Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models

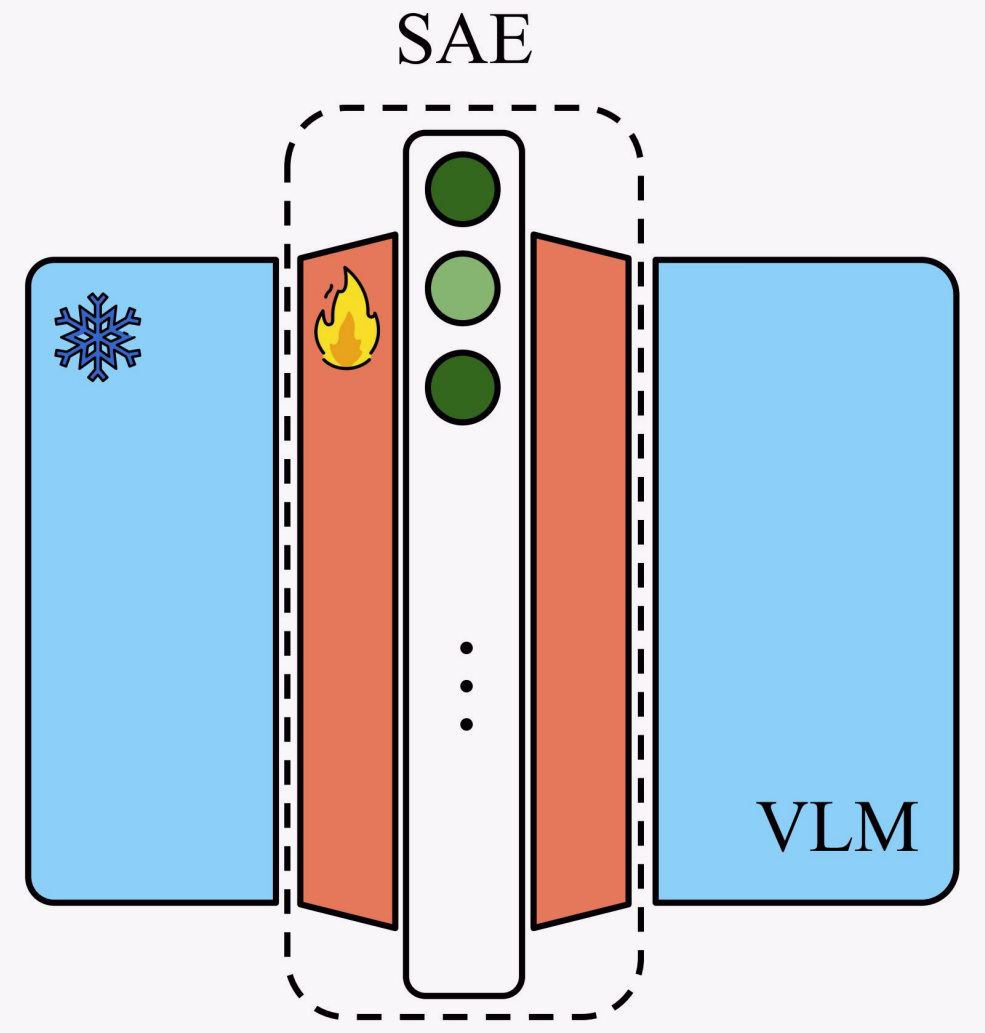
Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, Zeynep Akata

Can SAEs find disentangled representations in VLMs?

Using our new **Monosemanticity Score** we quantitatively find:

YES!

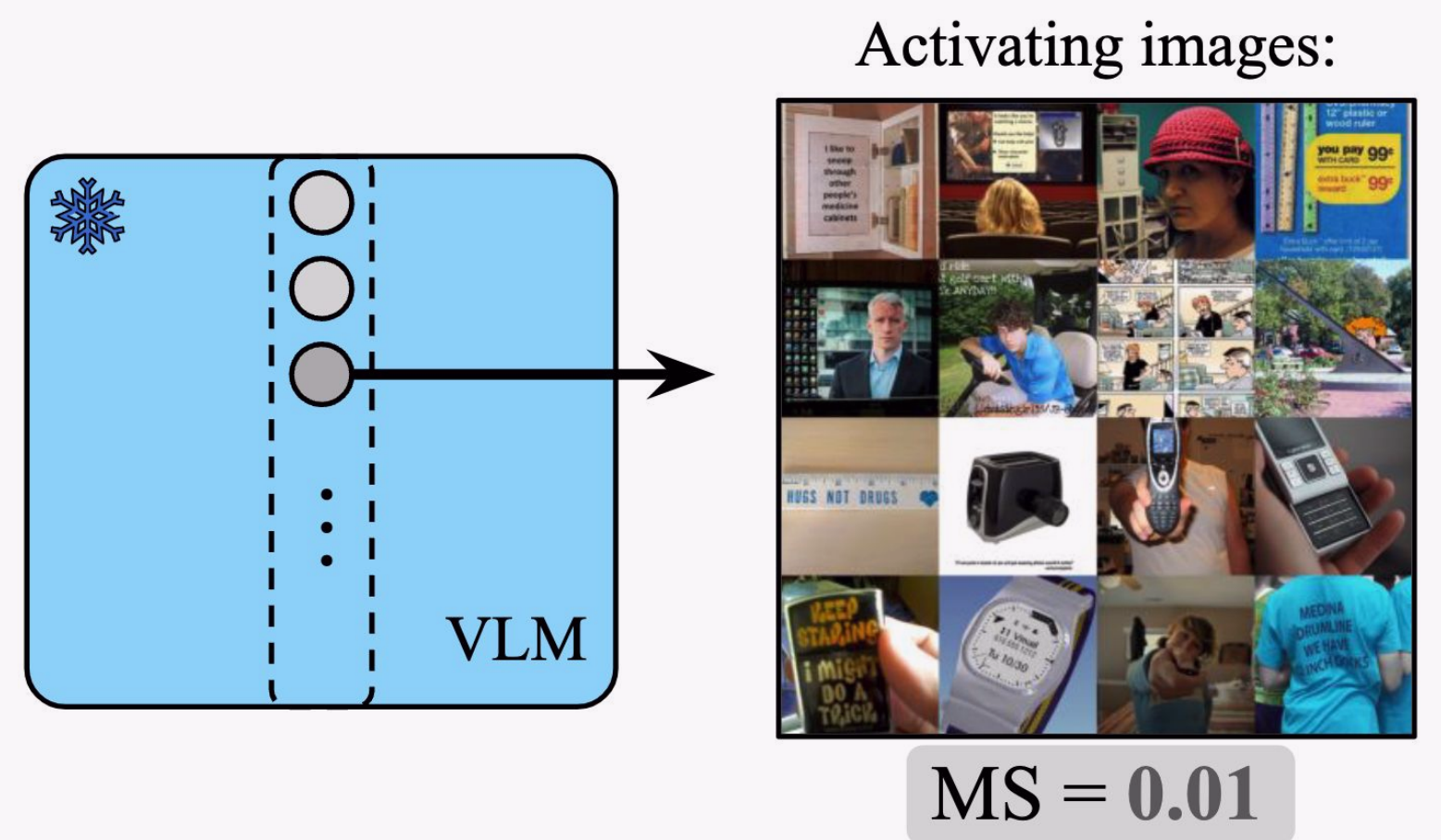
Sparse Autoencoder (SAE)

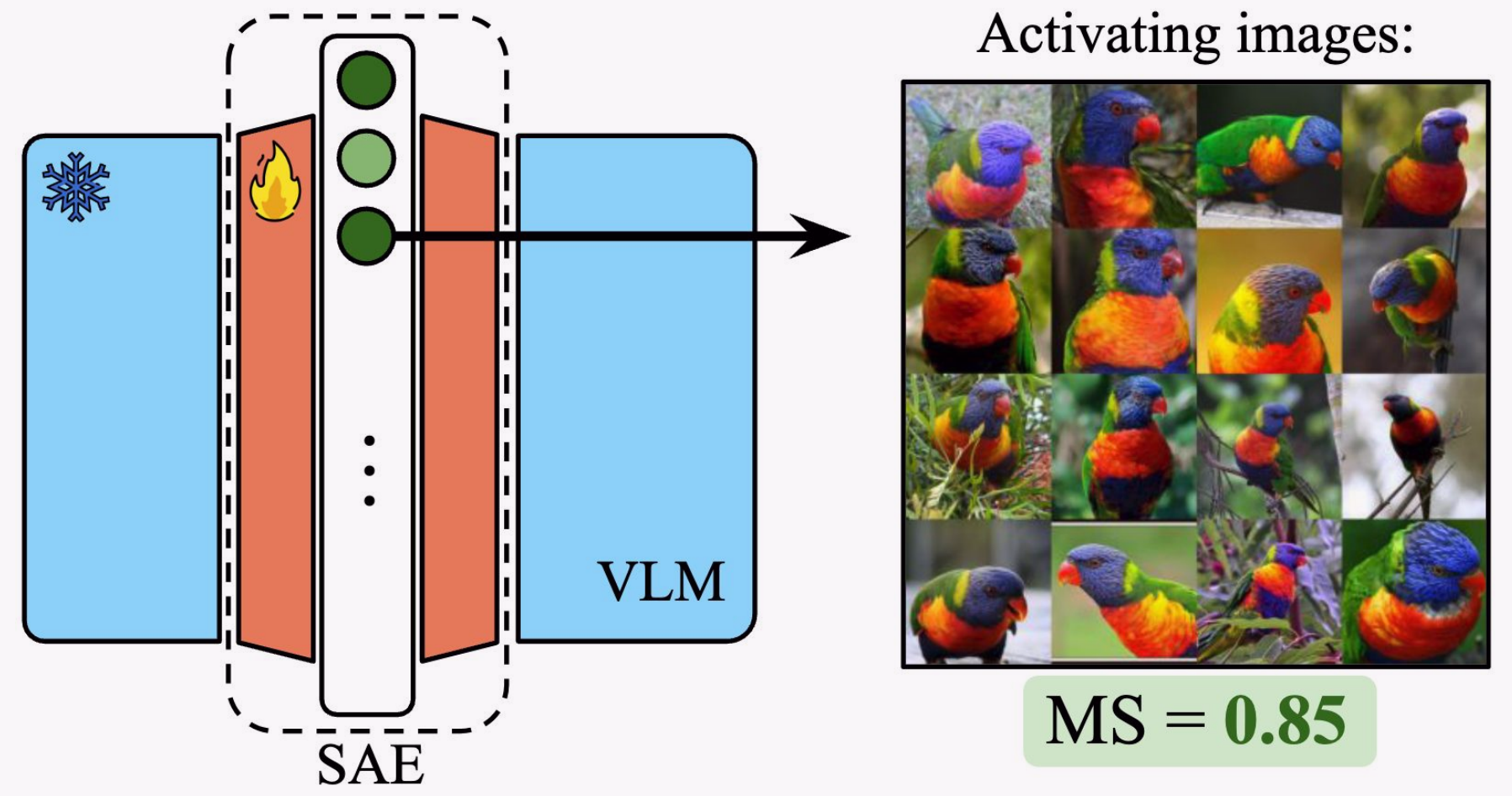


Reconstructs activations via sparsely activated high dimensional space

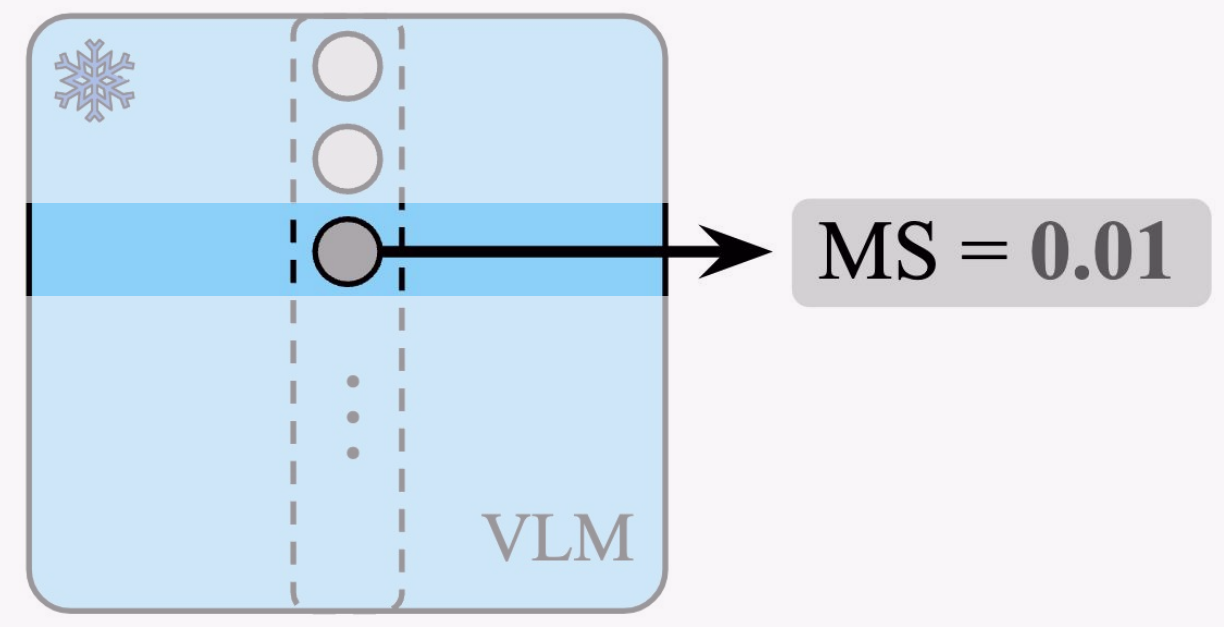
Expected to disentangle representations into human understandable concepts

Problem #1: Sparser = more monosemantic?



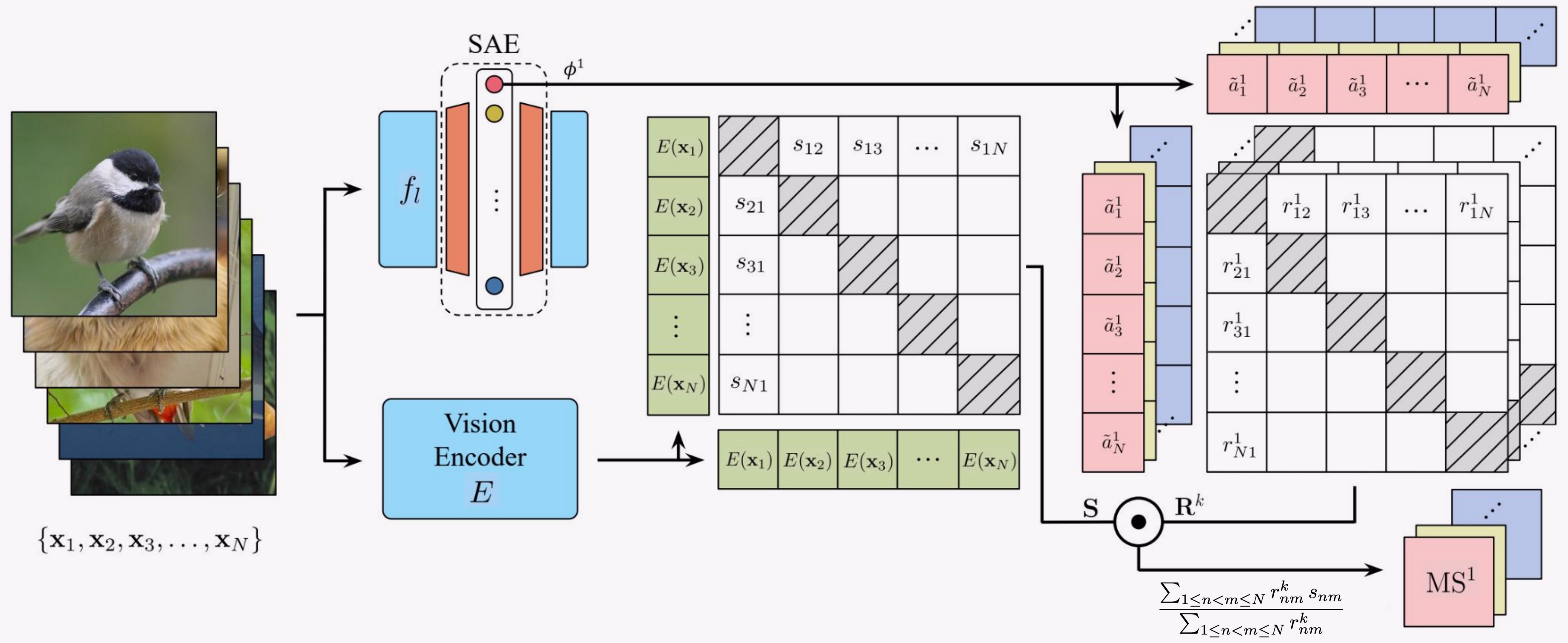


Problem #2: Lack of per neuron metric



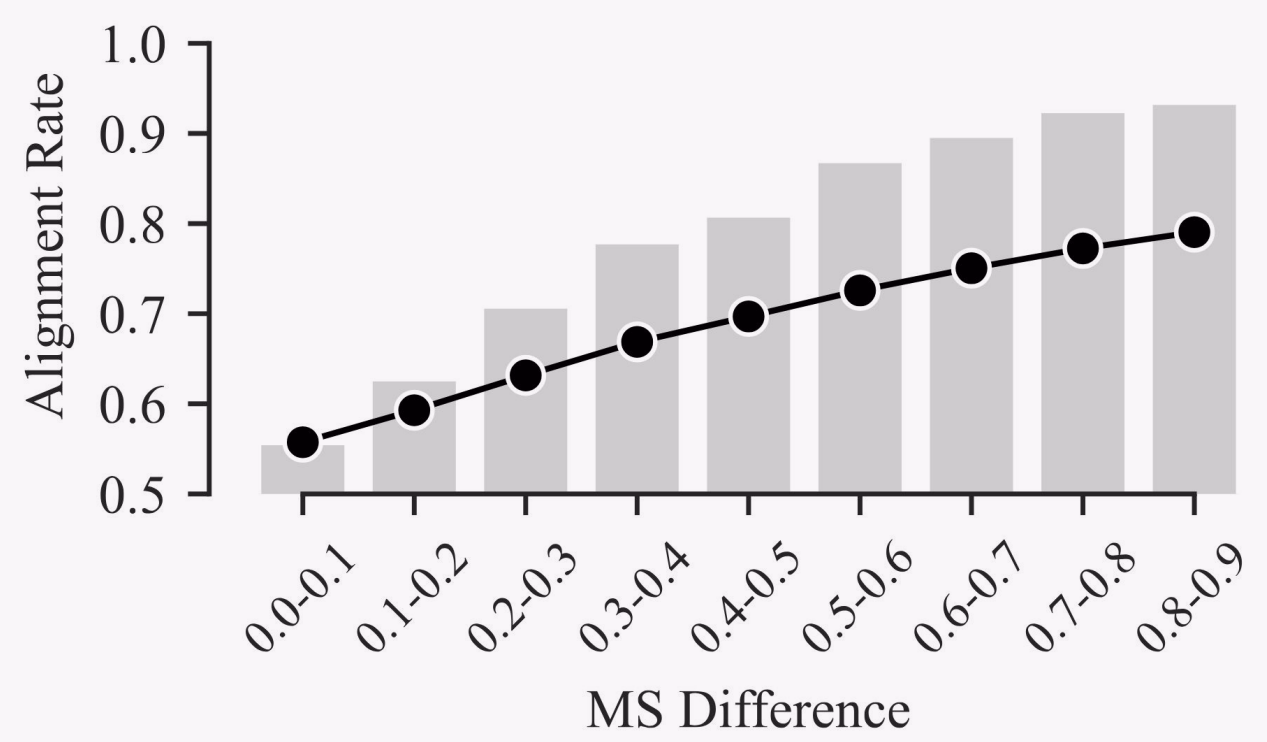
Solution: **Monosemanticity Score (MS)**

Computation of MS


SAE
ϕ^1
f_l
Vision Encoder E
$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_N\}$
$E(\mathbf{x}_1)$ $E(\mathbf{x}_2)$ $E(\mathbf{x}_3)$ ⋯ $E(\mathbf{x}_N)$
s_{12} s_{13} ⋯ s_{1N}
s_{21}
s_{31}
s_{N1}
$\tilde{a}^1_1$ $\tilde{a}^1_2$ $\tilde{a}^1_3$ ⋯ $\tilde{a}^1_N$
r^1_{12} r^1_{13} ⋯ r^1_{1N}
r^1_{21}
r^1_{31}
r^1_{N1}
$\mathbf{S}$ ⊙ $\mathbf{R}^k$
$\frac{\sum_{1\le n<m\le N} r^k_{nm} s_{nm}}{\sum_{1\le n<m\le N} r^k_{nm}}$
MS^1

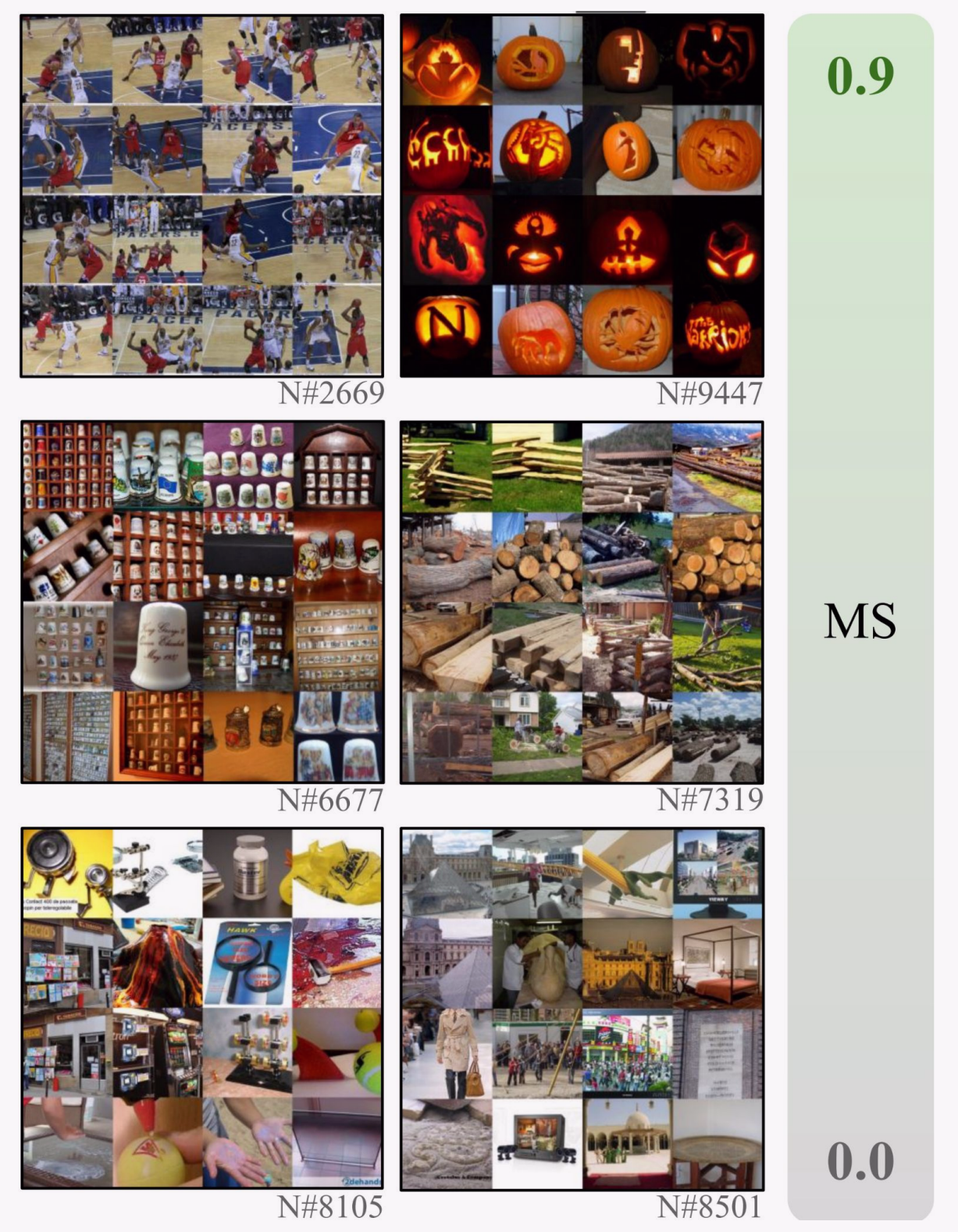
(A) Extracting embeddings and activations **(B)** Pairwise similarity matrix **(C)** Activation-weighted average

Human alignment



Percentage of times humans agreed with the MS on which neuron's top activating images look more focused, relative to the MS difference between the neurons

MS of sample neurons



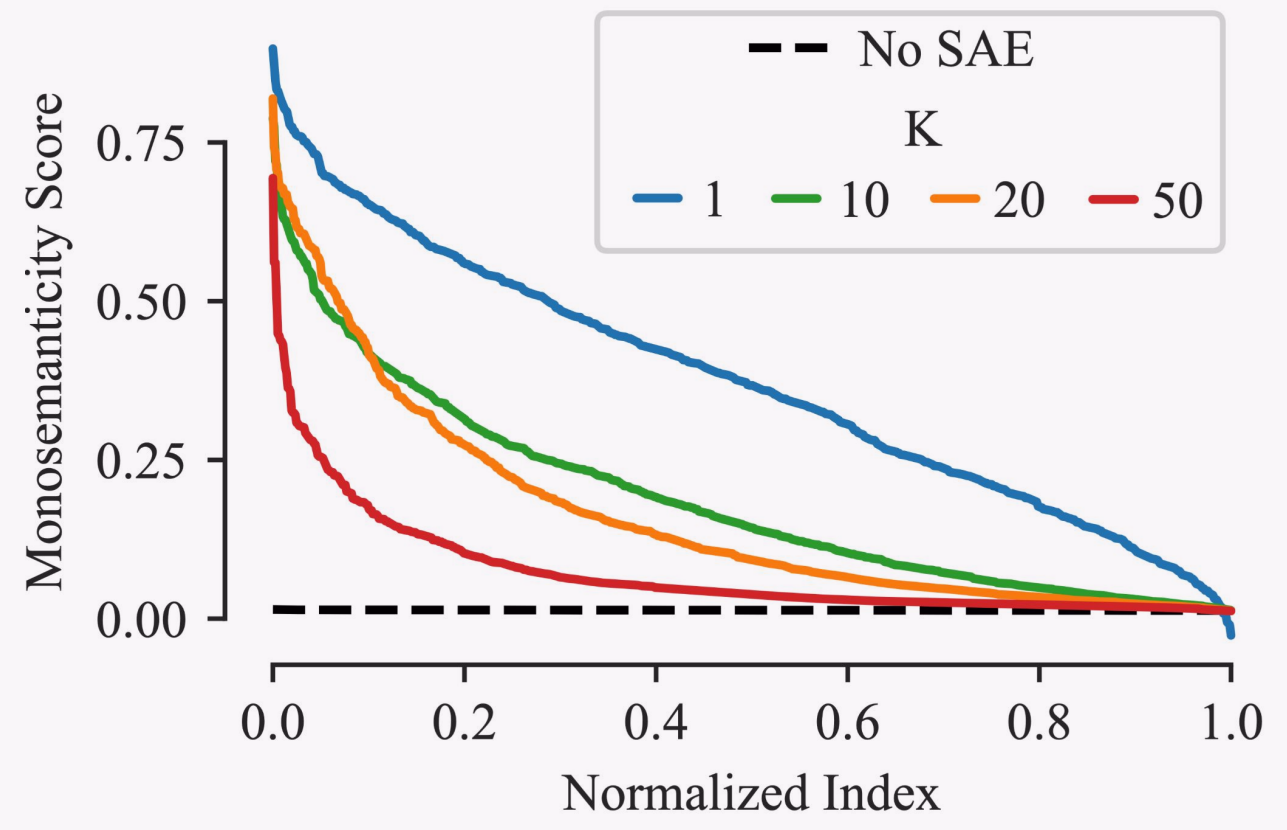
Analysis using **Monosemanticity Score**

- SAE gives more monosemantic neurons
- Sparsity drives monosemanticity
- Wider latent enables better reconstruction by increasing monosemantic neuron count

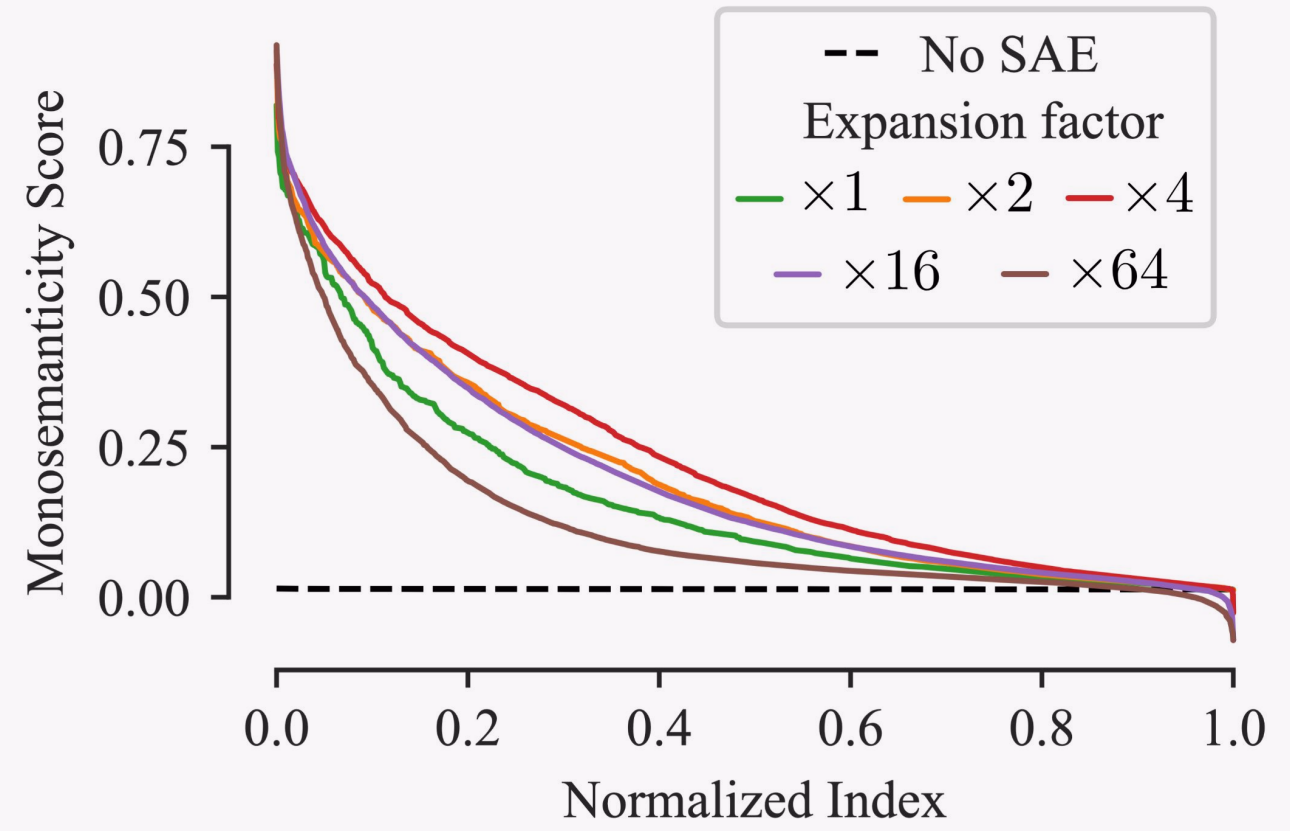
The highest MS observed across original ('No SAE') and SAE neurons

SAE type	Layer	No SAE	Expansion factor					
			×1	×2	×4	×8	×16	×64
BatchTopK [4]	11	0.01	0.61	0.73	0.71	0.87	0.90	1.00
	17	0.01	0.65	0.79	0.86	0.86	0.93	1.00
	22	0.01	0.66	0.79	0.80	0.88	0.92	1.00
	23	0.01	0.73	0.72	0.83	0.89	0.93	1.00
	last	0.01	0.57	0.78	0.78	0.81	0.85	1.00
Matryoshka [5, 27]	11	0.01	0.84	0.90	0.95	1.00	0.89	1.00
	17	0.01	0.86	0.84	0.93	0.94	0.96	1.00
	22	0.01	0.83	0.83	0.87	0.94	1.00	1.00
	23	0.01	0.82	0.84	0.89	0.93	0.96	1.00
	last	0.01	0.82	0.91	0.89	0.93	0.91	1.00

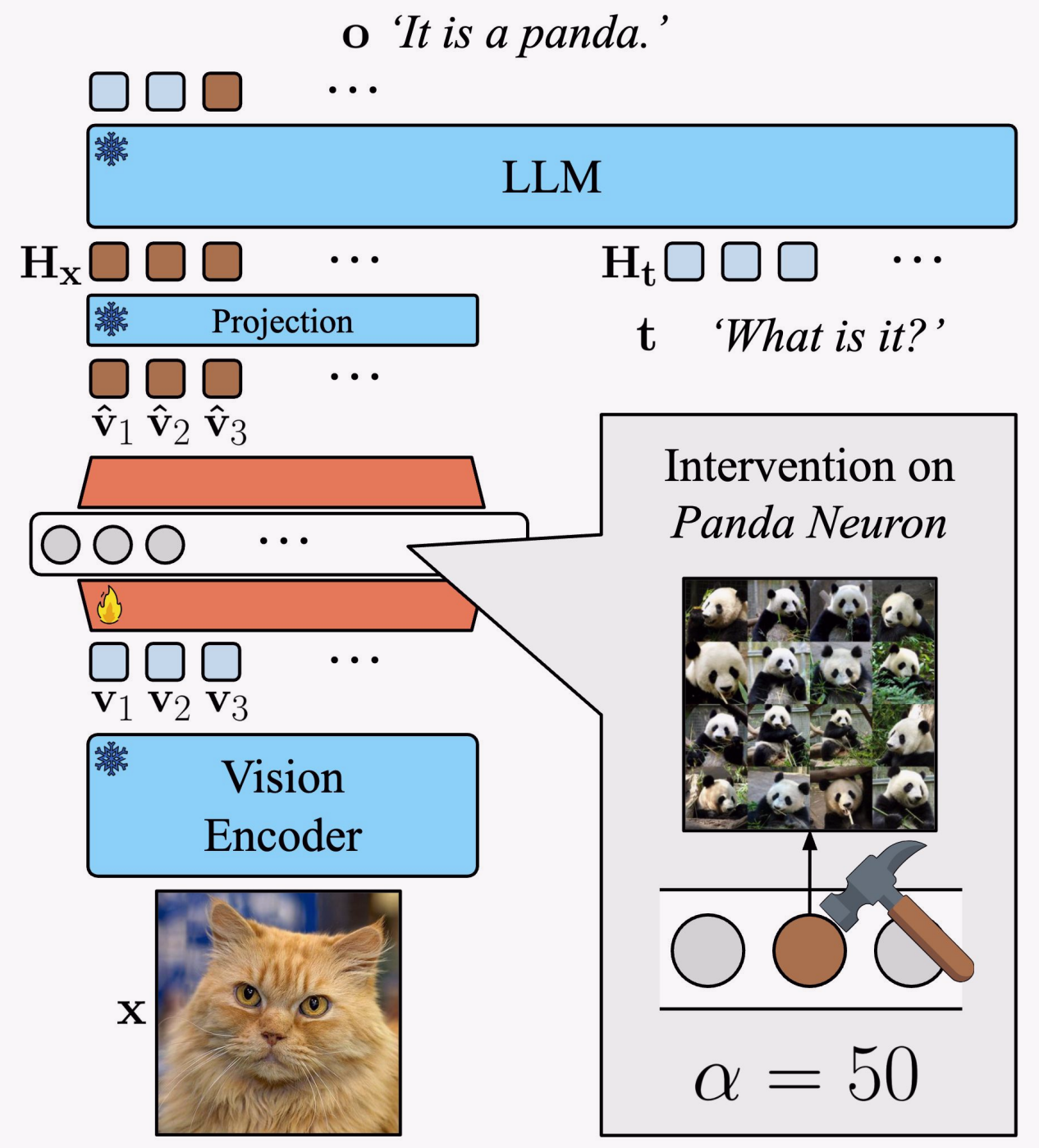
MS of neurons in SAEs trained with different sparsites K

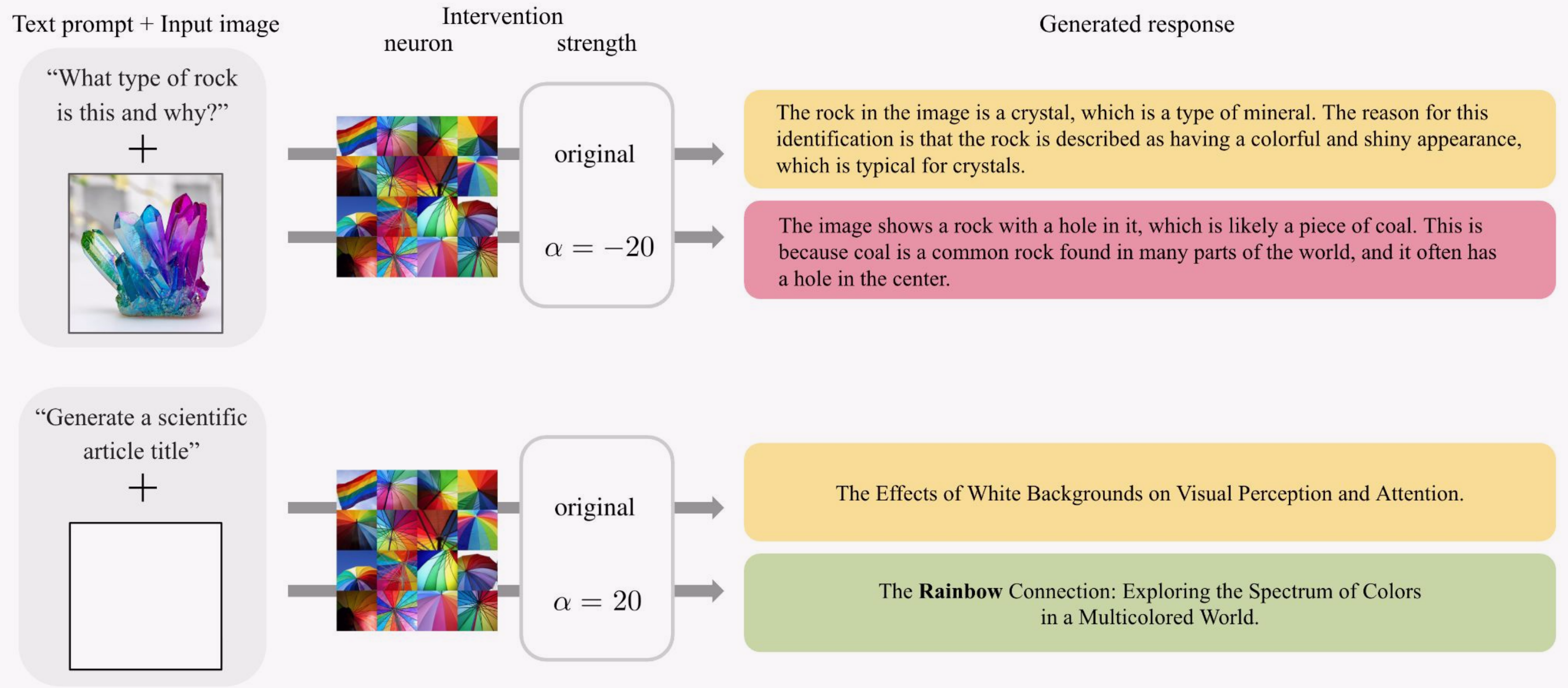


MS of neurons in SAEs trained with different expansion factors



Applying monosemanticity to steer LLaVA


o 'It is a panda.'
LLM
$\mathbf{H}_\mathbf{x}$
$\mathbf{H}_\mathbf{t}$
Projection
t 'What is it?'
$\hat{\mathbf{v}}_1$ $\hat{\mathbf{v}}_2$ $\hat{\mathbf{v}}_3$
$\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$
Vision Encoder
x
Intervention on Panda Neuron
$\alpha = 50$


Text prompt + Input image
Intervention
neuron
strength
Generated response
"What type of rock is this and why?"
+
original
$\alpha = -20$
"Generate a scientific article title"
+
original
$\alpha = 20$

The rock in the image is a crystal, which is a type of mineral. The reason for this identification is that the rock is described as having a colorful and shiny appearance, which is typical for crystals.

The image shows a rock with a hole in it, which is likely a piece of coal. This is because coal is a common rock found in many parts of the world, and it often has a hole in the center.

The Effects of White Backgrounds on Visual Perception and Attention.

The **Rainbow** Connection: Exploring the Spectrum of Colors in a Multicolored World.

(a) Concept insertion

	Ours	**DiffMean**
Desired concept appeared	48.7	**53.1**
Base prompt followed	**85.8**	66.2
Both criteria satisfied	**42.4**	35.8

(b) Concept suppression

	Ours	**DiffMean**
Desired concept removed	**64.4**	64.0
Unrelated concept kept	**81.4**	38.7
Both criteria satisfied	**52.5**	33.3