A. Generating concepts

A.1. General concepts

We compile our concept set for natural images by parsing product metadata from the Amazon-Berkeley Objects dataset [?] and captions from the MS COCO dataset [?] to extract relevant nouns, noun phrases, verbs, and adjectives. We then clean the set to remove punctuation and phrases longer than four words. This results in more than 190k diverse and descriptive concepts ranging from colors, textures, and patterns that lower-layer neurons tend to typically identify; to objects, parts, scenes, and actions that the deeper neurons in a network might learn.

A.2. Medical concepts

Our concept set is collected from two sources. The first is the class labels of chest X-ray datasets such as [?], [?], and [?], which gives us a list of findings that can be identified from chest X-rays. We also follow the concept generation procedure from [?] and [?] in order to get more fine-grained concepts to describe a neuron's behaviour, where we query GPT-3.5 to provide short, descriptive indicators of the classes the model is trained for. For normal chest X-rays, we additionally include a "No findings" concept.

Based on [?], we use the following prompt to GPT-3.5 to generate descriptive indicators of the 14 disease classes present in the NIH Chest X-ray dataset [?]: *"Can you provide concise radiology descriptors for {class}? List in bullet points with no extra context."*. Using this prompt with the class *Atelectasis*, for example, we get the descriptions:
- Complete or partial lung collapse
- Airless pulmonary parenchyma
- Volume loss
- Crowded vessels
- Mediastinal shift
- Linear or platelike opacities

B. Prompt Ensembling

For our experiments in Section **??** describing neurons in ResNet-50, we perform prompt ensembling using prompts from [?] [?], with a few additions that worked well in practice. Our final list of prompts is:

- 'a photo of a {}.'
- 'a blurry photo of a {}.'
- 'a bad photo of a {}.'
- 'This is a photo of a {}'
- 'This is a blurry photo of a {}'
- 'This is a bad photo of a {}'
- 'There is a {} in the scene'
- 'There is the {} in the scene'
- 'a photo of a {} in the'scene'
- 'a blurry photo of a {} in the scene'
- 'a bad photo of a {} in the scene'
- 'a photo of a small {}.'
- 'a photo of a medium {}.'
- 'a photo of a large {}.'
- 'This is a photo of a small {}.'
- 'This is a photo of a medium {}.'
- 'This is a photo of a large {}.'
- 'There is a small {} in the scene.'
- 'There is a medium {} in the scene.'
- 'There is a large {} in the scene.'

For our experiments in Section **??**, we use the following prompts while generating images for our automated evaluation method. This set of prompts was selected by testing a much larger set of prompts and manually comparing generated images, keeping the ones that produced the highest quality images and/or most diversity:

- '{} in action'
- '{} in sunlight'
- 'a photo of {}'
- 'RAW photo, {}'
- '{} in a scene'
- 'A close up of {}'
- 'Close up photo of {}, soft lightning, Fujifilm XT3'
- '{} in its element'
- 'Spotlight on {}.'
- 'There is a {} in the image.'

For our experiments in Section **??** describing ResNet50 trained with medical images, we use the following prompt:

- 'a chest x-ray containing {}'

C. Effect of varying η used to generate image masks

In this section, we measure the sensitivity of the generated descriptions to the value of η used in equ. **??**. In Table 1, we vary the value of η and show cosine similarity (in text embedding space) of the resulting concepts with the baseline concepts generated with $\eta = 0.10$ (used in the main paper), as well as the percentage neuron descriptions that stay exactly the same. We can see that our results are not sensitive to the specific choice of η, with around 80% of the neurons being assigned exactly the same concept with different values of η. These results are the average over 50 neurons each from Layer 1 and Layer 4 of ResNet50. In general the descriptions are mostly unchanged, and the descriptions that do change are often semantically similar as shown in Figure 1 and the high cosine similarities in Table 1.

Table 1. Similarity of concepts generated using different values of η in eq. **??** as compared to the results reported in the main paper ($\eta = 0.10$). We note good similarity scores even for Layer 1 neurons, which are the most sensitive to η since they look at finer-grained spatial information and are thus affected the most by the shape of the masks generated using η.

η	Avg. Cosine Similarity		Exact matches (%)	
	Layer 4	Layer 1	Layer 4	Layer 1
0.05	0.9800	0.9668	86.0	76.0
0.15	0.9741	0.9712	86.0	78.0
0.20	0.9790	0.9692	88.0	78.0

D. Automated evaluation with $\tau = 0.001$

In this section we present the results of our automatic evaluation results using a stricter threshold for Activation-Fraction, with $\tau = 0.001$ as described in Section **??**. This implies that a generated image counts as highly activating only if it is in the top-0.1% of activations for the neuron. The results are shown in Table 2. As expected, the results are lower across the board than in Table **??**, where the cutoff was top-1%. However, a sizable fraction (10-38%) of generated images is still able to reach top-0.1%. Trends between methods are similar to those in Table **??**, but we can see our method outperforms baselines with a larger margin in this more challenging setting.

E. Qualitative examples of neuron descriptions for ResNet50

In figures 3 to 6, we display additional qualitative examples of neuron descriptions and top activating images for layers 1 to 4 of ResNet50. It is interesting to note how the activation patterns of "checkerboard" (Layer 2, neuron 445) and "spirals" (Layer 3, neuron 286) are different from the visual concepts humans might use to identify these patterns.

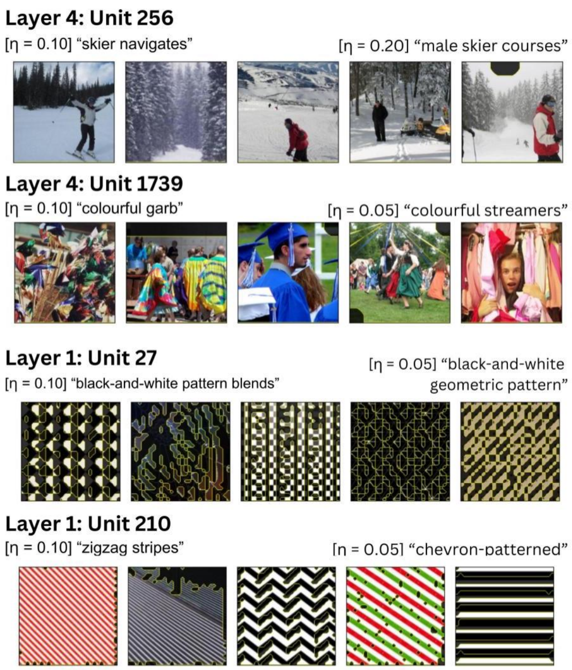

Layer 4: Unit 256
[η = 0.10] "skier navigates"
[η = 0.20] "male skier courses"
Layer 4: Unit 1739
[η = 0.10] "colourful garb"
[η = 0.05] "colourful streamers"
Layer 1: Unit 27
[η = 0.10] "black-and-white pattern blends"
[η = 0.05] "black-and-white geometric pattern"
Layer 1: Unit 210
[η = 0.10] "zigzag stripes"
[η = 0.05] "chevron-patterned"

Figure 1. Examples of semantically similar concepts generated by different values of η that are not exact matches.

F. Qualitative examples of activation-fraction scores

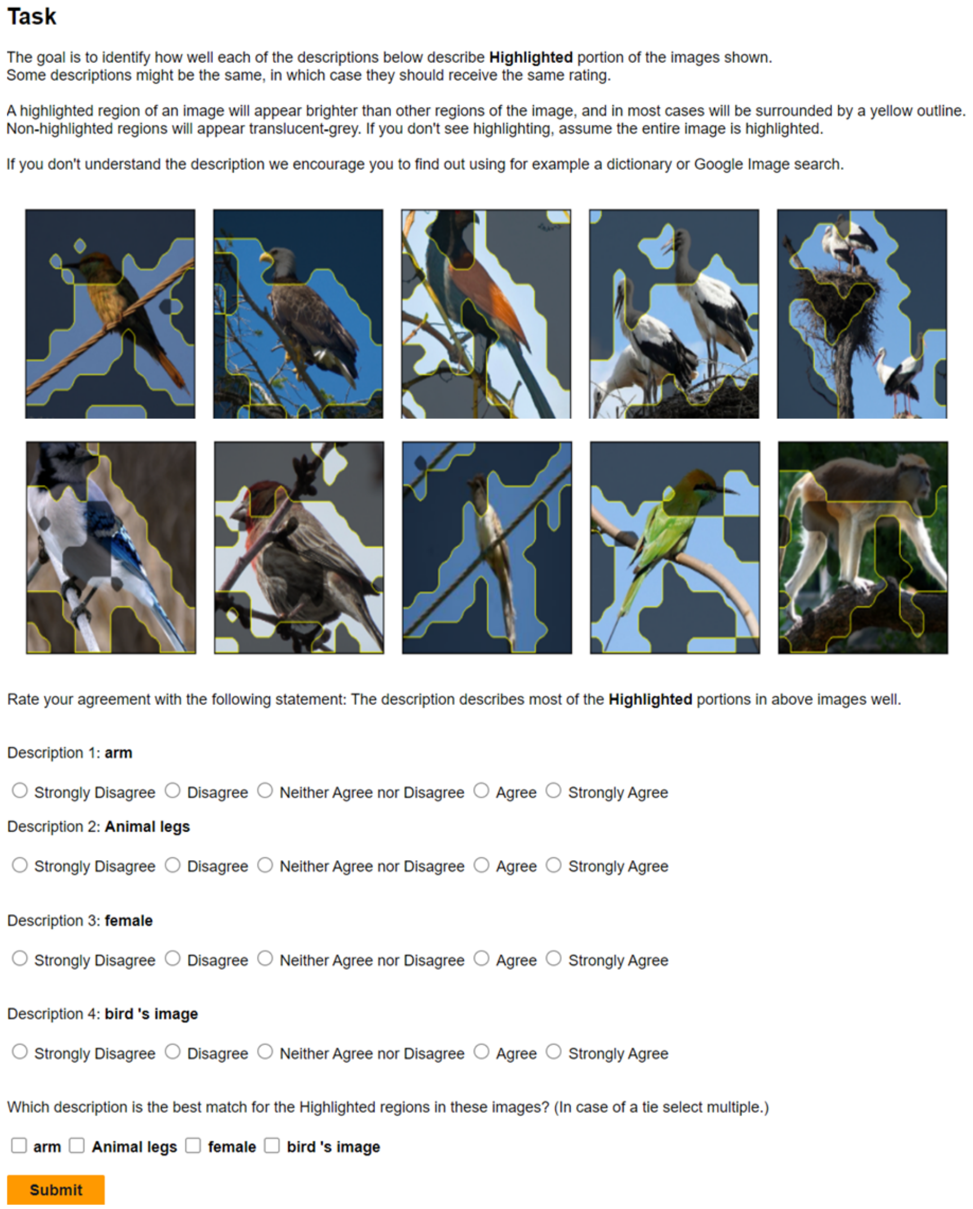
In figures 10 to 7, we display the explanations for 4 randomly selected neurons in each of the layers of ResNet-50 (ImageNet), along with most highly activating inputs, and randomly chosen stable diffusion generated images for that explanation. We can see activation fraction scores are highly aligned with whether the generated images look like (at least one of) the highly activating images. They can also help reveal spurious correlations, where the description captures a feature of highly activating images, but it is not the feature that causally causes the neuron to activate. For example, see layer1 neuron 191 in Figure 10.

G. Examples of neuron descriptions for medical ResNet50 and their relation to final layer neurons

In this section, we look at the relation between class labels for medical ResNet50 and the neurons from layer 4 that highly influence them. Looking specifically at high scoring neuron concepts, we see that for some final layer neurons, the class labels correspond well with the identified concepts

Table 2. Automatic evaluation of neuron concepts for ResNet50 [?] trained on ImageNet [?], using images generated by Stable Diffusion v2-1. These scores represent the average fraction of the generated images that are in the top-0.1% most highly activating inputs for that neuron, as well as standard error of the mean.

Layers	Net-Dissect [?]	MILAN [?]	CLIP-dissect [?]	SAND (Ours)
Layer 1	0.177 ± 0.028	0.143 ± 0.029	0.154 ± 0.032	**0.239 ± 0.040**
Layer 2	0.110 ± 0.019	0.080 ± 0.018	0.143 ± 0.025	**0.158 ± 0.023**
Layer 3	0.091 ± 0.027	0.077 ± 0.021	0.102 ± 0.020	**0.149 ± 0.031**
Layer 4	0.119 ± 0.029	0.082 ± 0.026	0.228 ± 0.033	**0.381 ± 0.043**
All layers	0.124	0.096	0.157	**0.232**



Task

The goal is to identify how well each of the descriptions below describe **Highlighted** portion of the images shown.
Some descriptions might be the same, in which case they should receive the same rating.

A highlighted region of an image will appear brighter than other regions of the image, and in most cases will be surrounded by a yellow outline.
Non-highlighted regions will appear translucent-grey. If you don't see highlighting, assume the entire image is highlighted.

If you don't understand the description we encourage you to find out using for example a dictionary or Google Image search.

Rate your agreement with the following statement: The description describes most of the **Highlighted** portions in above images well.

Description 1: **arm**

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Description 2: **Animal legs**

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Description 3: **female**

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Description 4: **bird 's image**

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Which description is the best match for the Highlighted regions in these images? (In case of a tie select multiple.)

□ **arm** □ **Animal legs** □ **female** □ **bird 's image**

Figure 2. The interface we used for our Mechanical Turk experiments.

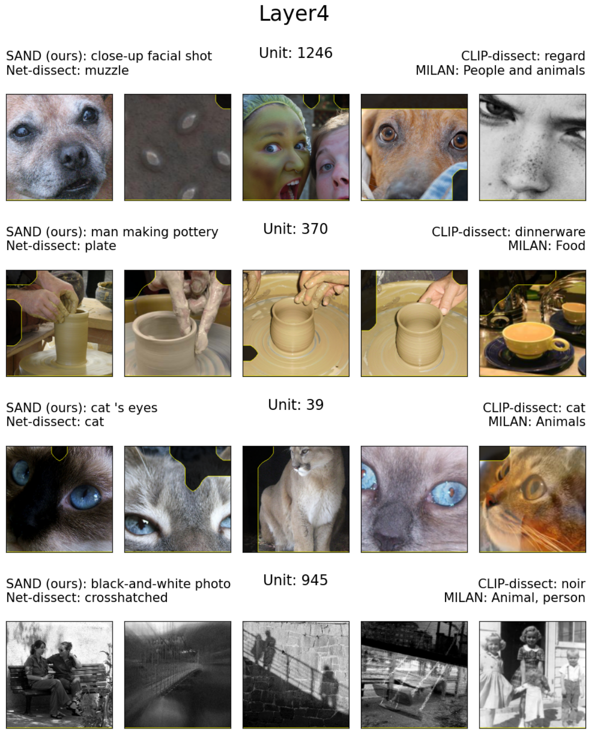


Figure 3. Example Layer 4 neurons from Resnet50, with their top activating images and descriptions assigned by various methods.

in layer 4. For example, the neuron for the class 'Pleural thickening' (thickening of pleural lining around lungs) is highly activated by neurons 519, 577, and 1126 of layer 4, which SAND assigns a relevant concept – 'Irregular, nodular thickening'. Similarly, 'Cardiomegaly' (enlarged heart) is highly activated by neurons 474 and 1320, which are labeled as 'Enlarged heart silhouette' by SAND (Figure **??**). In some cases, we also see partial matches between a class label and its highly activating neuron in layer 4 – such as the class 'Pneumonia' and neuron 451 ('scarring or thickening') – which could happen if the neuron is semantically contributing to multiple classes. We display example neurons in Figure 11.

We also occasionally find that the relation between the concepts learned by layer 4 neurons and their corresponding highly-weighted final layer neurons is not as straightforward. For example, the neuron for the class 'Atelectasis' (complete or partial collapse of lung) is highly activated by layer 4 neurons for 'Visualization of abdominal contents in chest' (neuron 331) and 'Enlarged Cardiomediastinum' (neuron 1396), which are not related to it. Neuron 257, which is connected to the class 'Consolidation' (swelling or hardening of lung tissue that has filled with liquid instead of air) visualises 'abdominal contents in chest', instead of the lung region (Figure 12).

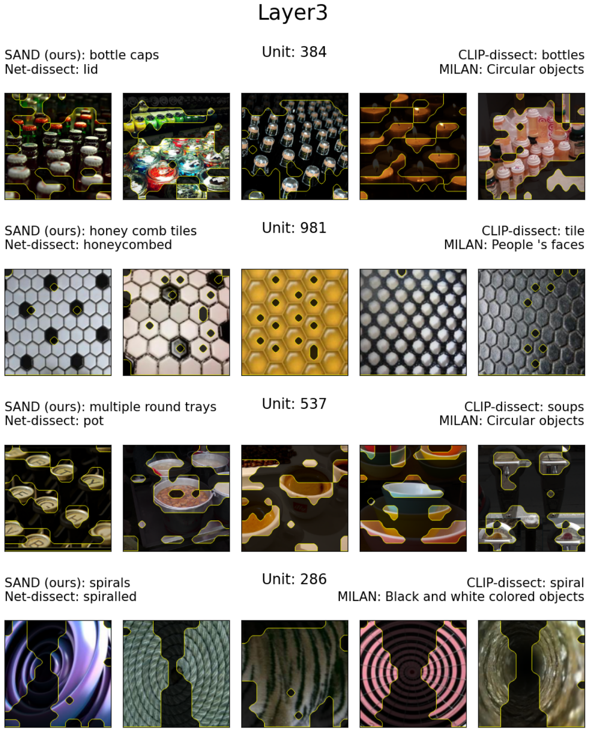


Figure 4. Example Layer 3 neurons from Resnet50, with their top activating images and descriptions assigned by various methods.

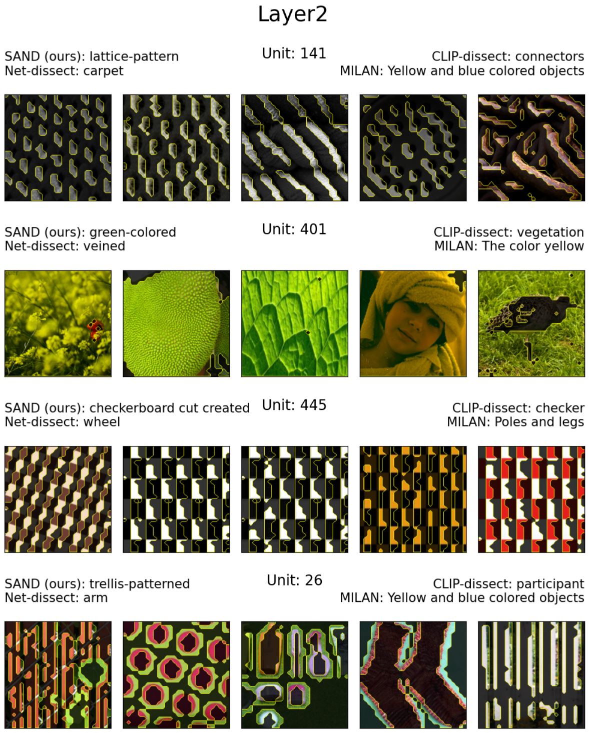


Figure 5. Example Layer 2 neurons from Resnet50, with their top activating images and descriptions assigned by various methods.

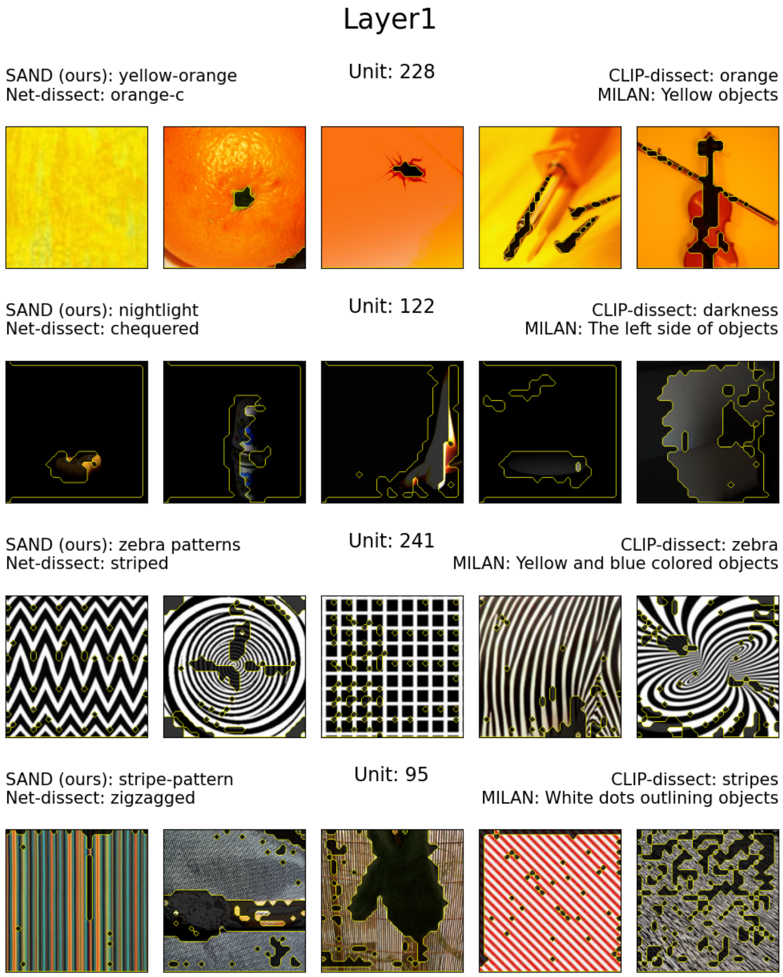


Figure 6. Example Layer 1 neurons from Resnet50, with their top activating images and descriptions assigned by various methods.

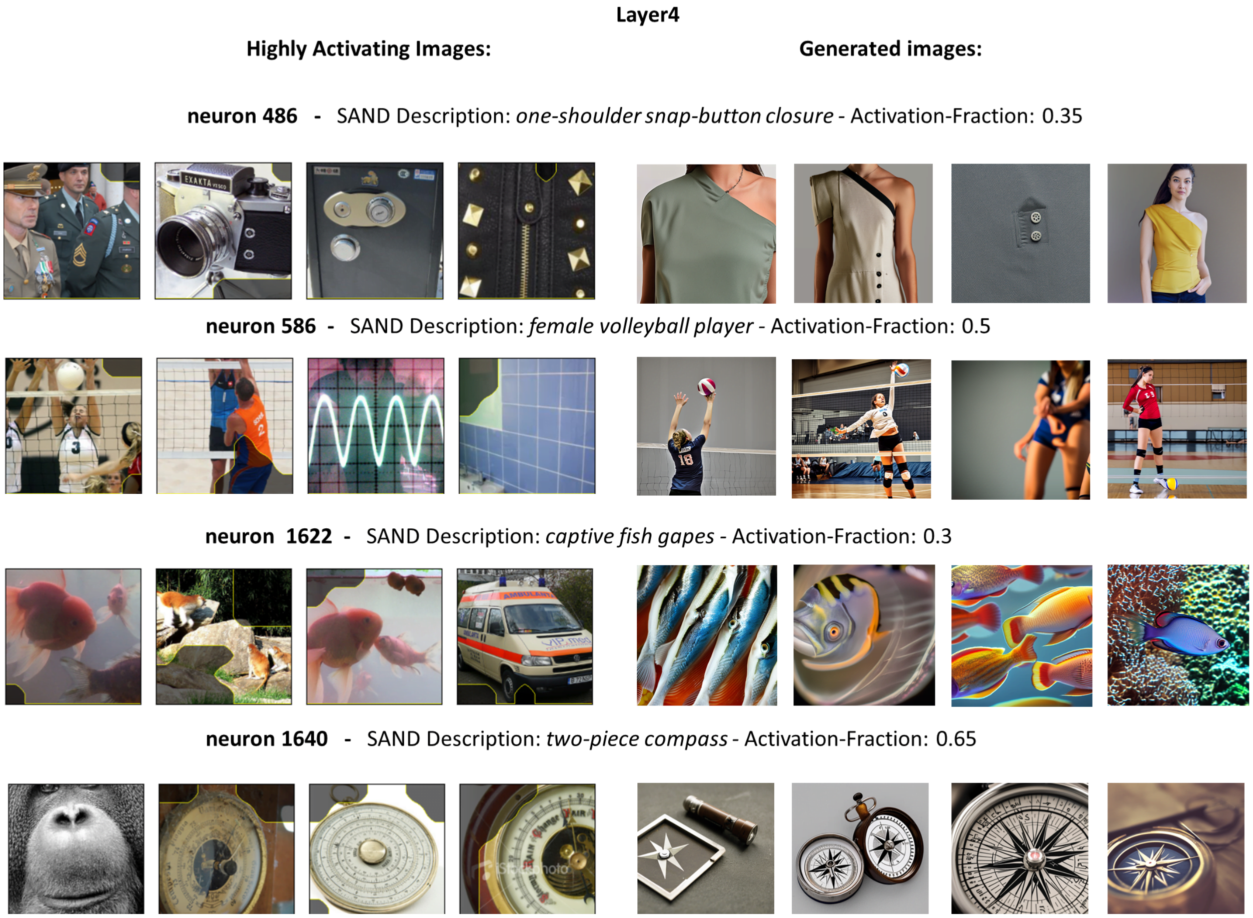

Layer4
Highly Activating Images:
Generated images:
neuron 486 - SAND Description: one-shoulder snap-button closure - Activation-Fraction: 0.35
neuron 586 - SAND Description: female volleyball player - Activation-Fraction: 0.5
neuron 1622 - SAND Description: captive fish gapes - Activation-Fraction: 0.3
neuron 1640 - SAND Description: two-piece compass - Activation-Fraction: 0.65

Figure 7. Randomly chosen Layer 4 neurons from Resnet50(ImageNet), with their top activating images, and random subset of stable-diffusion images generated based on the description.

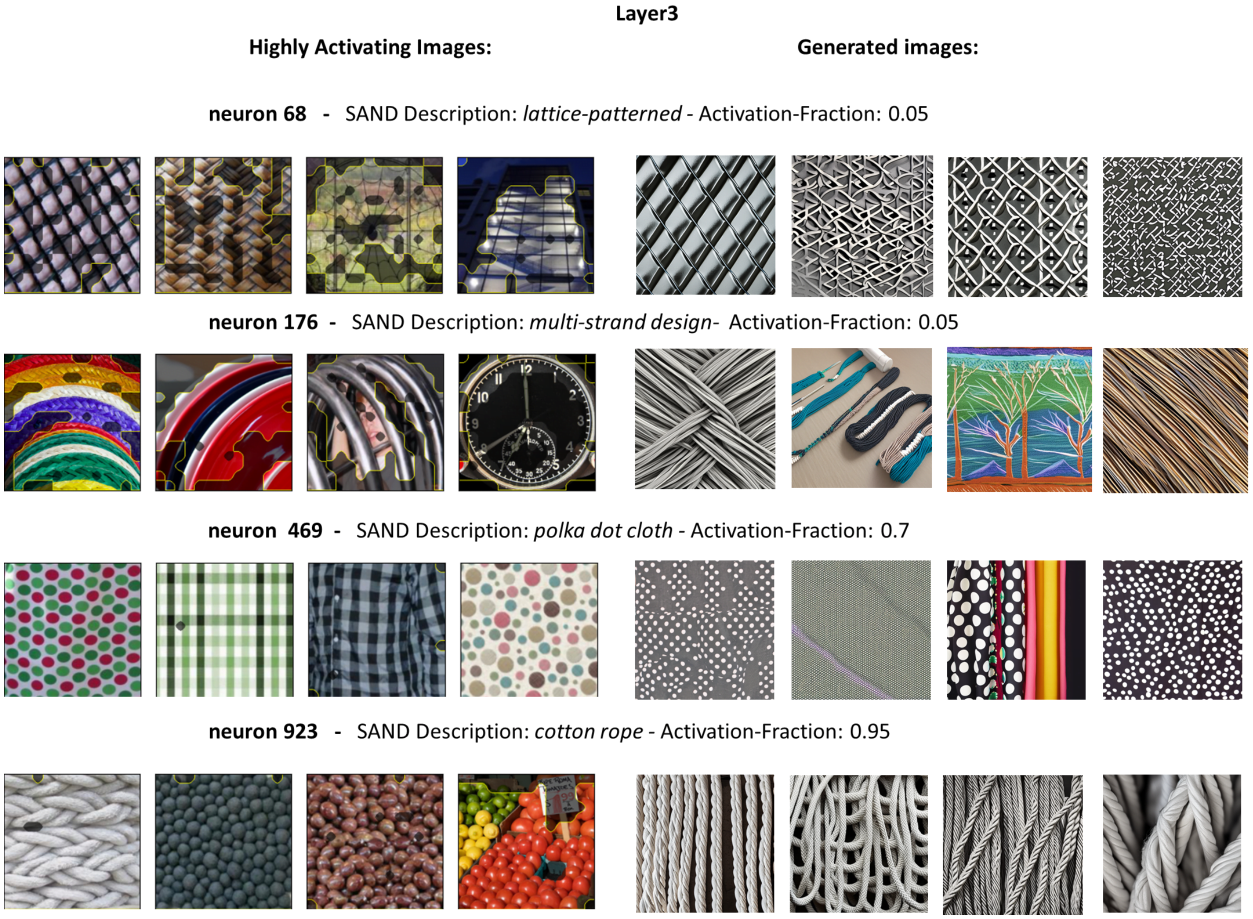

Layer3
Highly Activating Images:
Generated images:
neuron 68 - SAND Description: lattice-patterned - Activation-Fraction: 0.05
neuron 176 - SAND Description: multi-strand design- Activation-Fraction: 0.05
neuron 469 - SAND Description: polka dot cloth - Activation-Fraction: 0.7
neuron 923 - SAND Description: cotton rope - Activation-Fraction: 0.95

Figure 8. Randomly chosen Layer 3 neurons from Resnet50(ImageNet), with their top activating images, and random subset of stable-diffusion images generated based on the description.

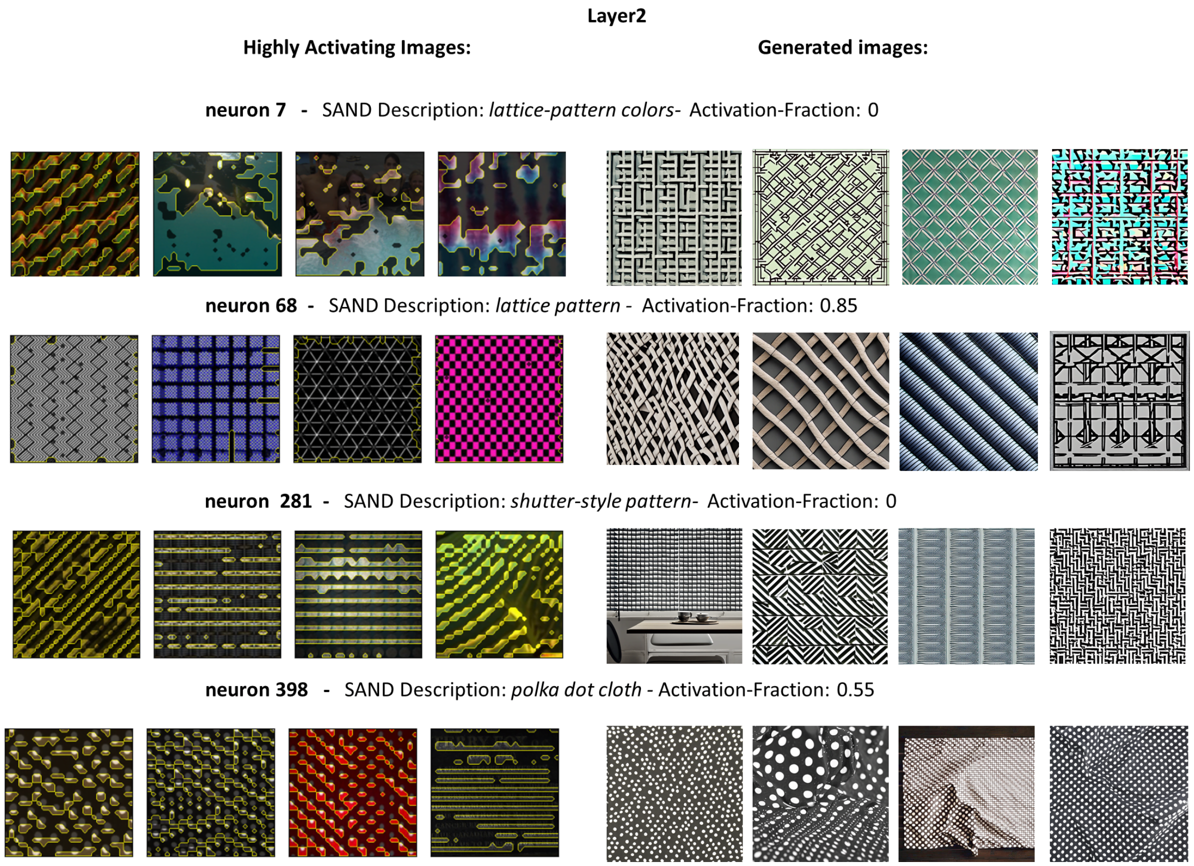

Layer2
Highly Activating Images:
Generated images:
neuron 7 - SAND Description: lattice-pattern colors- Activation-Fraction: 0
neuron 68 - SAND Description: lattice pattern - Activation-Fraction: 0.85
neuron 281 - SAND Description: shutter-style pattern- Activation-Fraction: 0
neuron 398 - SAND Description: polka dot cloth - Activation-Fraction: 0.55

Figure 9. Randomly chosen Layer 2 neurons from Resnet50(ImageNet), with their top activating images, and random subset of stable-diffusion images generated based on the description.

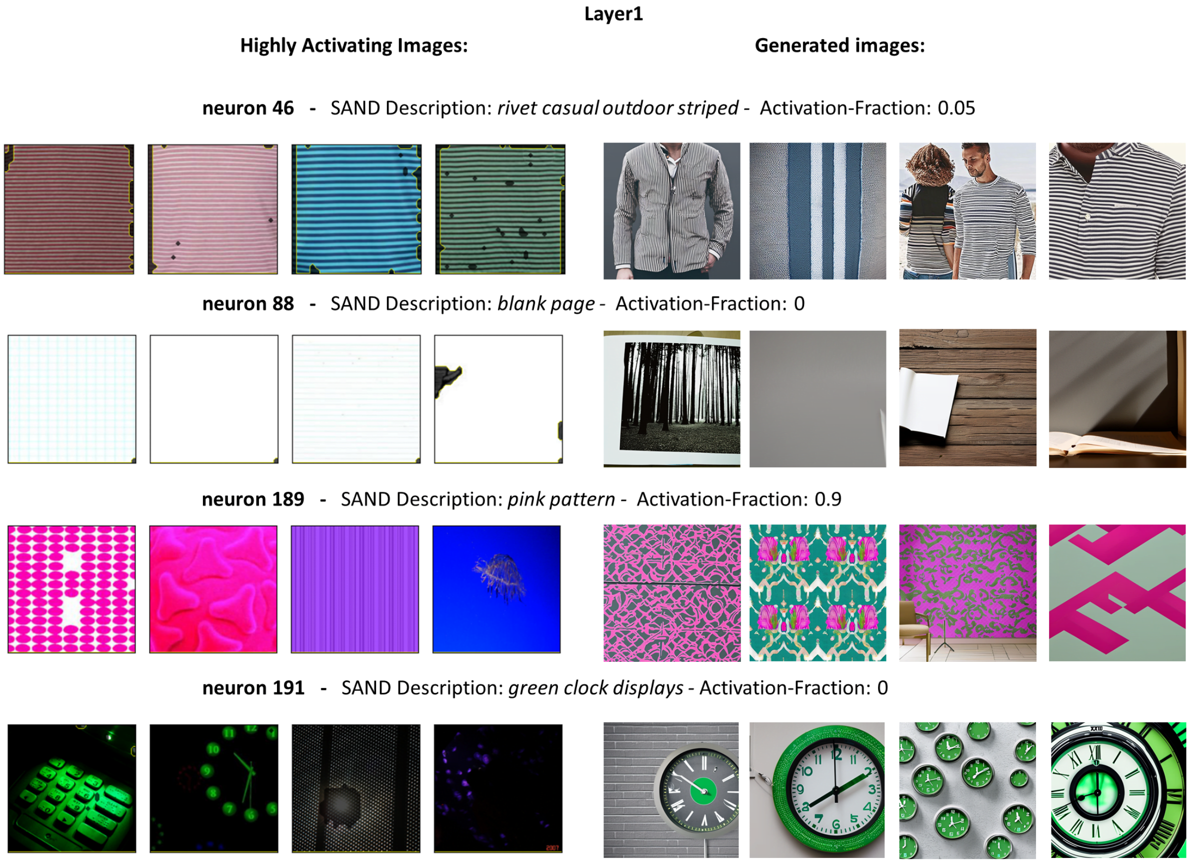
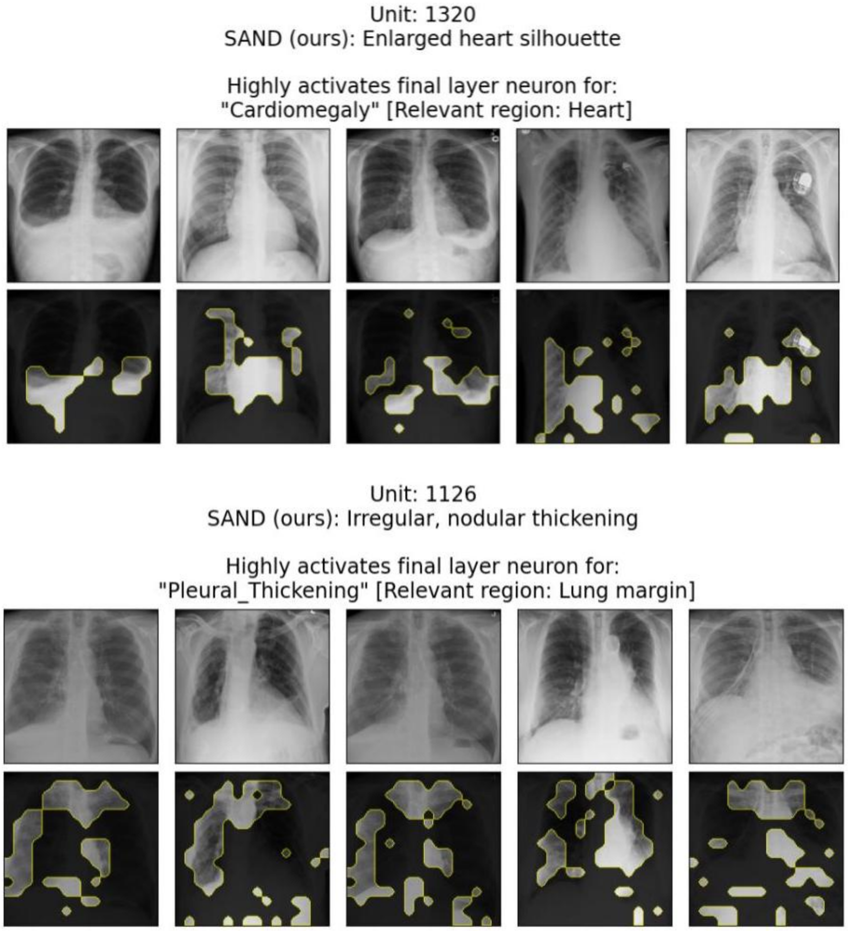

Layer1
Highly Activating Images:
Generated images:
neuron 46 - SAND Description: rivet casual outdoor striped - Activation-Fraction: 0.05
neuron 88 - SAND Description: blank page - Activation-Fraction: 0
neuron 189 - SAND Description: pink pattern - Activation-Fraction: 0.9
neuron 191 - SAND Description: green clock displays - Activation-Fraction: 0

Figure 10. Randomly chosen Layer 1 neurons from Resnet50(ImageNet), with their top activating images, and random subset of stable-diffusion images generated based on the description.



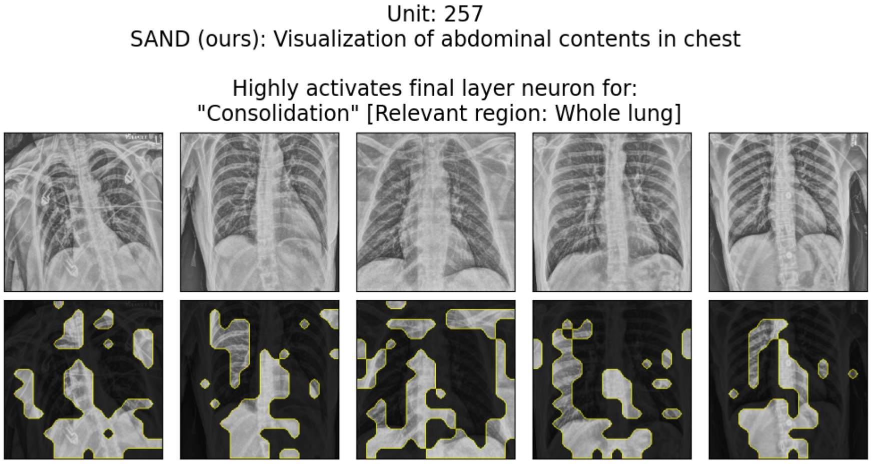
Unit: 1320
SAND (ours): Enlarged heart silhouette

Highly activates final layer neuron for:
"Cardiomegaly" [Relevant region: Heart]

Unit: 1126
SAND (ours): Irregular, nodular thickening

Highly activates final layer neuron for:
"Pleural_Thickening" [Relevant region: Lung margin]

Figure 11. Examples of Layer 4 neuron concepts from medical ResNet50 that match the final layer class they highly influence.



Unit: 257
SAND (ours): Visualization of abdominal contents in chest

Highly activates final layer neuron for:
"Consolidation" [Relevant region: Whole lung]

Figure 12. Example of a Layer 4 neuron concept from medical ResNet50 that is not a complete match for the final layer class it highly influences.