CENG 7880 - Trustworthy and Responsible AI - Fall 2025

Concept Bottleneck Models

Dr. Emre Akbaş

Department of Computer Engineering

Today

- Concept bottleneck models
  - The original CBM paper (ICML 2020)
  - Label-free CBM (ICLR 2023).
  - LaBo: Language in a bottle (CVPR 2023)
- Concept/information leakage
- Top-down vs bottom-up CBMs
- DiscoverThenName (ECCV 2024)

Motivation

Most visual recognition models operate as **black-boxes**.

Motivation

Most visual recognition models operate as **black-boxes**.

Internal reasoning is **opaque**; decisions **cannot be inspected or interrogated**.

Yet **transparency** is essential for fairness, accountability, interpretability and intervenability.

Motivation

Most visual recognition models operate as **black-boxes**.

Internal reasoning is **opaque**; decisions **cannot be inspected or interrogated**.

Yet **transparency** is essential for fairness, accountability, interpretability and intervenability.

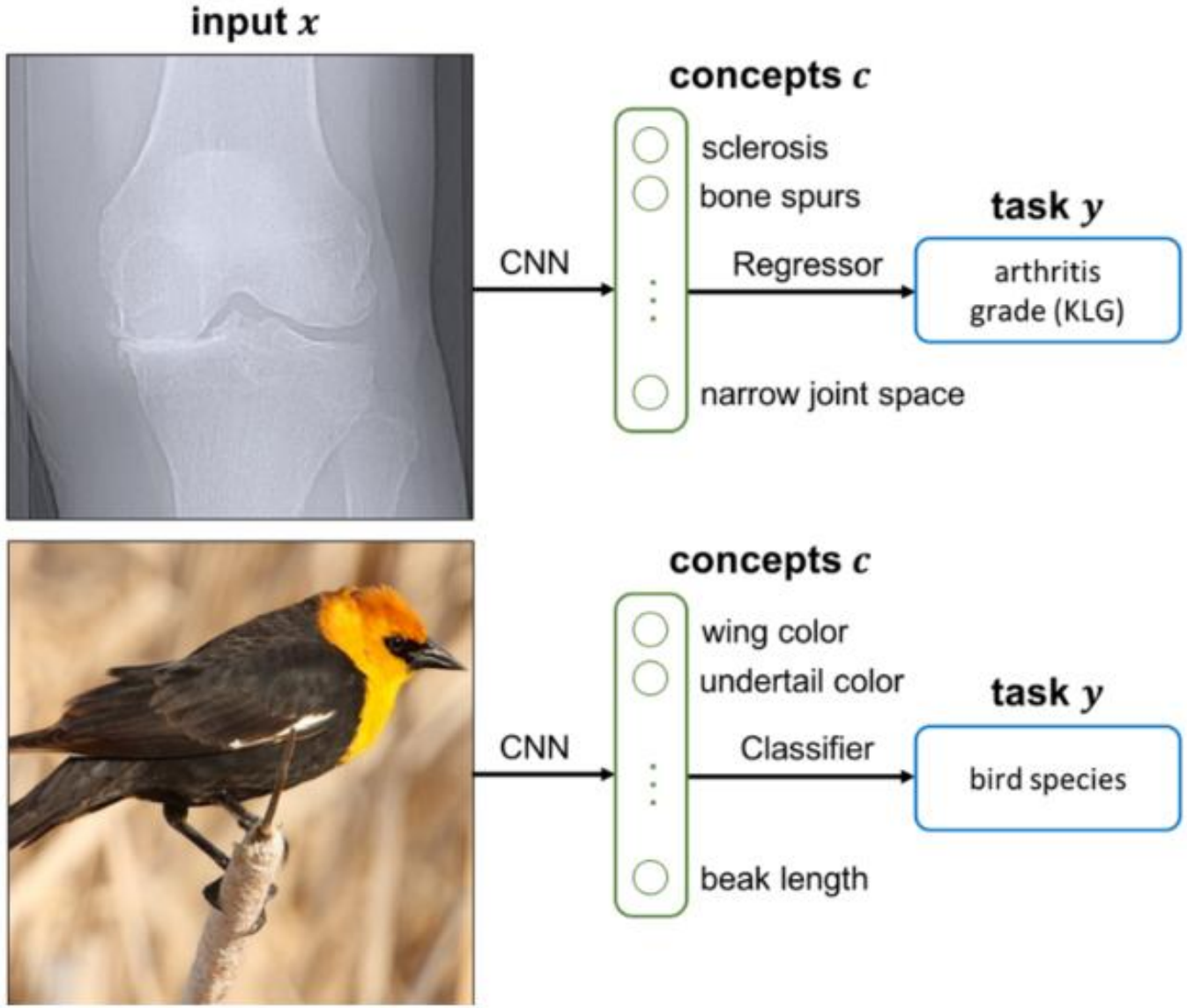
Concept bottleneck models (CBMs) offer a structured solution:

A CBM bases its **decision** on **human-understandable concepts**, enabling:

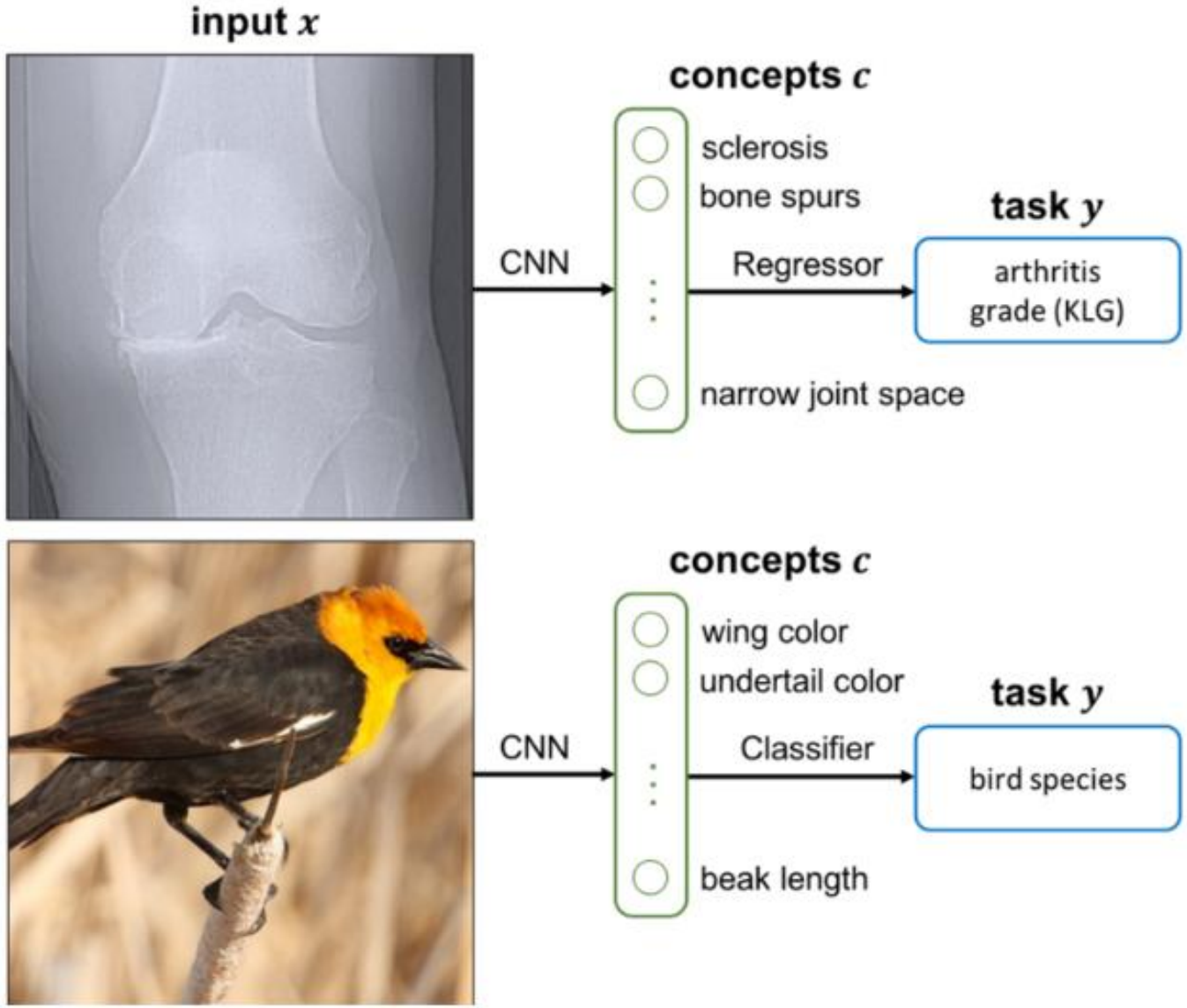
- explanation,
- diagnosis and
- control.

Concept Bottleneck Model (CBM)

Koh et al. Concept bottleneck models. In ICML 2020.



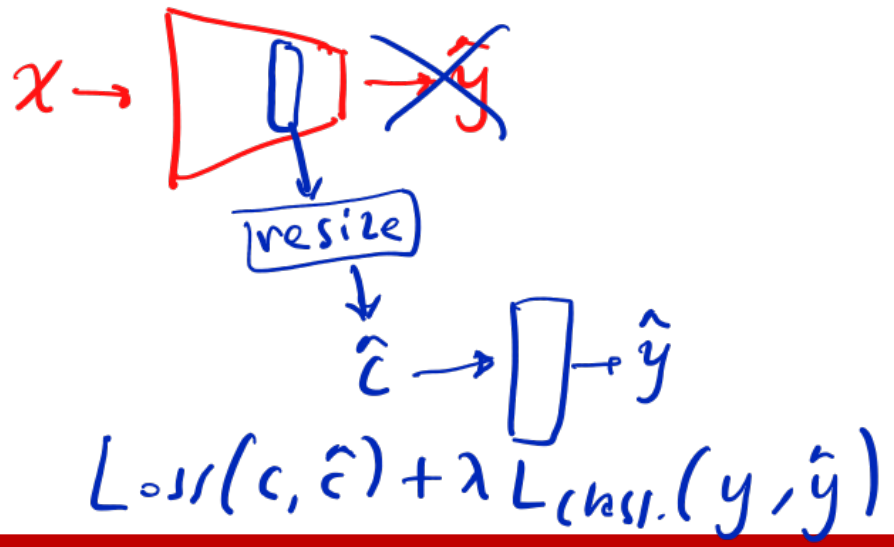
Concept Bottleneck Model (CBM)



Koh et al. Concept bottleneck models. In ICML 2020.

Attribute based classifiers since 2009
(see related work in Koh et al.)

Modern CBMs' promise is to convert any vision encoder to an explainable model.



Concept Bottleneck Model (CBM)

$$f(x) = g(h(x))$$

x : image

$h : x \rightarrow \hat{c}$ concept activations $\in \mathbb{R}^K$ K : # of concepts

$g : \hat{c} \rightarrow \hat{y}$ label

$$\mathcal{L}_{concept}(h(x), c) + \lambda \mathcal{L}_{task}(g(h(x)), y)$$

Concept Bottleneck Model (CBM)

Koh et al. Concept bottleneck models. In ICML 2020.

They assume each training image is annotated with concepts, i.e. not just *(x, y)* but *(x,c,y)*.

They used two datasets: X-ray grading (OAI) and bird classification (CUB).

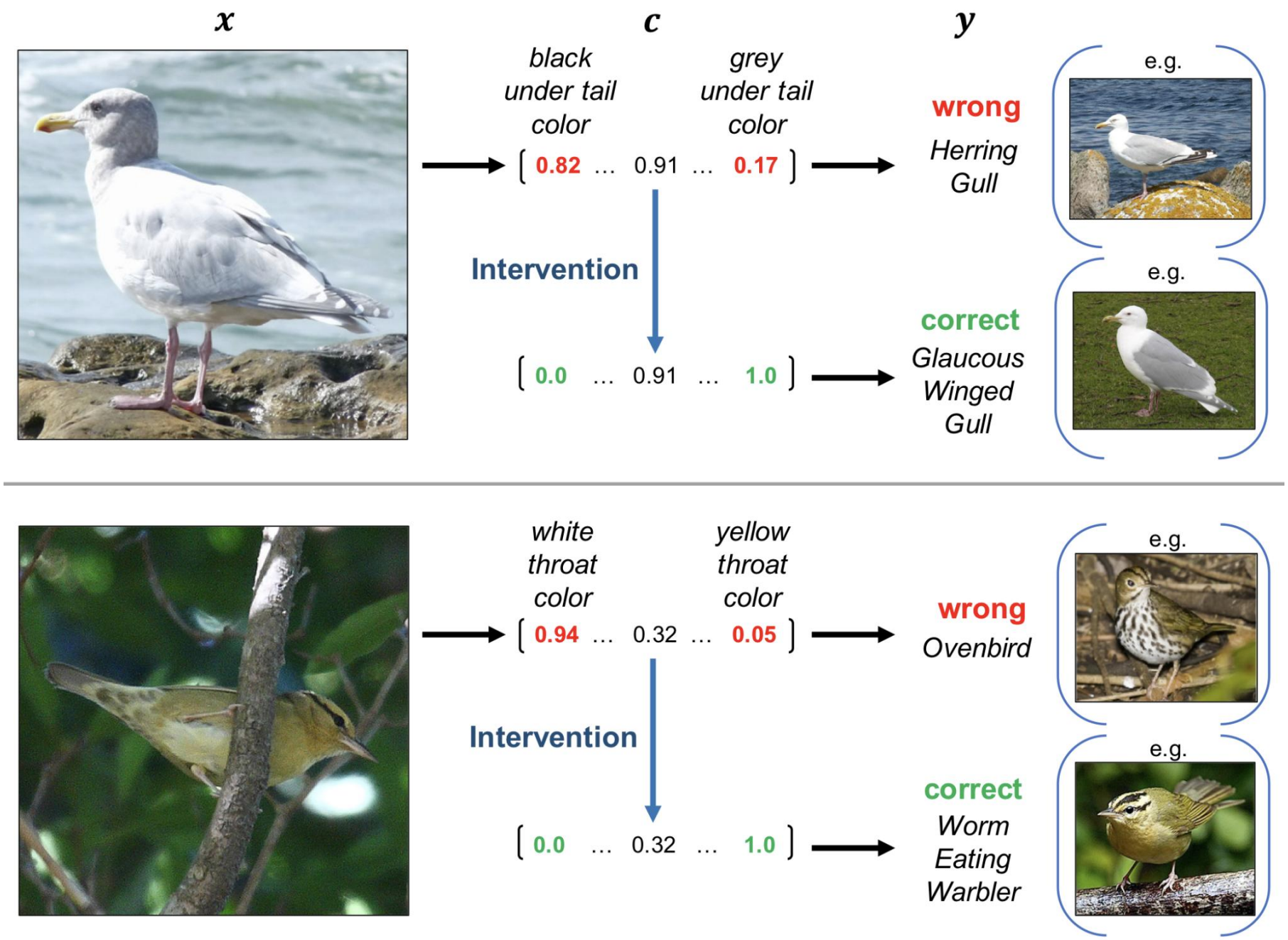
Model	y RMSE (OAI)	y Error (CUB)
Independent	0.435± 0.024	0.240±0.012
Sequential	0.418± 0.004	0.243±0.006
Joint	0.418± 0.004	0.199±0.006
Standard	0.441± 0.006	0.175±0.008
NO BOTTLENECK	0.443 ± 0.008	0.173±0.003
Multitask	0.425± 0.010	0.162±0.002

CBM

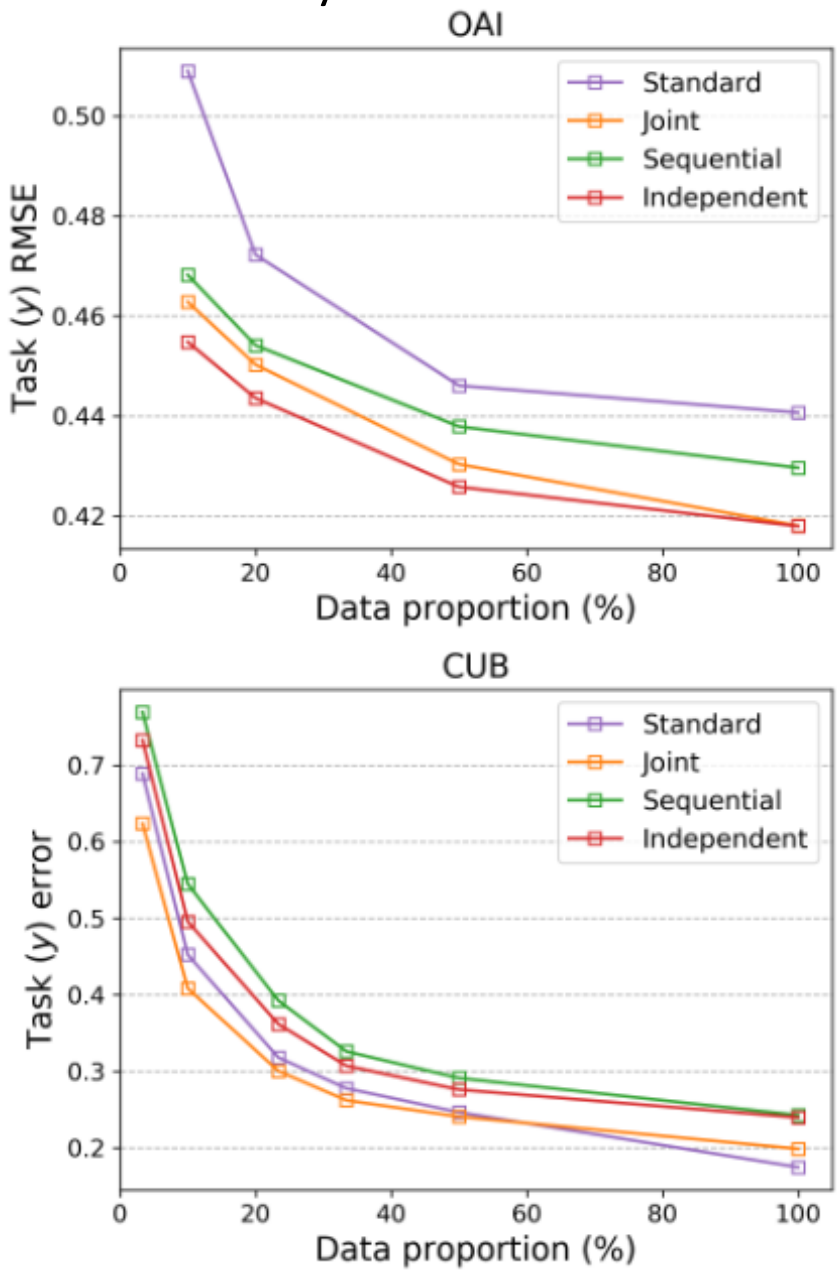
blackbox

Test-time intervention

Test-time intervention



Data efficiency



Robustness to background shift

Train:
Black-billed Cuckoo on
Forest Path background

Test:
Black-billed Cuckoo on
Coffee Shop background

Figure 5. In the TravelingBirds dataset, we change the image backgrounds associated with each class from train to test time (illustrated above for a single class).

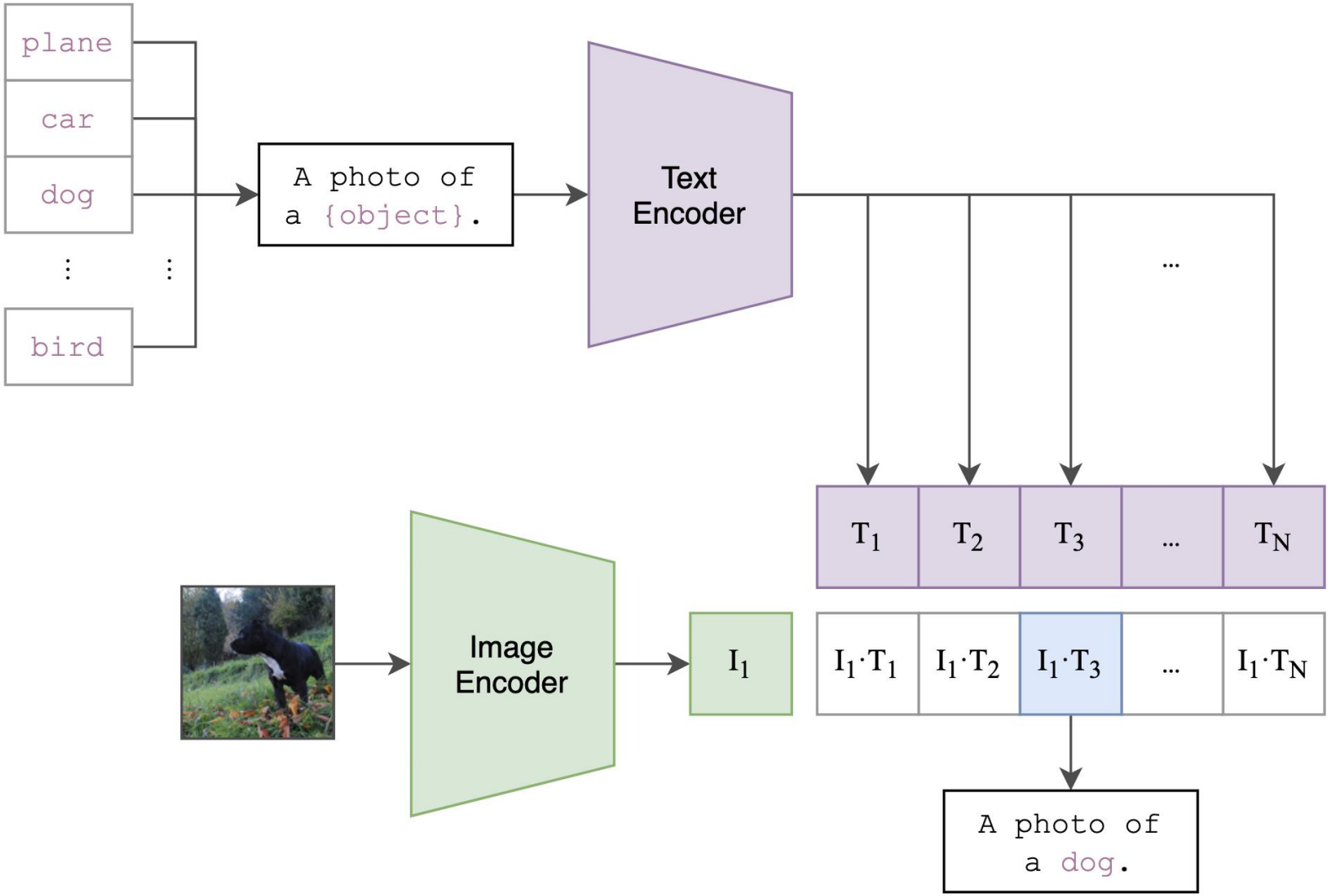
MODEL	y ERROR	c ERROR
STANDARD	0.627±0.013	-
JOINT	0.482±0.018	0.069±0.002
SEQUENTIAL	0.496±0.009	0.072±0.002
INDEPENDENT	0.482±0.008	0.072±0.002

Major weakness: assuming each example in the dataset is annotated with concepts.

Major weakness: assuming each example in the dataset is annotated with concepts.

Subsequent CBMs address this weakness with the help of VLMs (vision-language models) such as CLIP or LMMs (large multimodal models) such as Llava.

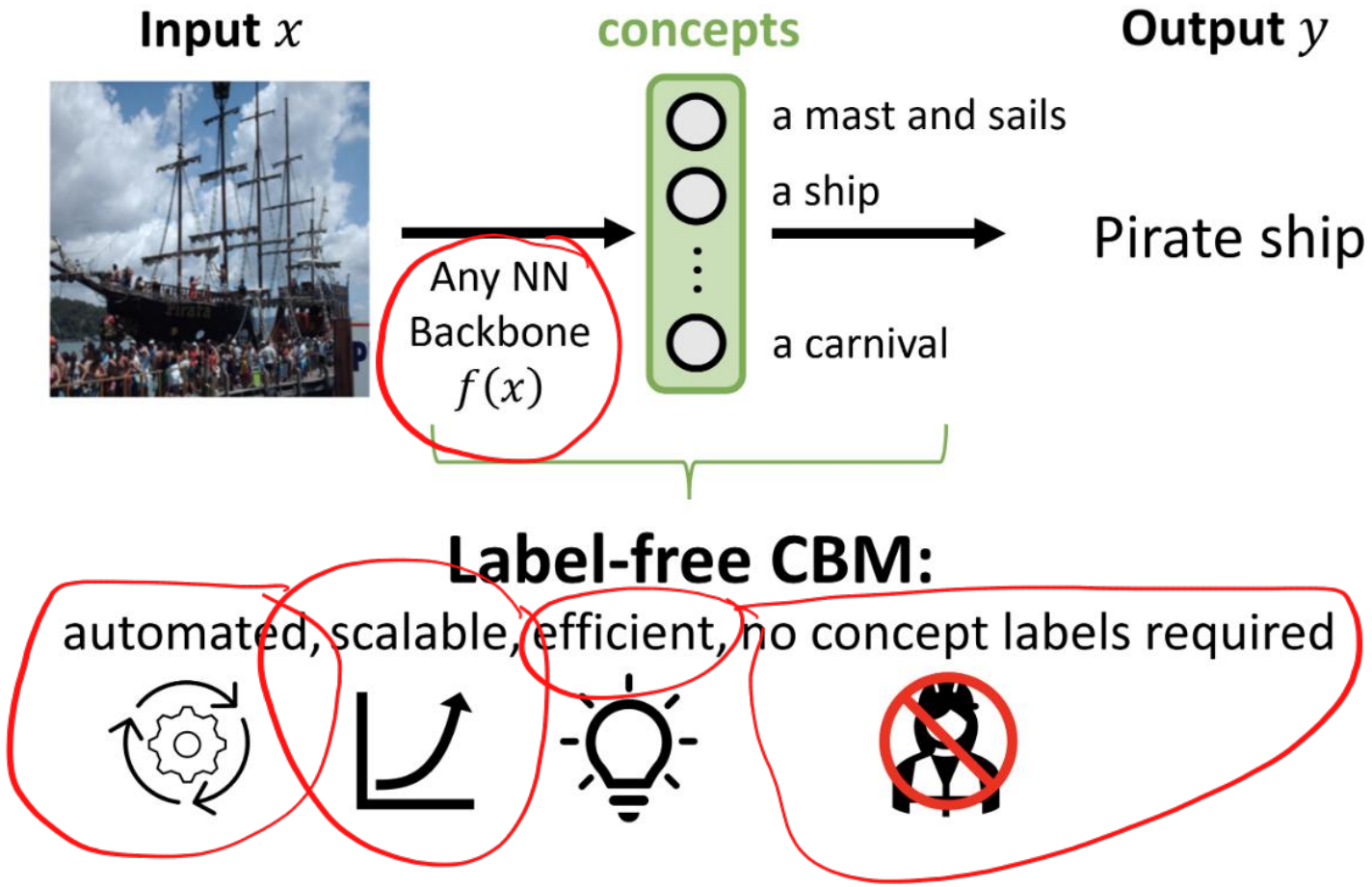
CLIP



Radford et al. Learning transferable visual models from natural language supervision. In ICML 2021.

Label-free CBM (LFCBM)

Oikarinen et al. Label-free concept bottleneck models. In ICLR 2023.



Label-free CBM (LFCBM)

Label-free CBM

Step 2: Compute embedding f(x) & concept matrix P

Input imgs $x_1, \cdots, x_n$

Backbone $f(x)$

Embeddings $f(x_1)$ … $f(x_N)$

Step 3: Compute CBL coefficient W_c by max similarity between f_c and P

transform $f_c(x) = W_c f(x)$

Concept bottleneck layer (CBL)

concept 1
concept 2
⋮
concept M

$f_c(x_1)$ … $f_c(x_N)$

Step 4: Train a sparse FC layer W_F on $f_c(x)$

FC layer $z = W_F f_c(x)$

Predicted class $\hat{y}$

$\hat{y} = \text{argmax}_{i \in \{1,2,\ldots,d_z\}} z_i$

Bridge, car, …, cloud

$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_N$

Step 1: Generate and filter concept set

GPT 3 → Filtering

Target concepts

Antartica
a cashier
a seat
butter
dirt
markings

Image Enc E_I

Text Enc E_T

Inner product

Concept Matrix

$E_I(x_1) \cdot E_T(t_1)$	…	$E_I(x_1) \cdot E_T(t_M)$
$E_I(x_2) \cdot E_T(t_1)$	…	$E_I(x_2) \cdot E_T(t_M)$
…	…	…
$E_I(x_N) \cdot E_T(t_1)$	…	$E_I(x_N) \cdot E_T(t_M)$

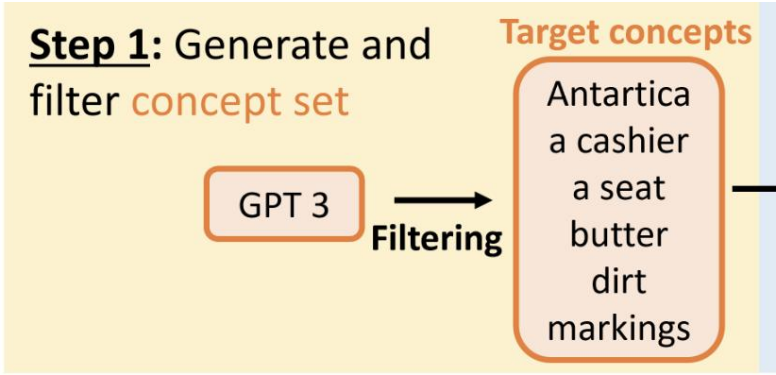
P

(Notations)

$x_i \in \mathrm{R}^{d_0}$
$f(x) \in \mathrm{R}^{d}$
$W_c \in \mathrm{R}^{M \times d}$
$f_c(x) \in \mathrm{R}^{M}$
$z \in \mathrm{R}^{d_z}$
$W_F \in \mathrm{R}^{d_z \times M}$

Oikarinen et al. Label-free concept bottleneck models. In ICLR 2023.

Label-free CBM (LFCBM)



STEP 1: CONCEPT SET CREATION AND FILTERING

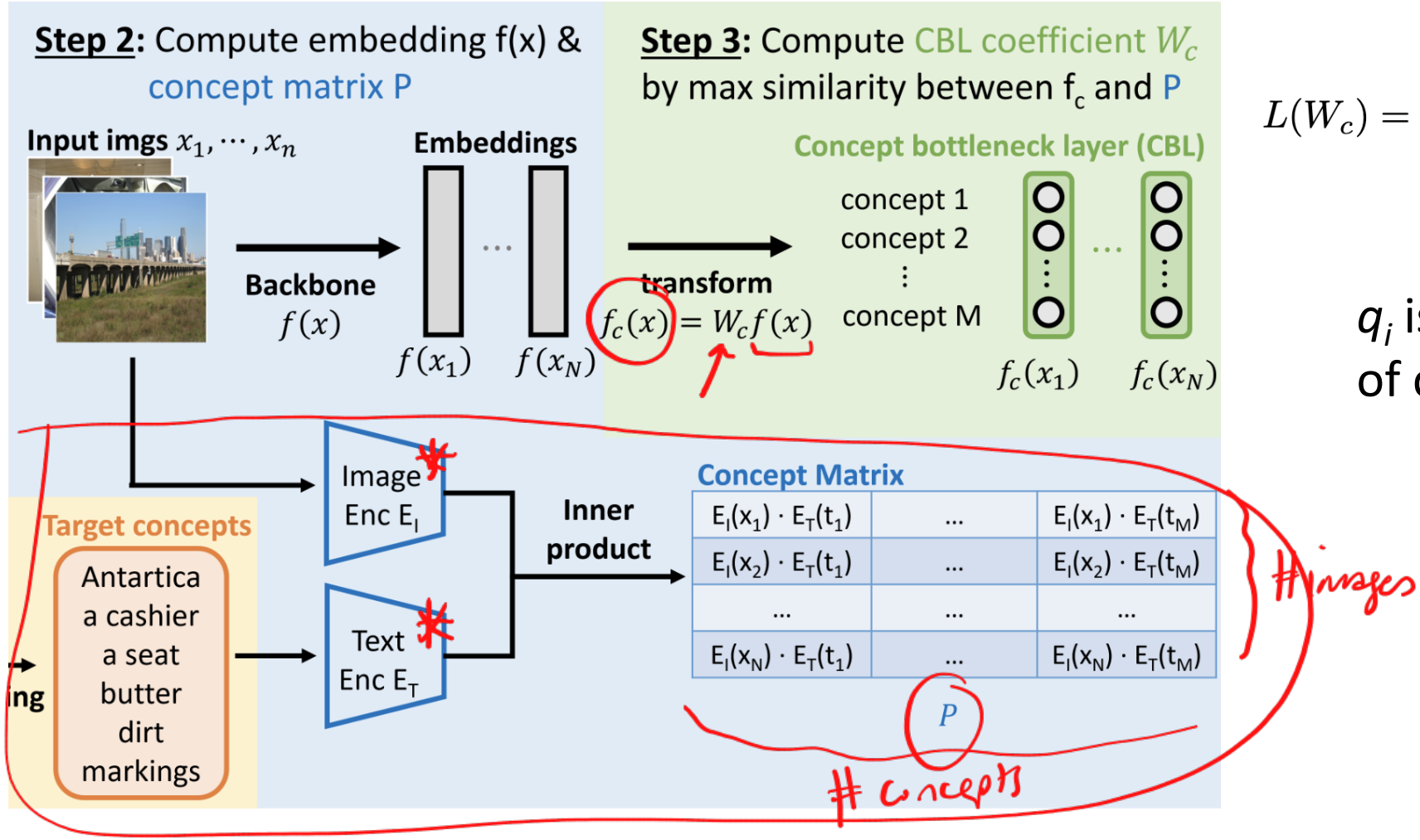
They ask GPT-3 the following:

- 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}:

Oikarinen et al. Label-free concept bottleneck models. In ICLR 2023.

Label-free CBM (LFCBM)

STEP 2 AND 3: LEARNING THE CONCEPT BOTTLENECK LAYER (CBL)



$$L(W_c) = \sum_{i=1}^{M} -\text{sim}(t_i, q_i) := \sum_{i=1}^{M} -\frac{\bar{q_i}^3 \cdot \bar{P_{:,i}}^3}{||\bar{q_i}^3||_2 ||\bar{P_{:,i}}^3||_2}.$$

q_i is the vector of similarities of concept *i* to all images

Oikarinen et al. Label-free concept bottleneck models. In ICLR 2023.

Label-free CBM (LFCBM)

STEP 4: LEARNING THE SPARSE FINAL LAYER

Step 4: Train a sparse FC layer W_F on $f_c(x)$

Predicted class $\hat{y}$

$\hat{y} = \text{argmax}_{i \in \{1,2,\ldots,d_z\}} z_i$

Bridge, car, ..., cloud

FC layer $z = W_F f_c(x)$

$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_N$

(Notations)

$x_i \in \mathrm{R}^{d_0}$

$f(x) \in \mathrm{R}^{d}$

$W_c \in \mathrm{R}^{M \times d}$

$f_c(x) \in \mathrm{R}^{M}$

$z \in \mathrm{R}^{d_z}$

$W_F \in \mathrm{R}^{d_z \times M}$

$$\min_{W_F, b_F} \sum_{i=1}^{N} L_{ce}(W_F f_c(x_i) + b_F, y_i) + \lambda R_\alpha(W_F)$$

Oikarinen et al. Label-free concept bottleneck models. In ICLR 2023.

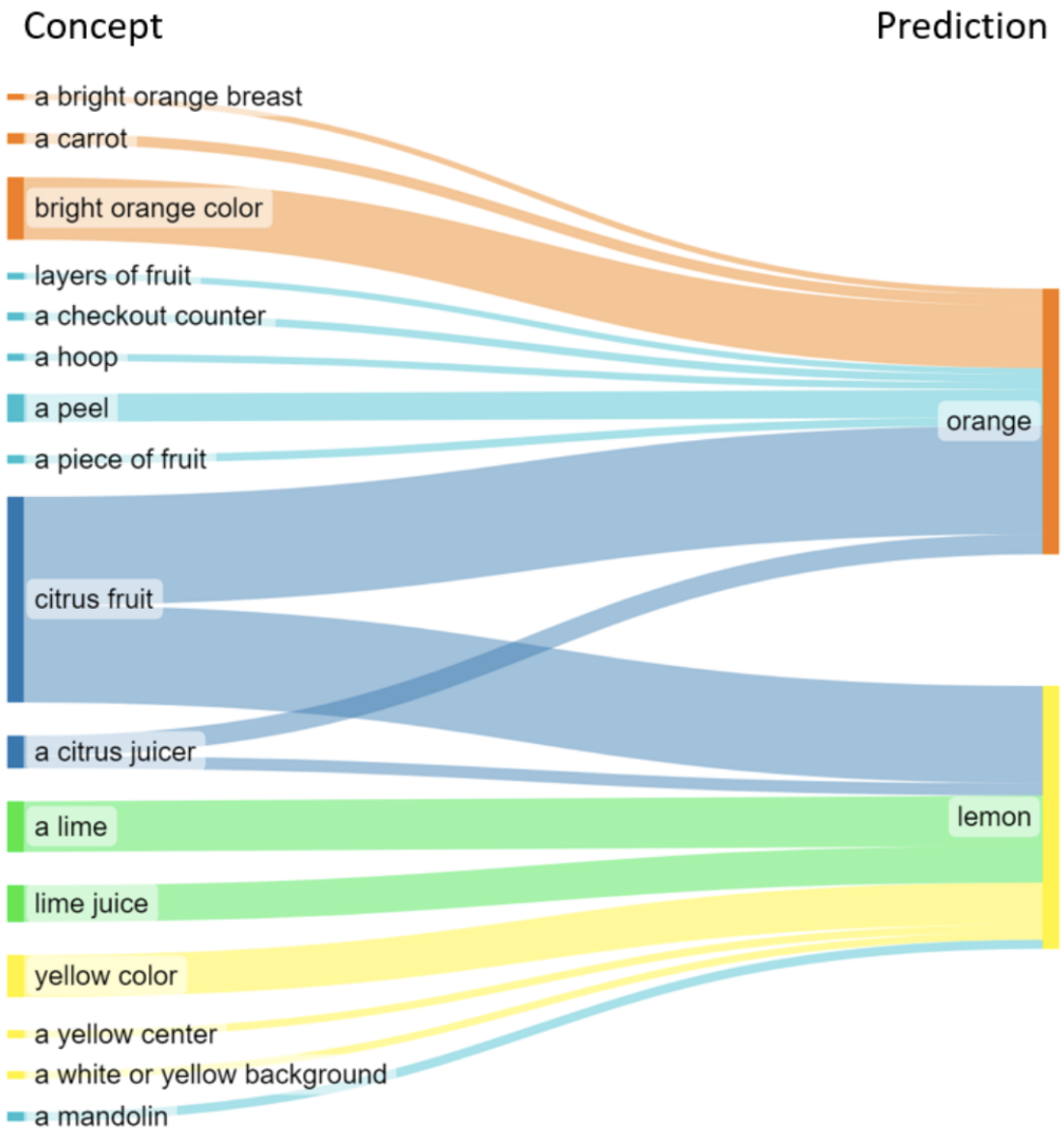
Results

Model	Sparse final layer	Dataset: CIFAR10	CIFAR100	CUB200	Places365	ImageNet
Standard	No	88.80%*	70.10%*	76.70%	48.56%	76.13%
Standard (sparse)	Yes	82.96%	58.34%	**75.96%**	38.46%	**74.35%**
P-CBM	Yes	70.50%*	43.20%*	59.60%*	N/A	N/A
P-CBM (CLIP)	Yes	84.50%*	56.00%*	N/A	N/A	N/A
Label-free CBM **(Ours)**	Yes	**86.40%** ± 0.06%	**65.13%** ± 0.12%	74.31% ± 0.29%	**43.68%** ± 0.10%	71.95% ± 0.05%

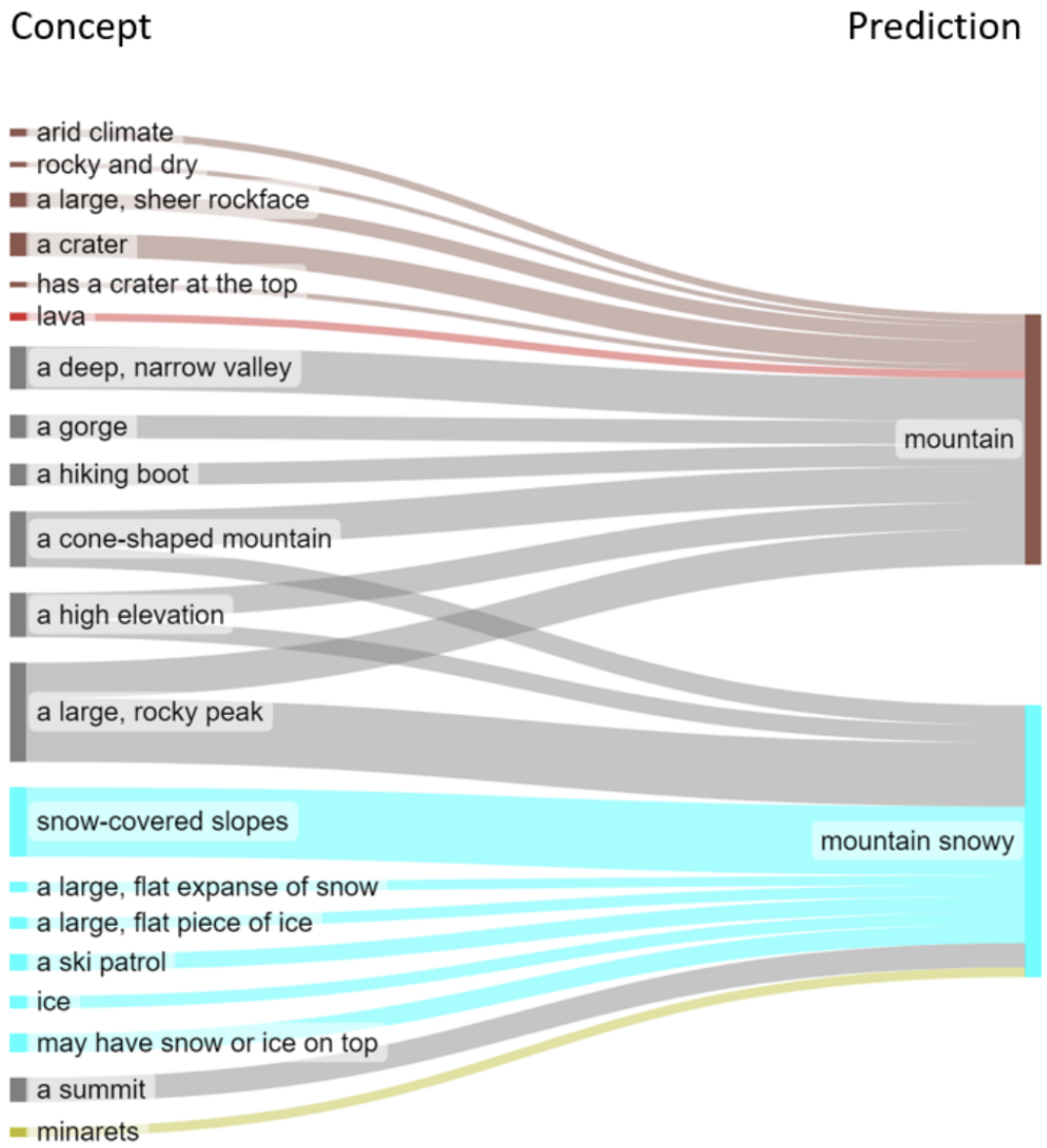
Table 2: Accuracy comparison, best performing sparse model bolded. We can see our method outperforms Posthoc-CBM and performs similarly to a sparse standard model. The results for our method are mean and standard deviation over three training runs. *Indicates reported accuracy.

Oikarinen et al. Label-free concept bottleneck models. In ICLR 2023.

ImageNet CBM
Orange vs Lemon



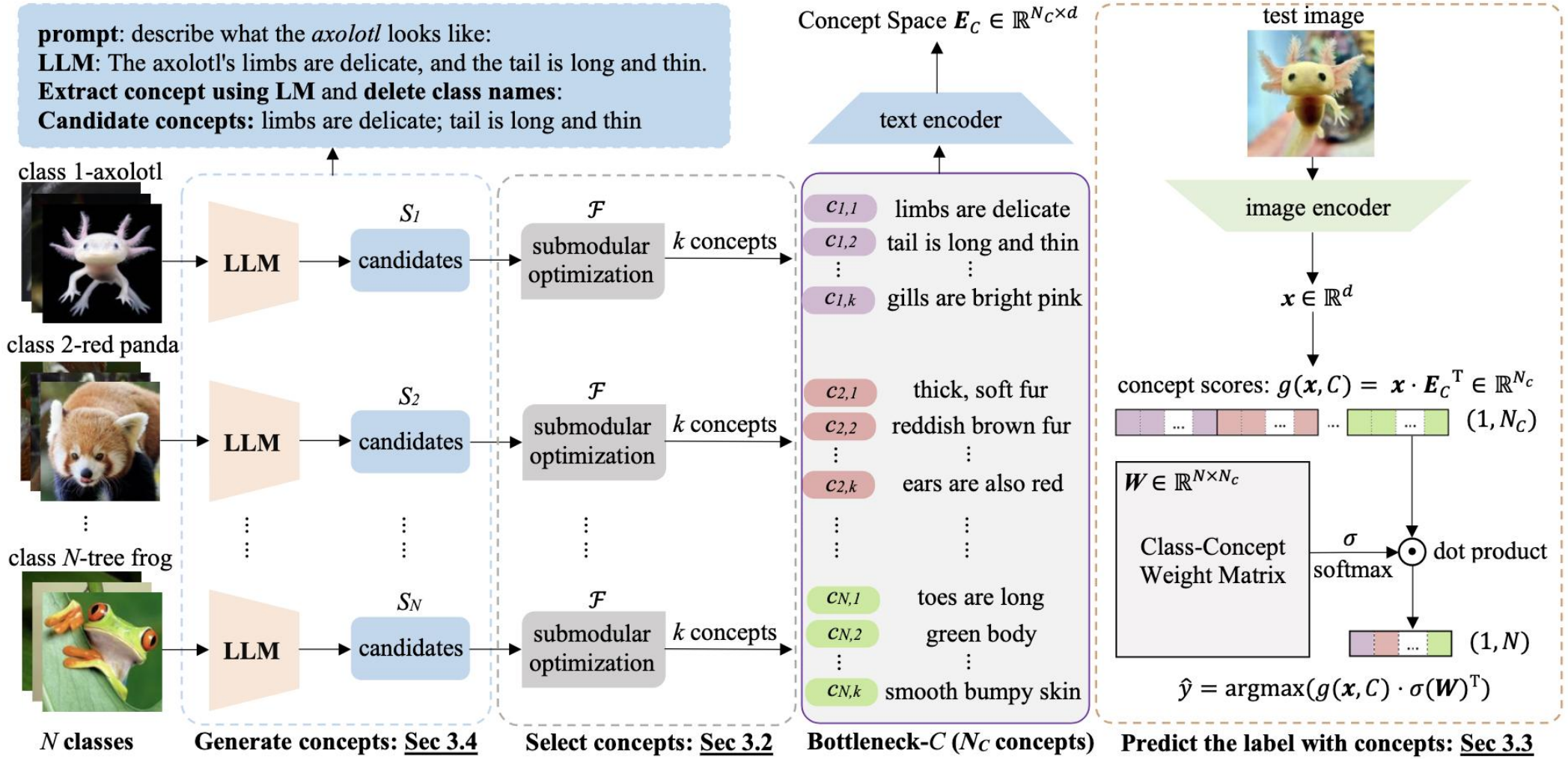
Places365 CBM
Mountain vs Mountain Snowy



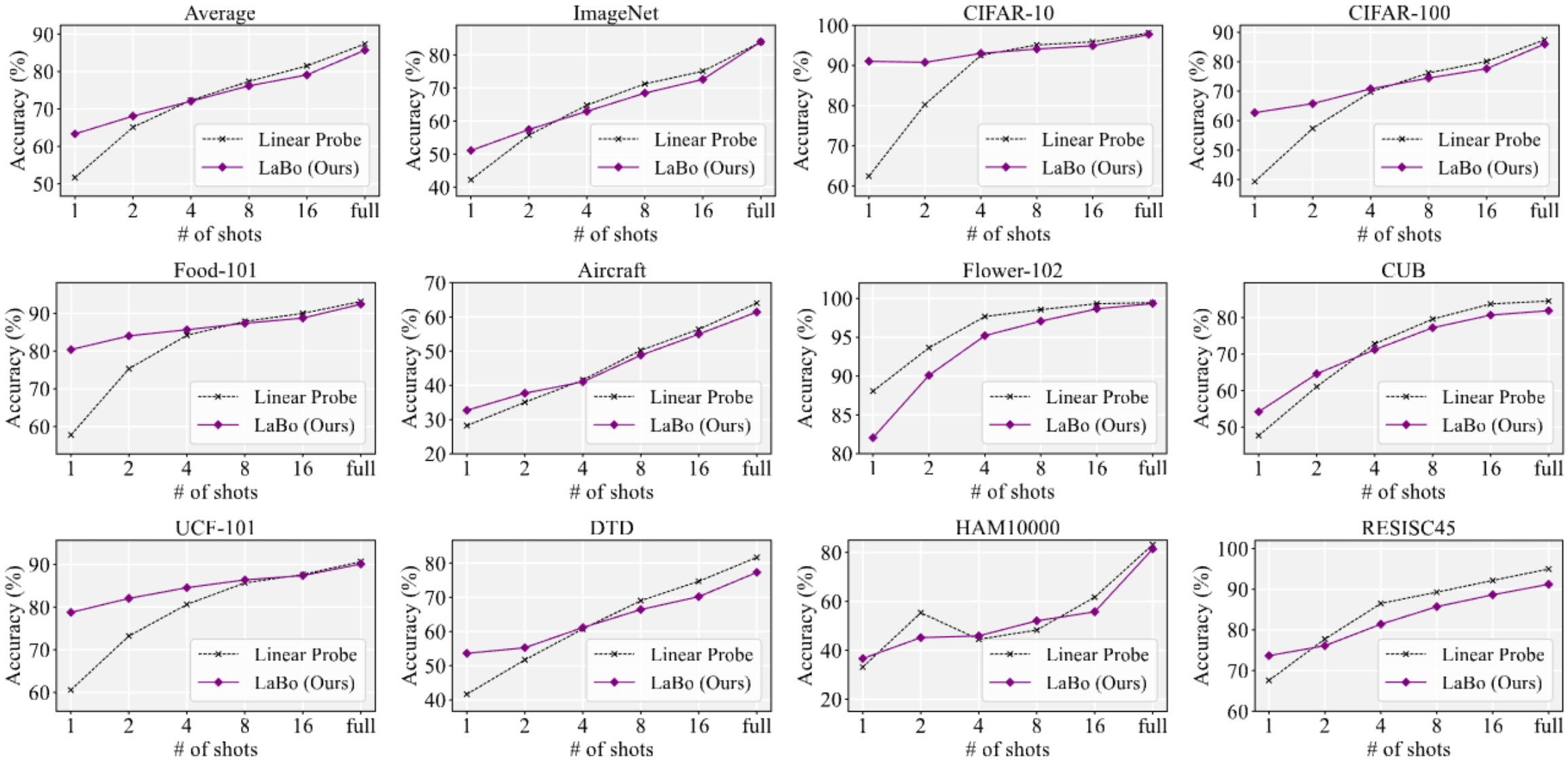
Many other CBMs

LaBo

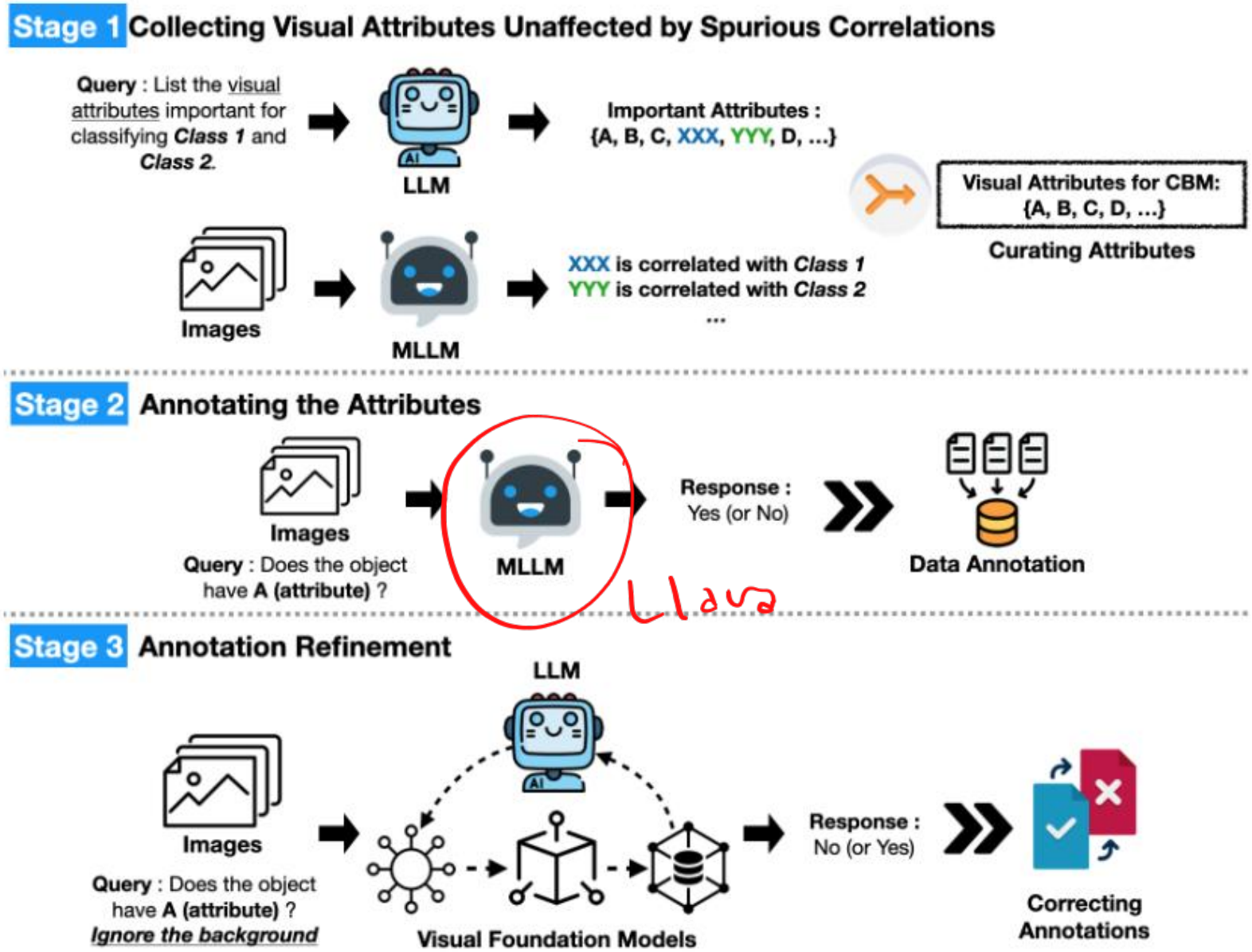
A very active area of research. A few notable examples:



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

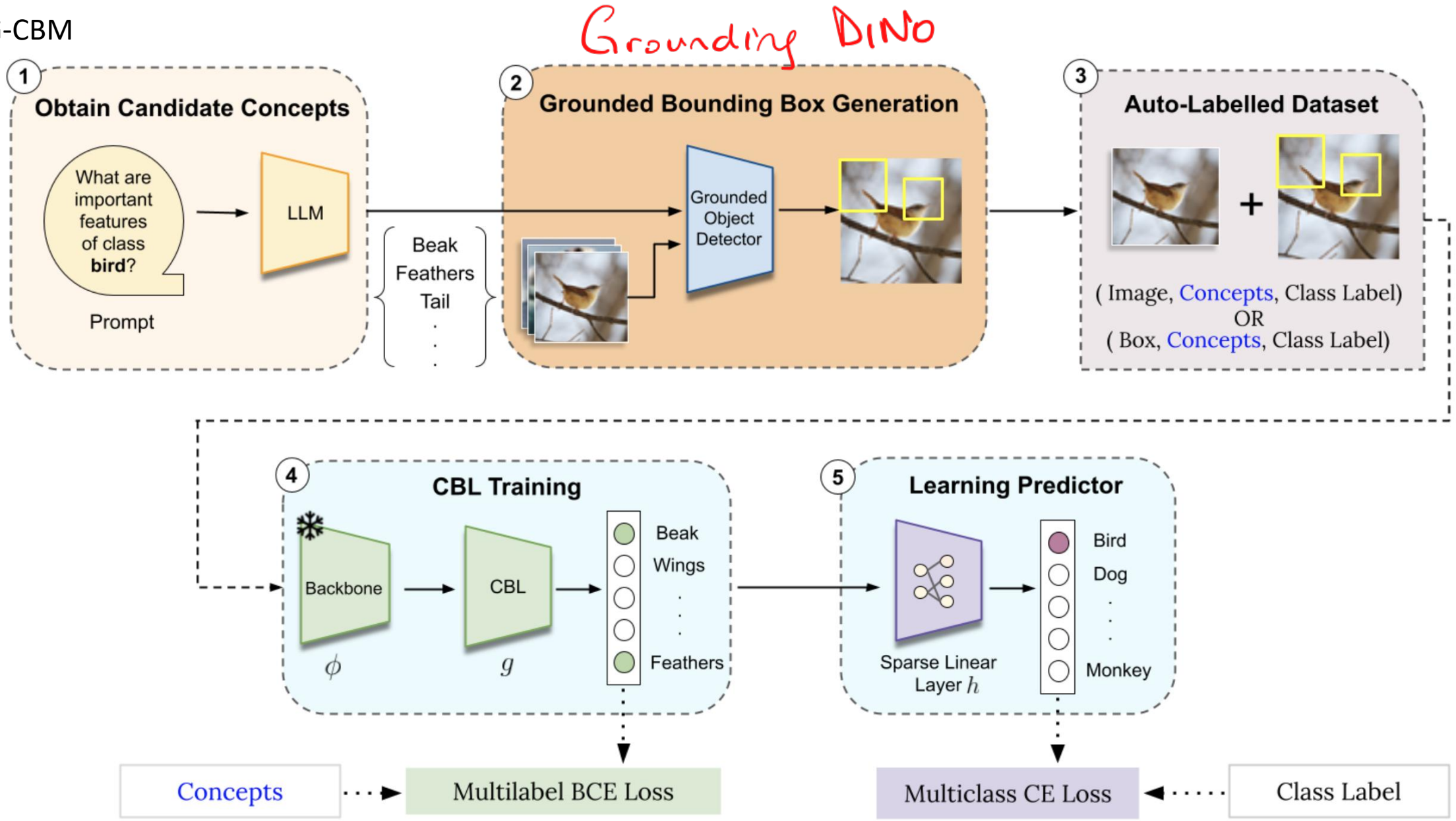


Yang et al. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In CVPR 2023.



Kim et al. Constructing concept-based models to mitigate spurious correlations with minimal human effort. In ECCV 2024.

VLG-CBM



Srivastava et al. VLG-CBM: Training concept bottleneck models with vision-language guidance. In NeurIPS 2024.

Concept/information leakage problem

CBMs show **unexpectedly high accuracy** when using **irrelevant concepts** for the task.

ImageNet Roman Law concept

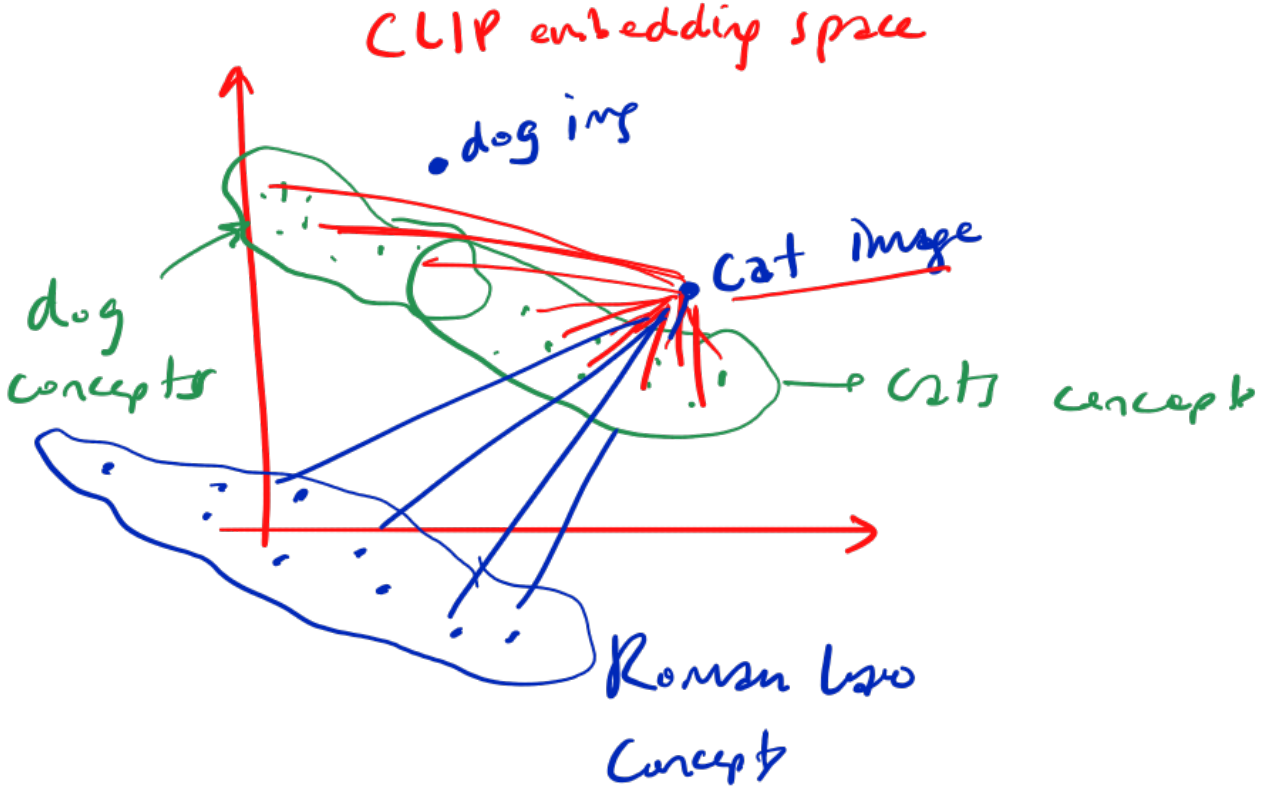
Mahinpei et al. *Promises and pitfalls of black-box concept learning models*. ICML 2021 Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI.

Margeloiu et al. *Do concept bottleneck models learn as intended?* In ICLR 2021 Workshop on Responsible AI.

Makonnen et al. *Measuring leakage in concept-based methods: An information theoretic approach*. In ICLR 2025 Workshop XAI4Science.

+ many other papers working on addressing this problem.

Concept/information leakage problem



Attempts

(1) Restrict by similarity threshold

(2) use top-k (k small) concepts per image

Top-down vs bottom-up CBMs

Top-down CBM: externally specified concepts by humans or models.

Bottom-up CBM: concepts are discovered during training from image content.

Top-down vs bottom-up CBMs

Top-down CBM: externally specified concepts by humans or models.

- Concept set is designed before training CBM.
- Flexible approach.

Bottom-up CBM: concepts are discovered during training from image content.

Top-down vs bottom-up CBMs

Top-down CBM: externally specified concepts by humans or models.

- Concept set is designed before training CBM.
- Flexible approach.

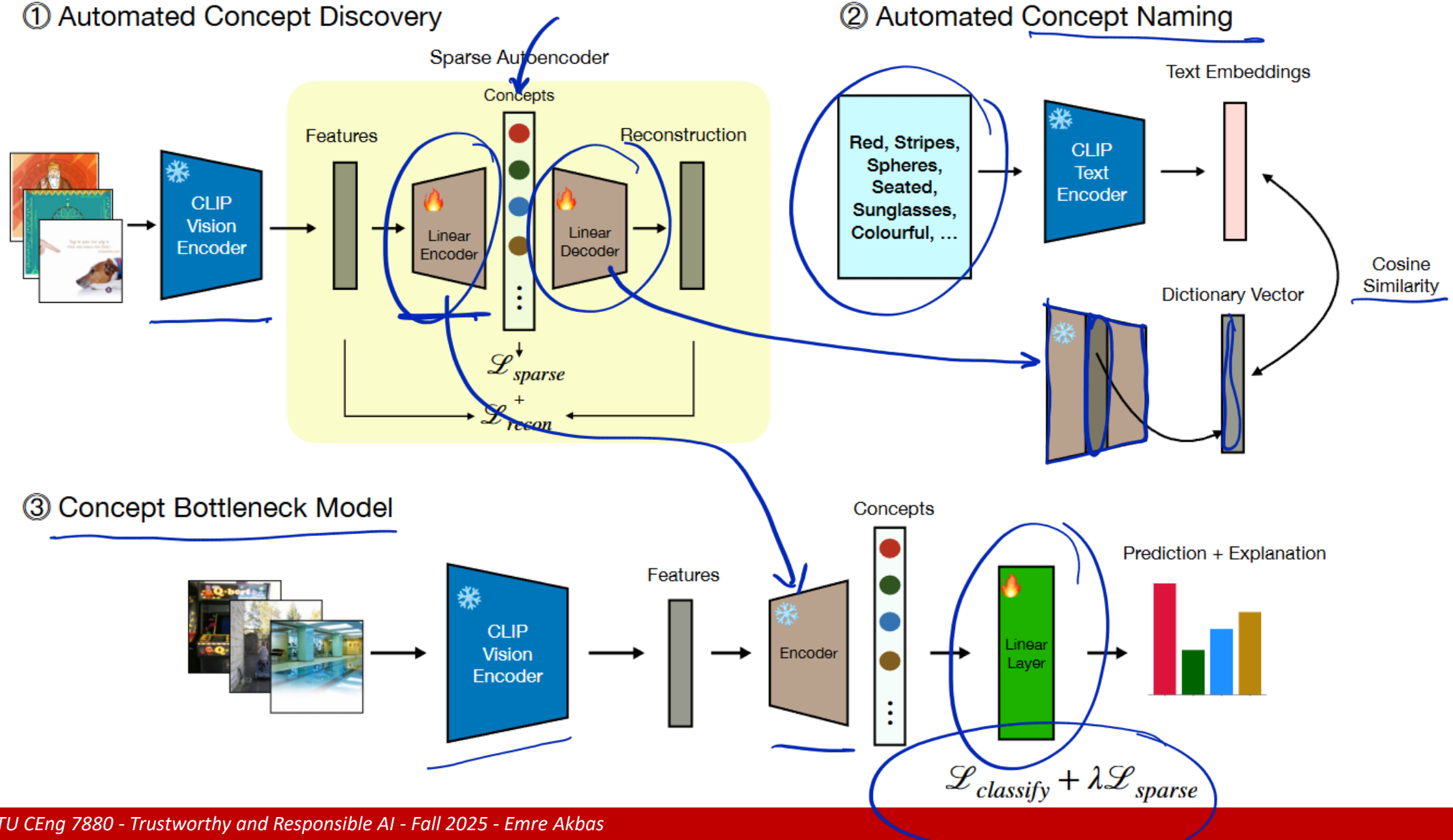
Bottom-up CBM: concepts are discovered during training from image content.

- Concepts emerge automatically during training.
- A sparse autoencoder (or similar dictionary-learning method) is trained to reconstruct CLIP embeddings or backbone features.
- Units in the learned dictionary correspond to interpretable directions in latent space; they often align with semantic attributes.
- Concept naming is done post-hoc.

An example bottom-up CBM →

DiscoverThenName

Rao et al. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In ECCV 2024.



Summary

CBMs convert any vision encoder to an explainable model.

They provide

- explanation,
- diagnosis and
- control.