

1  
2  
3  
4  
5  
6  
  
7  
  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18

---

# Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models

## Appendix

---

### Contents

<b>A</b>	<b>Broader Impact</b>	<b>2</b>
<b>B</b>	<b>More details on steering</b>	<b>2</b>
<b>C</b>	<b>User study</b>	<b>3</b>
<b>D</b>	<b>Benchmark</b>	<b>8</b>
<b>E</b>	<b>Additional results on monosemanticity</b>	<b>8</b>
E.1	Unnormalized plots . . . . .	8
E.2	Detailed statistics and more models . . . . .	9
E.3	Matryoshka hierarchies . . . . .	12
<b>F</b>	<b>Reconstruction of SAEs</b>	<b>13</b>
<b>G</b>	<b>Uniqueness of concepts</b>	<b>14</b>
<b>H</b>	<b>Additional qualitative results</b>	<b>14</b>

## 19 A Broader Impact

20 Our work contributes to the field of interpretability and alignment, which are essential components  
 21 for building safe AI systems. Our MonoSemanticity score provides a new way to evaluate the  
 22 effectiveness of recently popular dictionary learning methods, such as sparse autoencoders (SAEs),  
 23 by incorporating human judgment into the evaluation process. This makes it easier to assess and  
 24 build trust in systems that use SAEs.

25 In addition, we show that SAEs can be highly effective in steering applications. They can be used to  
 26 encourage or discourage specific behaviors in models, or to help models recognize or ignore certain  
 27 concepts, including potentially dangerous ones. This is especially useful for ensuring that models  
 28 produce desired outputs and remain aligned with human values and goals.

## 29 B More details on steering

30 We illustrate in Figure A1 how we steer LLaVA-like models. We separately train SAEs on top of  
 31 the pretrained CLIP vision encoder to reconstruct the *token embeddings*  $\mathbf{v}_i$ , and then attach it back  
 32 after the vision encoder during inference. Intervening on a neuron within the SAE layer steers the  
 33 reconstructed tokens  $\hat{\mathbf{v}}_i$  towards the activated concept, which then steers the LLM’s generated output.  
 34 We present in Figure A2 additional examples of LLaVA prompted to generate scientific titles, and the  
 35 outputs before and after intervening on SAE neurons. Increasing the activation of specific neurons  
 36 will modify the outputs to include elements from images highly activating the corresponding neuron.

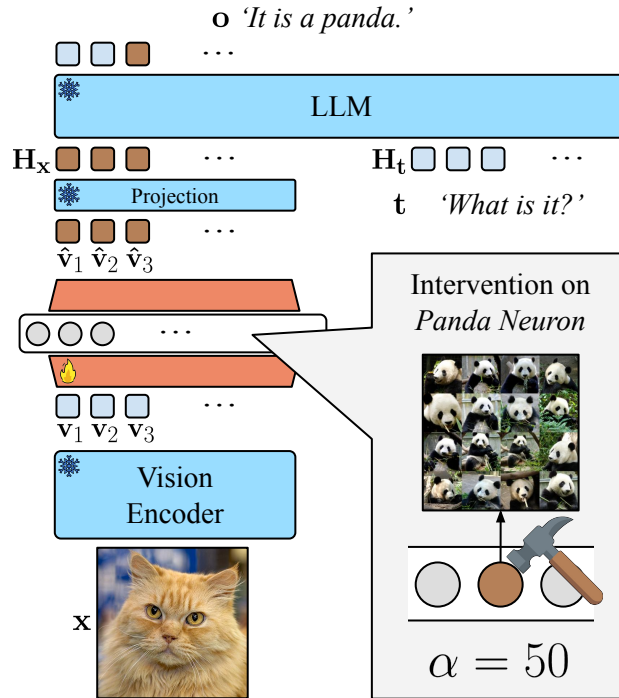


Figure A1: LLaVA-like models can be steered towards *seeing* a concept (e.g. *panda*) not present in the input image  $x$ . By attaching SAE after vision encoder and intervening on its neuron representing that concept, we effectively manipulate the LLM’s response. Such flexible and precise steering is possible thanks to the extensive concept dictionary identified through the SAE.

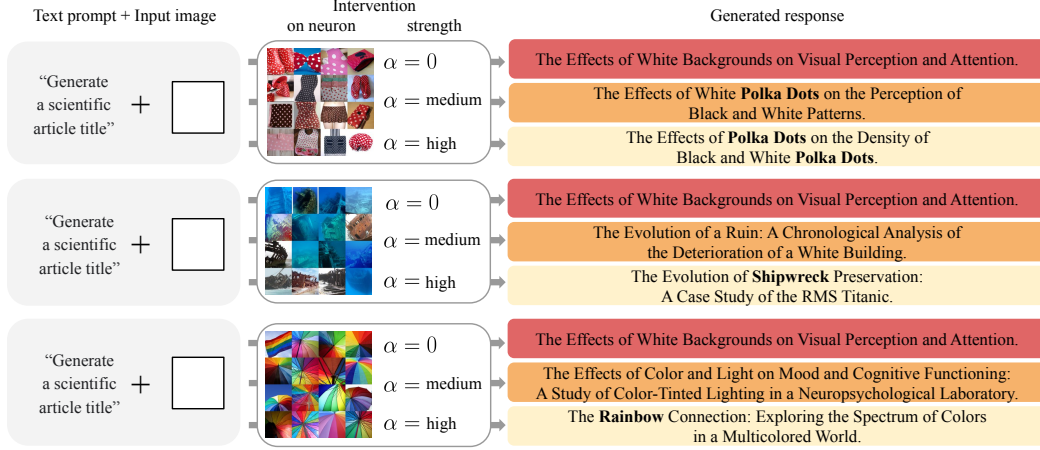


Figure A2: Effects of neuron interventions on MLLM-generated scientific article titles. Steering magnitudes are categorized as “0”, “medium”, and “high” based on the intervention strength. The neurons are visualized with the highest activating images from which we deduce their associated concepts: “polka dots”, “shipwreck”, and “rainbow”.

## 37 C User study

38 To validate the alignment of our MonoSemanticity score (MS) with human judgment, we conducted  
39 a user study. Example questions are shown in Figures A3, A4, and A5. Each question shows two  
40 grids of images:

$$((x_i)_{i=1}^{16}, (y_i)_{i=1}^{16}),$$

41 where each grid contains the 16 images with the highest activations for two neurons  $k_x$  and  $k_y$ ,  
42 respectively. Formally, for each  $i = 1, \dots, 16$ ,

$$x_i := \mathbf{x}_n \in \mathcal{I}, \quad \text{rank}_n(a_n^{k_x}) = i,$$

$$y_i := \mathbf{x}_m \in \mathcal{I}, \quad \text{rank}_m(a_m^{k_y}) = i,$$

44 where  $a_n^k$  is the activation of neuron  $k$  on image  $\mathbf{x}_n$ , and  $\text{rank}_n(a_n^k)$  is the rank of image  $\mathbf{x}_n$  when  
45 sorting all images by their activation values in descending order. Images  $\mathcal{I}$  come from training set of  
46 the ImageNet.

47 For each neuron pair  $(k_x, k_y)$ , we asked three human annotators the question: “Which set of images  
48 looks more similar and focused on the same thing?” Each annotator gave an answer  $r_j \in \{k_x, k_y\}$   
49 for  $j = 1, 2, 3$ . The final human choice was decided by majority vote:

$$R_{(k_x, k_y)}^{\text{user}} \in \{k_x, k_y\}.$$

50 At the same time, we answered the question using Monosemanticity Score:

$$R_{(k_x, k_y)}^{\text{MS}} := \begin{cases} k_x & \text{if } \text{MS}^{k_x} > \text{MS}^{k_y} \\ k_y & \text{otherwise} \end{cases}$$

51 We say the MS and users are *aligned* if their decision is the same:

$$\delta_{(k_x, k_y)} := \begin{cases} 1 & \text{if } R_{(k_x, k_y)}^{\text{user}} = R_{(k_x, k_y)}^{\text{MS}} \\ 0 & \text{otherwise} \end{cases}$$

52 The overall *alignment score* is the fraction of all neuron pairs where the MS and humans are aligned:

$$\text{Alignment Score} = \frac{1}{|\mathcal{Q}|} \sum_{(k_x, k_y) \in \mathcal{Q}} \delta_{(k_x, k_y)},$$

53 where  $\mathcal{Q}$  is the set of all neuron pairs evaluated.

Table A1: Alignment Scores (AS) obtained from user study. To compute the MS, we use embeddings of image encoder  $E$ , either DINOv2 ViT-B or CLIP ViT-B. Results are grouped by MS distance between neurons in the question. We made sure that every group is represented by enough pairs.

(a) MS distances computed using DinoV2 embeddings.

MS Distance (based on DinoV2)	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9
Number of pairs	126	134	116	125	116	114	122	84	63
AS ( $E = \text{DINOv2 ViT-B}$ )	0.56	0.66	0.71	0.81	0.85	0.93	0.94	0.96	1.00
AS ( $E = \text{CLIP ViT-B}$ )	0.60	0.66	0.74	0.82	0.87	0.87	0.94	0.96	1.00

(b) MS distances computed using CLIP embeddings.

MS Distance (based on CLIP)	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5
Number of pairs	231	224	215	197	132
AS ( $E = \text{CLIP ViT-B}$ )	0.55	0.81	0.93	0.96	0.93
AS ( $E = \text{DINOv2 ViT-B}$ )	0.53	0.77	0.92	0.96	0.93

54 In total, we collected 1,000 user pair rankings with the help of 71 annotators on the Mechanical  
55 Turk platform. The number of answers per annotator ranged from 1 to 205, with a median of 24.  
56 Annotators were compensated at a rate of \$0.02 per answer.

57 The neurons used in the study were randomly selected from the last layer of CLIP ViT-L, BatchTopK  
58 SAE ( $\varepsilon = 4, K = 20$ ) trained on the last layer of CLIP ViT-L, Matryoshka SAE ( $\varepsilon = 4, K = 20$ )  
59 trained on the last layer of CLIP ViT-L, and BatchTopK SAE ( $\varepsilon = 4, K = 20$ ) trained on the last  
60 layer of SigLIP SoViT-400m.

61 In addition to the plot presenting the user study results in the main paper, we also provide Table A1,  
62 which reports the exact values obtained, along with the sizes of each group categorized by MS  
63 distances between neuron pairs. When designing the questions, we balanced the number of pairs  
64 within each distance interval. Our goal is to evaluate MS computed using embeddings from two  
65 different image encoders  $E$ , namely DINOv2 ViT-B and CLIP ViT-B. As a result, the group sizes are  
66 not perfectly equal due to necessary trade-offs. Nevertheless, all groups are sufficiently large and of  
67 comparable size.



Click to show/hide instructions

You will be presented with a pair of image sets, each represented as a grid of 16 images. Your task is to decide which set contains images that look more similar and focused on the same thing.

Example:

Set [A]

Set [B]

Which set of images looks more alike and focused on the same thing?

☒ Set [A]

☐ Set [B]

Explanation behind answer: Set [A].

Set [A] contains similar images of uniformed people standing in formation.

Set [B] contains varied images such as a hoodie, baseball, or dog.

Set [A]

Set [B]

☐ Set [A]

☐ Set [B]

Submit

Which set of images looks more alike and focused on the same thing?

☐ Set [A]

☐ Set [B]

Submit

Figure A3: Example question used in the user study. Best viewed horizontally.

Click to show/hide instructions

You will be presented with a pair of image sets, each represented as a grid of 16 images. Your task is to decide which set contains images that look more similar and focused on the same thing.

Example:

Set [A]

Set [B]

Which set of images looks more alike and focused on the same thing?

☒ Set [A]

☐ Set [B]

Explanation behind answer: Set [A].

Set [A] contains similar images of uniformed people standing in formation.

Set [B] contains varied images such as a hoodie, baseball, or dog.

Set [A]

Set [B]

☐ Set [A]

☐ Set [B]

Submit

Figure A4: Example question used in the user study. Best viewed horizontally.

6

Click to show/hide instructions

You will be presented with a pair of image sets, each represented as a grid of 16 images. Your task is to decide which set contains images that look more similar and focused on the same thing.

Example:

Set [A]



Set [B]



Which set of images looks more alike and focused on the same thing?

☒ Set [A]

Explanation behind answer: Set [A].

☐ Set [B]

Set [A] contains similar images of uniformed people standing in formation.

☐ Set [A]

Set [B] contains varied images such as a hoodie, baseball, or dog.

Set [A]



Set [B]



Which set of images looks more alike and focused on the same thing?

☐ Set [A]

Set [B]

☐ Set [A]

Set [B]

Submit

Figure A5: Example question used in the user study. Best viewed horizontally.

## 68 D Benchmark

69 While MS shows very good results in our user study, we anticipate the development of improved  
 70 alternatives in the future. To facilitate such advancements, we will release our collected data as a  
 71 benchmark for evaluating neuron monosemanticity.

72 The benchmark will include the following files:

- 73 • `pairs.csv` – Contains 1000 pairs of neurons  $(r_x, r_y)$ , along with user preferences  $R_{(k_x, k_y)}^{\text{user}}$   
 74 and MS values computed using two different image encoders: DINOv2 ViT-B and CLIP ViT-  
 75 B. Each row includes the following columns: `k_x`, `k_y`, `R_user`, `MS_x_dino`, `MS_y_dino`,  
 76 `MS_x_clip`, `MS_y_clip`.
- 77 • `top16_images.csv` – Lists the 16 most activating images from the ImageNet training set  
 78 for each neuron used in the study. Columns: `k`, `x_1`, ..., `x_16`.
- 79 • `activations.csv` – Provides activation values of all 50,000 ImageNet validation images  
 80 for each neuron. Columns: `k`, `a_1`, ..., `a_50000`.

81 With this data and by following our evaluation procedure, researchers will be able to compare their  
 82 methods directly to MS under same conditions. They will have access to the same underlying  
 83 information, specifically the complete set of neuron activations on the ImageNet validation set.

## 84 E Additional results on monosemanticity

### 85 E.1 Unnormalized plots

86 Monosemanticity scores across all neurons, without normalized index, are shown in Figure A6. We  
 87 observe that neurons cover a wider range of scores as we increase the width of the SAE layer.  
 88 Furthermore, for a given threshold of monosemanticity, the number of neurons having a score higher  
 than this threshold is also increasing with the width.

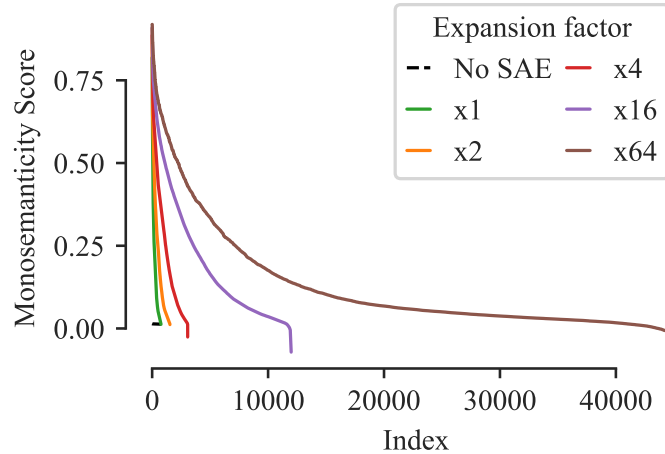


Figure A6: MS in decreasing order across neurons. Results are shown for a layer without SAE (“No SAE”), and with SAE using different expansion factors ( $\times 1$ ,  $\times 2$ ,  $\times 4$ ,  $\times 16$  and  $\times 64$ ).

89

## 90 E.2 Detailed statistics and more models

91 We report in Tables A2 and A5, the average ( $\pm$  std), best and worst monosemanticity scores across  
 92 neurons for the two SAE variants, attached at different layers and for increasing expansion factors.  
 93 Although average scores remain similar when increasing expansion factor, we observe a high increase  
 94 between the original layer and an SAE with expansion factor  $\varepsilon = 1$ . The best scores get consistently  
 95 better as expansion factor gets increased.

96 Until now, our analysis has focused on SAEs trained on CLIP ViT-L activations, evaluated using  
 97 the MS score computed from embeddings produced by the DINOv2 image encoder  $E$ . To broaden  
 98 this investigation, we now consider SAEs trained on activations from SigLIP SoViT-400m. As an  
 99 alternative image encoder  $E$ , we adopt CLIP ViT-B for evaluation.

100 Tables A3 and A6 show average, best and worst MS computed using CLIP ViT-B as the vision  
 101 encoder  $E$ . Even though less distinctively than in original setup, the neurons from SAEs still score  
 102 better compared to the ones originally found in the model.

103 In Tables A4 and A7, we report MS statistics for SAEs trained for SigLIP SoViT-400m model  
 104 computed using CLIP ViT-B as the vision encoder  $E$ . The results highly resemble the ones for CLIP  
 105 ViT-L model.

Table A2: The average MS of neurons in a CLIP ViT-L model. DINOv2 ViT-B is used as the image encoder  $E$ .

SAE type	Layer	No SAE	Expansion factor					
			x1	x2	x4	x8	x16	x64
BatchTopK	11	0.0135 $\pm$ 0.0003	0.03 $\pm$ 0.06	0.04 $\pm$ 0.06	0.04 $\pm$ 0.06	0.03 $\pm$ 0.05	0.03 $\pm$ 0.05	0.03 $\pm$ 0.05
	17	0.0135 $\pm$ 0.0004	0.05 $\pm$ 0.07	0.07 $\pm$ 0.09	0.08 $\pm$ 0.11	0.07 $\pm$ 0.10	0.07 $\pm$ 0.10	0.06 $\pm$ 0.10
	22	0.0135 $\pm$ 0.0003	0.14 $\pm$ 0.12	0.18 $\pm$ 0.15	0.20 $\pm$ 0.17	0.21 $\pm$ 0.17	0.21 $\pm$ 0.18	0.17 $\pm$ 0.18
	23	0.0135 $\pm$ 0.0003	0.15 $\pm$ 0.13	0.18 $\pm$ 0.16	0.20 $\pm$ 0.17	0.21 $\pm$ 0.17	0.20 $\pm$ 0.18	0.17 $\pm$ 0.18
	last	0.0135 $\pm$ 0.0002	0.12 $\pm$ 0.11	0.17 $\pm$ 0.15	0.19 $\pm$ 0.17	0.19 $\pm$ 0.16	0.16 $\pm$ 0.16	0.13 $\pm$ 0.15
Matryoshka	11	0.0135 $\pm$ 0.0003	0.05 $\pm$ 0.10	0.06 $\pm$ 0.10	0.05 $\pm$ 0.09	0.05 $\pm$ 0.09	0.04 $\pm$ 0.08	0.03 $\pm$ 0.06
	17	0.0135 $\pm$ 0.0004	0.09 $\pm$ 0.14	0.10 $\pm$ 0.15	0.11 $\pm$ 0.16	0.11 $\pm$ 0.15	0.10 $\pm$ 0.15	0.06 $\pm$ 0.10
	22	0.0135 $\pm$ 0.0003	0.17 $\pm$ 0.17	0.21 $\pm$ 0.18	0.23 $\pm$ 0.19	0.23 $\pm$ 0.19	0.23 $\pm$ 0.19	0.18 $\pm$ 0.19
	23	0.0135 $\pm$ 0.0003	0.17 $\pm$ 0.16	0.21 $\pm$ 0.19	0.22 $\pm$ 0.18	0.22 $\pm$ 0.18	0.20 $\pm$ 0.18	0.12 $\pm$ 0.16
	last	0.0135 $\pm$ 0.0002	0.16 $\pm$ 0.17	0.20 $\pm$ 0.18	0.23 $\pm$ 0.19	0.22 $\pm$ 0.19	0.19 $\pm$ 0.19	0.13 $\pm$ 0.16

Table A3: The average MS of neurons in a CLIP ViT-L model. CLIP ViT-B is used as the image encoder  $E$ .

SAE type	Layer	No SAE	Expansion factor					
			x1	x2	x4	x8	x16	x64
BatchTopK	11	0.4837 $\pm$ 0.0067	0.52 $\pm$ 0.05	0.53 $\pm$ 0.06	0.53 $\pm$ 0.05	0.53 $\pm$ 0.05	0.53 $\pm$ 0.05	0.53 $\pm$ 0.06
	17	0.4840 $\pm$ 0.0079	0.55 $\pm$ 0.07	0.56 $\pm$ 0.08	0.57 $\pm$ 0.08	0.56 $\pm$ 0.05	0.56 $\pm$ 0.08	0.56 $\pm$ 0.09
	22	0.4816 $\pm$ 0.0053	0.60 $\pm$ 0.09	0.61 $\pm$ 0.09	0.62 $\pm$ 0.09	0.63 $\pm$ 0.09	0.62 $\pm$ 0.10	0.60 $\pm$ 0.11
	23	0.4814 $\pm$ 0.0045	0.60 $\pm$ 0.09	0.61 $\pm$ 0.10	0.62 $\pm$ 0.10	0.62 $\pm$ 0.10	0.61 $\pm$ 0.10	0.59 $\pm$ 0.12
	last	0.4812 $\pm$ 0.0042	0.59 $\pm$ 0.08	0.60 $\pm$ 0.10	0.61 $\pm$ 0.10	0.61 $\pm$ 0.10	0.59 $\pm$ 0.10	0.56 $\pm$ 0.10
Matryoshka	11	0.4837 $\pm$ 0.0067	0.54 $\pm$ 0.08	0.55 $\pm$ 0.08	0.55 $\pm$ 0.08	0.54 $\pm$ 0.08	0.53 $\pm$ 0.07	0.52 $\pm$ 0.06
	17	0.4840 $\pm$ 0.0079	0.57 $\pm$ 0.09	0.58 $\pm$ 0.09	0.58 $\pm$ 0.10	0.58 $\pm$ 0.10	0.57 $\pm$ 0.10	0.54 $\pm$ 0.09
	22	0.4816 $\pm$ 0.0053	0.61 $\pm$ 0.09	0.62 $\pm$ 0.09	0.63 $\pm$ 0.10	0.62 $\pm$ 0.11	0.62 $\pm$ 0.11	0.59 $\pm$ 0.12
	23	0.4814 $\pm$ 0.0045	0.60 $\pm$ 0.09	0.62 $\pm$ 0.10	0.62 $\pm$ 0.10	0.61 $\pm$ 0.11	0.60 $\pm$ 0.11	0.54 $\pm$ 0.11
	last	0.4812 $\pm$ 0.0042	0.59 $\pm$ 0.09	0.61 $\pm$ 0.10	0.62 $\pm$ 0.11	0.61 $\pm$ 0.11	0.59 $\pm$ 0.12	0.54 $\pm$ 0.12

Table A4: The average MS of neurons in a SigLIP SoViT-400m model. CLIP ViT-B is used as the image encoder  $E$ .

SAE type	Layer	No SAE	Expansion factor					
			x1	x2	x4	x8	x16	x64
BatchTopK	11	0.4805 $\pm$ 0.0014	0.50 $\pm$ 0.03	0.51 $\pm$ 0.04	0.51 $\pm$ 0.05	0.51 $\pm$ 0.06	0.52 $\pm$ 0.06	0.52 $\pm$ 0.07
	16	0.4809 $\pm$ 0.0024	0.51 $\pm$ 0.04	0.52 $\pm$ 0.05	0.52 $\pm$ 0.06	0.53 $\pm$ 0.07	0.53 $\pm$ 0.07	0.53 $\pm$ 0.08
	21	0.4810 $\pm$ 0.0052	0.52 $\pm$ 0.05	0.53 $\pm$ 0.06	0.53 $\pm$ 0.06	0.53 $\pm$ 0.07	0.54 $\pm$ 0.08	0.53 $\pm$ 0.08
	last	0.4811 $\pm$ 0.0048	0.61 $\pm$ 0.09	0.61 $\pm$ 0.09	0.62 $\pm$ 0.09	0.62 $\pm$ 0.09	0.62 $\pm$ 0.10	0.60 $\pm$ 0.11
Matryoshka	11	0.4805 $\pm$ 0.0014	0.50 $\pm$ 0.03	0.50 $\pm$ 0.05	0.50 $\pm$ 0.05	0.50 $\pm$ 0.06	0.51 $\pm$ 0.07	0.51 $\pm$ 0.07
	16	0.4809 $\pm$ 0.0024	0.51 $\pm$ 0.05	0.52 $\pm$ 0.06	0.52 $\pm$ 0.07	0.52 $\pm$ 0.07	0.52 $\pm$ 0.07	0.51 $\pm$ 0.07
	21	0.4810 $\pm$ 0.0052	0.52 $\pm$ 0.05	0.53 $\pm$ 0.06	0.53 $\pm$ 0.06	0.53 $\pm$ 0.07	0.52 $\pm$ 0.07	0.51 $\pm$ 0.07
	last	0.4811 $\pm$ 0.0048	0.61 $\pm$ 0.09	0.62 $\pm$ 0.10	0.62 $\pm$ 0.10	0.62 $\pm$ 0.10	0.60 $\pm$ 0.11	0.58 $\pm$ 0.11

Table A5: Comparison of the best / worst MS of neurons in a CLIP ViT-L model. DINOv2 ViT-B is used as the image encoder  $E$ .

SAE type	Layer	No SAE	Expansion factor					
			$\times 1$	$\times 2$	$\times 4$	$\times 8$	$\times 16$	$\times 64$
BatchTopK	11	0.01 / 0.01	0.61 / -0.02	0.73 / -0.08	0.71 / -0.06	0.87 / -0.07	0.90 / -0.10	1.00 / -0.11
	17	0.01 / 0.01	0.65 / 0.01	0.79 / -0.02	0.86 / -0.07	0.86 / -0.08	0.93 / -0.08	1.00 / -0.12
	22	0.01 / 0.01	0.66 / 0.01	0.79 / 0.01	0.80 / 0.01	0.88 / -0.08	0.92 / -0.06	1.00 / -0.11
	23	0.01 / 0.01	0.73 / 0.01	0.72 / 0.01	0.83 / 0.01	0.89 / -0.02	0.93 / -0.06	1.00 / -0.10
	last	0.01 / 0.01	0.57 / 0.01	0.78 / 0.01	0.78 / 0.01	0.81 / -0.01	0.85 / -0.04	1.00 / -0.10
Matryoshka	11	0.01 / 0.01	0.84 / -0.06	0.90 / -0.07	0.95 / -0.08	1.00 / -0.11	0.89 / -0.10	1.00 / -0.10
	17	0.01 / 0.01	0.86 / -0.04	0.84 / -0.05	0.93 / -0.07	0.94 / -0.09	0.96 / -0.08	1.00 / -0.14
	22	0.01 / 0.01	0.83 / 0.01	0.83 / 0.01	0.87 / -0.02	0.94 / -0.06	1.00 / -0.11	1.00 / -0.11
	23	0.01 / 0.01	0.82 / 0.01	0.84 / 0.01	0.89 / -0.04	0.93 / -0.04	0.96 / -0.06	1.00 / -0.11
	last	0.01 / 0.01	0.82 / 0.01	0.91 / 0.01	0.89 / -0.03	0.93 / -0.05	0.91 / -0.07	1.00 / -0.12

Table A6: Comparison of the best / worst MS of neurons in a CLIP ViT-Large model. CLIP ViT-B is used as the image encoder  $E$ .

SAE type	Layer	No SAE	Expansion factor					
			$\times 1$	$\times 2$	$\times 4$	$\times 8$	$\times 16$	$\times 64$
BatchTopK	11	0.50 / 0.47	0.80 / 0.41	0.87 / 0.38	0.90 / 0.28	0.91 / 0.27	0.95 / 0.24	1.00 / 0.20
	17	0.50 / 0.47	0.84 / 0.37	0.87 / 0.33	0.94 / 0.35	0.94 / 0.28	0.96 / 0.24	1.00 / 0.14
	22	0.50 / 0.47	0.82 / 0.39	0.85 / 0.38	0.89 / 0.37	0.93 / 0.29	0.93 / 0.15	1.00 / 0.15
	23	0.50 / 0.47	0.81 / 0.41	0.84 / 0.40	0.89 / 0.35	0.91 / 0.27	0.93 / 0.24	1.00 / 0.08
	last	0.50 / 0.47	0.80 / 0.40	0.84 / 0.40	0.87 / 0.36	0.87 / 0.31	0.89 / 0.25	1.00 / 0.17
Matryoshka	11	0.50 / 0.47	0.90 / 0.39	0.95 / 0.31	0.97 / 0.23	1.00 / 0.22	0.94 / 0.18	1.00 / 0.19
	17	0.50 / 0.47	0.94 / 0.33	0.93 / 0.35	0.96 / 0.29	0.96 / 0.22	0.97 / 0.14	1.00 / 0.11
	22	0.50 / 0.47	0.88 / 0.40	0.87 / 0.33	0.89 / 0.29	0.94 / 0.23	1.00 / 0.15	1.00 / 0.06
	23	0.50 / 0.47	0.85 / 0.40	0.86 / 0.35	0.90 / 0.35	0.91 / 0.19	0.93 / 0.17	1.00 / 0.14
	last	0.50 / 0.47	0.85 / 0.41	0.88 / 0.40	0.89 / 0.31	0.91 / 0.26	0.92 / 0.17	1.00 / 0.09

Table A7: Comparison of the best / worst MS of neurons in a SigLIP SoViT-400m model. CLIP ViT-B is used as the image encoder  $E$ .

SAE type	Layer	No SAE	Expansion factor					
			$\times 1$	$\times 2$	$\times 4$	$\times 8$	$\times 16$	$\times 64$
BatchTopK	11	0.49 / 0.48	0.61 / 0.41	0.83 / 0.29	0.88 / 0.27	0.90 / 0.23	1.00 / 0.12	1.00 / 0.15
	16	0.53 / 0.47	0.74 / 0.38	0.75 / 0.34	0.93 / 0.25	0.94 / 0.20	0.93 / 0.22	1.00 / 0.18
	21	0.54 / 0.47	0.76 / 0.38	0.77 / 0.35	0.83 / 0.25	0.89 / 0.17	0.95 / 0.20	1.00 / 0.11
	last	0.50 / 0.47	0.83 / 0.41	0.86 / 0.40	0.88 / 0.37	0.92 / 0.33	0.93 / 0.20	1.00 / 0.11
Matryoshka	11	0.49 / 0.48	0.70 / 0.40	0.93 / 0.29	0.77 / 0.27	0.93 / 0.18	0.91 / 0.22	1.00 / 0.16
	16	0.53 / 0.47	0.78 / 0.40	0.84 / 0.29	0.91 / 0.19	0.93 / 0.18	1.00 / 0.19	1.00 / 0.16
	21	0.54 / 0.47	0.85 / 0.39	0.81 / 0.37	0.83 / 0.25	0.93 / 0.24	0.94 / 0.21	1.00 / 0.15
	last	0.50 / 0.47	0.87 / 0.40	0.87 / 0.38	0.89 / 0.30	0.91 / 0.25	0.94 / 0.15	1.00 / 0.15

106 In Figure A7 we plot MS across single neurons. We consider setups in which (a) neurons of CLIP  
 107 ViT-L are evaluated with DINOv2 as the image encoder  $E$ , (b) neurons of CLIP ViT-L are evaluated  
 108 with CLIP ViT-B as  $E$ , and (c) neurons of SigLIP SoViT-400m are evaluated with CLIP ViT-B as  
 109  $E$ . In all three cases SAE neurons are more monosemantic compared to the original neurons of the  
 110 models. It shows that MS results are consistent across different architectures being both explained  
 and used as  $E$ .

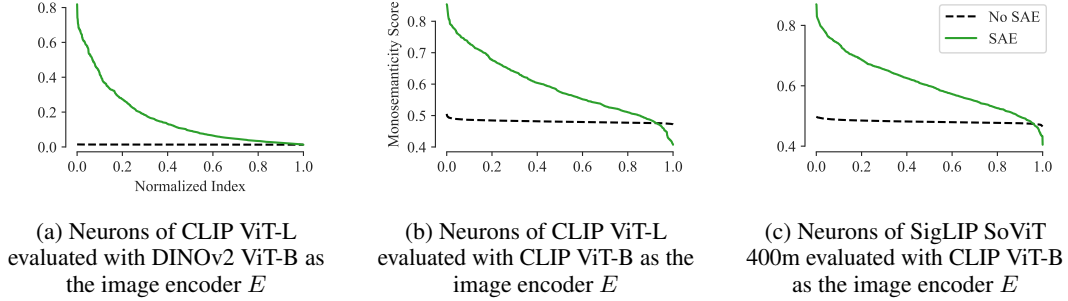


Figure A7: MS in decreasing order across neurons. Results are shown for the last layers of two different models, without SAE (black dashed line), and with SAE being trained with expansion factor 1 (green solid line). MS is computed with distinct image encoders  $E$ .

111

112 In Figures A8 and A9, we plot again MS scores across neurons for SAEs trained with different  
 113 expansion factors and sparsity levels, but using CLIP ViT-B as the image encoder  $E$ . We observe very  
 114 similar patterns when compared to the MS computed using DINOv2 ViT-B. Both higher expansion  
 115 factor and lower sparsity helps find more of the monosemantic units.

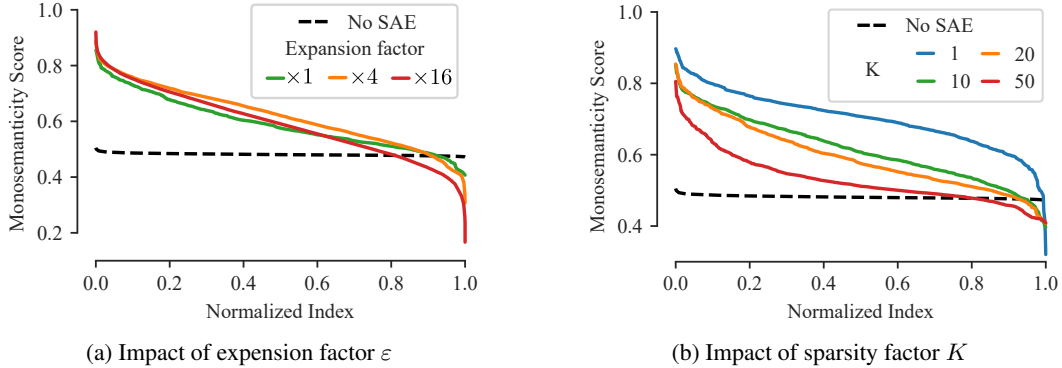


Figure A8: Monosemanticity Scores (computed using CLIP ViT-B) in decreasing order across neurons, normalized by width. Results are shown for the last layer of the model, without SAE (“No SAE”, in black dashed line), and with SAE using either (a) different expansion factors (in straight lines, for  $\varepsilon = 1$ , for  $\varepsilon = 4$  and for  $\varepsilon = 16$ ) or (b) different sparsity levels, with straight lines for  $K = 1$ , for  $K = 10$ , for  $K = 20$ , and for  $K = 50$ .

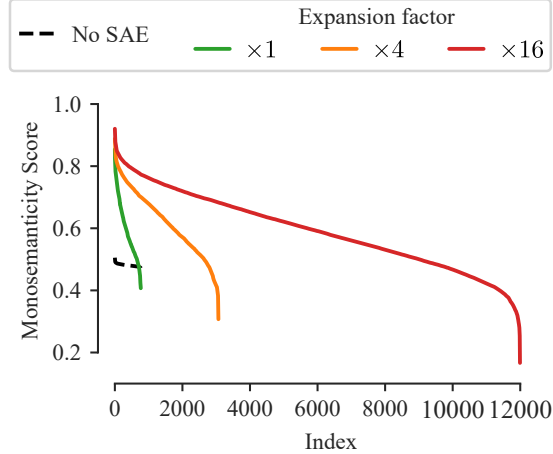


Figure A9: Monosemanticity Scores (computed using CLIP ViT-B) in decreasing order across neurons without normalizing by width. Results are shown for a layer without SAE (“No SAE”), and with SAE using different expansion factors ( $\times 1$ ,  $\times 4$  and  $\times 16$ ).

### 116 E.3 Matryoshka hierarchies

117 We train and evaluate the SAE on embeddings extracted from iNaturalist [1] dataset using an  
 118 expansion factor  $\varepsilon = 2$  and groups of size  $\mathcal{M} = \{3, 16, 69, 359, 1536\}$ . These group sizes correspond  
 119 to the numbers of nodes of the first 5 levels of the species taxonomy tree of the dataset, i.e. the  
 120 respective number of “kingdoms”, “phylums”, “classes”, “orders”, and “families”.

121 To measure the granularity of the concepts, we map each neuron to the most fitting depth in the  
 122 iNaturalist taxonomy tree to compare the hierarchy of concepts within the Matryoshka SAE with  
 123 human-defined ones. To obtain this neuron-to-depth mapping, we select the top-16 activating images  
 124 per neuron, and compute the average depth of the Lowest Common Ancestors (LCA) in the taxonomy  
 125 tree for each pair of images. For instance, given a neuron with an average LCA depth of 2, we can  
 126 assume that images activating this neuron are associated to species from multiple “classes” of the  
 127 same “phylum”. We report the average assigned LCA depth of neurons across the Matryoshka group  
 128 level in Table A8. We notice that average LCA depths are correlated with the level, suggesting that  
 129 the Matryoshka hierarchy can be aligned with human-defined hierarchy. We additionally aggregate  
 130 statistics of MS of neurons for each level. Average and maximum MS also correlates with the level,  
 confirming that the most specialized neurons are found in the lowest levels.

Table A8: Average LCA depth and monosemanticity (MS) scores across neurons at each level in the Matryoshka nested dictionary.

Level		0	1	2	3	4
Depth		3.33	2.92	3.85	3.86	4.06
MS	Avg.	0.06	0.08	0.09	0.16	0.24
	Max.	0.11	0.30	0.29	0.69	0.76
	Min.	0.04	0.03	0.03	0.03	-0.05

131



## F Reconstruction of SAEs

In Table A9 and Table A10, we report respectively  $R^2$  and sparsity ( $L_0$ ), for the two SAE variants we compare in Section 4.2. As BatchTopK activation enforces sparsity on *batch-level*, during test-time it is replaced with  $\text{ReLU}(\mathbf{x} - \gamma)$ , with  $\mathbf{x}$  is the input and  $\gamma$  is a vector of thresholds estimated for each neuron, as the average of the minimum positive activation values across a number of batches. For this reason the test-time sparsity may slightly differ from  $K$  fixed at the value of 20 in our case.

We report in Table A11 the detailed metrics ( $R^2$ ,  $L_0$  and statistics of MS) obtained for SAEs trained with different  $K$  values considered in Section 4.2.

Table A9: Comparison of  $R^2$  (in %) by different SAEs trained with  $K = 20$  for a CLIP ViT-L model.

SAE type	Layer	No SAE	Expansion factor					
			x1	x2	x4	x8	x16	x64
BatchTopK	11	100	74.7	75.0	75.1	75.0	74.7	73.5
	17	100	70.4	71.9	72.6	72.9	72.9	72.5
	22	100	68.7	72.6	74.9	76.0	76.8	77.4
	23	100	67.2	71.5	74.0	75.3	76.0	76.8
	last	100	70.1	74.6	77.1	78.2	78.6	79.1
Matryoshka	11	100	72.8	73.9	74.5	75.1	75.2	74.5
	17	100	67.3	69.5	70.7	71.8	72.6	72.7
	22	100	65.5	69.6	71.5	74.0	75.4	76.6
	23	100	63.9	68.5	71.0	73.1	74.8	74.6
	last	100	66.8	71.6	74.1	76.0	77.6	78.2

Table A10: Comparison of true sparsity measured by  $L_0$ -norm for different SAEs trained with  $K = 20$  for a CLIP ViT-L model.

SAE type	Layer	No SAE	Expansion factor					
			x1	x2	x4	x8	x16	x64
BatchTopK	11	1024	19.7	19.5	19.4	19.6	20.0	22.9
	17	1024	19.4	19.4	19.2	19.6	19.5	22.3
	22	1024	19.6	19.7	19.7	19.8	20.3	23.0
	23	1024	19.8	19.8	19.9	20.1	20.3	22.2
	last	768	19.9	19.9	19.9	20.1	20.2	22.2
Matryoshka	11	1024	19.4	19.5	19.4	19.6	19.8	21.3
	17	1024	19.3	19.3	19.3	19.4	19.5	20.5
	22	1024	19.7	19.7	19.6	19.8	19.9	22.0
	23	1024	19.7	19.8	19.8	19.9	20.6	25.1
	last	768	20.0	19.9	19.8	19.9	20.2	22.5

Table A11: Statistics for SAEs trained with different sparsity constraint  $K$  on activations of the last layer with expansion factor 16. “No SAE” row contains results for raw activations before attaching the SAE.

$K$	$L_0$	$R^2(\%)$	MS		
			Min	Max	Mean
1	0.9	31.3	-0.03	0.90	$0.37 \pm 0.20$
10	9.9	60.6	0.01	0.79	$0.19 \pm 0.16$
20	20.0	66.8	0.01	0.82	$0.16 \pm 0.17$
50	50.1	74.9	0.01	0.69	$0.07 \pm 0.08$
No SAE	—	—	0.01	0.01	$0.01 \pm 0.00$



Figure A10: Images highly activating the neuron we intervene on in Figure 6, which we manually labeled as “Pencil Neuron”.

## G Uniqueness of concepts

The sparse reconstruction objective regularizes the SAE activations to focus on different concepts. To confirm it in practice, we collect top-16 highest activating images for each neuron of SAE and compute Jaccard Index  $J$  between every pair of neurons. The images come from training set. We exclude 10 out of 12288 neurons for which we found less than 16 activating images and use Matryoshka SAE trained on the last layer with expansion factor of 16. We find that  $J > 0$  for 16000 out of 75368503 pairs ( $> 0.03\%$ ) and  $J > 0.5$  for only 20 pairs, which shows very high uniqueness of learned concepts.

## H Additional qualitative results

We illustrate in Figure A10 the highly activating images for the “Pencil” neuron, which we used for steering in Figure 6. In Figures A11 and A12 we provide more randomly selected examples of neurons for which we computed MS using two different image encoders. In both cases we see a clear correlation between score and similarity of images in a grid.

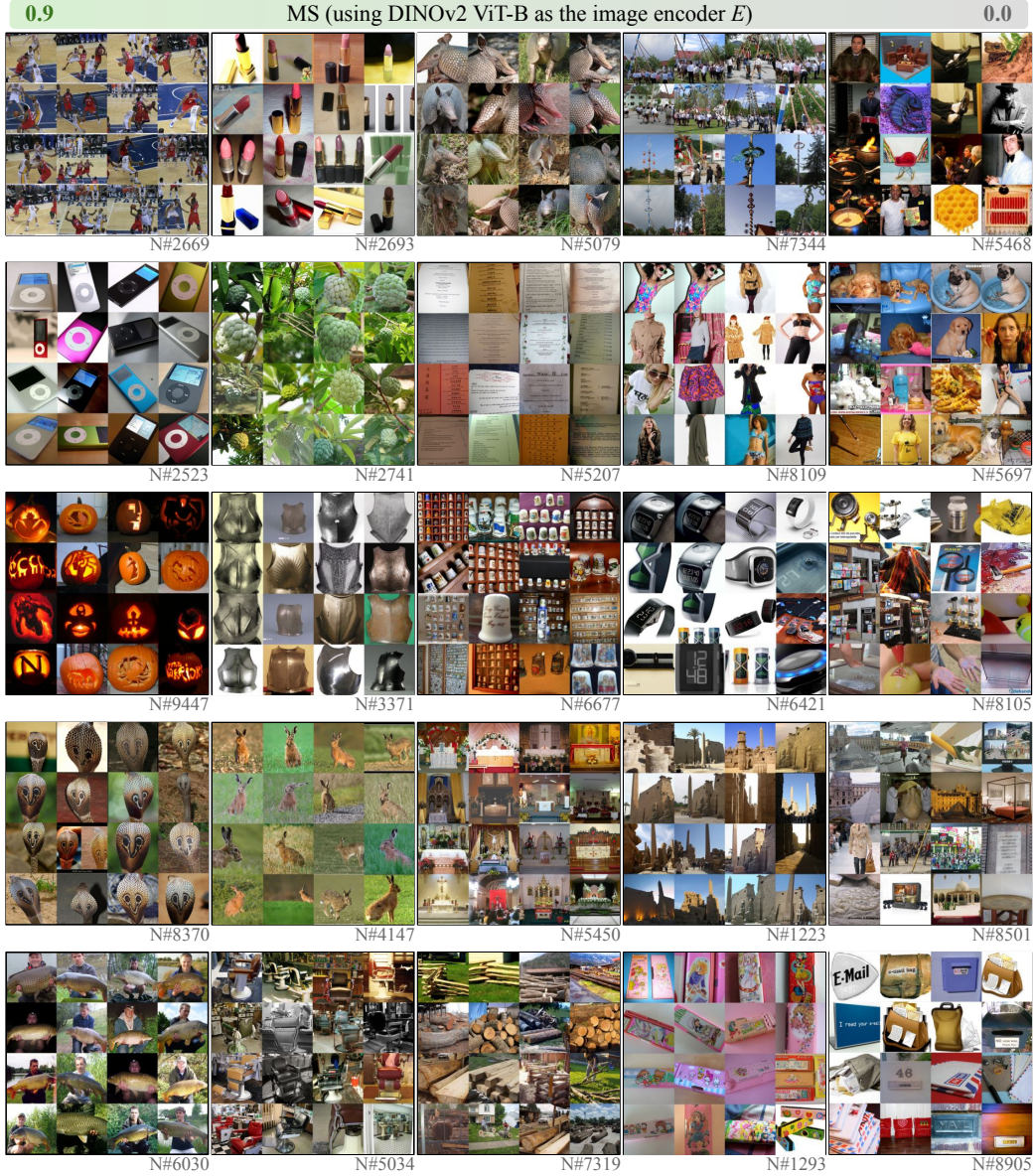


Figure A11: Qualitative examples of highest activating images for different neurons from high (left) to low (right) MS score. As the metric gets higher, highest activating images are more similar, illustrating the correlation with monosemanticity. DINOv2 ViT-B is used as the image encoder  $E$ .



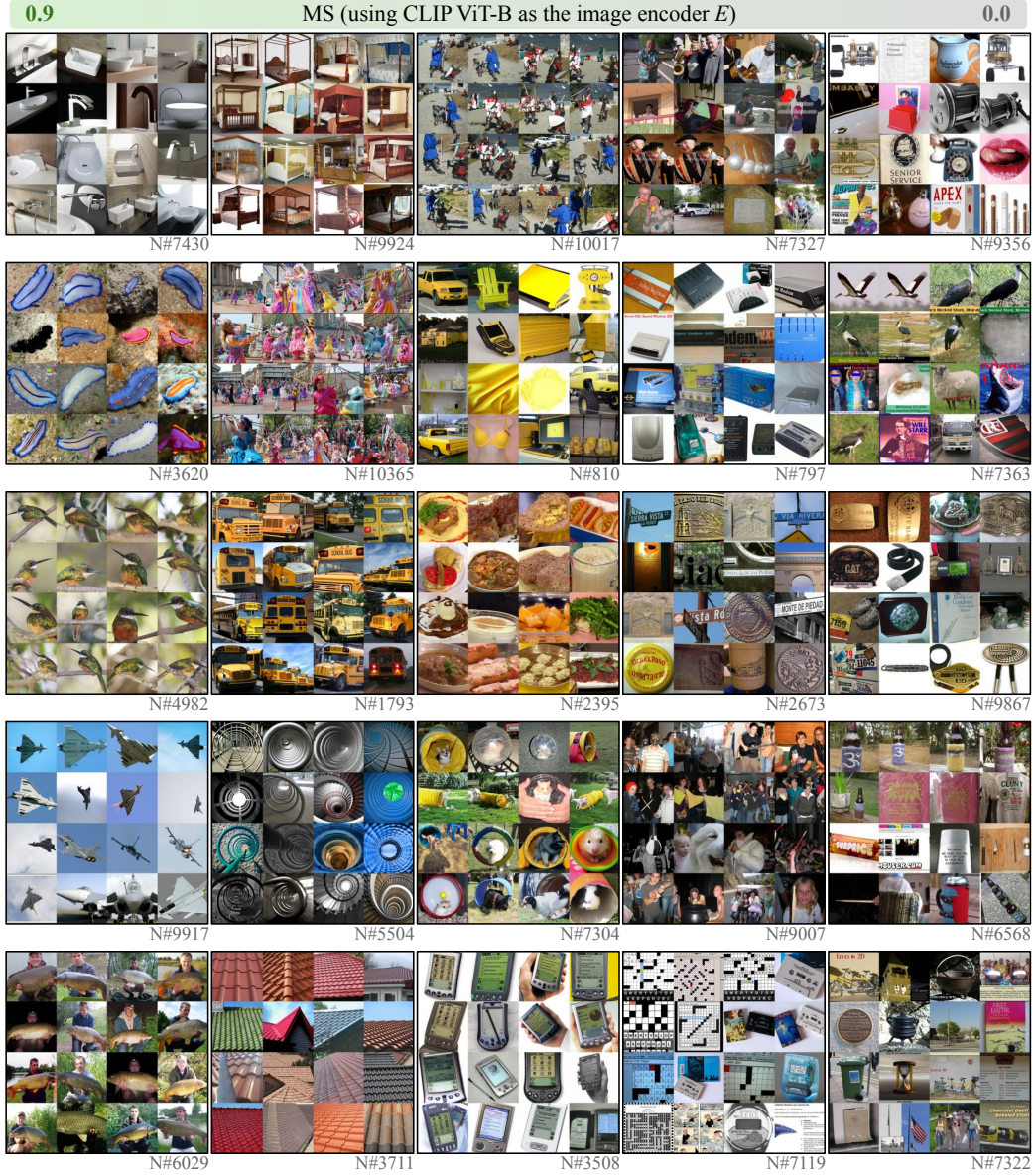


Figure A12: Qualitative examples of highest activating images for different neurons from high (left) to low (right) MS score. As the metric gets higher, highest activating images are more similar, illustrating the correlation with monosemanticity. CLIP ViT-B is used as the image encoder  $E$ .

## 153 **References**

- 154 [1] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha.  
155 Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF*  
156 *conference on computer vision and pattern recognition*, pages 12884–12893, 2021.