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

MAX PLANCK INSTITUTE FOR INFORMATICS

Sukrut Rao*[1,2], Sweta Mahajan*[1,2], Moritz Böhle[1], Bernt Schiele[1]

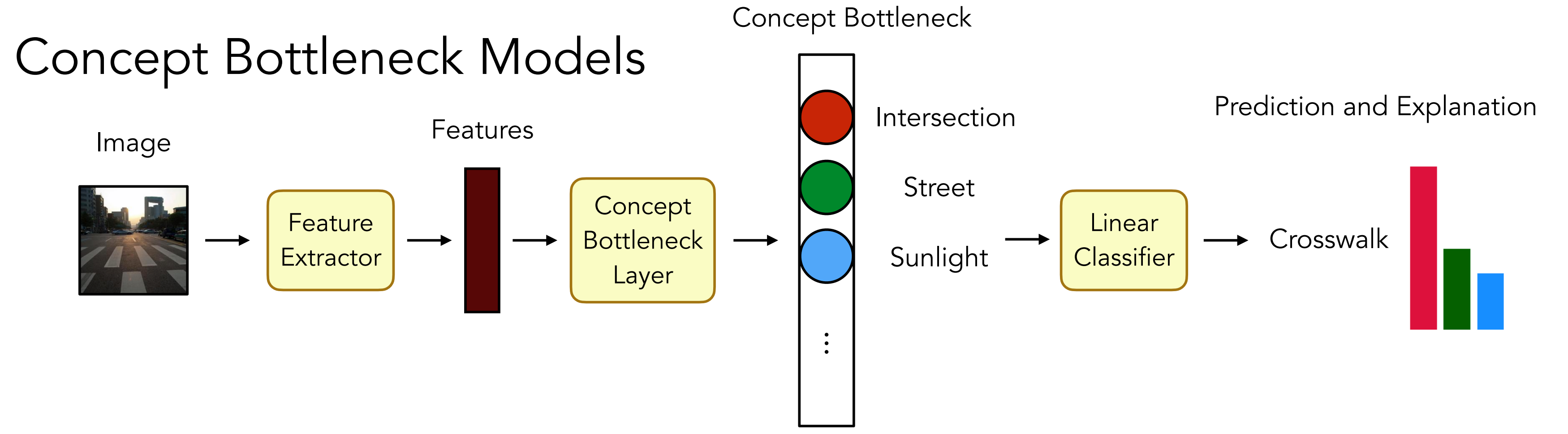
[1]Max Planck Institute for Informatics, Saarland Informatics Campus

[2]RTG Neuroexplicit Models of Language, Vision, and Action, Saarbrücken

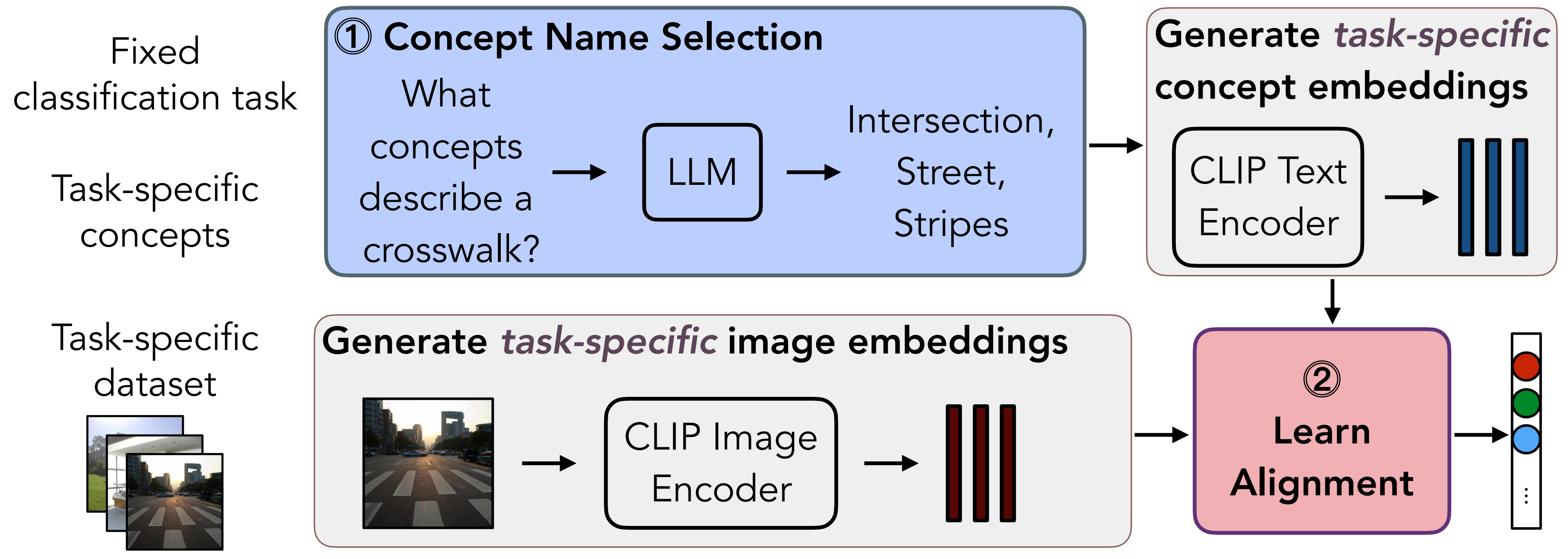
SIC Saarland Informatics Campus

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION MILANO 2024

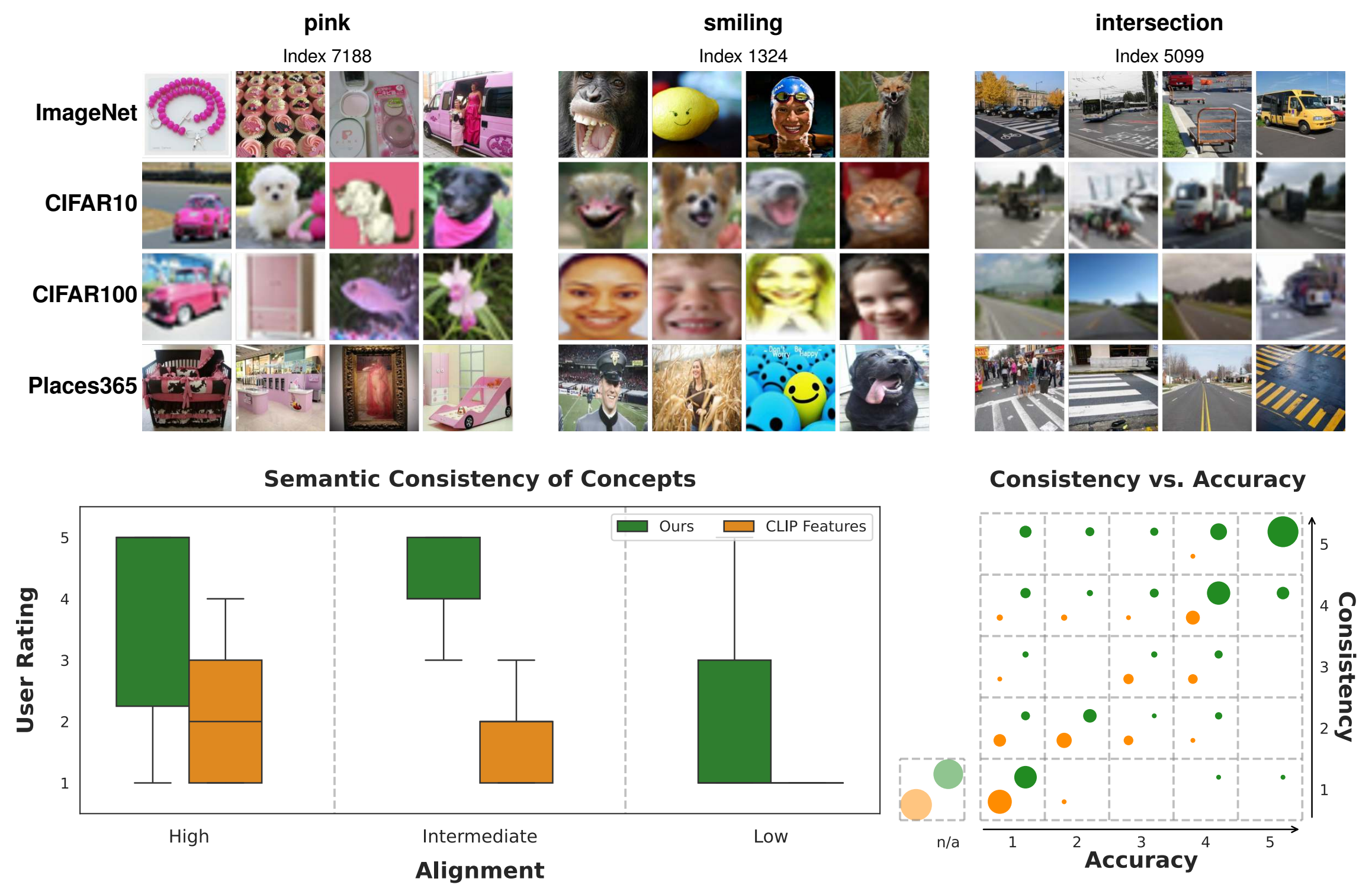
Background



Typical approach: Select concept 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

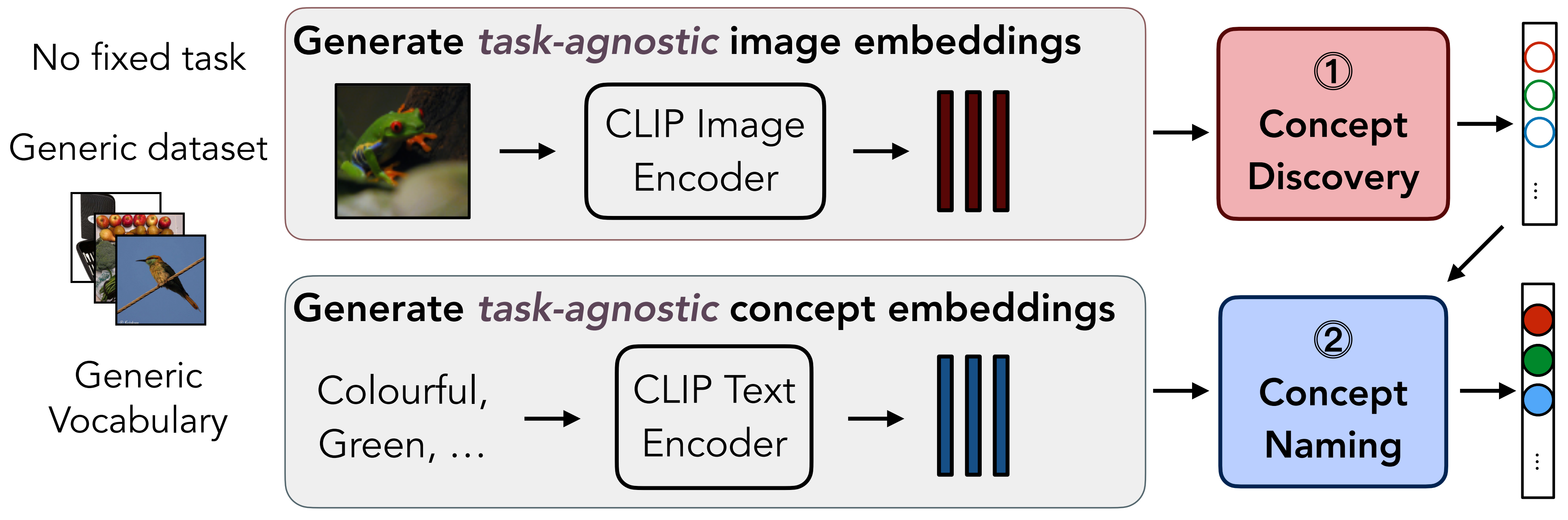
Consistent and interpretable concepts



Granularity controllable by vocabulary

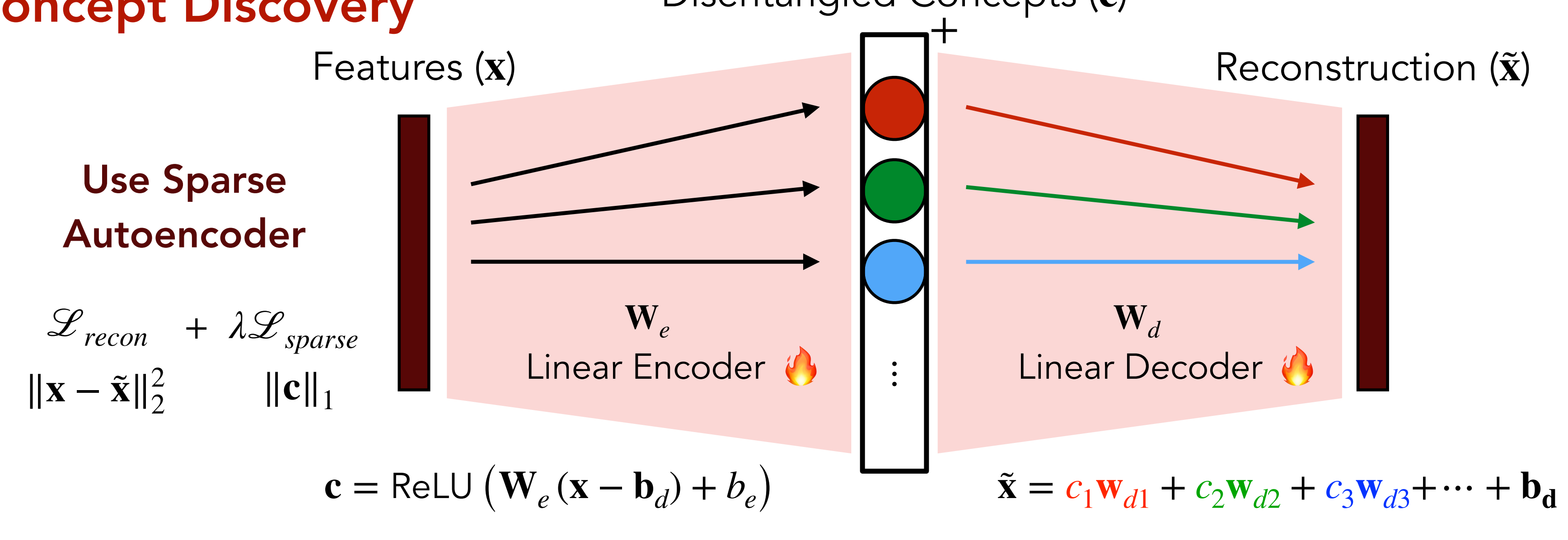
branches ← tree → tree in field
Index 8167
ornaments ← tree → christmas tree
Index 7446

Ours: Discover concepts, then assign names



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

① Concept Discovery



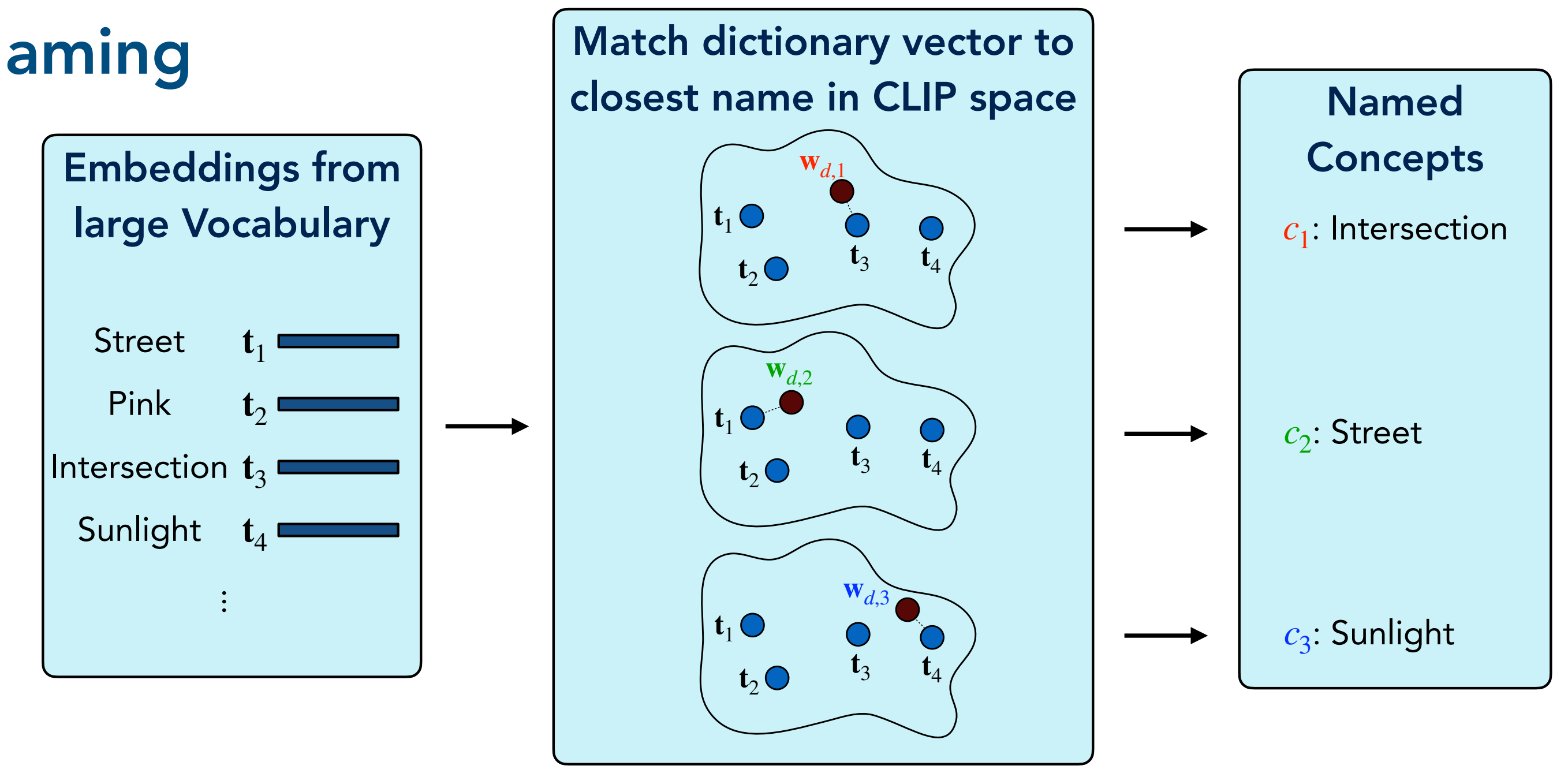
$$\mathcal{L}_{recon} + \lambda \mathcal{L}_{sparse}$$

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2 \qquad \|\mathbf{c}\|_1$$

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

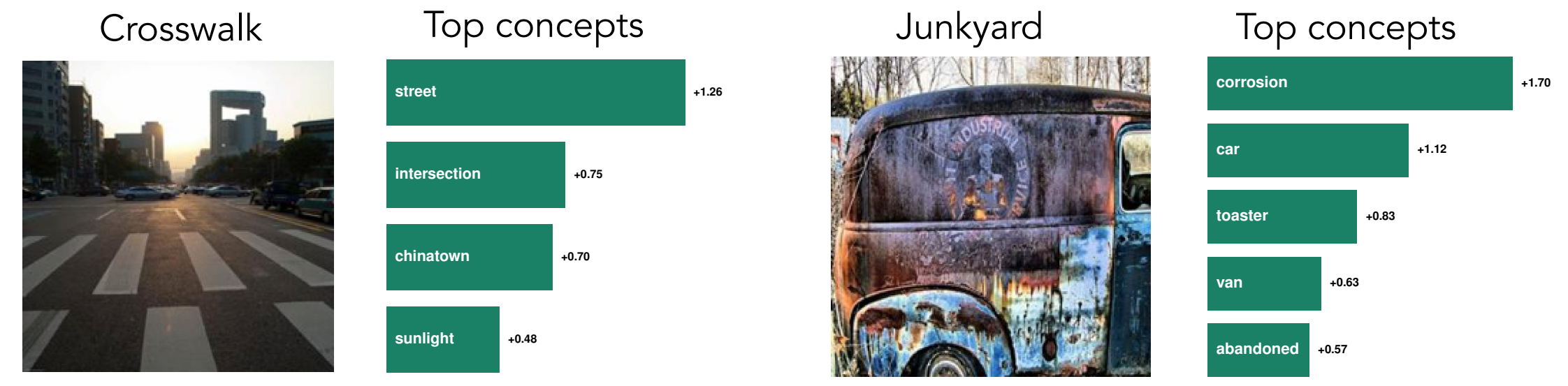
$$\tilde{\mathbf{x}} = c_1\mathbf{w}_{d1} + c_2\mathbf{w}_{d2} + c_3\mathbf{w}_{d3} + \cdots + \mathbf{b_d}$$

② Concept Naming

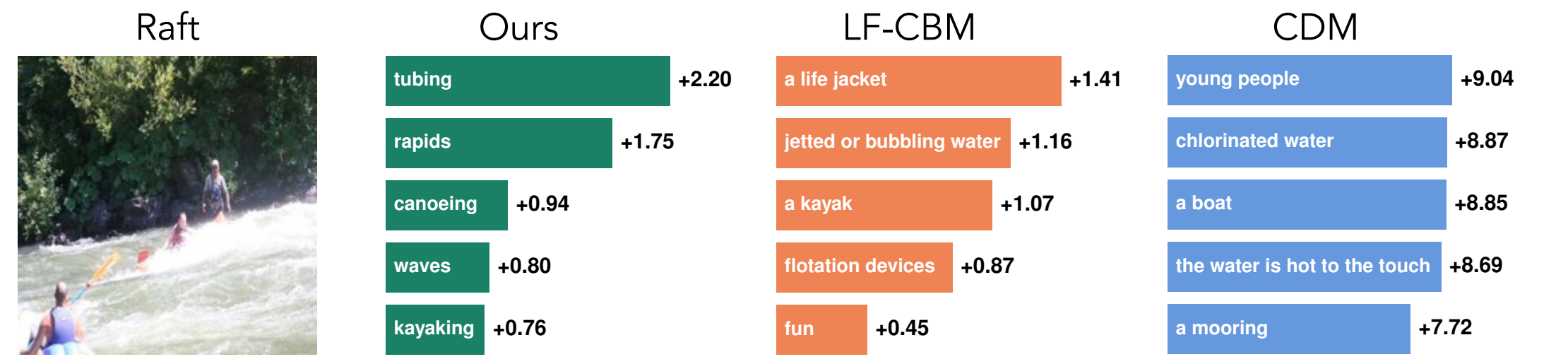


Concept Bottleneck Model: DN-CBM

Explanations for model decisions



Similar explanations to prior work



Class-level explanations

Competitive classification accuracy

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**

Effective Interventions

Pruned Concepts	
Landbird: Bird	sparrow, parrot, crow
Landbird: Non-bird	forest, clic
Waterbird: Bird	gull, ducks
Waterbird: Non-bird	landing, beach,

Model	Overall	Worst Groups		Training Groups	
		Landbird on Water	Waterbird on Land	Landbird on Land	Waterbird on Water
Before Intervention	82.8	71.3	57.5	**98.6**	**93.3**
Only Bird Concepts	**89.4 (+6.6)**	**86.6 (+15.3)**	**71.3 (+13.8)**	96.8 (-1.8)	91.4 (-1.9)
Only Non-bird Concepts	60.8 (-22.0)	28.5 (-42.8)	28.8 (-28.7)	95.0 (-3.6)	85.8 (-7.5)

References: Concept Bottlenecks (Koh et al., ICML 2020), CLIP (Radford et al., ICML 2021), Sparse Autoencoders (Bricken et al., Transformer Circuits Thread 2023), LF-CBM (Oikarinen et al., ICLR 2023), LaBo (Yang et al., CVPR 2023), CDM (Panousis et al., ICCVW 2023), DCLIP (Menon et al., ICLR 2023), Waterbirds (Petryk et al., CVPR 2022).

Acknowledgements: Funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action" - project number 471607914.