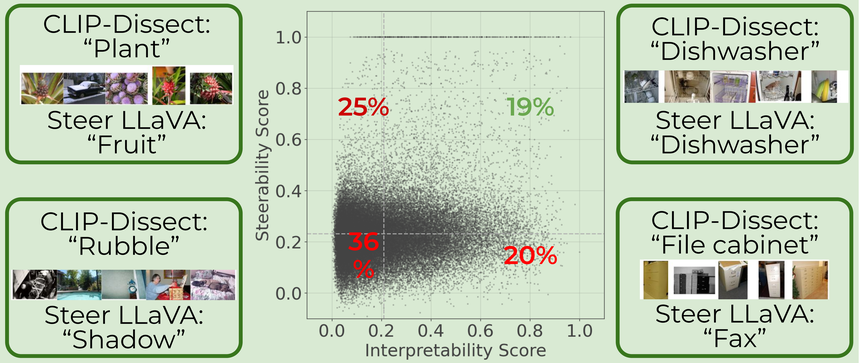
Interpretable and Steerable Concept Bottleneck Sparse Autoencoders

Akshay Kulkarni[1], Tsui-Wei (Lily) Weng[1], Vivek Narayanaswamy[2], Shusen Liu[2], Wesam A. Sakla[2], Kowshik Thopalli[2]
[1]UC San Diego, [2]LLNL

Lawrence Livermore National Laboratory
UC San Diego
CVPR JUNE 3-7, 2026 DENVER COLORADO

Sparse Autoencoder Analysis



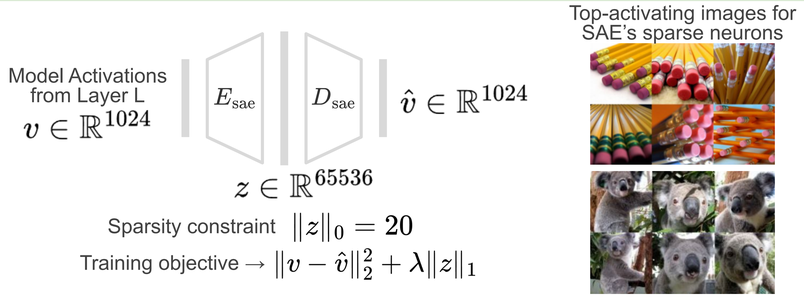
Most SAE neurons are low utility!

For an ImageNet-pretrained 65k-dim SAE

Concept set	Coverage
Broden	1.1k/1.2k (96%)
3k English words	1.8k/3.0k (62%)
VLG-CBM ImageNet	3.4k/4.7k (73%)
DECIDER ImageNet	4.3k/7.8k (55%)
20k English words	5.6k/20k (28%)

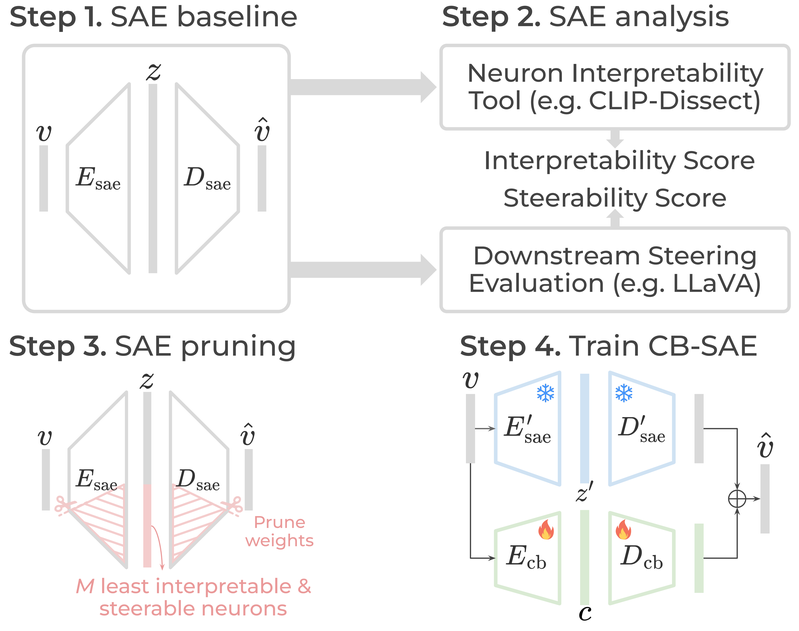
User-specified concepts missing!

Background: Sparse Autoencoders

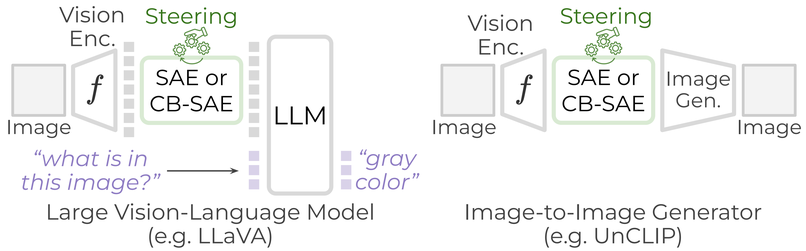


SAEs can perform unsupervised concept discovery with sparse neurons or decoder columns representing concepts

Our Proposed Approach: CB-SAE

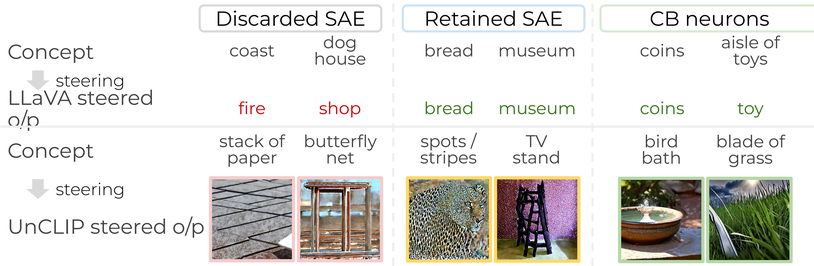


Usecase: Steering Generative Models



SAEs/CB-SAEs can steer multiple downstream tasks

Result I: Qualitative



Result II: Quantitative

Downstream Task	Steered Model	Method	Interpretability CD	Interpretability MS	Steerability
Image + Text → Text Generation	LLaVA-1.5-7B (CLIP-ViT-L + Vicuna-7B)	SAE	0.154	0.517	0.198
		CB-SAE (*Ours*)	**0.244**	**0.556**	**0.261**
	LLaVA-MORE (DINOv2-L + Gemma2-9B)	SAE	0.194	0.553	0.179
		CB-SAE (*Ours*)	**0.291**	**0.598**	**0.192**
Image → Image Generation	UnCLIP (CLIP-ViT-L + SD-2.1)	SAE	0.058	0.540	0.642
		CB-SAE (*Ours*)	**0.092**	**0.594**	**0.659**

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC. and was supported by the LLNL-LDRD Program under Project No. 25- SI-001 and DOE ECRP 51917/SCW1885. LLNL-CONF2013863.

Interpretable and Steerable Concept Bottleneck Sparse Autoencoders

Akshay Kulkarni[1], Tsui-Wei (Lily) Weng[1], Vivek Narayanaswamy[2], Shusen Liu[2], Wesam A. Sakla[2], Kowshik Thopalli[2]
[1]UC San Diego, [2]LLNL

Lawrence Livermore National Laboratory
UC San Diego
CVPR JUNE 3-7, 2026 DENVER COLORADO

Background: Sparse Autoencoders

Model Activations from Layer L $v \in \mathbb{R}^{1024}$

E_{sae} D_{sae} $\hat{v} \in \mathbb{R}^{1024}$

$z \in \mathbb{R}^{65536}$

Sparsity constraint $\|z\|_0 = 20$

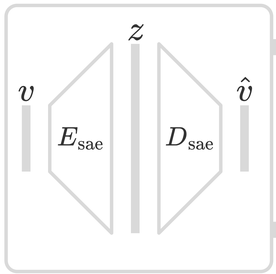
Training objective → $\|v - \hat{v}\|_2^2 + \lambda\|z\|_1$

Top-activating images for SAE's sparse neurons

SAEs can perform unsupervised concept discovery with sparse neurons or decoder columns representing concepts

Our Proposed Approach: CB-SAE

Step 1. SAE baseline



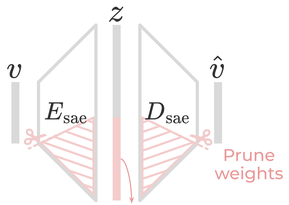
Step 2. SAE analysis

Neuron Interpretability Tool (e.g. CLIP-Dissect)

Interpretability Score
Steerability Score

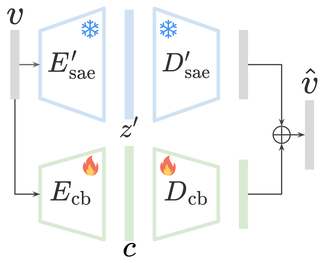
Downstream Steering Evaluation (e.g. LLaVA)

Step 3. SAE pruning



M least interpretable & steerable neurons

Step 4. Train CB-SAE



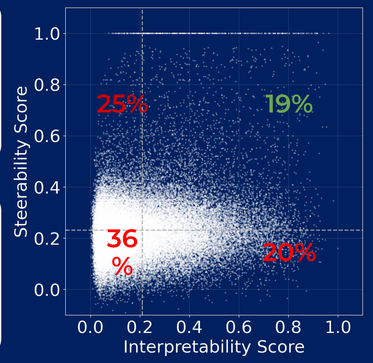
Sparse Autoencoder Analysis

CLIP-Dissect: "Plant"

Steer LLaVA: "Fruit"

CLIP-Dissect: "Rubble"

Steer LLaVA: "Shadow"



CLIP-Dissect: "Dishwasher"

Steer LLaVA: "Dishwasher"

CLIP-Dissect: "File cabinet"

Steer LLaVA: "Fax"

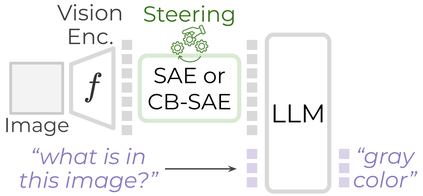
Most SAE neurons are low utility!

For an ImageNet-pretrained 65k-dim SAE

Concept set	Coverage
Broden	1.1k/1.2k (96%)
3k English words	1.8k/3.0k (62%)
VLG-CBM ImageNet	3.4k/4.7k (73%)
DECIDER ImageNet	4.3k/7.8k (55%)
20k English words	5.6k/20k (28%)

User-specified concepts missing!

Usecase: Steering Generative Models



Large Vision-Language Model (e.g. LLaVA)

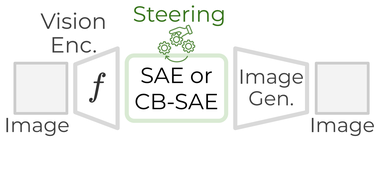
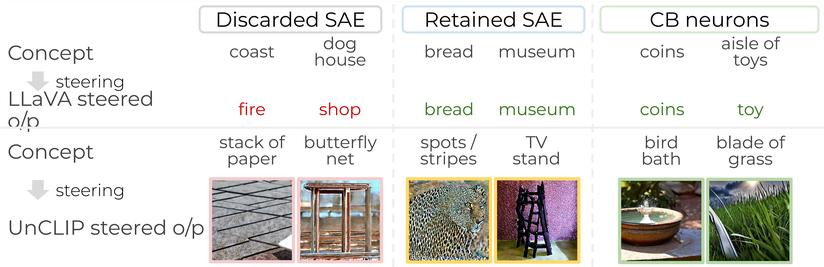


Image-to-Image Generator (e.g. UnCLIP)

SAEs/CB-SAEs can steer multiple downstream tasks

Result I: Qualitative



Result II: Quantitative

Downstream Task	Steered Model	Method	Interpretability CD	Interpretability MS	Steerability
Image + Text → Text Generation	LLaVA-1.5-7B (CLIP-ViT-L + Vicuna-7B)	SAE	0.154	0.517	0.198
		CB-SAE (*Ours*)	**0.244**	**0.556**	**0.261**
	LLaVA-MORE (DINOv2-L + Gemma2-9B)	SAE	0.194	0.553	0.179
		CB-SAE (*Ours*)	**0.291**	**0.598**	**0.192**
Image → Image Generation	UnCLIP (CLIP-ViT-L + SD-2.1)	SAE	0.058	0.540	0.642
		CB-SAE (*Ours*)	**0.092**	**0.594**	**0.659**

Interpretable and Steerable Concept Bottleneck Sparse Autoencoders

Akshay Kulkarni[1], Tsui-Wei (Lily) Weng[1], Vivek Narayanaswamy[2], Shusen Liu[2], Wesam A. Sakla[2], Kowshik Thopalli[2]
[1]UC San Diego, [2]LLNL

CVPR JUNE 3-7, 2026 DENVER COLORADO

Background: Sparse Autoencoders

Model Activations from Layer L $v \in \mathbb{R}^{1024}$

E_{sae} D_{sae} $\hat{v} \in \mathbb{R}^{1024}$

$z \in \mathbb{R}^{65536}$

Sparsity constraint $\|z\|_0 = 20$

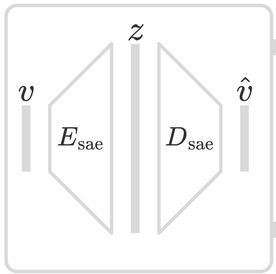
Training objective $\rightarrow \|v - \hat{v}\|_2^2 + \lambda\|z\|_1$

Top-activating images for SAE's sparse neurons

SAEs can perform unsupervised concept discovery with sparse neurons or decoder columns representing concepts

Our Proposed Approach: CB-SAE

Step 1. SAE baseline



Step 2. SAE analysis

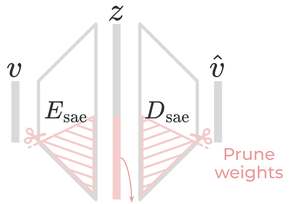
Neuron Interpretability Tool (e.g. CLIP-Dissect)

Interpretability Score

Steerability Score

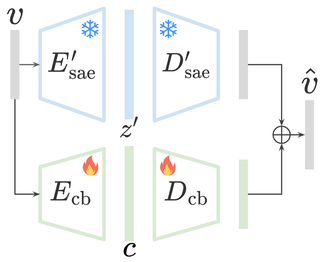
Downstream Steering Evaluation (e.g. LLaVA)

Step 3. SAE pruning



M least interpretable & steerable neurons

Step 4. Train CB-SAE



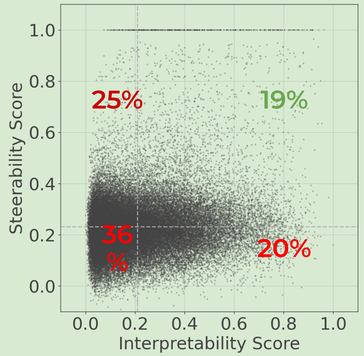
Sparse Autoencoder Analysis

CLIP-Dissect: "Plant"

Steer LLaVA: "Fruit"

CLIP-Dissect: "Rubble"

Steer LLaVA: "Shadow"



CLIP-Dissect: "Dishwasher"

Steer LLaVA: "Dishwasher"

CLIP-Dissect: "File cabinet"

Steer LLaVA: "Fax"

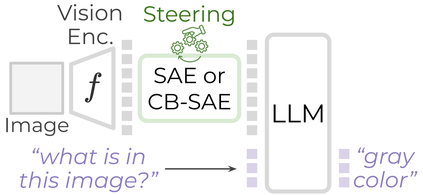
Most SAE neurons are low utility!

For an ImageNet-pretrained 65k-dim SAE

Concept set	Coverage
Broden	1.1k/1.2k (96%)
3k English words	1.8k/3.0k (62%)
VLG-CBM ImageNet	3.4k/4.7k (73%)
DECIDER ImageNet	4.3k/7.8k (55%)
20k English words	5.6k/20k (28%)

User-specified concepts missing!

Usecase: Steering Generative Models



Large Vision-Language Model (e.g. LLaVA)

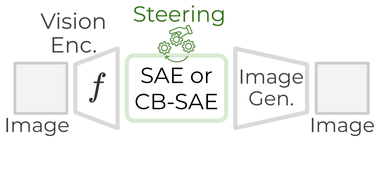
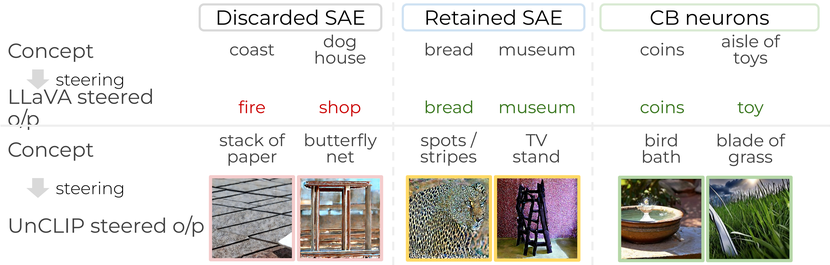


Image-to-Image Generator (e.g. UnCLIP)

SAEs/CB-SAEs can steer multiple downstream tasks

Result I: Qualitative



Result II: Quantitative

Downstream Task	Steered Model	Method	Interpretability CD	Interpretability MS	Steerability
Image + Text → Text Generation	LLaVA-1.5-7B (CLIP-ViT-L + Vicuna-7B)	SAE	0.154	0.517	0.198
		CB-SAE (*Ours*)	**0.244**	**0.556**	**0.261**
	LLaVA-MORE (DINOv2-L + Gemma2-9B)	SAE	0.194	0.553	0.179
		CB-SAE (*Ours*)	**0.291**	**0.598**	**0.192**
Image → Image Generation	UnCLIP (CLIP-ViT-L + SD-2.1)	SAE	0.058	0.540	0.642
		CB-SAE (*Ours*)	**0.092**	**0.594**	**0.659**