Qualitative examples for Paper #707

We present some qualitative examples of what kinds of concepts the SAEs that we train in the paper fire for, showing concepts from all 4 models. We showcase concepts of different abstractions, starting from more surface-level concepts and going to more abstract concepts. These concepts can all be accessed by putting their number into the `Specific ID` field in https://vlm-concept-visualization.com/ after selecting the correct VLM from the drop-down.

1 Example concepts

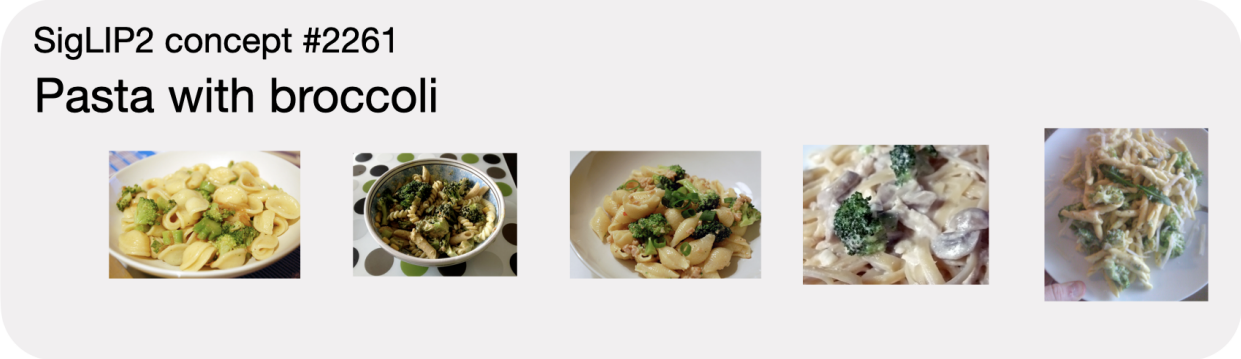

Figure 1: **Coherent concept, with similar visual profile**. Concept #2261 in SigLIP2 fires for pictures of pasta and broccoli, which is an interpretable concept, where all of the images look pretty similar in terms of color, texture, and shape composition

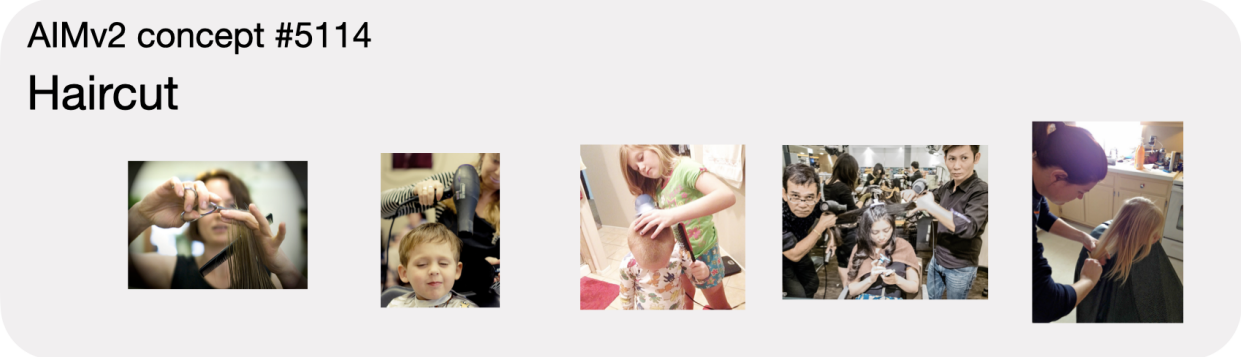

Figure 2: **Coherent concept, with different visual profiles**. Concept #5114 in AIMv2 fires for pictures of people getting haircuts. Though these pictures have fewer surface-level elements that are consistent between them, they all depict the same higher-level interpretable event.

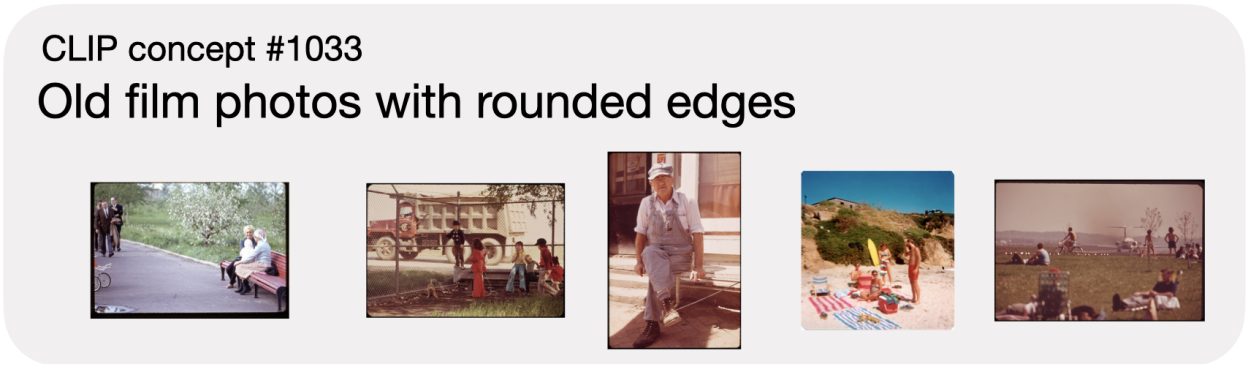

Figure 3: **Similar visual profiles, different semantics**. Concept #1033 in CLIP fires for pictures that have vintage film coloring, rounded edges, and are outdoors. Though the surface-level visual profiles of the top activating examples for this concepts are related, there aren't discernible higher-level semantics that connect them.

Figure 4: **Higher-level abstract semantic concept**. Concept #5812 in SigLIP fires for pictures of many disorganized things. The concepts which are most related to it by BridgeScore are text concepts detailing the idea of many things, rather than any other more surface-level semantics of the images (like, "toys"), hinting that the model has learned a **cross-modal representation of the abstract notion of "several"**

SigLIP concept #2826

Blue

A woman wearing a blue outfit and a blue veil.

A man and women wearing some odd blue things.

A blue chrome motorcycle with a dark blue seat.

A blue park bench painted blue sits on a blue sidewalk.

Two blue chairs sitting next to a blue wooden table.

Figure 5: **Text concept that is surface-level and visually-directed**. Concept #2826 in SigLIP fires for text that describes blue objects. This is a surface-level text feature, describable by the identity of one token. It is also a feature of the text that relates directly to the visual modality.

SigLIP concept #1926

Hypotheticals, opinions, modals

Maybe the man picked that dog because they have the same color hair.

She's probably not going to get to that frisbee in time.

Perhaps they shouldn't be playing outside on such a smoggy day.

She didn't expect that there would be this many birds to feed.

Only an idiot would put mustard on a pair of sugar frosted donuts.

Figure 6: **Abstract linguistic text concept**. Concept #1926 in SigLIP fires for text that contains uncertainty and/or modal verbs like "would" or "should". This concept is picking up on abstract linguistic features about hypothetical. There is no clear relationship in the visual modality between these captions (unlike the "blue" text), and instead there is a language feature that connects them.