SUSTech
Southern University of Science and Technology

神经计算与控制实验室
Neural Computing & Control Lab

Brain Intelligence and Artificial Intelligence

人脑智能与机器智能

Lecture 15 – Concept representation in human and AI

Quanying Liu （刘泉影）

SUSTech, BME department

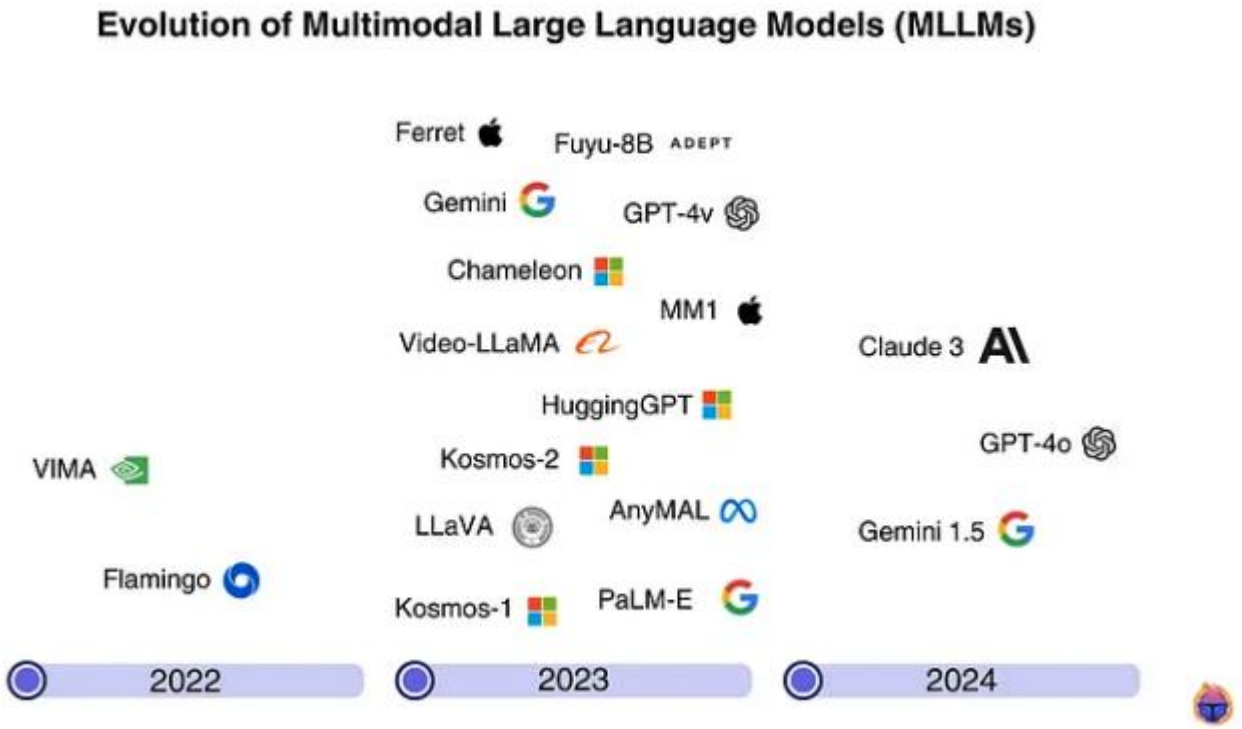
Email: liuqy@sustech.edu.cn

Lecture 15 – Concept representation in human and AI

- Cognitiove processes in Large Models & Humans
- **Computer vision**: CNN model & latent representation
- **Human vision**: Concept Bottleneck Model (CBM) & Interpretable concept representation
- **NCC LAB's work**: CoCoG & CoCoG-2
- **Future directions**

Multimodal LLMs (MLLMs)

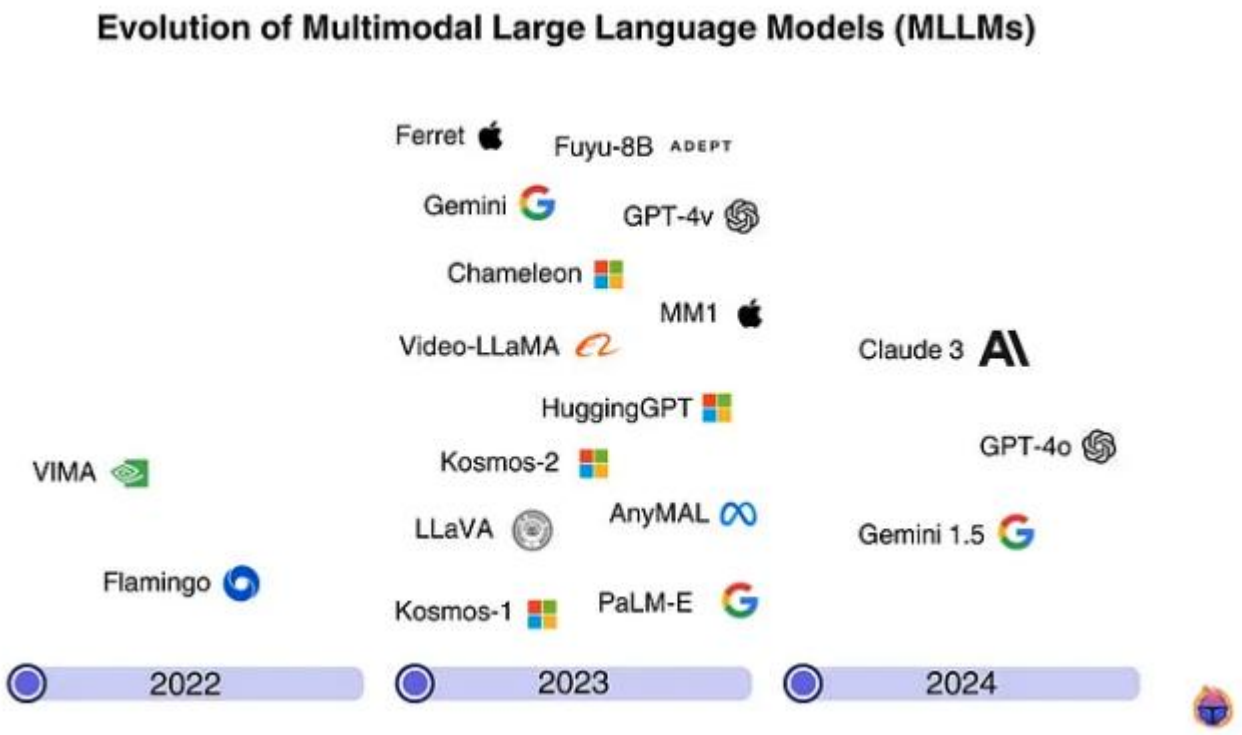
- To fuse multi-modal data (text, image, audio, video...) for learning
- To improve task performance in many complex dowmstream tasks



https://medium.com/@tenyks_blogger/multimodal-large-language-models-mllms-transforming-computer-vision-76d3c5dd267f

Multimodal LLMs (MLLMs)

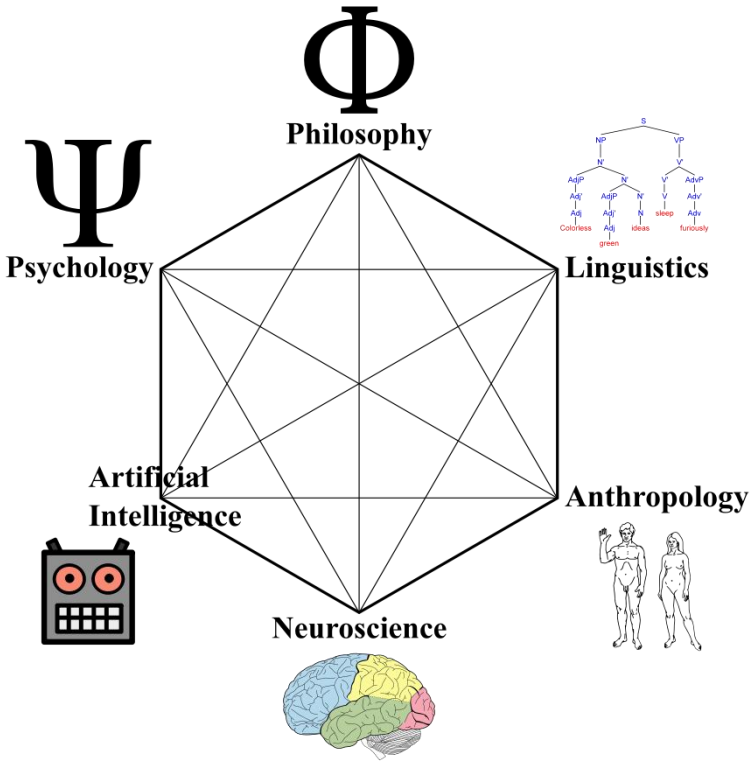
- To fuse multi-modal data (text, image, audio, video...) for learning
- To improve task performance in many complex dowmstream tasks



https://medium.com/@tenyks_blogger/multimodal-large-language-models-mllms-transforming-computer-vision-76d3c5dd267f

Cognitive Science

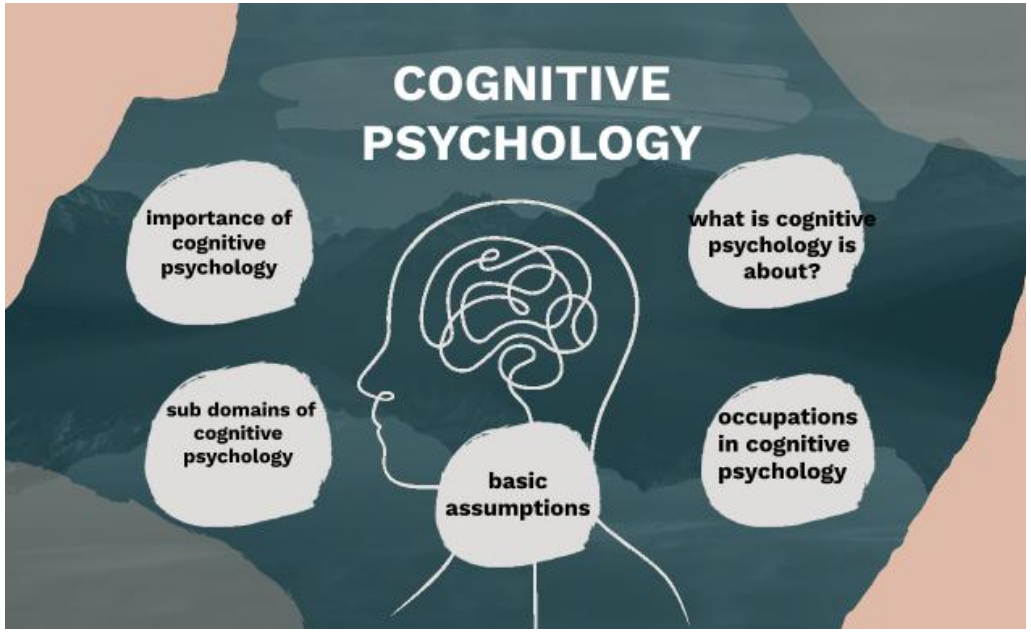
- to study the *mind* and its *processes* in humans and animals
- to examine the nature, the tasks, and the functions of *cognition*



https://en.wikipedia.org/wiki/Cognitive_science

Use methodology from cognitive science to understand cognition (or intelligence) of MLLMs

evaluate / interpret



Something **we do not know**

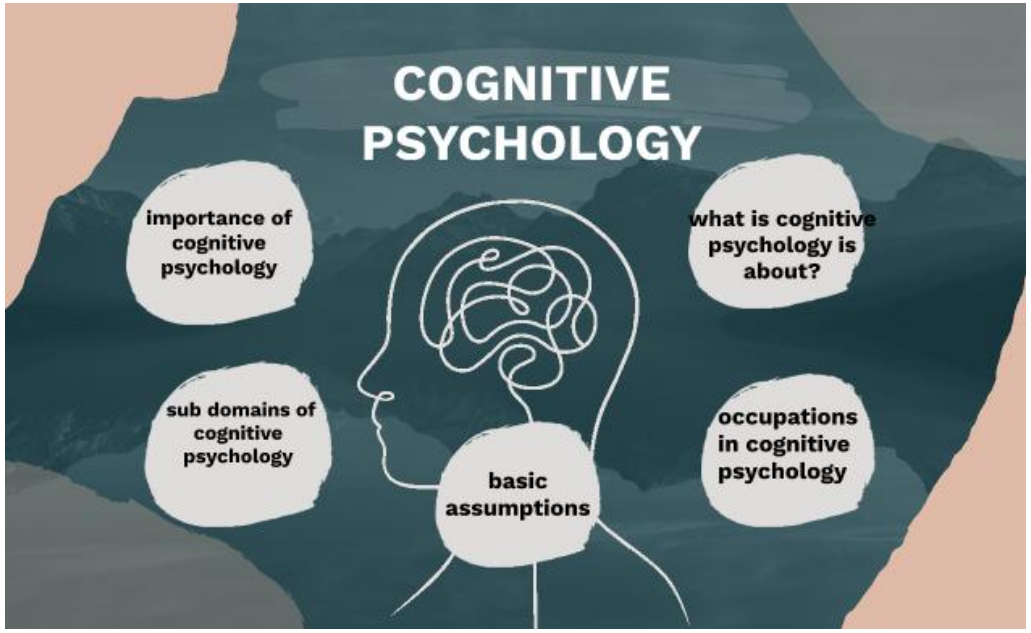
- MLLMs' mind and cognitive processes
- The intelligence capability of MLLMs
- Safety of MLLMs in human society

Something **we have known**

- Cognitive task design
- Data collection (e.g., behavioral data...)
- Theory from cognitive science, psychology

Use LLMs as a tool to understand the brain function (or cognitive process) in human

align
/
model



Something **we have known**

- MLLMs' network strcuture
- MLLMs' latent representation
- MLLMs' input & output

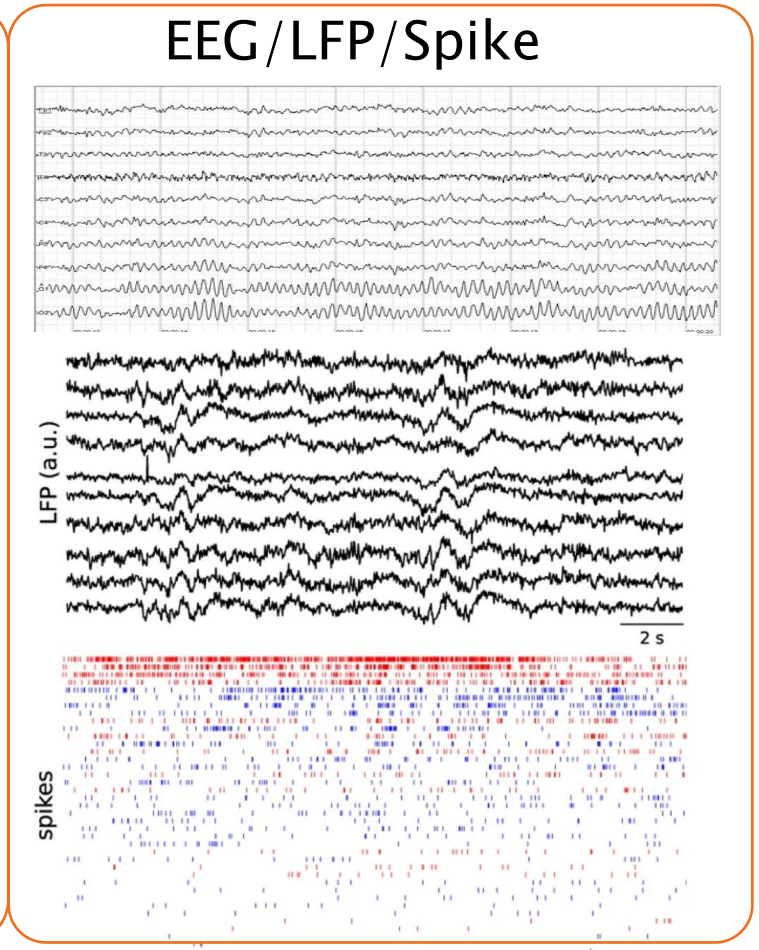
Something **we do not known**

- Human' s brain circuits for each function
- Human' s latent representation
- The neural machnisms underlying behaviors

MLLMs for Cognitive Science

1. To integrate multiple data modalities (behavioral data, neural data, ...)

- Multimodal models provide a framework for integrating multimodal data in cognitive science

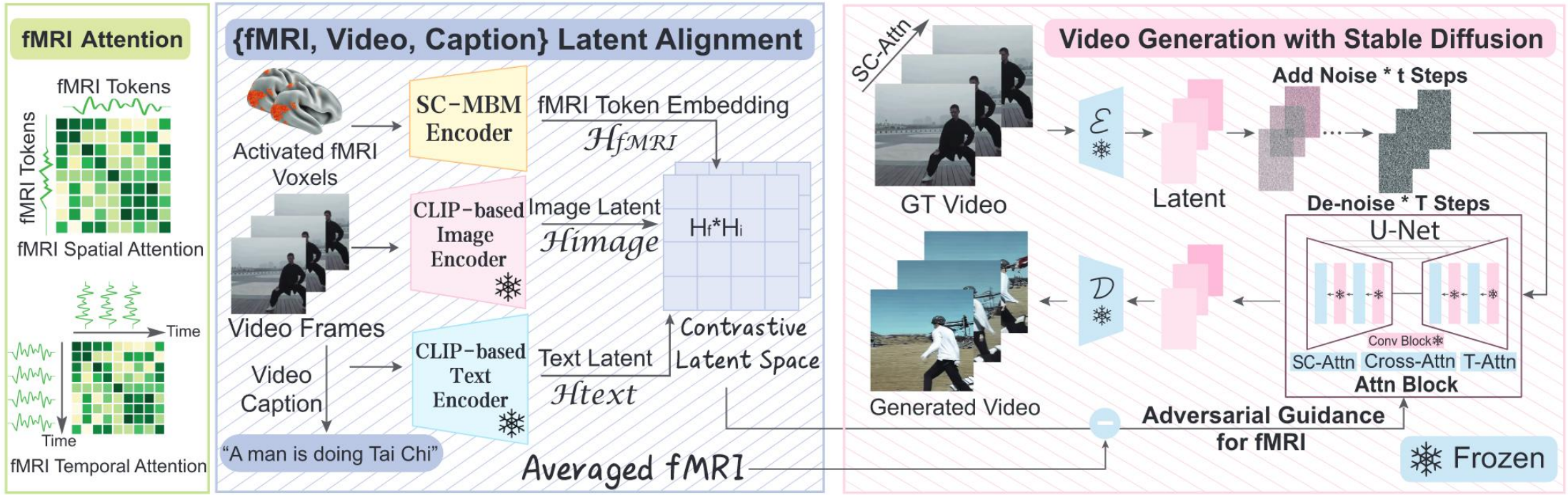


www.wikipedia.org
Teleńczuk et al., Nature (2017)
Singer, U. et al., arXiv (2022)

MLLMs for Cognitive Science

1. To integrate multiple data modalities (behavioral data, neural data, ...)

- MLLMs provide a pretrained latent representation to align with neural embeddings
- MLLMs deal with insufficient data in cognitive science and neuroscience



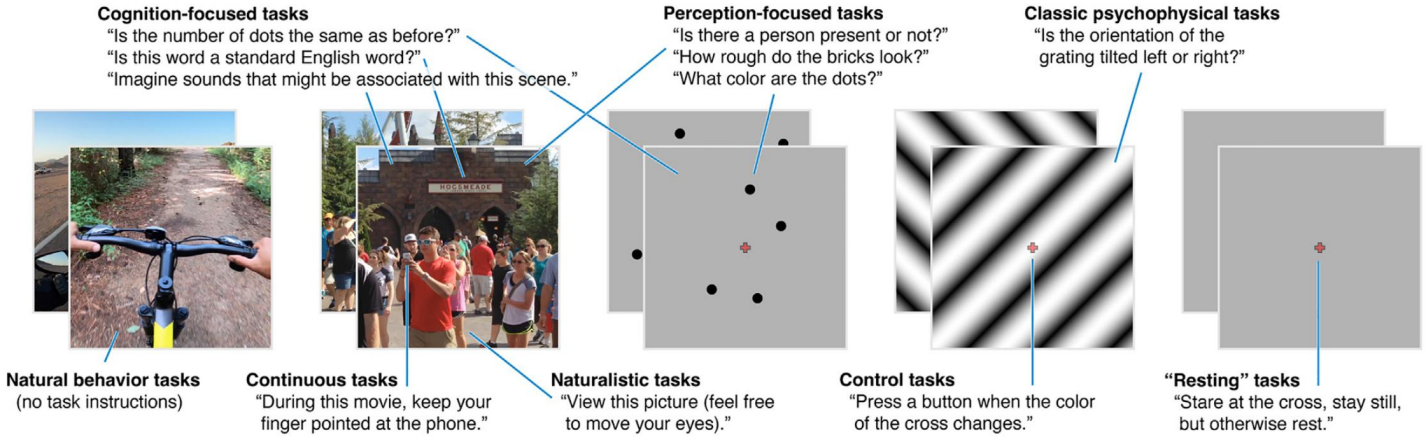
Chen, Z., Qing, J., & Zhou, J. H. (2023). *Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity* (arXiv:2305.11675).

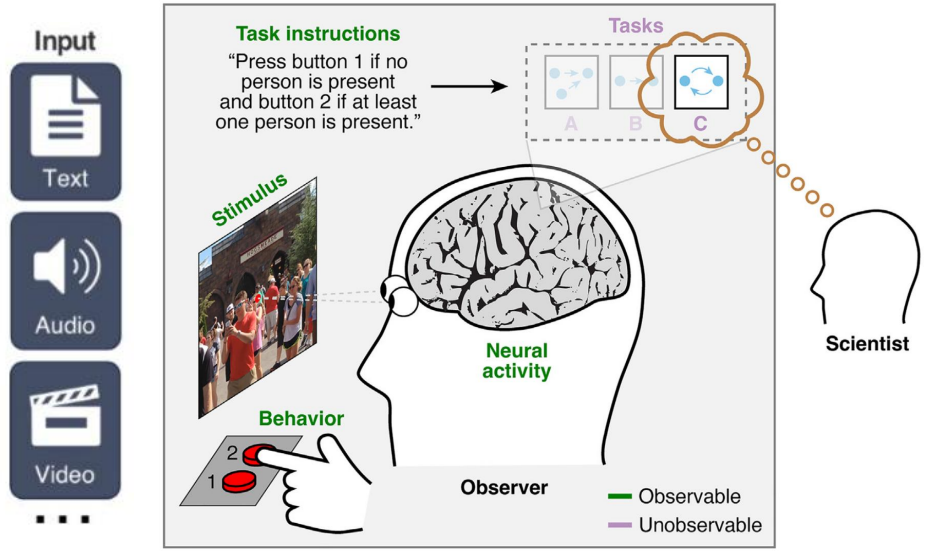
MLLMs for Cognitive Science

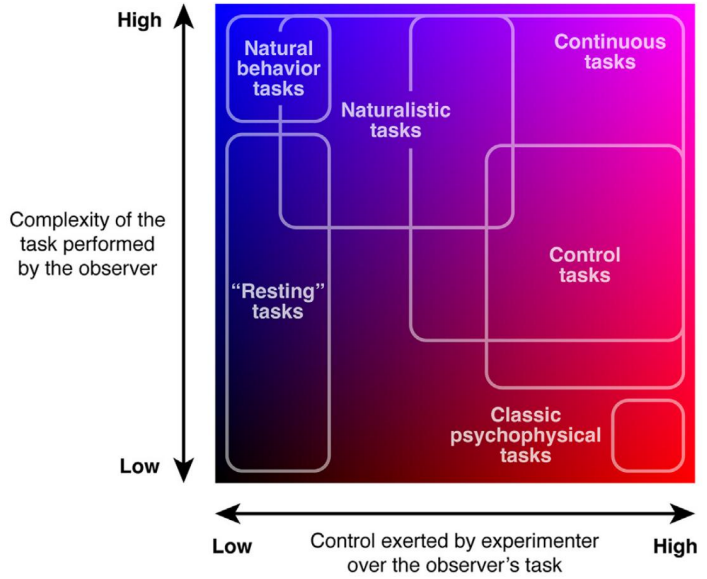
1. To integrate multiple data modalities

2. To mimic cognitive process of humans

- Humans naturally perceive multimodal sensory stimulus (eg., auditory, visual and olfactory inputs).
- There is a growing trend of using "**naturalistic**" stimulus (rather than the bars & dots) in cognitive science.







Kay, Kendrick, Kathryn Bonnen, Rachel N. Denison, Mike J. Arcaro, and David L. Barack. "Tasks and Their Role in Visual Neuroscience." *Neuron* 111, no. 11 (June 2023): 1697–1713.

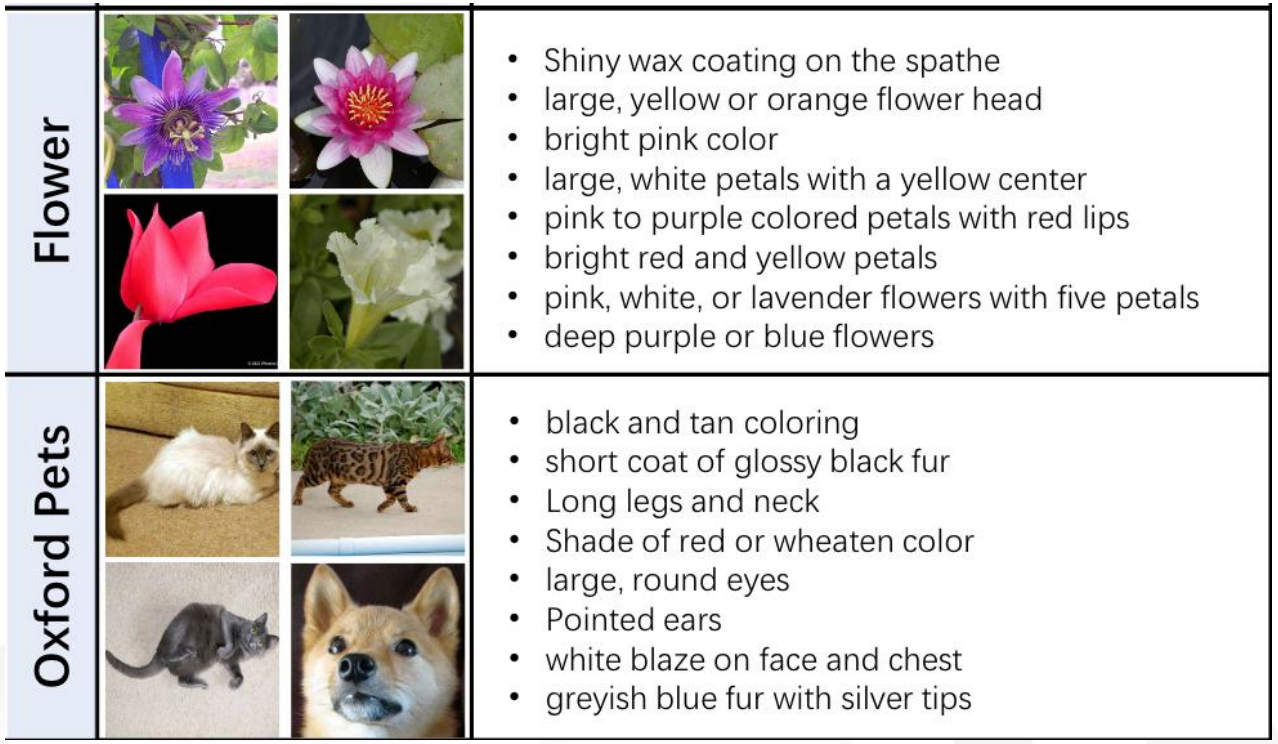
MLLMs for Cognitive Science

1. To integrate multiple data modalities
2. To mimic cognitive process of humans
3. To offer new tools for explaining human behavior.

An example to explain the classification results using multimodal models

The image class ‘flower’ (detected by AI models) has some shared attributes.

The same tool can be also used to uncover the concept representation in humans.



Yan, A., Wang, et al. (2023). *Learning Concise and Descriptive Attributes for Visual Recognition*

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

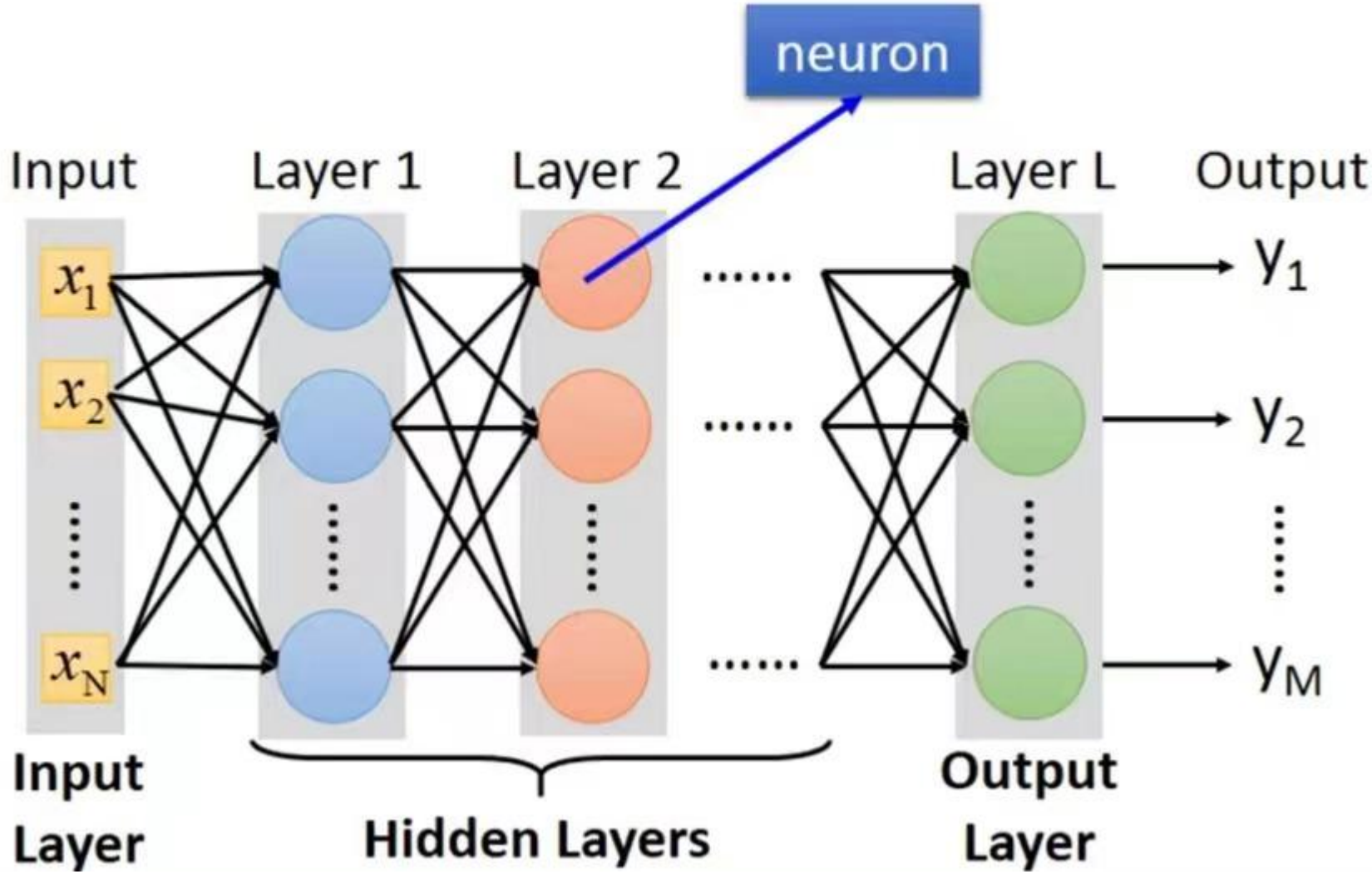
神经计算与控制实验室
Neural Computing & Control Lab

Convolutional Neural Network

Latent representation in CNN

Fully Connect Feedforward Network

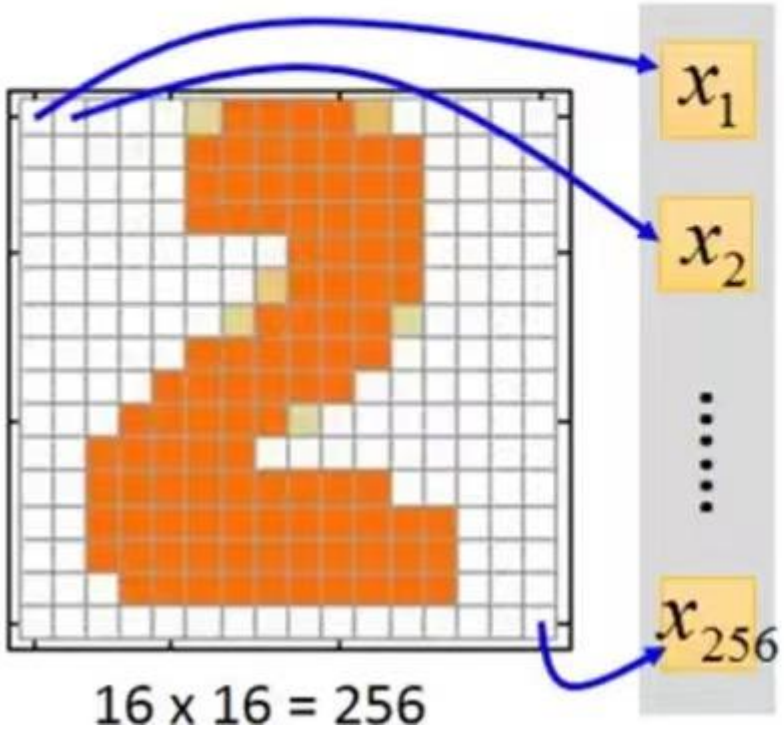
Sigmoid functior $S(x) = \frac{1}{1+e^{-x}}$



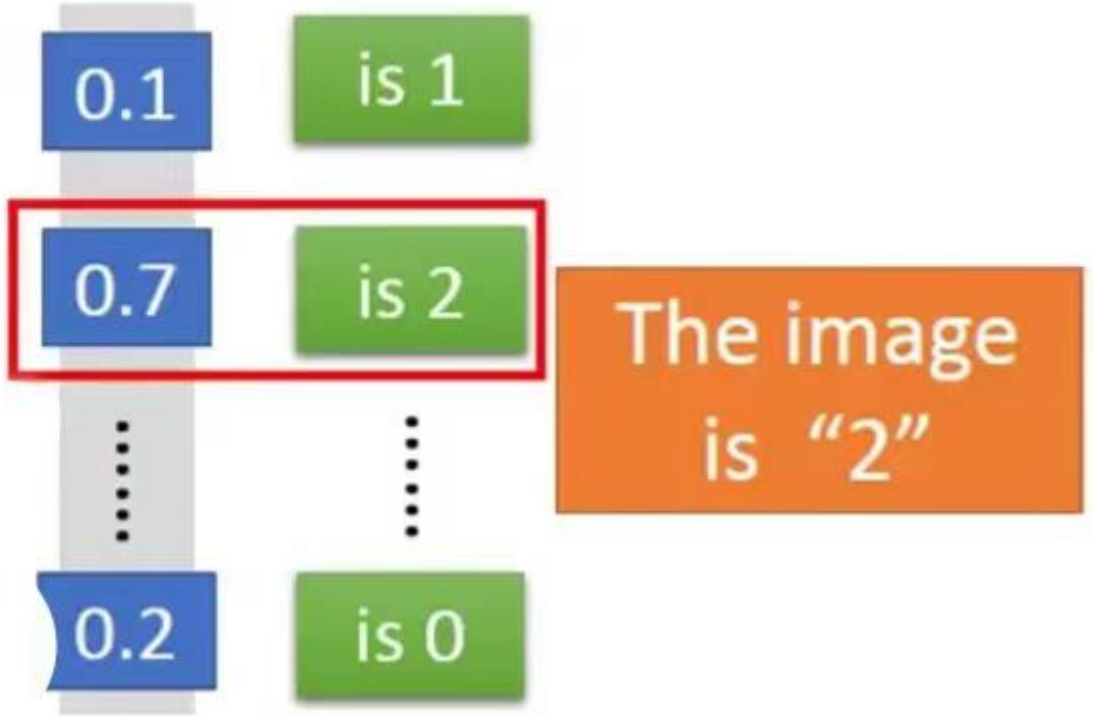
Example Application

softmax functio $S_i = \frac{e^i}{\sum_j e^j}$

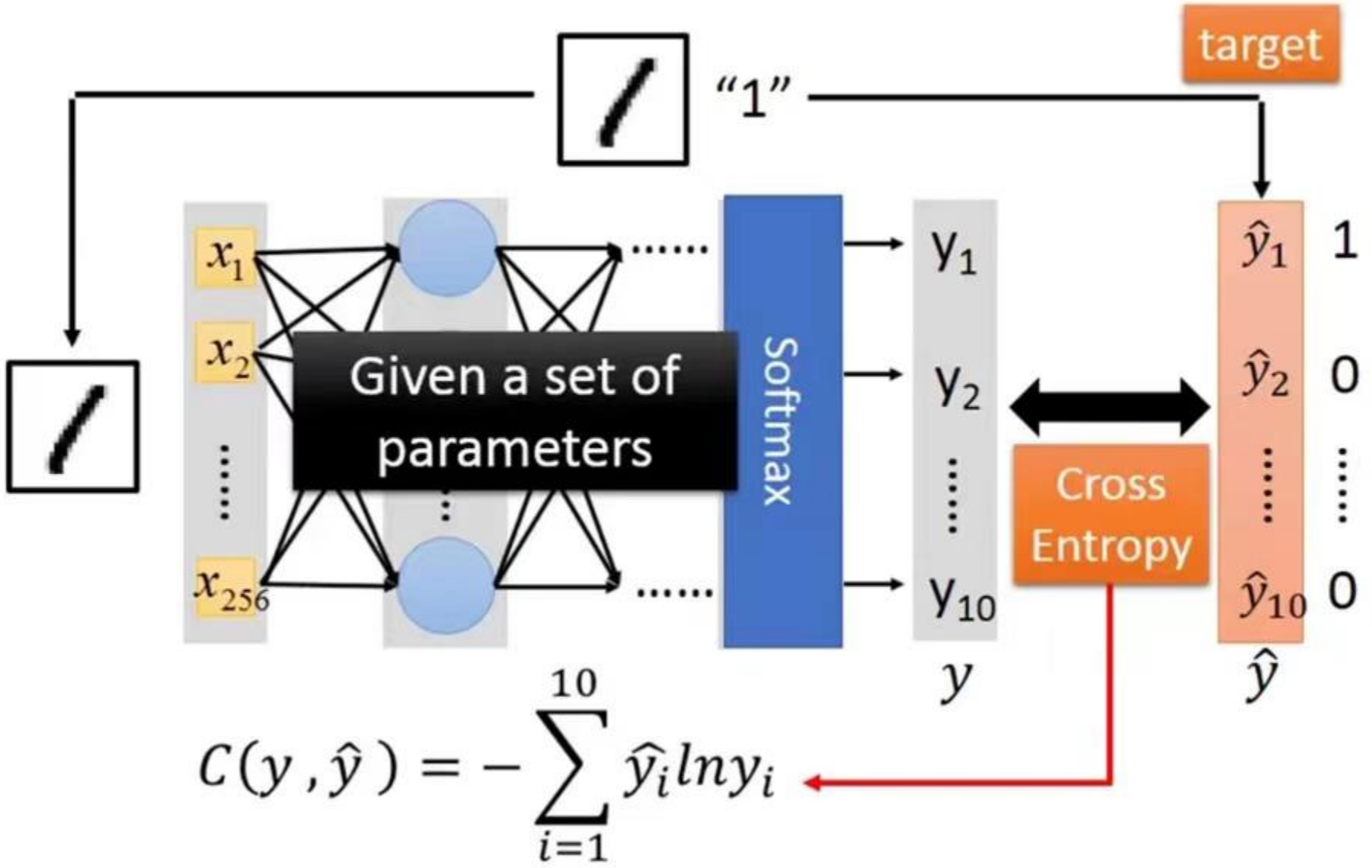
Input



Output

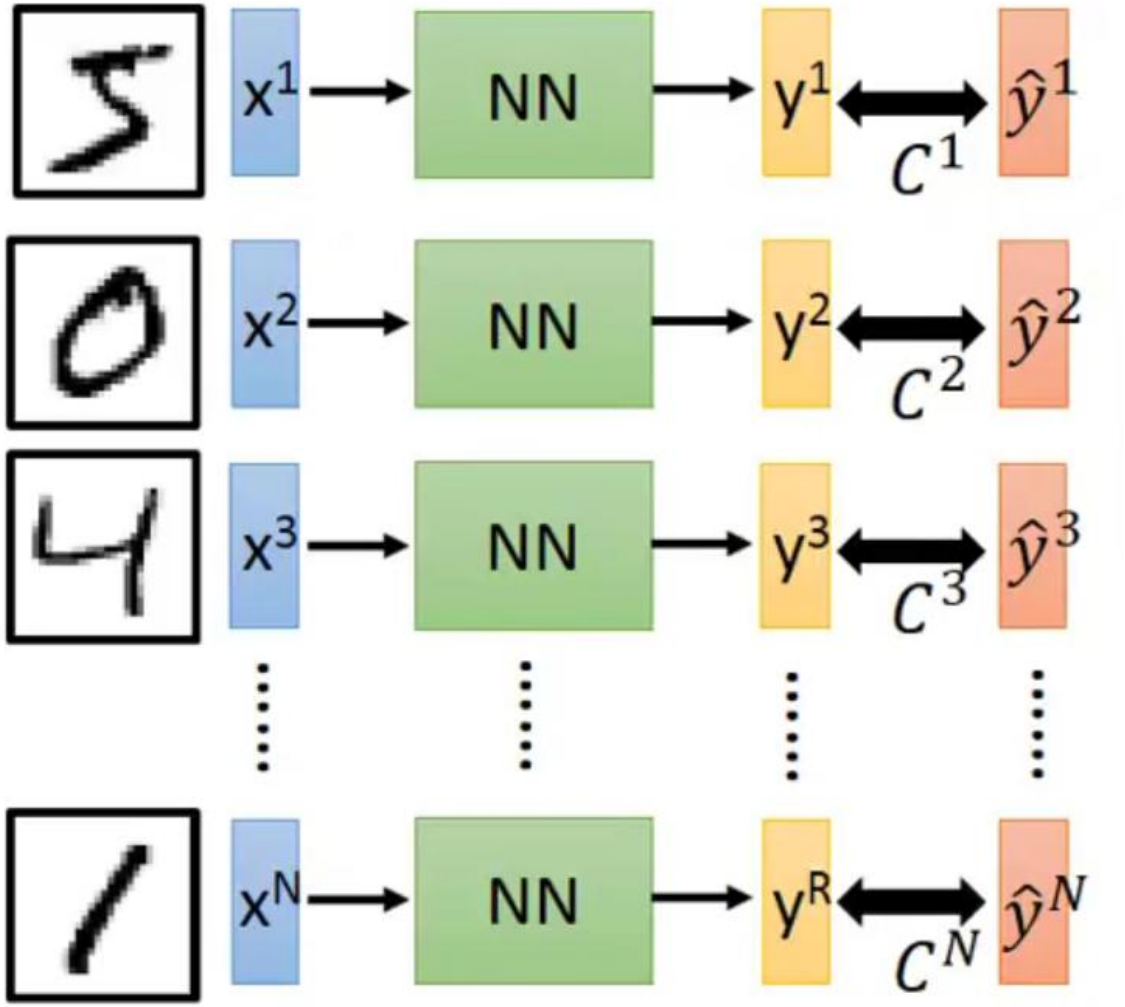


Loss for an Example



Total Loss

For all training data ...



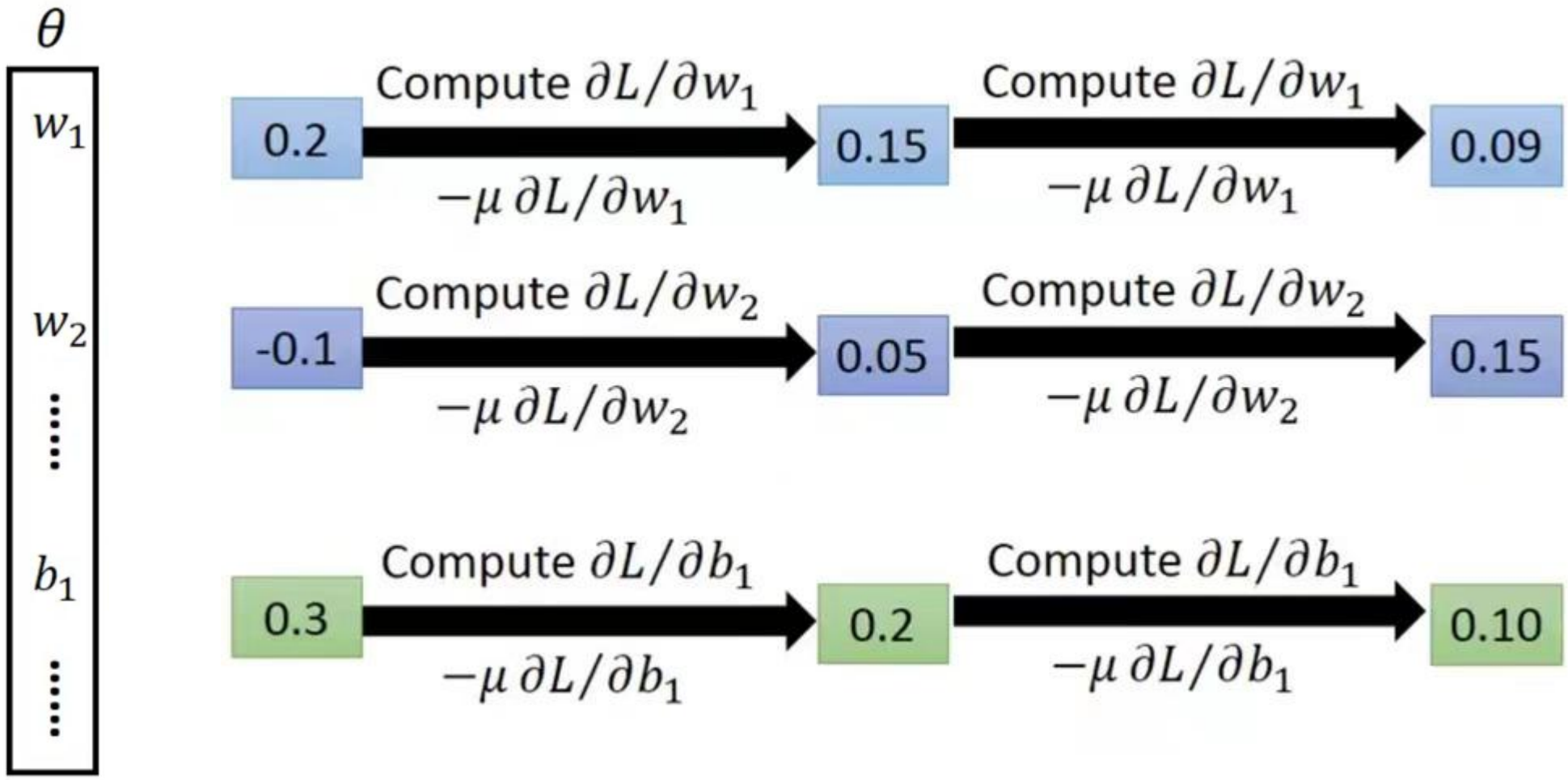
Total Loss:

$$L = \sum_{n=1}^{N} C^n$$

Find **a function in function set** that minimizes total loss L

Find **the network parameters** θ^* that minimizes total loss L

Gradient Descent



This is the “learning” of machines in deep learning.

Image Classification

100 x 100

Model

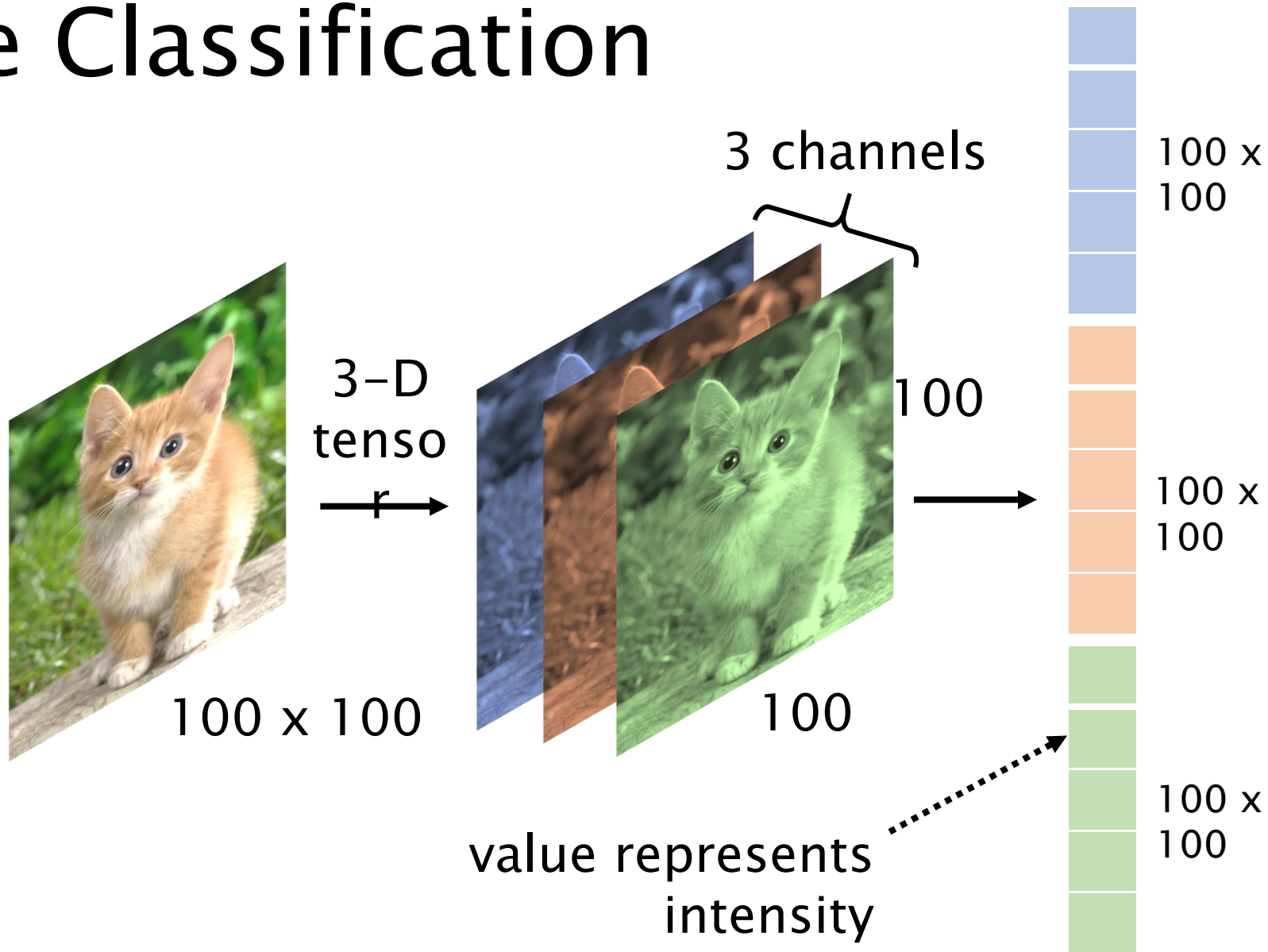
$$\begin{bmatrix} \vdots \\ 0.2 \\ 0.7 \\ 0.1 \\ \vdots \end{bmatrix} \quad y'$$

Cross entropy

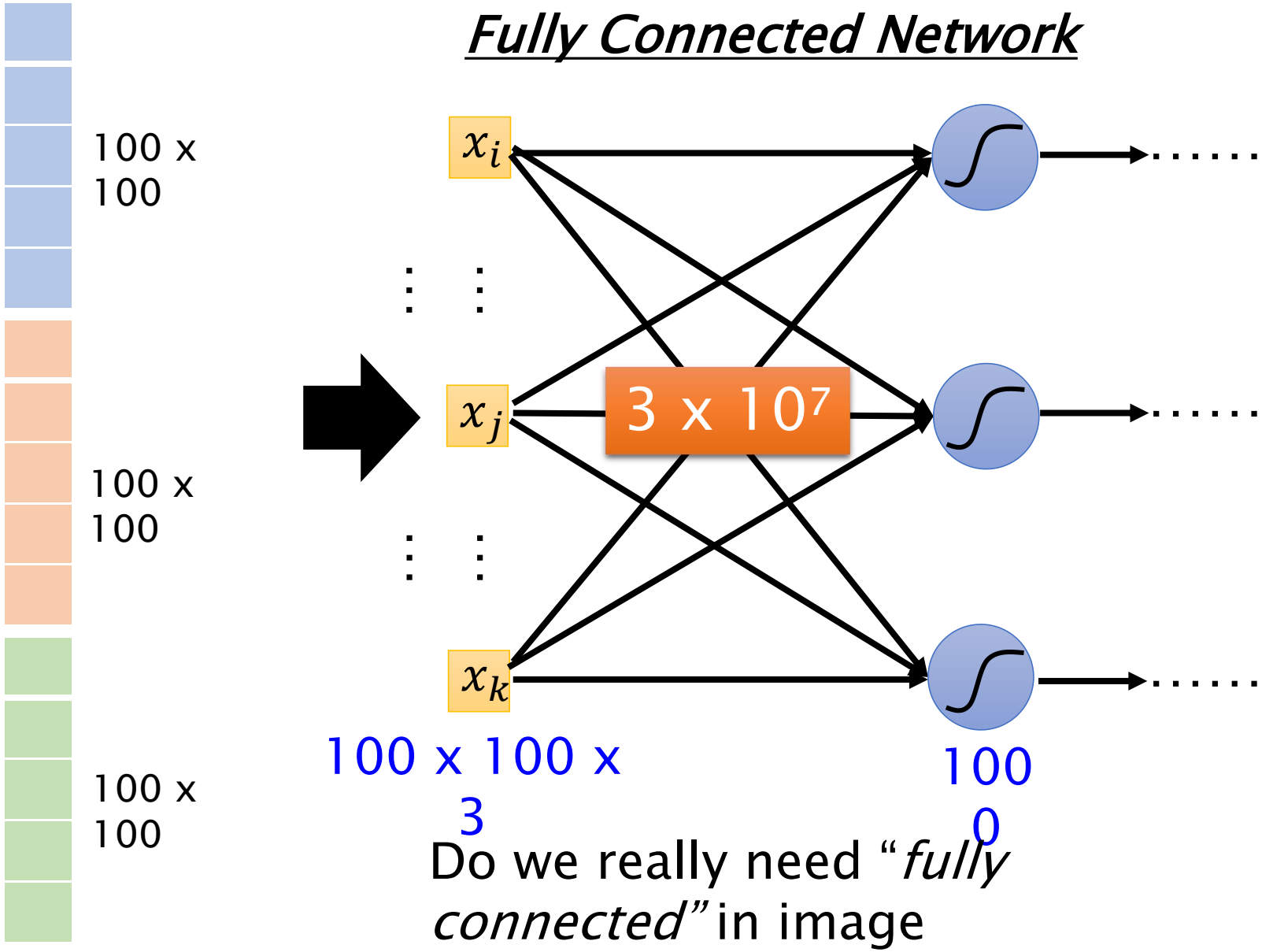
$$\begin{matrix} \\ \text{dog} \\ \text{cat} \\ \text{tree} \\ \\ \end{matrix} \begin{bmatrix} \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \quad \hat{y}$$

(All the images to be classified have the same size.)

Image Classification



Fully Connected Network



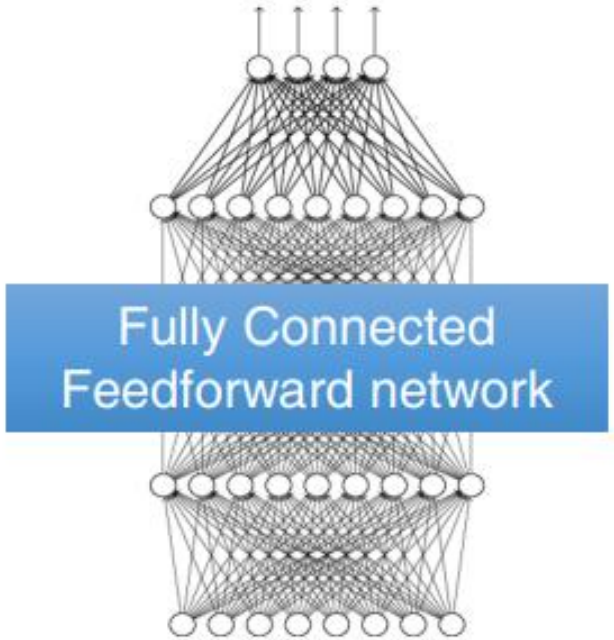
Do we really need "*fully connected*" in image processing?

Convolutional Layer

Neuron Version Story	***Filter Version Story***
Each neuron only considers a receptive field.	There are a set of filters detecting small patterns.
The neurons with different receptive fields share the parameters.	Each filter convolves over the input image.

They are the same story.

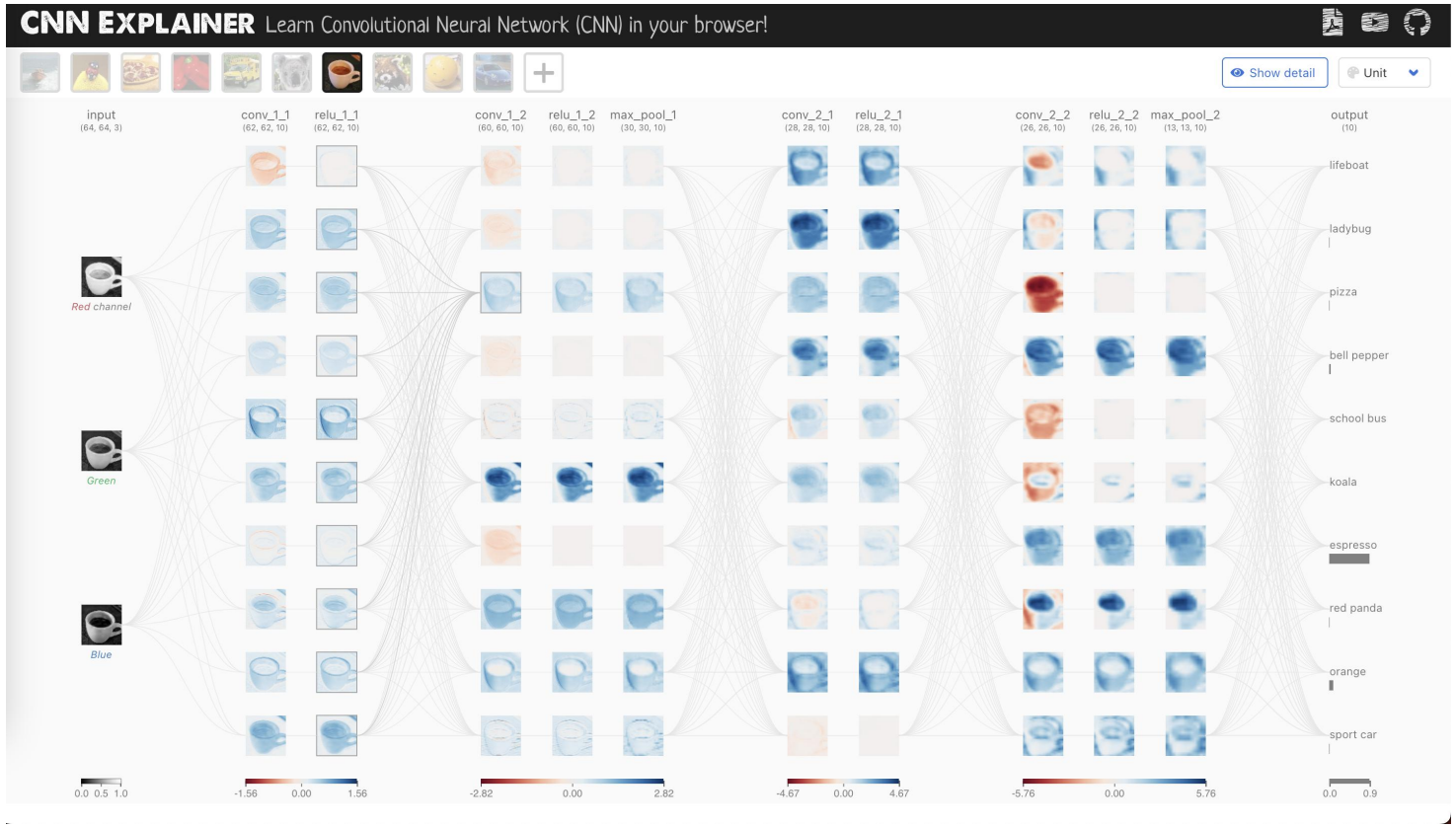
The whole CNN



CNN explainer: Learn CNN in your browser

website: https://poloclub.github.io/cnn-explainer/

github: https://github.com/poloclub/cnn-explainer



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

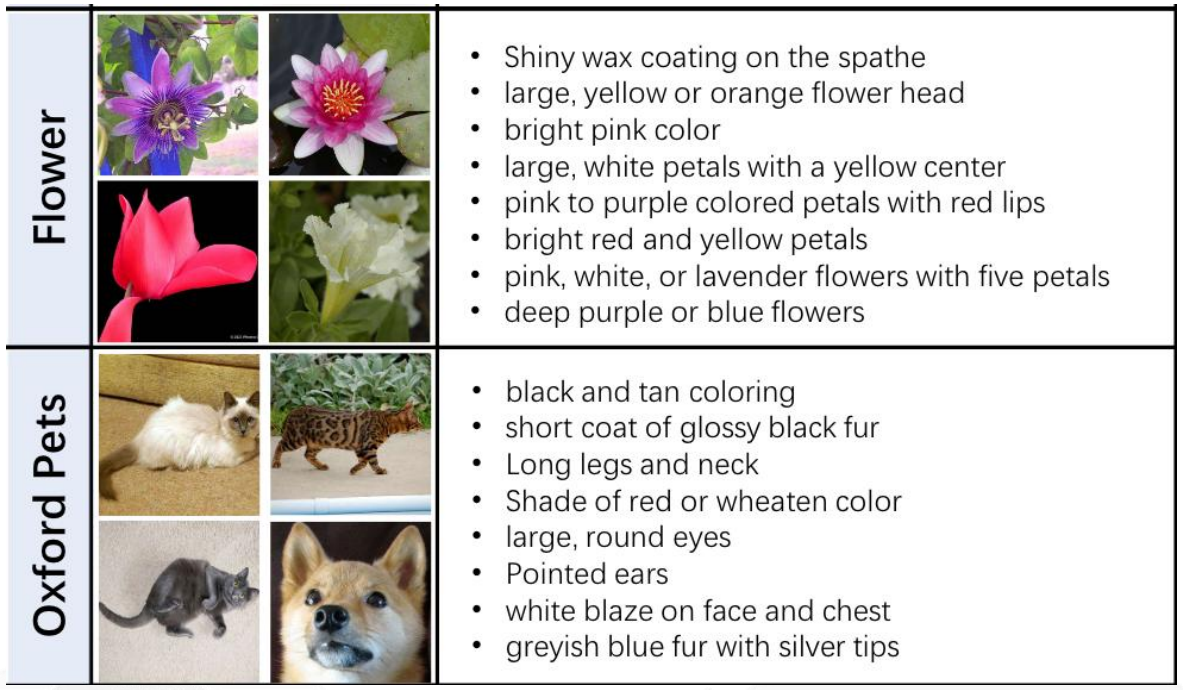
神经计算与控制实验室
Neural Computing & Control Lab

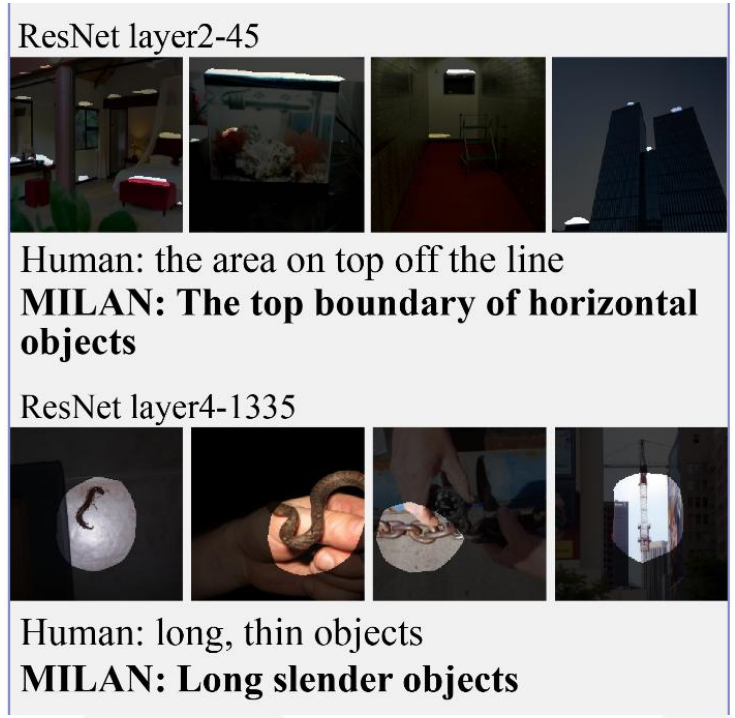
Concept Bottleneck Model

A brief introduction and its applications

Language-guided interpretability

Using language to provide interpretability of AI models (e.g., classifications, neural activities, etc.)

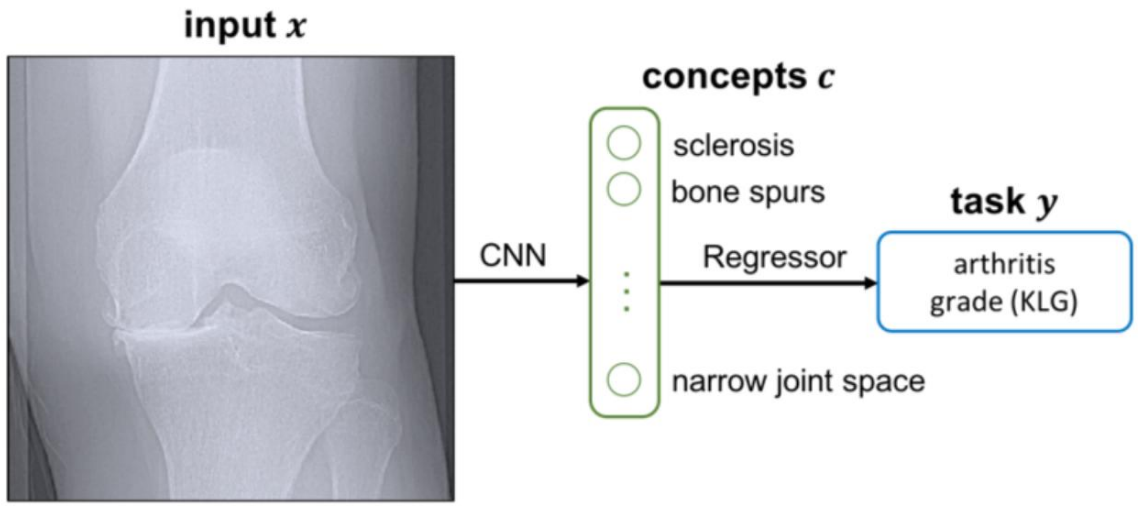




Hernandez, Evan, et al. "Natural language descriptions of deep visual features." ICLR. 2021.

Compared with end-to-end AI models

- We don't know why the model makes decisions
- When the model makes incorrect decisions, it's hard for us to find the reason and make changes



Would it be helpful if we introduced concepts to the model?

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Models

- Concept Bottleneck Model (ICML 2020)
- CLIP-DISSECT (ICLR 2023)
- Label-Free Concept Bottleneck Models (ICLR 2023)

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Oikarinen, T., & Weng, T.-W. (2023). *CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks* (arXiv:2204.10965). arXiv. http://arxiv.org/abs/2204.10965

Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T.-W. (2023). *Label-Free Concept Bottleneck Models* (arXiv:2304.06129). arXiv. http://arxiv.org/abs/2304.06129

Motivations of Concept Bottleneck Models (CBM)

Interpretability

We aim to understand the processes for the AI model's decisions, enabling humans to *comprehend* the underlying mechanisms while potentially enhancing robustness.

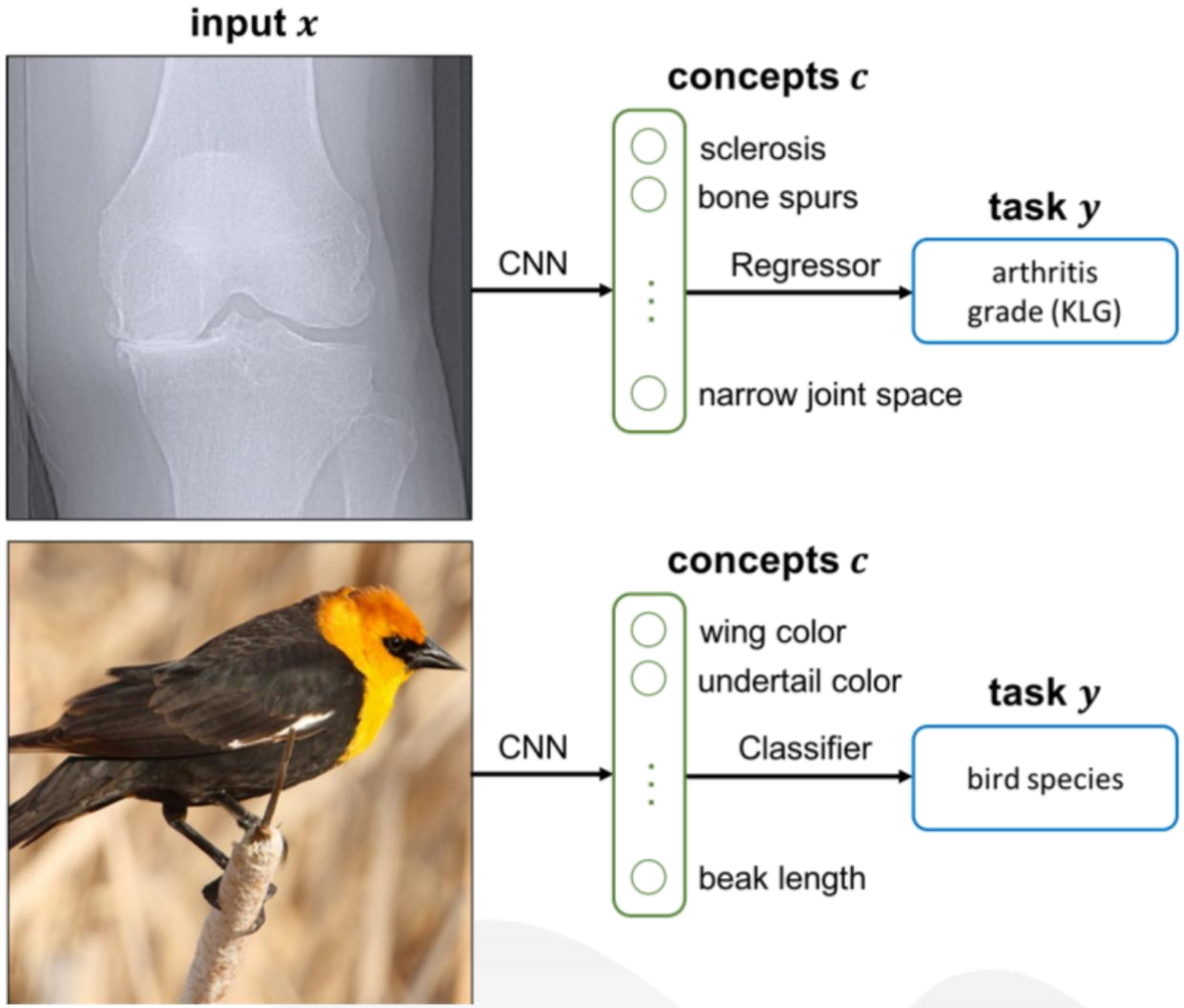
Interventions

When the AI model makes a **mistake**, we would like to know the reasons for the model's error *at the conceptual level*. It offers an interface for **human experts to intervene and guide the model.**

Task accuracy

Lastly, such an interpretable AI model shall not degrade too much, compared to the end-to-end AI models.

Structure of CBM



Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

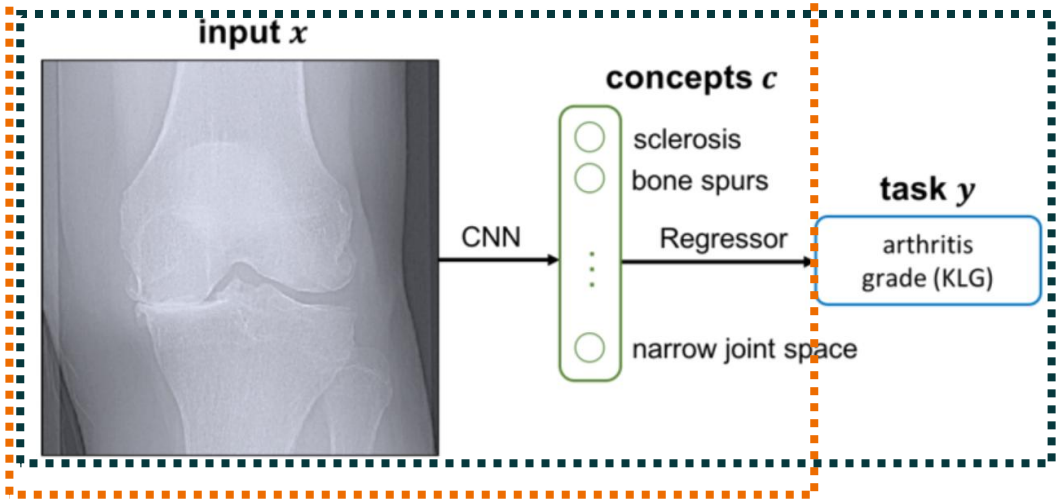
Different learning strategy of CBM

1. The *independent bottleneck* learns $\hat{f}$ and $\hat{g}$ independently: $\hat{f} = \arg\min_f \sum_i L_Y(f(c^{(i)}); y^{(i)})$, and $\hat{g} = \arg\min_g \sum_{i,j} L_{C_j}(g_j(x^{(i)}); c_j^{(i)})$. While $\hat{f}$ is trained using the true c, at test time it still takes $\hat{g}(x)$ as input.

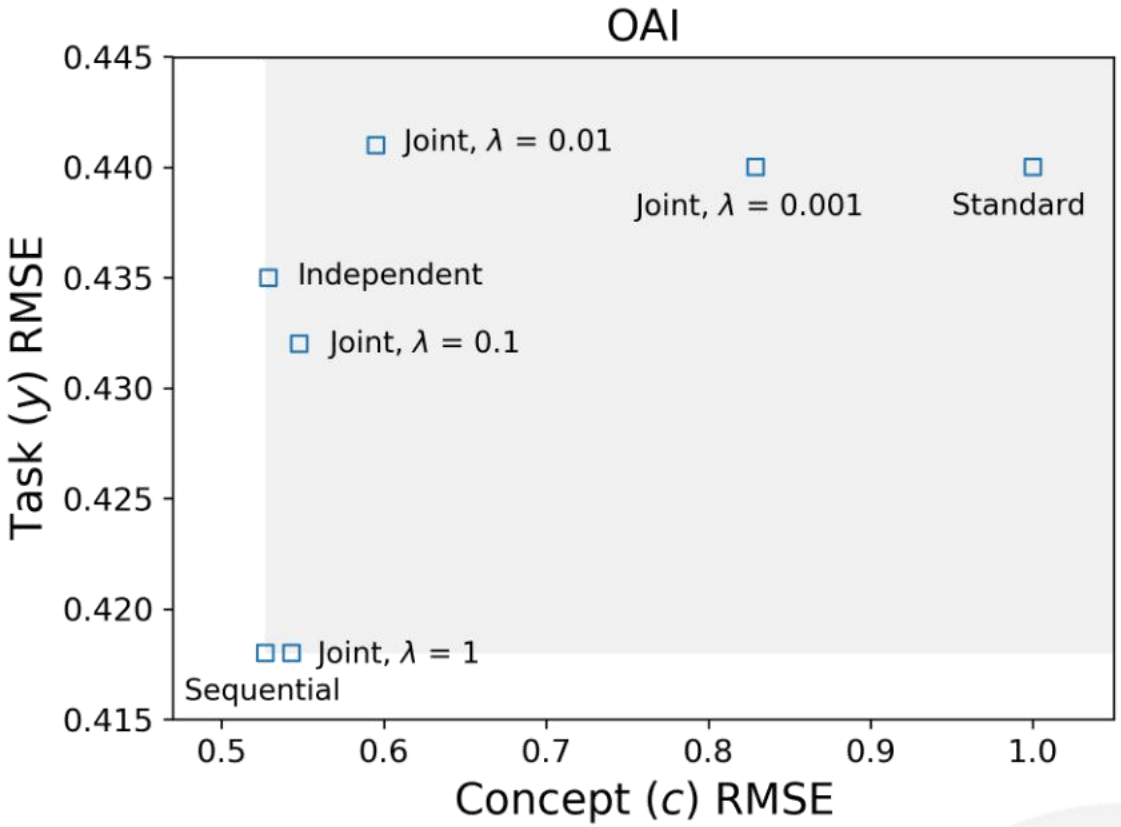
2. The *sequential bottleneck* first learns $\hat{g}$ in the same way as above. It then uses the concept predictions $\hat{g}(x)$ to learn $\hat{f} = \arg\min_f \sum_i L_Y(f(\hat{g}(x^{(i)})); y^{(i)})$.

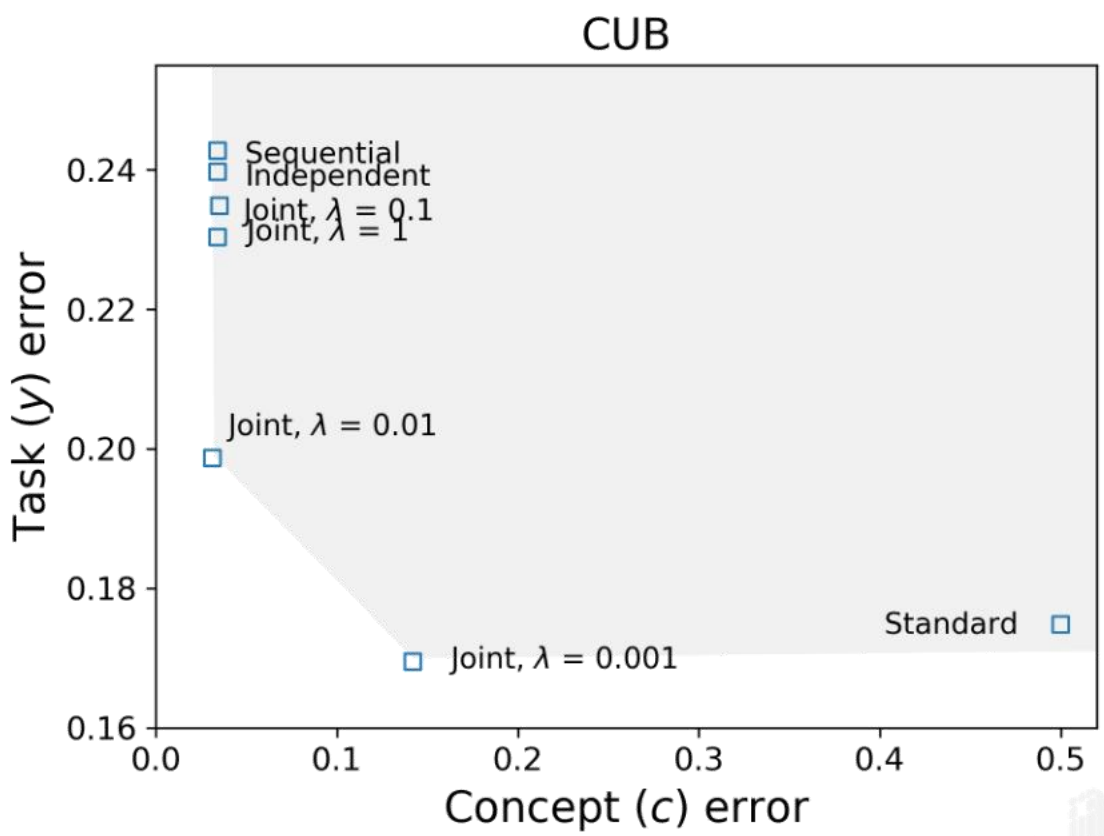
3. The *joint bottleneck* minimizes the weighted sum $\hat{f}, \hat{g} = \arg\min_{f,g} \sum_i \left[L_Y(f(g(x^{(i)})); y^{(i)}) + \sum_j \lambda L_{C_j}(g(x^{(i)}); c^{(i)}) \right]$ for some $\lambda > 0$.

4. The *standard model* ignores concepts and directly minimizes $\hat{f}, \hat{g} = \arg\min_{f,g} \sum_i L_Y(f(g(x^{(i)})); y^{(i)})$.

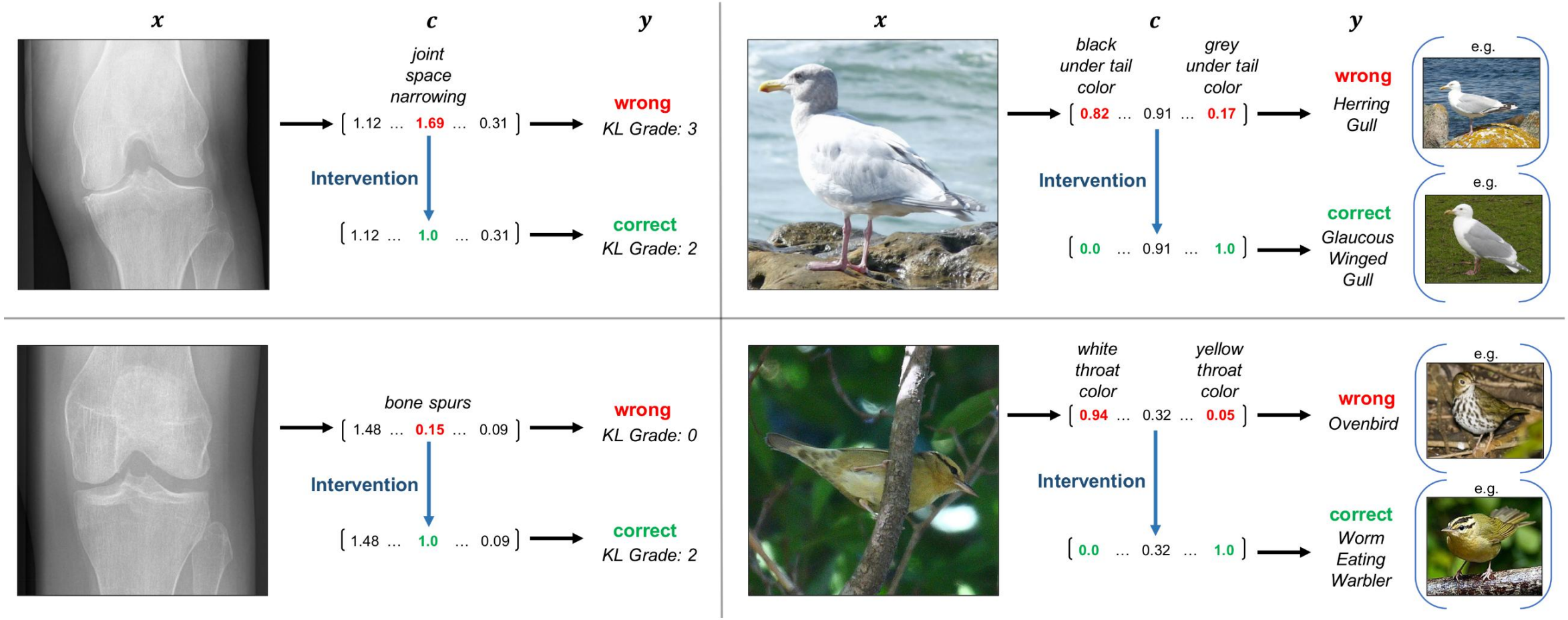


CBM do well on both task and concept prediction for Different learning strategy



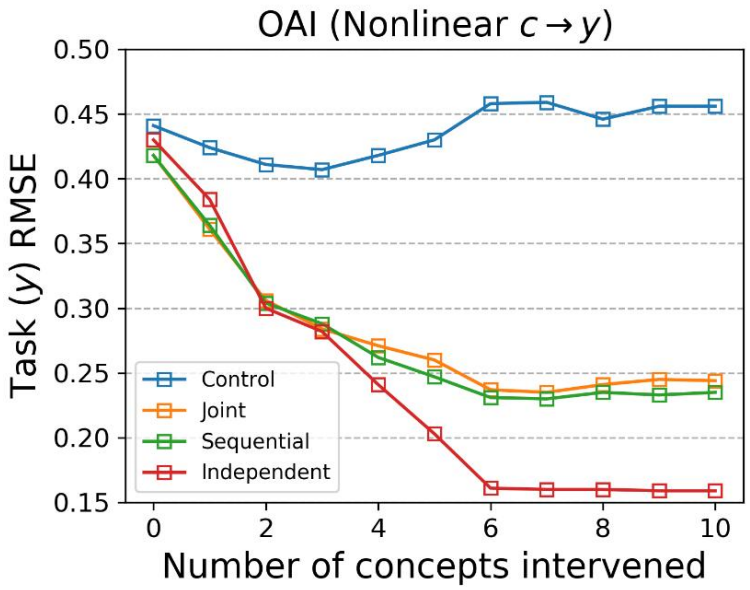


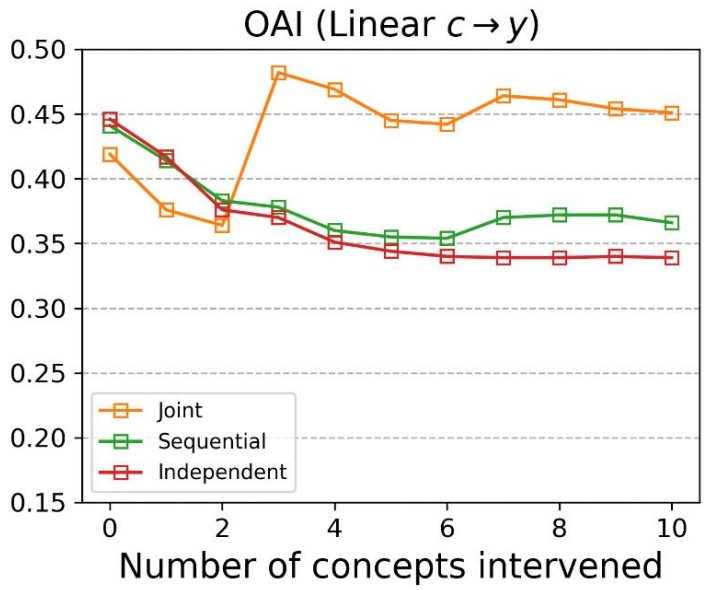
Successful examples of test-time intervention

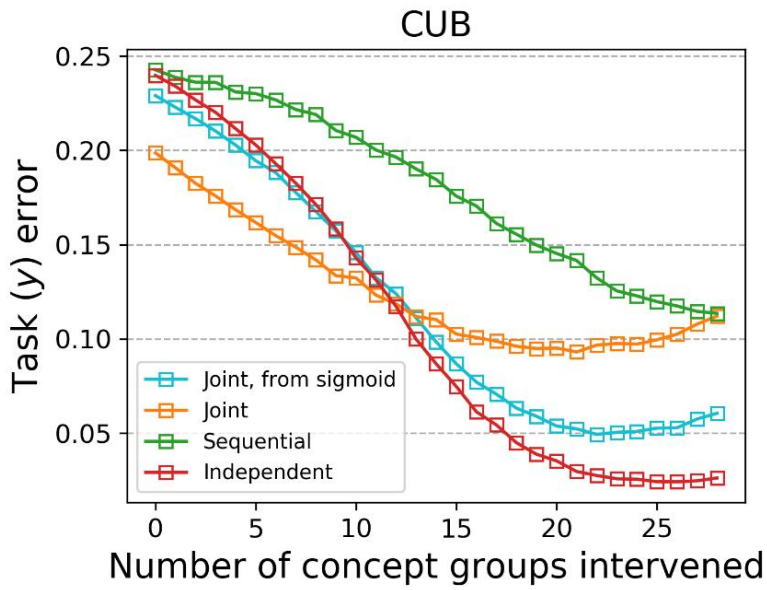


Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Intervention substantially improves task accuracy







Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Robustness to background shifts

Train:
Black-billed Cuckoo on
Forest Path background

Test:
Black-billed Cuckoo on
Coffee Shop background

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Limitations of vanilla CBM

- They need to collect **labeled data** for each of the **predefined concepts**, which is time consuming and labor intensive

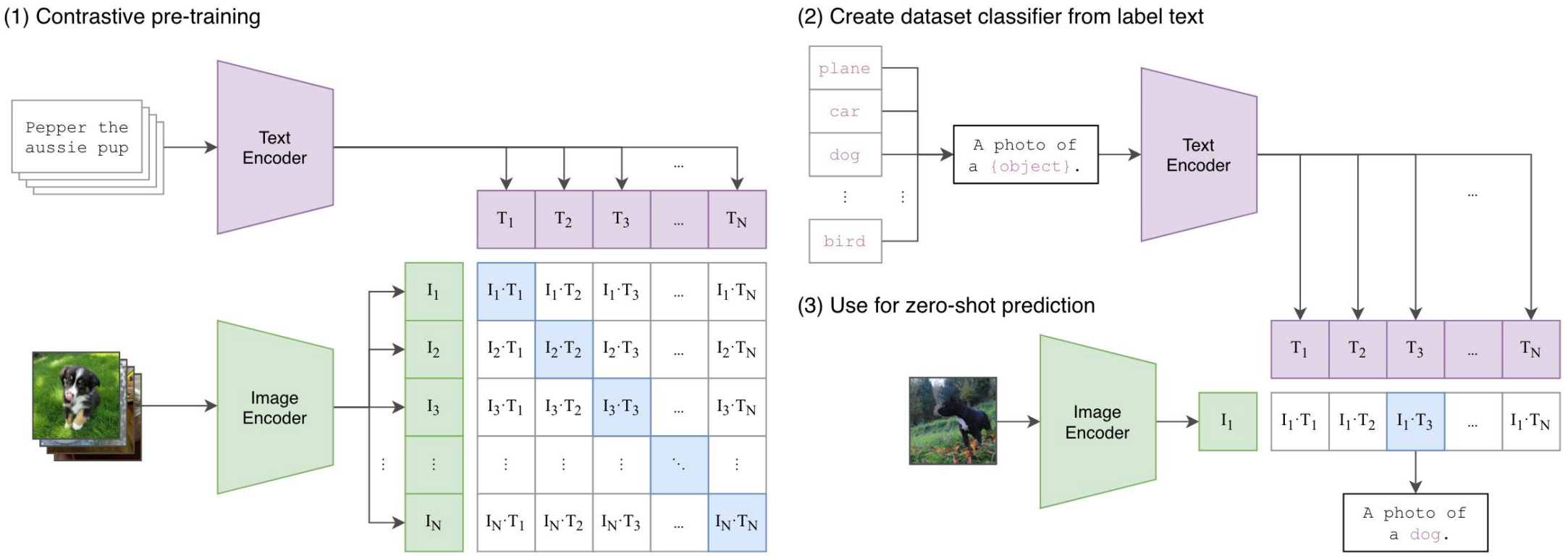
- The **accuracy** of a CBM is often significantly *lower* than that of a standard neural network, especially on more complex datasets

Would it be helpful if we introduced multimodal model?

Oikarinen, T., & Weng, T.-W. (2023). *CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks* (arXiv:2204.10965). arXiv. http://arxiv.org/abs/2204.10965
Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T.-W. (2023). *Label-Free Concept Bottleneck Models* (arXiv:2304.06129). arXiv.

Interpretability by pretrained visual-language model

CLIP: visual-language contrastive learning



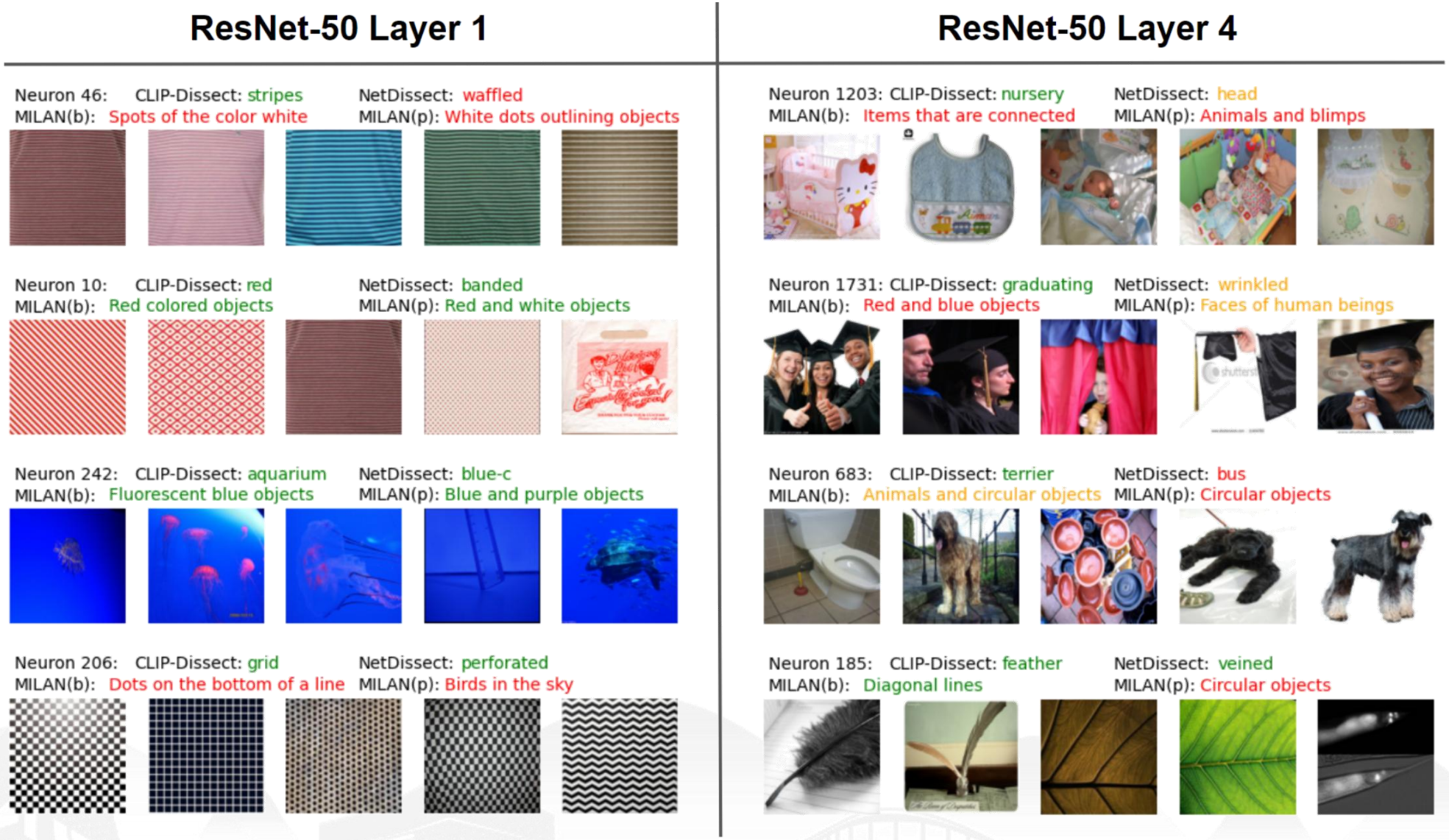
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Models

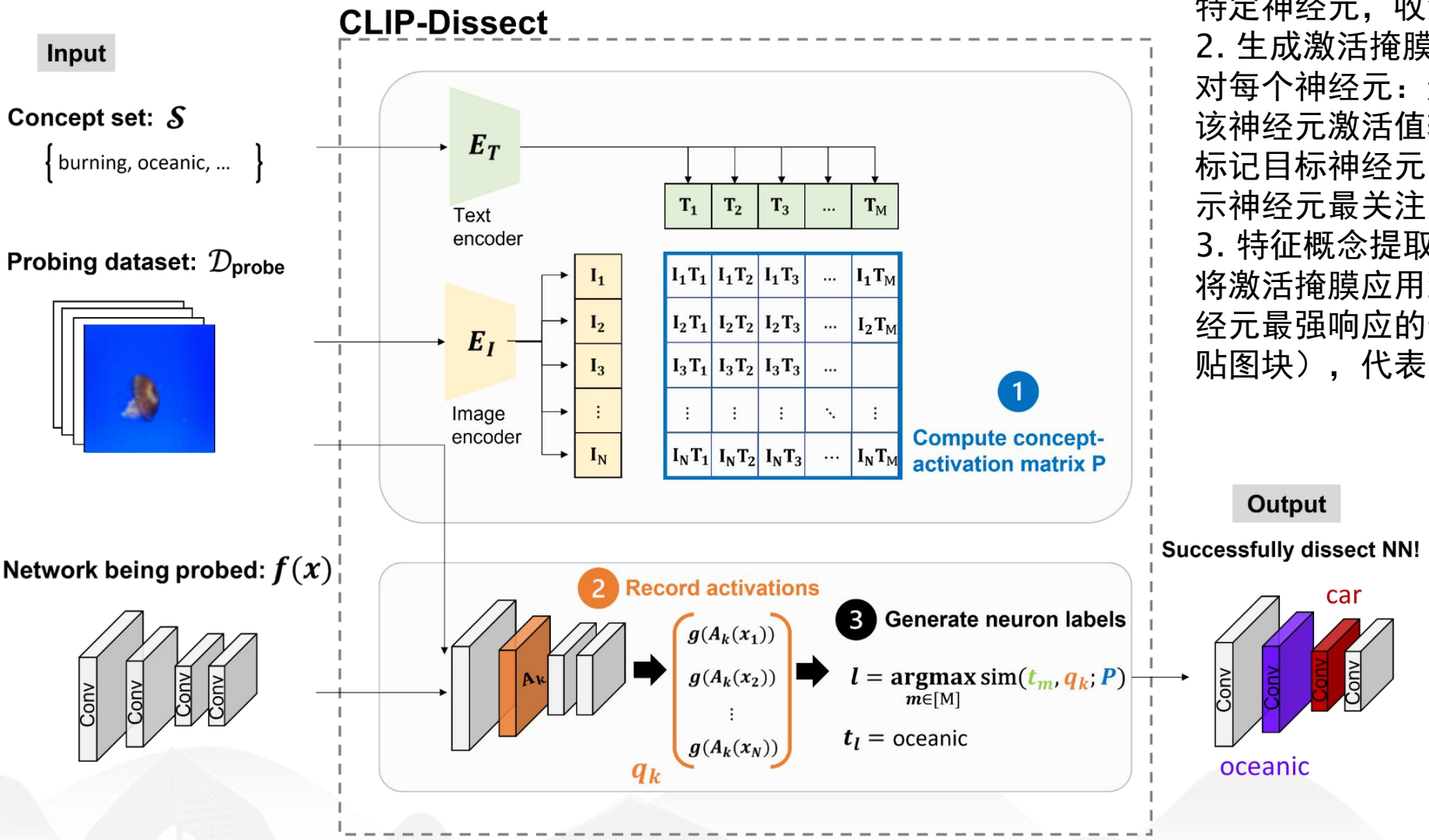
- Concept Bottleneck Model (ICML 2020)
- CLIP-DISSECT (ICLR 2023)
  - generalized concept-based model (neuron-level interpretability)
  - use CLIP to provides accurate descriptions for neurons
- Label-Free Concept Bottleneck Models (ICLR 2023)
  - transform any neural network into an interpretable CBM without labeled concept data

Oikarinen, T., & Weng, T.-W. (2023). *CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks* (arXiv:2204.10965). arXiv. http://arxiv.org/abs/2204.10965

Labels generated by CLIP-Dissect



Overview of CLIP-Dissect



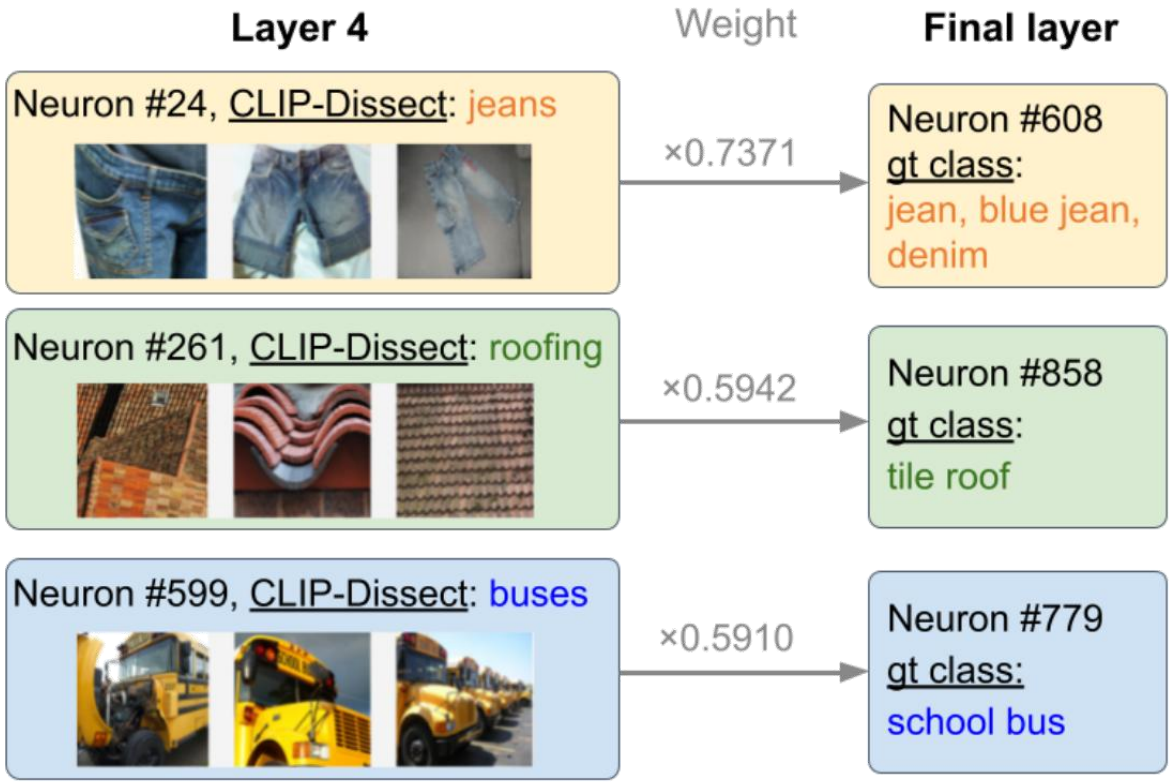
1. 神经元激活提取
对每个输入图像，将图像输入深度视觉模型。提取目标层中每个神经元的激活值。将激活值存储为矩阵，记录神经元对输入图像的响应。对于特定神经元，收集其在整个图像集上的激活分布。

2. 生成激活掩膜
对每个神经元：通过激活值设置阈值，筛选出对该神经元激活值较高的图像区域。生成激活掩膜，标记目标神经元的“感兴趣区域”。激活掩膜用于指示神经元最关注的图像部分。

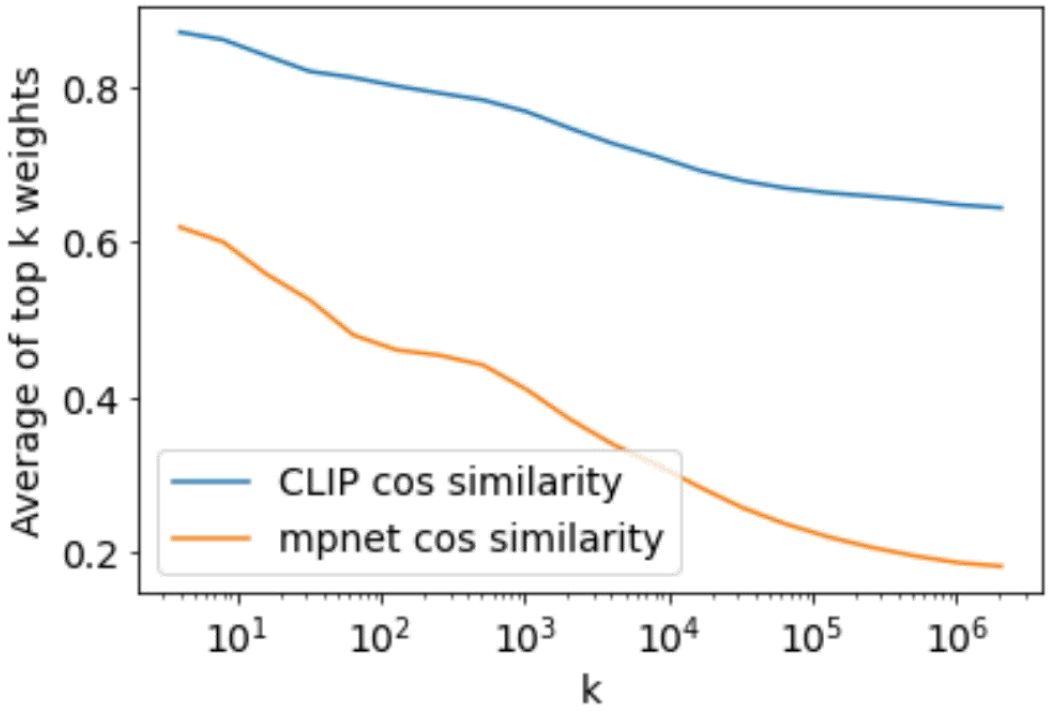
3. 特征概念提取
将激活掩膜应用到图像，裁剪出目标区域（即神经元最强响应的部分）。生成多个裁剪区域（如贴图块），代表了神经元表征的视觉特征。

4. CLIP 模型编码
利用 CLIP 模型的多模态编码能力：对裁剪的激活区域图像输入 CLIP 的视觉编码器，生成图像嵌入。对 CLIP 的文本编码器输入一组潜在描述（例如 "a striped pattern", "a cat", "blue sky"）。
比较图像嵌入与文本嵌入之间的余弦相似度，找出最相关的文本描述。

Use case of CLIP-Dissect



(a) Visualization of 3 highest weights of final layer.



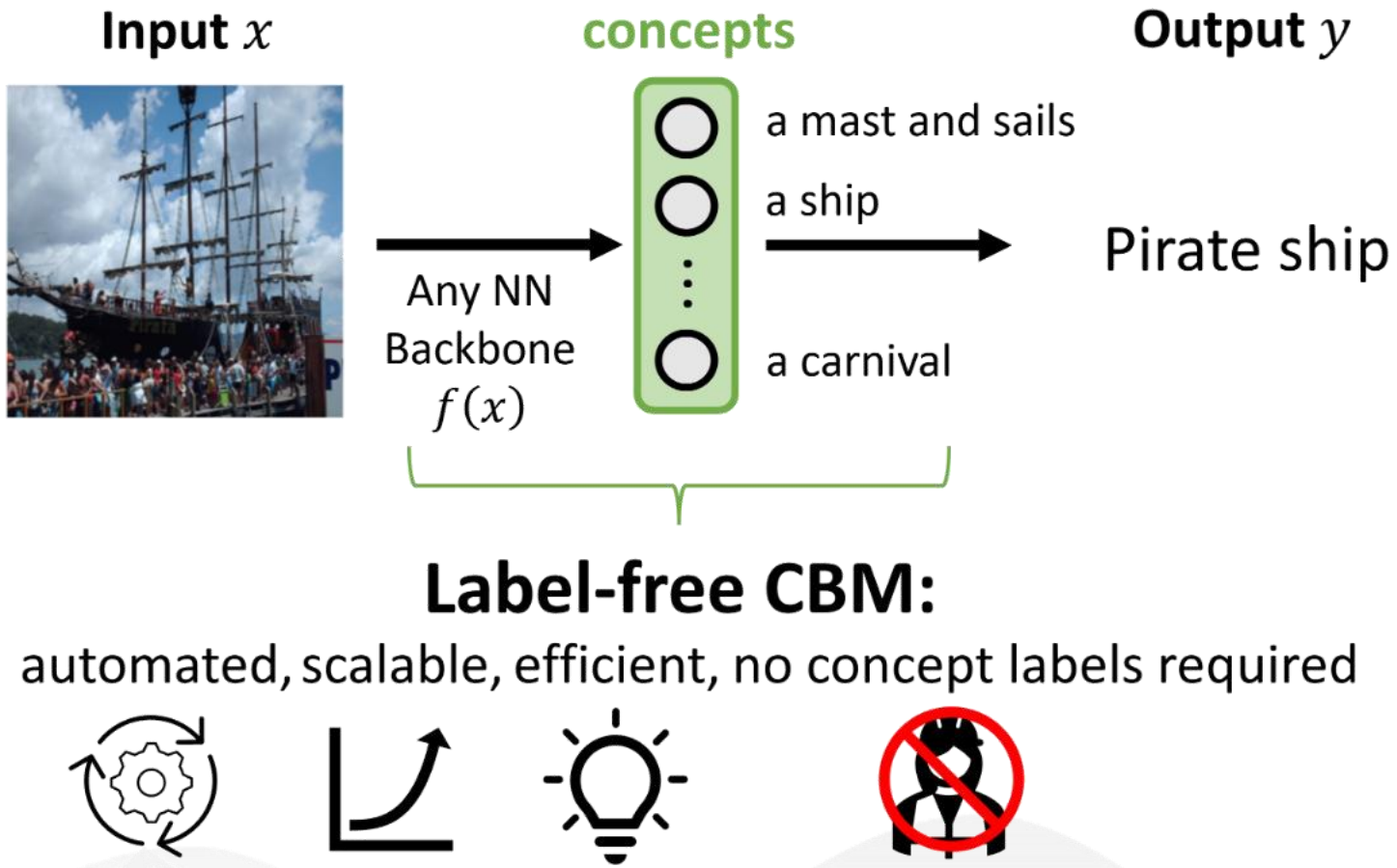
(b) Average cosine similarity between concepts.

Models

- Concept Bottleneck Model (ICML 2020)
- CLIP-DISSECT (ICLR 2023)
  - generalized concept-based model (neuron-level interpretability)
  - use CLIP to provides accurate descriptions for neurons
- Label-Free Concept Bottleneck Models (ICLR 2023)
  - transform any neural network into an interpretable CBM without labeled concept data

Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T.-W. (2023). *Label-Free Concept Bottleneck Models* (arXiv:2304.06129). arXiv. http://arxiv.org/abs/2304.06129

Label-free concept bottleneck models



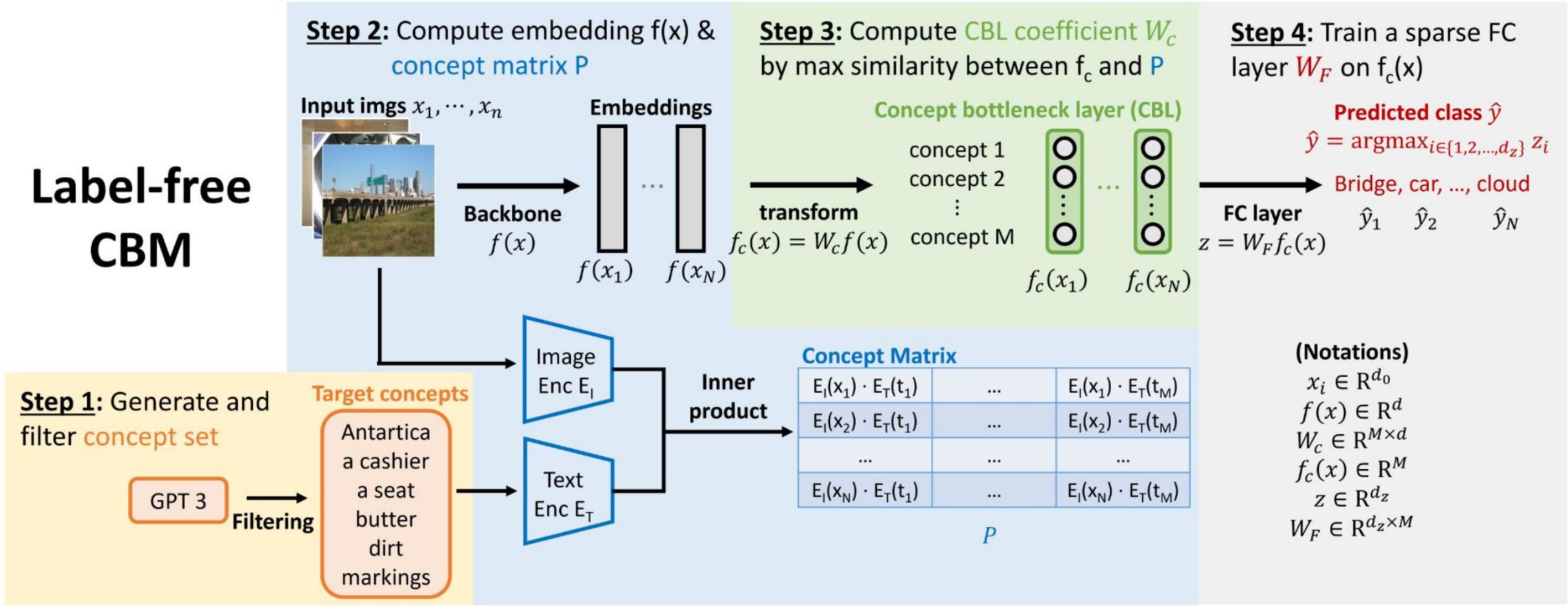
传统的概念瓶颈模型（Concept Bottleneck Models，CBMs）依赖于手动标注的概念标签作为中间表征，这些概念通常是人类可解释的（如“颜色”、“纹理”、“形状”等）。

然而，标注概念标签是昂贵且耗时的，尤其是在规模较大的数据集上。

因此，研究提出了无需人工标注概念标签的无标签概念瓶颈模型（LFCBMs）。

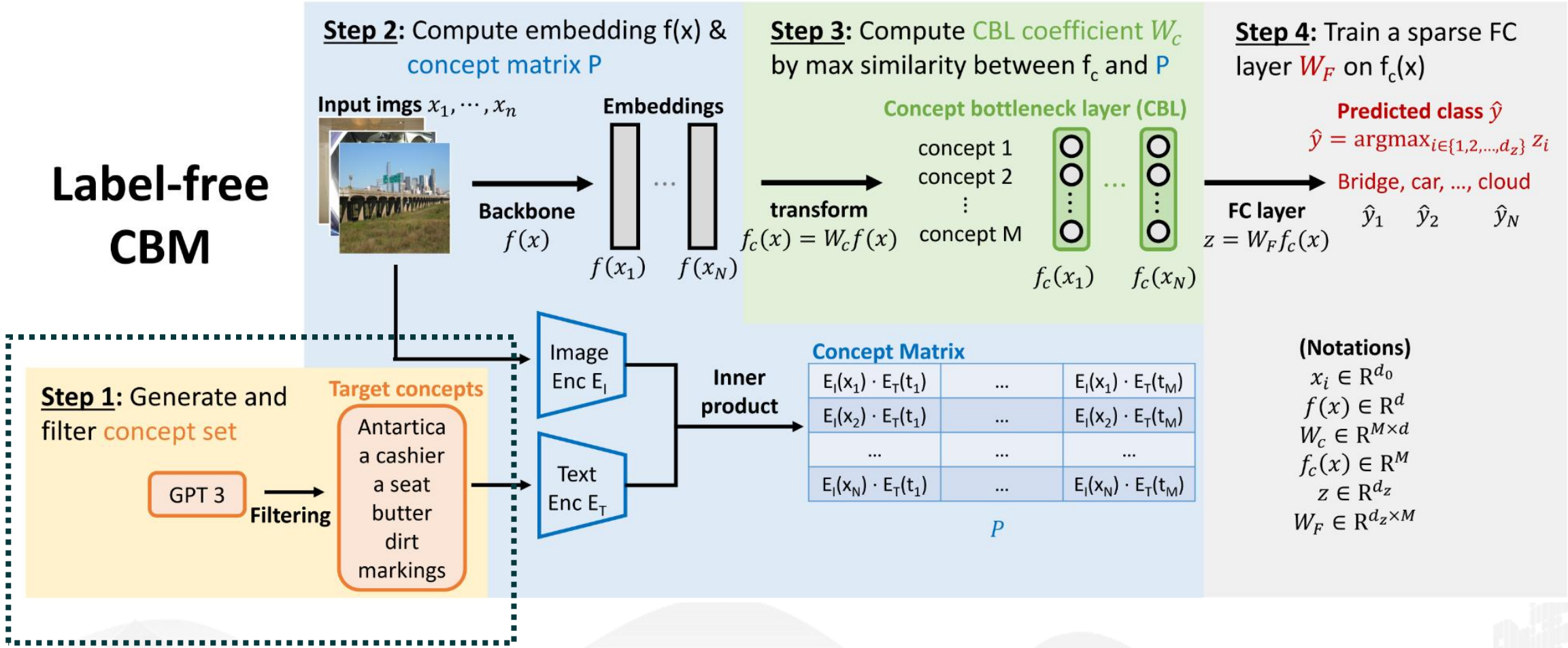
Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T.-W. (2023). *Label-Free Concept Bottleneck Models* (arXiv:2304.06129). arXiv. http://arxiv.org/abs/2304.06129

Overview of pipeline for creating label-free CBM



Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T.-W. (2023). *Label-Free Concept Bottleneck Models* (arXiv:2304.06129). arXiv. http://arxiv.org/abs/2304.06129

Step 1: concept set creation and filtering

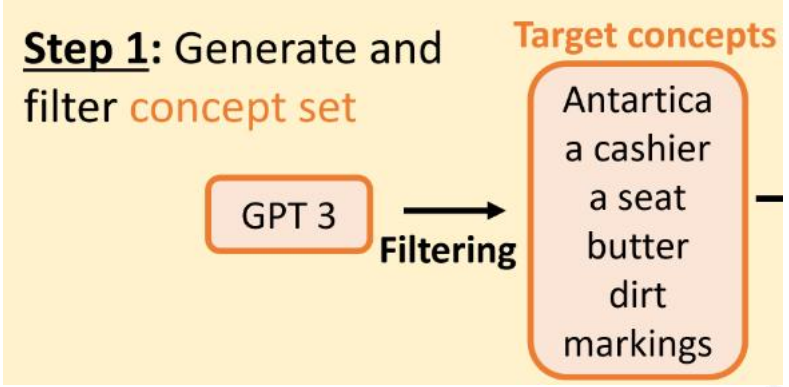


Step 1: concept set creation and filtering

A. Initial concept set creation

Ask GPT-3:

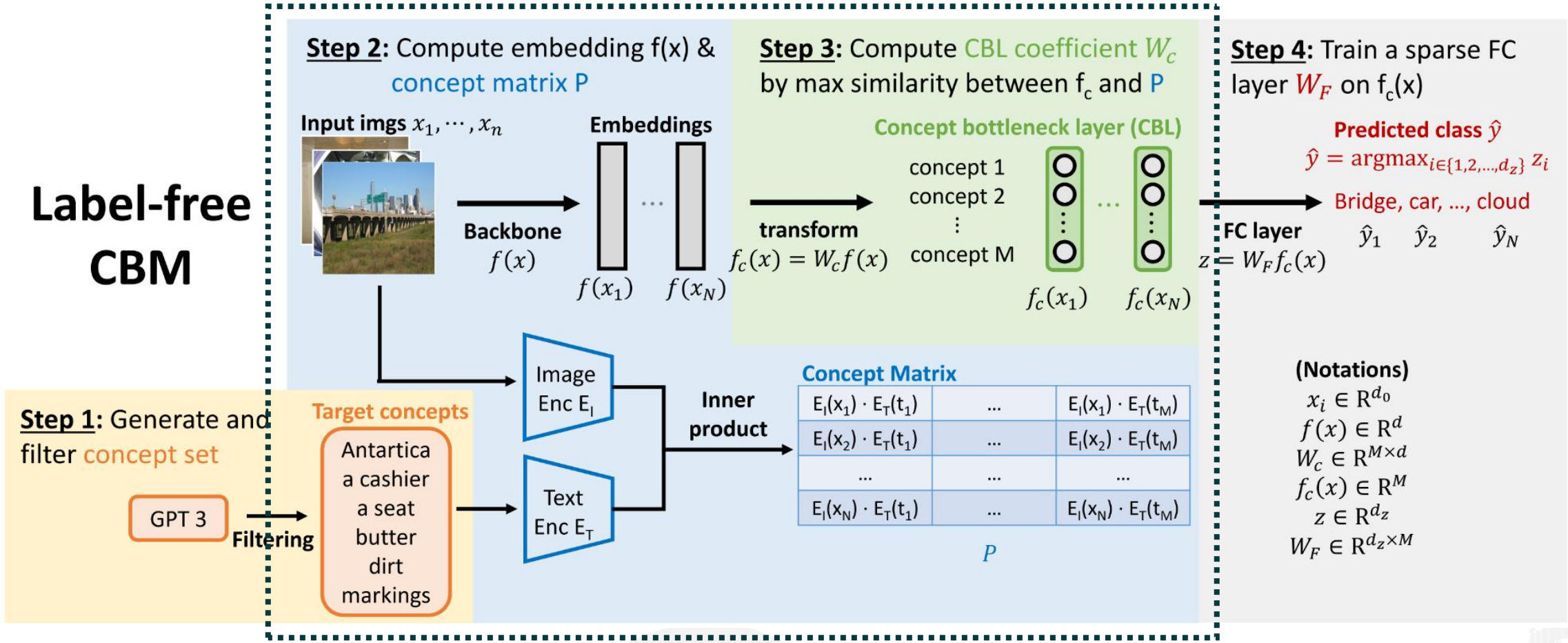
- *List the most important features for recognizing something as a {class}:*
- *List the things most commonly seen around a {class}:*
- *Give superclasses for the word {class}:*



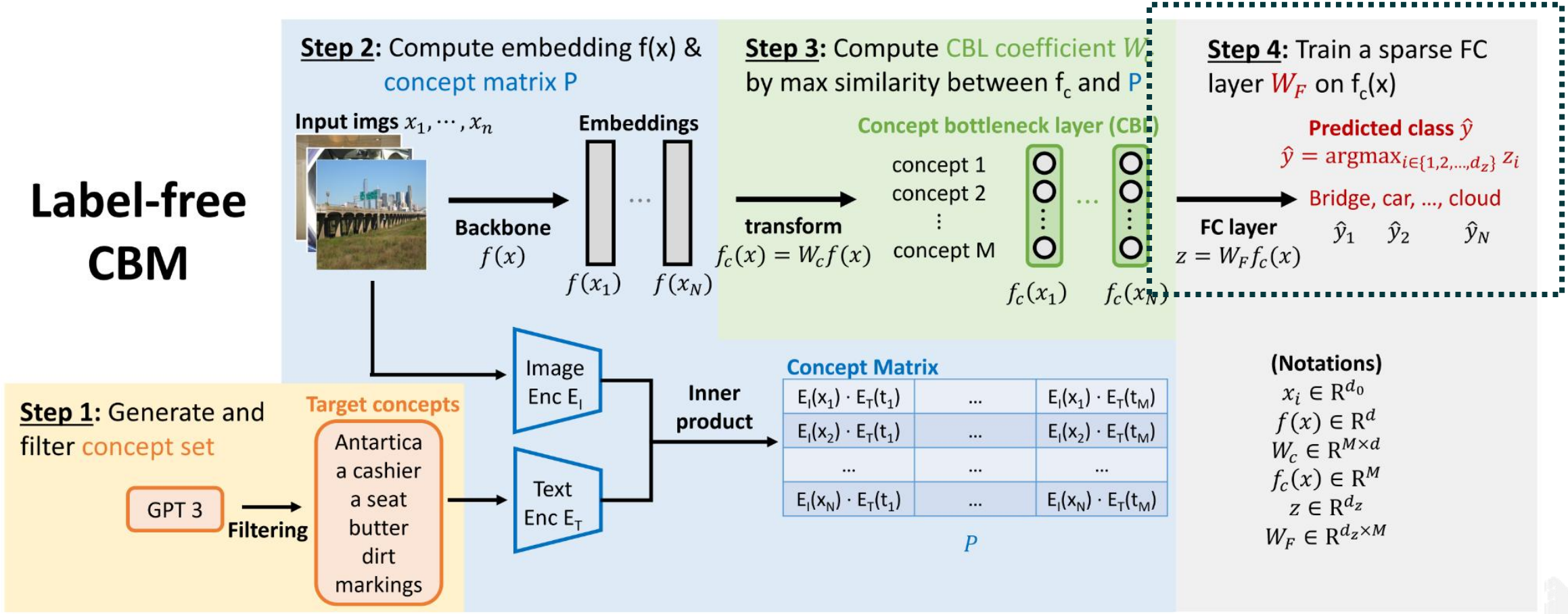
B. Concept set filtering

- *Concept length (<30)*
- *Remove concepts too similar to classes*
- *Remove concepts too similar to each other*
- *Remove concepts not present in training data*
- *Remove concepts we can't project accurately*

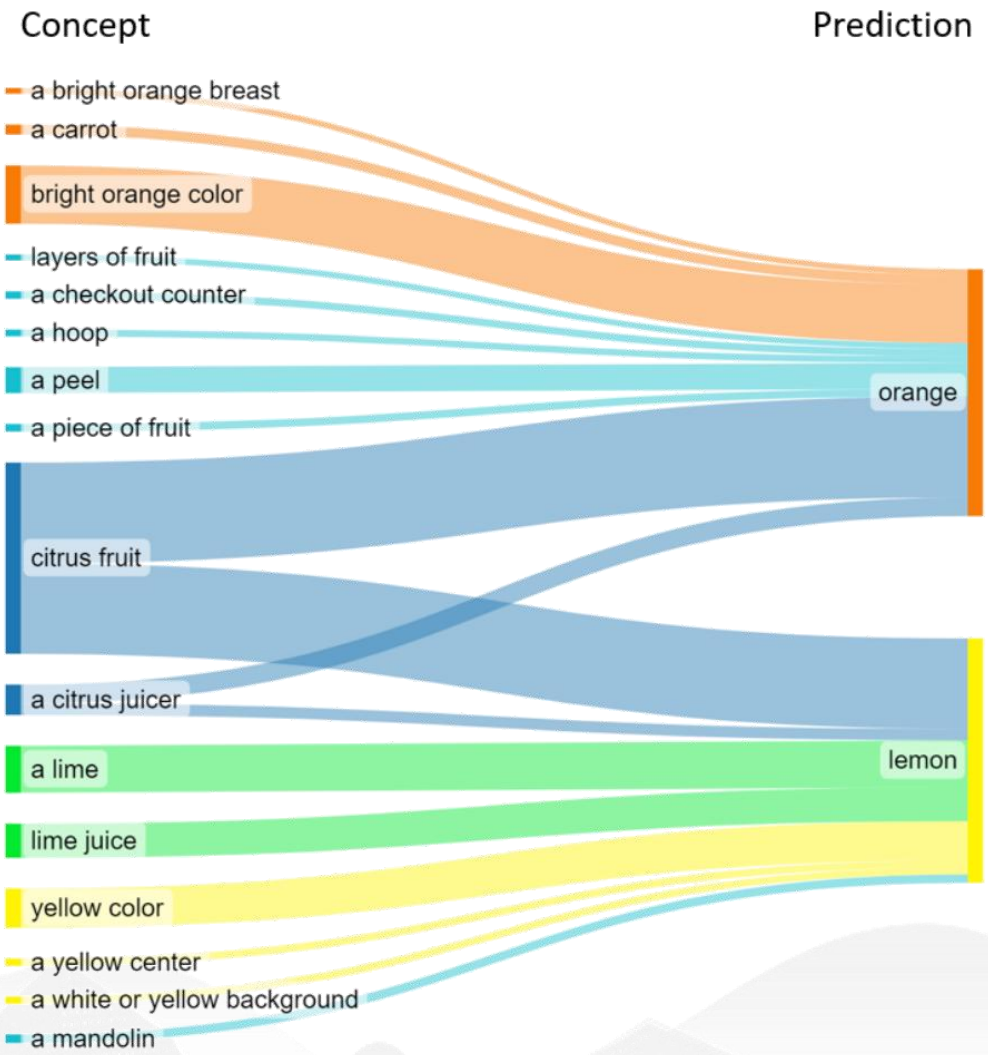
Step 2 and 3: learning the **concept bottleneck layer (CBL)**

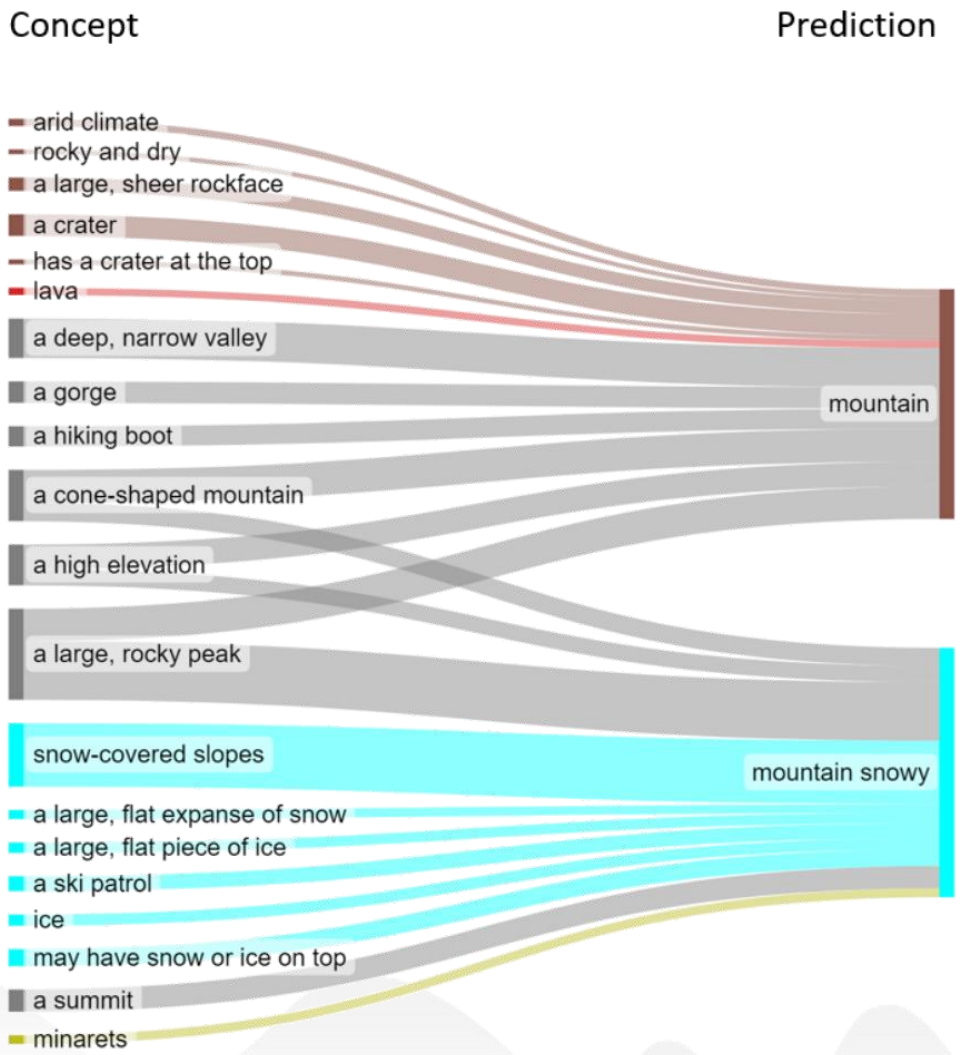


Step 4: learning the **sparse** final layer



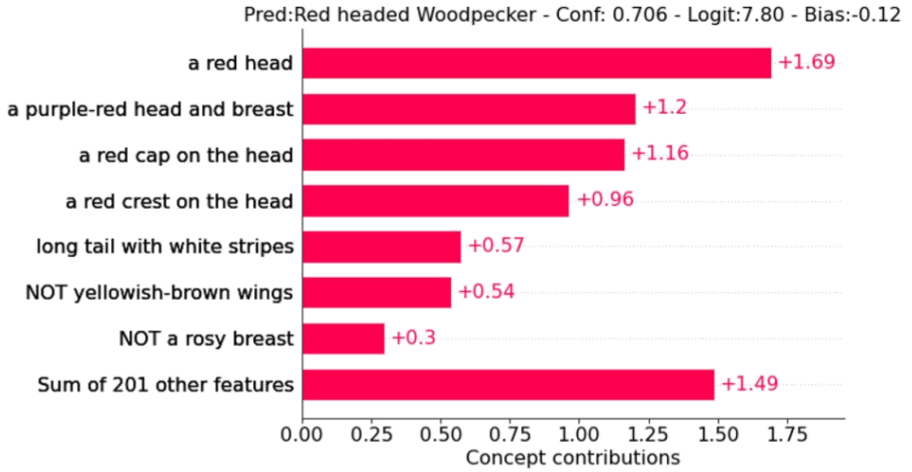
Final layer weights of Label-free CBM



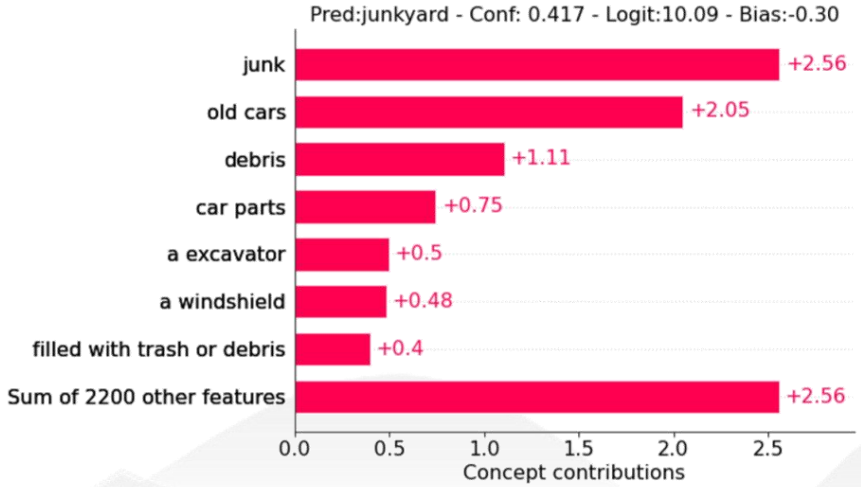


Interpretable concepts for each decision

CUB



Places 365



概念可视化：
对瓶颈层的每个维度 c 的高激活区域进行可视化，展示其所对应的概念模式。

模型行为分析：
通过调节瓶颈层的输入观察模型行为变化，从而验证瓶颈层表征的语义有效性。

Case study: manually improving an ImageNet model

Types of model errors

- *Type 1: Incorrect/ambiguous label*
- *Type 2: No sufficient concept in CBL*
- *Type 3: Incorrect concept activations*
- *Type 4: Incorrect final layer weight*

Editing final layer weights (Type 4)

- *Find an input where the model makes a Type 4 error*
- *Identify a concept to edit*
- *Change weights by sufficient magnitude*

Type 1: Incorrect/ambiguous label

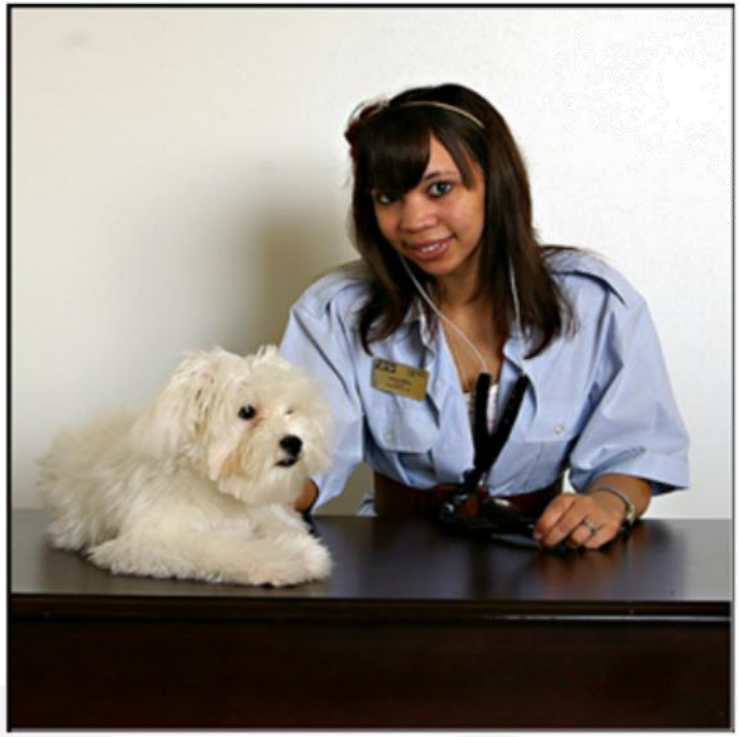
Type 1

Incorrect label

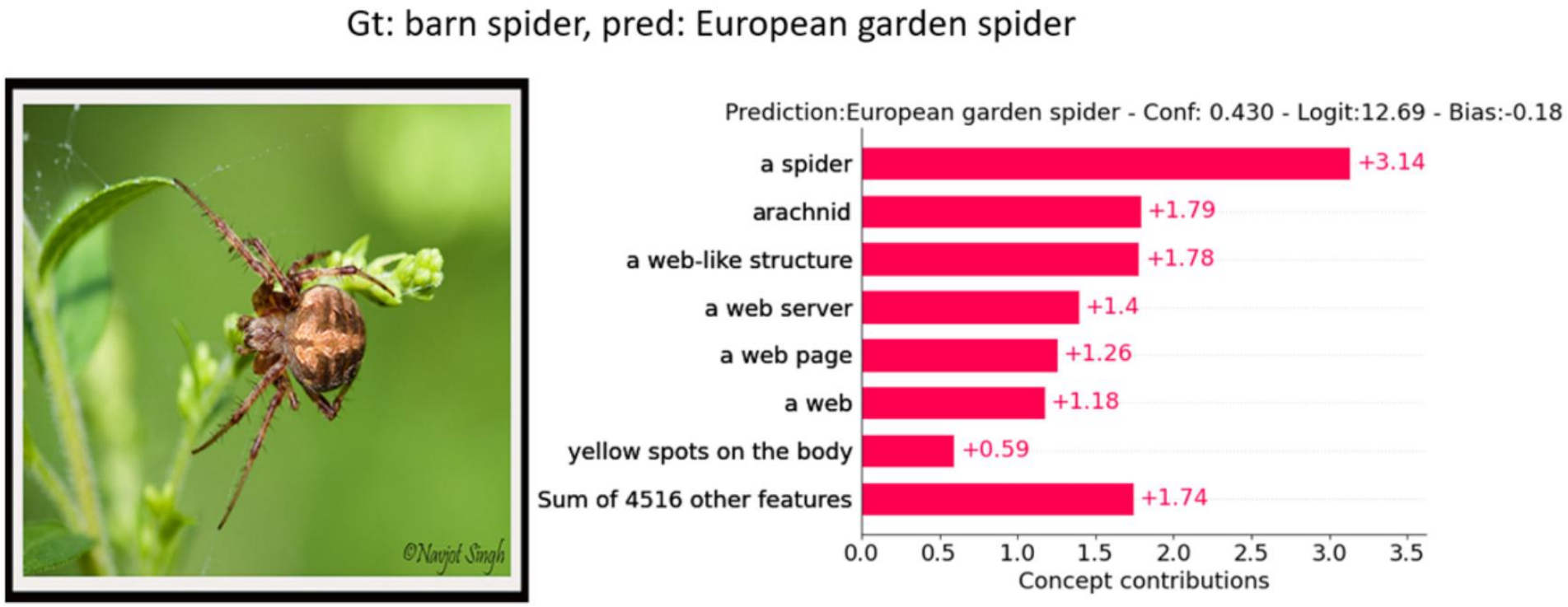
Gt: patas monkey, pred: lion

Ambiguous label

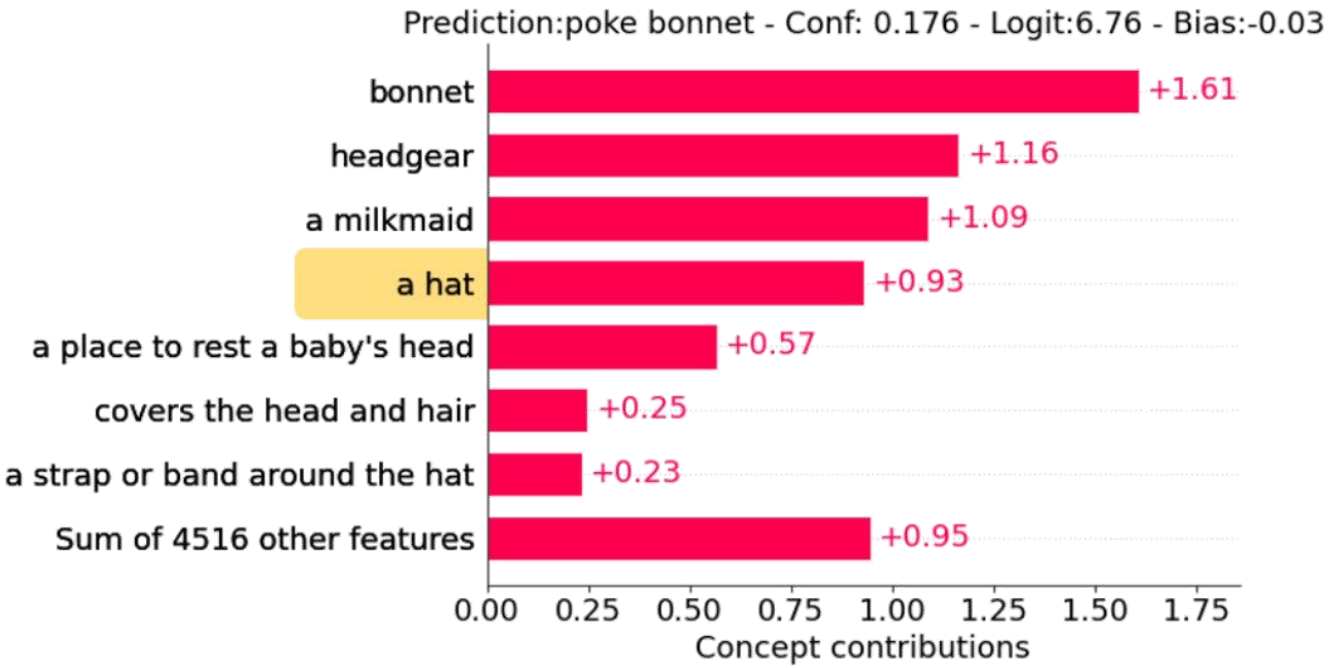
Gt: stethoscope, pred: maltese (dog)



Type 2: No sufficient concept in CBL



Type 3: Incorrect concept activations

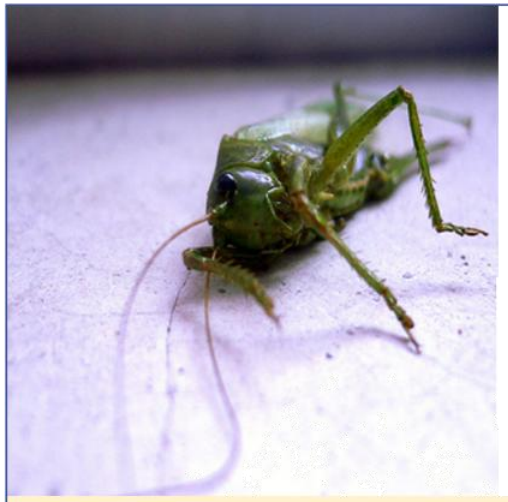


Intervene: activation of “a hat” 2.70 -> 0. New prediction: strainer ✓

过滤器

Type 4 error: Incorrect final layer weight

- *Find an input where the model makes a Type 4 error*
- *Identify a concept to edit*
- *Change weights by sufficient magnitude*



Gt: Grasshopper

Orig pred: Cricket Insect

New pred: Grasshopper

Other inputs:
+17 predictions corrected
-11 turned incorrect

Edit W_F:
"a green color" → "Grasshopper" : 0 → 0.197
"a green color" → "Cricket Insect" : 0 → -0.197

Gt: Hamper

Orig pred: Shopping Basket

New pred: Hamper

Other inputs:
+4 predictions corrected
-2 turned incorrect

Edit W_F:
"made of rope or string" → "Hamper": 0 → 0.206
"made of rope or string" → "Shopping Basket" : 0 → -0.206

Interim summary

We introduce **concept bottleneck model** along with its variations

1. which can have great task accuracy while supporting intervention and interpretation
2. allowing us to reason about these models in terms of high-level concepts that humans are familiar with
3. enabling more effective human editing through test-time intervention

Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." CVPR. 2023

Shin, Sungbin, et al. "A closer look at the intervention procedure of concept bottleneck models." ICML, 2023.

Chauhan, Kushal, et al. "Interactive concept bottleneck models." AAAI. 2023.

Sinha, Sanchit, et al. "Understanding and enhancing robustness of concept-based models." AAAI. 2023.

Lai, Songning, et al. "Faithful vision-language interpretation via concept bottleneck models." ICLR. 2023.

Webb, Taylor W., et al. "**The relational bottleneck as an inductive bias for efficient abstraction.**" Trends in Cognitive Sciences (2024).

Concept representation in human mental space

Large-scale behavioral studies show that humans represent natural objects in mental space using **sparse** and **semantic** concept embeddings.

- To understand human concept representation?
- To generate new images controlling a concept dimension?
- To test that such controllable generation change human behavior?

NCC lab's work: CoCoG & CoCoG-2

1. CoCoG: Controllable Visual Stimuli Generation Based on Human Concept Representations

- Background
  - Concept embedding
  - Similarity judgment
  - Language guided interpretability
- Our method
- Experiments

2. CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG: Controllable Visual Stimuli Generation based on Human Concept Representations. IJCAI

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. arXiv preprint arXiv:2407.14949.

Background – Concept representations

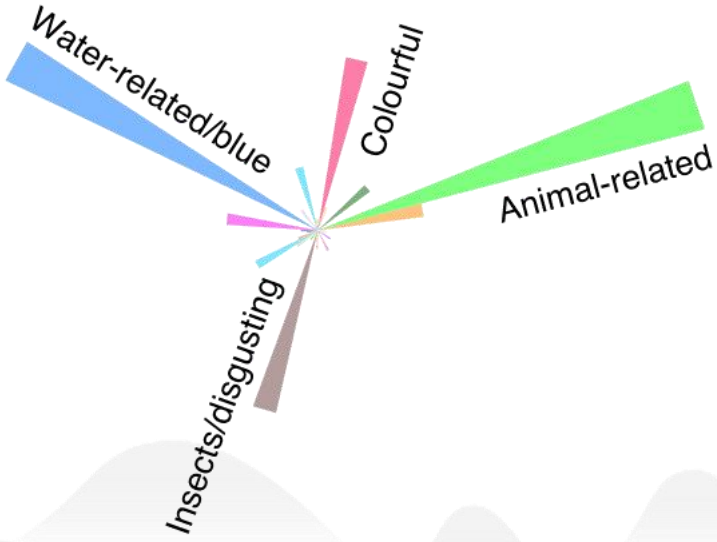
Concepts representation:
what are the properties of a visual object?

- food – eatable
- needle – pointy
- rainbow – colorful
- etc.

How to uncover concept representations in the human brain?

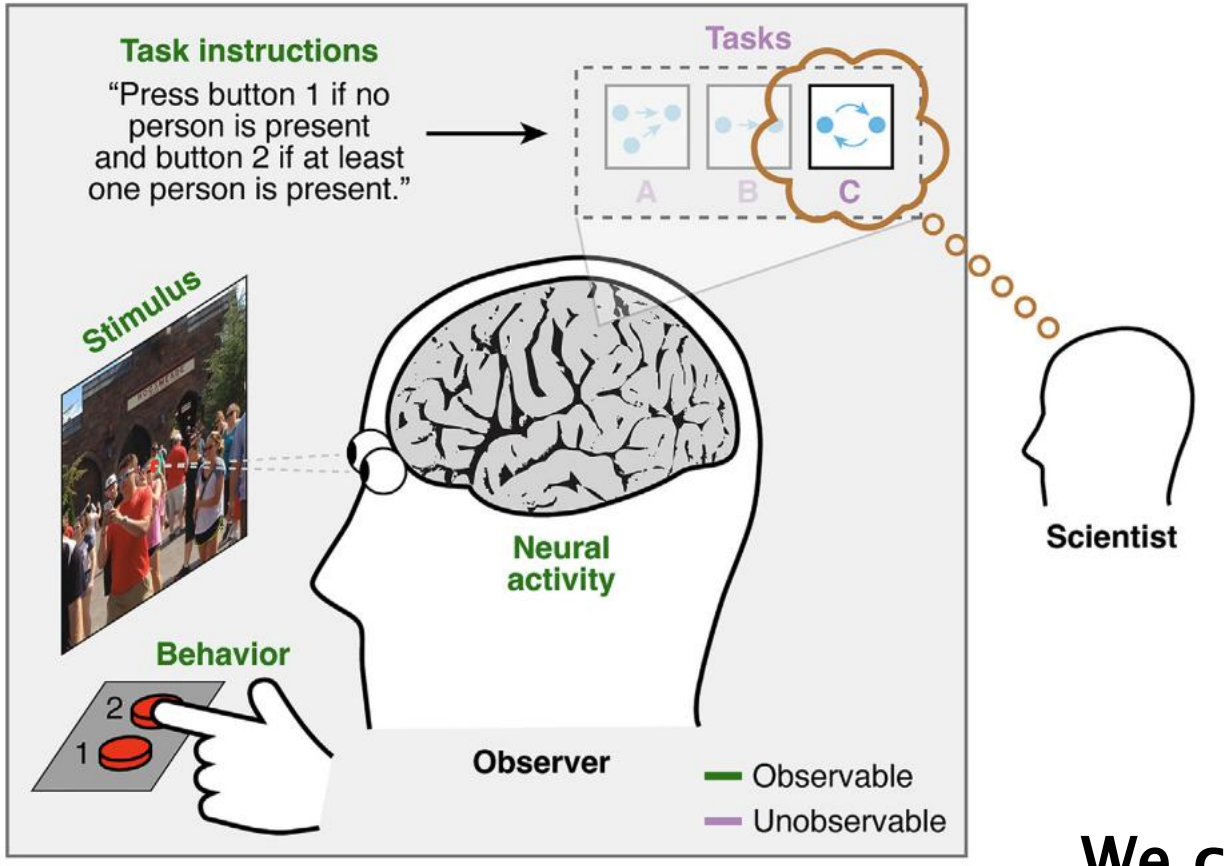
- Direct human lables

- Similarity judgment



Hebart, Martin N., et al. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements." Nat. Hum. Behav. (2020)

Concept embeddings: Linking sensory inputs with behaviors



(Kendrick et al. 2023, Neuron)

Key factors to influence concept embeddings in humans:

1. Task / context
   - Task paradigm
   - Your mood
   - Category
2. Sensory modality
   - Vision
   - Language
   - Smell

Fish

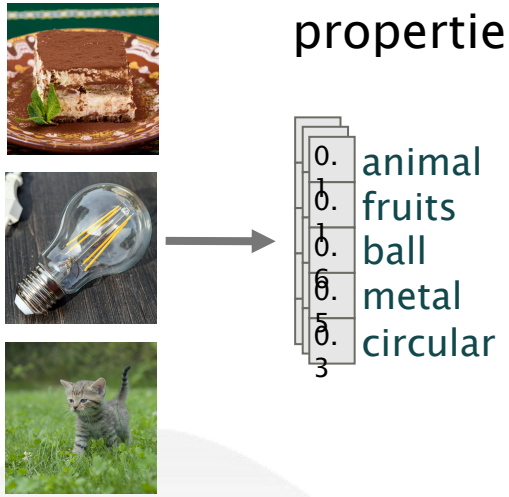
We can only observe the human behavior.

How do we study the concept embeddings in human mental space?

How to get concept embeddings in human

1. **Rating tasks (directly ask participants to rate)**
   - *What are the properties of an object?*
   - *What is the degree/extent/level of property **b** in object **a**?*

   Limits: need to specify the properties in advance; be subjective

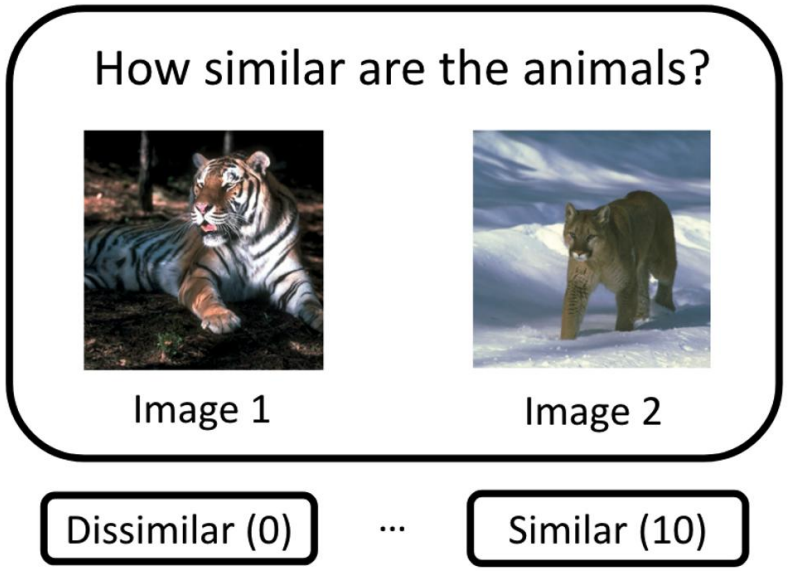


Marjieh, Raja, Nori Jacoby, Joshua C. Peterson, and Thomas L. Griffiths. "The Universal Law of Generalization Holds for Naturalistic Stimuli." arXiv, June 14, 2023.

How to get concept embeddings in human

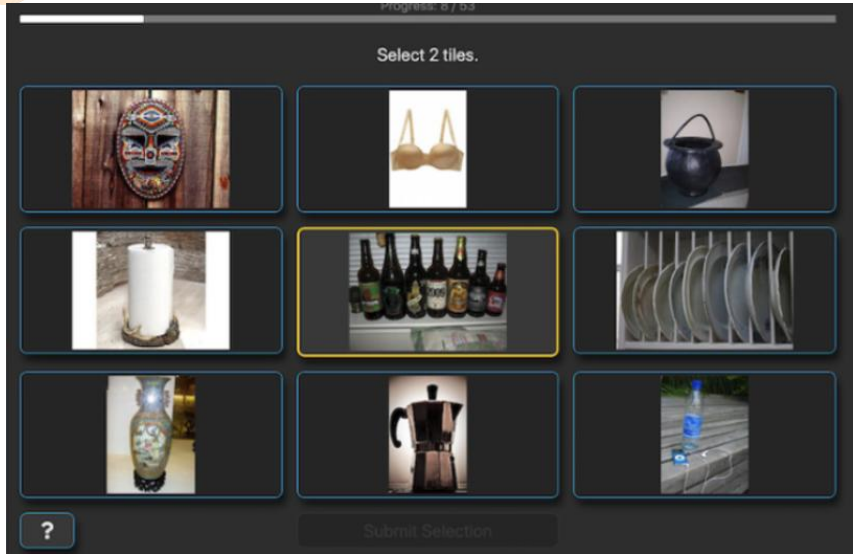
1. **Rating tasks (directly ask humans to rate)**
   - *What are the properties of an object?*
   - *What is the degree/extent/level of property **b** in object **a**?*

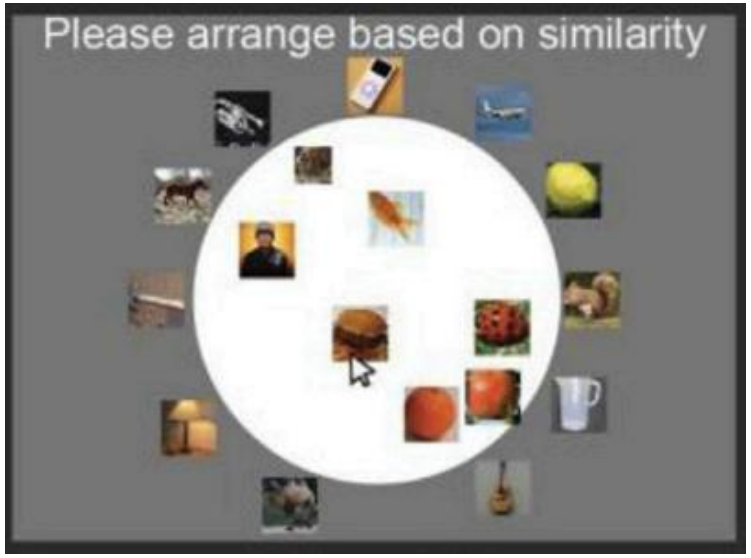
2. **Similarity judgment tasks (ask humans to compare)**
   - Implicitly characterizing the properties of objects
   - Only need information from behavioral judgments
   - in the context of other objects

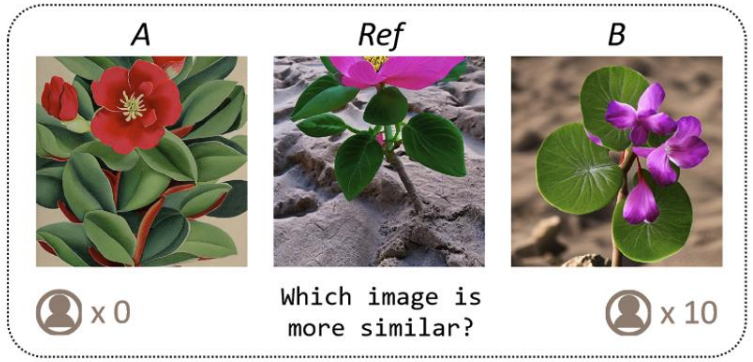


Marjieh, Raja, Nori Jacoby, Joshua C. Peterson, and Thomas L. Griffiths. "The Universal Law of Generalization Holds for Naturalistic Stimuli." arXiv, June 14, 2023.

Similarity judgment tasks







The task we used!

abacus

chopstick calculator

Q1: *which of the words at the bottom is closer in meaning with the one at the top?*

Q2: *do you know the meaning of {w1 / w2 / w3}?*

Q3: *did you find any of these words offensive/inappropriate?*

Q4: *how related are {w1 / w2 / w3} and {w1 / w2 / w3} to be?*

Radoslaw et al., 2019, NeuroImage; Brett et al., 2021, CVPR;
Borghesani et al., 2023, Sci. Data; Muttenthaler et al. 2023; Hebart et al., 2020, Nat. Hum. Behav

Triplet odd-one-out task

Which is the odd one out?

tiramisu

kitten

lightbulb

Triplet odd-one-out task

Which is the odd one out?

Mechanical or non-mechanical?

Triplet odd-one-out task

Which is the odd one out?

Animal or non-animal?

Our hypothesis

- The triplet similarity judgment task, i.e., odd-one-out, requires participants to select the two most similar items from a triplet.
- We can uncover human's concept space when performing similarity tasks.

- Two hypotheises:

1) Concepts represent the dimensions in mental space that are most sensitive to similarity judgments.
2) Manipulating these concepts can effectively influence human similarity judgments.

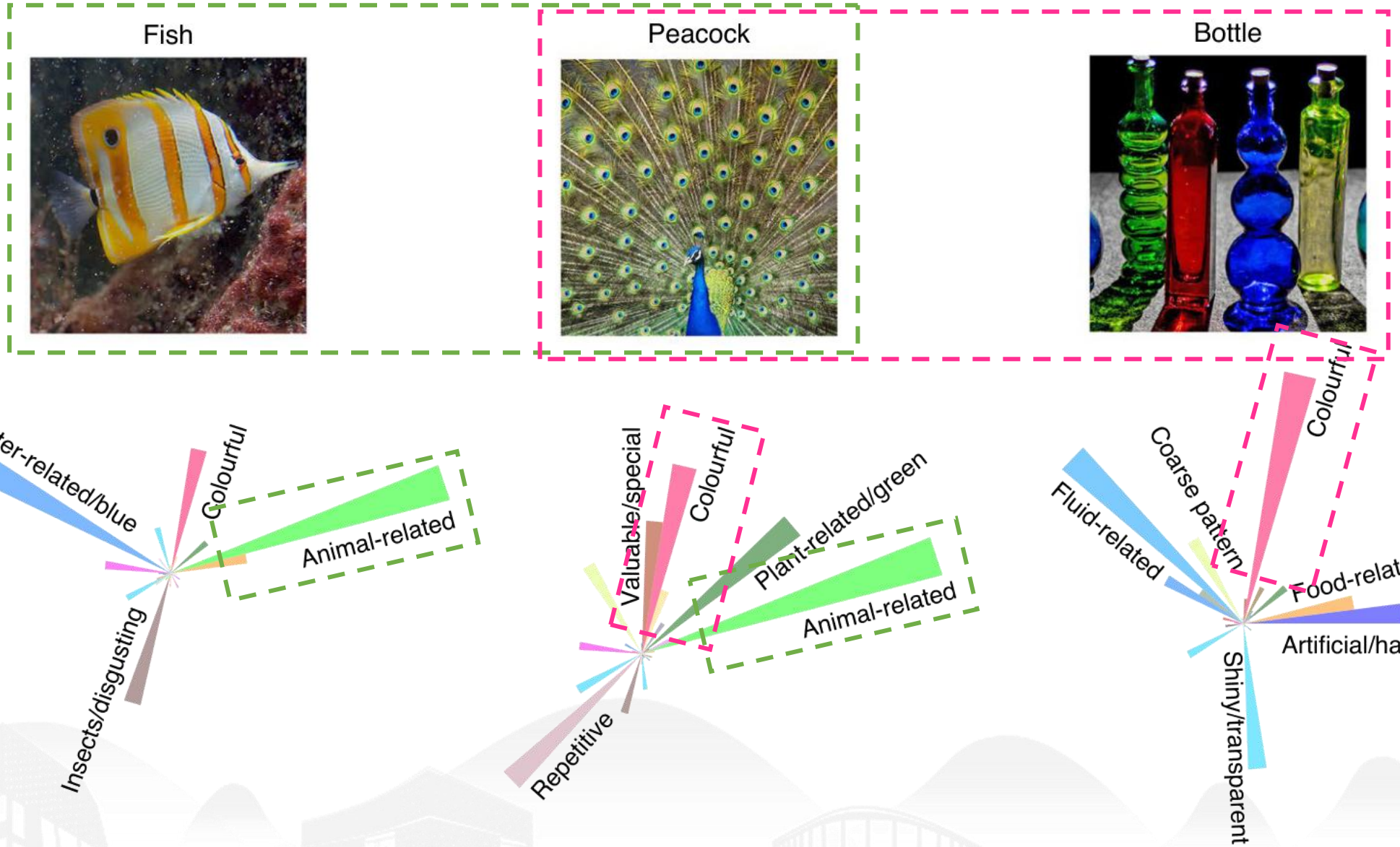
Hebart, Martin N., et al. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements." Nat. Hum. Behav. (2020)

Our hypothesis

- The triplet similarity judgment task, i.e., odd-one-out. participants select the two most similar items from a triplet.
- We can uncover human's concept space from there judgments.

Hebart, Martin N., et al. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements." Nat. Hum. Behav. (2020)

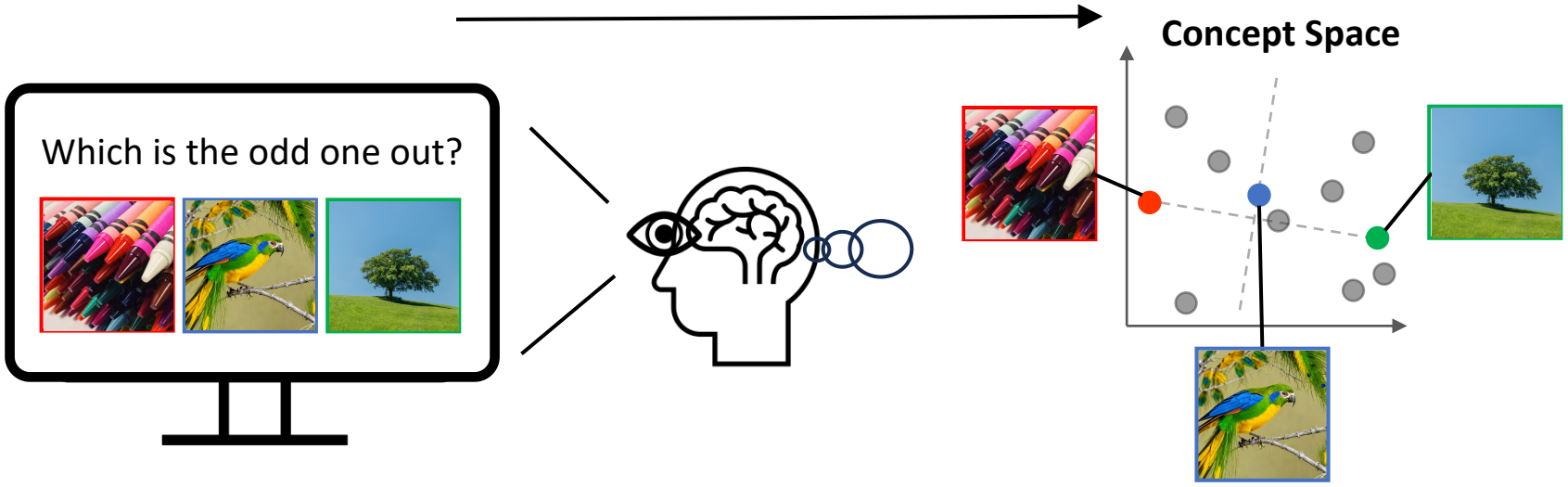
Our hypothesis



Hebart, Martin N., et al. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements." Nat. Hum. Behav. (2020)

Motivations

1. Uncover the concept space from visual stimuli in similarity judgment.

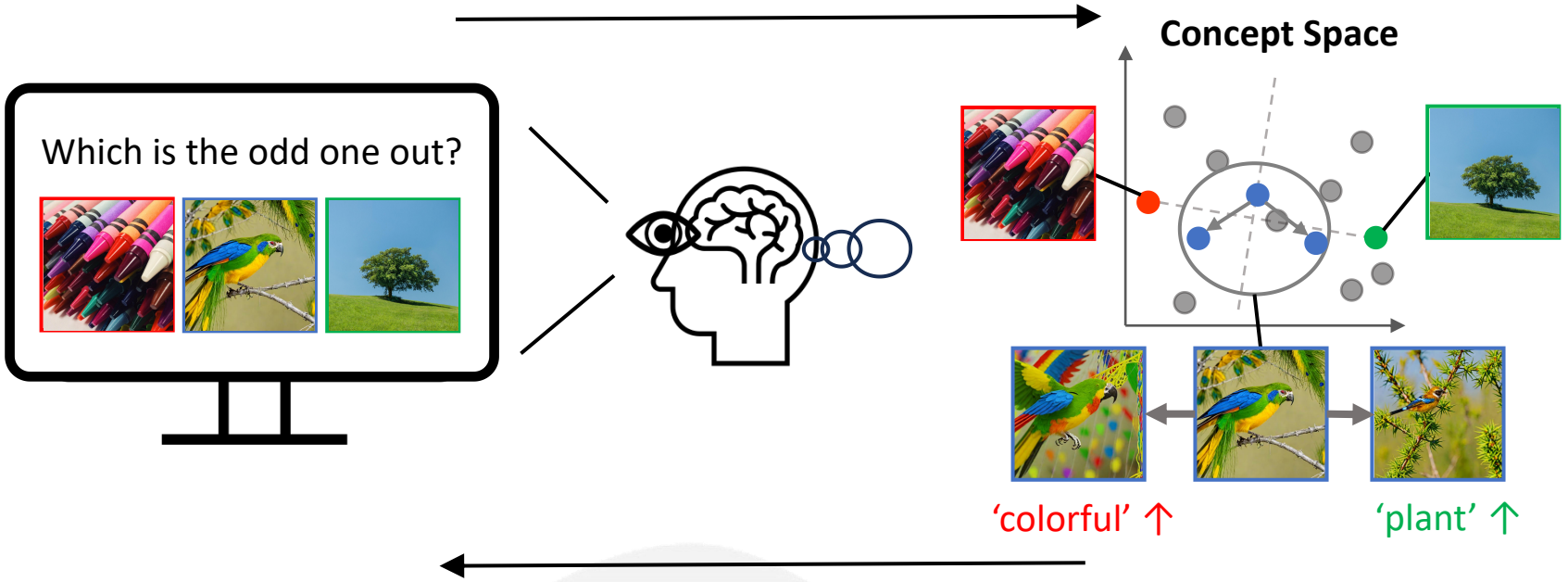


Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG: Controllable Visual Stimuli Generation based on Human Concept Representations. IJCAI

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. arXiv preprint arXiv:2407.14949.

Motivations

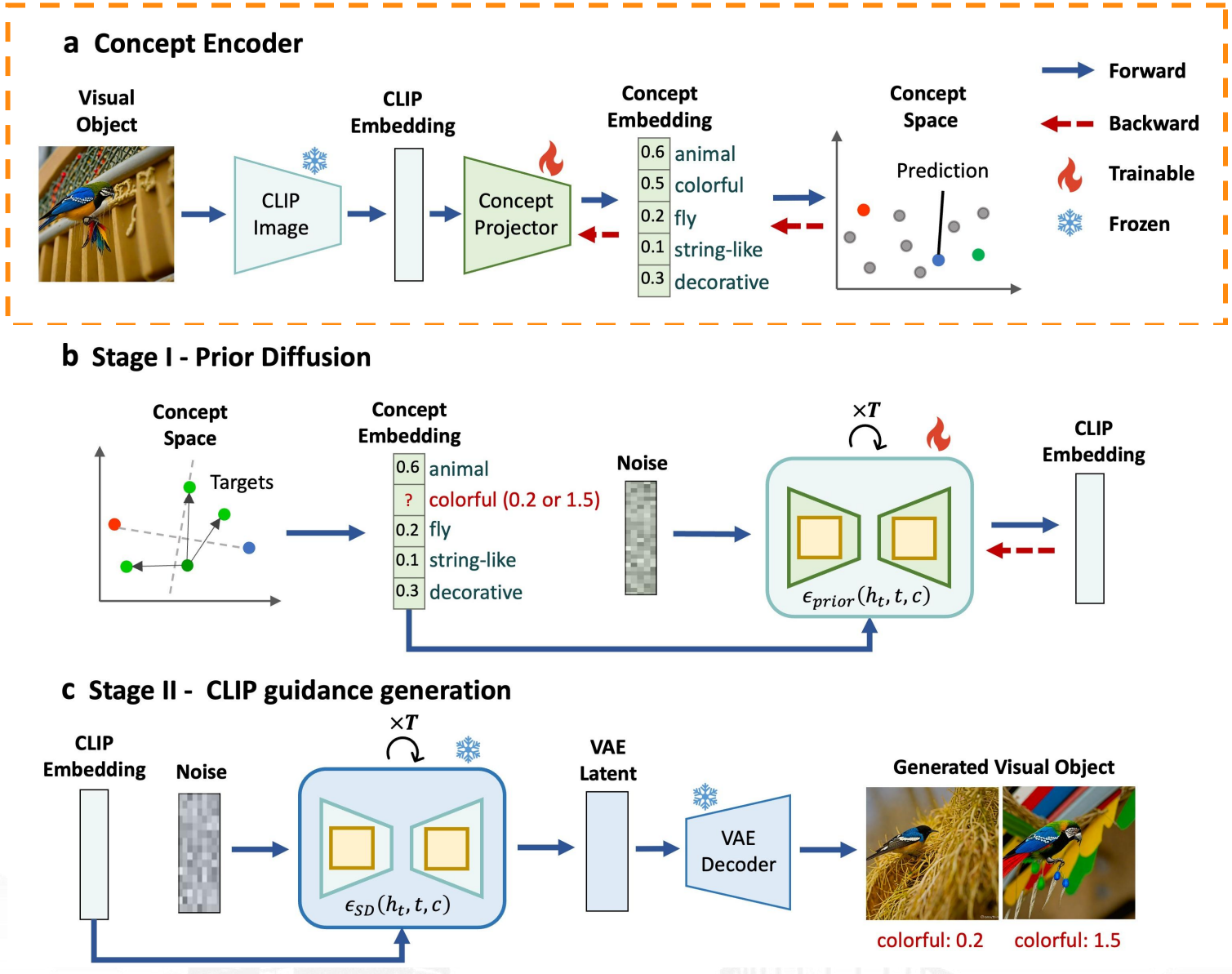
1. Uncover the concept space from visual stimuli in similarity judgment.
2. Manipulate concepts in visual stimuli to control similarity judgments.



Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG: Controllable Visual Stimuli Generation based on Human Concept Representations. IJCAI

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. arXiv preprint arXiv:2407.14949.

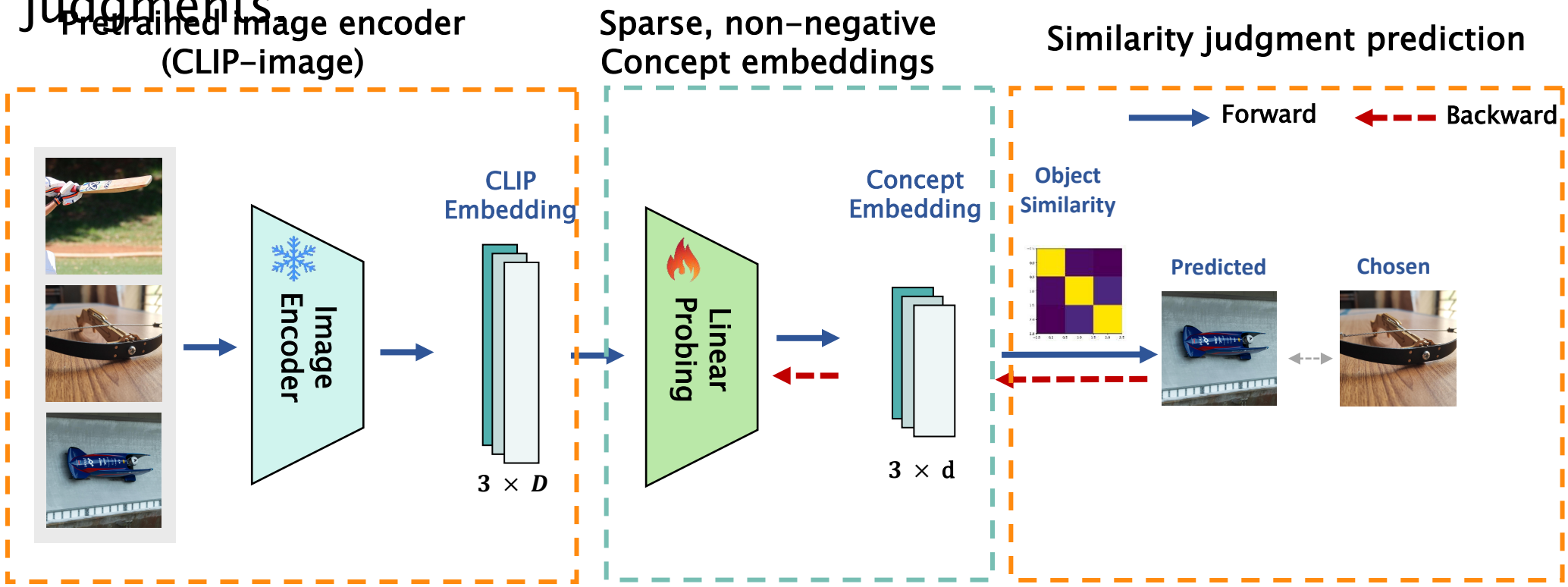
The framework of CoCoG



Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG: Controllable Visual Stimuli Generation based on Human Concept

Concept encoder

- Train encoder by fitting human behaviors in triplet similarity judgments



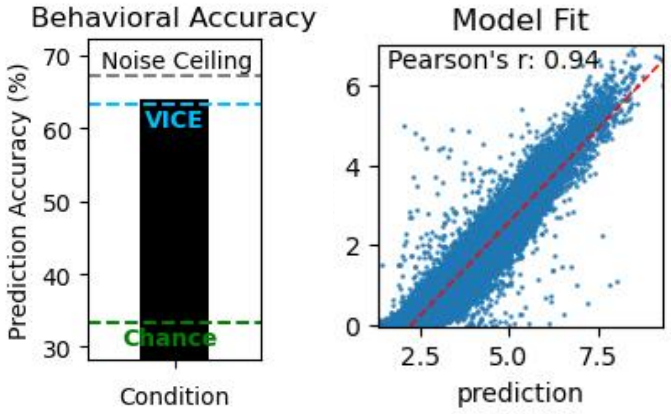
Similarity judgment prediction: $$p(\{a,b\}|\{i,j,k\}) := \frac{\exp(S_{ab})}{\exp(S_{ij}) + \exp(S_{ik}) + \exp(S_{jk})}$$

Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG: Controllable Visual Stimuli Generation based on Human Concept

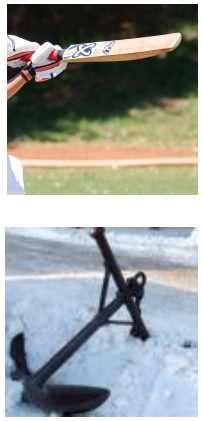
Concept encoder

- Train encoder by fitting human behaviors in triplet similarity judgments.
- SOTA performance with sparse and interpretable concept space.

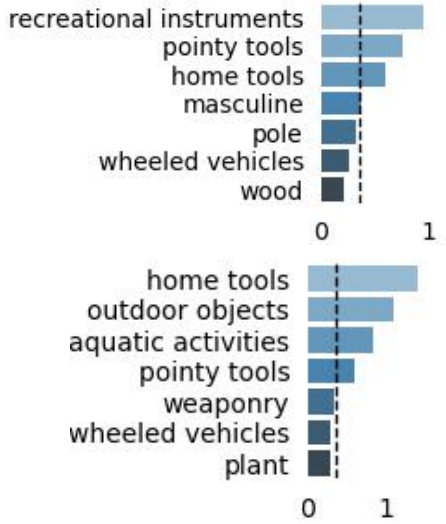
a

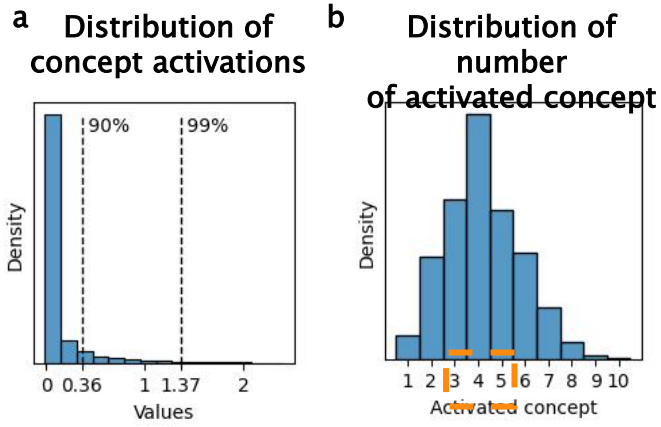


b

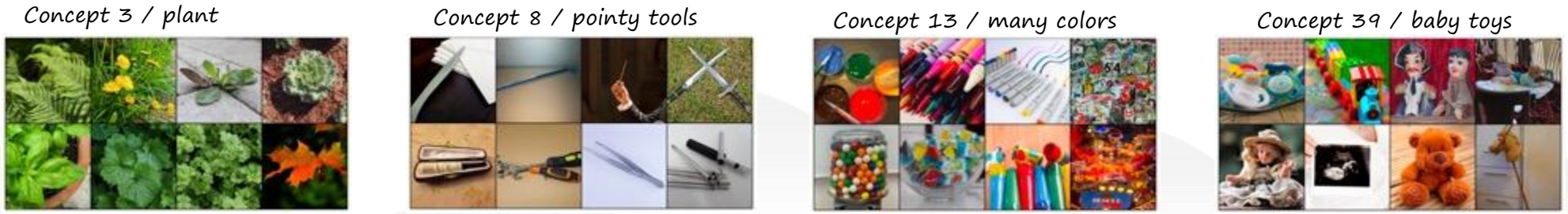


c



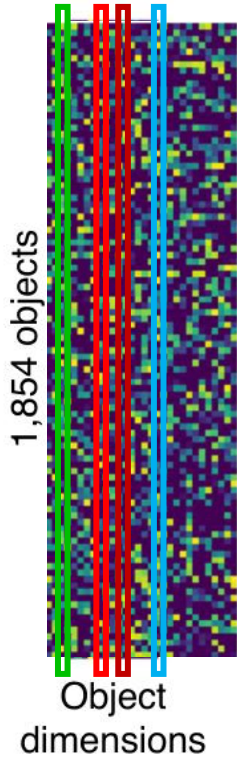


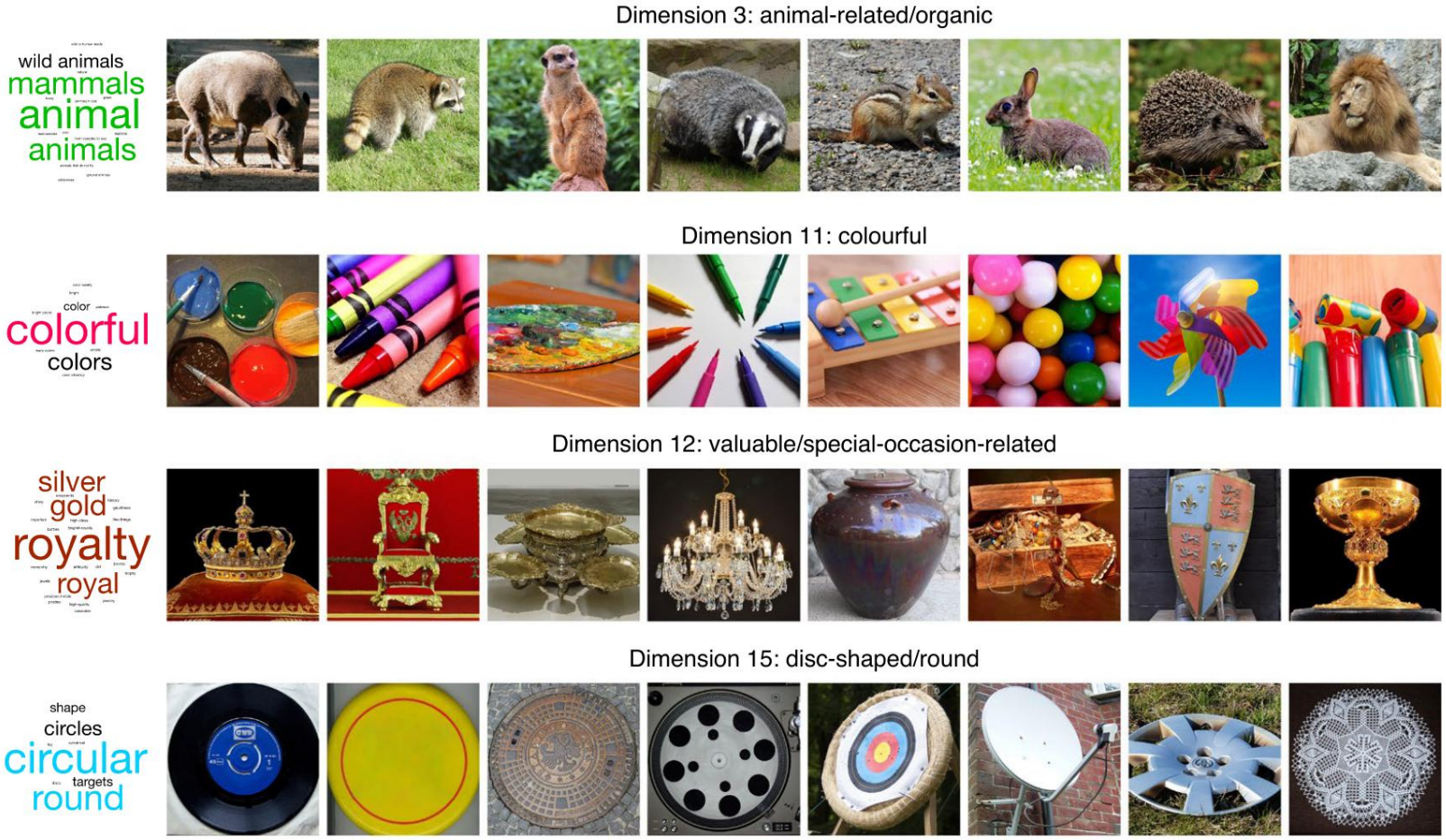
d



How to interpret the concept embeddings?

Each dimension (column) in concept embeddings is an attribute.

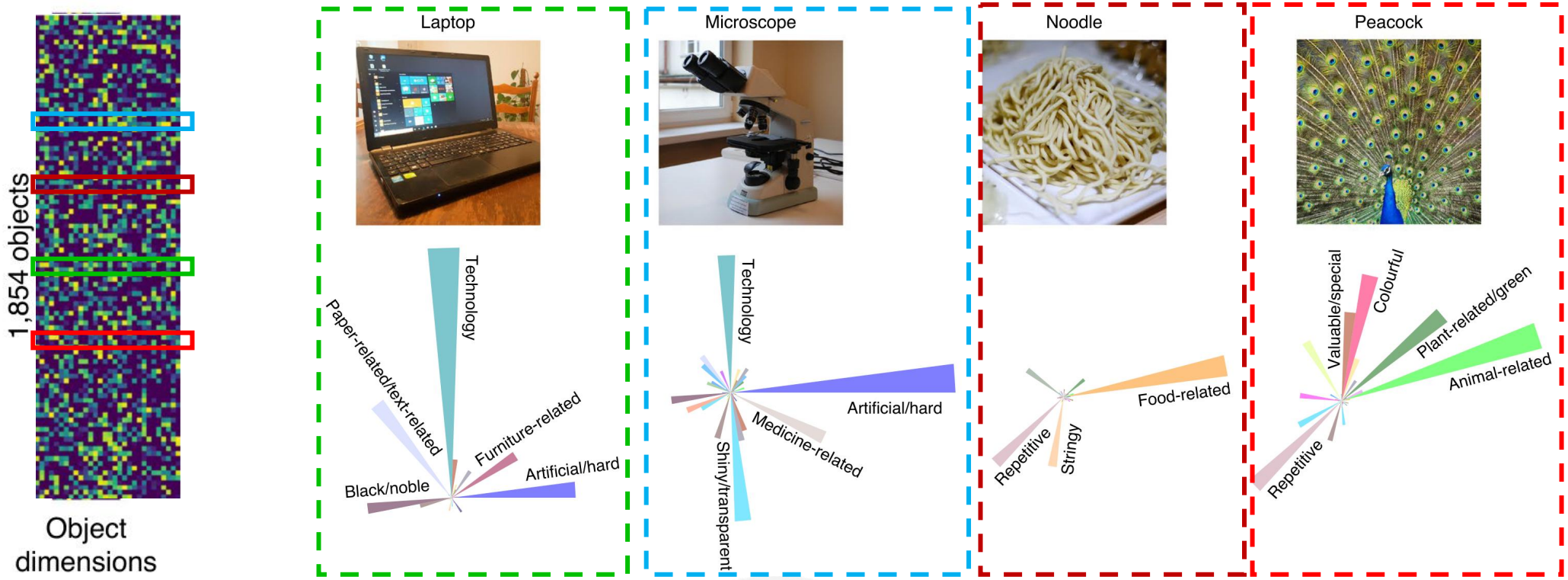




Hebart, Martin N., et al. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements." *Nature human behaviour* 4.11 (2020): 1173-1185.

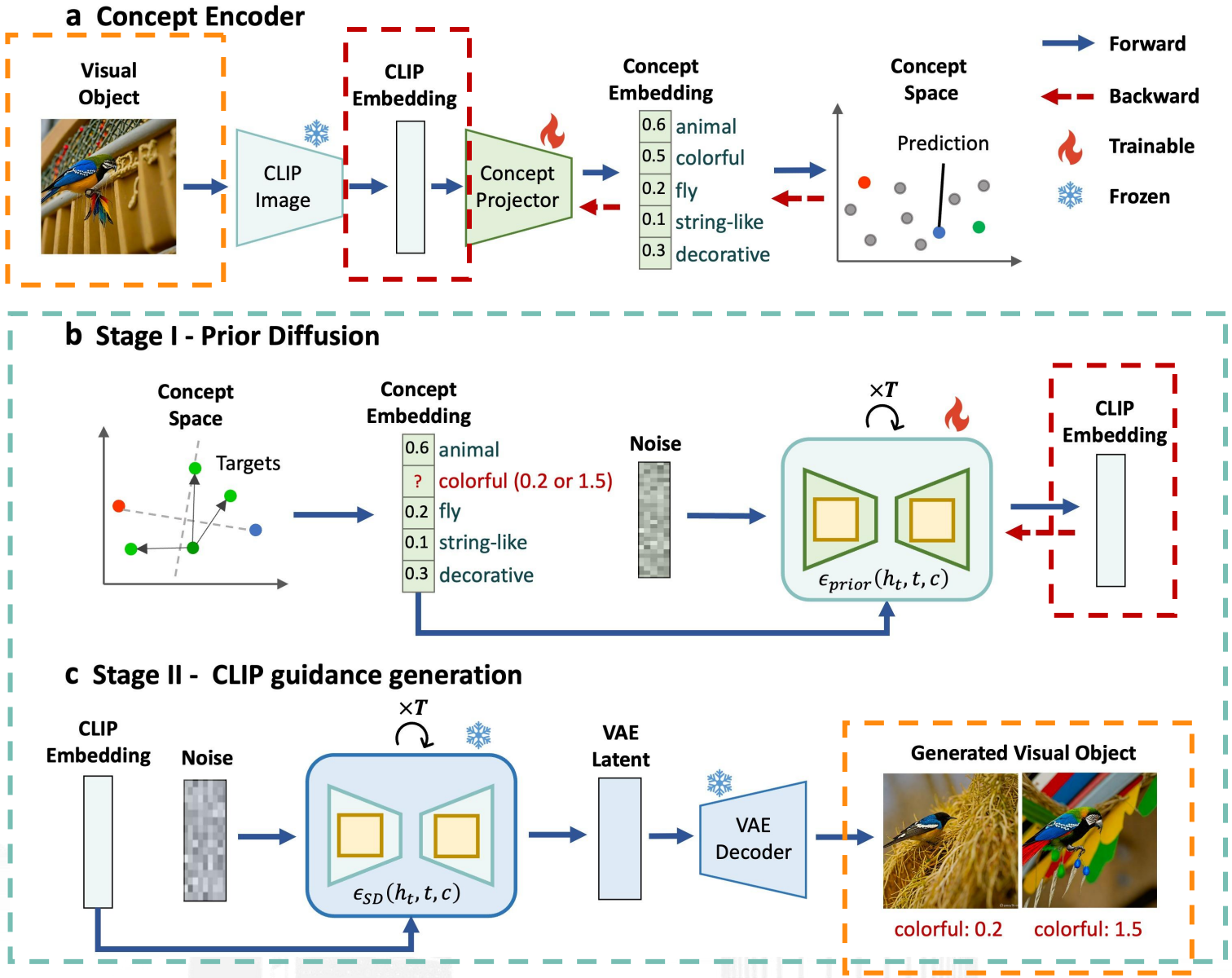
How to interpret the concept embedding?

Each row shows the weights of respective attributes.



Hebart, Martin N., et al. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements." *Nature human behaviour* 4.11 (2020): 1173-1185.

Concept-based controllable generation

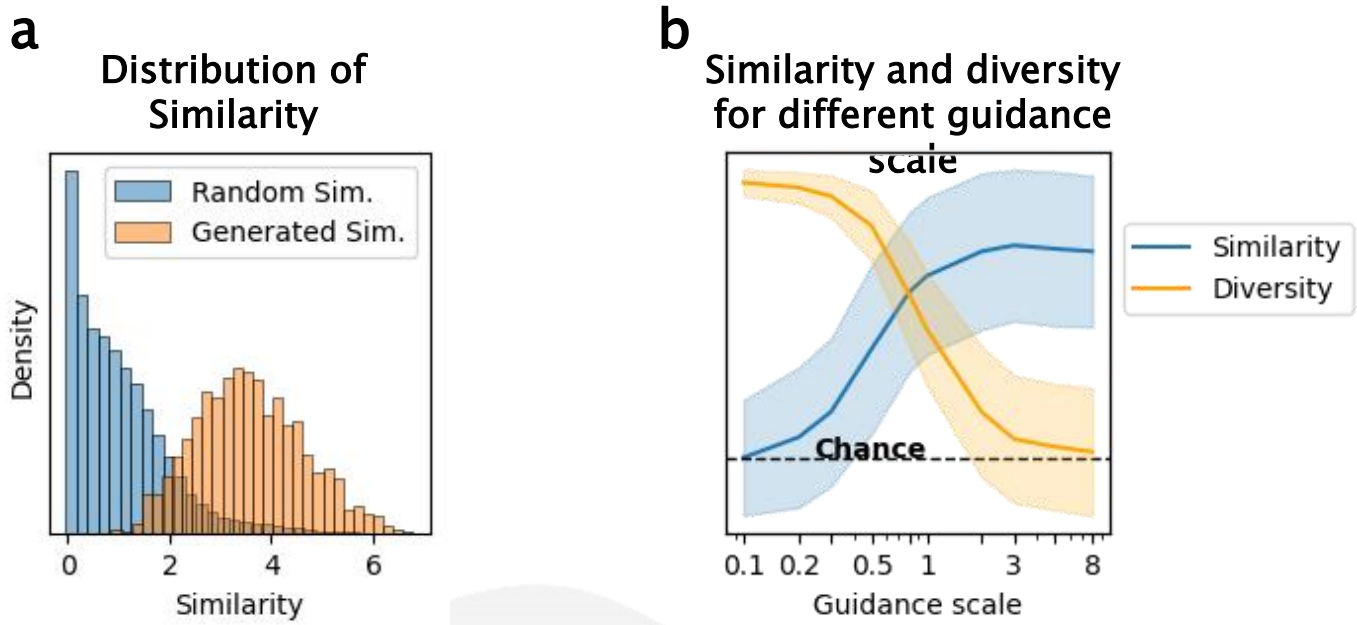


Concept-based controllable generation

- Concept decoder generate visual stimuli conditioned on target concepts.

Concept-based controllable generation

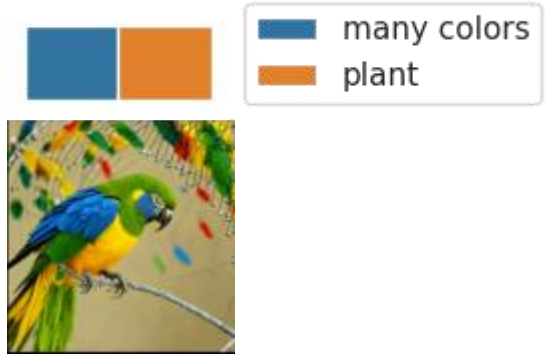
- Concept decoder generate visual stimuli conditioned on target concepts.
- The generated images ensure diversity while satisfying the target concept.



Concept-based controllable generation

- Concept decoder generate visual stimuli conditioned on target concepts.
- The generated images ensure diversity while satisfying the target concept.
- Manipulating similarity judgment behavior by key concepts intervention

Reference 1

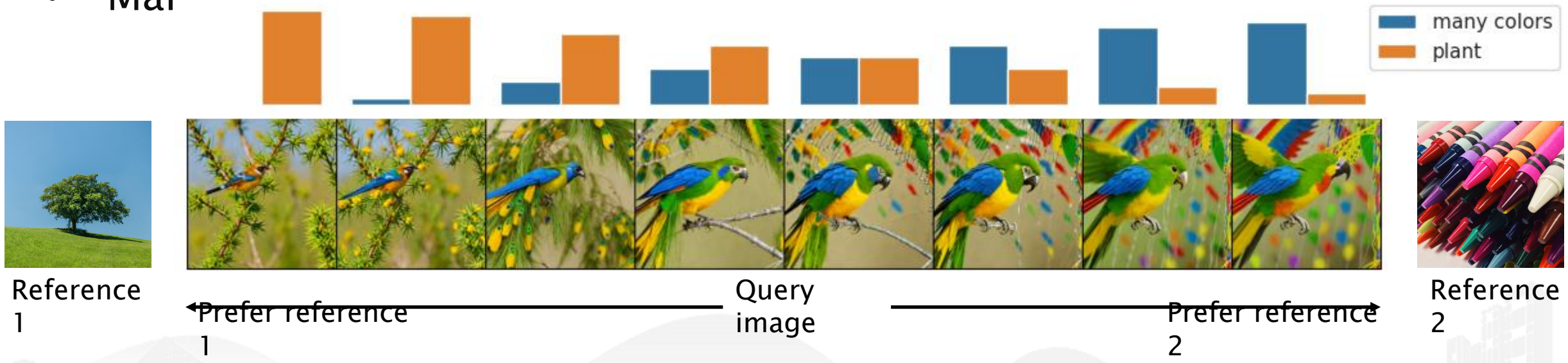


Query image

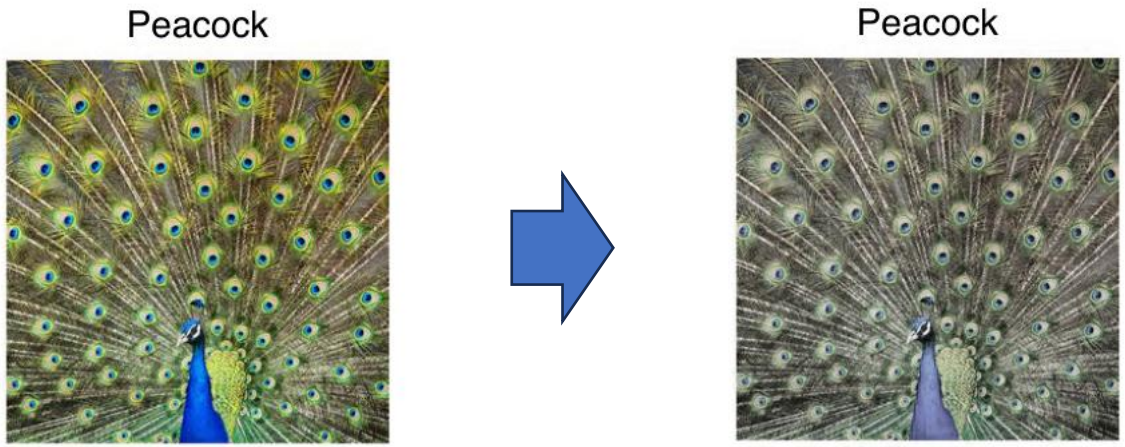
Reference 2

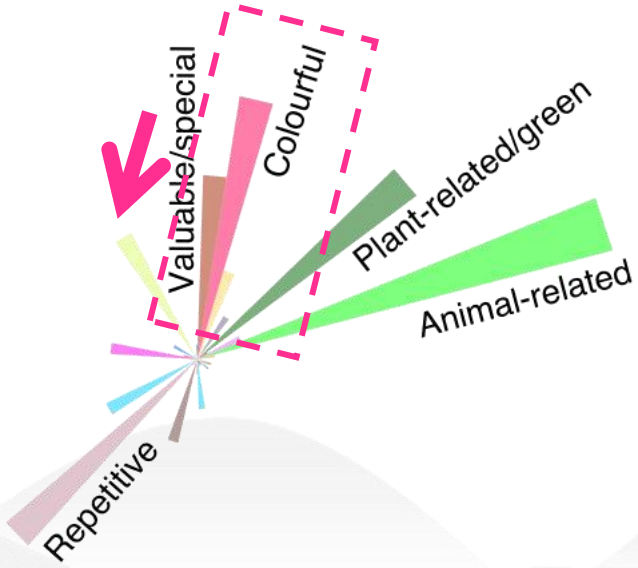
Concept-based controllable generation

- Concept decoder generate visual stimuli conditioned on target concepts.
- The generated images ensure diversity while satisfying the target concept.
- Manipulating similarity judgment behavior by key concepts intervention



Motivations – Active manipulation of concept





Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Generating Visual Stimuli

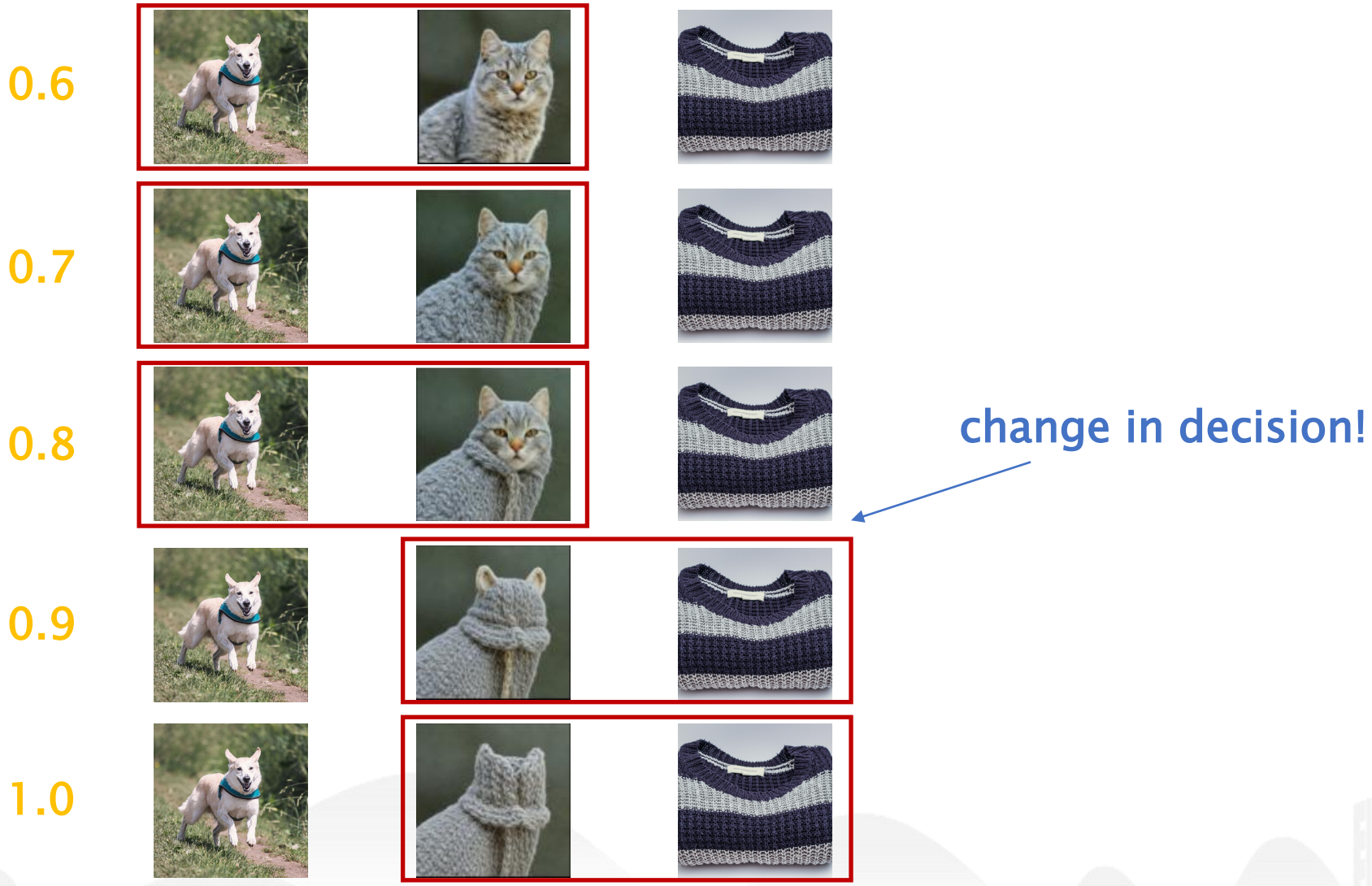
Does “**clothing**” influence this judgment?

“clothing”

CoCoG-2

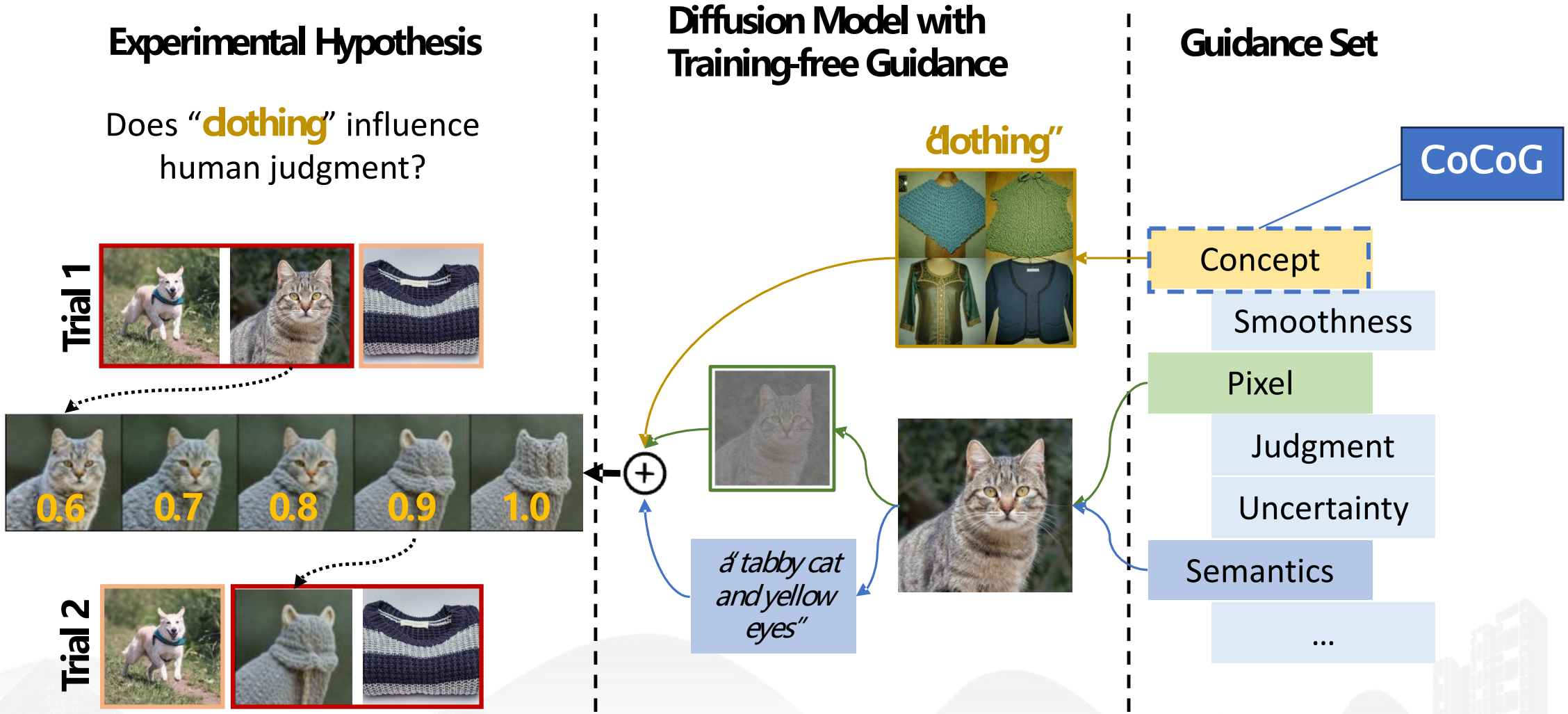
Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Generated Visual Stimuli change judgment



Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Framework

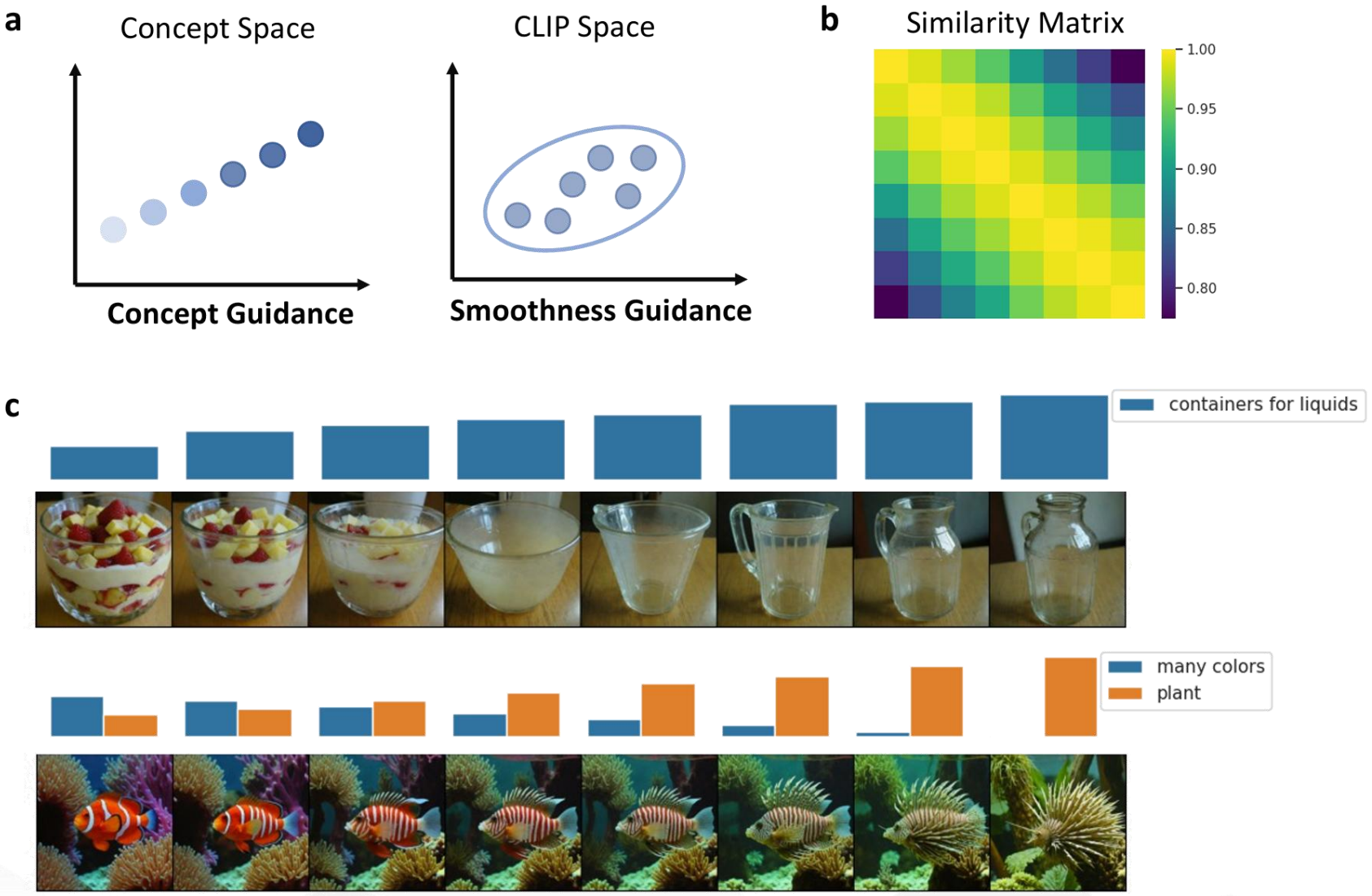


Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Experiment 1: Diversity

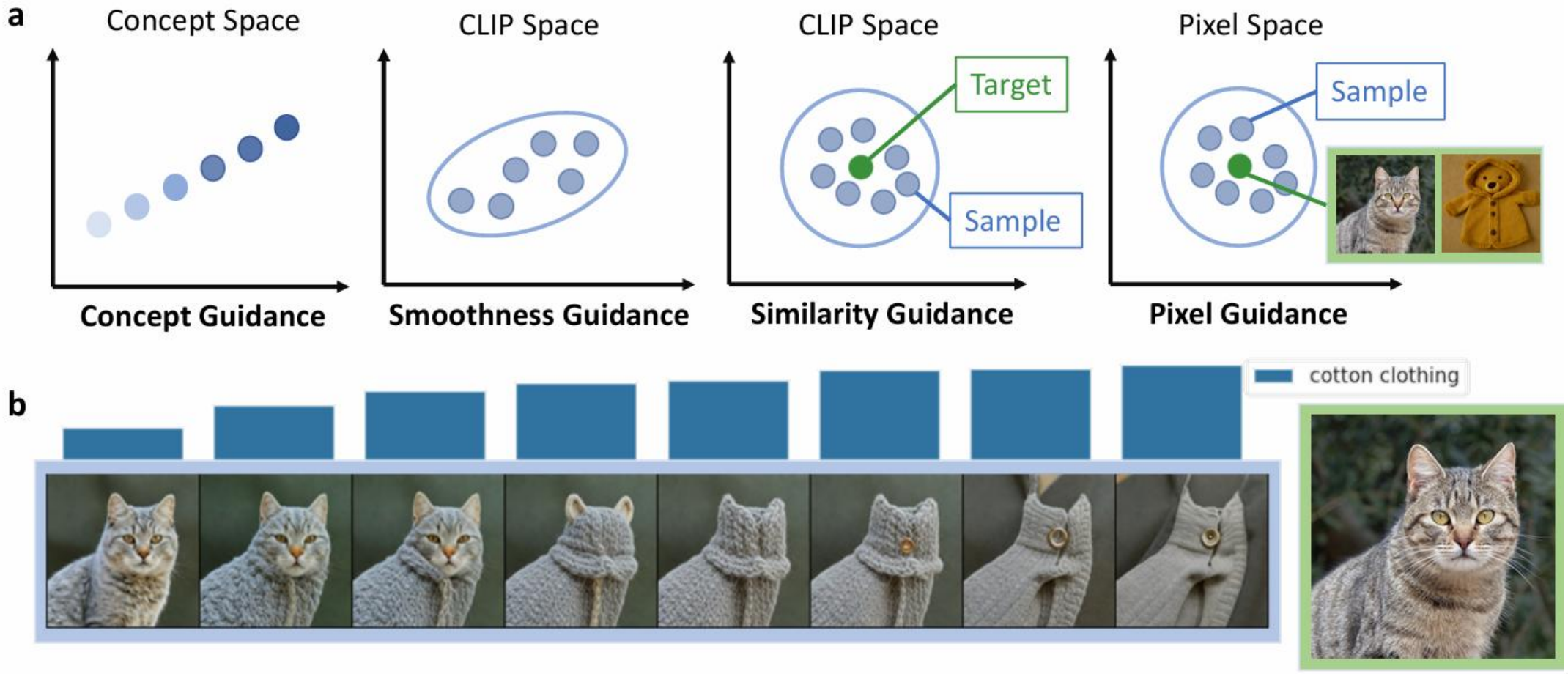
Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Experiment 2: Smoothness



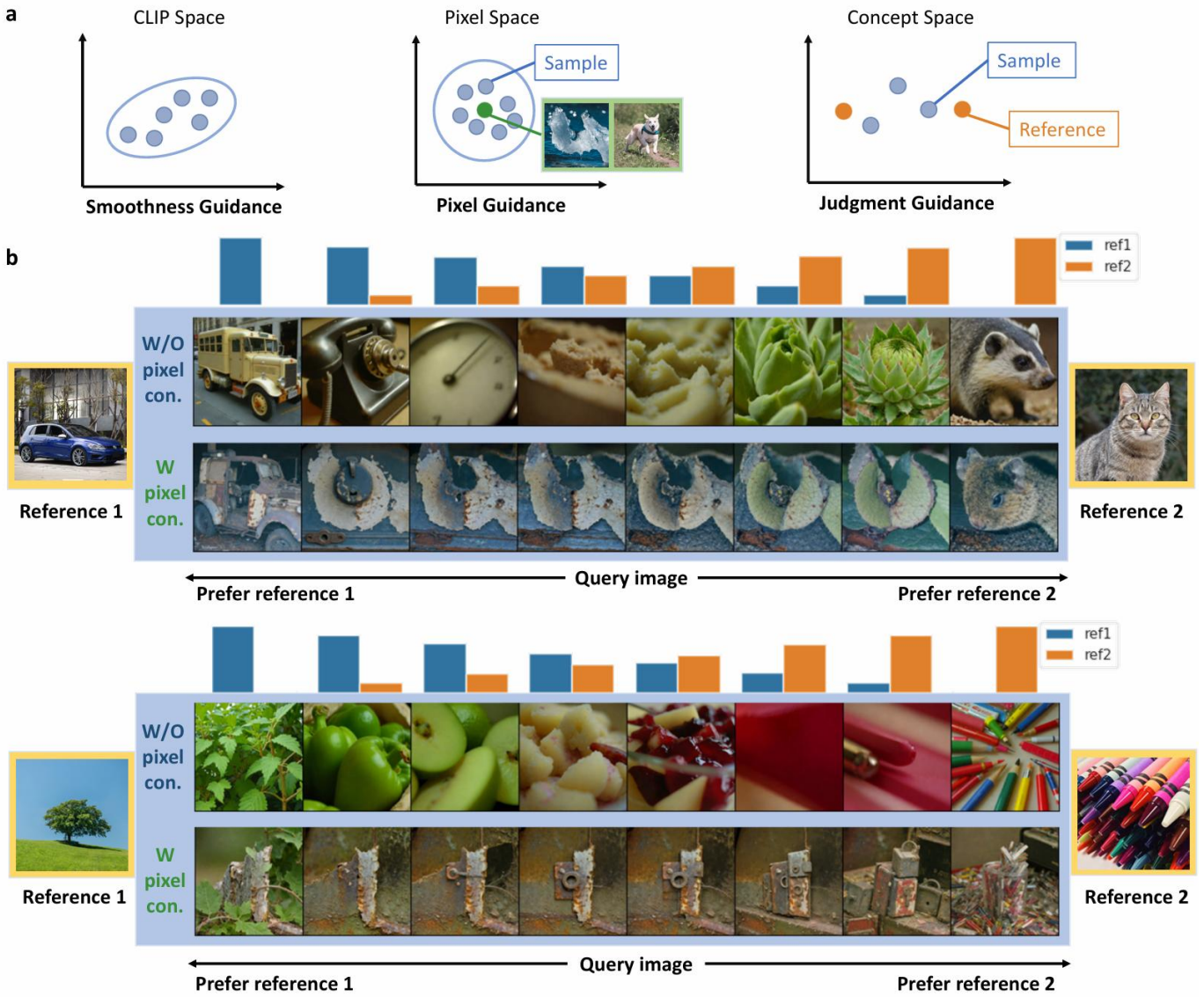
Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Experiment 3: Image Editing



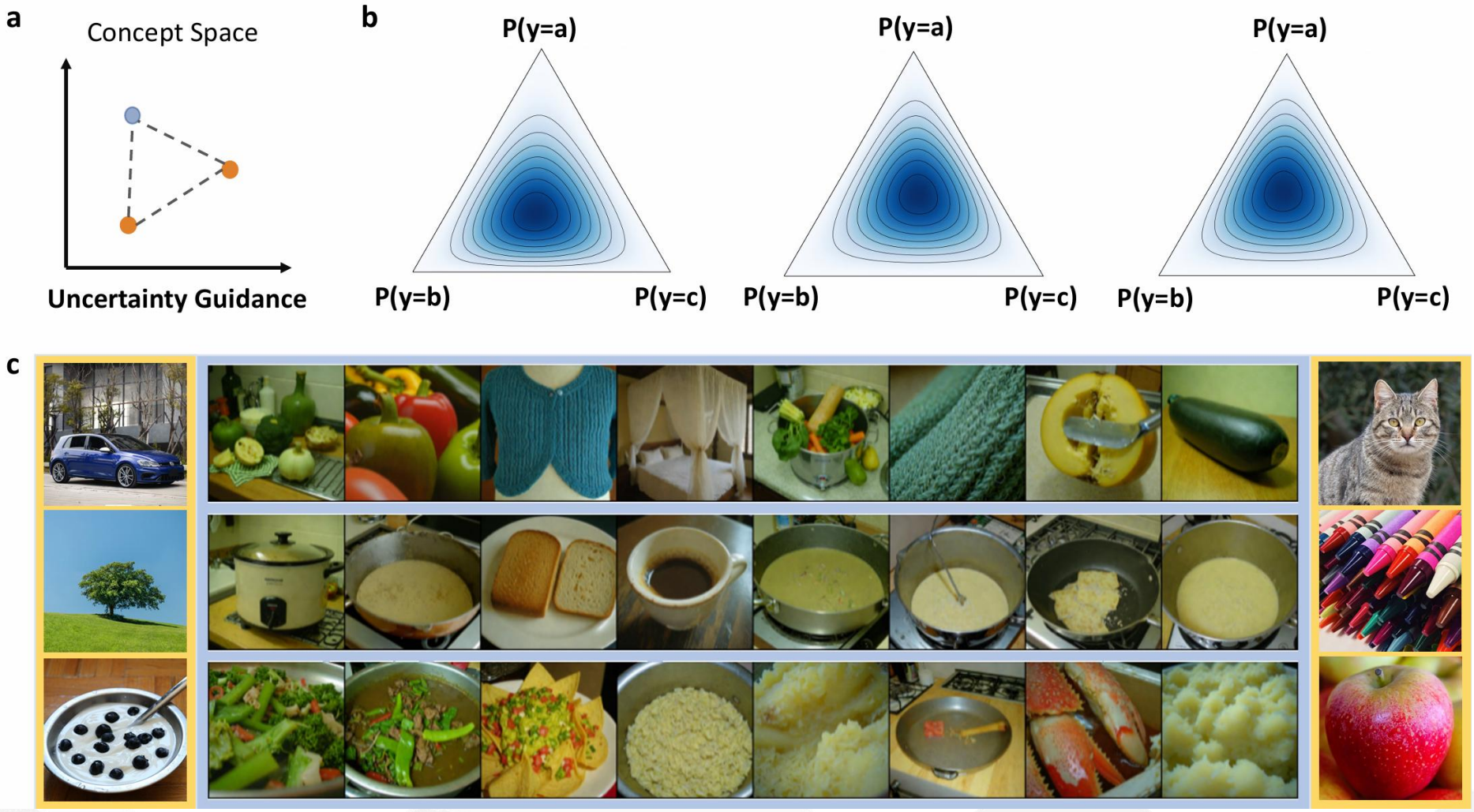
Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Experiment 4: Manipulation of Similarity Judgment



Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Experiment 5: High Uncertainty Trials (more informative)



Wei, C., Zou, J., Heinke, D., & Liu, Q. (2024). CoCoG-2: Controllable generation of visual stimuli for understanding human concept representation. IJCAI workshop on HBAI.

Future directions

1. Naturally **manipulate human behaviors** in **various perceptual decision tasks** (eg. emotion task & attention task & memory task) via intervening on concept embeddings
2. **Improve generalization & robustness of AI** by aligning the concept embeddings of AI and humans
3. **Human-in-the-loop design**
4. **Build human-like AI** by aligning the concept embeddings.

Lecture 15 – Concept representation in human and AI

- Cognitiove processes in Large Models & Humans
- **Computer vision**: CNN model & latent representation
- **Human vision**: Concept Bottleneck Model (CBM) & Interpretable concept representation
- **NCC LAB's work**: CoCoG & CoCoG-2
- **Future directions**