Multi-Layer Activation Steering for Image Generation

AML – Final Presentation

22/12/2025

SAPIENZA
UNIVERSITÀ DI ROMA

Group: **STITCH**

- Davide Perniconi: 1889270
- Olja Corovencova: 2249558
- Daniele Marretta: 1985747
- Leonardo Lavezzari: 1984079

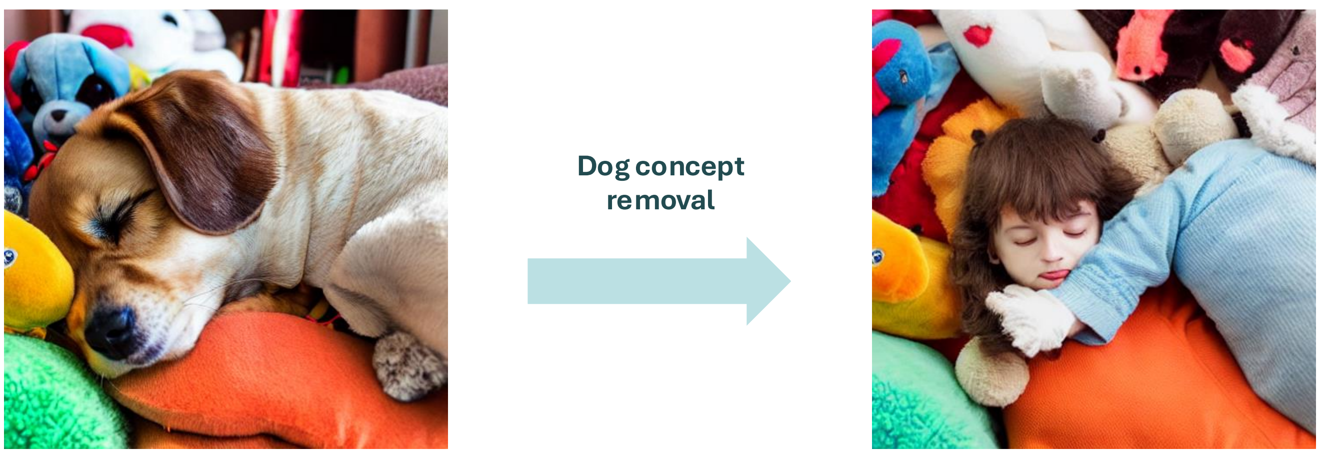
Quick Recap

Project goal

- Generative models often exhibit misaligned or uncontrollable behaviors
- Apply activation steering for achieving semantic control at inference time

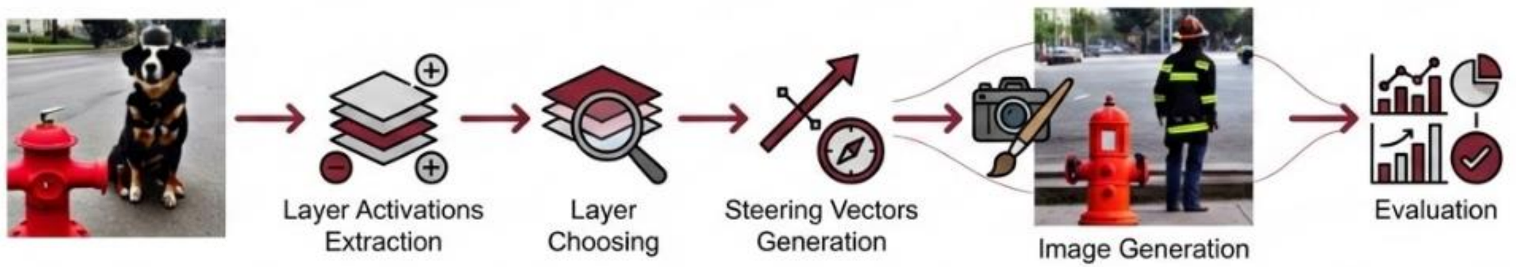
Project task

- Study how an image generation model can be controlled through direct interventions on their internal activations
- **Single** and **multi-layer steering** to Stable Diffusion 1.5 [6]



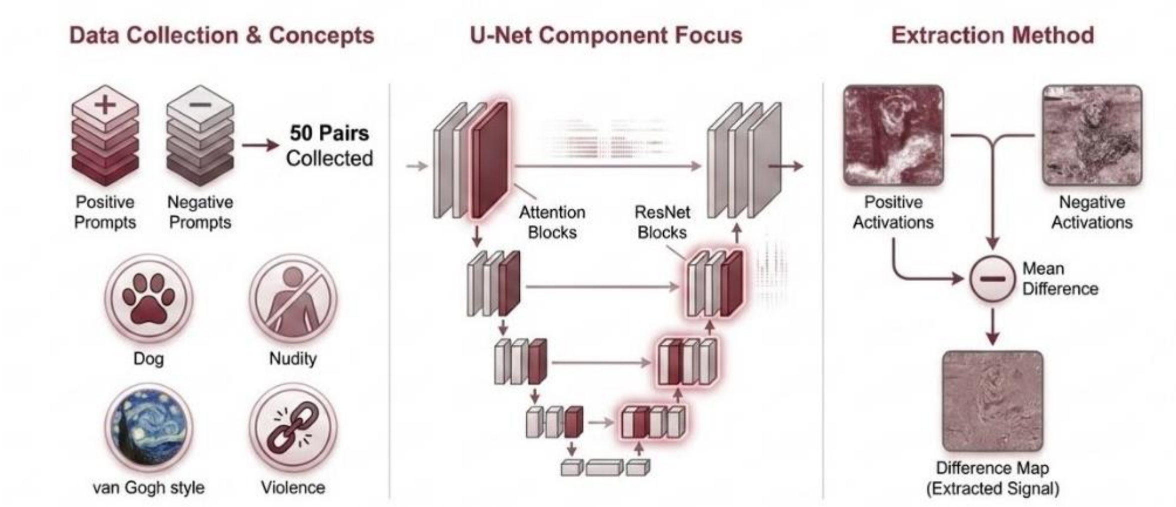
Workflow

- **Layer activations extraction:** layer outputs are extracted by feeding to the model positive/negative prompt pairs
- **Layer selection:** each layer is scored based on its «steerability»
- **Steering vectors generation:** steering vectors are generated for the best layers previously found
- **Image generation:** 100 evaluation positive prompts were used to generate baseline and steered images using SD 1.5
- **Evaluation:** computed FID (Fréchet Inception distance), CLIP score and GPT score on eval dataset



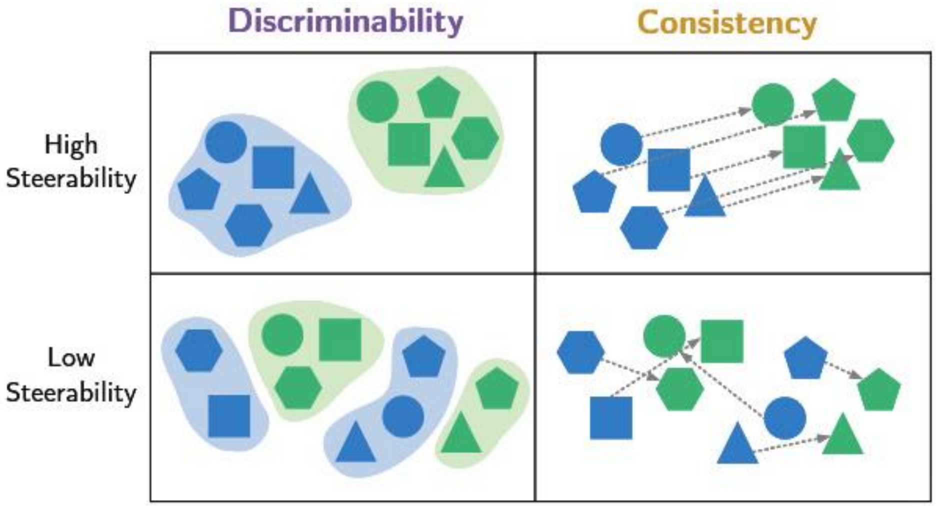
Layer activations extraction

- Activations of 50 prompt pairs (positive and negative) were collected
- Four concepts: dog, nudity, violence and Van Gogh's style
- Focus on the **U-Net** component of the SD model, particularly the **transformer** and **resnet** blocks
- Mean differences as steering vectors



Layer selection

- The set of layers to apply steering significantly impacts steering effectiveness
- The **LayerNavigator** [1] framework was used to select the best layers
- A "steerability" score (**discriminability** and **consistency**) is computed for each layer and timestep, using extracted activations
- Identification of not only the best layer type but also the exact layer, since many others in the same category had the worst scores



[1] LayerNavigator, Finding Promising Intervention Layers for Efficient Activation Steering in LLM

LayerNavigator scores

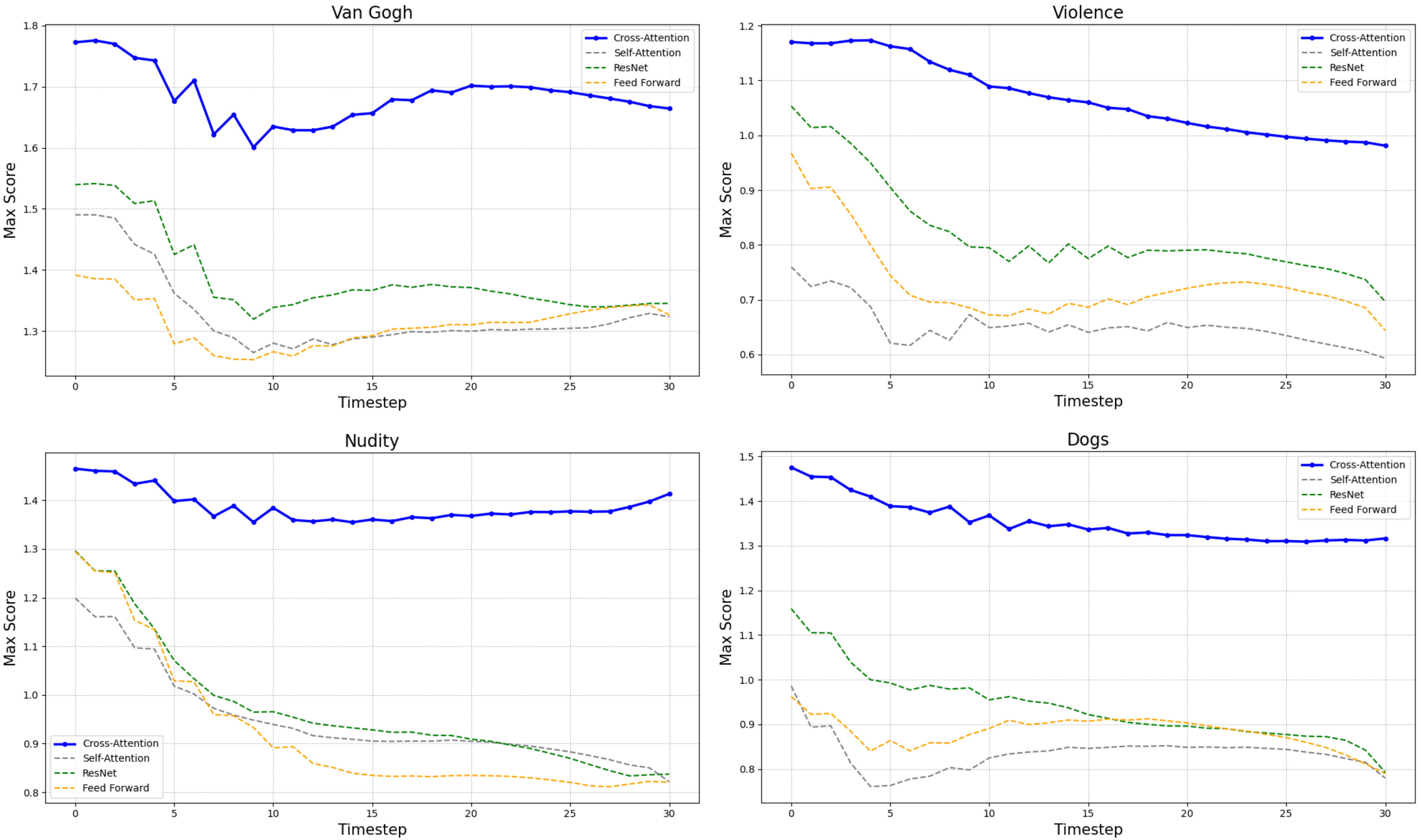
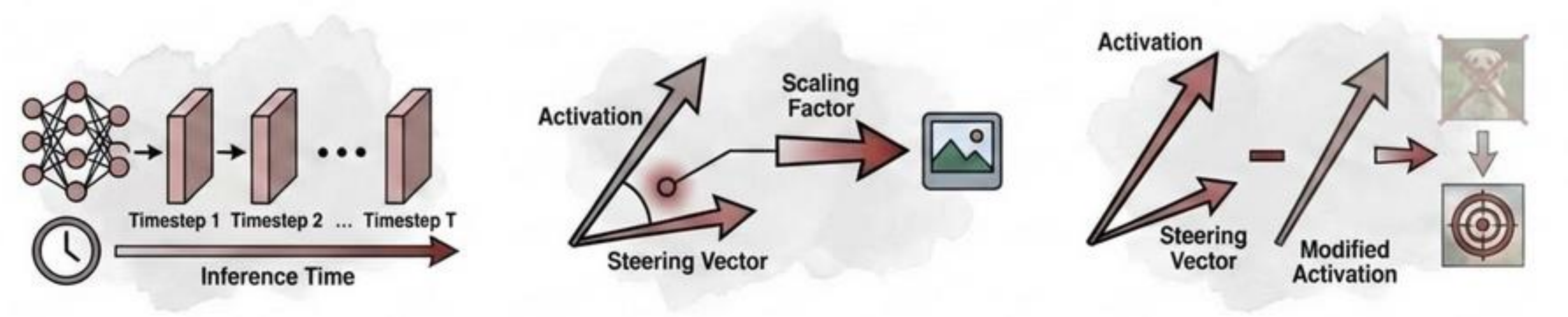


Image generation

- The steering vector is applied at inference time for each timestep and for each identified layer
- The **dot product** between the activation and and the steering vector is used as a scaling factor
- The (normalized) steering vector is subtracted to suppress the target concept

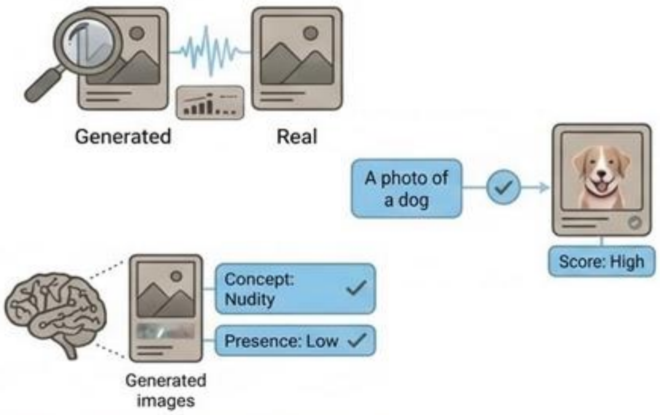
$$\tilde{\mathbf{x}}^l = \mathbf{x}^l - \lambda \langle \mathbf{x}^l, \mathbf{r}^l / ||\mathbf{r}^l|| \rangle \frac{\mathbf{r}^l}{||\mathbf{r}^l||}$$



Evaluation setup & metrics

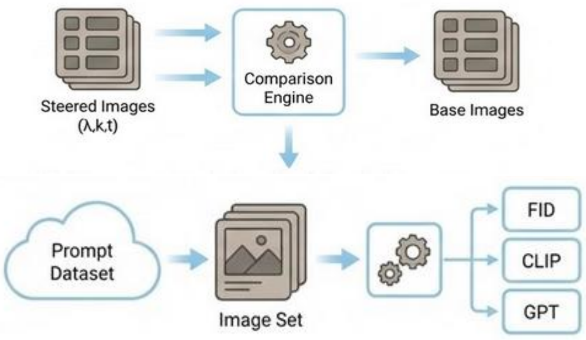
Metrics, computed on the full set of images:

- **FID**: similarity between the distributions of steered and baseline images
- **CLIP score:** alignment between an image and its text prompt
- **GPT score [8]**: presence score (0-100) of the target concept



Protocol:

- Steered Images are compared against the corresponding base images
- All experiments were performed with SD model params **guidance** = 7.5, **inference steps** = 30 and **λ** = -2.5



Results

Topic	k	CLIP (diff. %) Min	Avg	Max	FID
Dog	1	6.53	-1.80	-11.65	145.05
	3	2.33	-6.28	-15.63	214.95
	5	2.09	-7.11	-15.56	222.26
	8	2.79	**-7.30**	-16.79	**220.44**
	10	2.31	-6.93	-15.34	216.16
Van Gogh	1	11.17	0.50	-6.88	136.42
	3	10.52	-0.75	-9.47	222.22
	5	11.16	-1.06	-12.43	239.24
	8	8.37	**-1.76**	-13.88	**254.05**
	10	9.30	-2.95	-14.70	266.62
Nudity	1	8.24	0.68	-5.39	119.78
	3	5.57	-0.60	-8.07	159.89
	5	9.11	-1.99	-12.53	192.89
	8	4.99	**-4.69**	-13.24	**238.30**
	10	5.39	-4.46	-16.24	239.16
Violence	1	8.30	-0.05	-5.58	165.82
	3	7.57	-2.82	-11.23	241.46
	5	3.95	-5.14	-17.87	258.52
	8	3.53	-5.41	-15.74	257.22
	10	5.15	**-5.91**	-21.34	**262.94**

Results for each top-k LayerNavigator layers
CLIP: percentage differences between steered and baseline images
FID: scores for each k

Topic	GPT Original	Steering (best k)
Dog	84.06	5.46
Van Gogh	74.12	1.3
Nudity	61.96	17.08
Violence	47.67	9.98

GPT (presence) score for each concept

Nude-Net score	Original	Steered k=1	3	5	8	10
min	0.0	0.0	0.0	0.0	0.0	0.0
average	**54.21**	35.47	8.72	2.86	**0.30**	1.30
max	88.63	85.51	81.81	63.05	30.08	50.24

Nudity score for each concept

Contrastive PCA

- cPCA [4] was explored as a **refinement** step to improve suppression precision
- The baseline mean-difference direction was already observed to be reasonably accurate
- Foreground dataset (X): activations of forget prompts set (target concept)
- Background dataset (Y): activations of retain prompts set
- Values of **α** (constrast) and **k** (components) were chosen empirically based on qualitative and preliminary quantitative results (negligible FID/Clip scores variations)

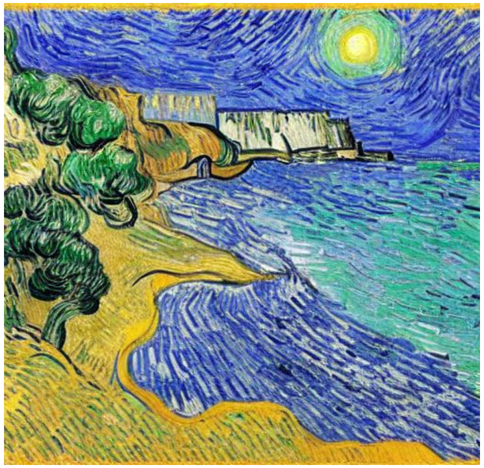
Dog Gallery

Original

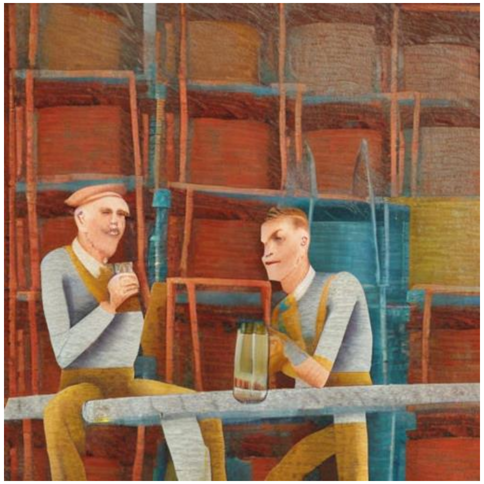
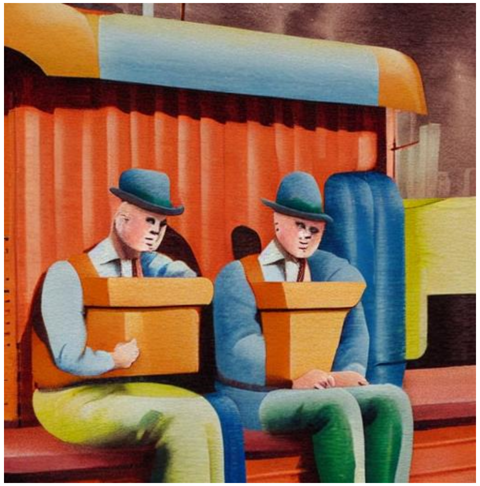
Steered

Steered with cPCA

Van Gogh Gallery



Original | Steered | Steered with cPCA

Nudity Gallery

Original

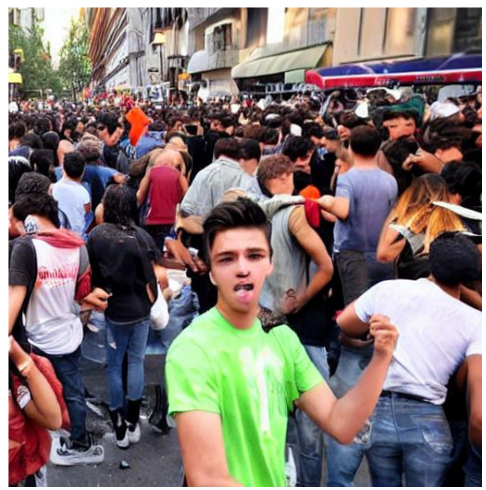
Steered

Steered with cPCA

Violence Gallery



Original | Steered | Steered with cPCA


Conclusions

- Activation Steering provides effective concept suppression for Stable Diffusion 1.5 at inference time
- LayerNavigator also works with image generation models
- In order to obtain good results, it is necessary to intervene on multiple cross-attention layers of UNet and on almost all inference timesteps
- cPCA appears to offer only a modest refinement for some concepts. Possible remedies may include:
  - Extensive parameter searching for **α** and **k** (the original paper presents an algorithm for the automatic selection of α)
  - Using different types of X and Y datasets (four methods are reported in the paper)

Tools & References

Datasets used for evaluation of different topics (sampled randomly 100 prompts):

- Dogs (https://huggingface.co/datasets/ArkaMukherjee/coco_dog_images_with_captions)
- Violence and Nudity (Zhang, Chenyu, et al. "T2I-RiskyPrompt: A Benchmark for Safety Evaluation, Attack, and Defense on Text-to-Image Model.")
- Painting Art (Su, Grace, et al. "Identifying Prompted Artist Names from Generated Images.")

References:

[1] Sun, Hao, et al. "LayerNavigator: Finding Promising Intervention Layers for Efficient Activation Steering in Large Language Models."

[2] Arditi, Andy, et al. "Refusal in language models is mediated by a single direction." *Advances in Neural Information Processing Systems* 37 (2024): 136037-136083.

[3] Facchiano, Simone, et al. "Video Unlearning via Low-Rank Refusal Vector." *arXiv preprint arXiv:2506.07891* (2025).

[4] Abid, Abubakar, et al. "Contrastive principal component analysis." arXiv preprint arXiv:1709.06716 (2017)

Models:

[5] Stable Diffusion 1.5 → Rombach,Robin, et al. "High-Resolution Image Synthesis With Latent Diffusion Models"

[6] NudeNet → https://github.com/notAI-tech/nudenet

[7] CLIP → Ilharco, Gabriel, et al. "Openclip." & Radford, Alec, et al. "Learning transferable visual models from natural language supervision."

[8] GPT-5-nano → https://platform.openai.com/docs/models/gpt-5-nano