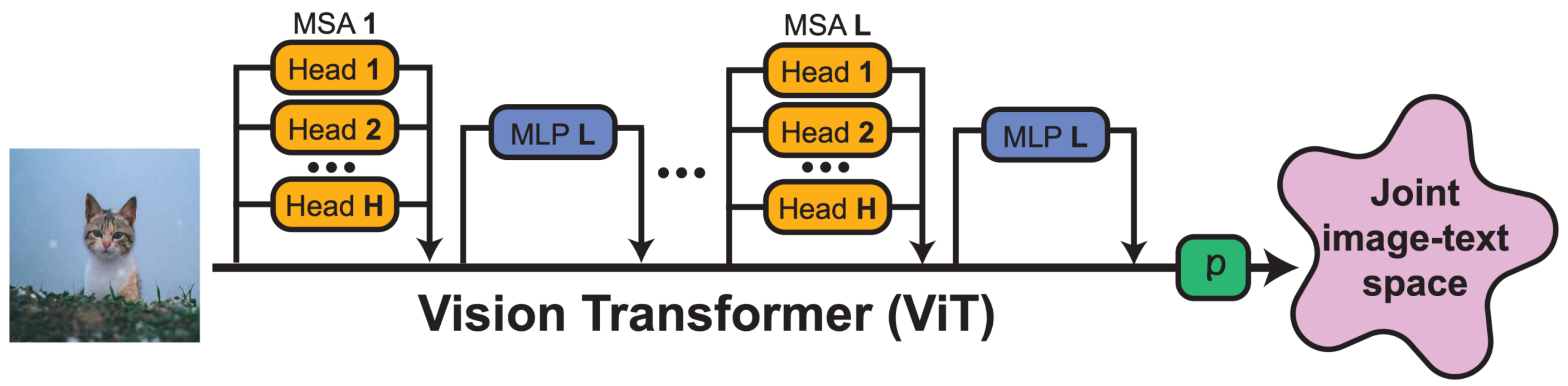
Interpreting the Second-Order Effects of Neurons in CLIP

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

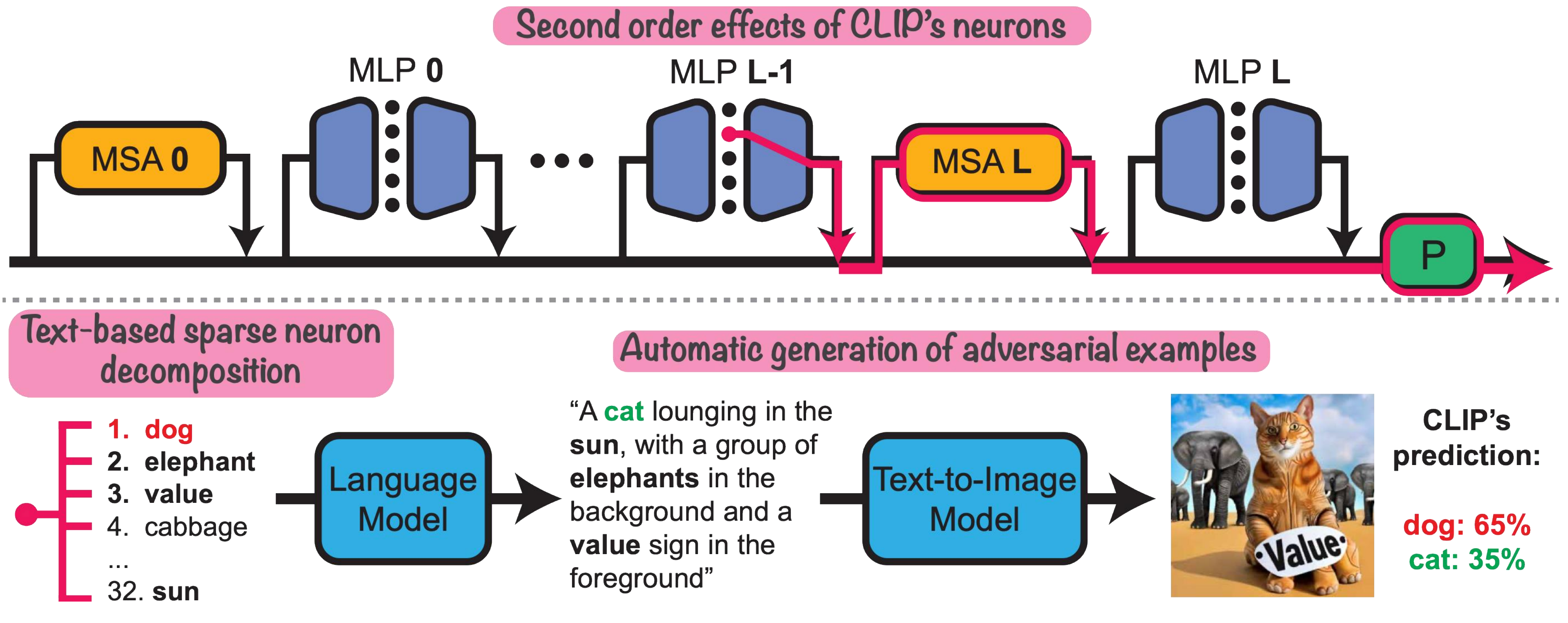
Yossi Gandelsman Alexei A. Efros Jacob Steinhardt

CLIP-VIT



We interpret the roles of neurons in CLIP

These tools allow us to automatically generate adversarial examples and to perform zero-shot segmentation

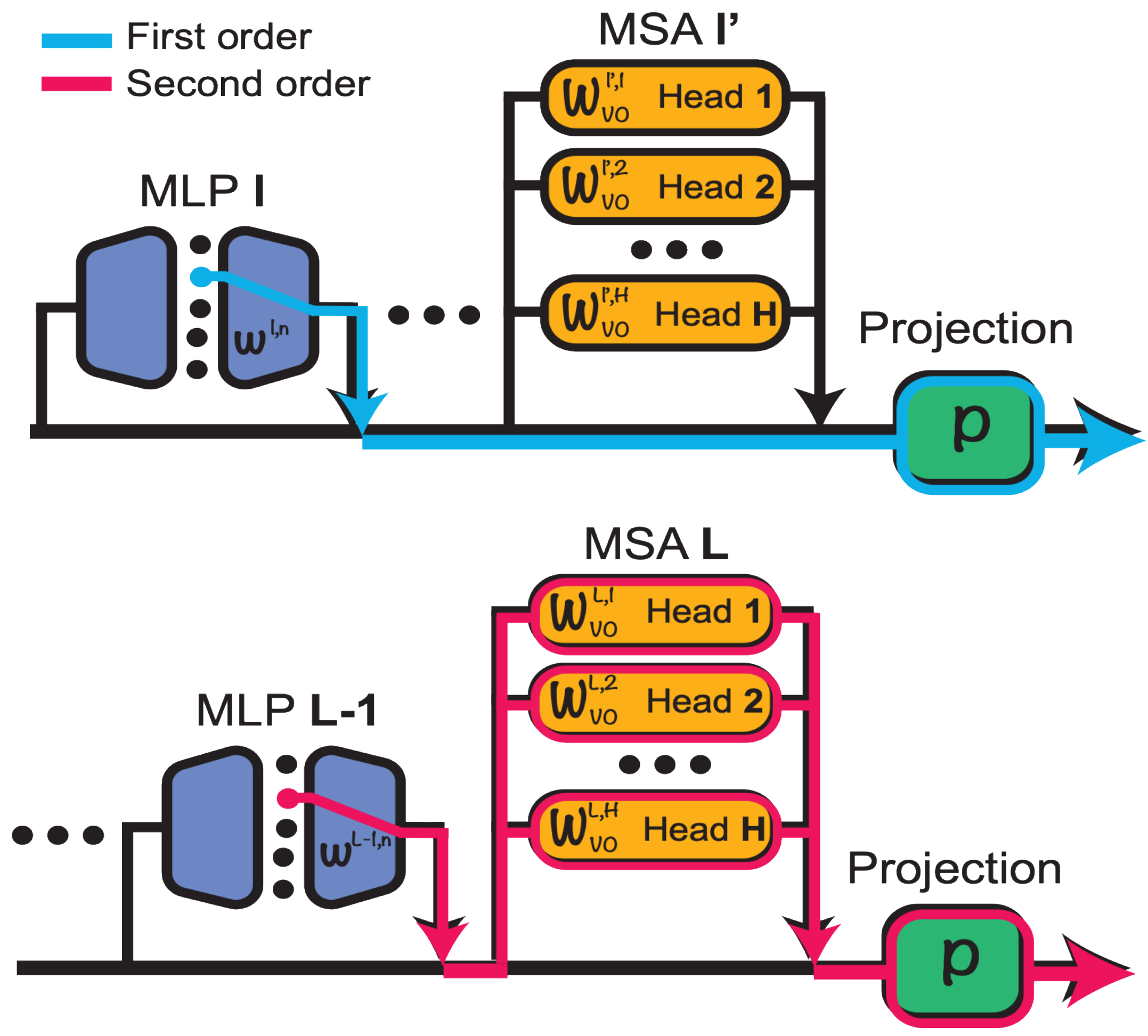


"A cat lounging in the sun, with a group of elephants in the background and a value sign in the foreground"

Interpreting neurons with text

Neuron	Description	5-top activated images
#4	+"snowy" +"frost" +"closings" +"advent"	
#391	+"woodworking" -"swelling" +"cedar" +"heirloom"	
#2137	+"refreshments" +"gelatin" +"sour" +"cosmopolitan"	
#2914	+"motorhome" +"yacht" +"cirrus" +"cabriolet"	

Second-order effects of neurons



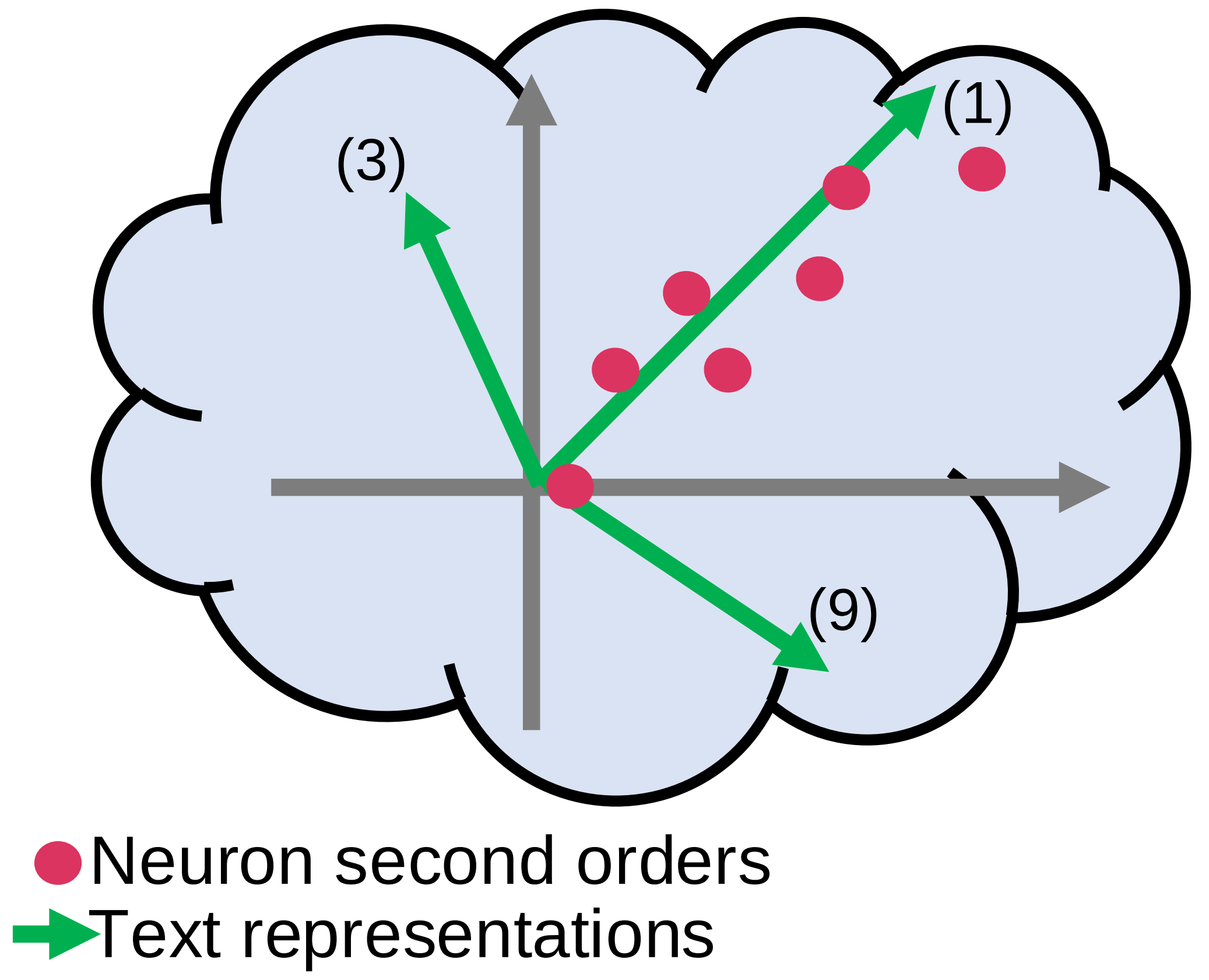
First-order effects of MLP layers are negligible in CLIP

We analyze the **second order** effects – from neurons through the later attention heads, to the output

$$\phi_n^l(I) = \sum_{l'=l+1}^{L} \sum_{h=1}^{H} \sum_{i=0}^{K} \underbrace{\left(p_i^{l,n}(I)a_i^{l',h}(I)\right)}_{\text{attention-weighted activations}} \underbrace{\left(PW_{VO}^{l',h}w^{l,n}\right)}_{\text{input-independent}}$$

Sparse text-based decomposition

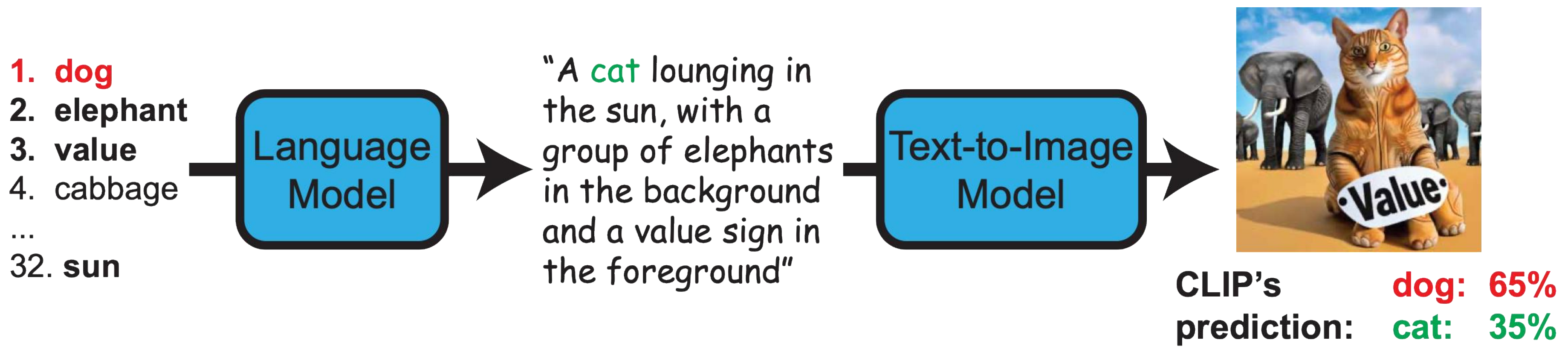
We use linear sparse decomposition technique to describe the second order of a neuron



Descriptions pool

(1) Dog
(2) Cat
(3) Blue
(4) Snowy
(5) Red
(6) April
(7) Two
(8) Seven
(9) Yellow
(10) Whale
...

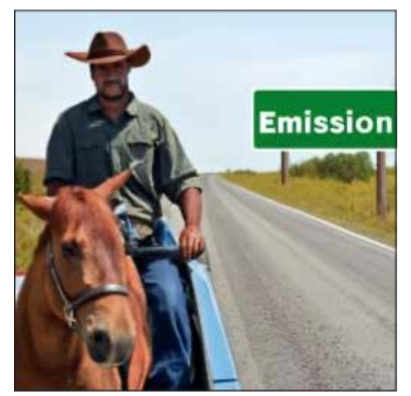
Automatic adversarial examples



"A cat lounging in the sun, with a group of elephants in the background and a value sign in the foreground"

Generated adversarial examples

horse → automobile



A horse is driven by a **trucker** on a road with a sign saying "**Emissions**" in the background.

dog → deer

A dog is running through a **fishery**, chasing after a **squirrel** and a **rabbit**, with a scenic view of the surrounding **forests**.

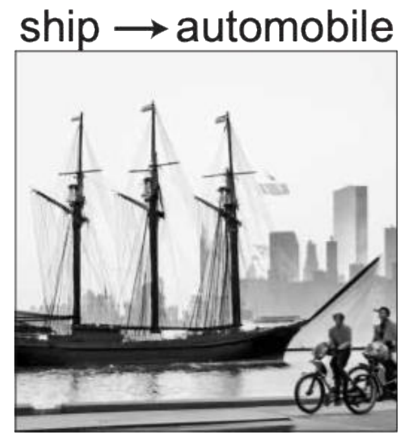
frog → bird

A **poodle** wearing a **ring** and holding a frog, standing next to a sign saying "**Thanksgiving**" in a festive atmosphere.

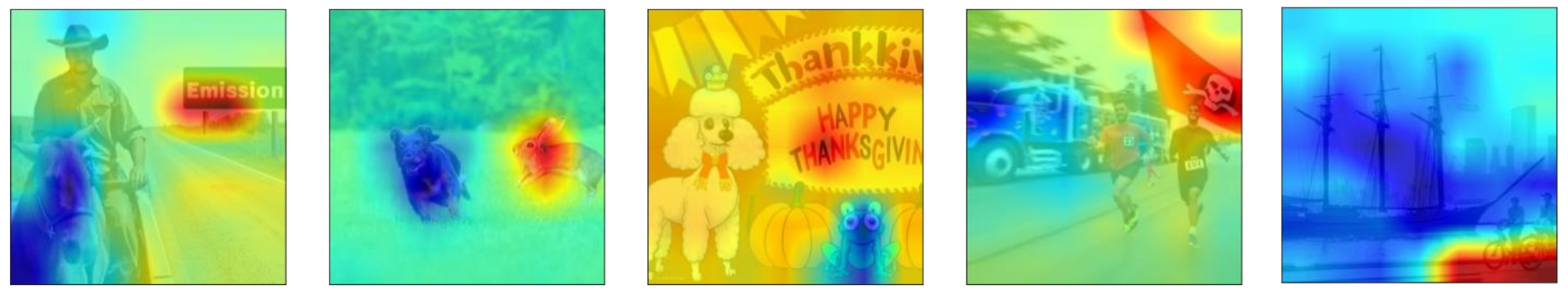
truck → ship

A **duo** of **trainers** are running alongside a truck, with a **pirate** flag waving in the background, and a **sea** of people cheering them on.

ship → automobile



A ship sailing through a **quartet** of **cyclists** on a **Brooklyn** road, with a **husband** and wife driving a **motorcycle** in the background.



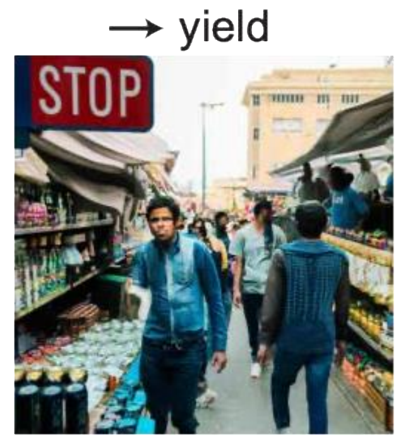
dog → deer

A dog is **walking** with a **patterned** leash through a forest with **rabbits** and **squirrels**, with a symmetrical **patterned** tree in the background.

dog → deer

A dog is running through a **forest**, chasing after a **squirrel**, with a **helicopter** flying overhead and a **patterned** stream in the distance.

stop sign → yield

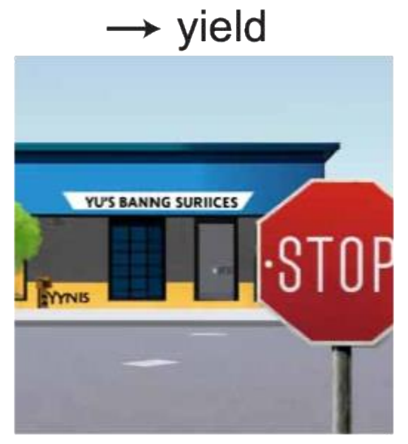


A group of people wandered through a market filled with **cans**, **eggs**, and **perfumes**, with a stop sign in the distance.

stop sign → yield

A stop sign stands in front of a building with a sign that says "**Yu**'s **Banking** Services".

stop sign → yield



A stop sign stands in front of a building with a sign that says "**Yu**'s **Banking** Services".

dog → cat

A dog is sitting on a **moonlight**, looking at a group of **owls** perched on a nearby branch.

frog → bird

A **writer** sitting on a **winged pony**, holding a **poodle** and wearing a **yuri**-themed hat, with a frog on its shoulder.

frog → bird

A frog riding on the back of an **elephant**, with auras of **purple** and orange surrounding them.

cat → vacuum cleaner

A cat is playing with a **hockey** stick near a **shovel** and a **venous** injection kit.

cat → vacuum cleaner

A cat is **brushing** its fur with a **blunt** comb, surrounded by drops of **ethanol** and a **dvr** recording in the corner.

Zero-Shot Segmentation

Averaging the activation maps of relevant neurons according to their descriptions

	Pix. Acc. ↑
Partial-LRP (Voita et al., 2019)	55.0
Rollout (Abnar & Zuidema, 2020)	61.8
LRP (Binder et al., 2016)	62.9
GradCAM (Selvaraju et al., 2017)	67.3
Chefer et al. (2021)	68.9
Raw-attention	69.6
TextSpan (Gandelsman et al., 2024)	76.5
Ours	**78.1**

performance on ImageNet-segmentation