One-Step is Enough: **Sparse Autoencoders for Text-to-Image Diffusion Models**

Viacheslav Surkov* · Chris Wendler* · Antonio Mari · Mikhail Terekhov · Justin Deschenaux · Robert West · Caglar Gulcehre · David Bau

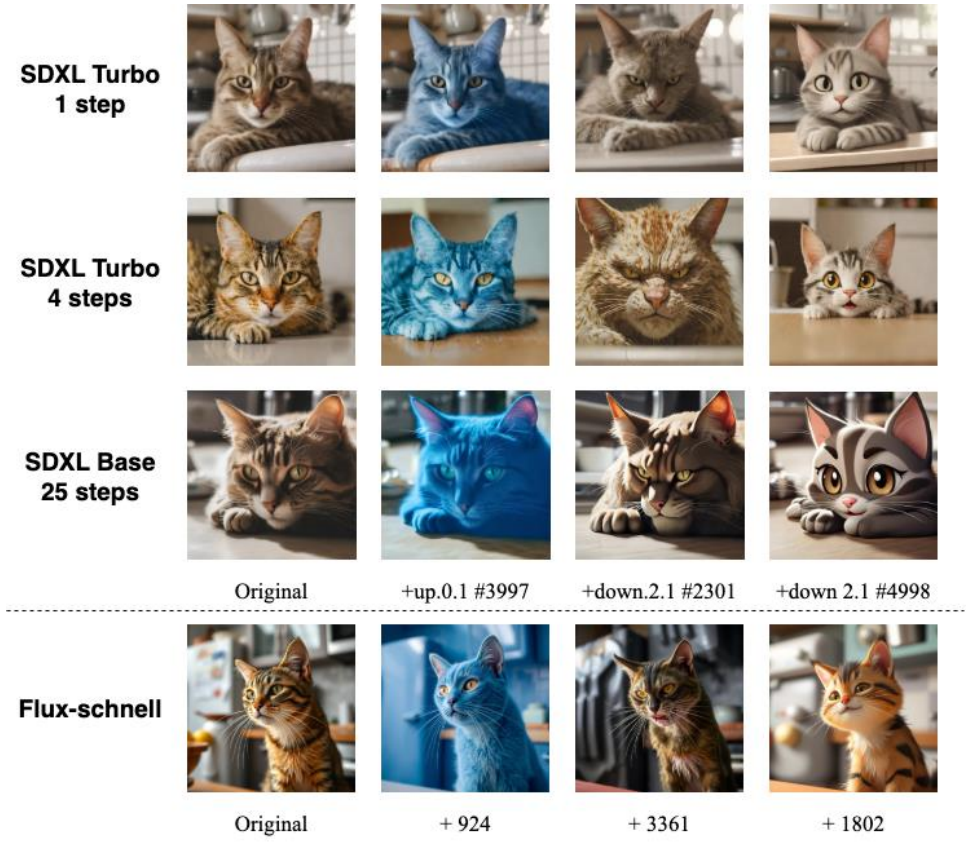
Why this

Make diffusion models less of a black box by lifting LLM-style sparse features into SDXL & FLUX – and editing them live

Why this?

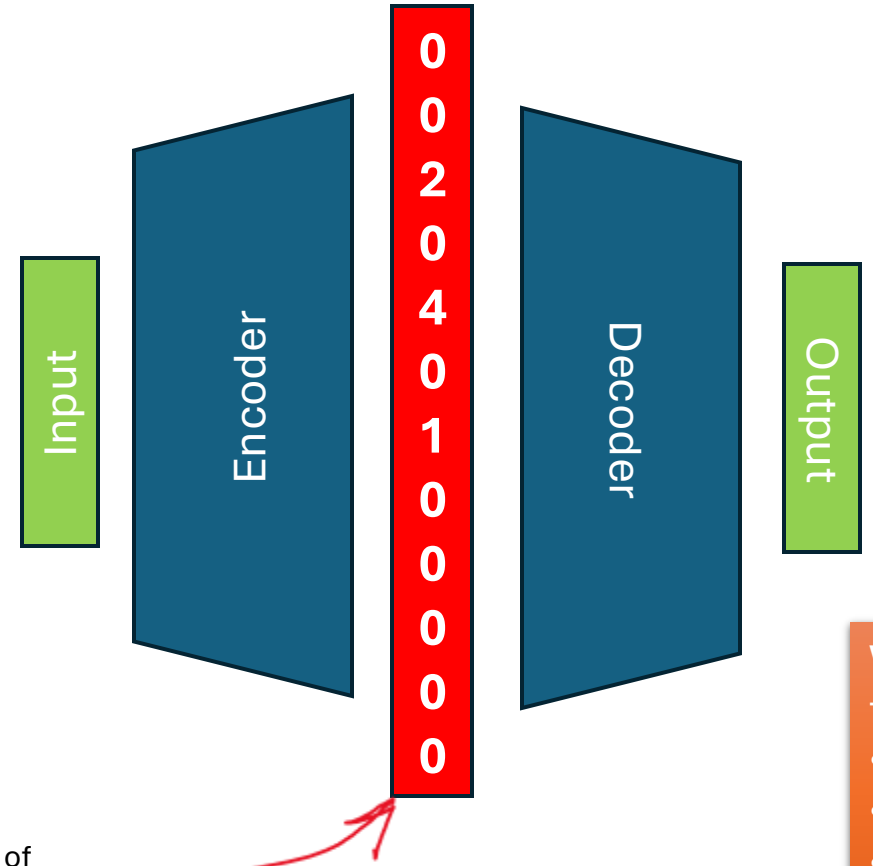
- Diffusion models are poorly understood internally
- Mechanistic work lags behind LLMs
- SAEs succeeded in LLMs at decomposing polysemantic reps into sparse, interpretable features

- ***Can this transfer to text-to-image?***



Where the SAEs live

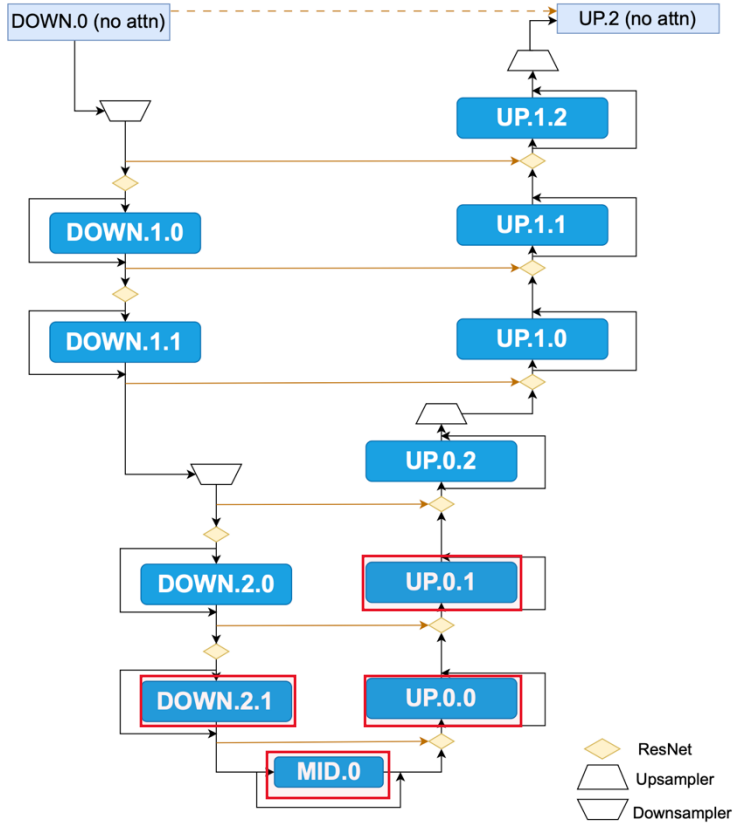
Sparse autoencoder



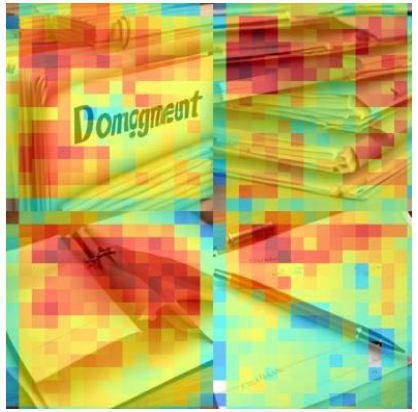
We selected 4 blocks for analysis

- **down.2.1**
- **mid.0**
- **up.0.0**
- **up.0.1**

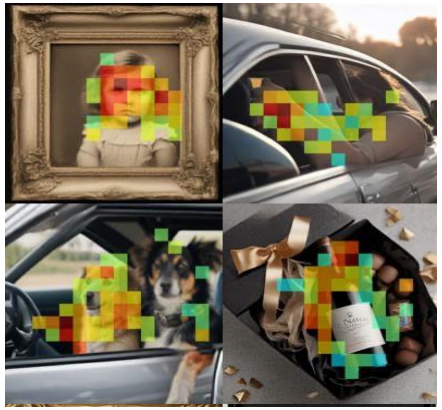
U-net of Stable Diffusion XL



Where the features are active

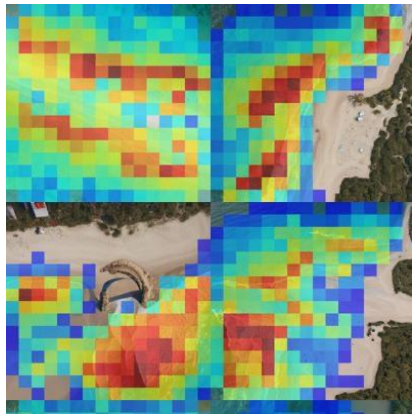


down.2.1 #0 (folders)

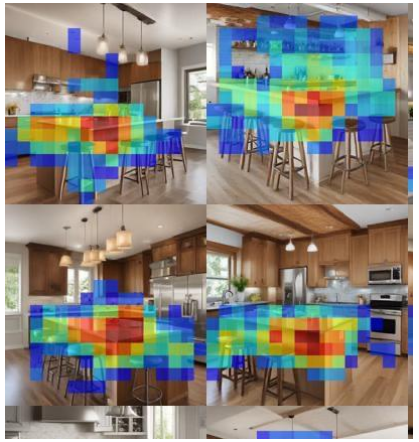


mid.0 #0 (inside)

up.0.0 #0 (glass stems)



up.0.1 #0 (water)



down.2.1 #1 (kitchen tables)

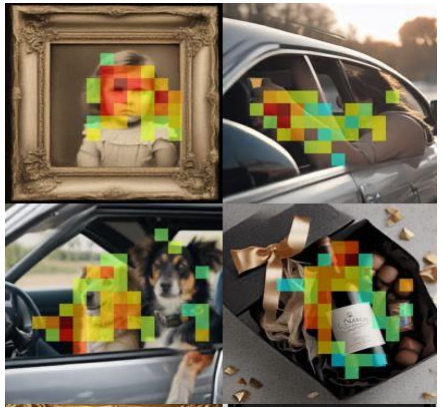
mid.1 #1 (highest in a group)

up.0.0 #1 (roof edges)

up.0.1 #1 (shelves)

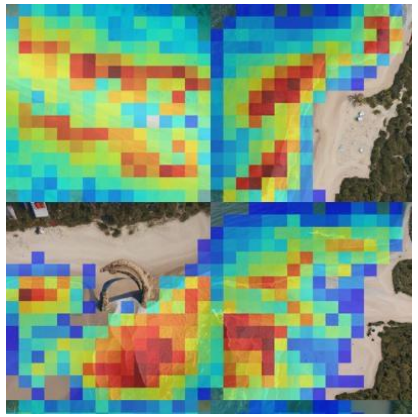
Where the features are active

down.2.1 #0 (folders)

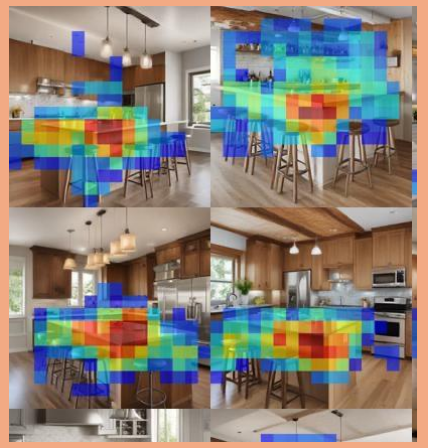


mid.0 #0 (inside)

up.0.0 #0 (glass stems)



up.0.1 #0 (water)



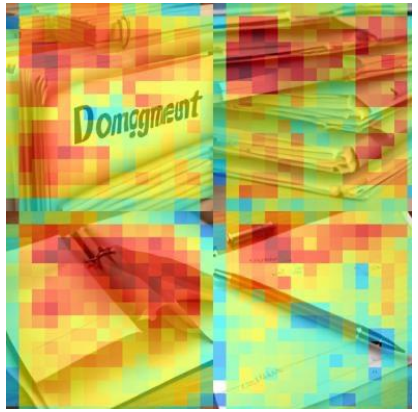
down.2.1 #1 (kitchen tables)

mid.1 #1 (highest in a group)

up.0.0 #1 (roof edges)

up.0.1 #1 (shelves)

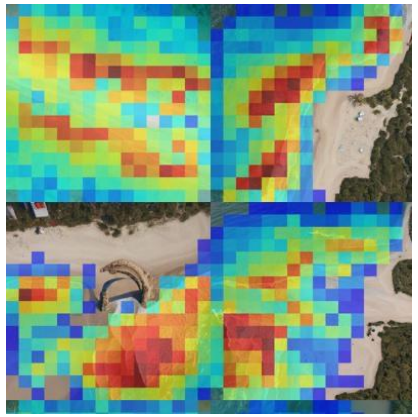
Where the features are active



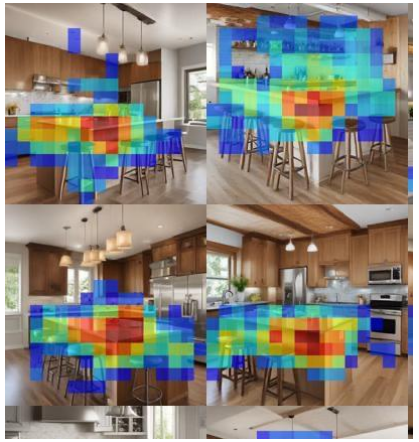
down.2.1 #0 (folders)

mid.0 #0 (inside)

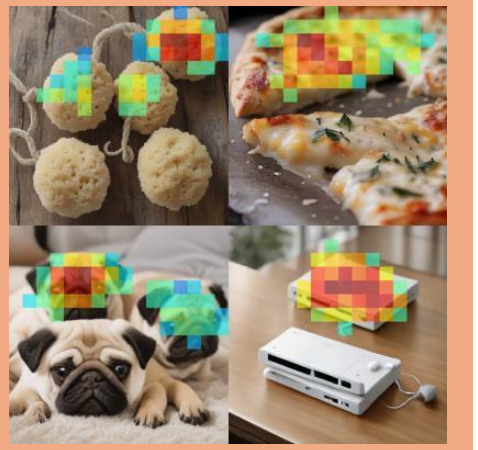
up.0.0 #0 (glass stems)



up.0.1 #0 (water)



down.2.1 #1 (kitchen tables)

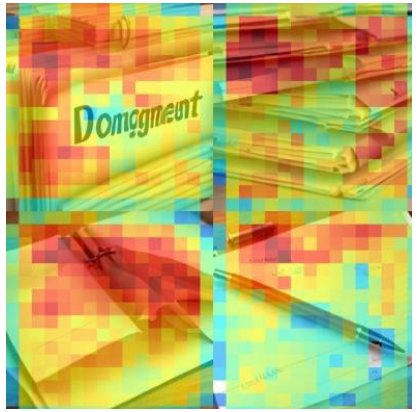


mid.1 #1 (highest in a group)

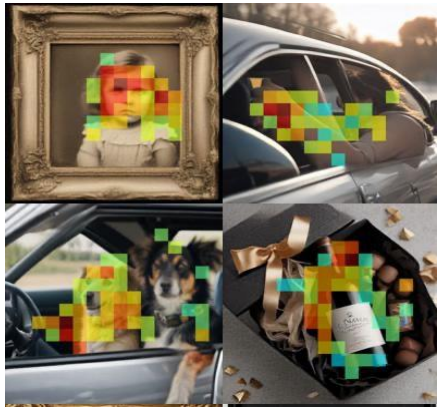
up.0.0 #1 (roof edges)

up.0.1 #1 (shelves)

Where the features are active

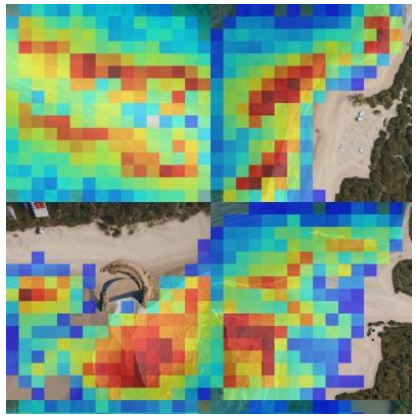


down.2.1 #0 (folders)

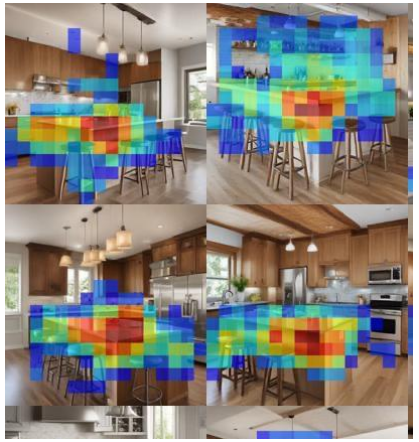


mid.0 #0 (inside)

up.0.0 #0 (glass stems)



up.0.1 #0 (water)



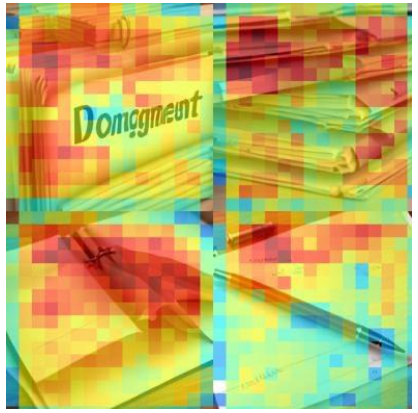
down.2.1 #1 (kitchen tables)

mid.1 #1 (highest in a group)

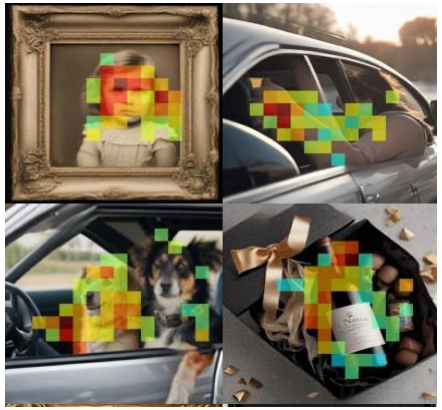
up.0.0 #1 (roof edges)

up.0.1 #1 (shelves)

Where the features are active

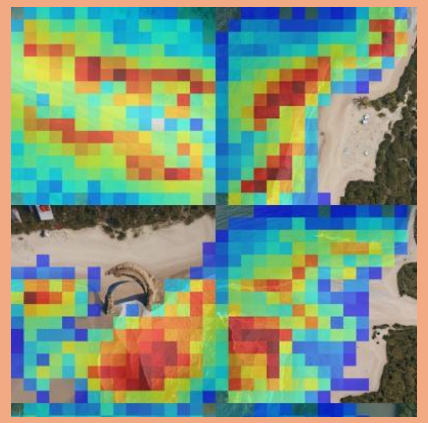


down.2.1 #0 (folders)

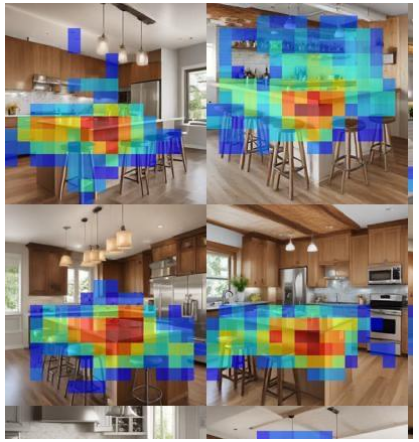


mid.0 #0 (inside)

up.0.0 #0 (glass stems)



up.0.1 #0 (water)



down.2.1 #1 (kitchen tables)

mid.1 #1 (highest in a group)

up.0.0 #1 (roof edges)

up.0.1 #1 (shelves)

Features are causal

Add greenery

down.2.1 #1802

down.2.1 #1802

down.2.1 #1802

Add collar

up.0.0 #4473

up.0.0 #4473

up.0.0 #4473

Change orientation

mid.0 #4227

mid.0 #4227

mid.0 #4227

Add tiger texture

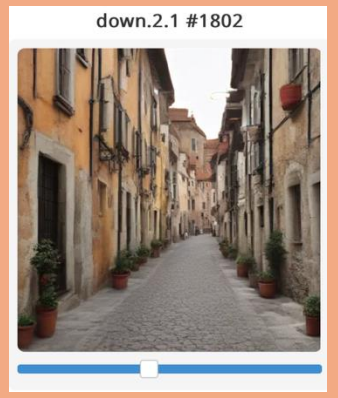
up.0.1 #4977

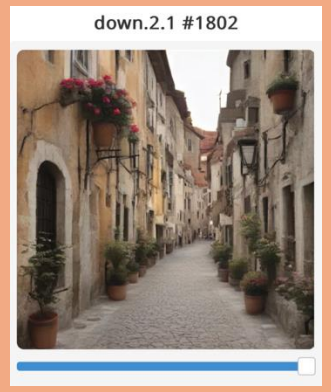
up.0.1 #4977

up.0.1 #4977

Features are causal

Add greenery





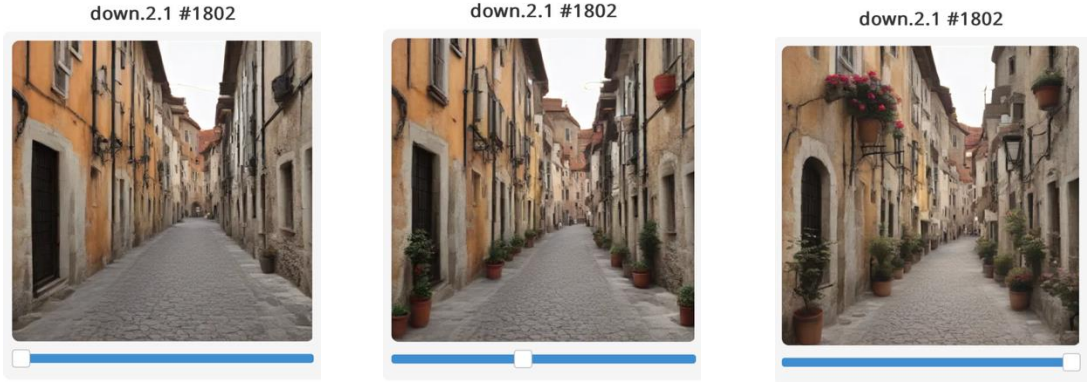
Add collar

Change orientation

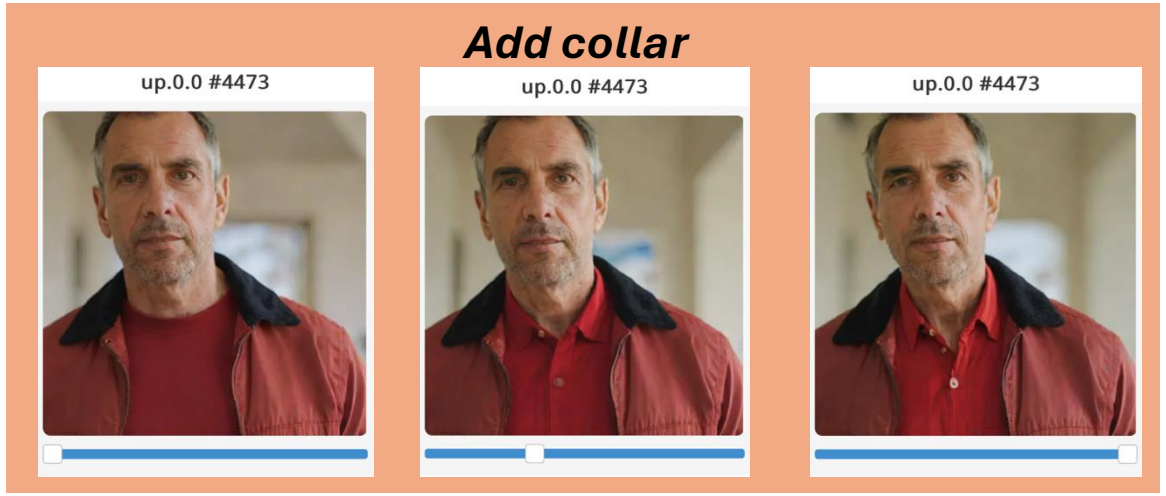
Add tiger texture

Features are causal

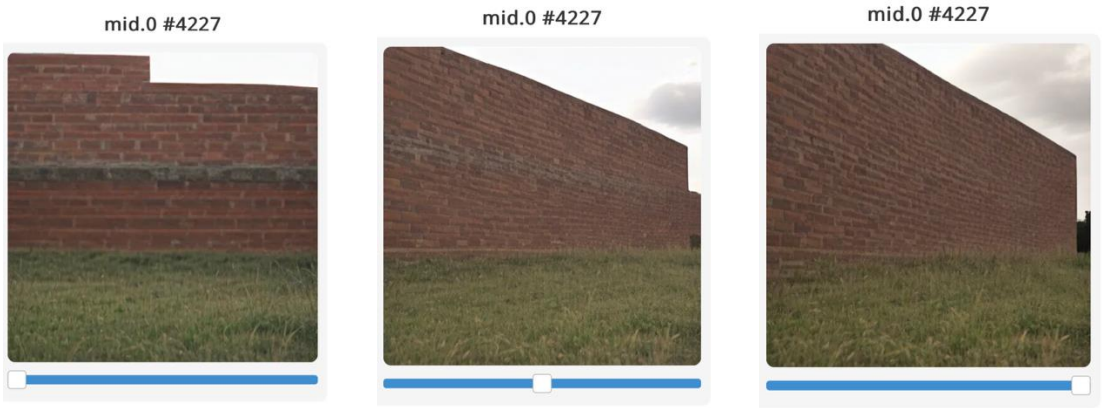
Add greenery



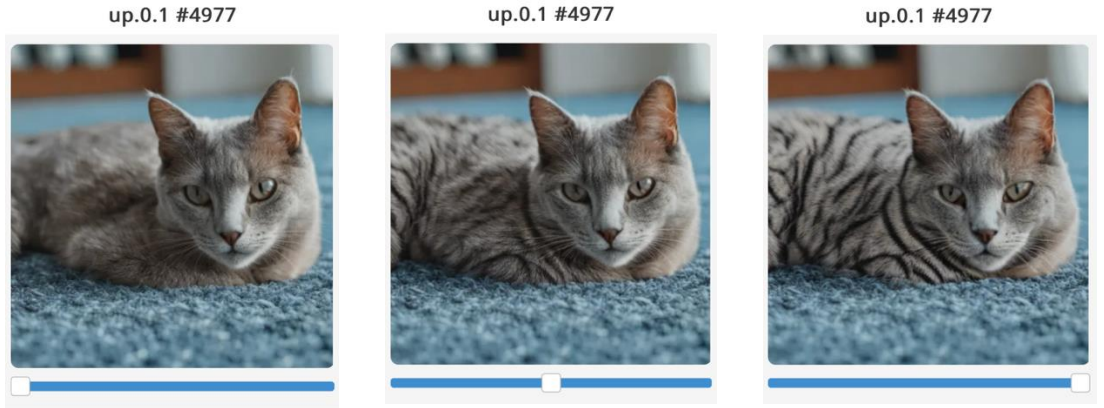
Add collar



Change orientation

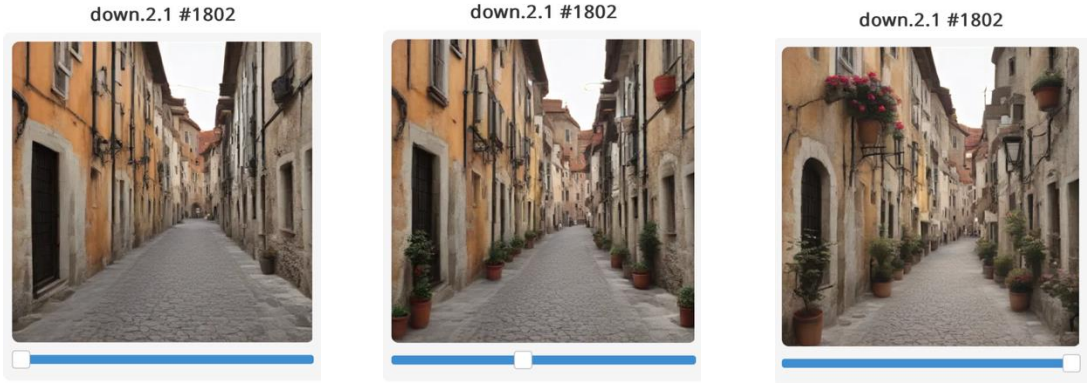


Add tiger texture

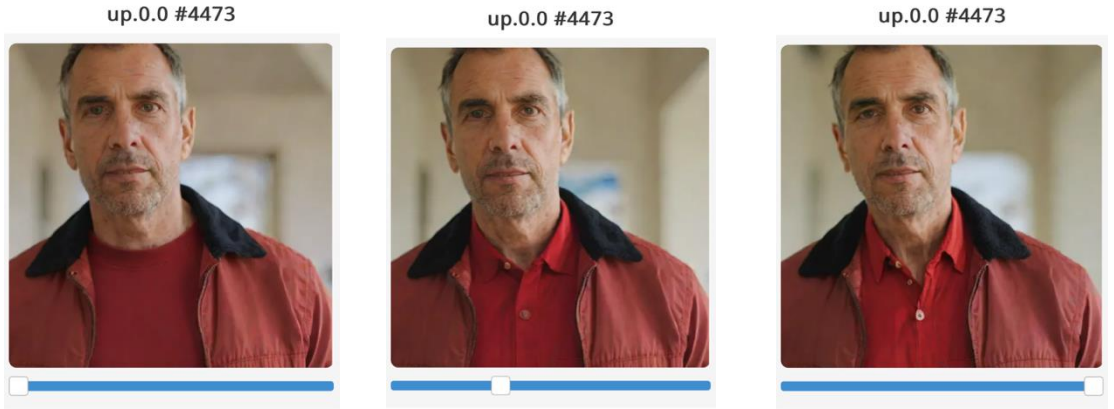


Features are causal

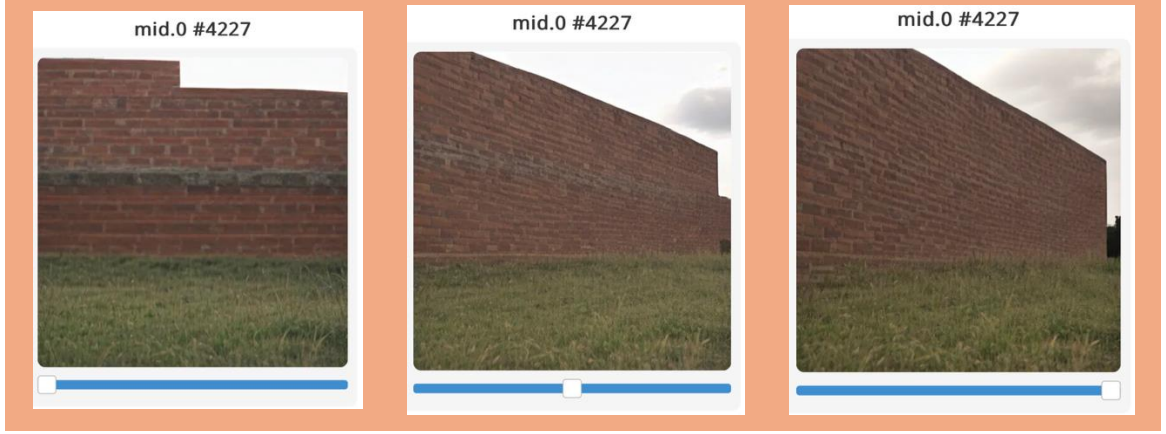
Add greenery



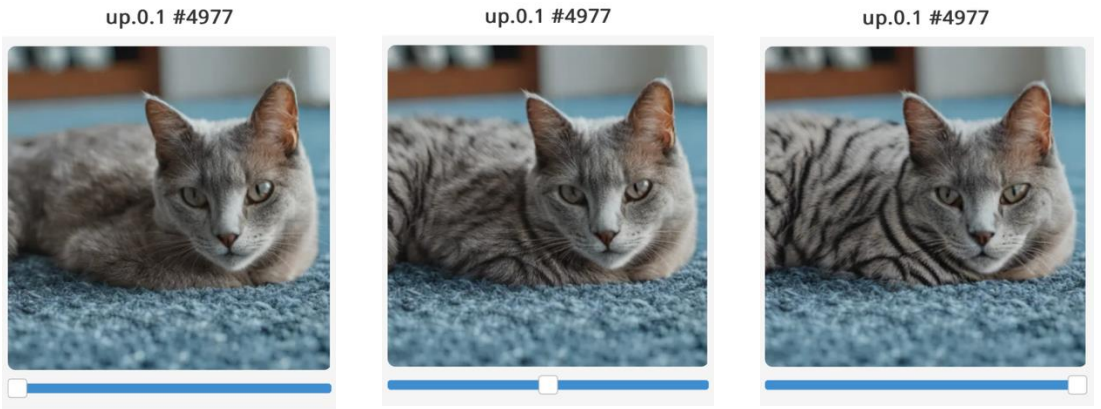
Add collar



Change orientation



Add tiger texture



Features are causal

Add greenery

down.2.1 #1802

down.2.1 #1802

down.2.1 #1802

Add collar

up.0.0 #4473

up.0.0 #4473

up.0.0 #4473

Change orientation

mid.0 #4227

mid.0 #4227

mid.0 #4227

Add tiger texture

up.0.1 #4977

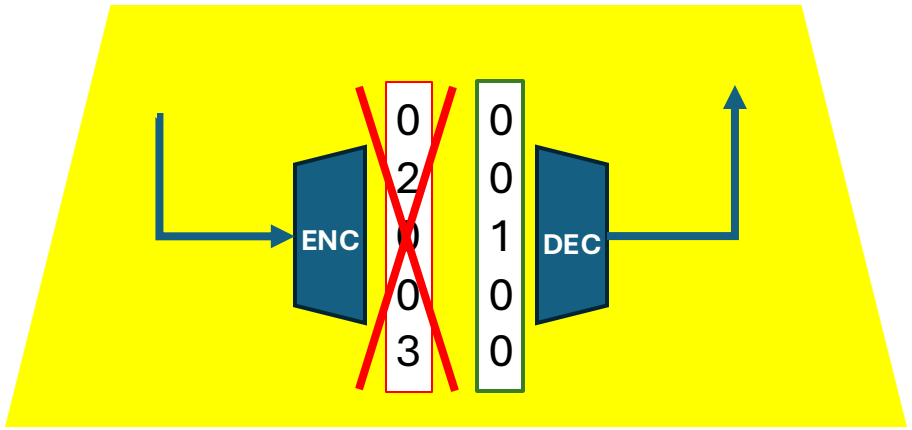
up.0.1 #4977

up.0.1 #4977

Features in isolation

Zero-prompt intervention

prompt=“”



down.2.1

up.0.1

Zore-prompt interventions are novel in mechanistic interpretability: they haven’t been explored in LLMs

“One-step is enough” generalization

- What if we apply single-step SAEs to multi-step models?
  - 4 step SDXL Turbo
  - 20 step SDXL Base

- ***Interventions work well without any retraining!***

What the blocks mean

SAEs reveal SDXL's blocks specializations

down.2.1
Change style
/ add object

up.0.1
Change color/texture

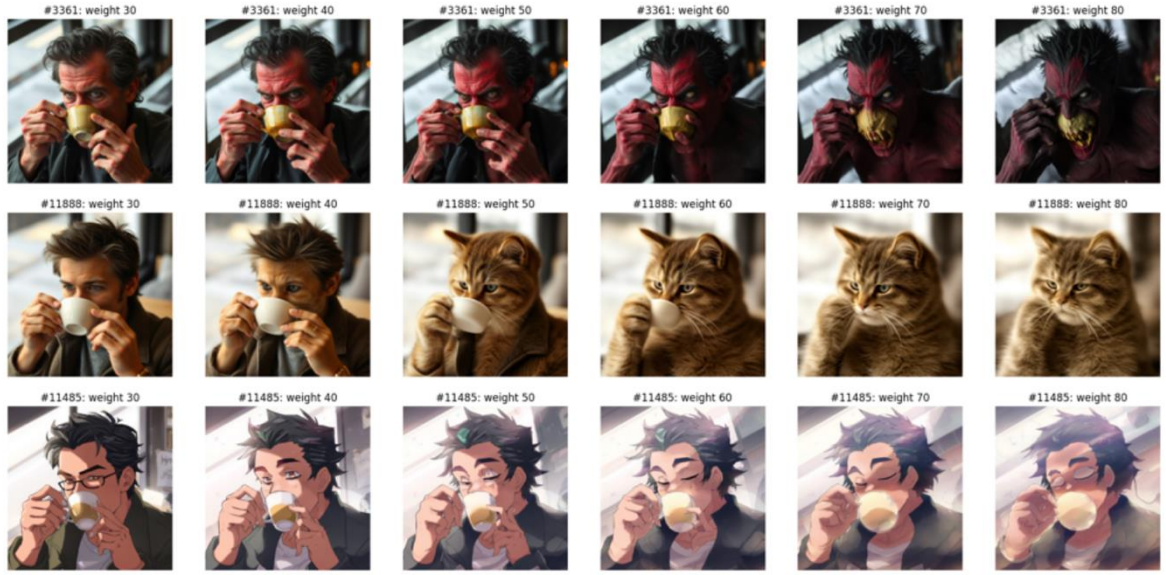
up.0.0
Change detail

Takeaways: SAEs give **causal knobs**, **1-step training generalizes**, and blocks show **clear roles**.

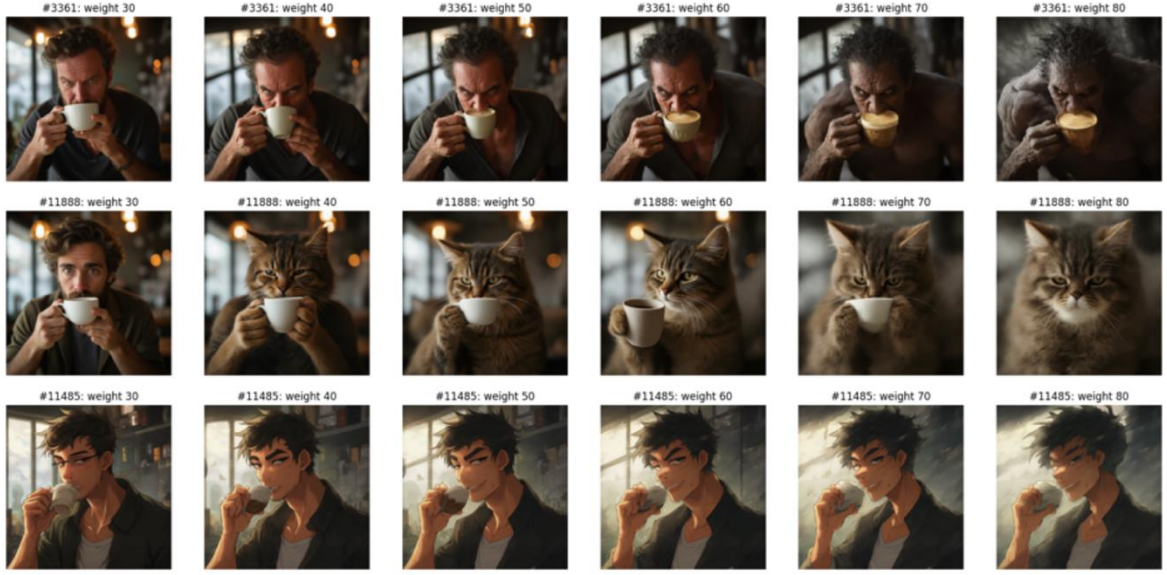
Bonus: Flux

SAEs generalize to DiTs

Flux-schnell (1 step)



Flux-dev (25 steps)

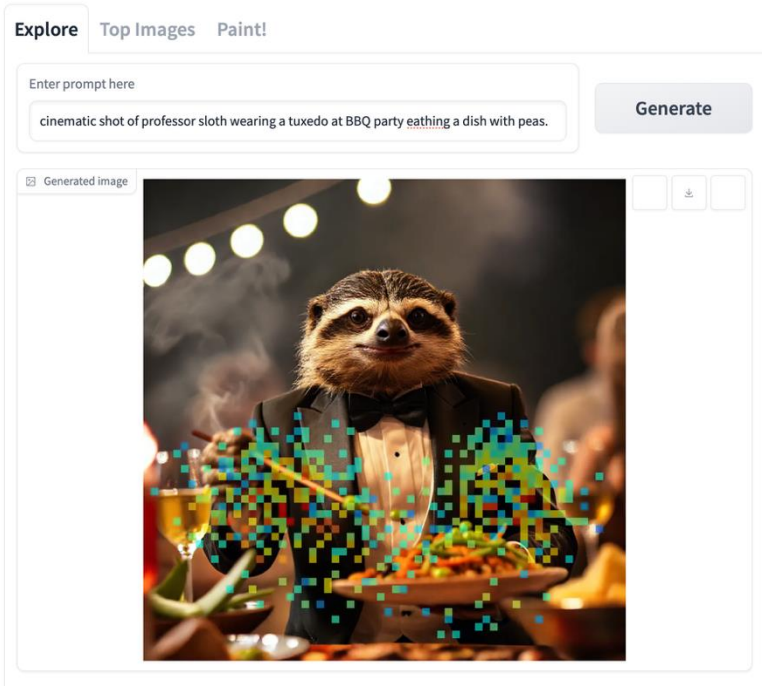


Play with the features – live

Project website

sdxl-unbox.epfl.ch

HF Demos



SDXL Turbo

SDXL Base

FLUX schnell