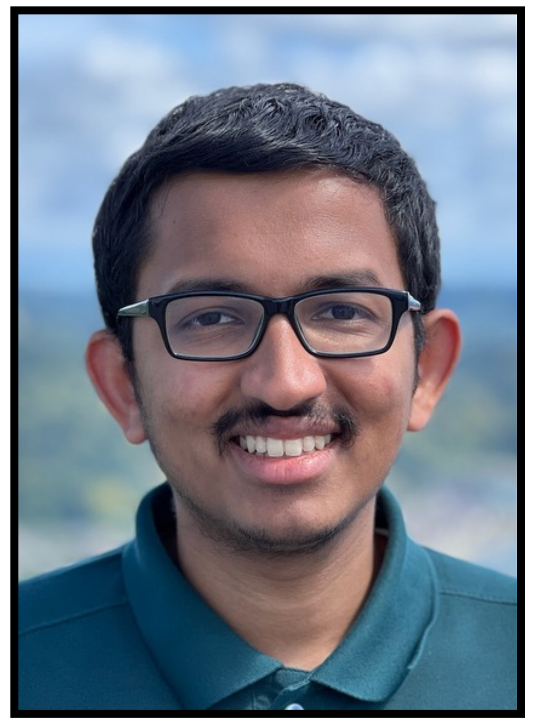
MAX PLANCK INSTITUTE
FOR INFORMATICS

SIC Saarland Informatics Campus

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

# Discover-then-Name: Task-Agnostic Concept Bottlenecks via Automated Concept Discovery



Sukrut Rao*

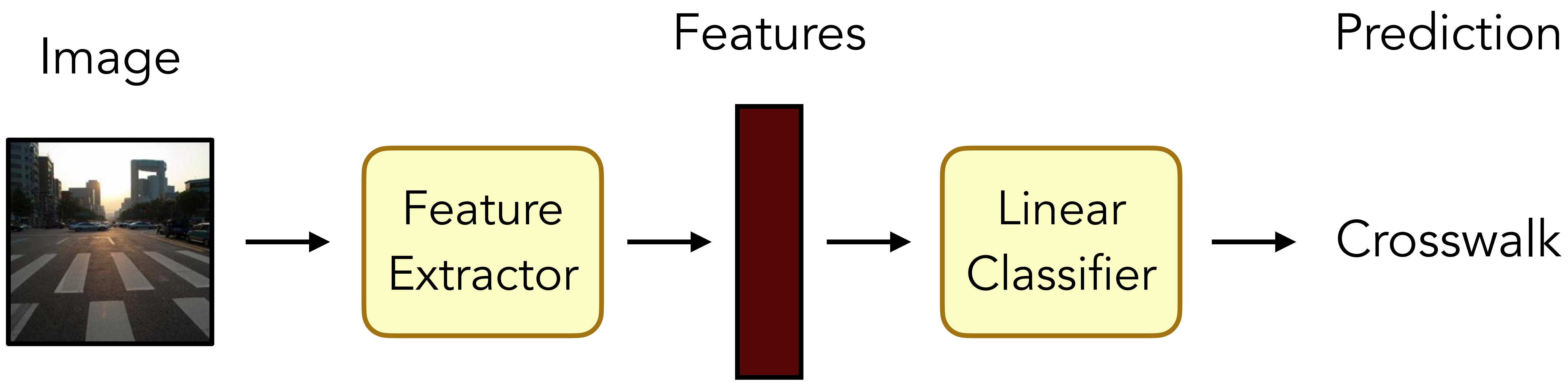
Sweta Mahajan*

Moritz Böhle

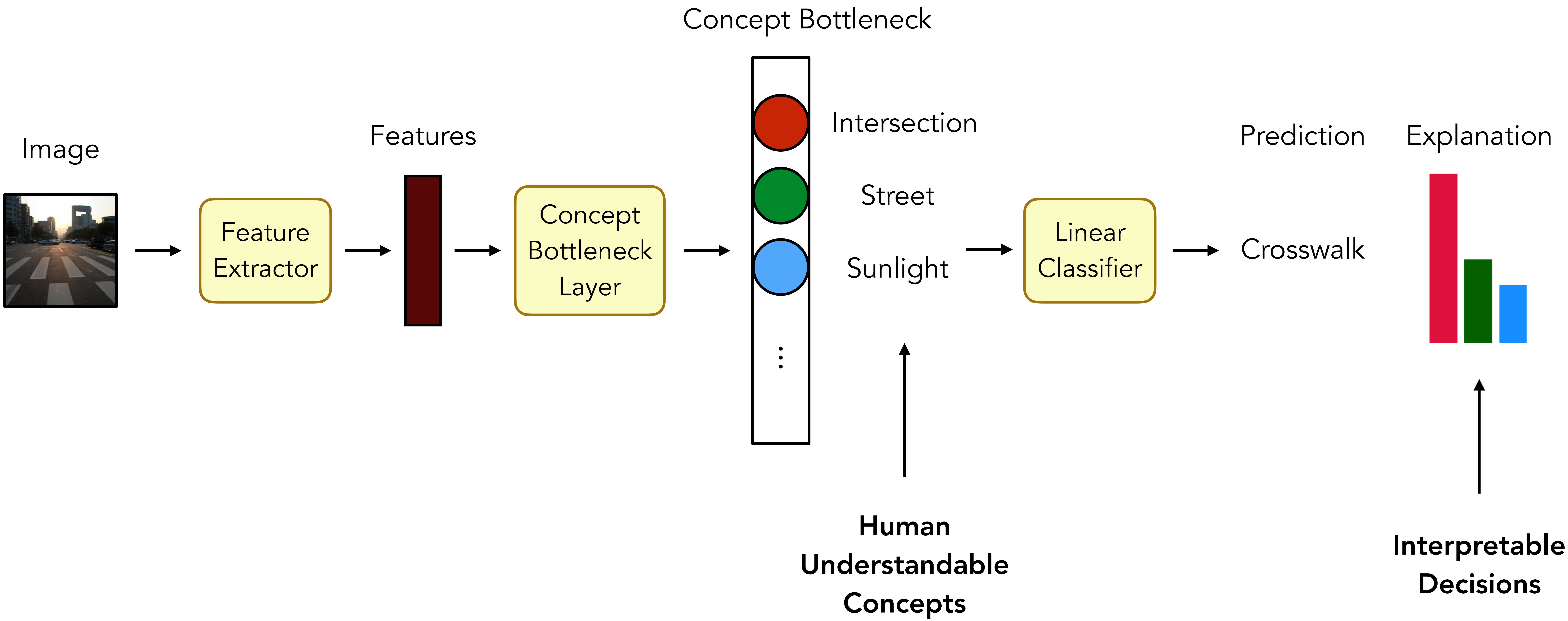
Bernt Schiele

Max Planck Institute for Informatics, Saarland Informatics Campus

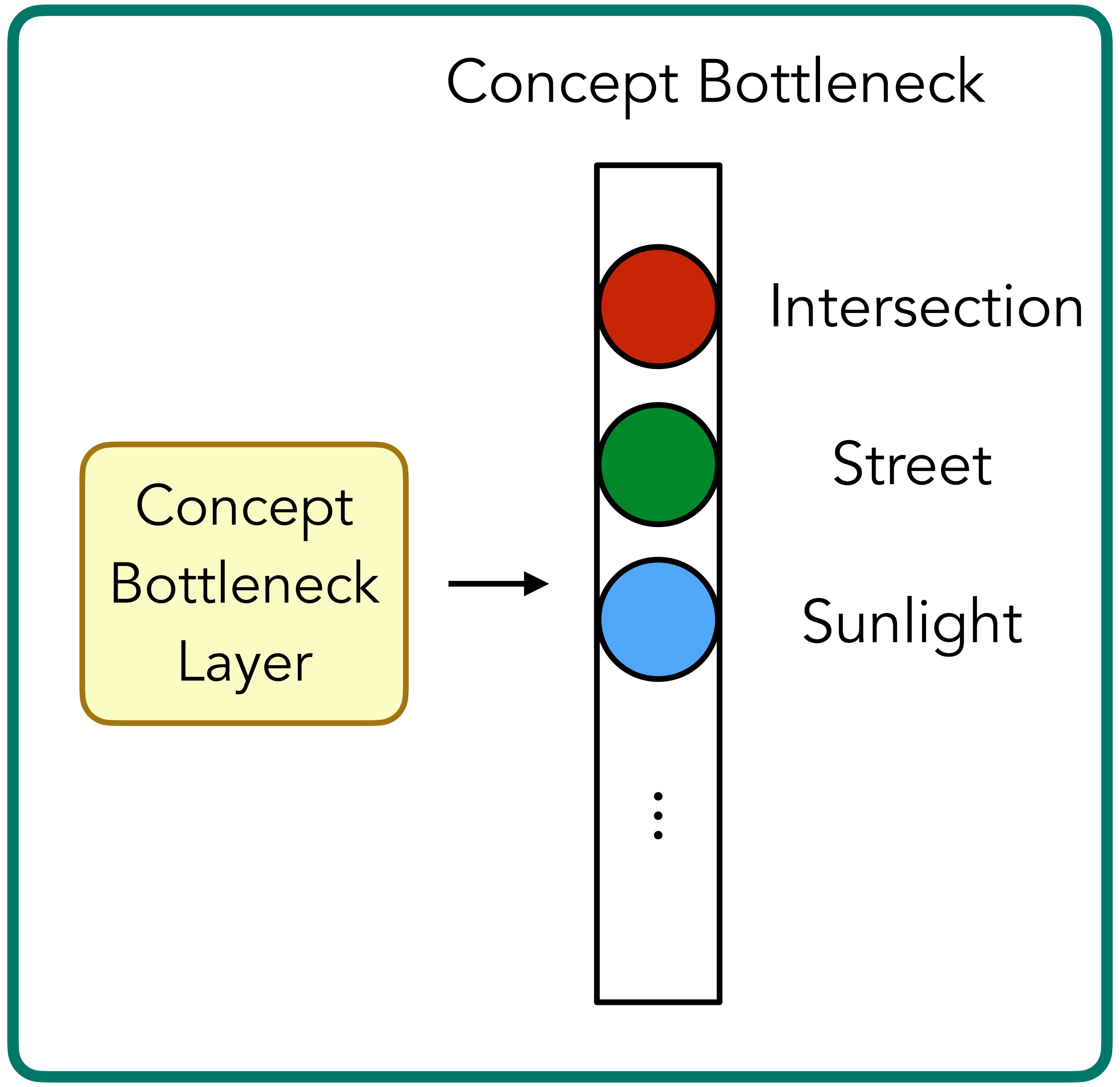
# Concept Bottleneck Models



# Concept Bottleneck Models



# Concept Bottleneck Models

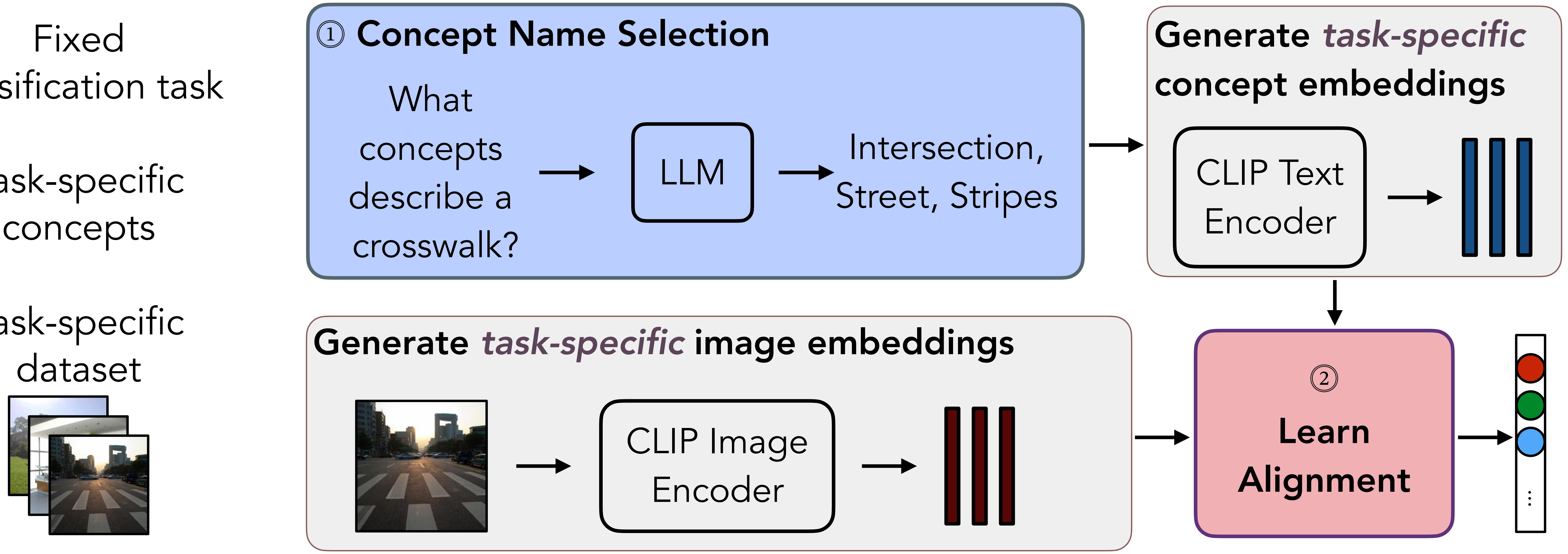


# Typical approach[1]: Select concepts names, learn mapping

Fixed classification task

Task-specific concepts

Task-specific dataset



① Concept Name Selection

What concepts describe a crosswalk? → LLM → Intersection, Street, Stripes

Generate *task-specific* concept embeddings

CLIP Text Encoder →

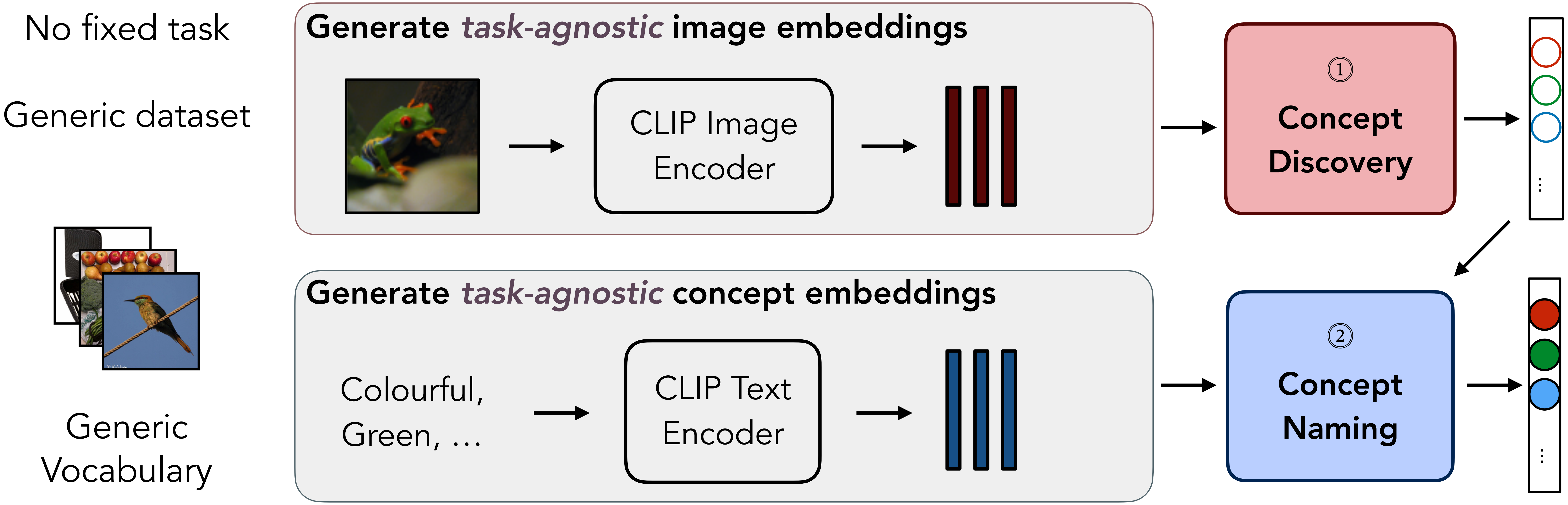
Generate *task-specific* image embeddings

→ CLIP Image Encoder →

② Learn Alignment

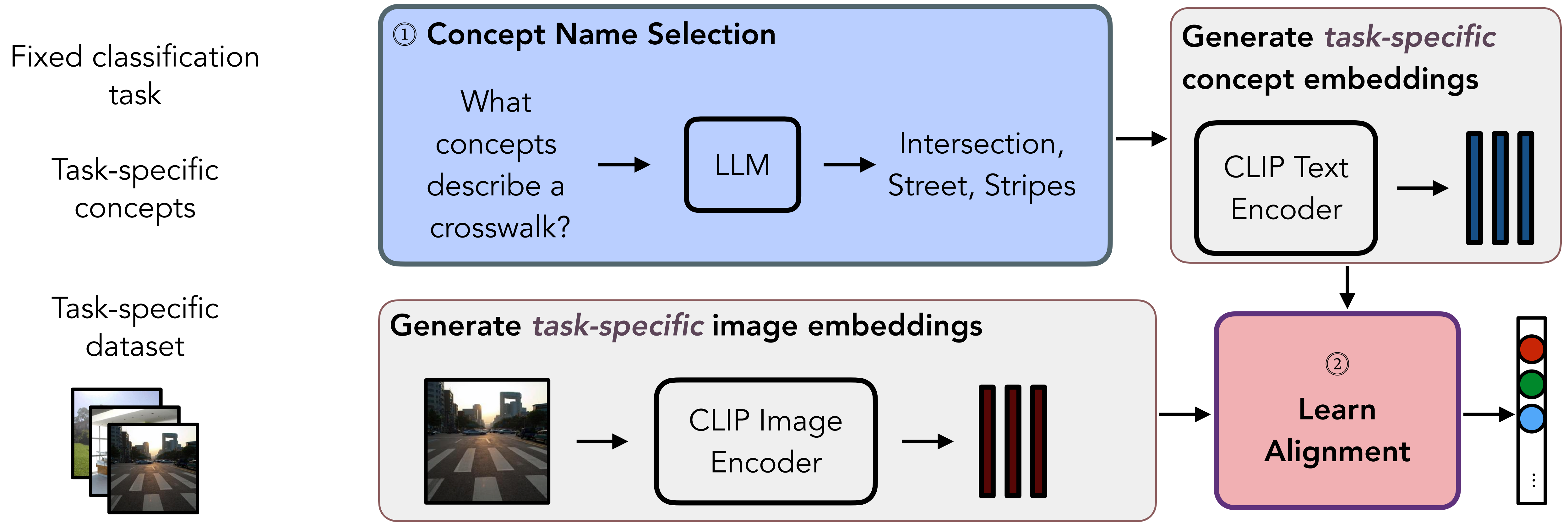
[1] Examples: Label-Free CBM [Oikarinen et al., 2023], LaBo [Yang et al., 2023], CDM [Panousis et al., 2023], DCLIP [Menon et al., 2023]

# Ours: Discover concepts, then assign names



# Overview

## Typical approach: Select concepts names, learn mapping



Fixed classification task

Task-specific concepts

Task-specific dataset

① Concept Name Selection

What concepts describe a crosswalk? → LLM → Intersection, Street, Stripes

Generate *task-specific* concept embeddings

CLIP Text Encoder

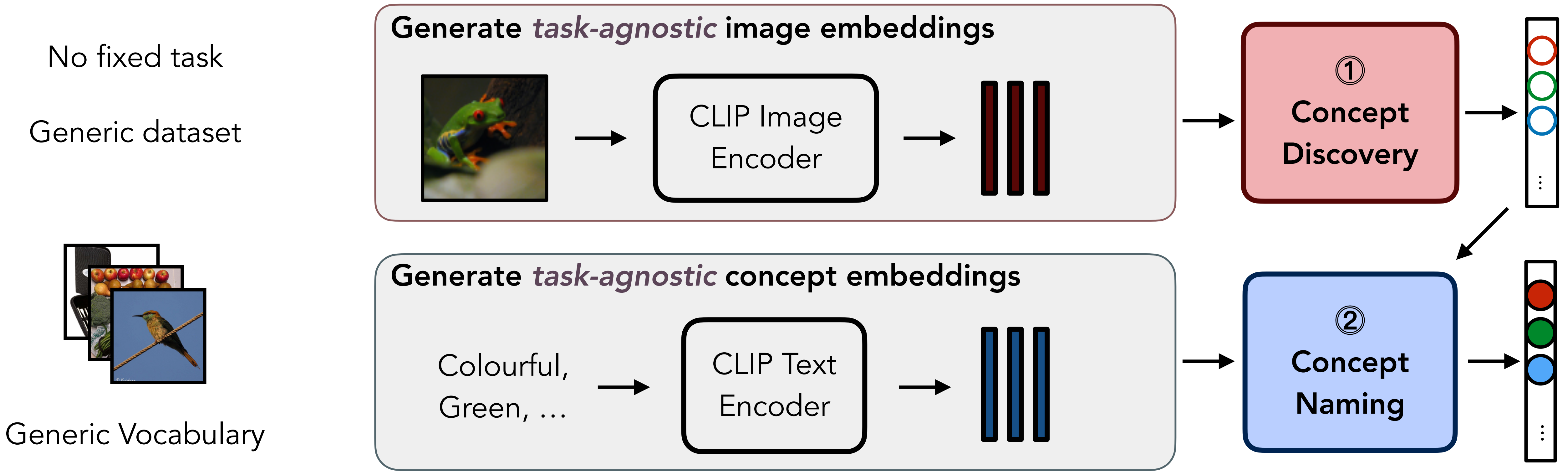
Generate *task-specific* image embeddings

CLIP Image Encoder

② Learn Alignment

- Need to query LLMs for concepts
- Concept bottleneck for single task
- Aligns to predefined concepts

## Ours: Discover concepts, then assign names



No fixed task

Generic dataset

Generic Vocabulary

Generate *task-agnostic* image embeddings

CLIP Image Encoder

① Concept Discovery

Generate *task-agnostic* concept embeddings

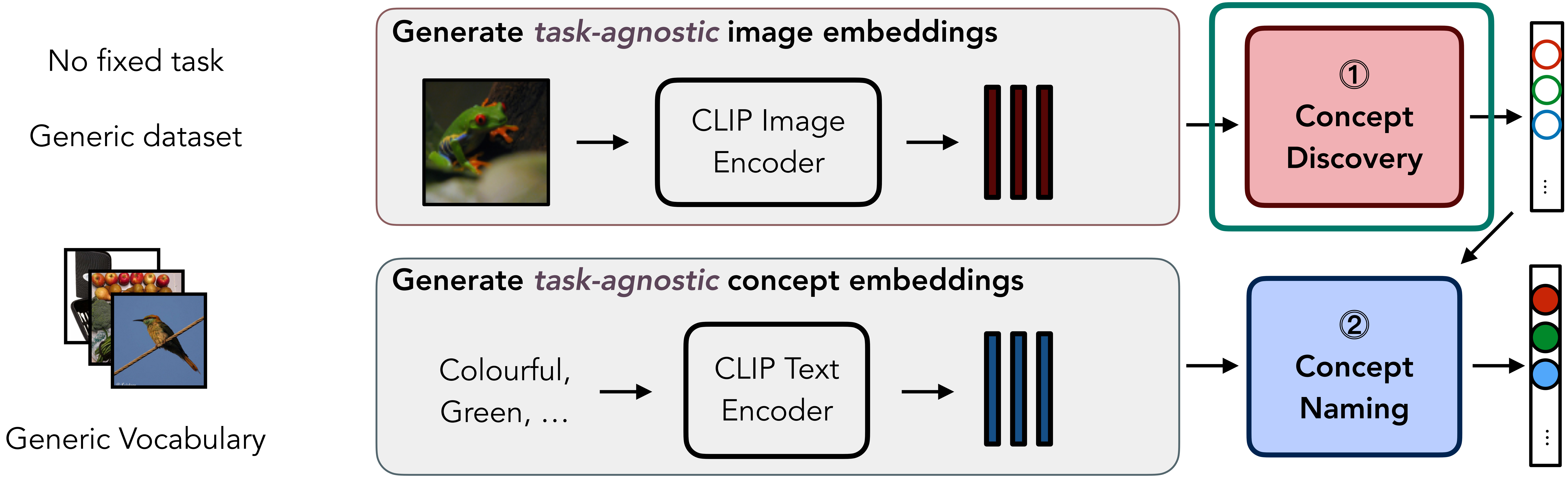
Colourful, Green, …

CLIP Text Encoder

② Concept Naming

- No LLM queries needed
- Single concept bottleneck for multiple datasets
- Identifies concepts used by the model

# Automated Concept Discovery and Naming



No fixed task

Generic dataset

Generic Vocabulary

**Generate *task-agnostic* image embeddings**

CLIP Image Encoder

① Concept Discovery

**Generate *task-agnostic* concept embeddings**

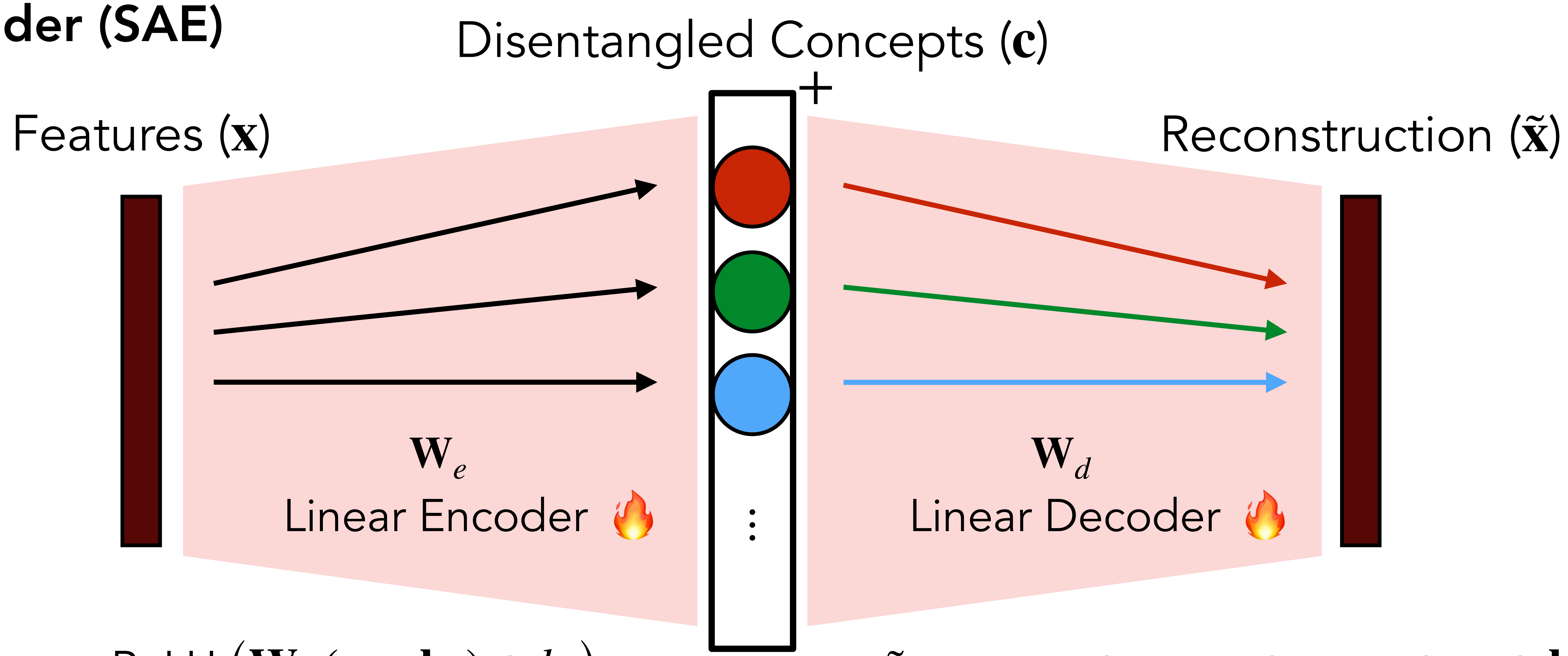
Colourful, Green, …

CLIP Text Encoder

② Concept Naming

# ① Concept Discovery

## Sparse Autoencoder (SAE)



$$\mathbf{c} = \text{ReLU}\left(\mathbf{W}_e(\mathbf{x}-\mathbf{b}_d)+b_e\right)$$

$$\tilde{\mathbf{x}} = c_1\mathbf{w}_{d1} + c_2\mathbf{w}_{d2} + c_3\mathbf{w}_{d3} + \cdots + \mathbf{b_d}$$

$$\underset{\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2}{\mathcal{L}_{recon}} + \lambda \underset{\|\mathbf{c}\|_1}{\mathcal{L}_{sparse}}$$

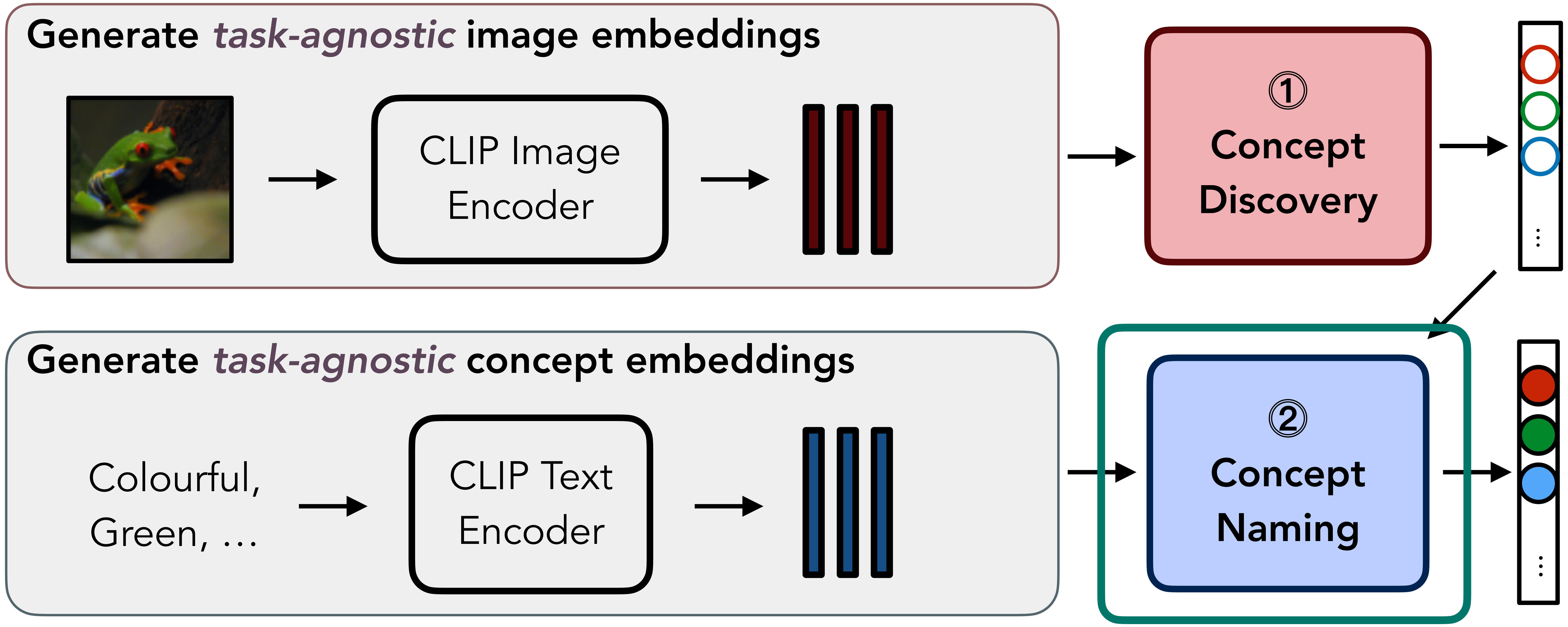
Sparse Autoencoder: Bricken et al. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread, 2023.

# Automated Concept Discovery and Naming

No fixed task

Generic dataset

Generic Vocabulary



**Generate *task-agnostic* image embeddings**

CLIP Image Encoder

① **Concept Discovery**

**Generate *task-agnostic* concept embeddings**

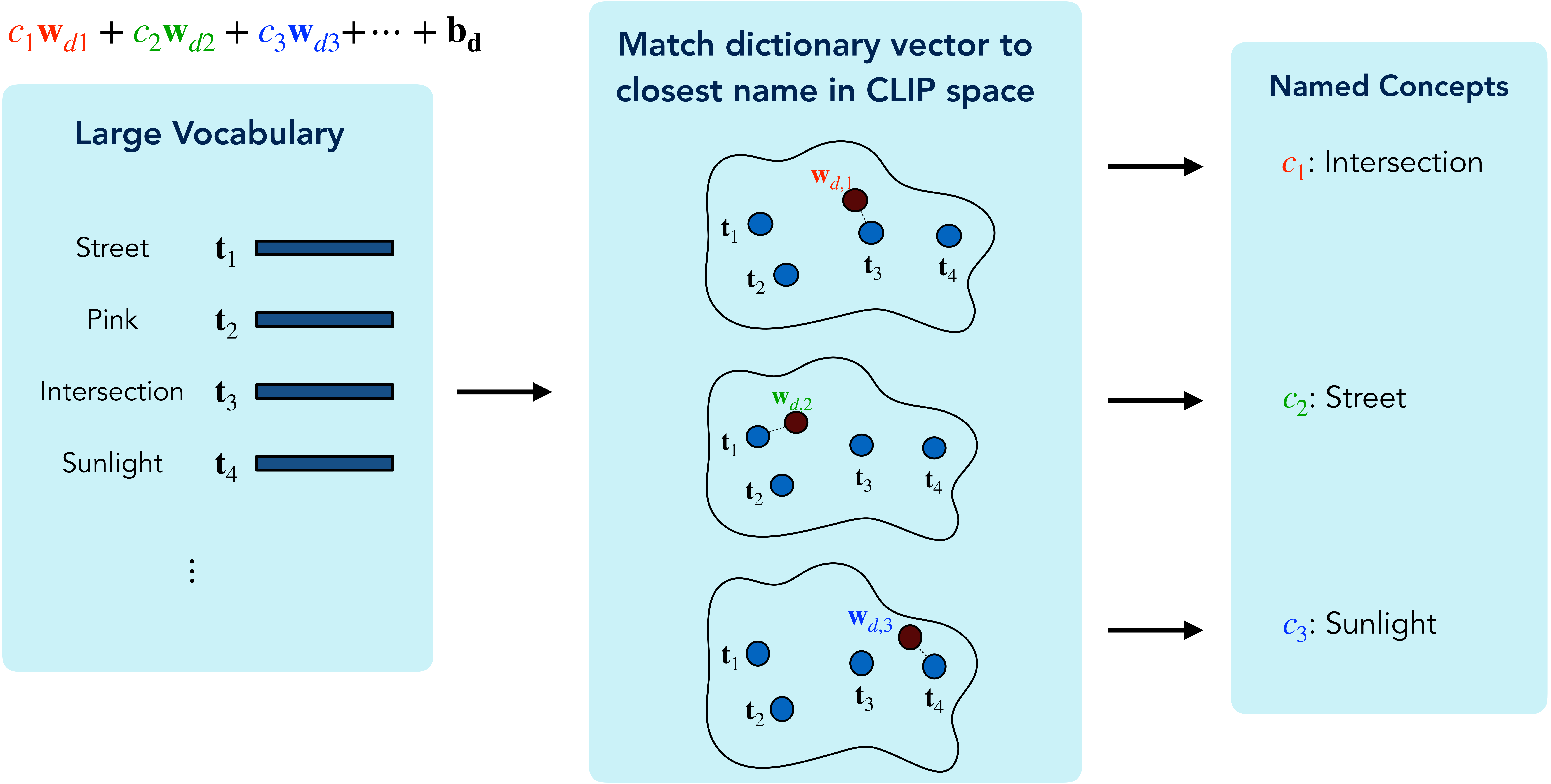
Colourful, Green, …

CLIP Text Encoder

② **Concept Naming**

# ② Concept Naming

$$\tilde{\mathbf{x}} = c_1\mathbf{w}_{d1} + c_2\mathbf{w}_{d2} + c_3\mathbf{w}_{d3} + \cdots + \mathbf{b_d}$$



## Large Vocabulary

Street $\mathbf{t}_1$

Pink $\mathbf{t}_2$

Intersection $\mathbf{t}_3$

Sunlight $\mathbf{t}_4$

⋮

## Match dictionary vector to closest name in CLIP space

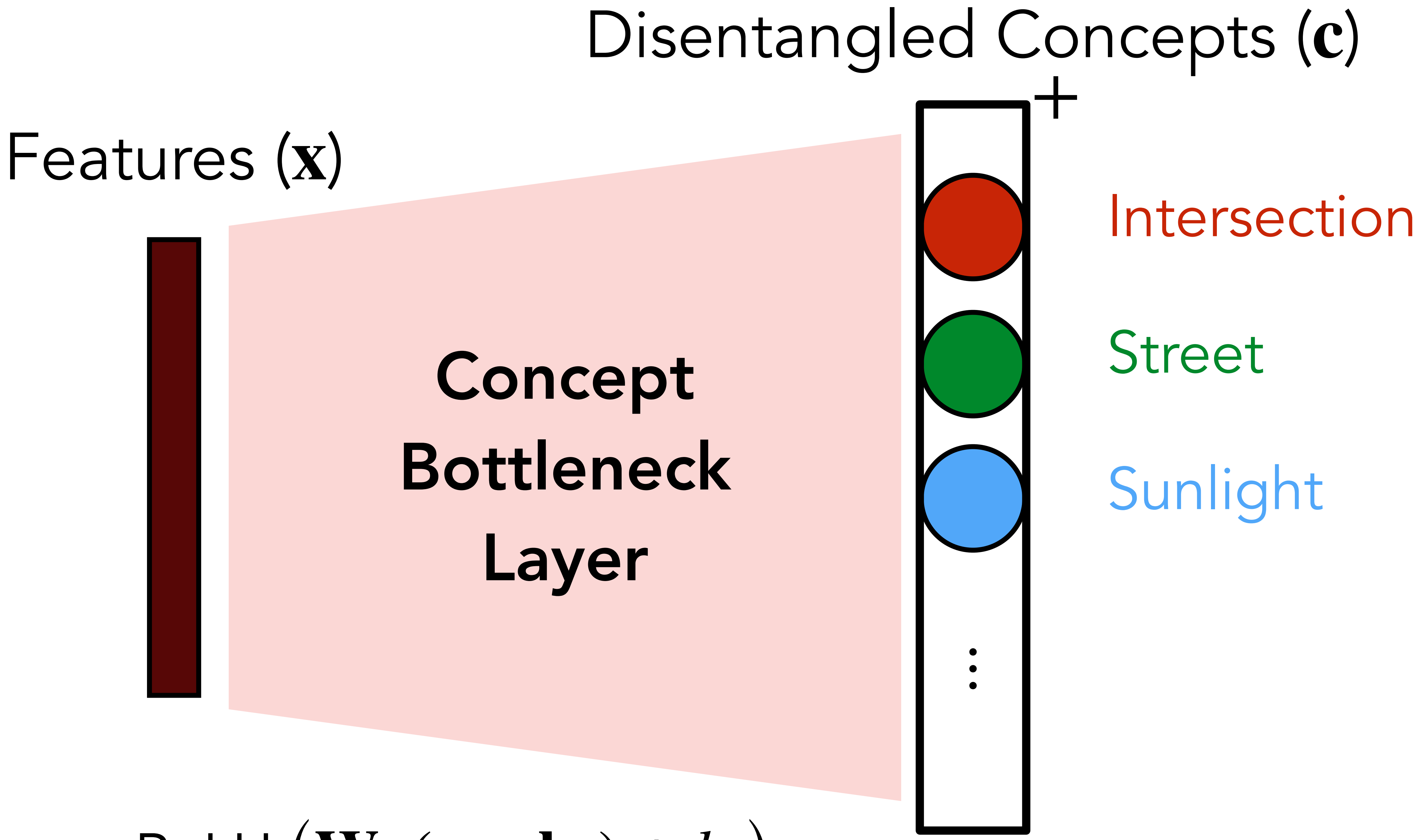
## Named Concepts

$c_1$: Intersection

$c_2$: Street

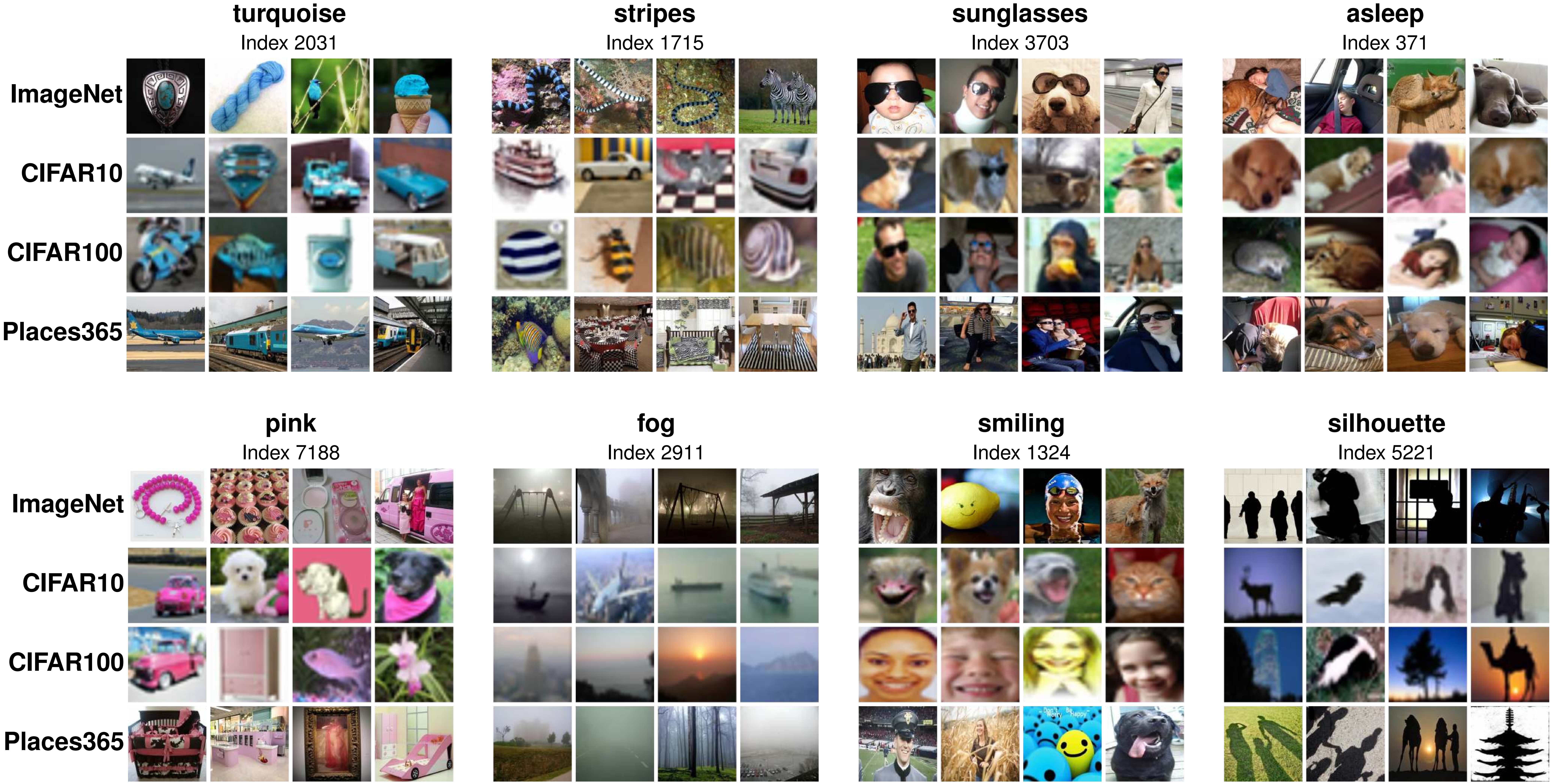
$c_3$: Sunlight

# Concept Bottleneck Layer

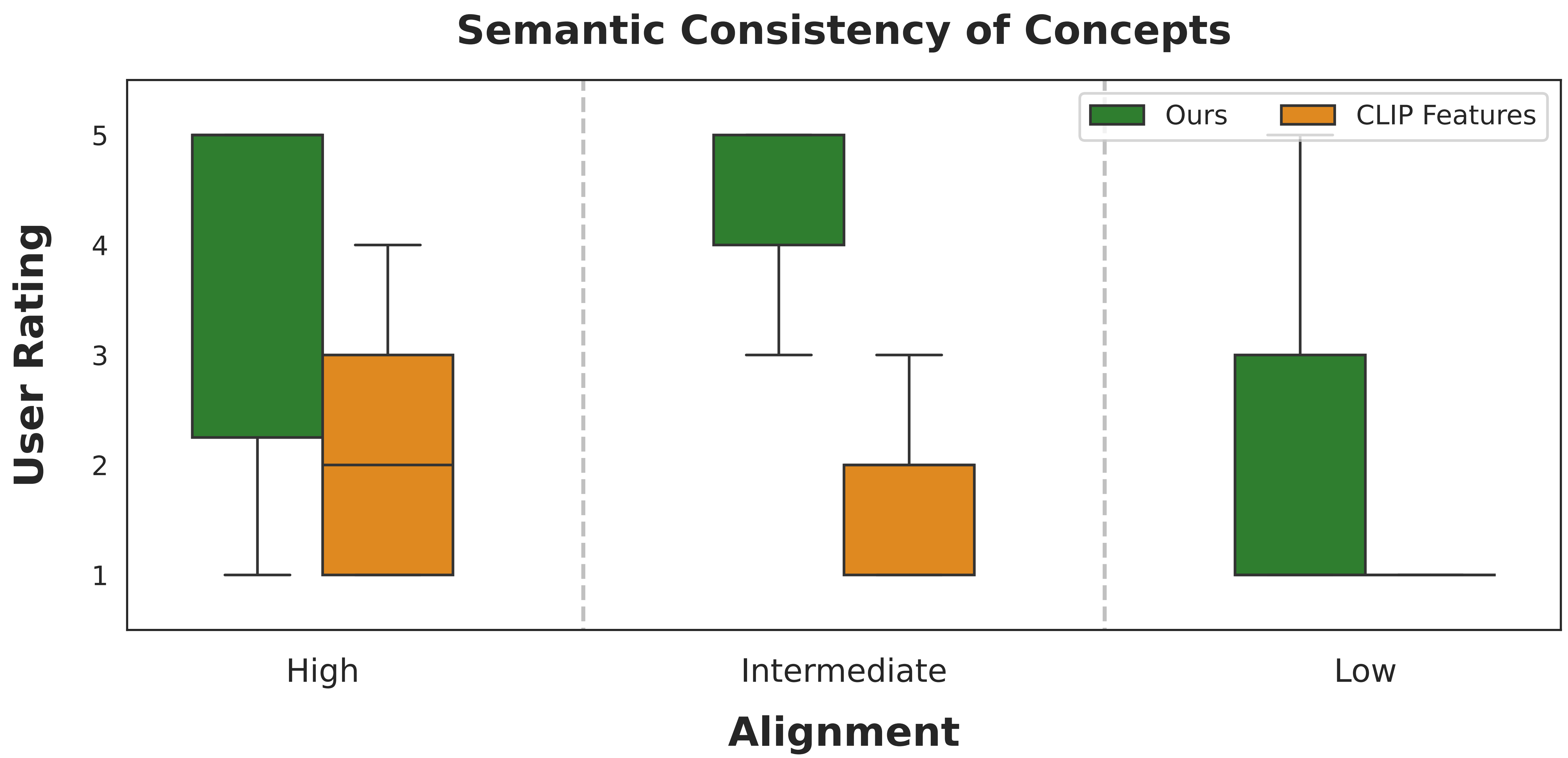


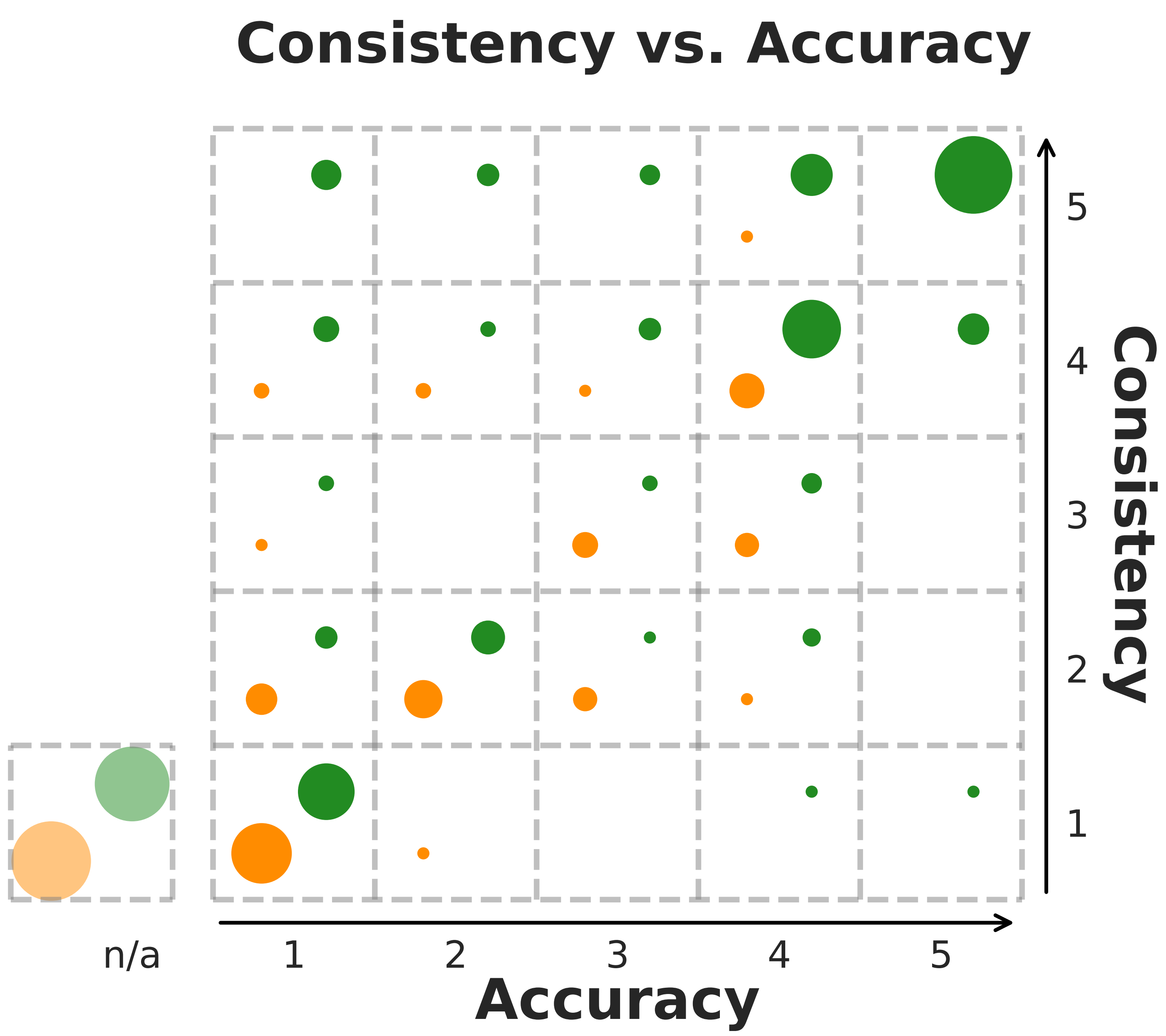
$$\mathbf{c} = \text{ReLU}\left(\mathbf{W}_e(\mathbf{x} - \mathbf{b}_d) + b_e\right)$$

# Consistent and Interpretable Concepts



# Consistent and Interpretable Concepts: User Study



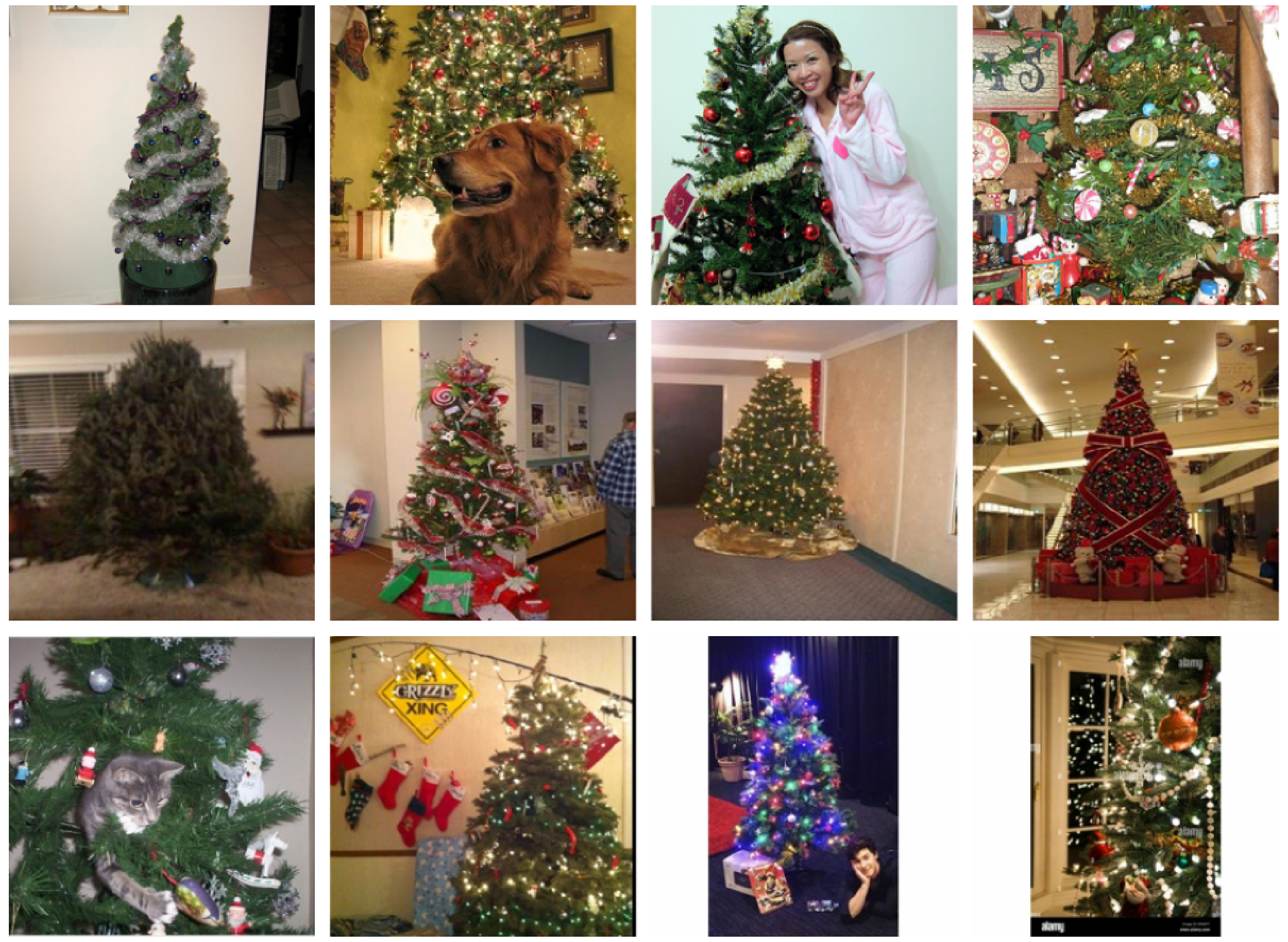


- Better semantic consistency than CLIP features
- High name accuracy for semantically consistent concepts

# Granularity Controllable by Vocabulary

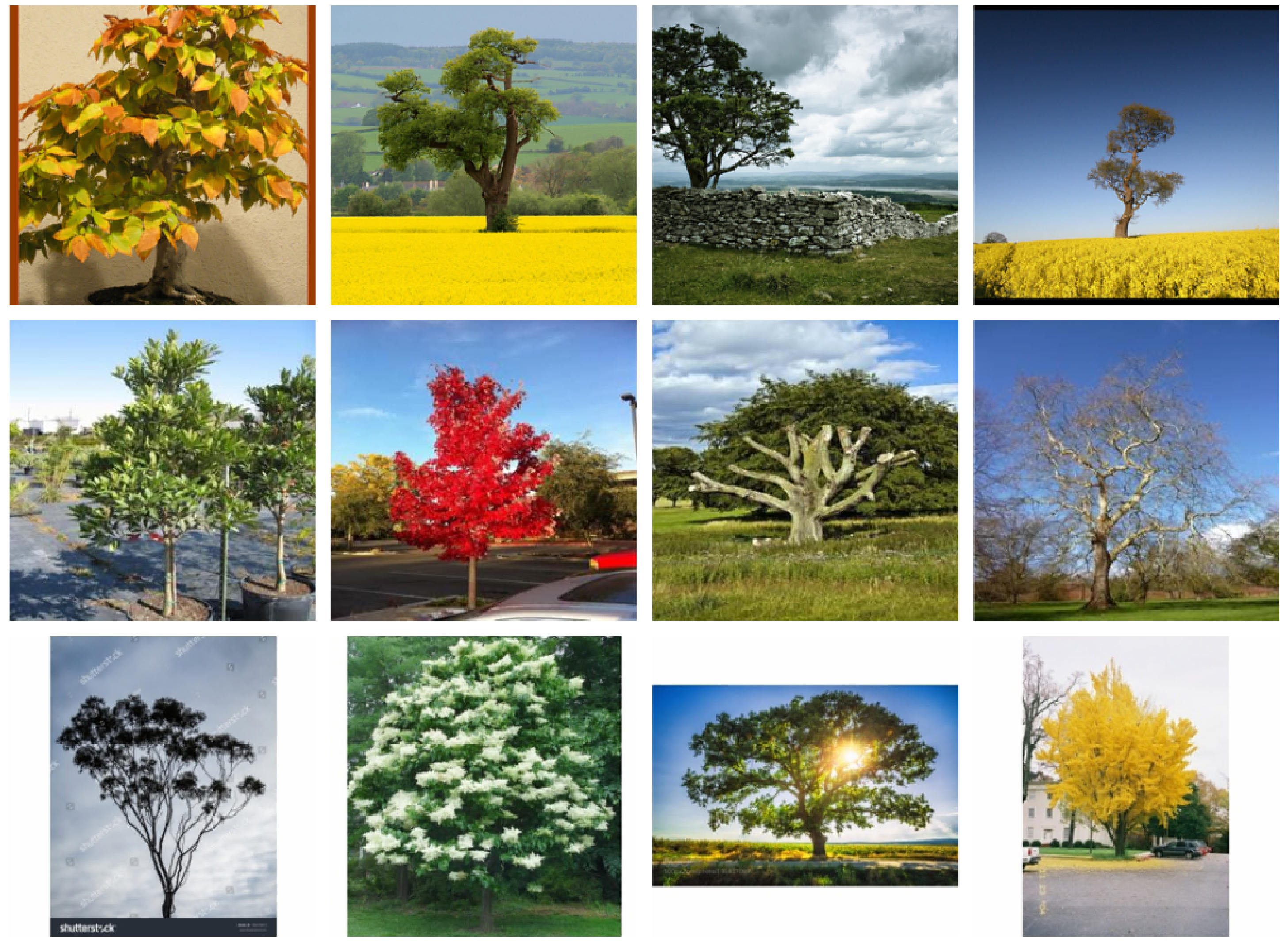
**tree → christmas tree**

Index 7446

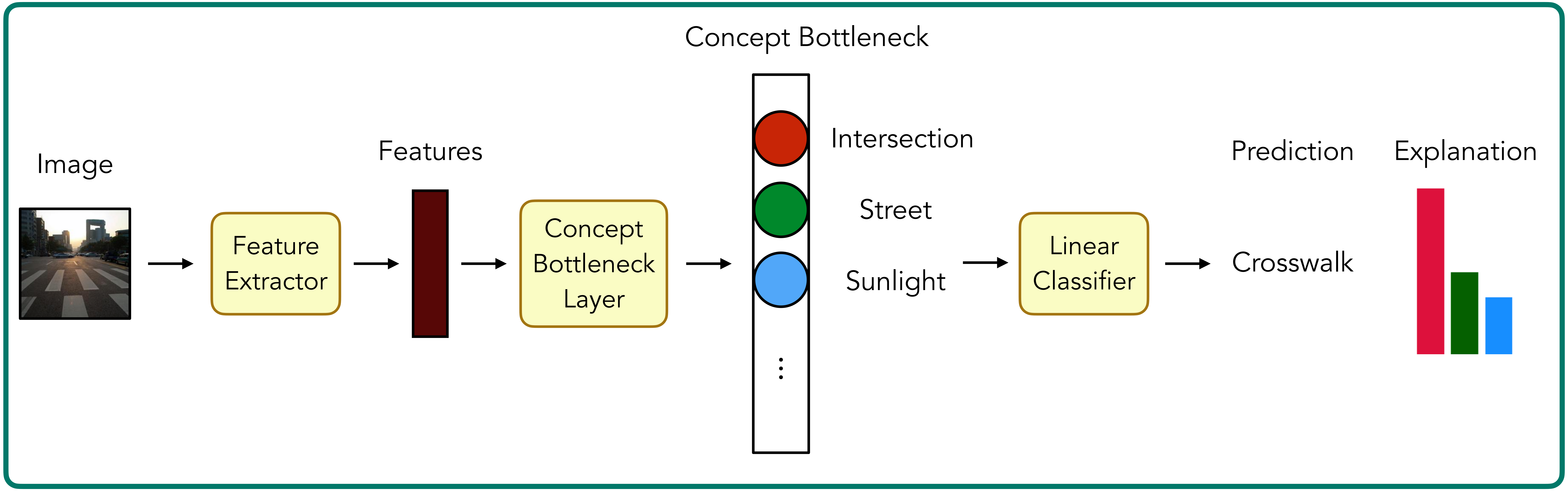


**tree → tree in field**

Index 8167



# Concept Bottleneck Models: DN-CBM



# DN-CBM: Results

- Classification Performance

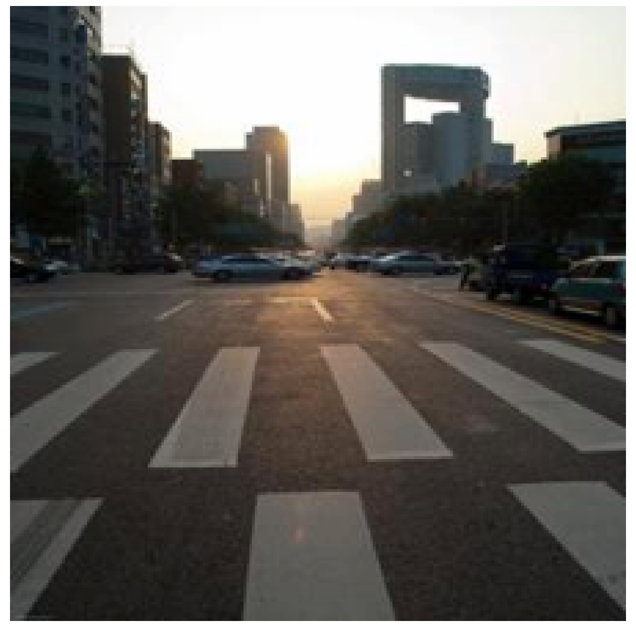
| Model | CLIP ResNet-50 | | | | CLIP ViT-B/16 | | | |
|---|---|---|---|---|---|---|---|---|
| | Places365 | ImageNet | CIFAR10 | CIFAR100 | Places365 | ImageNet | CIFAR10 | CIFAR100 |
| **Linear Probe** | 53.4 | 73.3 | 88.7 | 70.3 | 55.1 | 80.2 | 96.2 | 83.1 |
| **Zero Shot** | 38.7 | 59.6 | 75.6 | 41.6 | 41.2 | 68.6 | 91.6 | 68.7 |
| **LF-CBM** | 49.0 | 67.5 | 86.4 | 65.1 | 50.6 | 75.4 | 94.6 | 77.4 |
| **LaBo** | - | 68.9 | **87.9** | **69.1** | - | 78.9 | 95.7 | 81.2 |
| **CDM** | 52.7 | 72.2 | 86.5 | 67.6 | 52.6 | 79.3 | 95.3 | 80.5 |
| **DCLIP** | 37.9 | 59.6 | - | - | 40.3 | 68.0 | - | - |
| **DN-CBM (Ours)** | **53.5** | **72.9** | 87.6 | 67.5 | **55.1** | **79.5** | **96.0** | **82.1** |

CLIP [Radford et al., 2021], LF-CBM [Oikarinen et al., 2023], LaBo [Yang et al., 2023], CDM [Panousis et al., 2023], DCLIP [Menon et al., 2023].

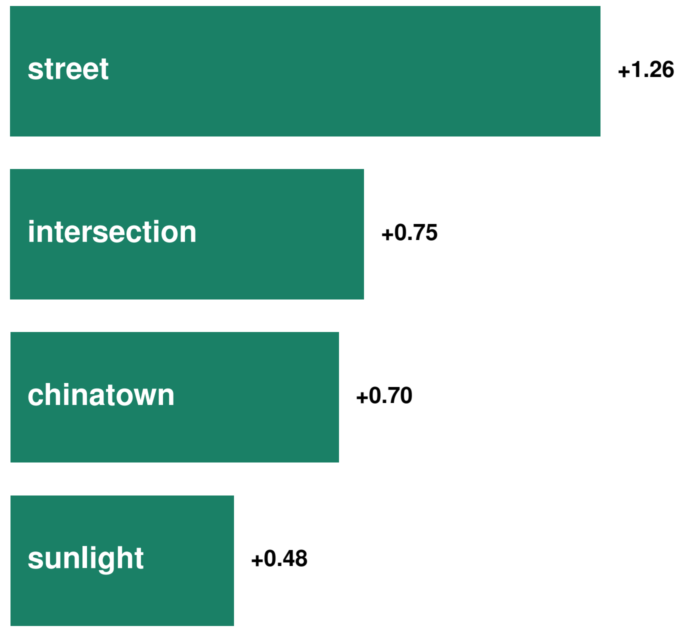
# DN-CBM: Results

- Classification Performance
- **Explanations for Decisions**

Crosswalk

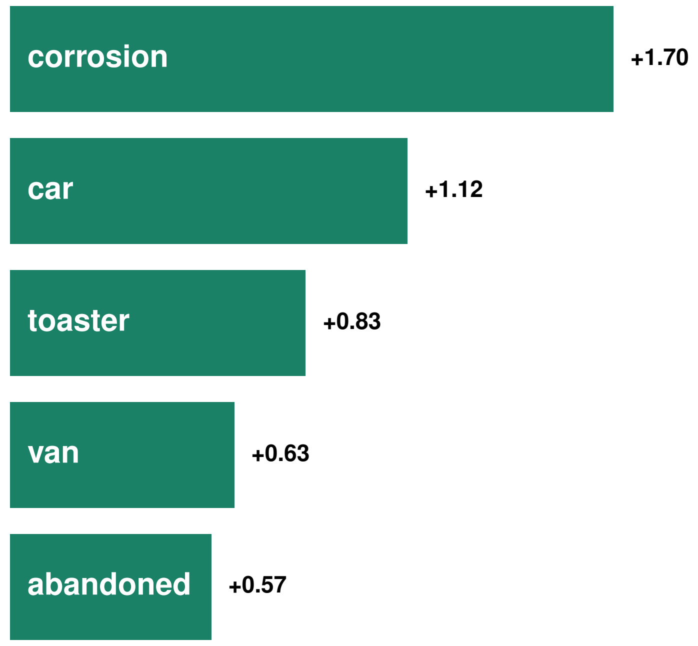


Top concepts



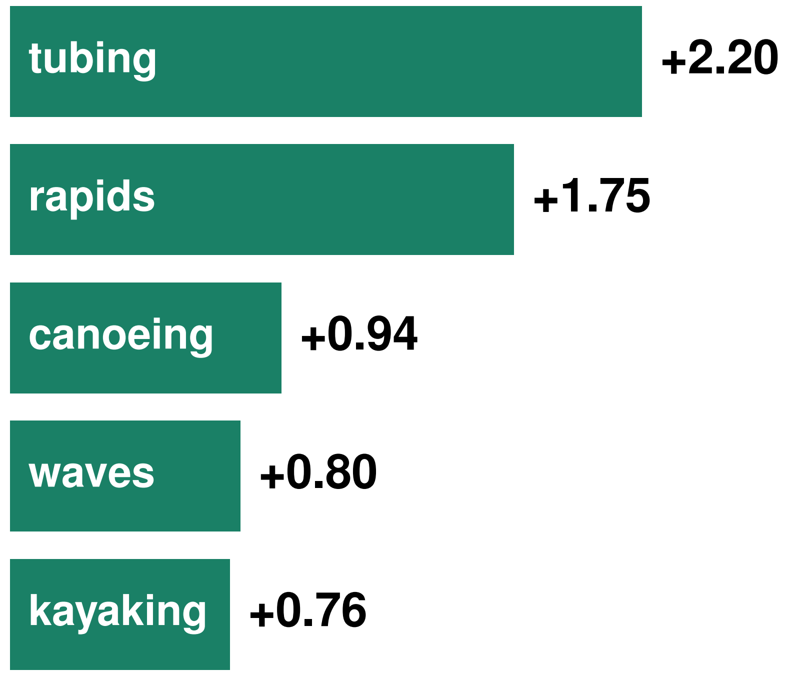
Junkyard

Top concepts

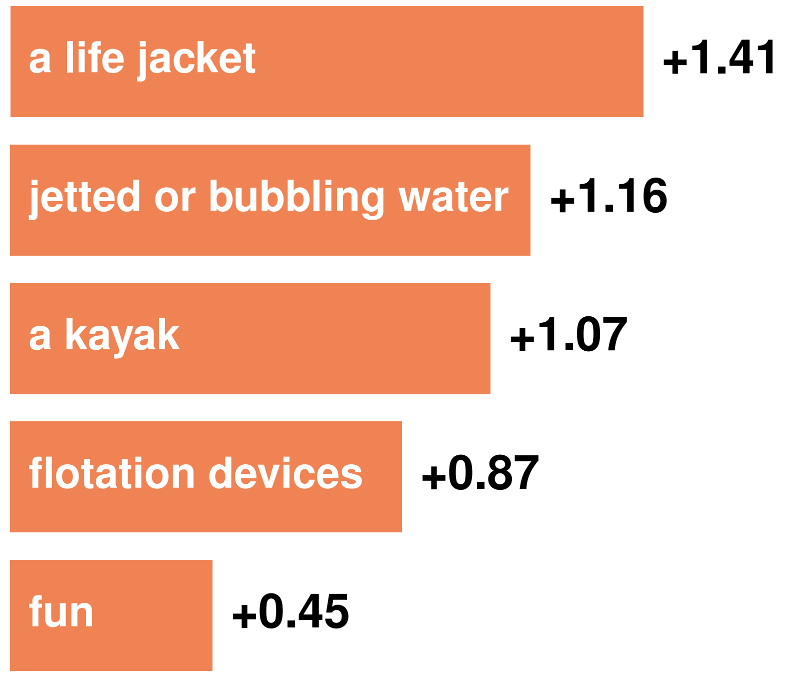


Raft

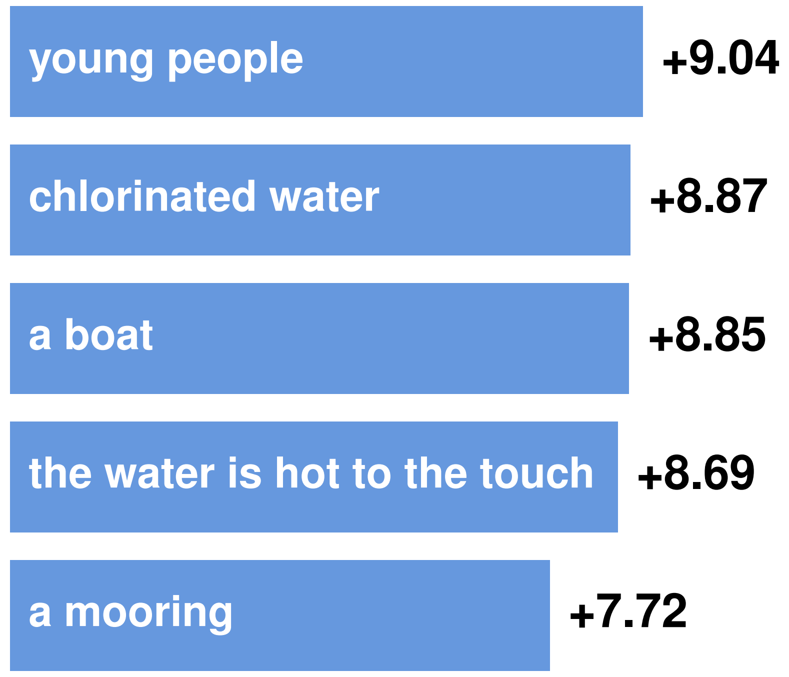
Ours



LF-CBM



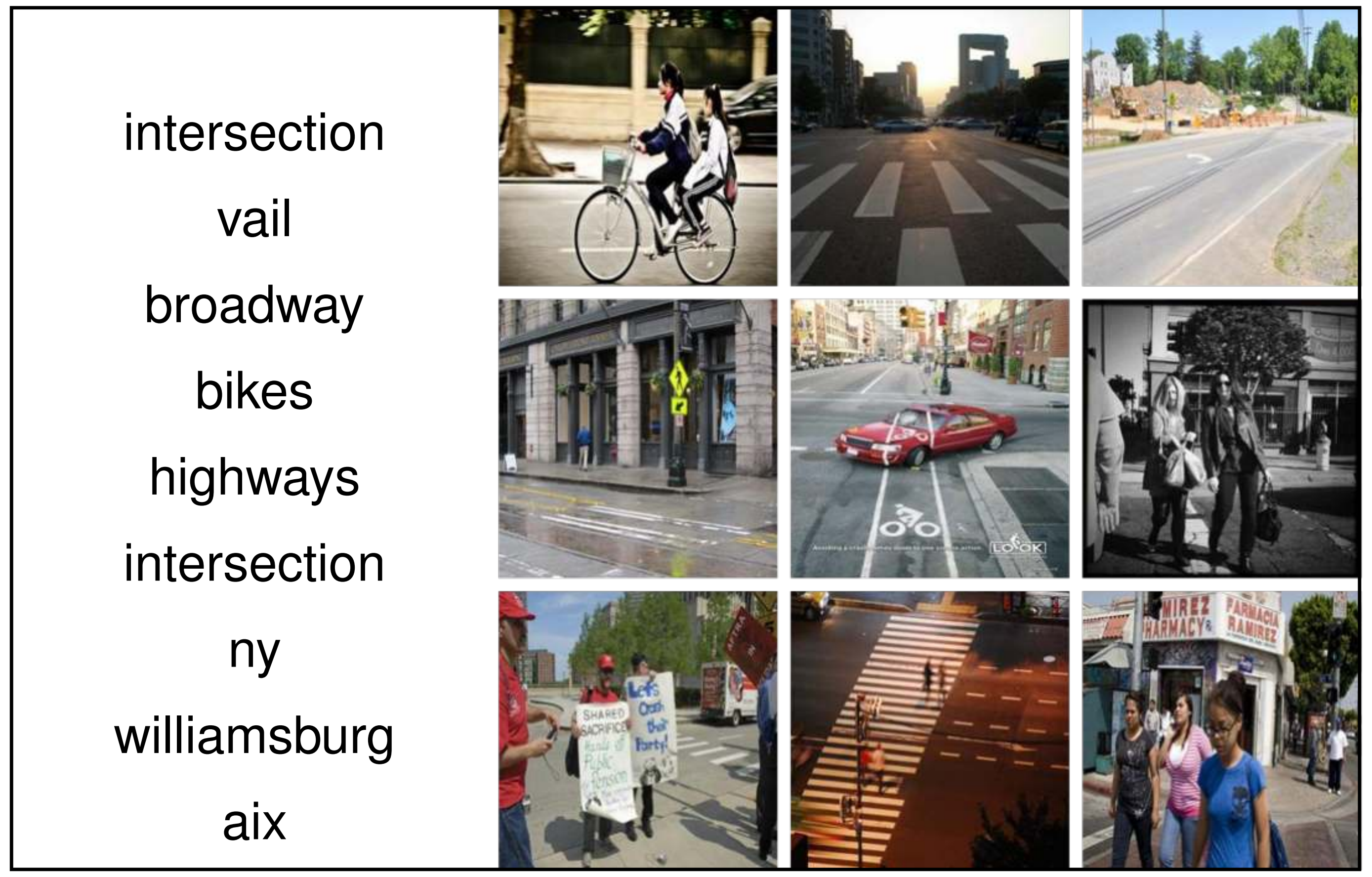
CDM



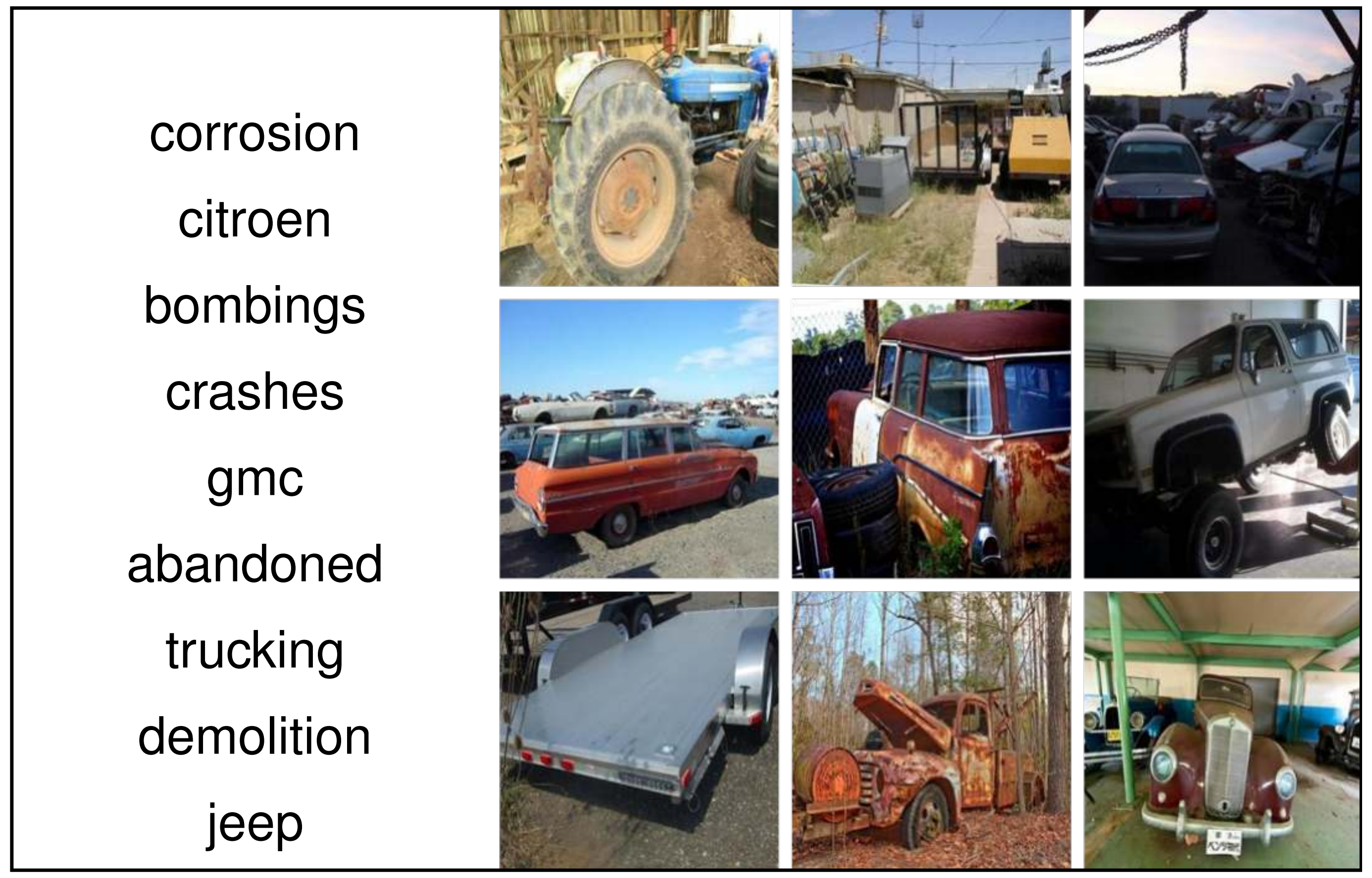
# DN-CBM: Results

- Classification Performance
- Explanations for Decisions
- **Class-level Explanations**

Crosswalk



Junkyard



# DN-CBM: Results

- Classification Performance
- Explanations for Decisions
- Class-level Explanations
- **Effective Interventions**

**Training Groups**

Landbird on Land

Waterbird on Water

**Test-only (Worst) Groups**

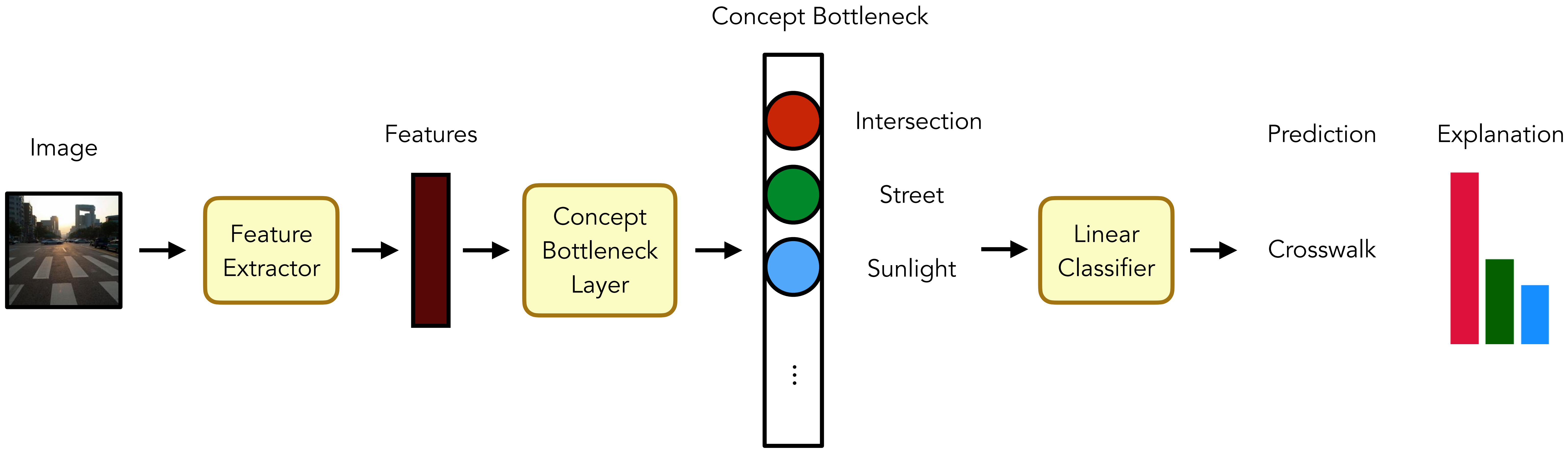
Landbird on Water

Waterbird on Land

| | **Bird concepts** | **Non-bird concepts** |
|---|---|---|
| **Landbird** | sparrow, parrot, crow | forest, clic |
| **Waterbird** | gull, ducks | landing, beach, canoeing |

| Model | Overall | Worst Groups | |
|---|---|---|---|
| | | Landbird on Water | Waterbird on Land |
| **Before Intervention** | 82.8 | 71.3 | 57.5 |
| **Only Bird Concepts** | **89.4 (+6.6)** | **86.6 (+15.3)** | **71.3 (+13.8)** |
| **Only Non-bird Concepts** | 60.8 (-22.0) | 28.5 (-42.8) | 28.8 (-28.7) |

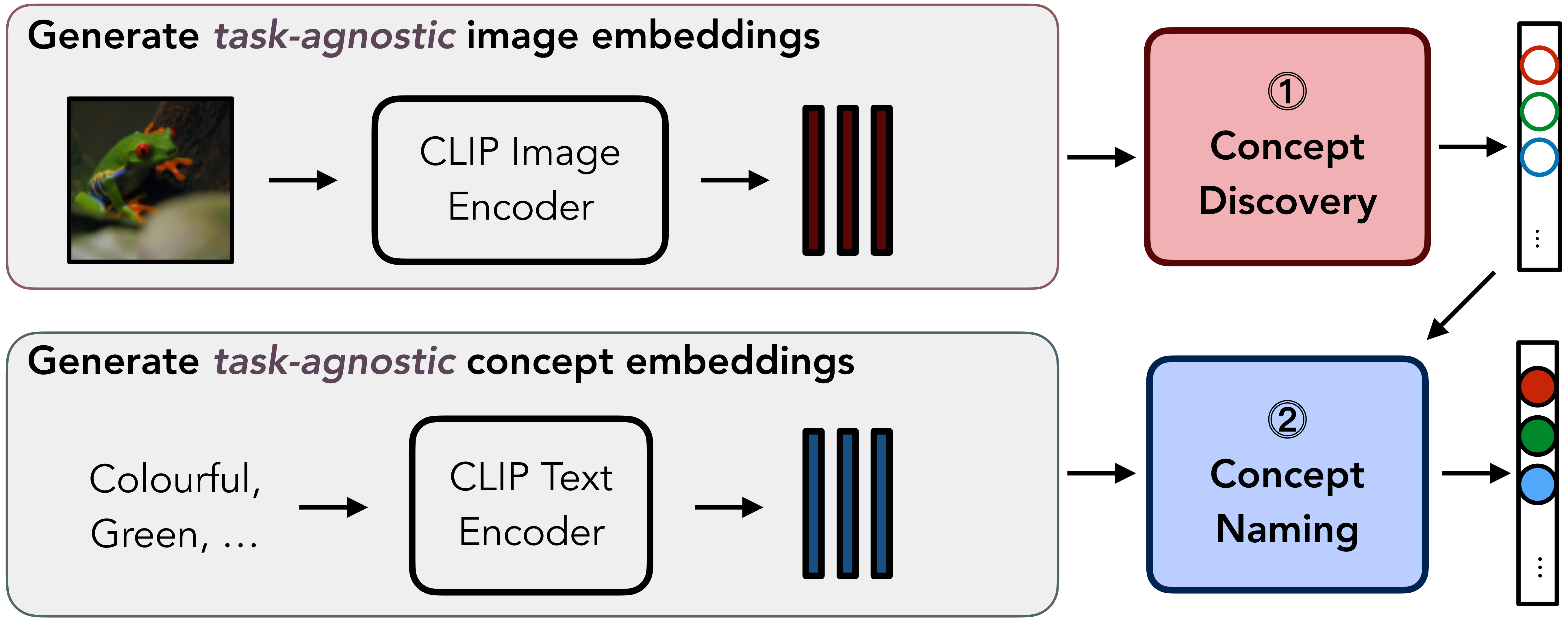
# Summary



No fixed task

Generic dataset

Generic Vocabulary



# Thank you!

- **Poster Session:** 7
- **Date and Time:** October 4, 2024, 10:30 AM – 12:30 PM

## Paper

https://arxiv.org/abs/2407.14499

## Code

https://github.com/neuroexplicit-saar/Discover-then-Name