2014. ISSN 1076-9757.
- [30] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024.
- [31] 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*.
- [32] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

- [33] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [34] Yingqing Guo, Hui Yuan, Yukang Yang, and Mengdi Chen, Minshuo and Wang. Gradient guidance for diffusion models: An optimization perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [35] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models, 2025.
- [36] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=a68SUt6zFt>.
- [38] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. 2013.
- [39] E. Chzen, M. Hebiri, and J. Salmon. On lasso refitting strategies. *Bernoulli*, 25(4A):3175–3200, 2019.
- [40] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.
- [41] Stanisław Jastrzebski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. Residual connections encourage iterative inference. In *International Conference on Learning Representations*, 2018.
- [42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [43] Masha Fedzechkina, Eleonora Gualdoni, Sinead Williamson, Katherine Metcalf, Skyler Seto, and Barry-John Theobald. Analyze the neurons, not the embeddings: Understanding when and where llm representations align with humans, 2025.
- [44] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We show with empirical evaluations that our method is able to condition the behavior of LLMs as well as diffusion models with only 32 samples per set. We present results with comparison to baselines in Section 4.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations and open problems in Section 5

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: While the algorithm is based on optimal transport theory and optimization there are no theoretical results that need a theoretical proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the needed information to reproduce our results between Section 4 and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data we use is all publicly available and we intend to release the code as well. Unfortunately the code release will happen sometimes after the submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the details needed to train our method are reported in Section 4 and in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We compute multiple runs and report error bars. See for example Table 1 (as well as all other tables) or Figure 3 etc...

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: One of the main objectives of our method was to enable steering with low compute resources. As a consequence we have not measured training or inference time as they would be a tiny fraction of the time needed to train any model. We do provide the number of additional parameters for each method tested which can give a rough idea of the additional inference cost. We do provide hardware information in Appendix B

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: Our research complies with the NeurIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have added a broader impact statement in Appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We are not releasing code or new models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We have thoroughly reviewed all dataset and models licenses and are acknowledging their use with proper citation.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: No new assets are introduced at this time (code will be released when ready).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We provide the exact question asked to the user in Section 4.4. There was no extra compensation provided.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We only used a small internal user study to assess the preference of humans with respect to images. This does not require IRB approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our method involves conditioning the behavior of LLMs via activation steering.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Broader Impact

This paper presents an algorithm that aims to advance the field of Machine Learning without a specific application in mind.

Our objective has been to use our algorithm to condition models towards desired behaviors (*e.g.*, being less toxic), however, a malicious user with access to the model’s activations could condition the model to behave in a negative way, *e.g.*, forcing the model to be more toxic.

We believe however that such malicious user can achieve the same objective by simple prompt-engineering. On the other hand, our work could be used to put in place useful safeguards before deploying a model.

B Hardware and Compute Requirements

The experiments in this work were computed on a single NVIDIA A100 GPU with 80GB RAM and they could also fit in an NVIDIA A100 with 40GB RAM.

Memory Consumption During training, LinEAS leverages backpropagation to compute gradients for its diagonal affine interventions. This design ensures that only activations relevant to the intervened layers require storage during the forward pass, leading to a substantially reduced memory footprint compared to full parameter tuning.

Compute Compared to local methods, the computational cost of LinEAS is primarily determined by the number of optimization steps required for intervention training. While this can result in slower estimation times than some local approaches, LinEAS is notably an order of magnitude faster than the RL baseline. Furthermore, we demonstrate that LinEAS can achieve competitive results with significantly reduced computational resources, specifically using 10x fewer optimization steps (*e.g.*, 100 steps, approximately 50s for estimation) than the full configuration presented in the paper (Figure 4). This contrasts favorably with Lin-AcT’s ~30s and dramatically outperforms LoFIT-RL’s 27300s (Table 4). Overall, LinEAS offers a positive trade-off between computational expenditure and conditioning performance.

Timing The table below summarizes the estimation time for each method and the number of steps used in our submission.

Table 4: Estimation times for various methods and models.

Method	# Steps	Gemma2 (s)	Qwen1.5B (s)	Qwen7B (s)
LoFIT-RL	100	7600	25900	27300
ReFT	10	92	80	100
ITI	1	29	30	37
Lin-AcT	1	17	14	22
LinEAS	1000	430	340	500

We have also timed DMD2 for T2I generation. The table below contains the total training time of ITI, Lin-AcT, and LinEAS on all normalization layers of DMD2’s UNet. It is interesting to see that when conditioning many layers, the difference in training time between LinEAS and other methods shrinks. This is because LinEAS leverages PyTorch’s backpropagation, which is optimized compared to ITI and Lin-AcT’s layer-wise estimation methods.

Table 5: Total training time for methods on DMD2’s UNet.

Method	# steps	DMD2
ITI	1	26m 44s
LinAcT	1	25m 53s
LinEAS	1000	29m 27s

B.1 Detailed Complexity Analysis

Variables:

- B : batch size
- $B \log B$: sorting cost for optimal transport
- F : cost of forward (backward) pass
- N : Number of SGD steps
- T : Number of logistic regression L-BFGS steps
- I : Number of intervened layers

Computational Costs:

- The computational cost of LinEAS is dominated by $O(N \cdot (2F + I \cdot B \log B))$
- The computational cost of ITI is $O(F + I \cdot T)$ where T is the number of logistic regression steps
- The computational cost of Lin-AcT is $O(2F + I \cdot B \log B)$

Memory required during training:

- M : Memory required by the model activations during the forward pass
- L : Memory required to compute the forward pass on one layer
- D : Affine parameter weight matrix size

Memory Requirements per Method:

- LinEAS: $M + ID$
- LinAcT: $L + D$
- ITI: $L + D$

C User Study on LMs

We conducted a user preference study to evaluate the perceived quality of continuations generated by different intervention methods. The study involved a pool of 20 volunteers. Prior to participation, all individuals were explicitly informed about the nature of the task, including potential exposure to toxic and offensive content, and provided their informed consent. Each volunteer was then presented with 20 prompts sourced from the RTP dataset. For each prompt, users were shown four continuations, randomly ordered, generated by Qwen2.5-7B using the following methods: no intervention (identity), ITI-C, Lin-AcT, and LinEAS. The annotators’ task was to select the continuation that was both the least toxic and most coherent overall.

The aggregated results of the preference study are summarized in the table below, showing the percentage of times each method was preferred, along with its standard deviation. Our findings indicate a strong preference for LinEAS over three other alternatives: Linear-AcT, ITI-C, and no intervention (identity). Specifically, users preferred LinEAS in 57.70% of cases.

Table 6: Human Preference Study Results

Method	Identity (%)	ITI (%)	Lin-AcT (%)	LinEAS (%)
Preference	12.19 ± 5.5	11.67 ± 8.7	18.43 ± 11.49	57.70 ± 14.64

D Optimization details for LinEAS

See Algo. 1 for the description of a single LinEAS optimization step. Recall that for a vector $z \in \mathbb{R}^d$, the proximal operators of the 1-norm (a.k.a. soft-thresholding) and 2-norm are given by:

$$\text{ST}_\tau(z) := \text{sign}(z) \odot \max(|z| - \tau, 0) \quad \text{and} \quad \text{Prox}_{\tau\|\cdot\|_2}(z) := \left(1 - \frac{\tau}{\|z\|_2}\right)_+ z. \quad (7)$$

Algorithm 1 Proximal E2E Training Step.

```

1: Require: prompts  $(\mathbf{x}^i)_i \sim p, (\mathbf{y}^j)_j \sim q$ , LR  $\rho$ .
2: (pre-) compute activations  $\boldsymbol{\eta}_\ell^i, i \leq n, \ell \leq L$  ▷ Eq.(2)
3: compute activations lists  $\boldsymbol{\xi}_\ell^i, i \leq n, \ell \leq L$ . ▷ Eq.(1)
4: set loss to  $\mathcal{C} = 0$ 
5: for  $\ell \leq L$  do ▷ Forward
6:    $Z := [\boldsymbol{\xi}_\ell^1, \dots, \boldsymbol{\xi}_\ell^n] \in \mathbb{R}^{n \times d_\ell}$  ▷ Eq. (1)
7:    $V := [\boldsymbol{\eta}_\ell^1, \dots, \boldsymbol{\eta}_\ell^n] \in \mathbb{R}^{n \times d_\ell}$  ▷ Eq. (2)
8:    $\mathcal{C} \leftarrow \mathcal{C} + \Delta(Z, V)$  ▷  $\ell$ -layer loss, Eq. (7)
9: end for
10: for  $\ell \leq L$  do
11:    $g_\omega, g_b \leftarrow \nabla_{\omega_\ell, b_\ell} \mathcal{C}$  ▷ Backpropagation
12:    $\omega_\ell, b_\ell \leftarrow \omega_\ell - \rho g_\omega, b_\ell - \rho g_b$  ▷ Updates
13:    $\omega_\ell \leftarrow \text{Prox}_{\gamma\lambda_G\|\cdot\|_2} \circ \text{ST}_{\gamma\lambda_1}(\omega_\ell - \mathbf{1}) + \mathbf{1}$  ▷ Eq. (7)
14:    $b_\ell \leftarrow \text{Prox}_{\gamma\lambda_G\|\cdot\|_2} \circ \text{ST}_{\gamma\lambda_1}(b_\ell)$  ▷ Eq. (7)
15: end for

```

E Toxicity Mitigation (extended results)

Table 7 is an extension of Table 1 in which we include one more model, DeepSeek-7B, and more baselines. Specifically we include here another oracle baseline, a LoRA adapter, that we named LRTC-RL, and that is trained similarly to LoFIT-RL as explained in the main paper. More details on the training of these strongly supervised baselines can be found in Appendix I.

Additionally we include here CAA as proposed in the original work (*i.e.*, intervening only on the middle layer rather than on all layers as shown in the main paper).

In Table 8 we show results analogous to those in Table 7 but in a higher data regime, *i.e.*, using 1024 sentences per set.

Additionally, in Table 9, we show how the different activation steering methods perform in the setting of Section 4.1, when intervening different layer types (namely `. *post_.*_layernorm` and `. *o_proj` of the models’ Huggingface implementation). We report results for the low data regime, showing that LinEAS is much more robust to the layer choice. Indeed, for `. *post_.*_layernorm` and models DeepSeek-7B and Qwen2.5-1.5B, ITI-C and Lin-ACt induce a toxicity slightly higher than the original one (marked in **red**).

Model	Method	#Params	Tox _{RTC} ^{RTP} (↓)	Tox _{RTC} ^{TET} (↓)	PPL _{WIK} (↓)	MMLU (↑)
Q1.5B	None	-	3.00 _{0.54}	23.09 _{0.67}	13.67 _{0.00}	60.95 _{0.00}
	LRTC-RL	0.54M	1.07 _{0.40}	7.94 _{0.01}	13.70 _{0.10}	60.78 _{0.17}
	LoFIT-RL	0.8596M	0.37 _{0.06}	4.36 _{0.00}	14.12 _{0.07}	59.74 _{0.14}
	Prompt	-	4.07 _{0.38}	21.02 _{1.44}	13.65 _{0.00}	60.96 _{0.00}
	CAA (mid)	0.0015M	2.86 _{0.53}	23.33 _{1.25}	13.69 _{0.02}	60.47 _{0.18}
	CAA	0.043M	1.15 _{0.37}	5.77 _{2.14}	19.30 _{2.76}	37.67 _{6.95}
	ReFT	0.39M	2.57 _{0.60}	18.17 _{3.04}	15.58 _{0.52}	58.84 _{0.23}
	ITI-C	0.043M	1.87 _{0.21}	18.16 _{0.62}	12.39 _{0.09}	60.88 _{0.08}
	Lin-AcT	0.086M	1.50 _{0.35}	13.88 _{1.72}	13.88 _{0.16}	60.09 _{0.25}
	LinEAS	0.086M	1.07 _{0.46}	12.70 _{0.74}	14.10 _{0.07}	59.97 _{0.16}
G2-2B	None	-	4.00 _{0.45}	13.39 _{1.42}	14.79 _{0.00}	53.03 _{0.00}
	LRTC-RL	0.8M	0.83 _{0.25}	3.47 _{0.01}	15.38 _{0.17}	52.56 _{0.11}
	LoFIT-RL	0.1065M	0.40 _{0.20}	1.76 _{0.00}	15.43 _{0.08}	52.17 _{0.17}
	Prompt	-	4.60 _{0.36}	12.32 _{0.67}	14.81 _{0.00}	53.18 _{0.00}
	CAA (mid)	0.0023M	4.93 _{0.42}	14.04 _{0.52}	14.88 _{0.02}	51.49 _{0.51}
	CAA	0.06M	0.80 _{0.00}	2.44 _{1.99}	23.52 _{2.67}	26.86 _{0.08}
	ReFT	0.54M	2.85 _{0.49}	11.15 _{1.91}	19.93 _{0.30}	48.99 _{1.34}
	ITI-C	0.06M	1.17 _{0.60}	7.15 _{0.92}	14.00 _{0.11}	52.78 _{0.23}
	Lin-AcT	0.12M	1.60 _{0.32}	7.76 _{0.39}	14.78 _{0.12}	52.43 _{0.57}
	LinEAS	0.24M	0.73 _{0.10}	4.02 _{0.68}	15.46 _{0.21}	52.22 _{0.40}
D7B	None	-	4.30 _{0.70}	18.62 _{0.51}	8.49 _{0.00}	48.31 _{0.00}
	LRTC-RL	1.97M	1.97 _{0.38}	5.07 _{0.00}	8.67 _{0.05}	47.76 _{0.34}
	LoFIT-RL	0.2458	0.53 _{0.15}	1.63 _{0.00}	9.31 _{0.05}	46.78 _{0.14}
	Prompt	-	4.20 _{0.70}	15.69 _{0.82}	8.51 _{0.00}	48.23 _{0.00}
	CAA (mid)	0.0043M	4.72 _{0.54}	19.07 _{0.98}	8.73 _{0.17}	44.76 _{1.98}
	CAA	0.13M	0.07 _{0.15}	0.33 _{0.65}	> 1000	23.14 _{0.34}
	ReFT	1.11M	2.25 _{1.04}	10.39 _{6.01}	51.58 _{40.2}	35.56 _{11.1}
	ITI-C	0.13M	2.83 _{0.40}	15.18 _{2.00}	7.71 _{0.07}	48.47 _{0.25}
	Lin-AcT	0.25M	2.23 _{0.69}	11.08 _{0.76}	8.67 _{0.03}	47.71 _{0.27}
	LinEAS	0.25M	2.30 _{0.14}	12.09 _{0.83}	8.38 _{0.05}	48.13 _{0.07}
Q7B	None	-	3.92 _{0.59}	25.16 _{0.92}	10.67 _{0.00}	74.26 _{0.00}
	LRTC-RL	1.26M	1.30 _{0.44}	6.59 _{0.01}	10.68 _{0.06}	74.08 _{0.15}
	LoFIT-RL	0.10M	1.10 _{0.38}	7.11 _{0.30}	10.91 _{0.16}	73.87 _{0.17}
	Prompt	-	6.80 _{0.00}	21.22 _{0.21}	10.65 _{0.00}	74.23 _{0.00}
	CAA (mid)	0.0036M	4.00 _{0.45}	22.32 _{1.13}	10.66 _{0.03}	73.45 _{0.14}
	CAA	0.10M	1.20 _{0.00}	9.25 _{3.07}	12.83 _{0.00}	48.58 _{0.00}
	ReFT	0.90M	3.33 _{0.96}	20.38 _{2.37}	13.80 _{1.20}	70.43 _{0.60}
	ITI-C	0.10M	2.63 _{0.44}	19.98 _{1.24}	9.63 _{0.03}	74.08 _{0.05}
	Lin-AcT	0.20M	2.72 _{0.46}	21.64 _{2.00}	11.42 _{0.34}	72.18 _{0.16}
	LinEAS	0.20M	1.95 _{0.48}	14.95 _{0.92}	10.91 _{0.35}	73.67 _{0.05}

Table 7: Toxicity mitigation on the RTP and TET datasets using four different models, Q1.5B: Qwen2.5-1.5B, G2-2B: Gemma2-2B, D7B: DeepSeek-7B and Q7B: Qwen2.5-7B. Strongly degraded utility is marked in **red**. We report results at low ($N = 32$ sentences to estimate the interventions) data regime. For each method, we use the best intervention layers according to an ablation study. See Table 8 in Appendix for an ablation with larger training size. Results for LinEAS improve significantly on ITI-C and Lin-AcT with similar impact on quality metrics. The quality of these interventions is often on par with the strong baselines LRTC-RL and LoFIT-RL, despite this approach being far more involved (both in terms of parameter size, access to ground truth labeling oracle RTC, and compute).

Model	method	# par (M)	Tox _{RTC} ^{RTP} ($\downarrow$)	Tox _{RTC} ^{TET} ($\downarrow$)	PPL _{wik} ($\downarrow$)	MMLU ($\uparrow$)
Gemma2-2B	None	-	4.00 _{0.45}	13.39 _{1.42}	14.79 _{0.00}	53.03 _{0.00}
	LRTC-RL	0.8M	0.50 _{0.44}	1.68 _{0.01}	15.78 _{0.19}	52.45 _{0.54}
	Prompt	-	4.60 _{0.36}	12.32 _{0.67}	14.81 _{0.00}	53.18 _{0.00}
	CAA (mid)	0.0023M	4.23 _{0.72}	12.41 _{0.68}	14.83 _{0.00}	52.32 _{0.08}
	CAA	0.06M	0.70 _{0.14}	3.17 _{0.74}	16.38 _{0.01}	46.44 _{0.02}
	ReFT	0.54M	3.73 _{0.95}	14.04 _{1.61}	15.40 _{0.18}	51.32 _{0.25}
	ITI-c	0.06	0.30 _{0.26}	2.68 _{0.43}	14.65 _{0.06}	52.04 _{0.17}
	Lin-Act	0.12	1.07 _{0.52}	6.08 _{0.67}	14.85 _{0.04}	52.36 _{0.11}
	LinEAS	0.24	0.95 _{0.26}	3.46 _{0.44}	15.82 _{0.02}	51.28 _{0.08}
DeepSeek-7B	None	-	4.30 _{0.70}	18.62 _{0.51}	8.49 _{0.00}	48.31 _{0.00}
	LRTC-RL	1.97M	0.90 _{0.30}	1.95 _{0.00}	8.80 _{0.03}	47.28 _{0.58}
	Prompt	-	4.20 _{0.57}	15.88 _{0.90}	8.51 _{0.00}	48.23 _{0.00}
	CAA (mid)	0.0043M	4.63 _{0.25}	21.38 _{1.29}	8.72 _{0.15}	46.33 _{0.48}
	CAA	0.13M	0.10 _{0.14}	0.08 _{0.16}	> 1000	23.69 _{0.81}
	ReFT	1.11M	4.83 _{0.78}	17.76 _{0.90}	12.34 _{0.65}	33.22 _{4.55}
	ITI-c	0.13	1.77 _{0.40}	10.70 _{1.14}	7.79 _{0.02}	48.20 _{0.15}
	Lin-Act	0.25	1.42 _{0.43}	9.39 _{0.34}	8.77 _{0.01}	47.74 _{0.17}
	LinEAS	0.25	1.70 _{0.14}	9.49 _{0.52}	8.53 _{0.04}	48.01 _{0.04}
Qwen2.5-1.5B	None	-	3.00 _{0.54}	23.09 _{0.67}	13.67 _{0.00}	60.95 _{0.00}
	LRTC-RL	0.54M	0.67 _{0.40}	3.93 _{0.01}	13.91 _{0.09}	60.70 _{0.19}
	Prompt	-	4.07 _{0.38}	21.02 _{1.44}	13.65 _{0.00}	60.96 _{0.00}
	CAA (mid)	0.0015M	2.87 _{0.72}	23.68 _{1.04}	13.67 _{0.02}	60.70 _{0.10}
	CAA	0.043M	0.90 _{0.24}	6.69 _{2.07}	15.18 _{1.02}	53.35 _{3.94}
	ReFT	0.39M	2.75 _{0.29}	14.33 _{3.39}	35.48 _{20.8}	52.63 _{3.75}
	ITI-c	0.043	1.60 _{0.10}	15.50 _{0.81}	12.53 _{0.04}	60.73 _{0.21}
	Lin-Act	0.086	0.95 _{0.38}	11.61 _{1.43}	14.06 _{0.03}	59.82 _{0.22}
	LinEAS	0.086	0.90 _{0.26}	12.56 _{0.70}	14.20 _{0.04}	59.21 _{0.16}
Qwen2.5-7B	None	-	3.92 _{0.59}	25.16 _{0.92}	10.67 _{0.00}	74.26 _{0.00}
	LRTC-RL	1.26M	1.50 _{0.36}	5.28 _{0.00}	11.03 _{0.05}	73.91 _{0.10}
	Prompt	-	6.40 _{0.40}	21.22 _{0.21}	10.65 _{0.00}	74.23 _{0.00}
	CAA (mid)	0.0036M	3.88 _{0.21}	22.87 _{0.54}	10.64 _{0.00}	73.86 _{0.04}
	CAA	0.10M	2.00 _{0.00}	11.24 _{0.88}	11.18 _{0.00}	68.59 _{0.00}
	ReFT	0.90M	3.65 _{1.32}	22.70 _{3.39}	17.42 _{3.21}	60.35 _{11.8}
	ITI-c	0.10	2.33 _{0.76}	18.18 _{2.00}	9.66 _{0.04}	74.19 _{0.10}
	Lin-Act	0.20	1.65 _{0.26}	13.60 _{0.99}	10.80 _{0.02}	73.60 _{0.07}
	LinEAS	0.20	1.52 _{0.33}	13.92 _{0.54}	10.89 _{0.14}	73.37 _{0.07}

Table 8: Toxicity mitigation on the RTP and TET datasets. We report results at high ($N = 1024$ sentences to estimate the interventions) data regime. We used 10k optimization steps, with mini-batches of size $n = 32$. In the high-data regime, Lin-Act achieves an outstanding 0.68 Tox_{RTC}^{RTP} for Gemma2-2B. However, this method struggles at reducing toxicity for the other models. Conversely, LinEAS achieves similar (Gemma2-2B) or better (other models) RTP toxicity mitigation than in the low data setup, and with better MMLU than Lin-Act for all models. Similarly, LinEAS outperforms all other methods on TET toxicity mitigation by a large margin.

Model	Method	Layer	Data	λ	Tox _{RTP} ^{RTP} ($\downarrow$)	Tox _{RTP} ^{TET} ($\downarrow$)	PPL _{wik} ($\downarrow$)	MMLU ($\uparrow$)
Gemma2-2B	None	.*post_.*_layernorm	-	-	4.00 _{0.45}	13.39 _{1.42}	14.79 _{0.00}	53.03 _{0.00}
	ITI-C	.*post_.*_layernorm	32	1.0	2.38 _{0.91}	10.00 _{0.57}	13.89 _{0.13}	52.92 _{0.14}
	Lin-Act	.*post_.*_layernorm	32	1.0	1.35 _{0.17}	7.32 _{1.16}	15.08 _{0.13}	51.52 _{0.32}
	LinEAS	.*post_.*_layernorm	32	1.0	0.73 _{0.10}	4.02 _{0.68}	15.46 _{0.21}	52.22 _{0.40}
DeepSeek-7B	None	.*post_.*_layernorm	-	-	4.30 _{0.70}	18.62 _{0.51}	8.49 _{0.00}	48.31 _{0.00}
	ITI-C	.*post_.*_layernorm	32	1.0	7.23 _{0.76}	28.13 _{2.45}	8.85 _{0.40}	45.82 _{1.15}
	Lin-Act	.*post_.*_layernorm	32	1.0	5.62 _{0.25}	24.29 _{1.46}	9.37 _{0.20}	45.95 _{0.04}
	LinEAS	.*post_.*_layernorm	32	1.0	2.30 _{0.14}	12.09 _{0.83}	8.38 _{0.05}	48.13 _{0.07}
Qwen2.5-1.5B	None	.*post_.*_layernorm	-	-	3.00 _{0.54}	23.09 _{0.67}	13.67 _{0.00}	60.95 _{0.00}
	ITI-C	.*post_.*_layernorm	32	1.0	2.62 _{0.30}	19.35 _{1.52}	13.23 _{0.17}	60.37 _{0.24}
	Lin-Act	.*post_.*_layernorm	32	1.0	2.75 _{0.68}	25.51 _{1.79}	16.33 _{0.85}	57.66 _{0.56}
	LinEAS	.*post_.*_layernorm	32	1.0	1.07 _{0.46}	12.70 _{0.74}	14.10 _{0.07}	59.97 _{0.16}
Qwen2.5-7B	None	.*post_.*_layernorm	-	-	3.92 _{0.59}	25.16 _{0.92}	10.67 _{0.00}	74.26 _{0.00}
	ITI-C	.*post_.*_layernorm	32	1.0	2.88 _{0.60}	19.41 _{1.11}	9.69 _{0.07}	74.13 _{0.05}
	Lin-Act	.*post_.*_layernorm	32	1.0	2.77 _{0.39}	20.57 _{1.36}	11.64 _{0.24}	72.21 _{0.08}
	LinEAS	.*post_.*_layernorm	32	1.0	1.88 _{0.19}	15.39 _{0.60}	10.83 _{0.25}	73.56 _{0.07}
Gemma2-2B	None	.*o_proj	-	-	4.00 _{0.45}	13.39 _{1.42}	14.79 _{0.00}	53.03 _{0.00}
	ITI-C	.*o_proj	32	0.5	1.17 _{0.60}	7.15 _{0.92}	14.00 _{0.11}	52.78 _{0.23}
	Lin-Act	.*o_proj	32	1.0	1.60 _{0.32}	7.76 _{0.39}	14.78 _{0.12}	52.43 _{0.57}
	LinEAS	.*o_proj	32	1.0	2.10 _{0.34}	8.78 _{0.56}	14.72 _{0.16}	53.27 _{0.41}
DeepSeek-7B	None	.*o_proj	-	-	4.30 _{0.70}	18.62 _{0.51}	8.49 _{0.00}	48.31 _{0.00}
	ITI-C	.*o_proj	32	0.5	2.83 _{0.40}	15.18 _{2.00}	7.71 _{0.07}	48.47 _{0.25}
	Lin-Act	.*o_proj	32	1.0	2.23 _{0.69}	11.08 _{0.76}	8.67 _{0.03}	47.71 _{0.27}
	LinEAS	.*o_proj	32	1.0	2.30 _{0.48}	8.46 _{0.54}	8.61 _{0.06}	46.35 _{0.37}
Qwen2.5-1.5B	None	.*o_proj	-	-	3.00 _{0.54}	23.09 _{0.67}	13.67 _{0.00}	60.95 _{0.00}
	ITI-C	.*o_proj	32	0.5	1.87 _{0.21}	18.16 _{0.62}	12.39 _{0.09}	60.88 _{0.08}
	Lin-Act	.*o_proj	32	1.0	1.50 _{0.35}	13.88 _{1.72}	13.88 _{0.16}	60.09 _{0.25}
	LinEAS	.*o_proj	32	1.0	1.50 _{0.29}	12.03 _{0.71}	14.04 _{0.10}	59.53 _{0.17}
Qwen2.5-7B	None	.*o_proj	-	-	3.92 _{0.59}	25.16 _{0.92}	10.67 _{0.00}	74.26 _{0.00}
	ITI-C	.*o_proj	32	0.5	2.97 _{0.21}	20.68 _{2.21}	9.56 _{0.02}	74.20 _{0.07}
	Lin-Act	.*o_proj	32	1.0	2.25 _{0.13}	16.04 _{0.95}	10.77 _{0.09}	73.56 _{0.08}
	LinEAS	.*o_proj	32	1.0	2.00 _{0.18}	13.58 _{0.76}	12.52 _{0.07}	71.34 _{0.10}

Table 9: **LinEAS is more robust to the layer choice.** Toxicity mitigation on the RTP and TET datasets, intervening on .*post_.*_layernorm and .*o_proj layers. We report results at low (32 sentences) regime, showing in **bold** the best toxicity result per model. When intervening on .*post_.*_layernorm, both ITI-C and Lin-Act show poorer performance, specially for DeepSeek-7B and Qwen2.5-1.5B, where the toxicity goes even above the original one (in red). For .*o_proj, ITI-C and Lin-Act perform better. Overall, LinEAS shows strong robustness to the layer choice.

F Order of Magnitude of Interventions Parameters

To inform the scale of regularization terms, we plot descriptive statistics of the values of $\mathbf{w}$ and $\mathbf{b}$, layer by layer.

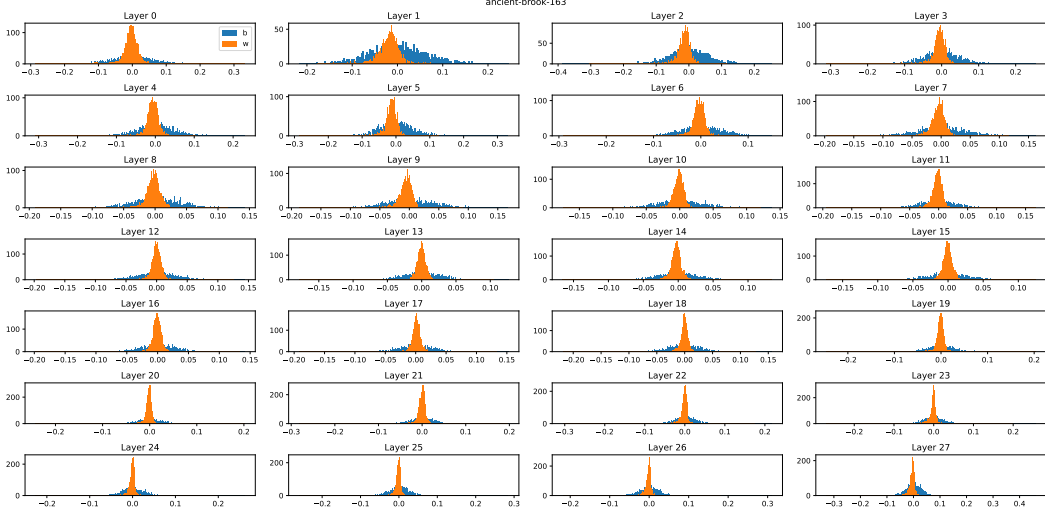


Figure 7: Distribution, layer by layer, of recentered scale parameters $\mathbf{w}$ and $\mathbf{b}$ biases, for a converged run of LinEAS, intervening on the 28 intervened layers of Gemma2-2B.

Sparsity and Refitting. When $\gamma \gg 0$, several coordinates of ω_ℓ and b_ℓ (parameters of T_ℓ) will collapse to 1 or 0, respectively. While this is desired, non-zero parameters typically suffer from shrinking, where the regularization terms $\mathcal{R}_1, \mathcal{R}_G$ dampen the effect of $\mathcal{C}$. A typical solution to this phenomenon is to perform m_{post} training steps updating only those non-collapsed parameters, a practice known as re-fitting in regression [38, 39]. We have not observed improvements when refitting parameters, and therefore do not use it. We observe that the entries of $\omega_\ell - 1$ and b_ℓ have similar scales (see Figure 7) and choose to use the same regularization strength.

G Effect of Sparsity on Toxicity Mitigation (extended results)

We complement the results shown in Figure 4, this time measuring the effect of sparsity on toxicity mitigation when the transport maps are optimized with 1024 (source and target) sentences. The results and trends discussed in Figure 4 also hold in this setup with more data.

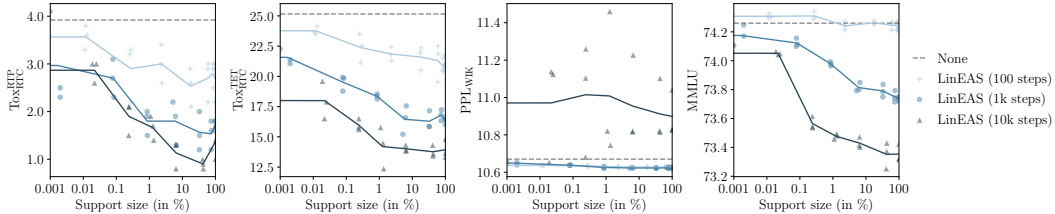


Figure 8: **Sparsity improves utility while mitigating toxicity, also in high data regime.** Toxicity results on Qwen2.5-7B using 1024 sentences, at different levels of sparsity γ that result in different support sizes (x axis). With a support of 1%-5% we maintain similar toxicity (left, center-left) while PPL_{WIK} decreases (center-right) and MMLU increases (right). Note that too long optimizations (30k steps) might harm utility, due to overfitting. Similarly, short optimizations (e.g., 300 steps) and strong sparsity leads to low conditioning (mild toxicity mitigation).

H Effect of Sparsity on T2I Generation

We ran a sweep over sparsity coefficients on DMD2 [30] and report it on Table 10. We find that T2I models are *more* sensitive than LMs to the sparsity penalty, almost saturating to either full support or no support outside the range [0.4, 0.8]. We hypothesize the UNet is more sensitive than the transformer because its activation maps are less redundant due to the changes in dimensionality as described by Veit et al. [40] and Jastrzebski et al. [41].

Table 10: Effect of the sparsity coefficient on DMD2. Performance metrics (support, IMGScore, and CLIPScore) at varying levels of sparsity. IMGScore generally increases with higher sparsity, while CLIPScore shows a slight increase.

sparsity	support(%)	IMGScore(%) $\uparrow$	CLIPScore(%) $\downarrow$
0.0	100 $\pm$ 0.0	71.4 $\pm$ 5.5	13.1 $\pm$ 3.2
0.4	92.7 $\pm$ 5.0	84.7 $\pm$ 6.0	14.5 $\pm$ 2.9
0.5	76.7 $\pm$ 14.0	89.5 $\pm$ 6.3	15.0 $\pm$ 2.9
0.6	51.6 $\pm$ 22.1	94.8 $\pm$ 3.6	15.5 $\pm$ 2.8
0.7	13.8 $\pm$ 11.4	98.9 $\pm$ 1.1	16.0 $\pm$ 2.8
0.8	3.3 $\pm$ 3.2	99.6 $\pm$ 0.4	16.1 $\pm$ 2.8
1.0	0.0 $\pm$ 0.0	100 $\pm$ 0.0	16.1 $\pm$ 2.9

I RLHF Implementation Details

We perform parameter-efficient adaptation of our baseline models with Huggingface’s implementation of LoRA in their PEFT library and Huggingface’s implementation of the PPO reinforcement learning algorithm in their TRL library. For each sample size in Table 1, we performed a hyperparameter search and chose the hyperparameters that yielded best validation toxicity scores at a perplexity close to LinEAS. Following Ouyang et al. [2], we fine-tune the models using proximal policy optimization (PPO) [42] on the same $N = 32$ data and use RTC (Roberta toxicity classifier) as our reward model. We instantiate the reward model with the Roberta toxicity classifier from Logacheva et al. [26] used for evaluation (RTC in Table 1); we use the base model without LoRA weights as the reference model and the LoRA model as the policy; we add an off-the-shelf value head from TRL to the policy to estimate the value function. We follow the original LoRA implementation and only fine-tune $\{k, q, v, o\}_{proj}$ layers while keeping the MLPs frozen. The summary of hyperparameters can be found in Table 11.

Hyperparameter	Values	Best 32 Samples	Best 1024 Samples
global_epochs	{10, 15, 20}	10	15
ppo_epochs	{1, 10, 20, 50}	20	20
learning_rate	$\{10^{-4}, 5 * 10^{-5}, 10^{-5}\}$	10^{-5}	10^{-5}
batch_size	{32, 64, 128}	32	128
mini_batch_size	{16, 32, 64}	32	64
lora_rank	{2, 4, 8}	2	2

Table 11: List of hyper-parameters used to train PPO for 32 and 1024 samples. Unless otherwise specified, we use TRL’s defaults. We could not use a mini-batch size greater than 64 due to memory constraints. We did not use gradient accumulation. We found the `learning_rate` and `global_epochs` to be the most important hyper-parameters. Low learning rate for few epochs leads to underfitting while high learning rate for many epochs tends to overfit.

J Composition of Interventions

We evaluate in this section the ability to compose two LinEAS maps pre-trained independently on two concepts. Our goal is to assess whether they can be composed, at the level of each activation, to induce *both* concepts. Achieving concept composition in activation steering is an open goal, and

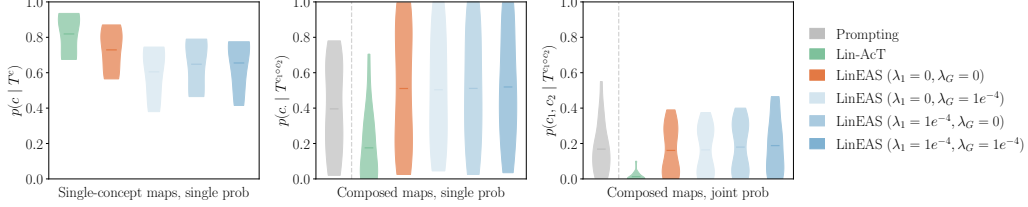


Figure 9: LinEAS obtains composable maps. We benchmark Lin-Act and LinEAS for various regularization strengths γ on a compositional task. We train each method to model distribution shifts towards a certain concept c taking in a list of 5 concepts. **(Top-left)** We measure, using LLM as a judge, whether sentences formulated *using that trained shift* do contain that concept. We obtain for each concept a probability, that we then aggregate using violin plots. **(Top-middle)** We pick two concepts c_1 and c_2 randomly, train separately their maps, and compose them. We then measure the probability of finding c_1 , and then c_2 , in the generated content, using the composition of both maps (in both orders). **(Top-right)** We now measure the probability of finding simultaneously *both* concepts in the generations. As can be seen, the composition of LinEAS maps (learnt separately for 2 concepts) yields an average $15\times$ (right) higher probability of including both 2 concepts in the same generation, with respect to Lin-Act, and a $3.1\times$ increase in generating at least one of the concepts (center). Interestingly, Lin-Act obtains slightly better probability when using only single-concept maps (left), but these fail at composition. Additionally, note that sparsity in LinEAS is beneficial for compositionality, increasing the joint concept presence from 0.16 to 0.19 when using *group lasso*.

our hypothesis is that sparse and end-to-end LinEAS maps affect minimally the model, facilitating composition.

Setup. For each of the concepts *day*, *night*, *elephant*, *football* and *fishing*, we generate $N = 50$ diverse sentences using Gemma2-27B that contain that concept, to form five target (q_i). Additionally, we ask Gemma2-27B to generate 50 diverse sentences about generic situations, forming a single source distribution p . We then learn five steering maps, from p to each of the q_i distributions, using both Lin-Act and LinEAS. For LinEAS we test all combinations of $\lambda_1 = 0, 1e^{-4}$ and $\lambda_G = 0, 1e^{-4}$, to assess the impact of sparsity. Equipped with these five concept maps, we compose them for each pair of concepts c_1, c_2 as follows: $T_{\ell}^{c_1 c_2}(z) := T_{\ell}^{c_2}(T_{\ell}^{c_1}(z))$. Note that $T_{\ell}^{c_1 c_2} \neq T_{\ell}^{c_2 c_1}$. Following Rodriguez et al. [10], we intervene on Gemma2-2B by generating 200 sentences that follow the prompt *Once upon a time* and we measure the presence of the concepts in the generations in a *LLM-as-a-judge manner*, querying Qwen2.5-7B-Instruct. More precisely, we query about: the presence of a concept c in each generated sentence, when using the map trained with concept c ; the presence of a concept c , when using the map trained with that concept *and* any other; the presence of both *two* concepts c_1, c_2 in the same sentence, yielding $p(c_1, c_2)$, using either $T_{\ell}^{c_1 c_2}$ or $T_{\ell}^{c_2 c_1}$.

Results, comparing interventions. Figure 9.(top) plots the three probabilities described above. In particular, the right plot shows $p(c_1, c_2 | T^{c_1 c_2})$, *i.e.*, the probability of observing both concepts in the same generated sentence, using the composed map. Lin-Act is able to generate concepts using single-concept maps (left plot) with average probability of 0.82 *vs.* 0.73 using LinEAS without sparsity. We also observe that increasing the sparsity (larger λ s) slightly diminishes the presence of concepts when using single-concept maps. However, we observe a drastically different picture when using *combined* maps: (middle) that probability goes from a Lin-Act average of 0.17 to around 0.52 with LinEAS ($3.1\times$ increase). Most importantly, the joint presence of concepts probability (right) goes from 0.013 for Lin-Act to 0.19 ($15\times$ increase) for LinEAS with *group lasso* regularization (both λ_1, λ_G used). See Appendix J for generation examples. These results show that LinEAS learns maps that are easier to compose than those from Lin-Act. Indeed, composition of Lin-Act maps is very brittle. While LinEAS achieves much stronger compositionality, our results show that there is still room for improvement on this important problem. We provide qualitative examples in Figure 1.

Results, prompting as baseline. We prompt the LLM to complete a generation with two concepts using *"Continue the following text, make sure concepts [c1] and [c2] appear in the continuation: Once upon a time"*. We observe that the mean probability of:

- generating either of the concepts (middle plot) is 39% with prompting / 52% for LinEAS.
- generating both concepts (right plot) is 17% for prompting / 19% for LinEAS.

In light of these results, we conclude that prompting is a strong baseline, since it has direct access to the concepts in textual form at generation time. However, LinEAS achieves better compositions without such direct access.

J.1 Composition of Interventions: Qualitative Results

We show qualitative results using composed maps as detailed in Appendix J. In Table 12 and Table 13 we include generations that were marked as containing both concepts using LLM-as-a-judge (Qwen2.5-7B-Instruct). We compare Lin-ACT and LinEAS (with $\gamma = 1e^{-4}$).

Table 12: Generations inducing both concepts *Fishing*, *Elephant*. Only the provided 2 generations for Lin-ACT were marked as containing both concepts using LLM-as-a-judge, out of 200. Conversely, 25 sentences generated by LinEAS contain both concepts. Note the stark difference in quality, LinEAS achieves high quality generations with composition.

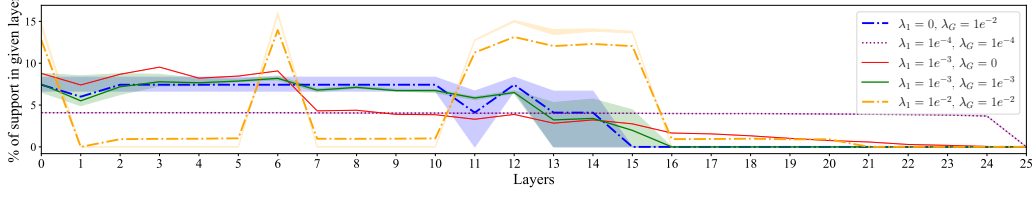
Composition	Method	Generation
Fishing ◦ Elephant	Lin-ACT	Once upon a time, the Tuffa as the bull elephant’s foot as she made a small putty as the water as she nosed the bank of the river as the water, as the sun as the swaying on the water as she stood in the shallows, the as the water [...]
Fishing ◦ Elephant	Lin-ACT	Once upon a time, a large, gray, as it was the bull as it snorted in the water as it struggled to break the line of the bank as it thrashed in the water, a small stream of water, as it made a final sprint, as the water, as it snorted noisily, the water as the river [...]
Fishing ◦ Elephant	LinEAS	Once upon a time, a man was fishing in the river when he saw a beautiful, pink baby elephant walking towards him. The elephant was thirsty and the man offered him a drink of water. The elephant was grateful and splashed some water on the man’s face.
Fishing ◦ Elephant	LinEAS	Once upon a time, a huge elephant’s tusk broke the water, his trunk splashing in the shallow river. "Hup!" he called, his tail swished against the muddy bank. I watched from the shore, my fishing rod dangling in the water.
Fishing ◦ Elephant	LinEAS	Once upon a time, the elephant’s trunk broke the water, his massive body rising and disappearing. The jungle rumbled in the distance, their long tusks scraping against the mud. It was a young calf, its small, wet back. The fisherman sat on the bank, his net swinging lazily in the shallow water. A small fish

Table 13: Generations inducing both concepts *Night*, *Football*.

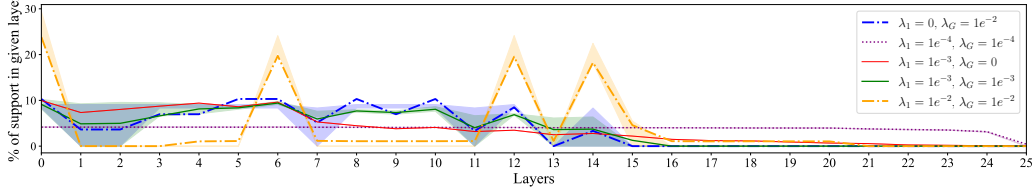
Composition	Method	Generation
Night $\circ$ Football	Lin-ACT	Once upon a time, a little girl’s dream came true. It was the first night of the 2013-14 season and the young forward had just scored her first goal for the first team. The ball had nestled in the net in the 15th minute of the game, and the 16-year-old couldn’t believe her luck.
Night $\circ$ Football	Lin-ACT	Once upon a time, there was a man who was so determined to win the game, he went out to play in the rain. It was a cold, wet night, and the rain was pouring down, but the man didn’t let it dampen his spirits. He was out to win the game, and he knew that he had to make a late, late goal to seal the victory.
Night $\circ$ Football	Lin-ACT	Once upon a time, I was a happy camper. I was in the middle of a long drive through the woods, the sound of the wind whistling through the trees. The air was cool and the sun was setting, casting a warm glow over the forest. As I drove, I couldn’t help but feel a sense of excitement. I had just won the race to the finish line, the ball bouncing off the net, and the ball was heading for the goal.
Night $\circ$ Football	LinEAS	Once upon a time, in the night, the lights of the field illuminated the players. The crowd roared in excitement, their voices echoing off the stands. It was the final whistle, and the opposing team celebrated, their coach shouting with joy.
Night $\circ$ Football	LinEAS	Once upon a time, the “ <i>Let’s go Rangers!</i> ” could be heard through the dark, cold night. The home crowd cheered and the ball flew past the goal, sending a shower of confetti into the air.
Night $\circ$ Football	LinEAS	Once upon a time, the sun set on the field of grass. The opposing teams were locked in a fierce battle, the crowd roaring with excitement. It was the opening goal, and the referee blew his whistle, signaling the end of the game.

K Group Sparsity Trade-offs

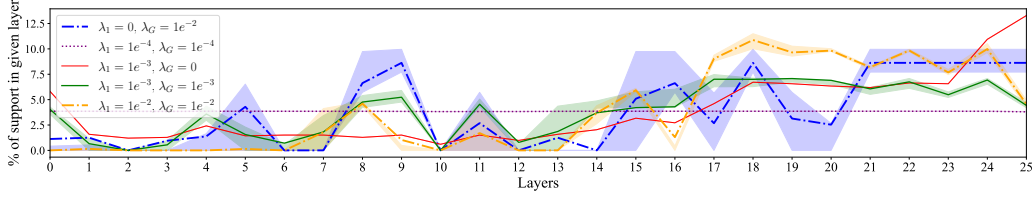
We consider the interventions learned with Gemma2-2B for the concepts *day*, *night*, *elephant*, *football*, and *fishing*, but set both $\lambda_1 > 0$ and $\lambda_G > 0$, enforcing a sparse group regularizer. Figure 10 shows how the proportion of intervened units, out of their entire support, is distributed across layers. Since we are plotting proportions of the support, differences between models in terms of the absolute number of impacted neurons are not reflected. Rather, the plot allows to grasp how regularization, and notably group regularization, impacts layer selection. An important finding of this experiment is that we do observe, for high group regularization regimes, that very specific layers are consistently selected to induce diverse concepts. Additionally, we posit that this grouping of neurons at certain layers impacts positively efficient transfer of concepts, notably when investigating compositionality.



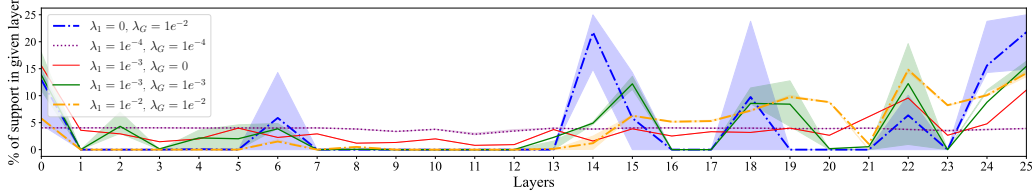
(a) LinEAS Distribution of support across layers (biases). Post Feed-Forward Layer Norms.



(b) LinEAS Distribution of support across layers (biases). Post Attention Layer Norms.



(c) LinEAS Distribution of support across layers (weights). Post Feed-Forward Layer Norms.



(d) LinEAS Distribution of support across layers (weights). Post Attention Layer Norms.

Figure 10: **LinEAS Distribution of support across layers.** Each subfigure shows the percentage of intervened units (out of the total support) across layers for different regularization strengths, averaged over 5 concepts (with 50% quantile range).

L Similarity of LinEAS Interventions with Human Judgment

We draw a set of 50 concepts, from the MEN dataset [29], a resource of 3,000 word pairs annotated with human similarity judgments (see also Fedzechkina et al. [43] for a study on LLM interpretability building on the same resource). With these 50 concepts, we can recover 20 word pairs annotated for their similarity in the MEN dataset. We train LinEAS interventions for each concept on Gemma2-2B. We then ask: do we recover similar interventions for similar concepts, and does sparsity help? We can answer positively to both: highly similar concepts have highly similar interventions; and enforcing sparsity through our scheme improves that correlation.

To measure this, we compute the average sparse support of interventions (shown as y-axis in Figure 11). We compute the similarity between the intervention vectors for each word pair, focusing on biases first: $s_{int}^b = \{\text{sim}(\mathbf{b}_{c1}, \mathbf{b}_{c2})\}_{\forall c1 \neq c2}$ (on the left in Figure 11). Finally, we consider the human similarity judgments s_{hum} reported in the MEN dataset for each word pair, and compute their correlation with the intervention similarity, $\text{corr}(s_{int}^b, s_{hum})$. We do the same for the weights, in this case through s_{int}^w (on the right in Figure 11 and). The numbers in the scatterplots in Figure 11 correspond to the LinEAS regularization parameters λ_1, λ_G .

Note that sparsity helps improve correlation beyond the non-sparse version of LinEAS (noted as 0;0, with a support of 100% in Figure 11). Overall, LinEAS shows a strong correlation with human judgment.

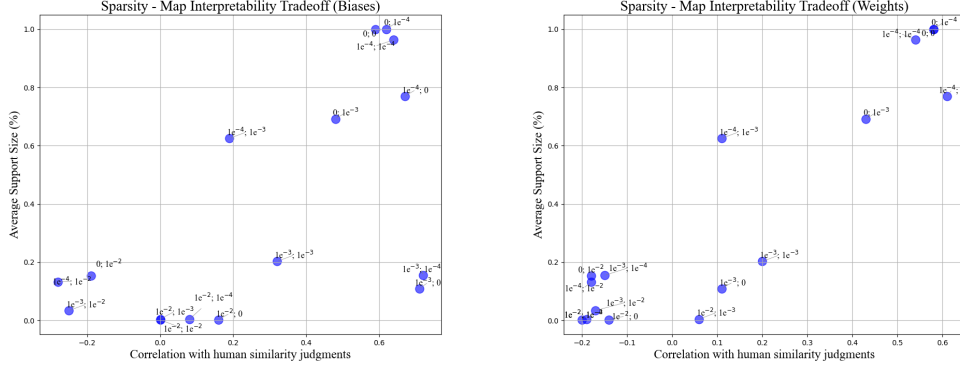


Figure 11: Scatter plots showing the correlation of interventions with human similarity judgment. We show the correlation with bias similarity in the left plot and with weight similarity on the right plot.

In Figure 12 we show the same results in form of matrix, showing that sparsity is indeed helpful to improve correlation with human alignment.

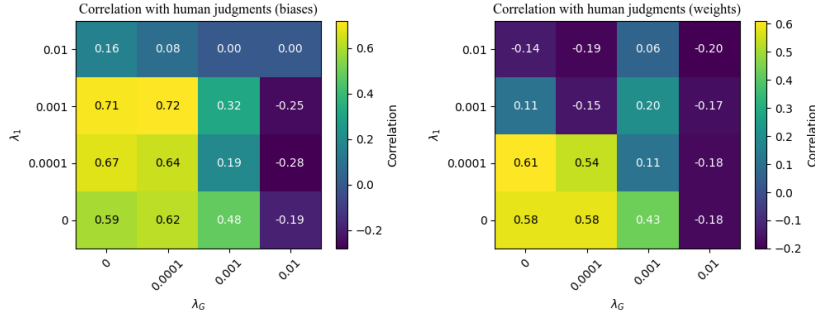


Figure 12: Heatmaps aggregating scatterplot results above but with a different view. We observe that highest correlations are obtained with a suitable regularization strength, for biases $\lambda_1 = 10^{-3}$, $\lambda_G = 10^{-4}$, for weights $\lambda_1 = 10^{-3}$, $\lambda_G = 0$.

M Additional Experiment: Inducing Truthfulness in LLMs

In this section, we complement and corroborate our insights from the experiments on toxicity mitigation in Section 4.1 with additional experiments on inducing truthfulness in LLMs, using the TruthfulQA benchmark [44]. In particular, we investigate how well LinEAS achieves to induce truthfulness on this benchmark in comparison to Lin-ACt, its strongest activation steering competitor from Section 4.1.

For LinEAS, we apply the intervention again to the post layernorm layers, while for Lin-ACt, we apply them to all layernorm layers as this was reported as optimal for Lin-ACt for TruthfulQA experiments in Rodriguez et al. [10]. We use 2-fold cross-validation on the 817 questions of the multiple choice part of the benchmark and learn the intervention on the concatenation of training fold questions concatenated with either incorrect (source) or correct (target) multiple-choice answer options. We report both MC1 and MC2 of TruthfulQA, and monitor overfitting on the TruthfulQA task by also evaluating MMLU 5-shot accuracy [28].

The results can be found in Table 14. We see that both methods can successfully induce truthfulness when presented with enough samples to learn the interventions, increasing the accuracy by up to

almost 5% (7%) on MC1 (MC2) in the high sample regime when 1024 samples are available. Overall the highest increases can be achieved with Lin-ACT, but only for the high sample regime. In the low sample regime, where only 32 samples are available, Lin-ACT tends to fail catastrophically: either it gets lower accuracies on MC1 and MC2 than even the unintervened model (Qwen2.5-7B, Qwen2.5-1.5B), or it fails completely on MMLU (Gemma2-2B). LinEAS on the other hand does well also in this low sample regime, and achieves second best overall performance on the Qwen models with only 32 samples available to learn interventions.

Model	Samples	Method	MC1 Acc. (%) ($\uparrow$)	MC2 Acc. (%) ($\uparrow$)	MMLU ($\uparrow$)
Qwen2.5-7B	-	None	37.82 _{0.00}	52.14 _{0.00}	74.26 _{0.00}
	32	Lin-ACT	32.17 _{0.66}	47.69 _{0.87}	59.22 _{1.74}
		LinEAS	40.10 _{0.37}	55.74 _{0.31}	73.88 _{0.07}
	1024	Lin-ACT	42.59 _{0.50}	58.92 _{0.82}	73.79 _{0.14}
		LinEAS	39.56 _{0.13}	55.20 _{0.15}	<u>73.88</u> _{0.02}
	Qwen2.5-1.5B	-	None	30.23 _{0.00}	43.70 _{0.00}
32		Lin-ACT	27.22 _{1.38}	42.64 _{3.17}	32.10 _{1.95}
		LinEAS	<u>32.26</u> _{0.65}	<u>46.07</u> _{0.63}	<u>60.34</u> _{0.12}
1024		Lin-ACT	32.90 _{0.36}	47.17 _{0.61}	60.17 _{0.18}
		LinEAS	31.77 _{0.11}	45.31 _{0.20}	60.41 _{0.01}
Gemma2-2B		-	None	21.18 _{0.00}	33.05 _{0.00}
	32	Lin-ACT	<u>24.94</u> _{1.15}	40.95 _{1.29}	27.59 _{1.09}
		LinEAS	24.21 _{0.85}	38.09 _{1.14}	<u>52.08</u> _{0.16}
	1024	Lin-ACT	25.65 _{0.53}	<u>39.73</u> _{0.49}	51.40 _{0.17}
		LinEAS	23.82 _{0.19}	37.80 _{0.53}	52.37 _{0.05}

Table 14: Results on TruthfulQA. We report results at low data (32 samples to estimate the interventions) and high (1024 samples) regimes. Results are averaged over five random seeds.

N Additional details and results on text to image generation

N.1 Additional qualitative results with LLM-generated prompts

We extend here the qualitative results mitigating and inducing styles on the 15 concepts described in Appendix O. Figures 13 to 15 compare Lin-ACT, and LinEAS. Both the diversity and (human) perceptual quality of the generations is higher with LinEAS. We observe stark differences between both methods’ generations. Figures 16 to 18 show additional generations with LinEAS with more granular strengths.

Note that all the generations start from the same prompt, with the concept of interest appended in the form of textual tags. It is interesting to see that LinEAS recovers an image conforms with the prompt *without the concept* at $\lambda = 1$. Also, observe how LinEAS’s generations are much more gradual than Lin-ACT.

We also comment on the surprising results obtained when inverting the LinEAS linear maps. We observe how the concept increases, and LinEAS shows much higher quality and coherence under this regime. This points to the fact that LinEAS is better exploiting the underlying structure in activation space.

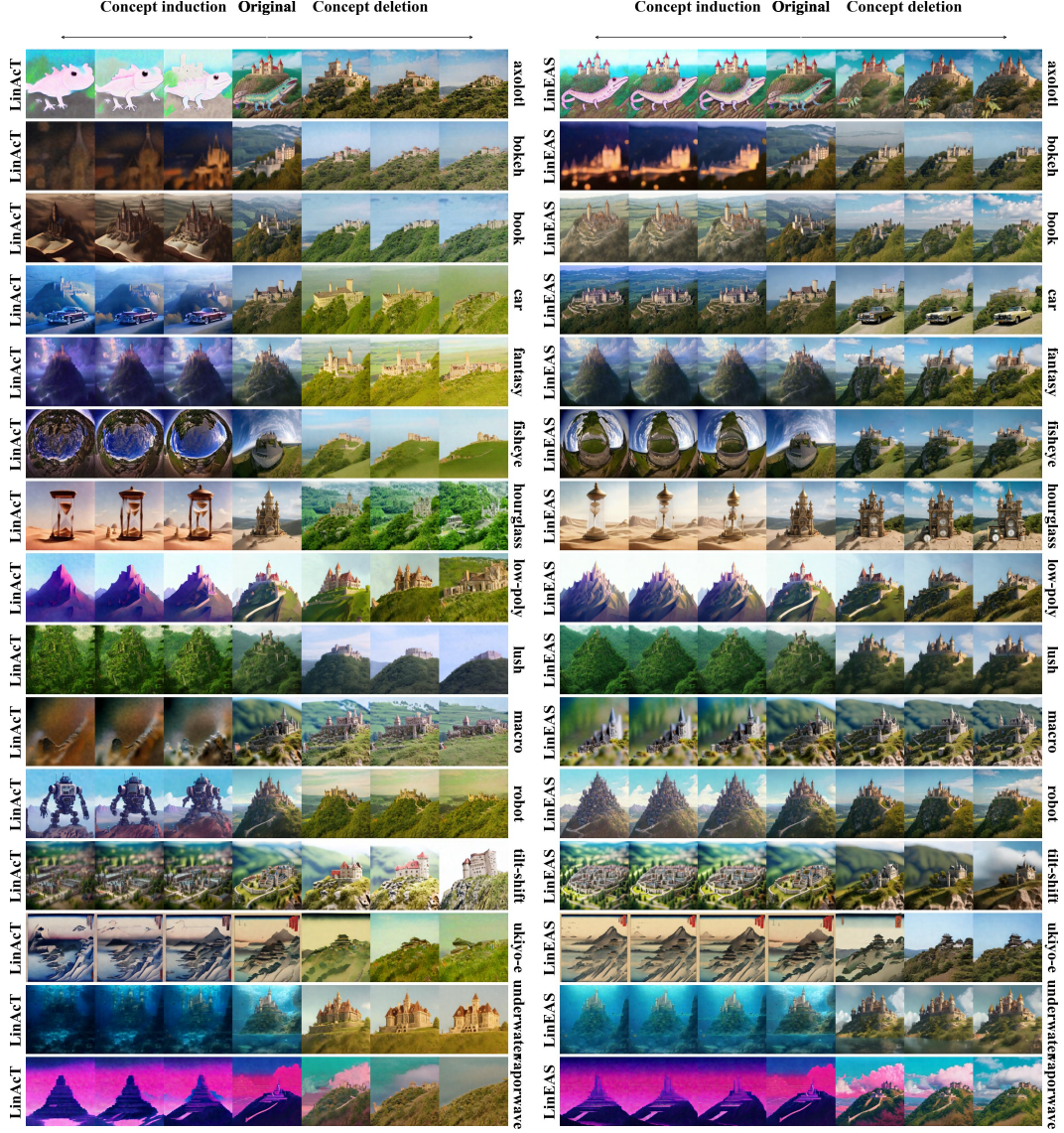


Figure 13: Images generated with Lin-Act (left panel) and LinEAS (right panel) with the prompt *A grand castle sits atop a hill overlooking a valley*. [concept tag] Each row contains a different conditioning concept to be mitigated by the steering method and each column a different intervention strength (λ). Each column contains a generation for $\lambda = 1.0, 0.8, 0.6, 0.4, 0.2, 0.0, 0.2, 0.4, 0.6, 0.8, 1.0$ respectively and the columns to the left of “original” contain generations using $T_{\ell}^{-1}(z)$.

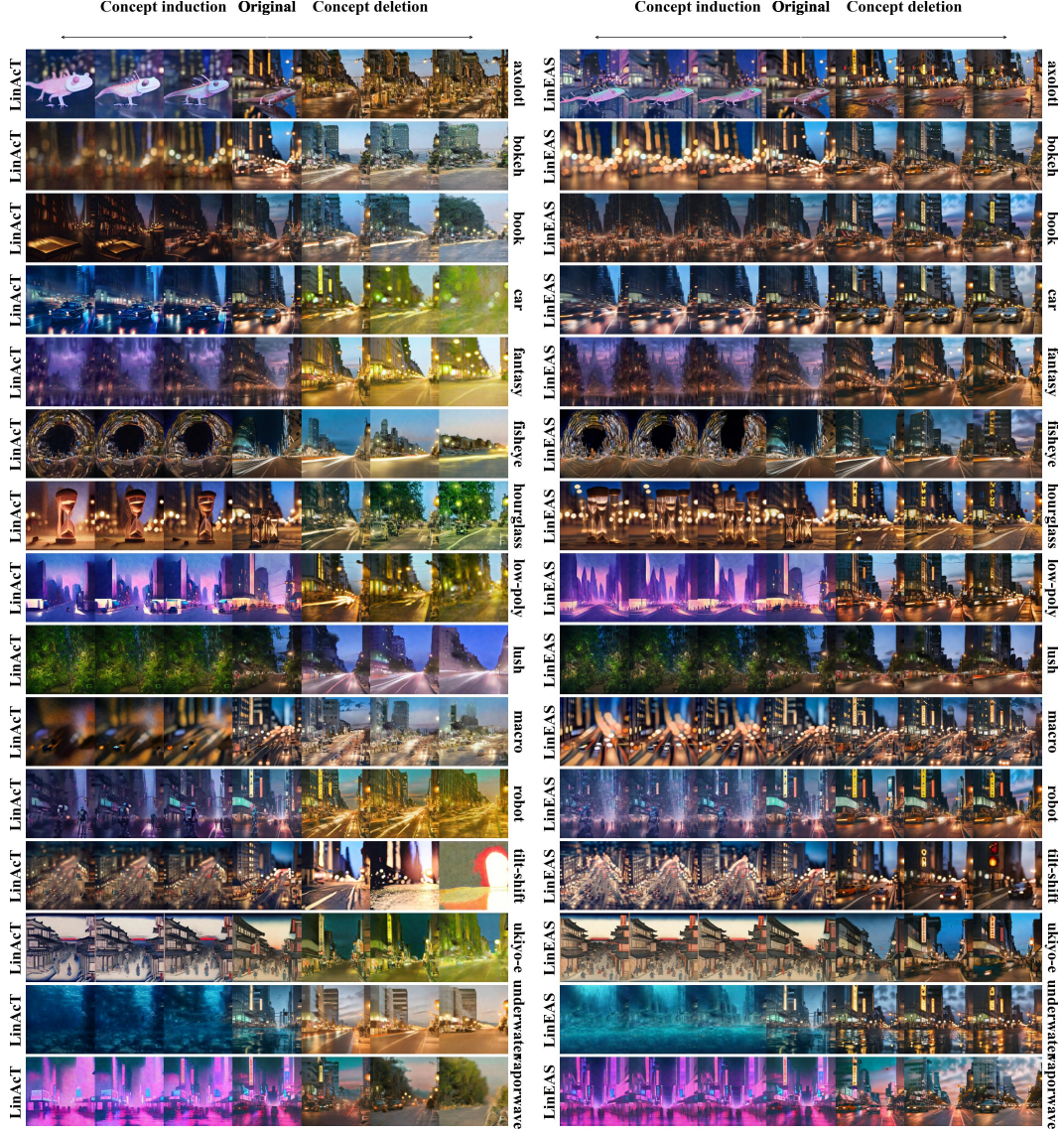


Figure 14: Images generated with Lin-AcT (left panel) and LinEAS (right panel) with the prompt *A bustling city street at twilight, lights blurring.* [concept tag] Each row contains a different conditioning concept to be mitigated by the steering method and each column a different intervention strength (λ). Each column contains a generation for $\lambda = [1.0, \dots, 0, \dots, 1.0]$ respectively and the columns to the left of “original” contain generations using $T_{\ell}^{-1}(z)$.

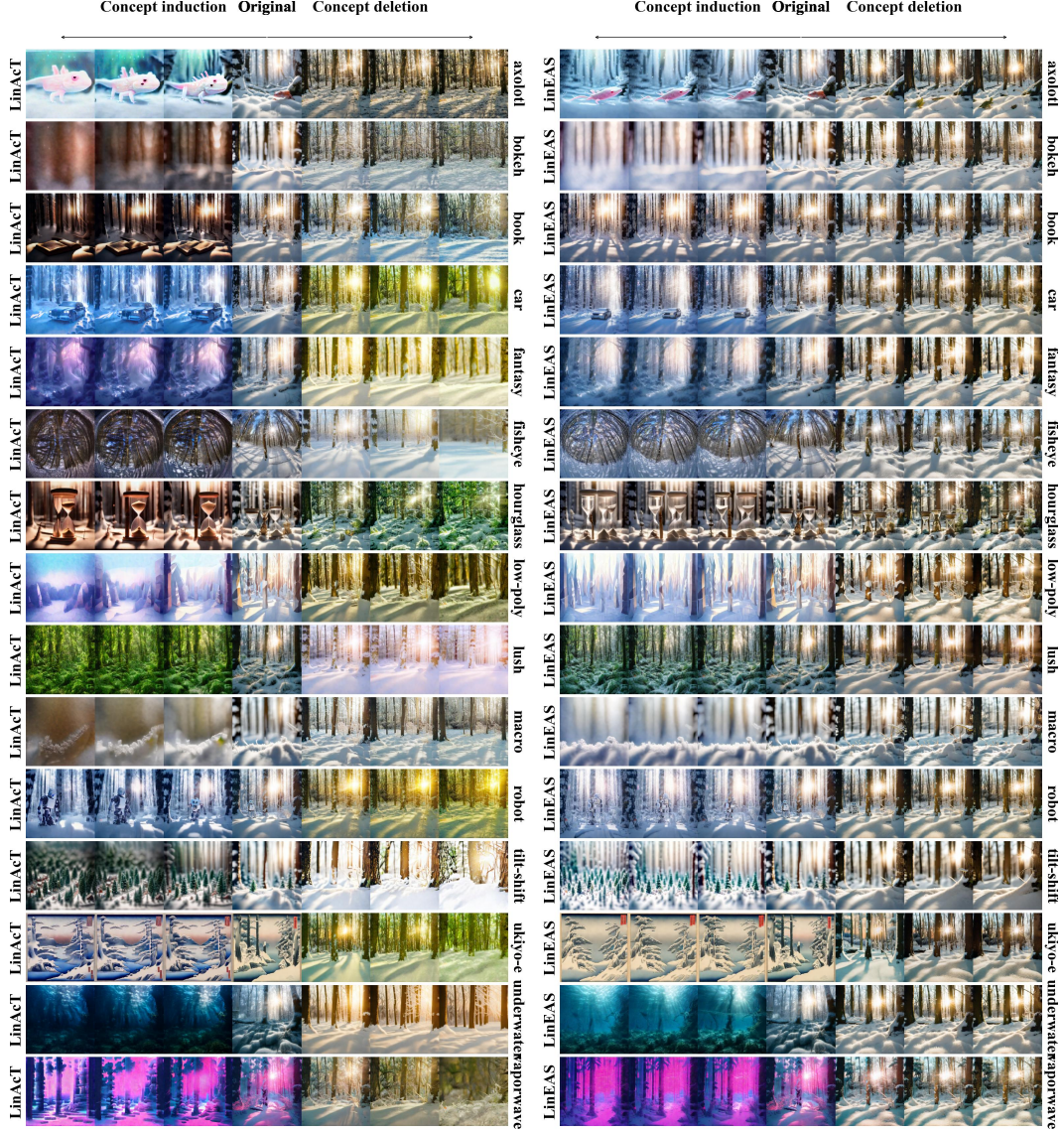


Figure 15: Images generated with Lin-Act (left panel) and LinEAS (right panel) with the prompt *A snow-covered forest with sunlight filtering through the trees. [concept tag]* Each row contains a different conditioning concept to be mitigated by the steering method and each column a different intervention strength (λ). Each column contains a generation for $\lambda = [1.0, \dots, 0, \dots, 1.0]$ respectively and the columns to the left of “original” contain generations using $T_{\ell}^{-1}(z)$.

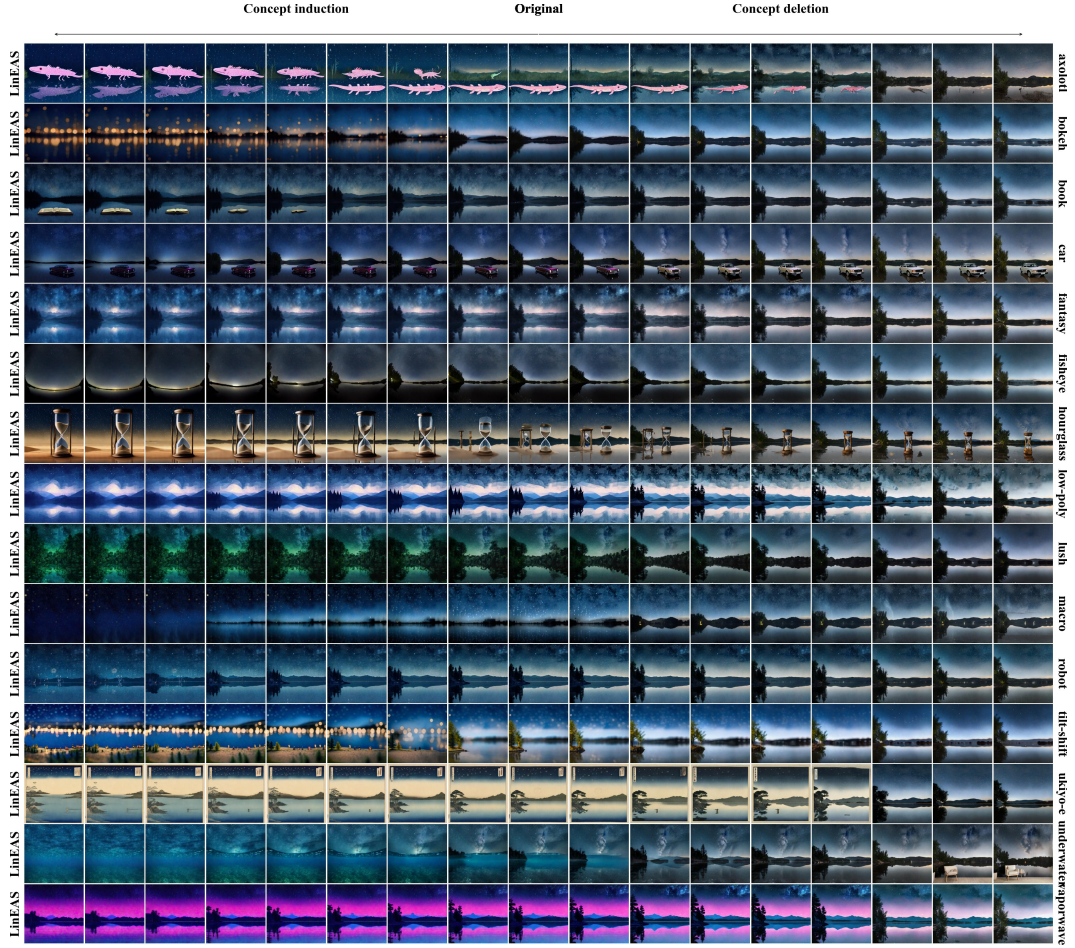


Figure 16: Images generated with LinEAS with the prompt *A starry night sky over a calm lake. [concept tag]* Each column contains images generated with a different steering strength for mitigating the concept corresponding to the row: $\lambda = [1.0, \dots, 0, \dots, 1.0]$. Images to the left of “original” were produced using the inverse steering map $T_{\ell}^{-1}(z)$.

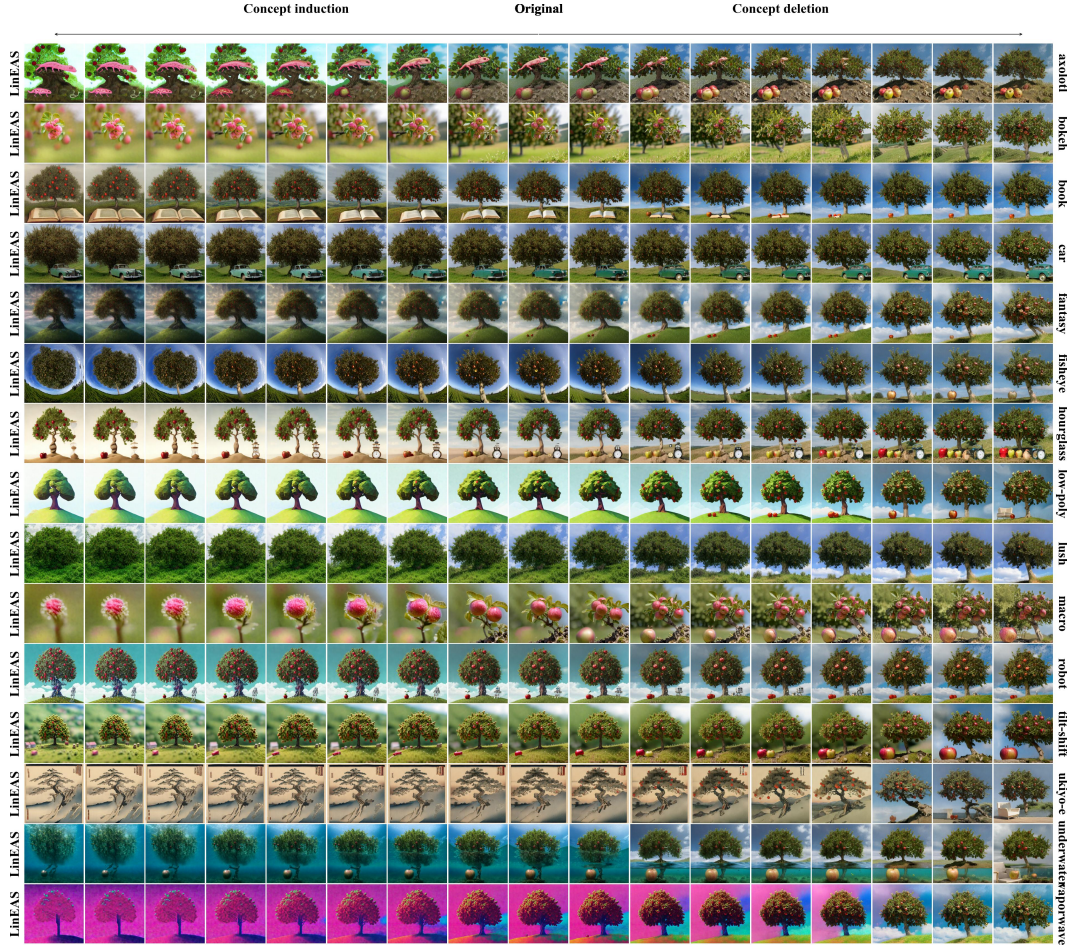


Figure 17: Images generated with LinEAS with the prompt *A fruiting apple tree on top of a hill.* [concept tag] Each column contains images generated with a different steering strength for mitigating the concept corresponding to the row: $\lambda = [1.0, \dots, 0, \dots, 1.0]$. Images to the left of “original” were produced using the inverse steering map $T_{\ell}^{-1}(z)$.

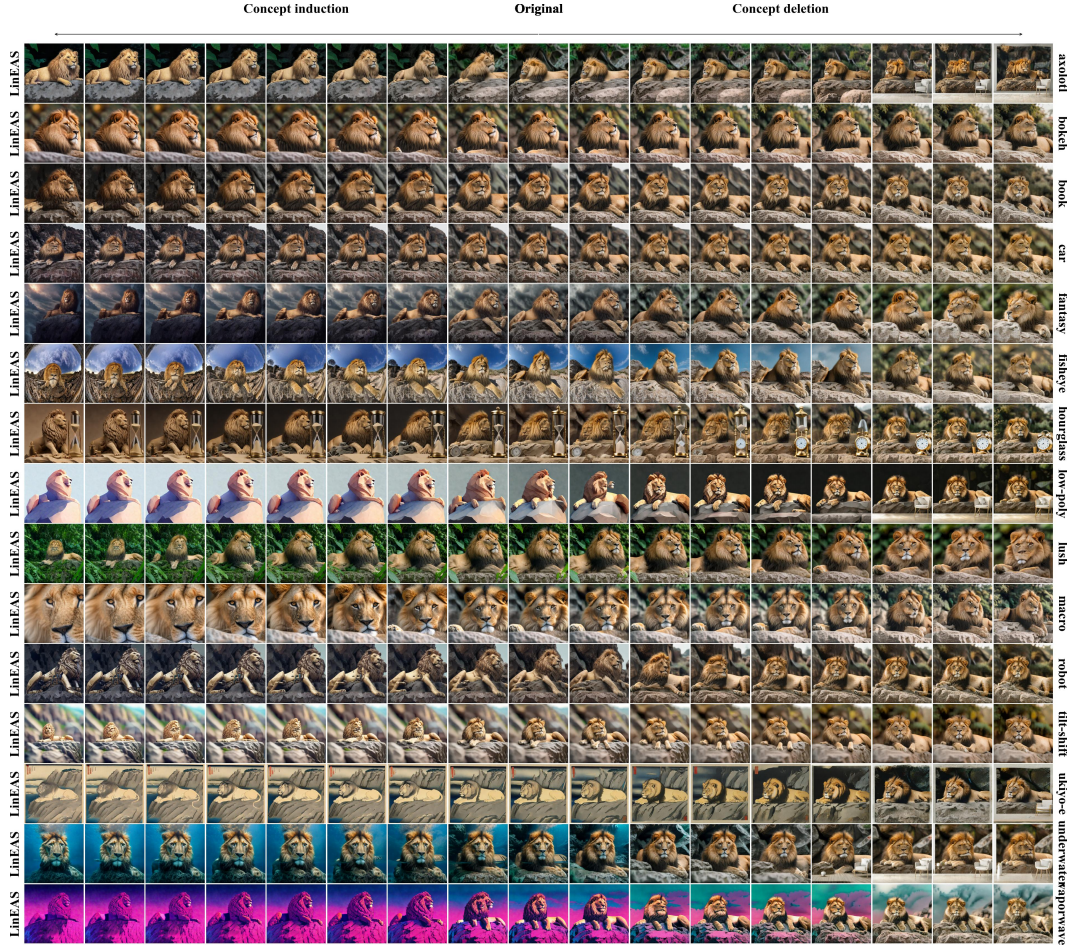


Figure 18: Images generated with LinEAS with the prompt *A majestic lion rests on a rocky outcrop.* [concept tag] Each column contains images generated with a different steering strength for mitigating the concept corresponding to the row: $\lambda = [1.0, \dots, 0, \dots, 1.0]$. Images to the left of “original” were produced using the inverse steering map $T_{\ell}^{-1}(z)$.

N.2 Detailed Quantitative Results

Here we report detailed per-concept ImgScores and ClipScores for ITI-C, Lin-AcT, and LinEAS both for the forward ($(T_\ell(z))$) and inverse ($(T_\ell^{-1}(z))$) application of the steering (Table 15). Remarkably, we find that inverting the steering operation tends to induce the concepts that these methods mitigate when they are not inverted.

Task	Intervention	$T_\ell(z)$		$T_\ell^{-1}(z)$	
		$\uparrow$ ImgScore	$\uparrow$ ClipScore	$\uparrow$ ImgScore	$\downarrow$ ClipScore.
Axolotl	ITI-C	0.13 _{0.11}	0.17 _{0.02}	0.26 _{0.17}	0.27 _{0.02}
	Lin-AcT	0.35 _{0.19}	0.15 _{0.02}	0.28 _{0.18}	0.29 _{0.03}
	LinEAS	0.56 _{0.23}	0.16 _{0.04}	0.60 _{0.19}	0.23 _{0.05}
Bokeh	ITI-C	0.38 _{0.19}	0.18 _{0.02}	0.33 _{0.18}	0.24 _{0.02}
	Lin-AcT	0.55 _{0.18}	0.18 _{0.02}	0.31 _{0.19}	0.24 _{0.02}
	LinEAS	0.75 _{0.14}	0.18 _{0.02}	0.62 _{0.22}	0.24 _{0.02}
Book	ITI-C	0.37 _{0.20}	0.19 _{0.01}	0.45 _{0.19}	0.21 _{0.02}
	Lin-AcT	0.62 _{0.16}	0.18 _{0.01}	0.49 _{0.19}	0.22 _{0.01}
	LinEAS	0.80 _{0.10}	0.17 _{0.01}	0.78 _{0.14}	0.19 _{0.02}
Car	ITI-C	0.18 _{0.17}	0.19 _{0.02}	0.43 _{0.23}	0.24 _{0.02}
	Lin-AcT	0.37 _{0.21}	0.18 _{0.02}	0.42 _{0.23}	0.23 _{0.02}
	LinEAS	0.78 _{0.13}	0.18 _{0.03}	0.81 _{0.12}	0.19 _{0.03}
Fantasy	ITI-C	0.12 _{0.14}	0.20 _{0.01}	0.35 _{0.19}	0.21 _{0.02}
	Lin-AcT	0.49 _{0.19}	0.20 _{0.02}	0.52 _{0.17}	0.23 _{0.01}
	LinEAS	0.72 _{0.16}	0.18 _{0.02}	0.73 _{0.15}	0.21 _{0.02}
Fisheye	ITI-C	0.40 _{0.16}	0.20 _{0.02}	0.21 _{0.13}	0.25 _{0.01}
	Lin-AcT	0.49 _{0.18}	0.20 _{0.02}	0.26 _{0.15}	0.25 _{0.02}
	LinEAS	0.68 _{0.14}	0.20 _{0.02}	0.51 _{0.16}	0.24 _{0.02}
Hourglass	ITI-C	0.16 _{0.18}	0.20 _{0.02}	0.57 _{0.26}	0.30 _{0.01}
	Lin-AcT	0.29 _{0.23}	0.17 _{0.02}	0.60 _{0.27}	0.29 _{0.01}
	LinEAS	0.54 _{0.23}	0.20 _{0.04}	0.74 _{0.19}	0.27 _{0.03}
Low-poly	ITI-C	0.12 _{0.12}	0.19 _{0.02}	0.50 _{0.17}	0.23 _{0.02}
	Lin-AcT	0.39 _{0.16}	0.19 _{0.02}	0.57 _{0.14}	0.24 _{0.02}
	LinEAS	0.57 _{0.19}	0.18 _{0.01}	0.67 _{0.14}	0.25 _{0.02}
Lush	ITI-C	0.29 _{0.21}	0.20 _{0.02}	0.44 _{0.17}	0.26 _{0.01}
	Lin-AcT	0.50 _{0.20}	0.21 _{0.02}	0.47 _{0.17}	0.25 _{0.01}
	LinEAS	0.71 _{0.15}	0.18 _{0.02}	0.75 _{0.12}	0.23 _{0.02}
Macro	ITI-C	0.31 _{0.16}	0.20 _{0.01}	0.18 _{0.15}	0.22 _{0.02}
	Lin-AcT	0.52 _{0.18}	0.18 _{0.02}	0.16 _{0.13}	0.22 _{0.02}
	LinEAS	0.68 _{0.17}	0.18 _{0.02}	0.54 _{0.21}	0.23 _{0.02}
Robot	ITI-C	0.36 _{0.20}	0.20 _{0.02}	0.32 _{0.17}	0.23 _{0.02}
	Lin-AcT	0.56 _{0.17}	0.18 _{0.02}	0.36 _{0.19}	0.25 _{0.02}
	LinEAS	0.74 _{0.15}	0.18 _{0.02}	0.71 _{0.15}	0.21 _{0.02}
Tilt-shift	ITI-C	0.11 _{0.11}	0.19 _{0.02}	0.44 _{0.20}	0.29 _{0.01}
	Lin-AcT	0.46 _{0.20}	0.21 _{0.02}	0.48 _{0.21}	0.29 _{0.01}
	LinEAS	0.65 _{0.16}	0.23 _{0.03}	0.65 _{0.15}	0.28 _{0.02}
Ukiyo-e	ITI-C	0.20 _{0.12}	0.21 _{0.02}	0.47 _{0.16}	0.28 _{0.02}
	Lin-AcT	0.29 _{0.17}	0.20 _{0.02}	0.64 _{0.15}	0.29 _{0.01}
	LinEAS	0.48 _{0.18}	0.18 _{0.03}	0.73 _{0.12}	0.28 _{0.01}
Underwater	ITI-C	0.33 _{0.19}	0.21 _{0.02}	0.43 _{0.19}	0.26 _{0.01}
	Lin-AcT	0.52 _{0.21}	0.18 _{0.02}	0.41 _{0.21}	0.26 _{0.02}
	LinEAS	0.59 _{0.21}	0.17 _{0.02}	0.56 _{0.20}	0.26 _{0.01}
Vaporwave	ITI-C	0.12 _{0.17}	0.18 _{0.01}	0.35 _{0.17}	0.26 _{0.01}
	Lin-AcT	0.33 _{0.18}	0.15 _{0.02}	0.56 _{0.14}	0.27 _{0.01}
	LinEAS	0.65 _{0.17}	0.16 _{0.02}	0.66 _{0.16}	0.26 _{0.02}

Table 15: ImgScore and ClipScore on mitigation for all concepts. Columns 3 and 4 contain results with the original steering direction ($T_\ell(z)$), *i.e.*, removing or mitigating the concepts. Columns 5 and 6 contain results applying the inverse steering operation ($T_\ell^{-1}(z)$), which effectively *induces* the concepts. LinEAS consistently achieves a higher ImgScore for similar ClipScore values.

O Image Prompts Dataset

In Section 4.4, we evaluate multiple activation steering methods on DMD2 [30]. To probe different aspects of T2I generation, we query an open-source LLM to generate a new dataset of prompts covering 3 different conditioning categories and 5 concepts per category. We also include a neutral category, which is used as the source in concept addition and target for removal. In Table 16, we show a sample of the dataset, and we include the full dataset in the supplementary material.

Supercategory	Concept	Example Prompts
Neutral		A beaver in its natural habitat.
		A playful dolphin jumping out of water.
		A train passing through mountains.
		A sturdy table with drawers.
Style	Vaporwave	Vaporwave neon cityscape at dusk Retro-futuristic arcade machine in a vaporwave setting
	Lush	In a lush, overgrown jungle, two young men sitting on a bench and a lady standing next to them. Seagulls in flight with a person feeding one, a lighthouse in the distance, surrounded by a lush, overgrown jungle.
	Low-poly	A mountain range at sunrise rendered in low poly style with crisp, angular facets. A futuristic cityscape with neon accents and low poly geometry that creates a digital vibe.
	Ukiyo-e	A majestic view of Mount Fuji, cherry blossoms in full bloom, woodblock print style, Ukiyo-e, Edo period aesthetics A stormy sea with giant waves crashing, a lone boat struggling against the current, traditional Japanese woodblock print, Ukiyo-e
	Fantasy	A majestic dragon flying over a glowing crystal mountain range, under a purple sky, fantasy art A knight in shining armor standing before a towering, ancient forest, mist swirling around, high fantasy
Objects	Robot	A sleek silver robot waves hello in a futuristic city. A tiny robot with glowing eyes explores a dark cave.
	Axolotl	A pale pink leucistic axolotl with feathery external gills, smiling serenely in a clear aquarium. A wild-type axolotl, dark and speckled, camouflaged amongst aquatic plants in its natural habitat.
	Book	A leather-bound antique book, its gold-leaf title faded, resting on a dusty mahogany desk. A stack of colorful children’s picture books, vibrant illustrations peeking from the edges.
	Car	A gleaming cherry-red classic 1950s convertible, chrome shining, cruising down a sun-drenched coastal highway, ocean on one side. A rugged, mud-splattered off-road 4x4 vehicle navigating a steep, rocky mountain trail, dust kicking up from its tires.
	Hourglass	An antique brass hourglass, its fine golden sand steadily flowing from the top bulb to the bottom, against a dark, moody background. A minimalist, modern glass hourglass with vibrant blue sand, casting a sharp shadow on a white surface.
Perspectives	Macro	Macro shot of a ladybug on a vibrant green leaf, its tiny black spots in sharp focus, dewdrop clinging nearby. Extreme macro shot of a honeybee’s multifaceted eye, revealing intricate hexagonal patterns.
	Fisheye	Fisheye lens perspective of a bustling street market, vibrant stalls and crowds curving dramatically around a central point. Extreme fisheye shot from the center of a packed concert crowd, hands raised, stage lights creating circular flares.
	Bokeh	Portrait of a smiling woman, her face in sharp focus, against a background of beautifully blurred city lights creating circular bokeh. A single red rose in perfect focus, its delicate petals detailed, with a creamy green bokeh background of garden foliage.
	Underwater	Underwater perspective of a vibrant coral reef teeming with colorful tropical fish, sunlight filtering through clear turquoise water. Sunken pirate shipwreck resting on the sandy ocean floor, schools of fish swimming through its decaying hull, underwater view.
	Tilt-shift	Tilt-shift perspective of a bustling city intersection, cars and pedestrians appearing like tiny toys, vibrant colors. A miniature-effect tilt-shift shot of a freight train winding through a verdant, rolling landscape.

Table 16: Sample of the dataset used for conditioning T2I models. The dataset is divided in 4 different categories: (1) neutral prompts used as source for concept addition and target for removal, (2) style prompts, (3) objects, and (4) perspectives.