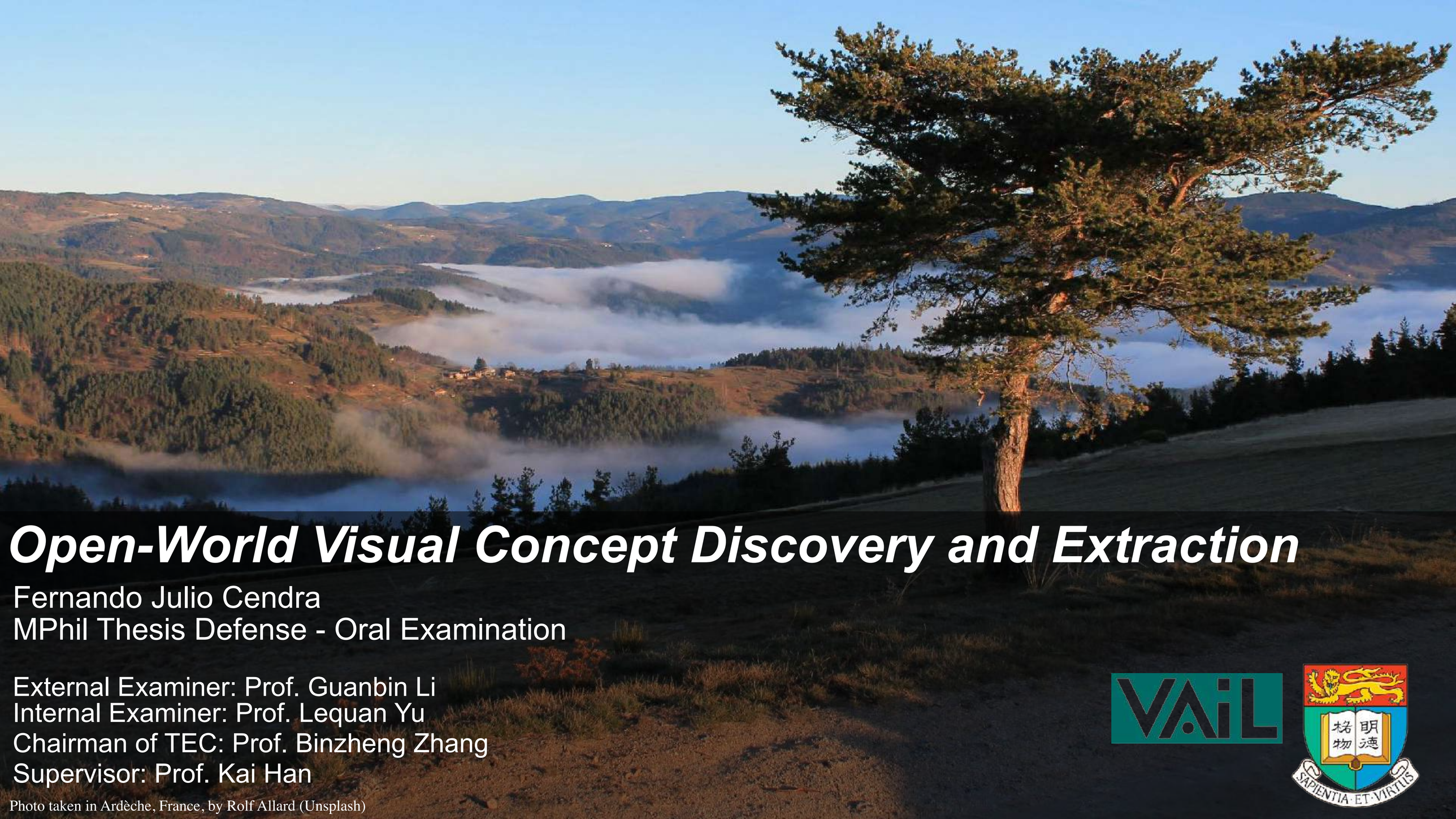


# *Open-World Visual Concept Discovery and Extraction*

Fernando Julio Cendra

MPhil Thesis Defense - Oral Examination

External Examiner: Prof. Guanbin Li
Internal Examiner: Prof. Lequan Yu
Chairman of TEC: Prof. Binzheng Zhang
Supervisor: Prof. Kai Han

Photo taken in Ardèche, France, by Rolf Allard (Unsplash)

# *Motivation:*

How baby observes and explores the world!

**Scene 1:** The baby is aware of his own cat and surroundings

Source: Jesse Drifter - Filmmaker, https://youtu.be/OEgGcfjG684 (My Son Met a Horse for the First Time… His Face Says It All 🐴🥹)

# ***Motivation:***

How baby observes and explores the world!

**Scene 2:** The baby is aware of this “thing” he never met before

Source: Jesse Drifter - Filmmaker, https://youtu.be/OEgGcfjG684 (My Son Met a Horse for the First Time… His Face Says It All 🐴🥹)

## Scene 1 (Known)

## Scene 2 (Novel)

## ***Motivation:***

Thus a robust perception systems should have…

- Awareness of new observations.
- Able to understand & distinguish new concepts.

Source: Jesse Drifter - Filmmaker, https://youtu.be/OEgGcfjG684 (My Son Met a Horse for the First Time… His Face Says It All 🐴🥹)

# *Enhancing perception systems with open-world capabilities*


- Awareness of new observations.
- Able to understand & distinguish new concepts.

# Outline of this seminar

(A) Perception task: Generalized Category Discovery (GCD)

➡ The ability to discover novel categories

(B) Perception task: Continual Category Discovery

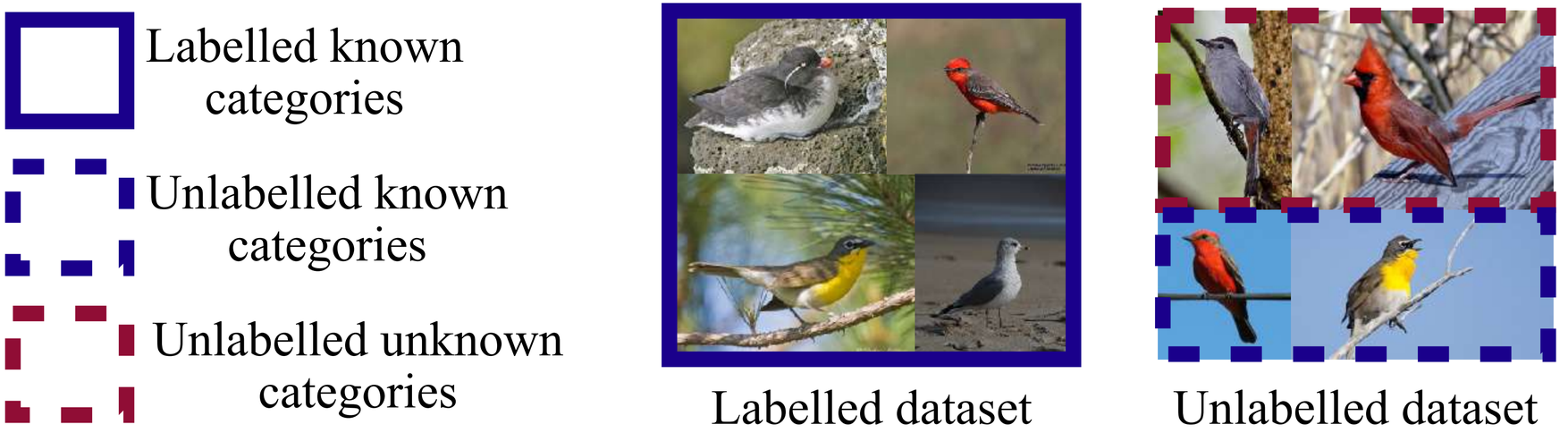
➡ The ability to continually discover novel categories

(C) Generation task: Intrinsic Concept Extraction

➡ The ability to learn & extract concepts

# Task A: Generalized Category Discovery

# Generalized Category Discovery: Intro




Given a labelled subset contain known classes, the task is to categorize the unlabelled images, which may belong to both known and novel classes.

# Generalized Category Discovery: Intro

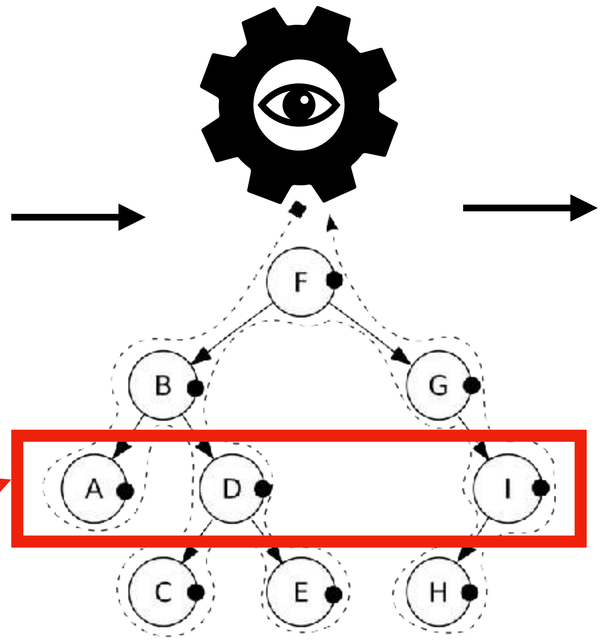
Why do we need these labelled dataset?

# Generalized Category Discovery: Intro

Why do we need these labelled dataset?

they serve as the “inductive biases” for the model to learn how to categorize the unlabelled data.

Labelled dataset




Cluster glanularity

Unlabelled dataset

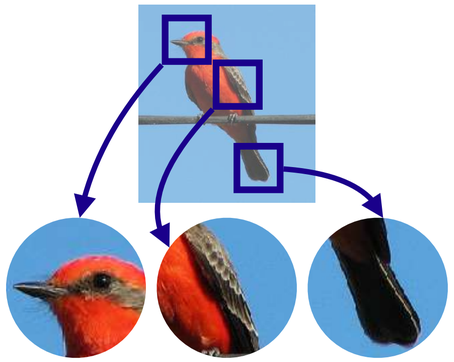
# Generalized Category Discovery: Intro

Is there other “feature” we can utilize from the data?

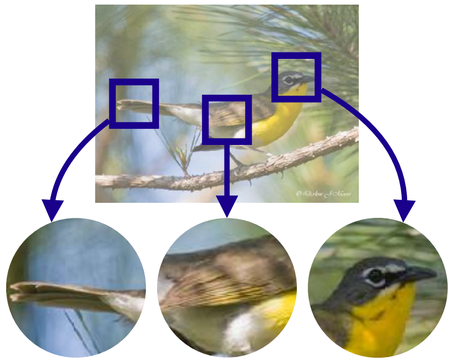
# Generalized Category Discovery: Intro

Is there other “feature” we can utilize from the data?

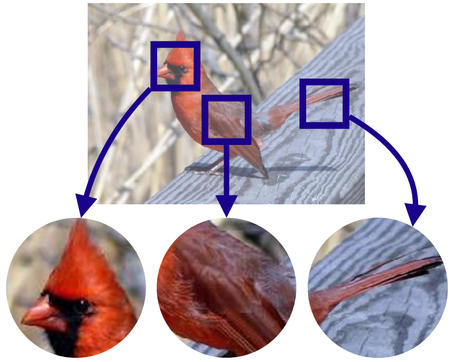
*Vermilion Flycatcher*



*Yellow-breasted Chat*



*Northern Cardinal*



Part-level observations

We hypothesize that part-level observations provide a strong signal for the baby to distinguish these known and novel categories

# Generalized Category Discovery: Intro

In fact, some previous GCD works argue that part-level observation is important:

- [GCD, Vaze et al., CVPR'22]:

> *"...We find this feature to be useful for this task, because which object parts are important for classification is likely to transfer well from the labelled to the unlabelled categories..."*

- [SPTNet, Wang et al., ICLR'24]:

> *"...We introduce spatial prompt tuning (SPT) as a method to focus on object parts and facilitate knowledge transfer between seen and unseen classes..."*

# Generalized Category Discovery: Intro

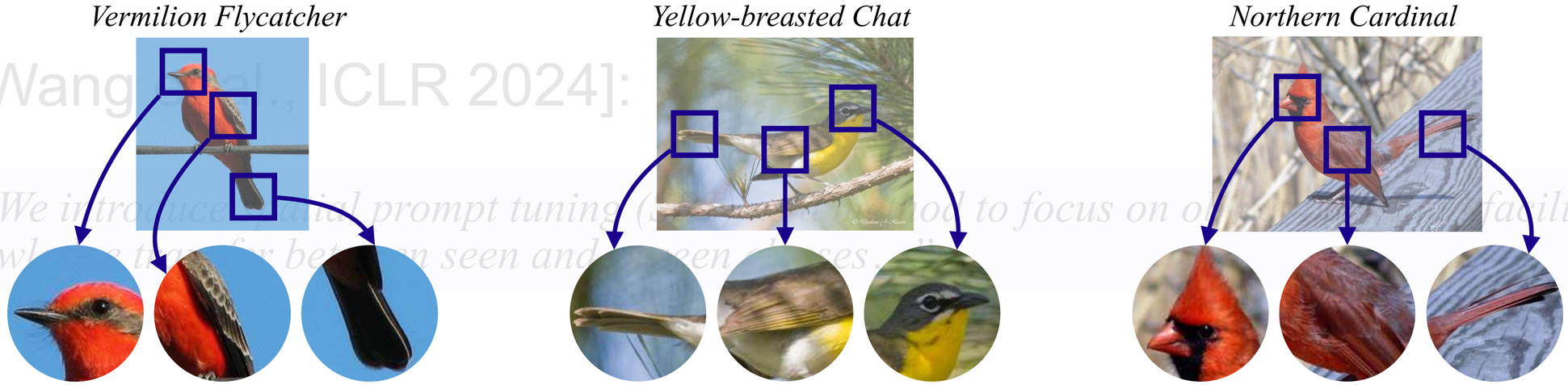
In fact, some previous GCD works argue that part-level observation is important:

- [Wang et al., ICLR 2024]:

*"...We introduce spatial prompt tuning (SPT) which encourages the model to focus on object parts and facilitate knowledge transfer between seen and unseen classes..."*

However, recent approaches predominantly rely on global representations derived from the [CLS] token of transformer models.

Moreover, these methods does not account for the inherent variability of objects parts, such as differing scales, orientations, or varying number of parts due to occlusions.



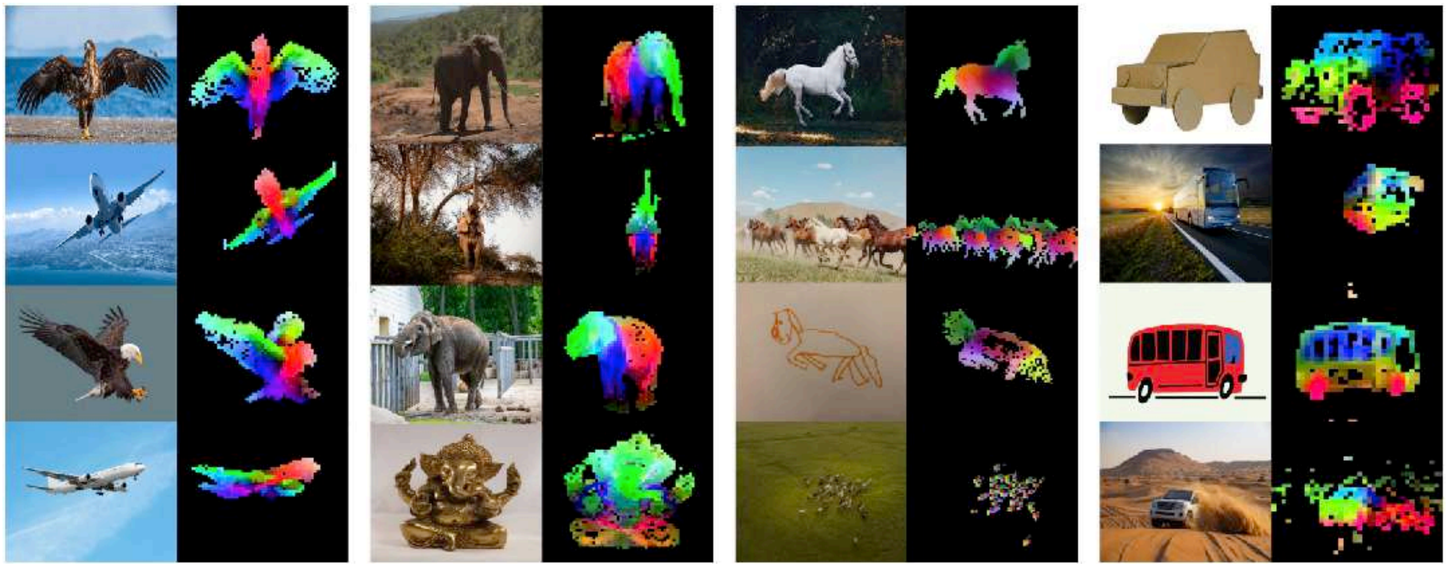
# Generalized Category Discovery: Intro

These issues necessistate an effective ***explicit control signal*** to fully harness the potential of part-level observations for GCD!

# Generalized Category Discovery: Intro

Vision foundation models demonstrate remarkable generalization across tasks.

- Additionally, these models excel at extracting high-dimensional patch token features that capture detailed and localized semantic information.
- Unlike the [CLS] token, patch tokens focus on specific parts. Thus, serving as an ideal **explicit control signal** that we need.

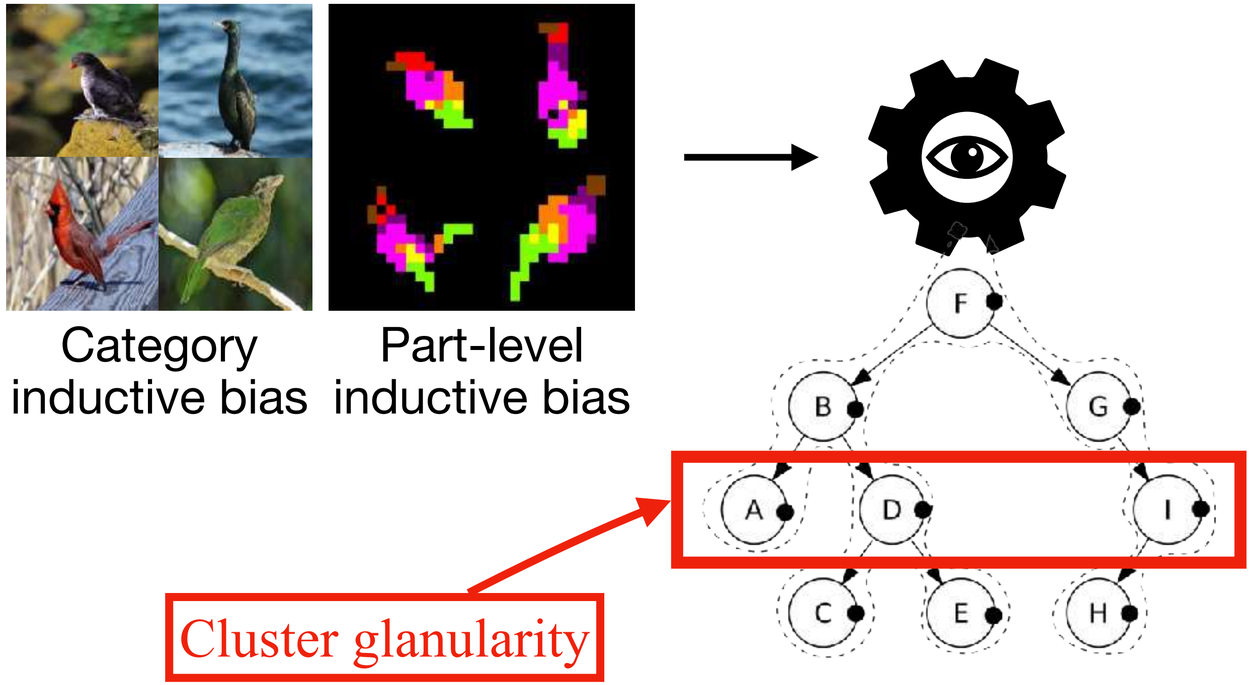


Visualization of the first PCA components.

Source: DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024

# Generalized Category Discovery: Method

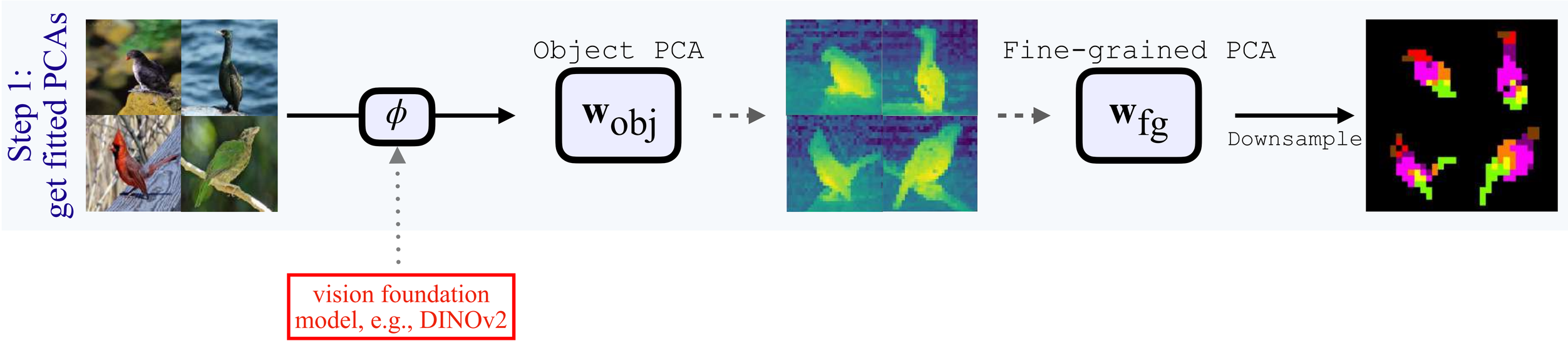
Let's use this idea to construct these control signal to guide our GCD model.



# Generalized Category Discovery: Method

First, given subset of the images, let's get the PCA projection for both:

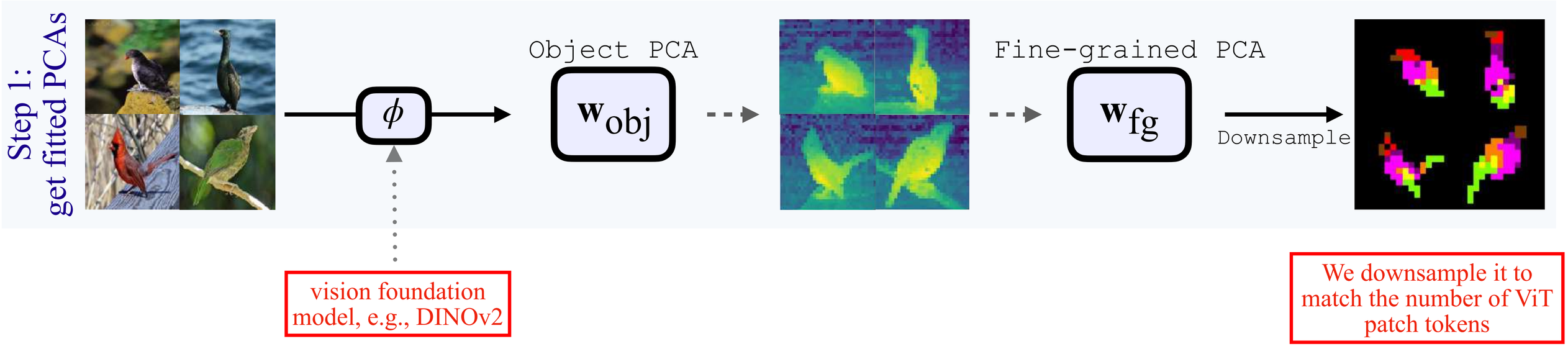
- Object PCA: projection to seperate foreground and background.
- Fine-grained PCA: projection to seperate foreground parts.



# Generalized Category Discovery: Method

First, given subset of the images, let's get the PCA projection for both:

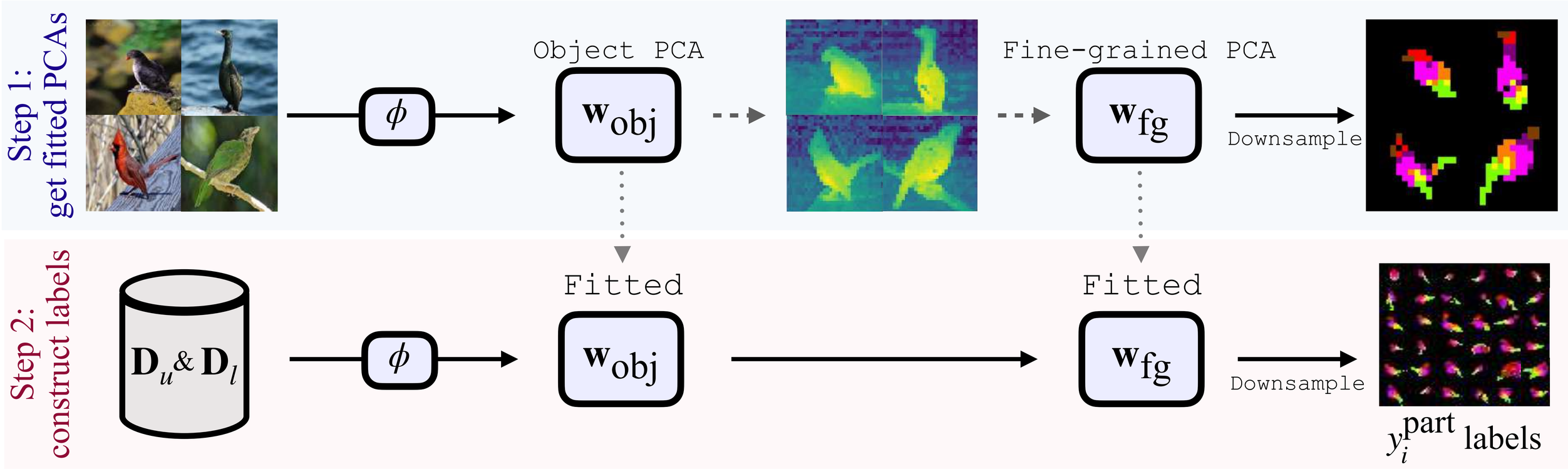
- Object PCA: projection to seperate foreground and background.
- Fine-grained PCA: projection to seperate foreground parts.



# Generalized Category Discovery: Method

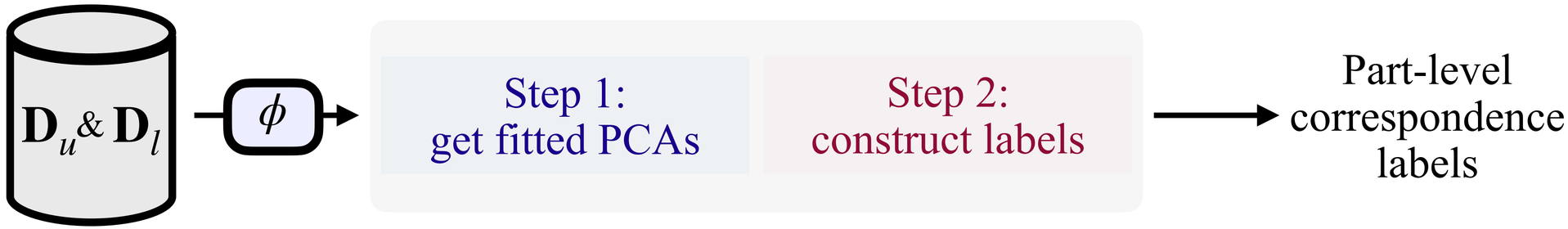
Second, given these PCA projections, we applied it for the rest of the dataset.

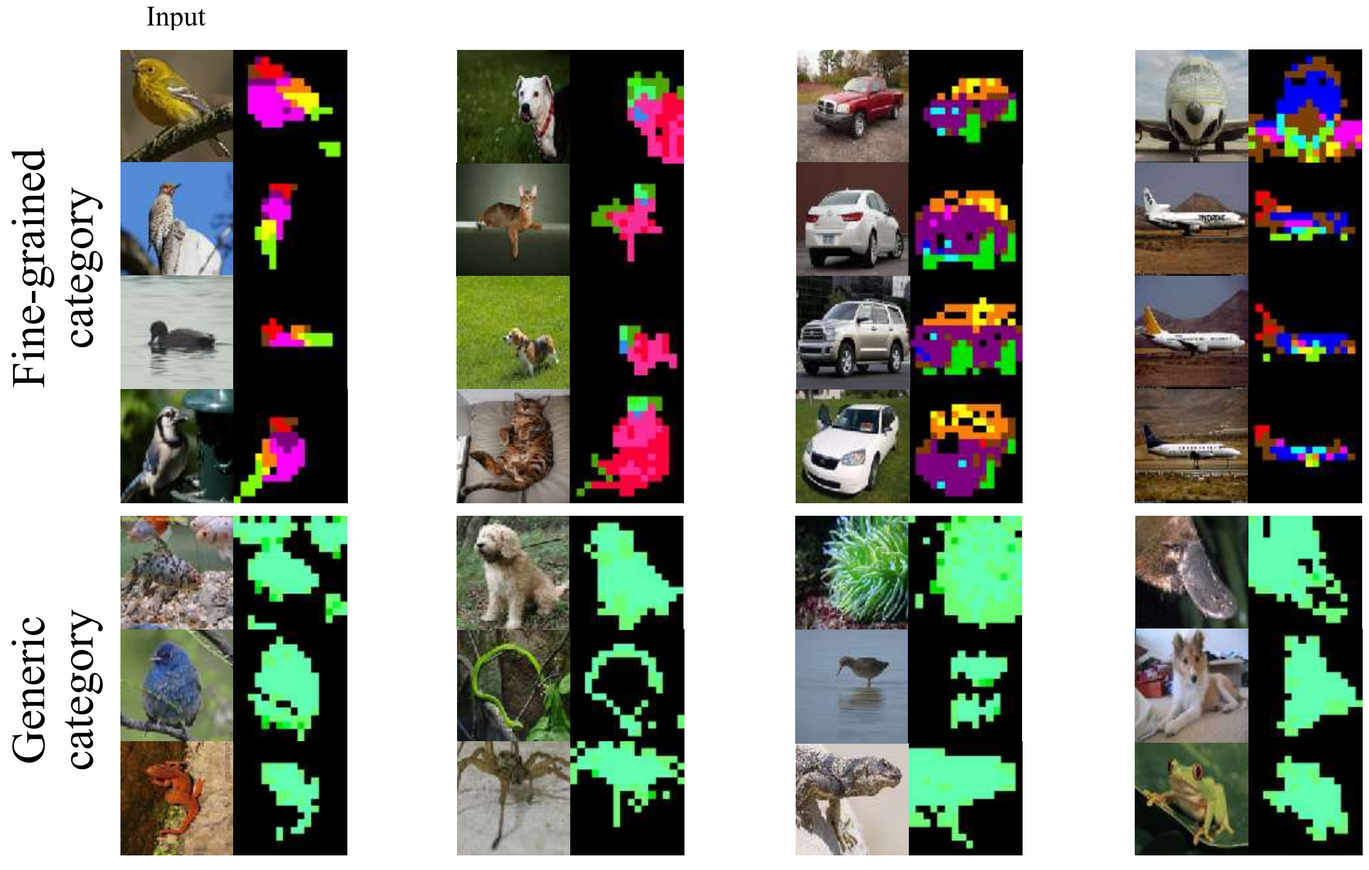
- Object PCA: projection to seperate foreground and background.
- Fine-grained PCA: projection to seperate foreground parts.



# Generalized Category Discovery: Method

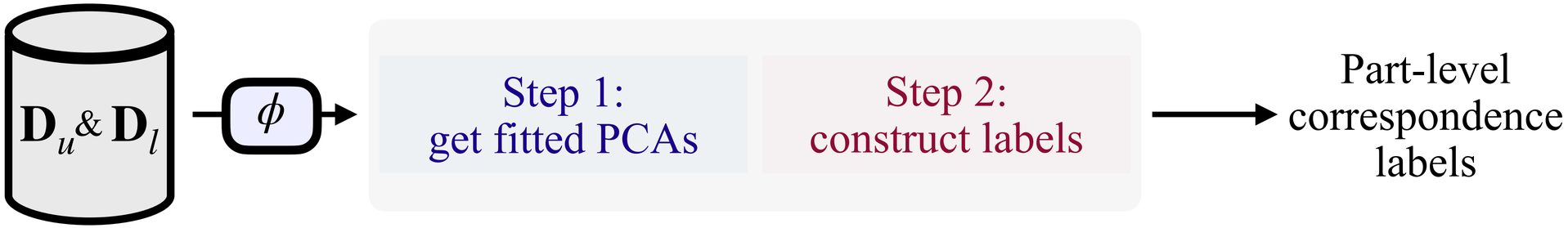
Visualization of the constructed part-level correspondence labels:

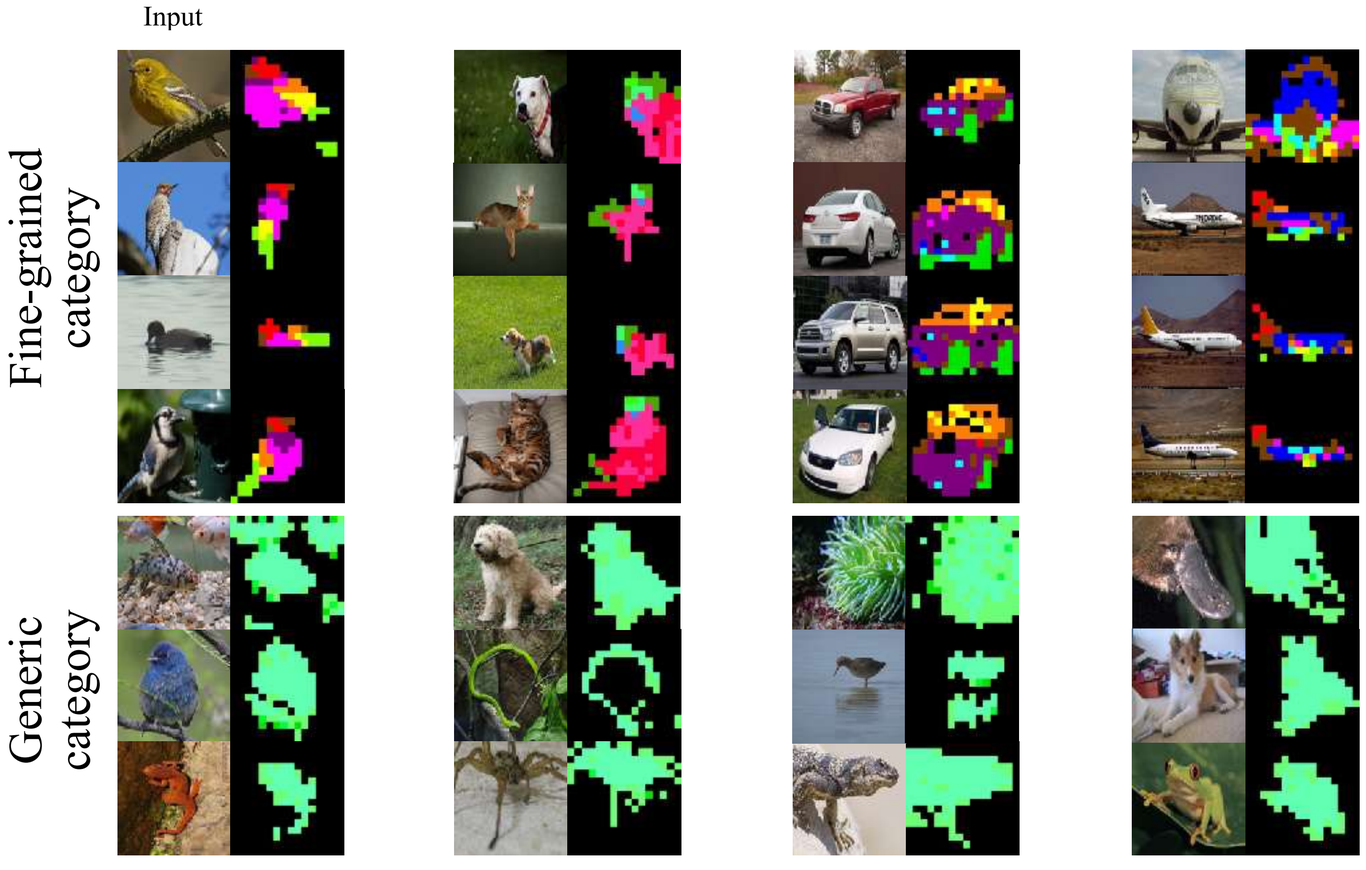


# Generalized Category Discovery: Method

Visualization of the constructed part-level correspondence labels:

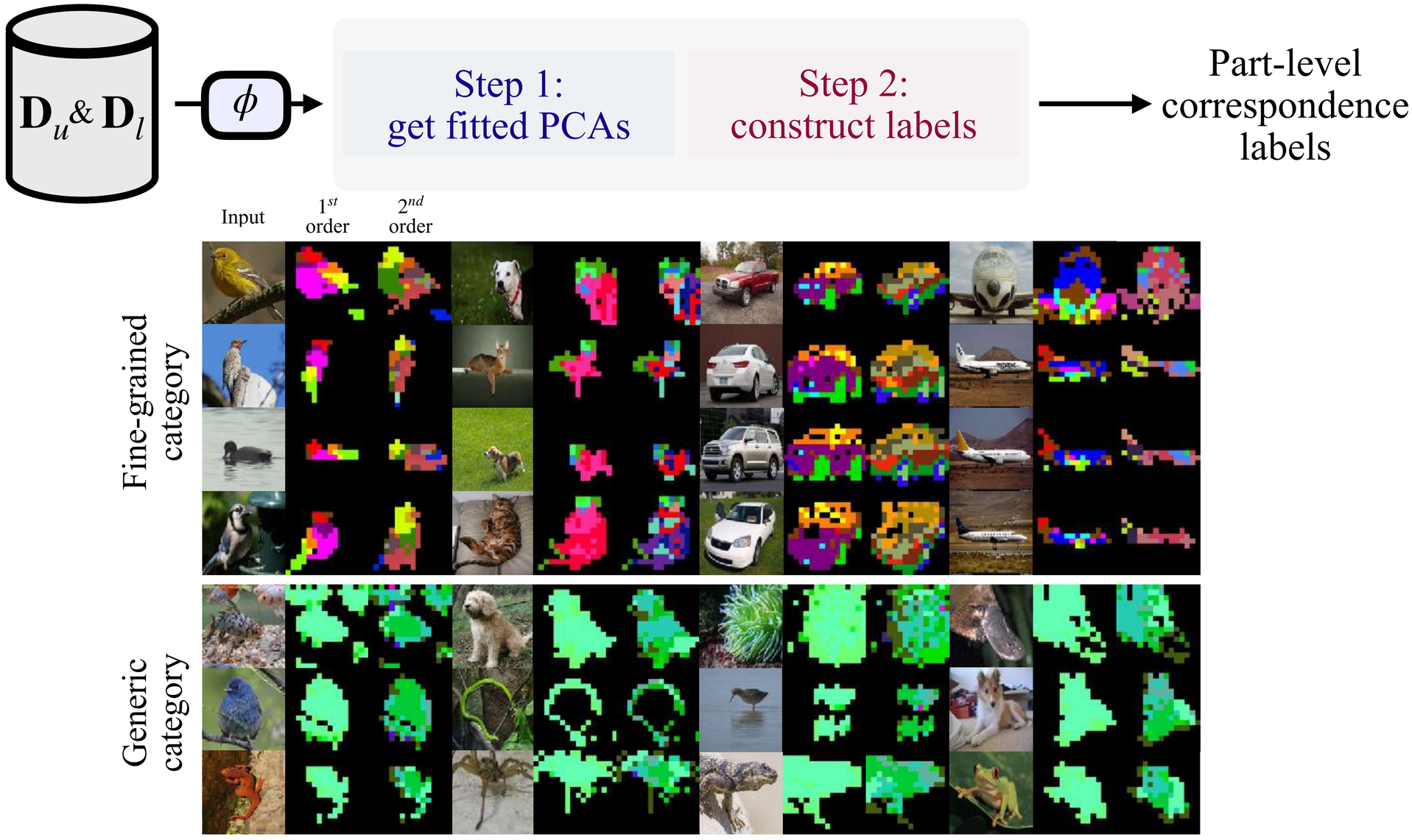



Here, we can actually perform further PCA projection on top of the current labels

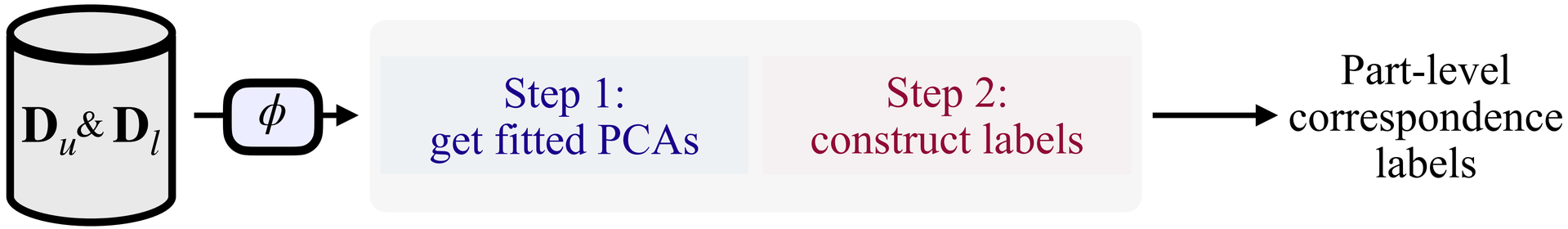
# Generalized Category Discovery: Method

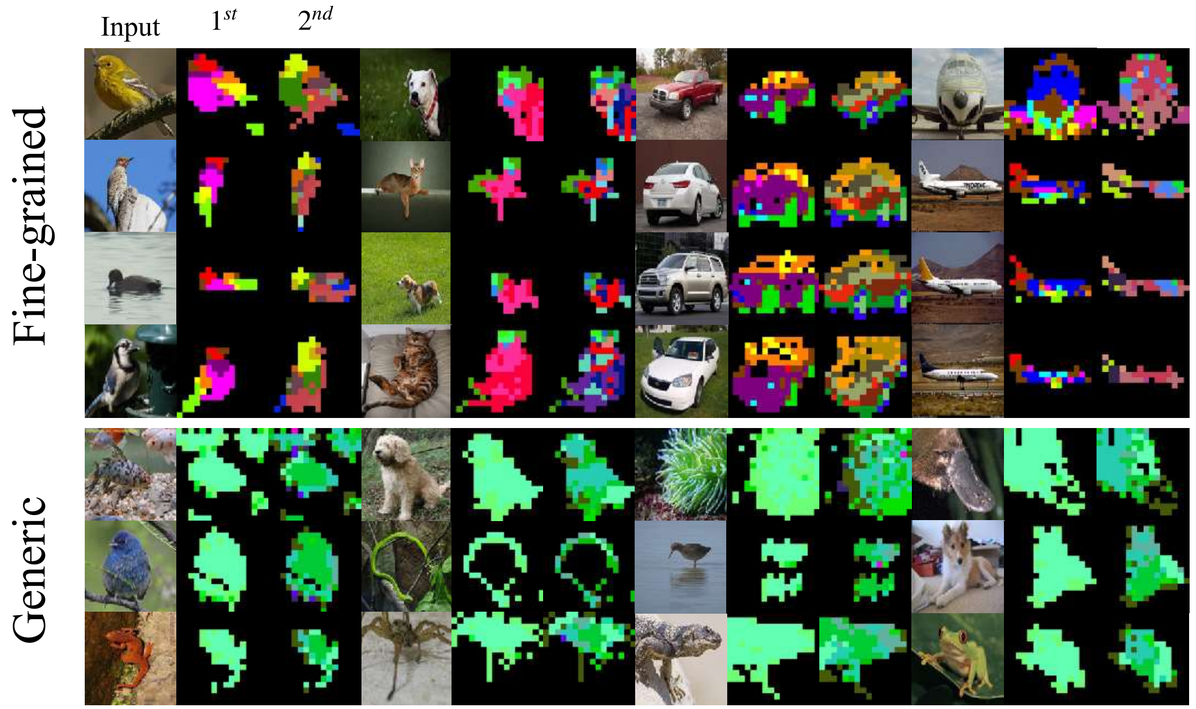
Visualization of the constructed part-level correspondence labels:



# Generalized Category Discovery: Method

Visualization of the constructed part-level correspondence labels:

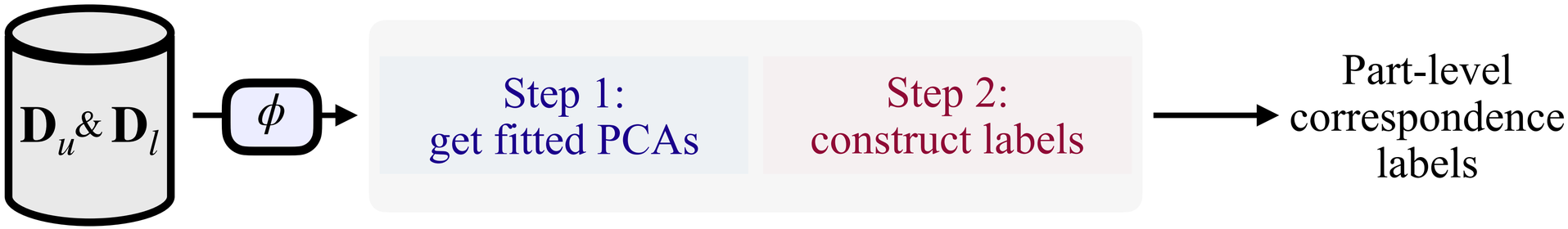



By having this ***explicit control signal***, we do not need to be worry about:

- differing part scales,
- Part orientations, or
- varying number of parts

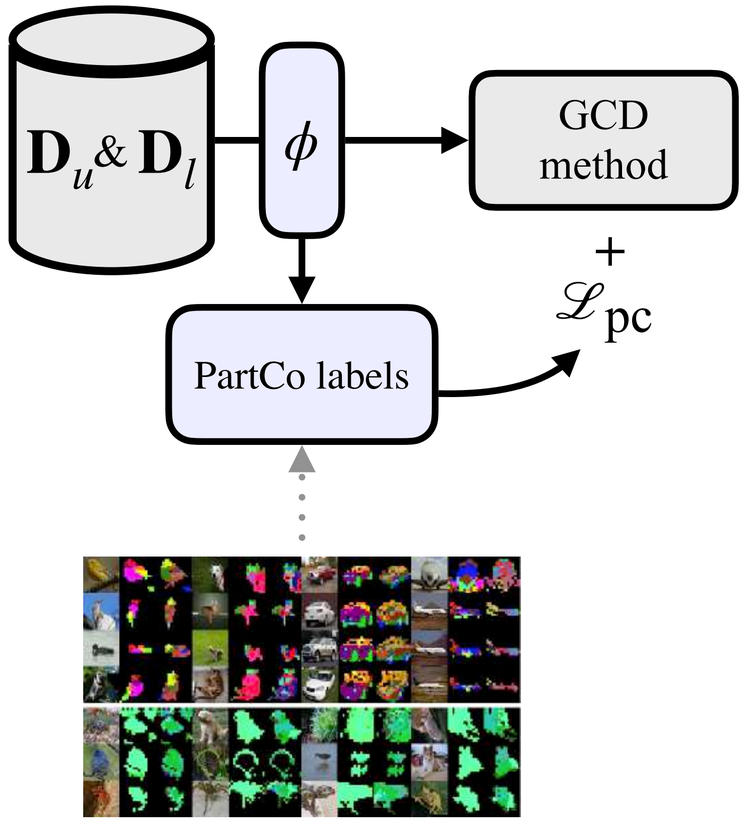
# Generalized Category Discovery: Method




Now, how can we make use of these labels?

# Generalized Category Discovery: Method

Introducing **PartCo** (Part Correspondence) framework

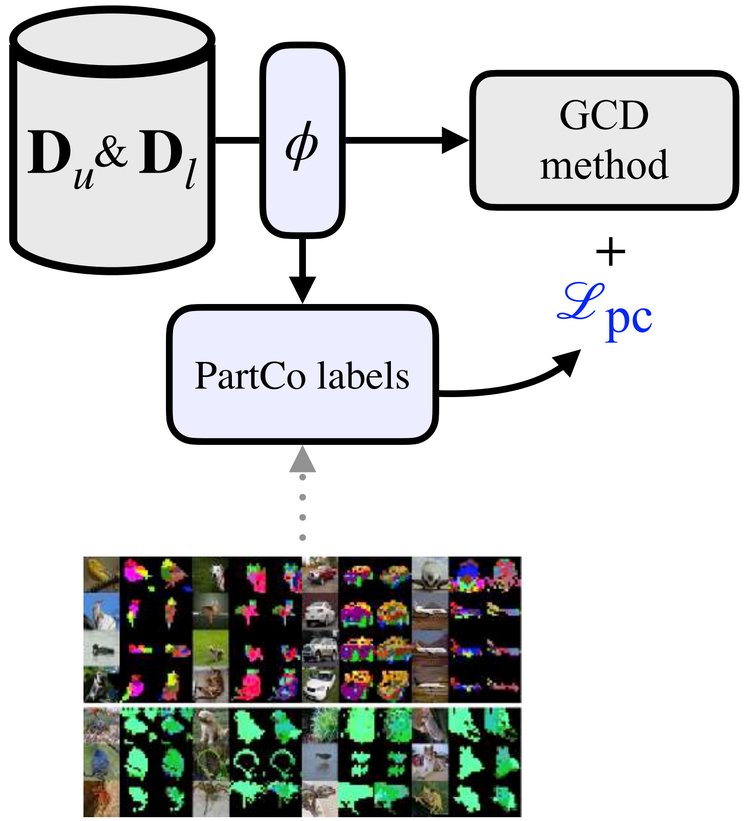


# Generalized Category Discovery: Method

Introducing **PartCo** (Part Correspondence) framework

A simple framework that can be adapt to any current GCD methods.

Specifically, we only add a part-level correspondence loss on top of current GCD methods to induce the part-level observation bias.

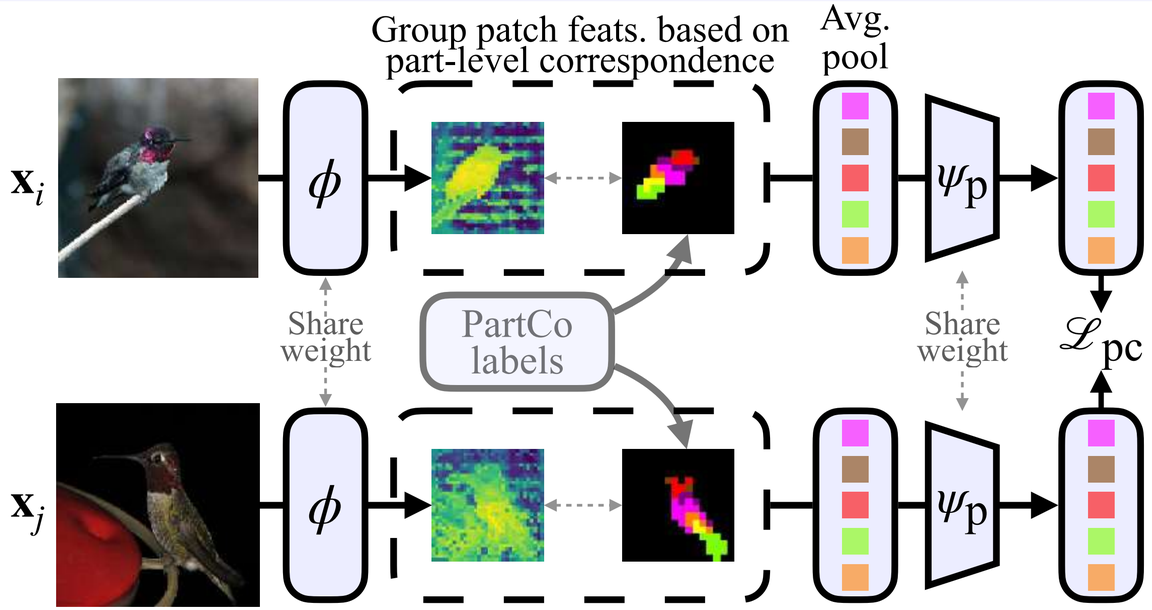


# Generalized Category Discovery: Method

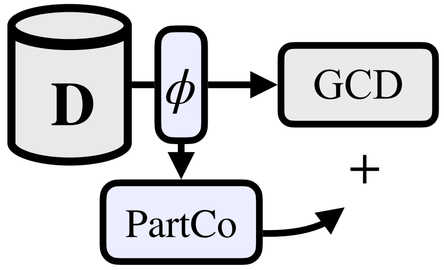
## Part Correspondence Loss

We employs a simple contrastive loss to enforce part-level correspondence.

Where, same semantic part from the same category are pulled together, while different parts are pushed apart.




(b) Part-level correspondent loss ($\mathscr{L}_{pc}$)



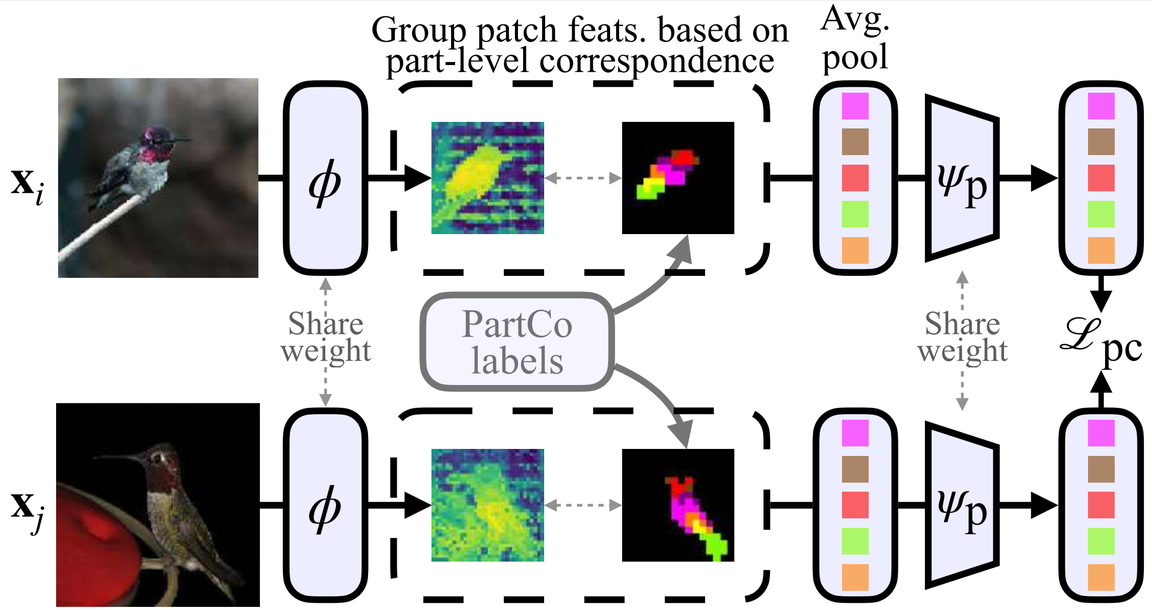

(a) PartCo framework

# Generalized Category Discovery: Method

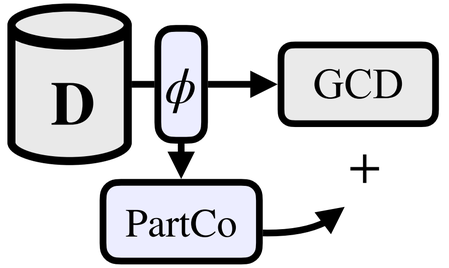
## Part Correspondence Loss

We employs a simple contrastive loss to enforce part-level correspondence.

Where, same semantic part from the same category are pulled together, while different parts are pushed apart.




(b) Part-level correspondent loss ($\mathscr{L}_{pc}$)




(a) PartCo framework

**Procedure:**

*Step 1:* get patch feats. → *Step 2:* group feats. via part label → *Step 3:* average pool each grouped feats → *Step 4:* calculate loss

# Generalized Category Discovery: Experiment

To evaluate our proposed approach, we evaluate it following below settings:

- Baseline models (PartCo-): SimGCD (ICCV' 23) & SelEx (ECCV'24 SOTA)
- Datasets: SSB finegrained & Generic datasets
- Metrics: All, Old, New clustering accuracy (%)

| Dataset | $\lvert\mathbf{D}_l\rvert$ | $M$ | $\lvert\mathbf{D}_u\rvert$ | $K$ |
|---|---|---|---|---|
| CUB [11] | 1.5K | 100 | 4.5K | 200 |
| Stanford-Cars [12] | 2.0K | 98 | 6.1K | 196 |
| FGVC-Aircraft [13] | 1.7K | 50 | 5.0K | 100 |
| CIFAR10 [14] | 12.5K | 5 | 37.5K | 10 |
| CIFAR100 [14] | 20.0K | 80 | 30.0K | 100 |
| ImageNet-100 [15] | 31.9K | 50 | 95.3K | 100 |




# Generalized Category Discovery: Experiment

| Method | Venue | Backbone | CUB | | | Stanford-Cars | | | FGVC-Aircraft | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| *k*-means | - | DINOv2 | 67.6 | 60.6 | 71.1 | 29.4 | 24.5 | 31.8 | 18.9 | 16.9 | 19.9 | 38.6 | 34.0 | 40.0 |
| SimGCD | ICCV'23 | DINOv2 | 71.5 | 78.1 | 68.3 | 71.5 | 81.9 | 66.6 | 63.9 | 69.9 | 60.9 | 69.0 | 76.6 | 65.3 |
| PartCo-SimGCD | **Ours** | DINOv2 | **81.1** | **82.4** | **80.5** | **78.9** | **91.5** | **72.8** | **76.5** | **80.9** | **74.4** | **78.8** | **84.9** | **75.9** |
| SPTNet | ICLR'24 | DINOv2 | 76.3 | 79.5 | 74.6 | 72.3 | 82.0 | 67.5 | 68.0 | 75.2 | 60.9 | 72.2 | 78.9 | 67.6 |
| PartCo-SPTNet | **Ours** | DINOv2 | **82.6** | **82.3** | **81.8** | **80.1** | **92.0** | **73.5** | **78.6** | **77.6** | **79.0** | **80.4** | **84.0** | **78.1** |
| FlipClass | NeurIPS'24 | DINOv2 | 79.3 | 80.7 | 78.5 | 78.0 | 88.0 | 73.2 | 71.1 | 75.1 | 69.1 | 76.1 | 81.3 | 73.6 |
| PartCo-FlipClass | **Ours** | DINOv2 | **85.2** | **86.3** | **84.7** | **80.5** | **92.7** | **74.6** | **77.1** | **78.5** | **76.4** | **80.9** | **85.8** | **78.6** |
| SelEx | ECCV'24 | DINOv2 | 87.4 | **85.1** | 88.5 | 82.2 | **93.7** | 76.7 | 79.8 | 82.3 | 78.6 | 83.1 | **87.0** | 81.3 |
| PartCo-SelEx | **Ours** | DINOv2 | **90.6** | 84.5 | **93.2** | **82.5** | 91.8 | **78.0** | **83.4** | **83.6** | **83.3** | **85.5** | 86.6 | **84.8** |

PartCo further enhance all baselines, e.g, SimGCD (~+10%), SPTNet (~+8%), FlipClass (~+5%), and SelEx (~+2.4%).

# Generalized Category Discovery: Experiment

*Table 1.* Comparison of GCD methods on the SSB benchmark. Results are reported in *ACC* across the 'All', 'Old' and 'New' categories.

| | | | CUB | | | Stanford-Cars | | | FGVC-Aircraft | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Venue | Backbone | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| *k*-means | - | DINOv2 | 67.6 | 60.6 | 71.1 | 29.4 | 24.5 | 31.8 | 18.9 | 16.9 | 19.9 | 38.6 | 34.0 | 40.0 |
| GCD | CVPR'22 | DINOv2 | 71.9 | 71.2 | 72.3 | 65.7 | 67.8 | 64.7 | 55.4 | 47.9 | 59.2 | 64.3 | 62.3 | 65.4 |
| $\mu$GCD | NeurIPS'23 | DINOv2 | 74.0 | 75.9 | 73.1 | 76.1 | 91.0 | 68.9 | 66.3 | 68.7 | 65.1 | 72.1 | 78.6 | 69.0 |
| CiPR | TMLR | DINOv2 | 78.3 | 73.4 | 80.8 | 66.7 | 77.0 | 61.8 | - | - | - | - | - | - |
| ProtoGCD | TPAMI | DINOv2 | 75.7 | 81.5 | 72.9 | 77.6 | 90.5 | 71.5 | 71.1 | 76.3 | 68.5 | 74.8 | 82.7 | 71.0 |
| APL | CVPR'25 | DINOv2 | 75.1 | 79.1 | 73.2 | 73.4 | 87.6 | 66.7 | 68.8 | 74.1 | 66.6 | 72.4 | 80.3 | 68.8 |
| AFGCD | ICCV'25 | DINOv2 | 76.5 | 77.3 | 76.0 | 75.5 | 86.2 | 70.3 | 68.1 | 75.9 | 64.1 | 73.4 | 79.8 | 70.1 |
| DebGCD | ICLR'25 | DINOv2 | 77.5 | 80.8 | 75.8 | 75.4 | 87.7 | 69.5 | 71.9 | 76.0 | 69.8 | 74.9 | 81.5 | 71.7 |
| MOS | CVPR'25 | DINOv2 | 81.1 | 82.1 | 80.6 | 75.5 | 89.6 | 68.7 | 69.7 | 78.1 | 65.5 | 75.4 | 83.3 | 71.6 |
| PartGCD | TMM | DINOv2 | 77.6 | 80.6 | 76.1 | 78.2 | 88.7 | 73.1 | 71.1 | 75.6 | 68.8 | 75.6 | 81.6 | 72.7 |
| SEAL | NeurIPS'25 | DINOv2 | 76.7 | 78.3 | 75.9 | 77.7 | 88.7 | 72.4 | 74.6 | 73.2 | 75.3 | 76.3 | 80.1 | 74.5 |
| RLCD | ICML'25 | DINOv2 | 78.7 | 79.5 | 78.3 | 79.5 | 91.8 | 73.5 | 72.6 | 77.3 | 70.3 | 76.9 | 82.9 | 74.0 |
| AllGCD | ICCV'25 | DINOv2 | 78.4 | 82.8 | 76.2 | 76.2 | 88.3 | 70.4 | - | - | - | - | - | - |
| Hyp-SimGCD | CVPR'25 | DINOv2 | 77.6 | 77.9 | 77.4 | 82.5 | 85.8 | 81.0 | 76.4 | 70.3 | 79.4 | 78.8 | 78.0 | 79.3 |
| ConGCD-SelEx | ICCV'25 | DINOv2 | 86.3 | 87.4 | 85.8 | 79.8 | 93.1 | 73.3 | 81.7 | 83.3 | 81.0 | 82.6 | 87.9 | 80.0 |
| NCGCD-SelEx | NeurIPS'25 | DINOv2 | 87.9 | 85.3 | 89.2 | 81.1 | 90.9 | 79.3 | 83.1 | 88.3 | 82.5 | 84.0 | 88.2 | 83.7 |
| Hyp-SelEx | CVPR'25 | DINOv2 | 90.7 | 85.3 | 93.4 | 83.8 | 93.3 | 79.2 | 83.4 | 82.0 | 84.1 | 86.0 | 86.9 | 85.5 |
| SimGCD | ICCV'23 | DINOv2 | 71.5 | 78.1 | 68.3 | 71.5 | 81.9 | 66.6 | 63.9 | 69.9 | 60.9 | 69.0 | 76.6 | 65.3 |
| PartCo-SimGCD | **Ours** | DINOv2 | **81.1** | **82.4** | **80.5** | **78.9** | **91.5** | **72.8** | **76.5** | **80.9** | **74.4** | **78.8** | **84.9** | **75.9** |
| SPTNet | ICLR'24 | DINOv2 | 76.3 | 79.5 | 74.6 | 72.3 | 82.0 | 67.5 | 68.0 | 75.2 | 60.9 | 72.2 | 78.9 | 67.6 |
| PartCo-SPTNet | **Ours** | DINOv2 | **82.6** | **82.3** | **81.8** | **80.1** | **92.0** | **73.5** | **78.6** | **77.6** | **79.0** | **80.4** | **84.0** | **78.1** |
| FlipClass | NeurIPS'24 | DINOv2 | 79.3 | 80.7 | 78.5 | 78.0 | 88.0 | 73.2 | 71.1 | 75.1 | 69.1 | 76.1 | 81.3 | 73.6 |
| PartCo-FlipClass | **Ours** | DINOv2 | **85.2** | **86.3** | **84.7** | **80.5** | **92.7** | **74.6** | **77.1** | **78.5** | **76.4** | **80.9** | **85.8** | **78.6** |
| SelEx | ECCV'24 | DINOv2 | 87.4 | **85.1** | 88.5 | 82.2 | **93.7** | 76.7 | 79.8 | 82.3 | 78.6 | 83.1 | **87.0** | 81.3 |
| PartCo-SelEx | **Ours** | DINOv2 | **90.6** | 84.5 | **93.2** | **82.5** | 91.8 | **78.0** | **83.4** | **83.6** | **83.3** | **85.5** | 86.6 | **84.8** |
| *k*-means | - | DINOv3 | 69.8 | 64.7 | 72.3 | 59.0 | 50.8 | 63.1 | 40.3 | 35.3 | 42.8 | 56.4 | 50.3 | 59.4 |
| SimGCD | ICCV'23 | DINOv3 | 75.9 | **83.8** | 72.0 | 73.9 | 79.4 | 71.3 | 71.4 | **84.4** | 65.0 | 73.7 | 82.5 | 69.4 |
| PartCo-SimGCD | **Ours** | DINOv3 | **78.8** | 83.4 | **76.6** | **78.5** | **86.5** | **74.6** | **75.3** | 81.6 | **72.2** | **77.5** | **83.8** | **74.5** |
| SelEx | ECCV'24 | DINOv3 | 83.5 | 78.1 | 86.2 | 78.8 | 90.3 | 73.3 | 73.9 | 77.3 | 72.4 | 78.7 | 81.9 | 77.3 |
| PartCo-SelEx | **Ours** | DINOv3 | **86.1** | **84.4** | **87.0** | **81.7** | **92.1** | **76.6** | **76.9** | **82.2** | **74.2** | **81.6** | **86.2** | **79.3** |

# Generalized Category Discovery: Experiment

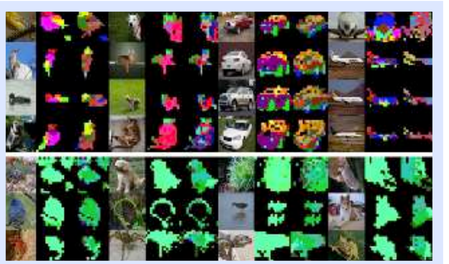
*Table 2.* Comparison of GCD methods on the generic benchmark. Results are reported in *ACC* across the 'All', 'Old' and 'New' categories.

| | | | CIFAR10 | | | CIFAR100 | | | ImageNet-100 | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Venue | Backbone | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| *k*-means | - | DINOv2 | 94.9 | 95.2 | 94.8 | 70.9 | 70.8 | 72.1 | 78.3 | 80.5 | 77.2 | 81.4 | 82.2 | 81.4 |
| GCD | CVPR'22 | DINOv2 | 97.8 | 99.0 | 97.1 | 79.6 | 84.5 | 69.9 | 78.5 | 89.5 | 73.0 | 85.3 | 91.0 | 80.0 |
| AMEND | WACV'24 | DINOv2 | 97.7 | 96.6 | 98.3 | 83.5 | 83.0 | 84.5 | 87.3 | 95.1 | 83.4 | 89.5 | 91.6 | 88.7 |
| CiPR | TMLR | DINOv2 | 99.0 | 98.7 | 99.2 | 90.3 | 89.0 | 93.1 | 88.2 | 87.6 | 88.5 | 92.5 | 91.8 | 93.6 |
| SPTNet | ICLR'24 | DINOv2 | 98.9 | 99.1 | 98.8 | 89.0 | 91.5 | 79.2 | 90.1 | 96.1 | 87.1 | 92.7 | 95.6 | 88.4 |
| FlipClass | NeurIPS'24 | DINOv2 | 99.0 | 98.2 | 99.4 | 91.7 | 90.4 | 94.2 | 91.0 | 96.3 | 88.3 | 93.9 | 95.0 | 94.0 |
| DebGCD | ICLR'25 | DINOv2 | 98.9 | 97.5 | 99.6 | 90.1 | 90.9 | 88.6 | 93.2 | 97.0 | 91.2 | 94.1 | 95.1 | 93.1 |
| SEAL | NeurIPS'25 | DINOv2 | 98.9 | 98.1 | 99.3 | 89.8 | 90.4 | 89.5 | 91.3 | 93.3 | 90.3 | 93.3 | 93.9 | 93.0 |
| RLCD | ICML'25 | DINOv2 | 99.0 | 98.9 | 99.1 | 91.2 | 91.2 | 91.2 | 92.1 | 96.2 | 90.0 | 94.1 | 95.4 | 93.4 |
| Hyp-SimGCD | CVPR'25 | DINOv2 | 98.9 | 97.7 | 99.5 | 91.5 | 90.0 | 94.6 | 91.9 | 96.2 | 89.8 | 94.1 | 94.6 | 94.6 |
| Hyp-SelEx | CVPR'25 | DINOv2 | 98.6 | 98.1 | 98.9 | 88.6 | 91.5 | 82.8 | 92.3 | 96.4 | 90.2 | 93.2 | 95.3 | 90.6 |
| SimGCD | ICCV'23 | DINOv2 | 98.7 | 96.7 | **99.7** | 88.5 | 89.2 | **87.2** | 89.9 | **95.5** | 87.1 | 92.4 | 93.8 | **91.3** |
| PartCo-SimGCD | **Ours** | DINOv2 | **99.0** | **98.7** | 99.2 | **89.0** | **92.0** | 83.0 | **90.1** | 92.0 | **89.2** | **92.7** | **94.2** | 90.4 |
| SelEx | ECCV'24 | DINOv2 | 98.5 | 98.8 | 98.5 | 87.7 | 90.8 | 81.5 | 90.9 | 96.2 | 88.3 | 92.4 | 95.3 | 89.4 |
| PartCo-SelEx | **Ours** | DINOv2 | **99.2** | **99.4** | **98.9** | **90.0** | **92.8** | **84.3** | **94.5** | **97.8** | **92.8** | **94.6** | **96.7** | **92.0** |
| *k*-means | - | DINOv3 | 94.1 | 95.1 | 93.6 | 65.5 | 66.4 | 63.5 | 78.3 | 78.3 | 78.3 | 79.3 | 79.9 | 78.5 |
| SimGCD | ICCV'23 | DINOv3 | 98.4 | **98.7** | 98.2 | 84.3 | 88.3 | 74.3 | 92.2 | 96.5 | 90.0 | 91.6 | **94.8** | 87.5 |
| PartCo-SimGCD | **Ours** | DINOv3 | **98.9** | 98.2 | **99.3** | **85.0** | **88.9** | **77.1** | **93.7** | **96.5** | **92.3** | **92.5** | 94.5 | **89.6** |
| SelEx | ECCV'24 | DINOv3 | 98.2 | 99.0 | 97.7 | 87.7 | 88.7 | 85.7 | 93.4 | **96.9** | 91.6 | 93.1 | 94.9 | 91.6 |
| PartCo-SelEx | **Ours** | DINOv3 | **98.3** | **99.1** | **97.9** | **90.2** | **91.7** | **87.0** | **93.8** | 96.6 | **92.4** | **94.1** | **95.8** | **92.4** |

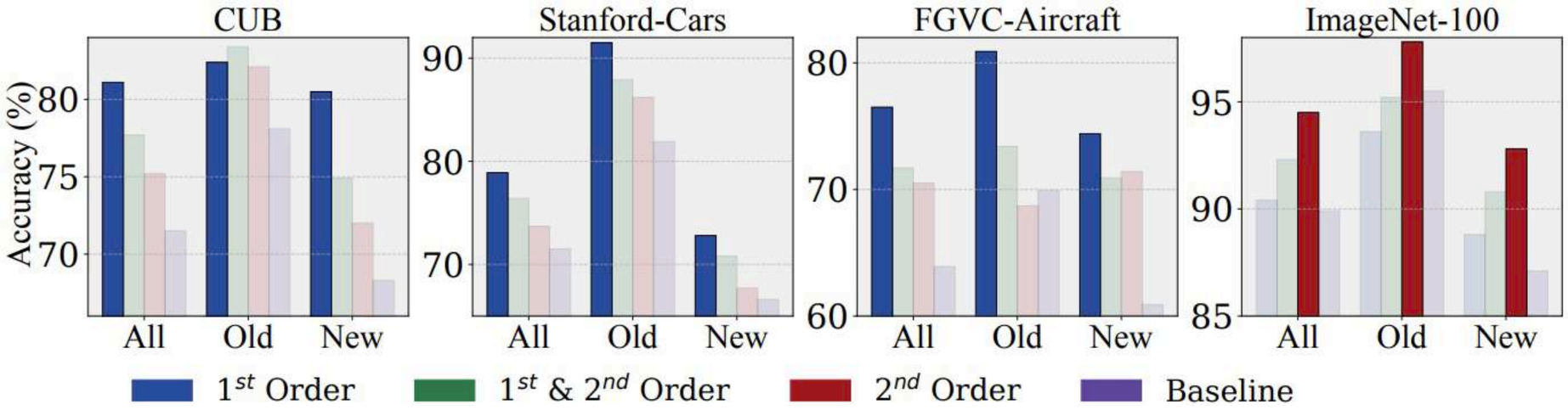
# Generalized Category Discovery: Experiment

We perform ablation study to assess how “finegrained” thus our labels should be. We find that:

1. For finegrained datasets: 1st order is enough.
2. For generic datasets: better to use 2nd order labels.



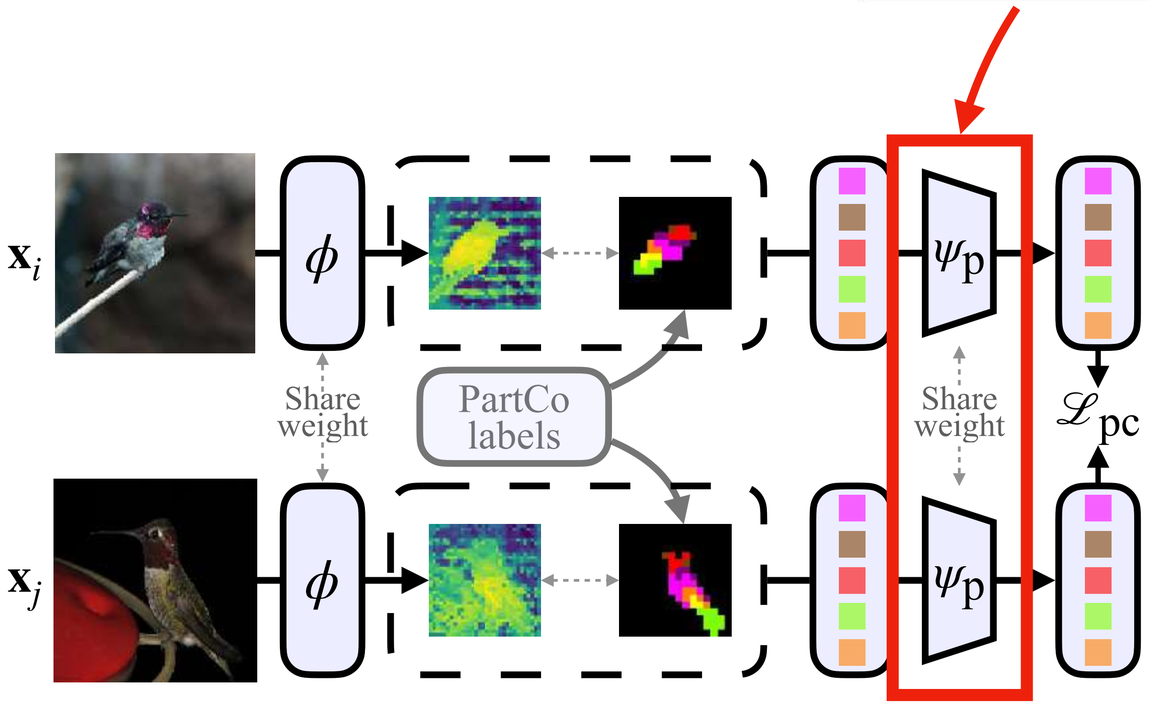
PartCo labels
(1st & 2nd order)



# Generalized Category Discovery: Experiment

We also perform study on the output dimension of our part-level projection layer. We find that output dim. of 128 results to better category discovery performance.

| | CUB | | | Stanford-Cars | | |
|---|---|---|---|---|---|---|
| Dim. | All | Old | New | All | Old | New |
| 64 | 79.3 | 76.7 | 80.6 | 76.9 | 88.3 | 71.4 |
| 128 | **81.1** | 82.4 | 80.5 | **78.9** | 91.5 | 72.8 |
| 256 | 79.8 | 80.7 | 79.4 | 76.4 | 90.5 | 69.7 |
| 512 | 78.2 | 83.1 | 75.7 | 75.6 | 87.5 | 69.9 |



# Generalized Category Discovery: Conclusion

**In this work we show that:**

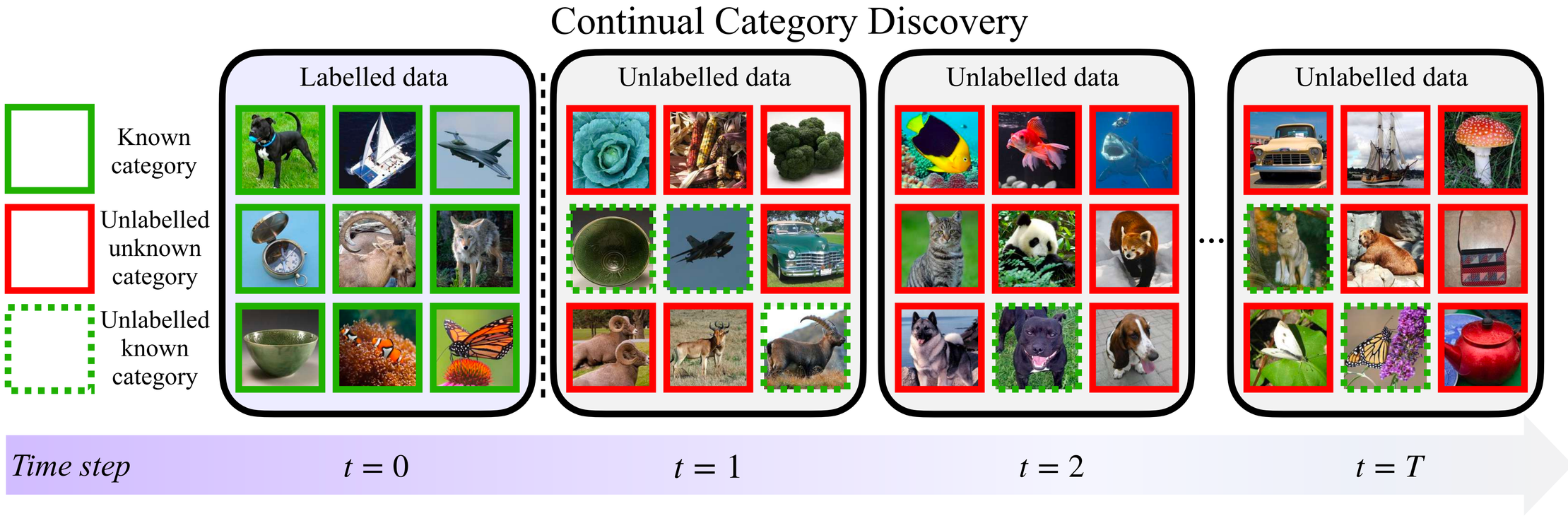
1. Part-level observation are indeed an effective “vehicle” to transfer knowledge between ‘seen’ and ‘unseen’ categories.

2. We developed a novel explicit part-level correspondence framework with minimal integration to any GCD methods.

3. Our method achieves SOTA performance on various GCD benchmarks.

# Task B: Continual Category Discovery

# Continual Category Discovery: Intro

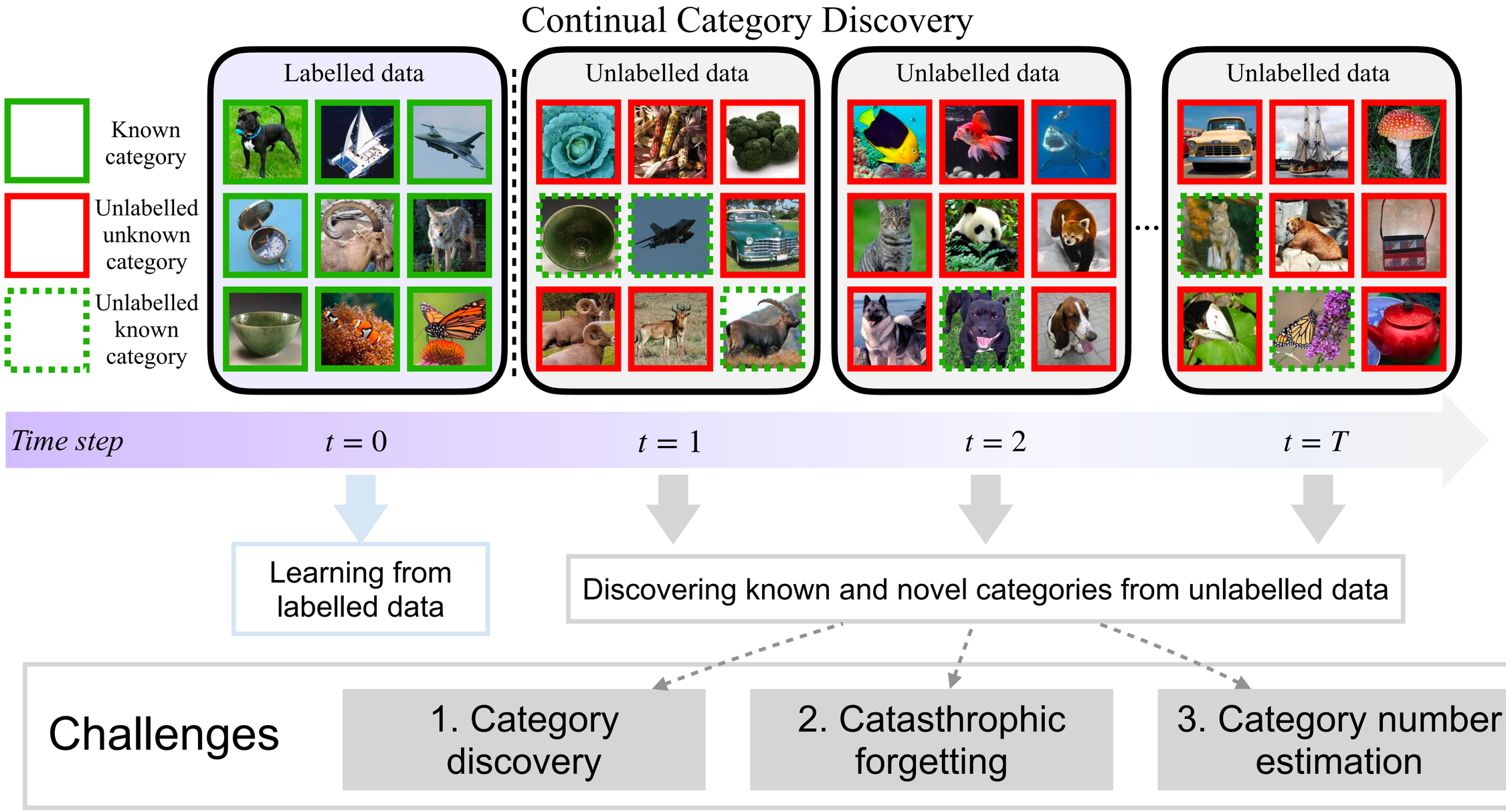



In CCD, we …

(1) first train the model with a labelled dataset

(2) subsequently discover novel categories from a continuous stream of unlabelled data.

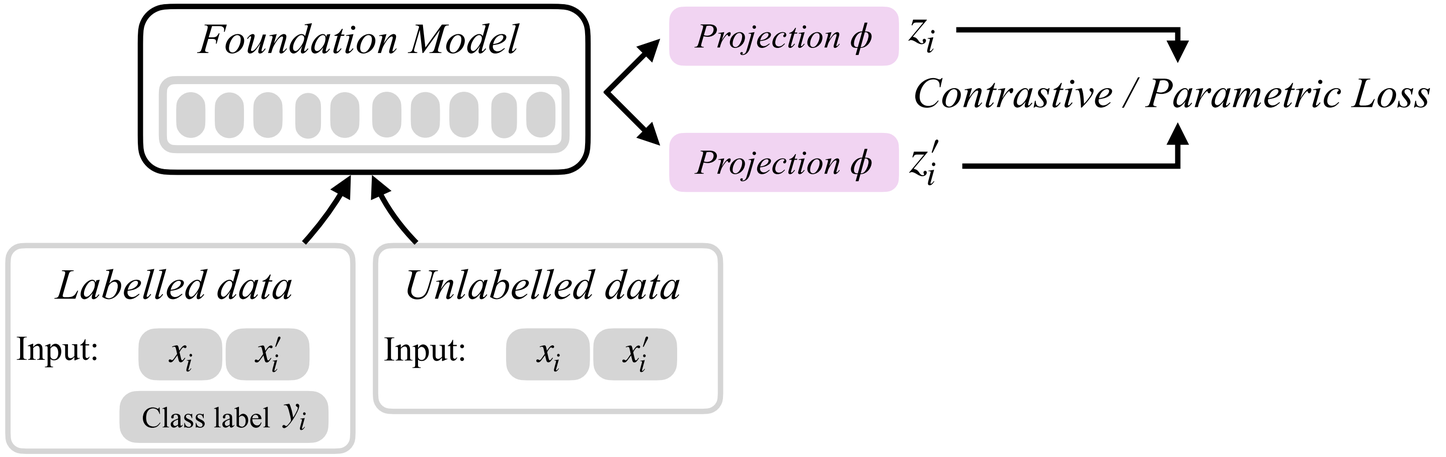
# Continual Category Discovery: Intro



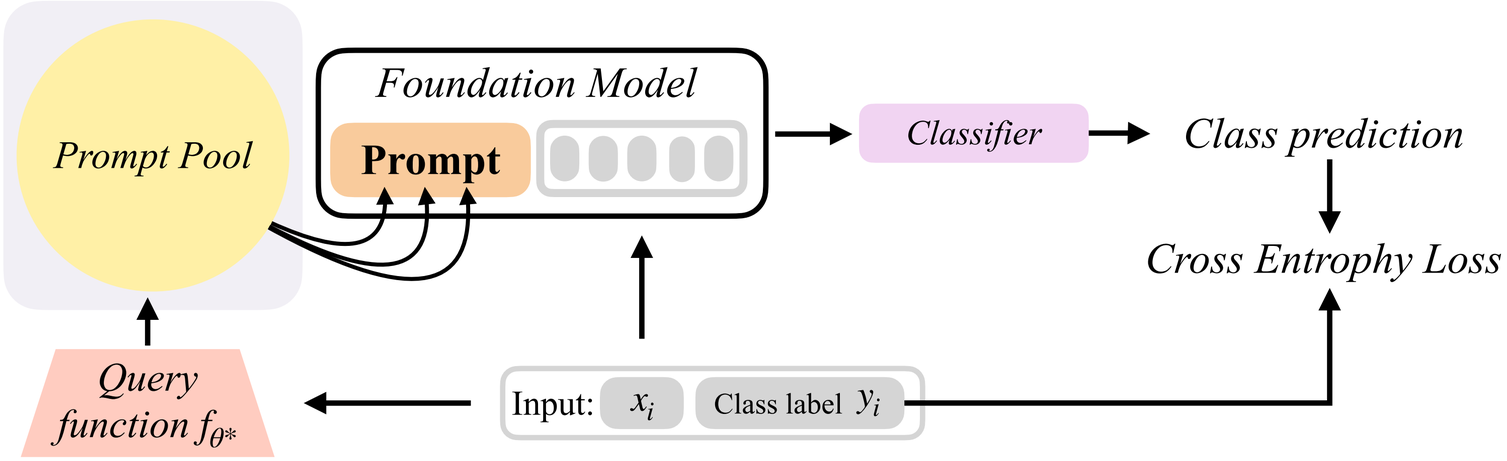
# Continual Category Discovery: Background

## Generalized Category Discovery (GCD)

- In GCD, given a dataset, a subset of which has class labels, the model is tasked to categorize all unlabelled images.




## Prompt-based Supervised Continual Learning

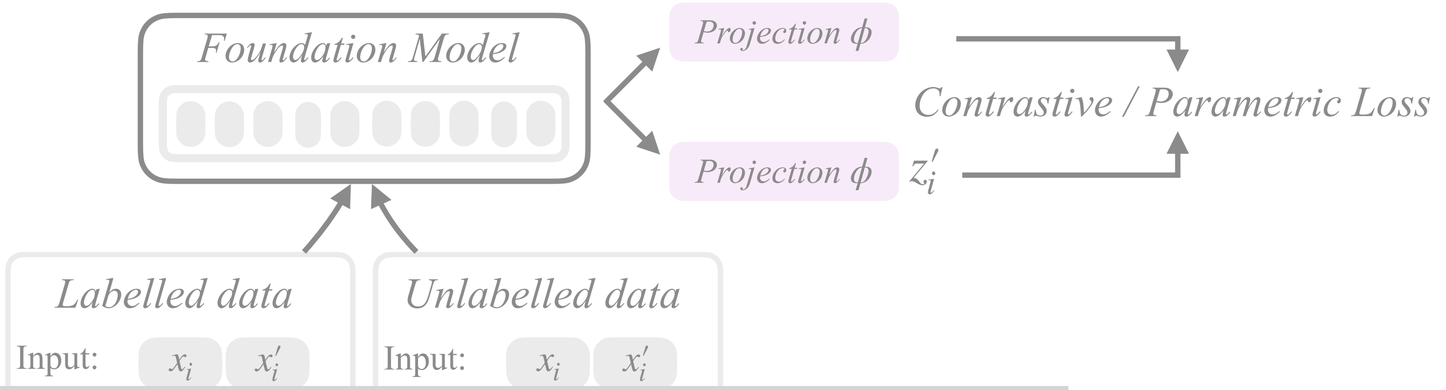



- The model leverages a prompt pool (e.g., L2P, Dual Prompt) to guide a foundation model in supervised continual learning.
- It extracts feature queries to retrieve top-k relevant propts from the pool, which, along with class labels, are used to supervised the model.

# Continual Category Discovery: Background

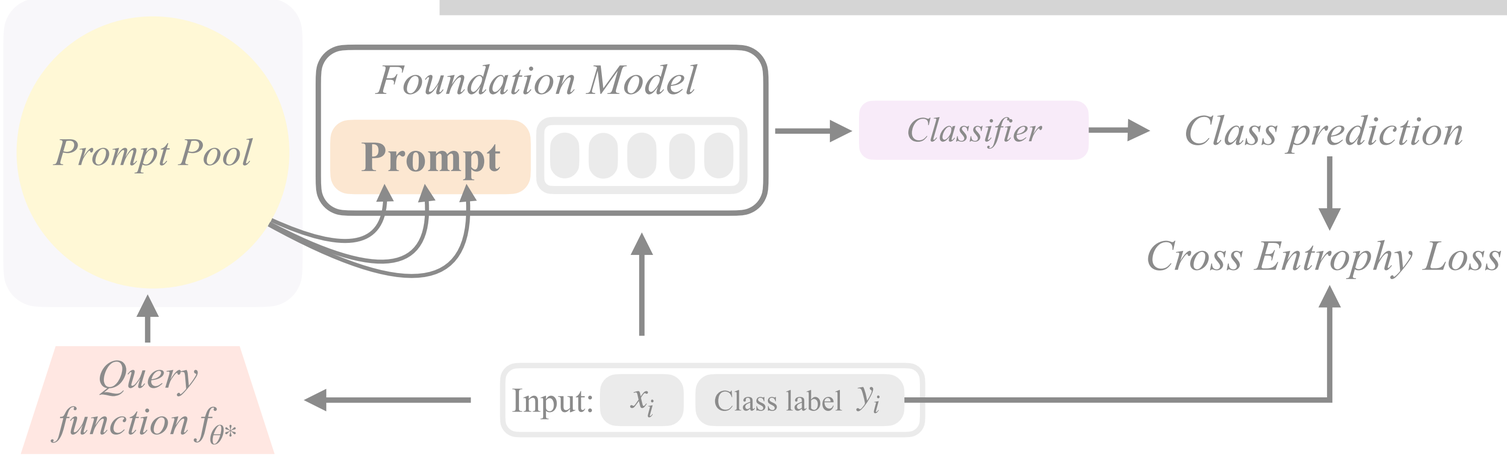
## Generalized Category Discovery (GCD)

- In GCD, given a dataset, a subset of which has class labels, the model is tasked to categorize all unlabelled images.




**Neither method is effective in CCD** because only unlabelled data is available during the discovery stage. Furthermore, the unlabelled data may contain categories not present in the initial labelled set.

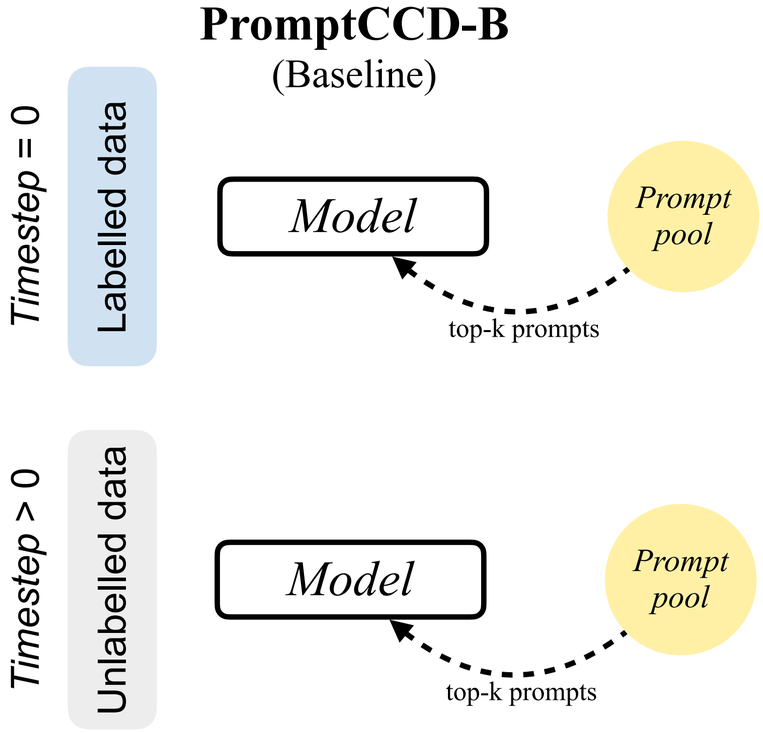
## Prompt-based Supe




- The model leverages a prompt pool (e.g., L2P, Dual Prompt) to guide a foundation model in supervised continual learning.
- It extracts feature queries to retrieve top-k relevant propts from the pool, which, along with class labels, are used to supervised the model.

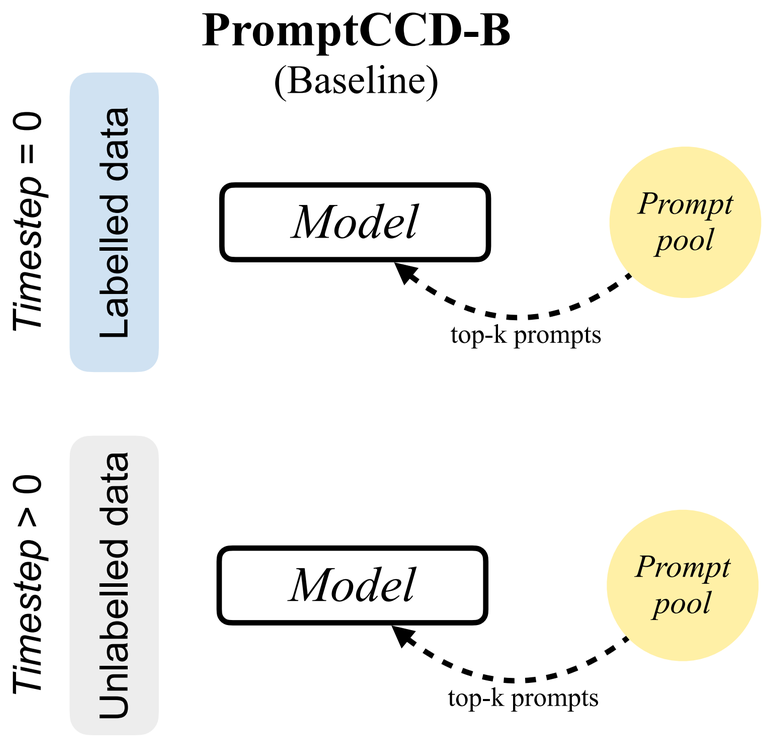
# Continual Category Discovery: Background

Therefore, we propose a baseline model for CCD, named, PromptCCD-B (baseline) to adapt vision foundation model with prompt pool modules, e.g., L2P & DP for CCD task:



# Continual Category Discovery: Background

Preliminary results of PromptCCD-B (baseline) by repurposing DINO model with recent prompt pool modules, e.g., L2P & DP for CCD task:




*CCD comparison results between frozen DINO & the PromptCCD-B with different variants of prompt pool designs.*

| Method | CIFAR100 | | | ImageNet-100 | | |
|---|---|---|---|---|---|---|
| | *All* | *Old* | *New* | *All* | *Old* | *New* |
| Frozen DINO | 56.45 | 68.10 | 53.21 | 67.25 | 73.25 | 65.44 |
| PromptCCD-B *w*/L2P (Ours) | 51.59 ± 6.3 | 67.27 ± 8.7 | 46.14 ± 6.1 | 66.14 ± 2.3 | 81.05 ± 1.5 | 61.36 ± 3.2 |
| PromptCCD-B *w*/DP (Ours) | **59.60** ± 1.2 | 78.93 ± 1.3 | 54.14 ± 1.6 | **70.64** ± 1.3 | 83.46 ± 0.4 | 67.24 ± 1.8 |

| Method | TinyImageNet | | | CUB | | |
|---|---|---|---|---|---|---|
| | *All* | *Old* | *New* | *All* | *Old* | *New* |
| Frozen DINO | 49.25 | 59.21 | 45.94 | 39.21 | 68.29 | 30.86 |
| PromptCCD-B *w*/L2P (Ours) | 56.66 ± 0.4 | 66.05 ± 0.8 | 53.69 ± 0.4 | 51.31 ± 1.0 | 72.43 ± 1.0 | 44.27 ± 1.4 |
| PromptCCD-B *w*/DP (Ours) | **58.61** ± 1.5 | 66.61 ± 0.6 | 55.84 ± 1.7 | **56.30** ± 1.1 | 78.64 ± 1.7 | 48.91 ± 1.1 |

# Continual Category Discovery: Background

## Limitation of current baseline

- **Lack of *explicit guidance*** (labels) may introduce bias when discovering categories from the continuous unlabelled data stream.
- **Fixed-size prompt pool limits scalability**. The ablation study shows that changing the number of pool size does not significantly impact the CCD performance, even when the size matches the total classes in the CUB dataset.
- There is **no mechanism to estimate the number of categories**, which is a crucial and under-explored challenge in CCD task.

*Ablation study on the prompt pool size using CUB datasets*

| | *w/* L2P | | | *w/* DP | | |
|---|---|---|---|---|---|---|
| Pool Size | *All* | *Old* | *New* | *All* | *Old* | *New* |
| 5 | 48.69 | 70.12 | 41.51 | 55.54 | 77.78 | 48.21 |
| 10 | 50.57 | 73.22 | 43.28 | 55.21 | 77.24 | 48.04 |
| 20 | 49.32 | 70.60 | 42.29 | 55.41 | 76.31 | 48.18 |
| 40 | 48.59 | 69.26 | 41.43 | 53.97 | 77.38 | 46.26 |
| 100 | 51.84 | 73.09 | 44.39 | 55.12 | 77.26 | 47.82 |
| 200 | 49.40 | 71.79 | 41.94 | 54.54 | 79.05 | 46.29 |

# Continual Category Discovery: Method

Our solution: to design a Gaussian Mixture Prompting Module (**GMP**)

# Continual Category Discovery: Method

Our solution: to design a Gaussian Mixture Prompting Module (**GMP**)

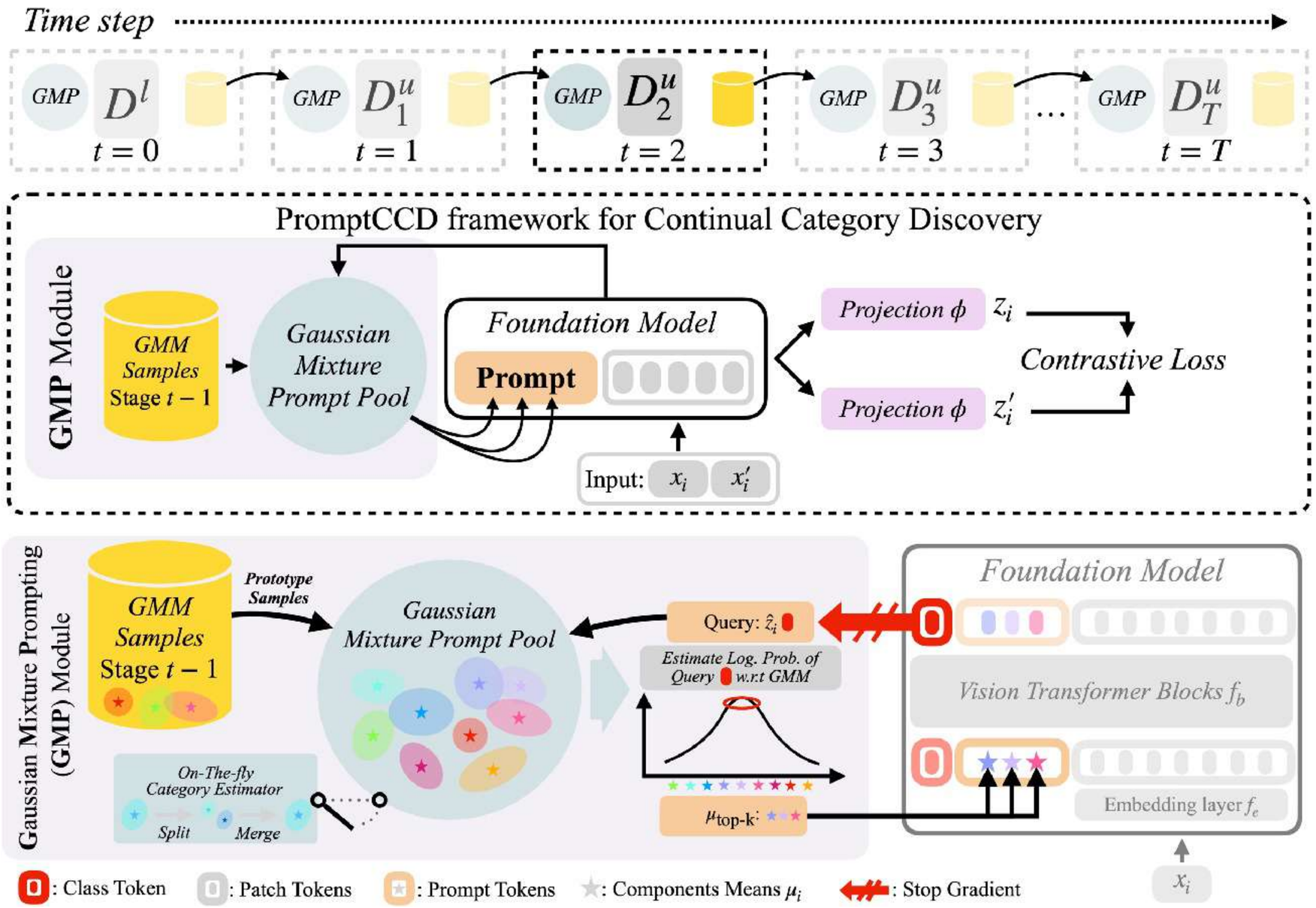

Our GMP learns a parameter-efficient, learnable GMM as a pool of prompts, leading to a new framework, called **PromptCCD**.

# Continual Category Discovery: Method


**GMP's prompts vs others**

- Unlike other prompt pools, GMP's prompt serves dual roles, i.e,. (1) **prompt to Instruct** and (2) **prompt as a class prototype** (*explicit guidance, which is not true for other prompt pool design*).
- GMP module **generates samples** to preserved the learned class prototypes for next incremental stage.
- GMP module enables **on-the-fly** category number estimation during discovery.

# Continual Category Discovery: Method



# Continual Category Discovery: Experiments

| Stages | C100 [28] | | IN-100 [43] | | Tiny [30] | | C-101 [12] | | Aircraft [35] | | SCars [27] | | CUB [48] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $C$ | # | $C$ | # | $C$ | # | $C$ | # | $C$ | # | $C$ | # | $C$ | # |
| Stage 0 ($D^l$) | 70 | 30.45$K$ | 70 | 77.46$K$ | 140 | 60.90$K$ | 71 | 4.70$K$ | 70 | 1.98$K$ | 130 | 4.62$K$ | 140 | 3.65$K$ |
| Stage 1 ($D_1^u$) | 80 | 5.95$K$ | 80 | 15.14$K$ | 160 | 11.90$K$ | 81 | 0.73$K$ | 80 | 0.37$K$ | 152 | 0.98$K$ | 160 | 0.71$K$ |
| Stage 2 ($D_2^u$) | 90 | 6.55$K$ | 90 | 16.66$K$ | 180 | 13.10$K$ | 91 | 0.65$K$ | 90 | 0.43$K$ | 174 | 1.13$K$ | 180 | 0.79$K$ |
| Stage 3 ($D_3^u$) | 100 | 7.05$K$ | 100 | 17.94$K$ | 200 | 14.10$K$ | 101 | 1.12$K$ | 100 | 0.55$K$ | 196 | 1.38$K$ | 200 | 0.85$K$ |

## CCD datasets & splits

- Above table shows the statistics of the CCD benchmarks datasets.
- While the table on right side shows how we split the classes for each stages.

| Class splits | $D^l$ | $D_1^u$ | $D_2^u$ | $D_3^u$ |
|---|---|---|---|---|
| $\{y_i \mid y_i \leq 0.7 * \lvert\mathcal{Y}\rvert\}$ | 87% | 7% | 3% | 3% |
| $\{y_i \mid 0.7 * \lvert\mathcal{Y}\rvert < y_i \leq 0.8 * \lvert\mathcal{Y}\rvert\}$ | 0% | 70% | 20% | 10% |
| $\{y_i \mid 0.8 * \lvert\mathcal{Y}\rvert < y_i \leq 0.9 * \lvert\mathcal{Y}\rvert\}$ | 0% | 0% | 90% | 10% |
| $\{y_i \mid 0.9 * \lvert\mathcal{Y}\rvert < y_i \leq \lvert\mathcal{Y}\rvert\}$ | 0% | 0% | 0% | 100% |

# Continual Category Discovery: Experiments

## Variants of PromptCCD

- ***Comparison with PromptCCD variants and frozen DINO:*** This table further highlights the need to adapt foundation models for CCD.

- ***Performance:*** Our PromptCCD w/GMP consistently outperforms other models.

*CCD results comparison between fronzen DINO, PromptCCD-B, and our proposed PromptCCD w/ GMP.*

| Method | CIFAR100 | | | ImageNet-100 | | |
|---|---|---|---|---|---|---|
| | *All* | *Old* | *New* | *All* | *Old* | *New* |
| Frozen DINO | 56.45 | 68.10 | 53.21 | 67.25 | 73.25 | 65.44 |
| PromptCCD-B $w$/L2P (Ours) | 51.59 ± 6.3 | 67.27 ± 8.7 | 46.14 ± 6.1 | 66.14 ± 2.3 | 81.05 ± 1.5 | 61.36 ± 3.2 |
| PromptCCD-B $w$/DP (Ours) | 59.60 ± 1.2 | 78.93 ± 1.3 | 54.14 ± 1.6 | 70.64 ± 1.3 | 83.46 ± 0.4 | 67.24 ± 1.8 |
| PromptCCD $w$/GMP (Ours) | **63.97** ± 1.4 | 76.67 ± 2.6 | 60.01 ± 1.7 | **75.38** ± 0.7 | 81.16 ± 0.7 | 73.71 ± 0.8 |

| Method | TinyImageNet | | | CUB | | |
|---|---|---|---|---|---|---|
| | *All* | *Old* | *New* | *All* | *Old* | *New* |
| Frozen DINO | 49.25 | 59.21 | 45.94 | 39.21 | 68.29 | 30.86 |
| PromptCCD-B $w$/L2P (Ours) | 56.66 ± 0.4 | 66.05 ± 0.8 | 53.69 ± 0.4 | 51.31 ± 1.0 | 72.43 ± 1.0 | 44.27 ± 1.4 |
| PromptCCD-B $w$/DP (Ours) | 58.61 ± 1.5 | 66.61 ± 0.6 | 55.84 ± 1.7 | 56.30 ± 1.1 | 78.64 ± 1.7 | 48.91 ± 1.1 |
| PromptCCD $w$/GMP (Ours) | **61.15** ± 1.0 | 66.29 ± 2.0 | 58.83 ± 1.0 | **56.65** ± 1.0 | 79.88 ± 2.5 | 48.96 ± 0.8 |

# Continual Category Discovery: Experiments

## CCD benchmark results

- Comparison with other representative CCD and GCD methods for CCD task.
- Experiment on different datasets (generic & finegrained) and foundation (pretrained) models.
- Evaluate using *cACC* metric.
- Overall, our PromptCCD w/ GMP outperforms other approaches across all datasets.

| Method | Pretrained Model | CIFAR100 | | | ImageNet-100 | | | TinyImageNet | | | Caltech-101 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *All* | *Old* | *New* | *All* | *Old* | *New* | *All* | *Old* | *New* | *All* | *Old* | *New* |
| ORCA | DINO | 60.91 | 66.61 | 58.33 | 40.29 | 45.85 | 35.40 | 54.71 | 63.13 | 51.93 | 76.77 | 82.80 | 73.20 |
| GCD | DINO | 58.18 | 72.27 | 52.83 | 69.41 | 81.56 | 65.65 | 55.20 | 65.87 | 51.61 | 78.27 | 86.60 | 72.92 |
| SimGCD | DINO | 25.56 | 38.76 | 20.43 | 31.38 | 40.47 | 27.44 | 33.40 | 29.11 | 34.74 | 33.65 | 37.53 | 31.62 |
| GCD *w*/replay | DINO | 49.93 | 73.15 | 41.47 | 72.04 | 83.75 | 69.01 | 56.33 | 67.54 | 52.60 | 76.51 | 86.14 | 72.48 |
| SimGCD *w*/replay | DINO | 40.13 | 66.72 | 30.91 | 47.53 | 67.86 | 39.18 | 37.45 | 58.15 | 30.36 | 49.38 | 52.72 | 47.99 |
| Grow & Merge | DINO | 57.43 | 63.68 | 55.31 | 67.84 | 75.10 | 66.60 | 52.14 | 59.68 | 49.96 | 75.75 | 83.66 | 71.59 |
| MetaGCD | DINO | 55.49 | 69.38 | 48.98 | 66.41 | 80.54 | 60.65 | 55.26 | 66.12 | 50.79 | 80.75 | 89.02 | 75.86 |
| PA-CGCD | DINO | 58.25 | 87.11 | 49.04 | 64.79 | 91.15 | 57.83 | 51.13 | 74.95 | 43.52 | 77.96 | 94.75 | 69.66 |
| PromptCCD *w*/GMP (Ours) | DINO | **64.17** | 75.57 | 60.34 | **76.16** | 81.76 | 74.35 | **61.84** | 66.54 | 60.26 | **82.44** | 89.08 | 79.72 |
| GCD | DINOv2 | 65.35 | 77.06 | 60.46 | 71.58 | 83.02 | 68.05 | 59.05 | 77.44 | 53.41 | 83.00 | 88.65 | 79.80 |
| MetaGCD | DINOv2 | 52.10 | 79.64 | 43.13 | 70.20 | 82.62 | 64.66 | 56.15 | 74.69 | 49.37 | 83.05 | 88.08 | 80.89 |
| PA-CGCD | DINOv2 | 54.36 | 79.19 | 45.65 | 74.82 | 88.20 | 72.02 | 52.10 | 68.07 | 46.32 | 83.06 | 94.07 | 77.55 |
| PromptCCD *w*/GMP (Ours) | DINOv2 | **69.73** | 78.01 | 66.16 | **76.28** | 82.61 | 74.53 | **68.20** | 75.56 | 65.23 | **83.86** | 87.93 | 81.42 |

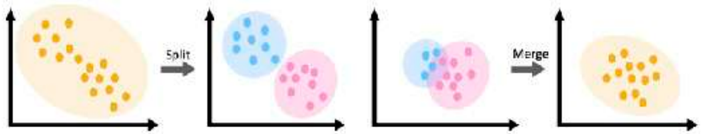
| Method | Pretrained Model | Aircraft | | | Stanford Cars | | | CUB | | | **Avg. results** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *All* | *Old* | *New* | *All* | *Old* | *New* | *All* | *Old* | *New* | *All* | *Old* | *New* |
| ORCA | DINO | 30.77 | 25.71 | 32.44 | 20.79 | 33.40 | 17.60 | 41.73 | 66.19 | 34.14 | 46.57 | 54.81 | 43.29 |
| GCD | DINO | 47.37 | 61.43 | 42.53 | 39.21 | 58.29 | 33.45 | 54.98 | 75.47 | 48.15 | 57.52 | 71.64 | 52.45 |
| SimGCD | DINO | 29.03 | 35.72 | 25.61 | 21.01 | 40.93 | 16.48 | 39.89 | 59.25 | 33.75 | 30.56 | 40.25 | 27.15 |
| GCD *w*/replay | DINO | 45.63 | 62.38 | 39.89 | 39.87 | 58.18 | 33.89 | 54.66 | 74.64 | 47.81 | 56.42 | 72.25 | 51.02 |
| SimGCD *w*/replay | DINO | 37.44 | 61.43 | 28.96 | 22.76 | 49.04 | 16.65 | 42.08 | 72.65 | 31.92 | 39.54 | 61.22 | 32.28 |
| Grow & Merge | DINO | 31.06 | 33.33 | 30.78 | 21.90 | 35.29 | 18.17 | 38.87 | 65.00 | 30.29 | 49.28 | 59.39 | 46.10 |
| MetaGCD | DINO | 44.63 | 59.05 | 39.39 | 35.98 | 56.97 | 29.96 | 44.59 | 74.40 | 35.40 | 54.73 | 70.78 | 48.72 |
| PA-CGCD | DINO | 48.24 | 73.09 | 40.60 | 43.88 | 80.43 | 33.54 | 52.48 | 77.26 | 44.74 | 56.68 | 82.68 | 48.42 |
| PromptCCD *w*/GMP (Ours) | DINO | **52.64** | 60.48 | 50.23 | **44.07** | 66.36 | 36.83 | **55.45** | 75.48 | 48.56 | **62.40** | 73.61 | 58.62 |
| GCD | DINOv2 | 57.87 | 63.80 | 55.39 | 58.52 | 71.65 | 53.80 | 66.70 | 83.33 | 60.81 | 66.01 | 77.85 | 61.67 |
| MetaGCD | DINOv2 | 54.90 | 64.29 | 52.08 | 57.16 | 71.87 | 52.01 | 62.19 | 82.50 | 55.13 | 62.25 | 77.67 | 56.75 |
| PA-CGCD | DINOv2 | 58.15 | 77.62 | 51.08 | 64.91 | 89.64 | 57.84 | 66.88 | 92.62 | 58.48 | 64.90 | 84.20 | 58.42 |
| PromptCCD *w*/GMP (Ours) | DINOv2 | **62.71** | 68.33 | 60.82 | **65.08** | 76.60 | 60.75 | **67.81** | 81.55 | 62.81 | **70.52** | 78.66 | 67.39 |

# Continual Category Discovery: Experiments

*CCD benchmark results*
*(when category number is unknown)*

| | Est. method | CIFAR100 | | | ImageNet-100 | | | TinyImageNet | | | CUB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category discovery at stage --→ | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Estimated category $C$ | GPC | 85 | 100 | 115 | 83 | 98 | 113 | 155 | 170 | 185 | 161 | 180 | 198 |
| Ground truth category $C$ | – | 80 | 90 | 100 | 80 | 90 | 100 | 160 | 180 | 200 | 160 | 180 | 200 |
| Methods | | *All* | *Old* | *New* | *All* | *Old* | *New* | *All* | *Old* | *New* | *All* | *Old* | *New* |
| GCD | GPC | 53.78 | 74.05 | 46.37 | 68.55 | 82.05 | 63.96 | 55.28 | 65.04 | 52.15 | 50.69 | 72.43 | 43.16 |
| Grow & Merge | GPC | 53.33 | 66.64 | 49.61 | 66.40 | 74.52 | 64.01 | 52.40 | 57.87 | 51.00 | 38.12 | 62.21 | 30.00 |
| MetaGCD | GPC | 47.55 | 70.79 | 38.57 | 63.48 | 80.82 | 56.28 | 56.21 | 68.33 | 50.99 | 44.30 | 70.69 | 35.83 |
| PA-CGCD | GPC | 55.66 | 90.21 | 44.99 | 66.74 | 91.28 | 58.97 | 50.55 | 72.44 | 43.38 | 52.27 | 76.38 | 44.24 |
| PromptCCD-U $w$/GMP (Ours) | GPC | **59.12** | 77.62 | 53.70 | **70.12** | 81.84 | 66.12 | **57.76** | 64.57 | 55.37 | **55.20** | 73.19 | 48.82 |

- When the class number in the unlabelled stream of data is unknown, we introduce PromptCCD-U, by incorporating a GMM-based ***on-the-fly*** category estimation method in our GMP module.




- As our GMP module is based on Gaussian Mixture Model (GMM), we incorporate the GMM-based category number estimation method in our framework. Specifically, the idea is to automatically **splitting** and **merging** the GMM's clusters during learning by assessing the cluster's compactness and separability using MCMC algorithm.

- For the sake comparison, we also compare our method with others by directly using the estimated category number from our model.

# Continual Category Discovery: Experiments

| PromptCCD | CIFAR100 Avg. *ACC* | | | ImageNet-100 Avg. *ACC* | | |
|---|---|---|---|---|---|---|
| | *All* | *Old* | *New* | *All* | *Old* | *New* |
| *w/o* GMP | 58.18 | 72.27 | 52.83 | 69.41 | 81.56 | 65.65 |
| GMP (random-k) | 59.98 | 73.81 $^{\uparrow 1.54}$ | 55.68 $^{\uparrow 2.85}$ | 68.30 | 80.09 $^{\downarrow 1.47}$ | 63.50 $^{\downarrow 2.15}$ |
| GMP (top-k) (Ours) | **64.17** | 75.57 $^{\uparrow 3.30}$ | 60.34 $^{\uparrow 7.51}$ | **76.16** | 81.76 $^{\uparrow 0.20}$ | 74.35 $^{\uparrow 8.70}$ |
| PromptCCD | TinyImageNet Avg. *ACC* | | | CUB Avg. *ACC* | | |
| | *All* | *Old* | *New* | *All* | *Old* | *New* |
| *w/o* GMP | 55.20 | 65.87 | 51.61 | 54.98 | 75.47 | 48.15 |
| GMP (random-k) | 55.69 | 63.95 $^{\downarrow 1.92}$ | 52.52 $^{\uparrow 0.91}$ | 51.46 | 73.10 $^{\downarrow 2.37}$ | 43.90 $^{\downarrow 4.25}$ |
| GMP (top-k) (Ours) | **61.84** | 66.54 $^{\uparrow 0.67}$ | 60.26 $^{\uparrow 8.65}$ | **55.45** | 75.48 $^{\uparrow 0.01}$ | 48.56 $^{\downarrow 0.41}$ |

| top-k Prompts | GMM Samples | C100 Avg. *ACC* | | | CUB Avg. *ACC* | | |
|---|---|---|---|---|---|---|---|
| | | *All* | *Old* | *New* | *All* | *Old* | *New* |
| 0 | 0 | 58.18 | 72.27 | 52.83 | 54.98 | 75.47 | 48.15 |
| 5 | 0 | 61.48 | 74.68 | 57.55 | 53.54 | 74.28 | 46.47 |
| 5 | 20 | 62.21 | 75.71 | 57.90 | 54.37 | 74.88 | 46.41 |
| 5 | 200 | 61.00 | 72.46 | 57.08 | 51.67 | 73.33 | 44.08 |
| 2 | 100 | 61.39 | 73.04 | 57.64 | 53.36 | 73.45 | 46.04 |
| <u>5</u> | <u>100</u> | **64.17** | 75.57 | 60.34 | **55.45** | 75.48 | 48.56 |
| 10 | 100 | 61.03 | 72.91 | 56.97 | 52.76 | 71.67 | 46.02 |

| Method | Prompt | *All* | *Old* | *New* |
|---|---|---|---|---|
| G&M | *w/o* GMP | 57.43 | 63.68 | 55.31 |
| G&M | *w/*GMP | **61.14** | **64.94** | **59.10** |
| MetaGCD | *w/o* GMP | 55.49 | 69.38 | 48.98 |
| MetaGCD | *w/*GMP | **58.99** | **69.69** | **54.29** |
| PromptCCD | *w/o* GMP | 58.18 | 72.27 | 52.83 |
| PromptCCD | *w/*GMP | **64.17** | **75.57** | **60.34** |

## Model analysis

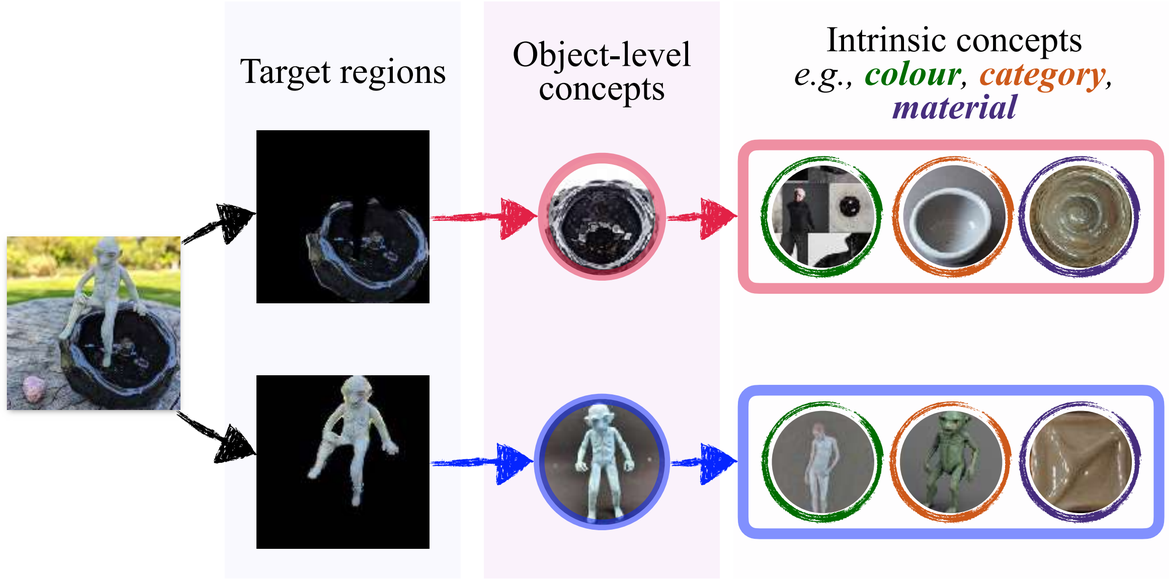
- ***Left table:*** to validate the effectiveness of using top-k prompts.
- ***Middle table:*** The choices of number for top-k prompts and GMM samples
- ***Right table:*** Improving other CCD methods with our GMP.

# Continual Category Discovery: Conclusion

- **PromptCCD:** A novel prompt learning framework for CCD, adapting vision foundation models to address CCD.
- **Gaussian Mixture Prompting (GMP):** A specialized prompting module for CCD utilizing learnable GMM as a pool of prompts. Additionally, GMP enables on-the-fly category estimation making it well suited for handling CCD task.
- **Performance:** PromptCCD achieves state of the art results in CCD benchmark.

# Task C: Intrinsic Concept Extraction

# Intrinsic Concept Extraction (ICE): Intro



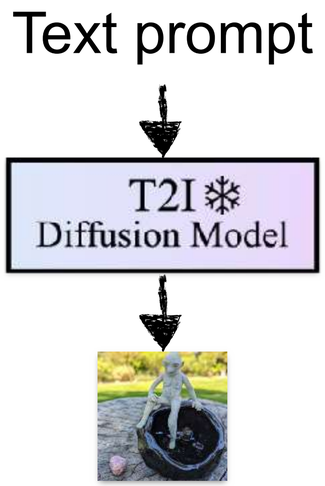

Given a single image, **ICE** aims at extracting object-level concepts and the underlying intrinsic attributes such as:

***Semantic category*, *colour*, and *material***

# Intrinsic Concept Extraction: Intro

## Harnessing Diffusion Models for ICE

Thus, instead of generating images from text prompt (left image), we would like leverage a T2I diffusion model exclusively for ICE (right image).

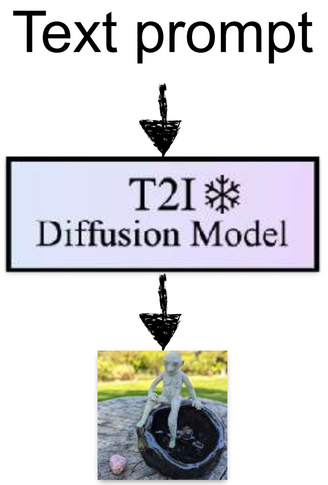



(a) Vanilla T2I generation

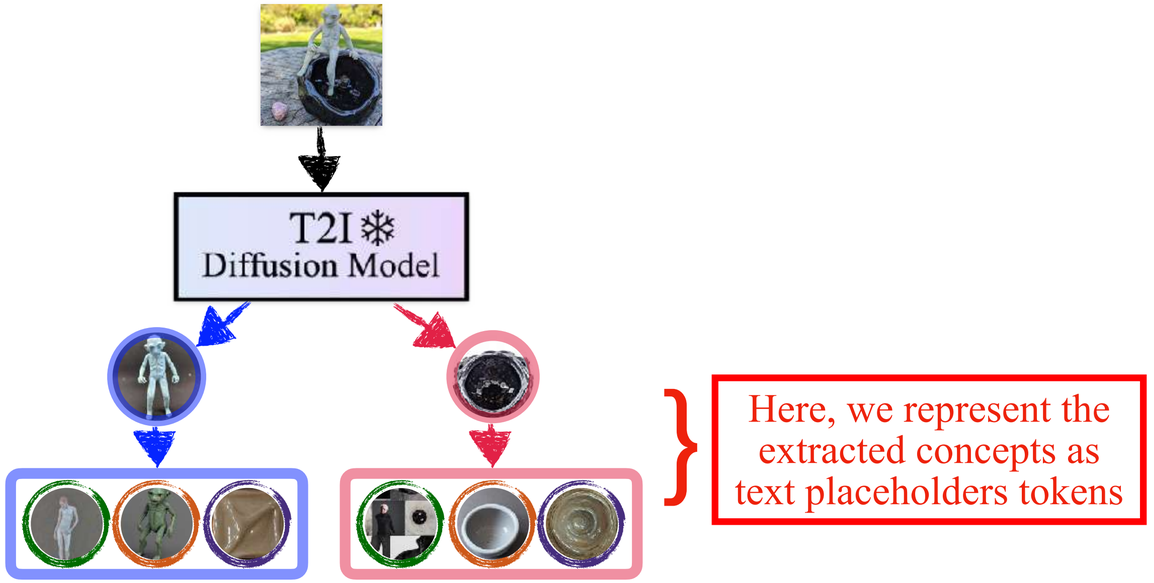
# Intrinsic Concept Extraction: Intro

## Harnessing Diffusion Models for ICE

Thus, instead of generating images from text prompt (left image), we would like leverage a T2I diffusion model exclusively for ICE (right image).




(a) Vanilla T2I generation




(b) ICE

# Intrinsic Concept Extraction: Intro

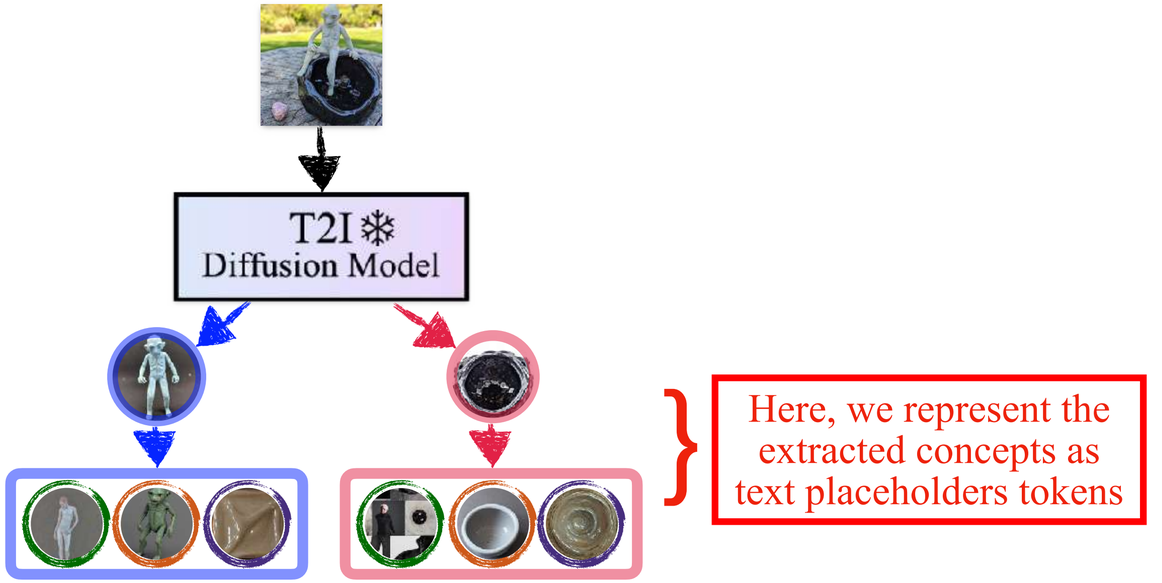
## Harnessing Diffusion Models for ICE

ICE task is under the umbrella of the image personalization task. But it differs on:

(1) handling multiple objects, (2) concepts coverage, and (3) fully unsupervised

Table 1. Comparison of ICE and relevant works.

| Method | Learned concept(s) | | | Framework details | |
|---|---|---|---|---|---|
| | Object-level concepts | Intrinsic concepts | Multi concepts | Single image | Extra information |
| Textual Inversion [9] | ✓ | ✗ | ✗ | ✗ | - |
| Dreambooth [28] | ✓ | ✗ | ✗ | ✗ | - |
| Inspiration Tree [34] | ✗ | ✓ | ✗ | ✗ | - |
| LangInt [17] | ✓ | ✓ | ✗ | ✓ | VQA-guided |
| Break-A-Scene [1] | ✓ | ✗ | ✓ | ✓ | Mask |
| MCPL [15] | ✓ | ✗ | ✓ | ✓ | Text |
| ConceptExpress [11] | ✓ | ✗ | ✓ | ✓ | - |
| **ICE (Ours)** | ✓ | ✓ | ✓ | ✓ | - |




(b) ICE

# Intrinsic Concept Extraction: Intro

## Why concept extraction?

ICE provides a detailed and interpretable representation of visual elements in an image, allowing for versatile generative applications

Extracted intrinsic concepts
*e.g.,* ***colour***, ***category***, ***material***

# Intrinsic Concept Extraction: Intro

## Why concept extraction?

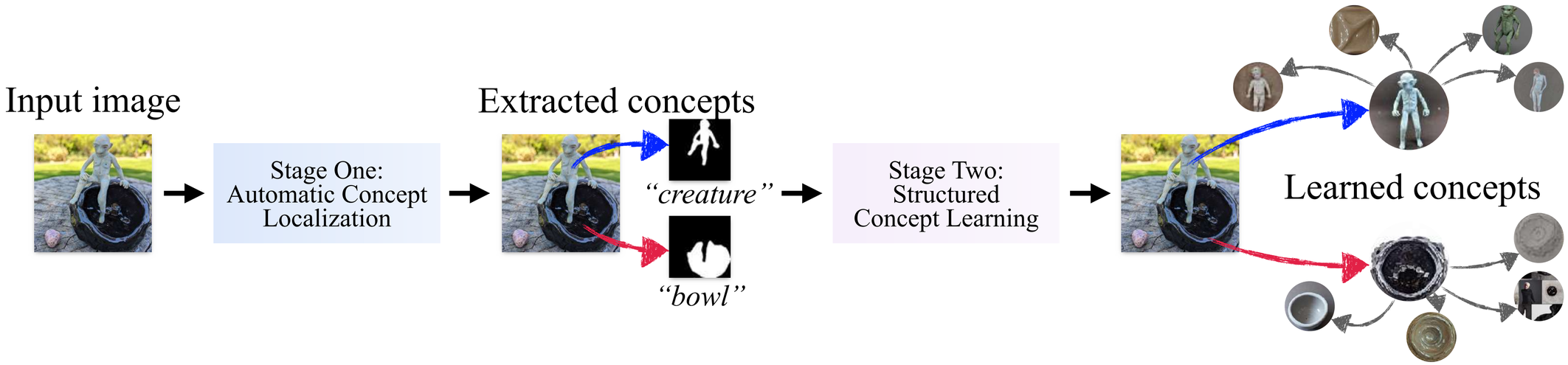
ICE provides a detailed and interpretable representation of visual elements in an image, allowing for versatile generative applications

# Intrinsic Concept Extraction: Method

## ICE framework

The proposed ICE framework operates through a two-stage architecture:

- **Stage One:** Automatic Concept Localization via frozen T2I Diffusion
- **Stage Two:** Structured Concept Learning via T2I Diffusion finetuning



# Intrinsic Concept Extraction: Method

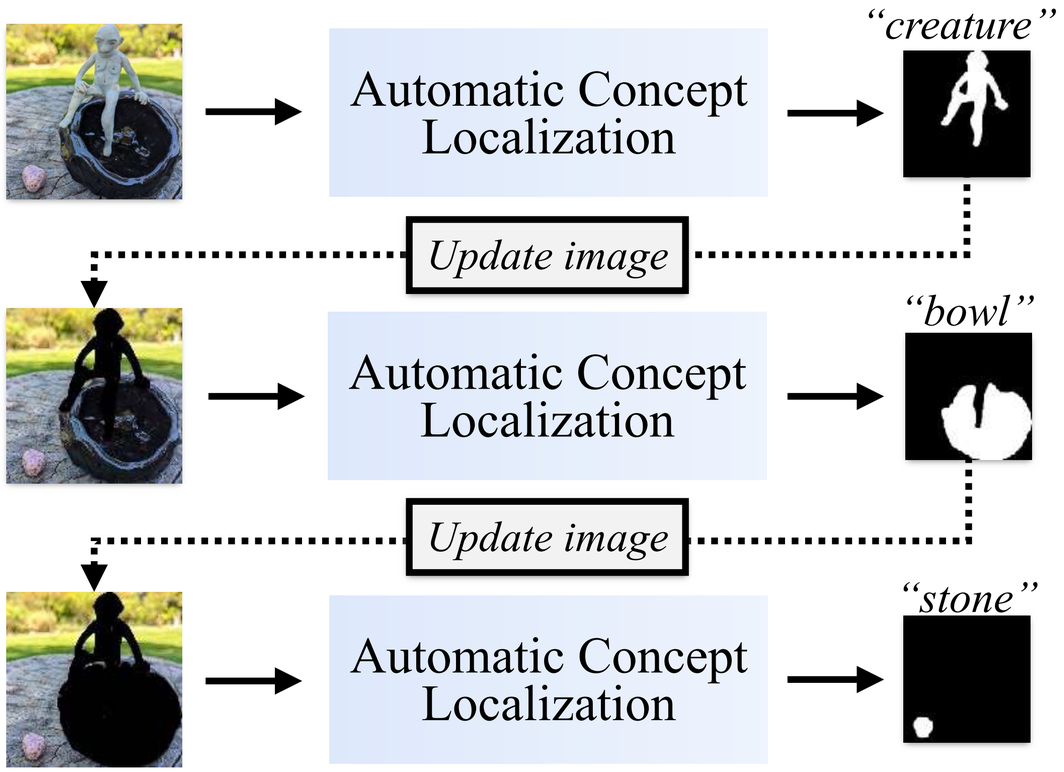
## **Stage One:** Automatic Concept Localization

**Input:** an image

**Output:**

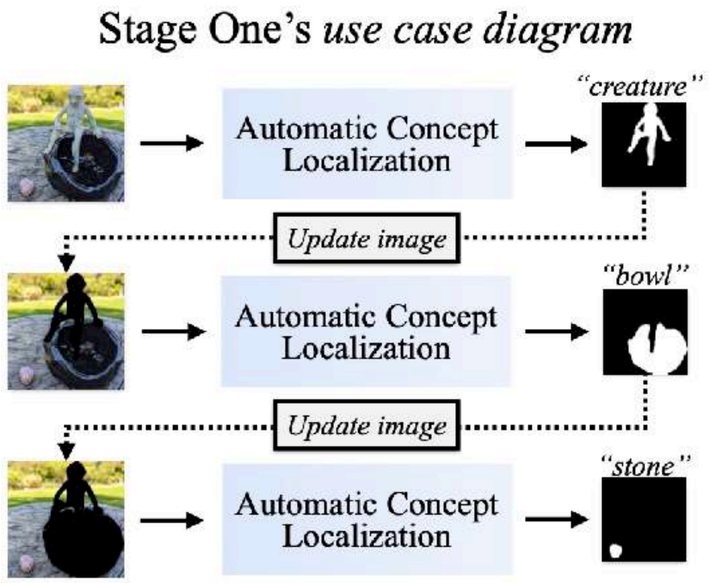
- Object masks
- Object categories

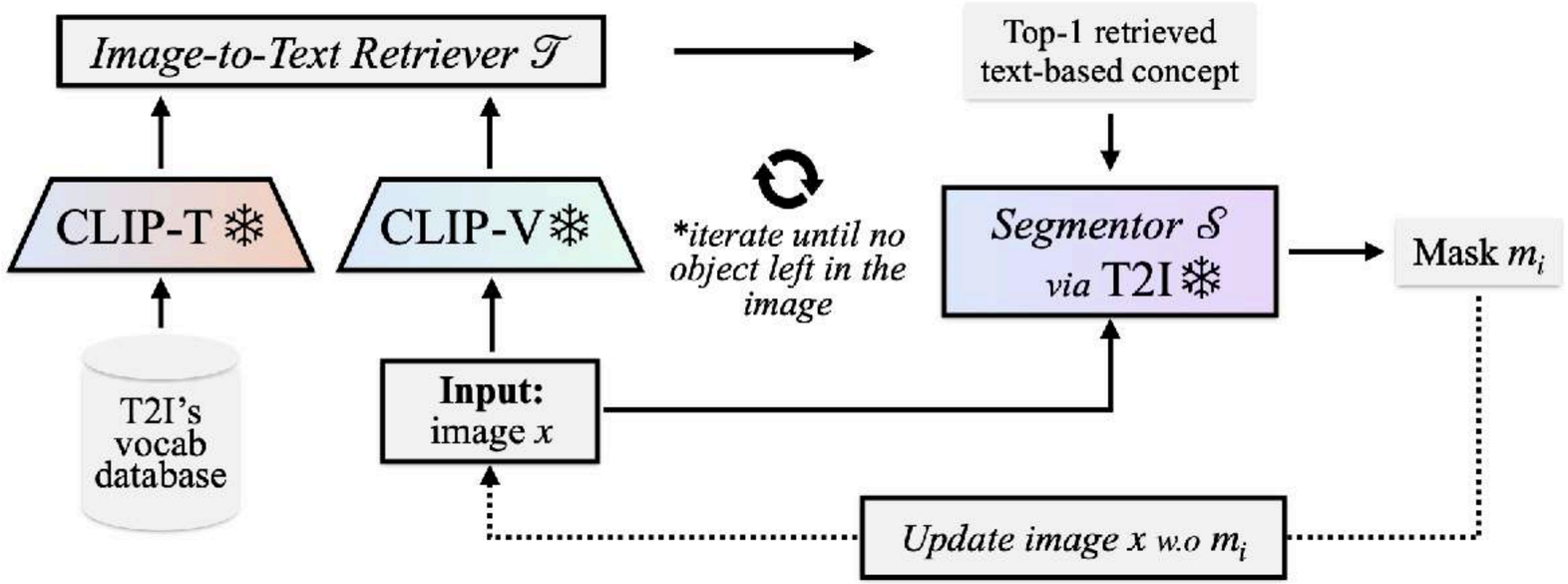
Stage One's *use case diagram*



# Intrinsic Concept Extraction: Method

## **Stage One:** Automatic Concept Localization




**Procedure:**

*Step 1:* Retrieve top-1 associated text based on input.

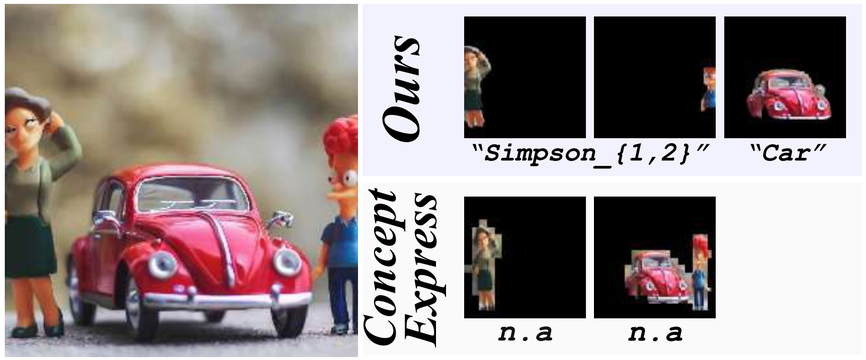
*Step 2:* Extract mask given top-1 text and input image via T2I model.

*Step 3:* Update input image by removing the masked region.

*Step 4:* Back to *Step 1*, until no "object" regions are left.

# Intrinsic Concept Extraction: Method

**Example:** Stage One outputs



*Ours*



"Simpson_{1,2}" "Car"

*Concept Express*



n.a n.a

*Ours*

"Fetch"

*Concept Express*

n.a n.a n.a

*Ours*

"Amphi-theater" "Buddha"

*Concept Express*

n.a n.a

*Ours*

"Fountains" "Figurine_{1,2,3}"

*Concept Express*

n.a n.a

*ConceptExpress (Hao et. al., ECCV'24 Oral)

# Intrinsic Concept Extraction: Method

## **Stage Two:** Structured Concept Learning

Now, given these extracted object masks and text concepts, we know like to learn both **object-level** & **intrinsic** concepts.

To do this, we divide this stage into two phases:

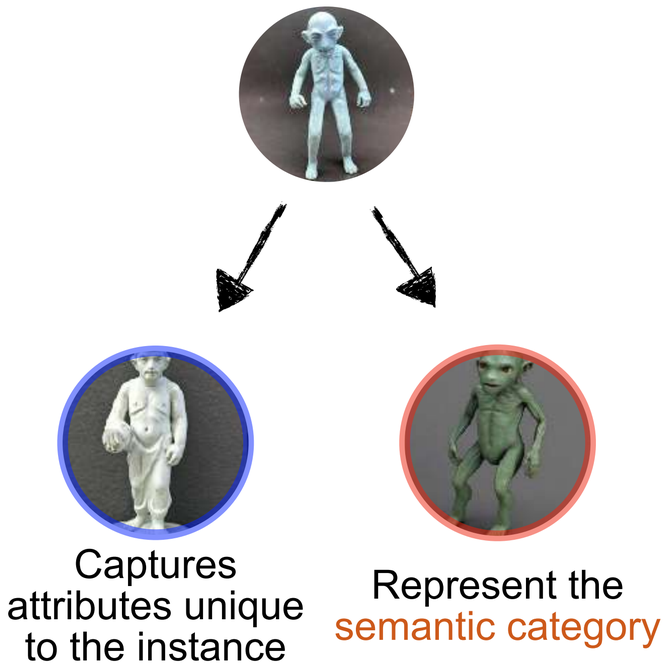
# Intrinsic Concept Extraction: Method

## **Stage Two:** Structured Concept Learning

Now, given these extracted object masks and text concepts, we would like to learn both **object-level** & **intrinsic** concepts.

To do this, we divide this stage into two phases:

*Phase one*: Learning object-level concepts




Here, we would like to learn to disentangle

Object level concept  to:

 Instance specific &  object specific

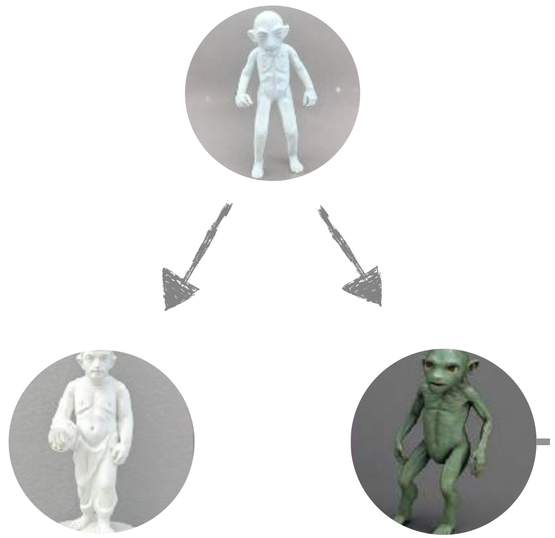
# Intrinsic Concept Extraction: Method

## **Stage Two:** Structured Concept Learning

Now, given these extracted object masks and text concepts, we would like to learn both **object-level** & **intrinsic** concepts.

To do this, we divide this stage into two phases:

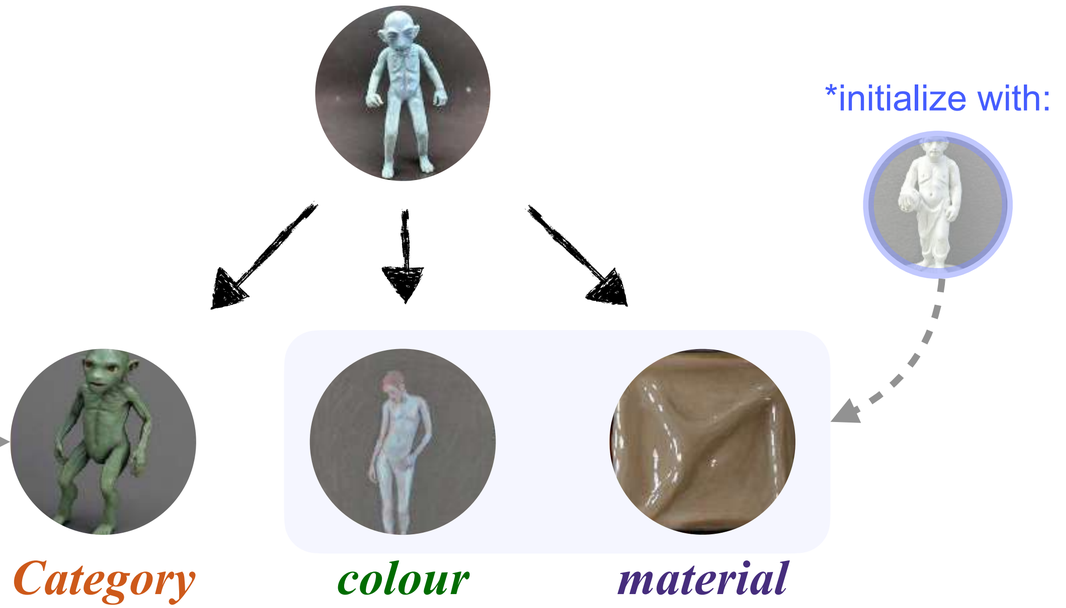
*Phase one*: Learning object-level concepts



Captures attributes unique to the instance

Represent the semantic category

*Phase two*: Learning intrinsic concepts

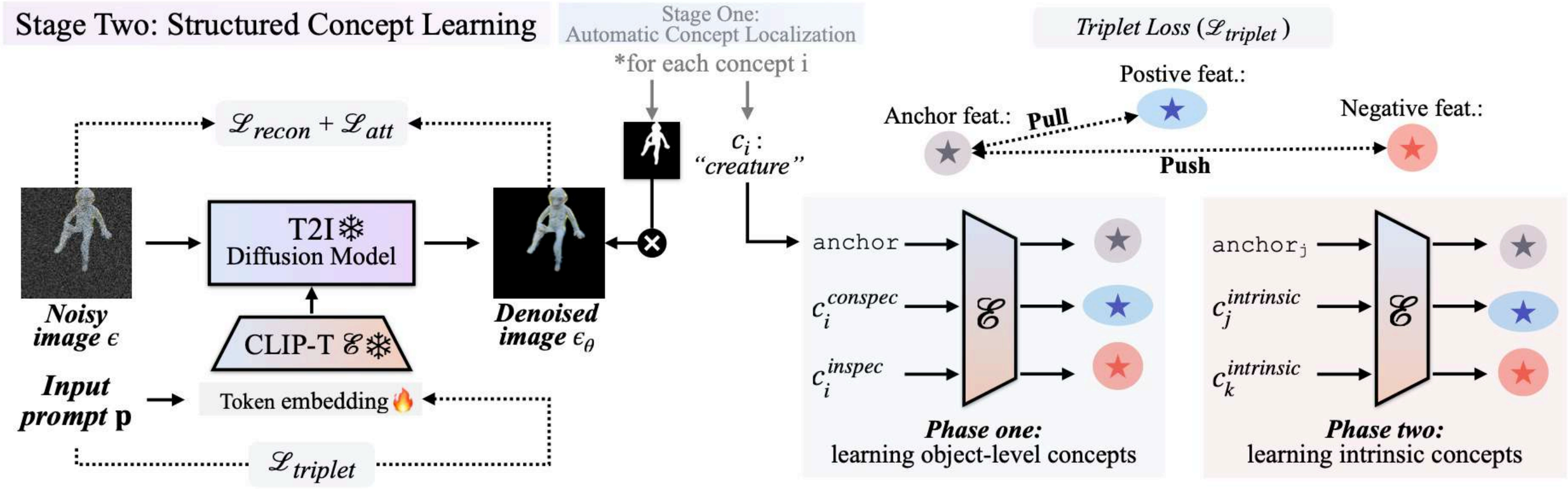



Represent the intrinsic concepts

# Intrinsic Concept Extraction: Method

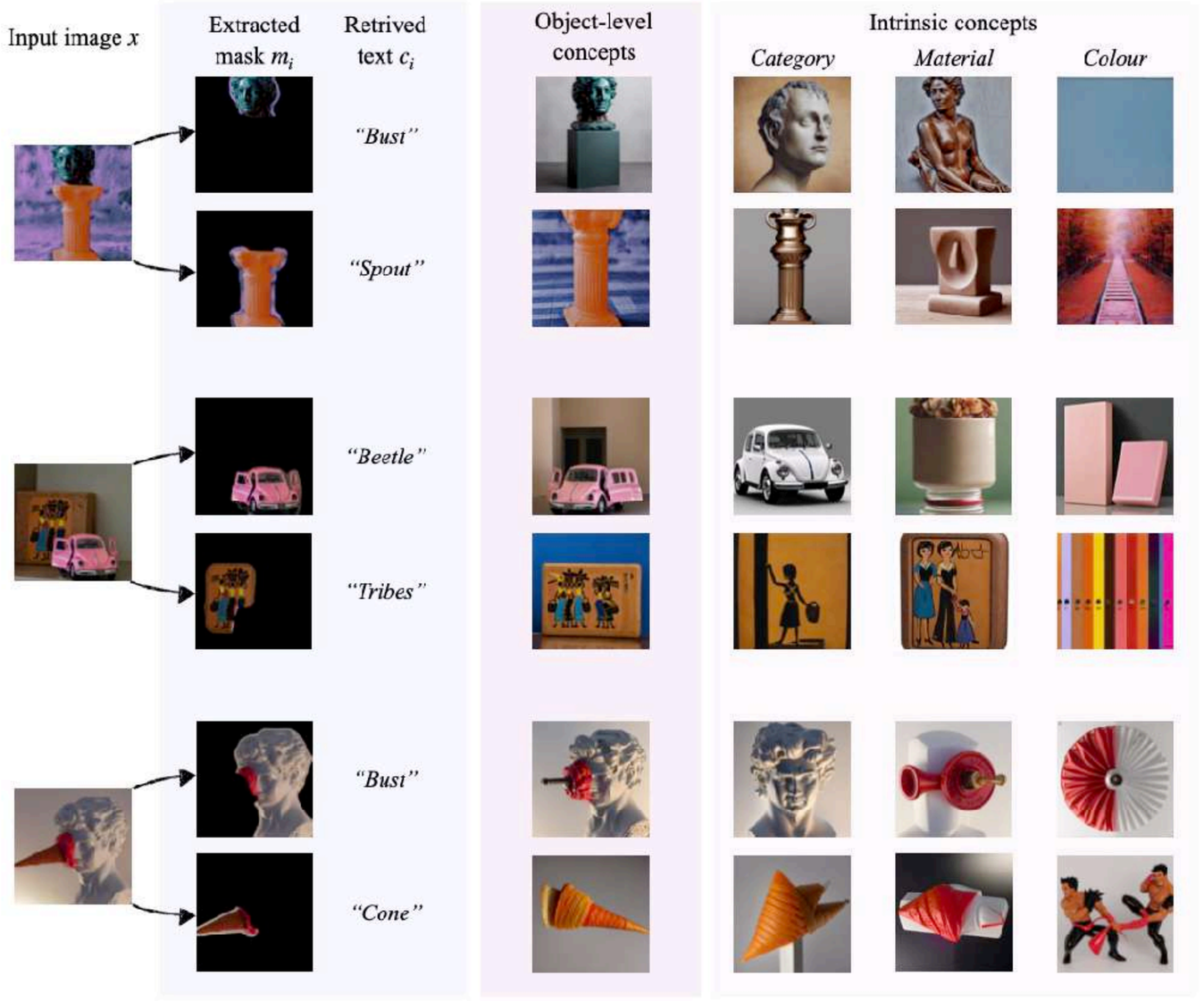
## **Stage Two:** Structured Concept Learning

Combine it with diffusion textual-inversion finetuning approach…



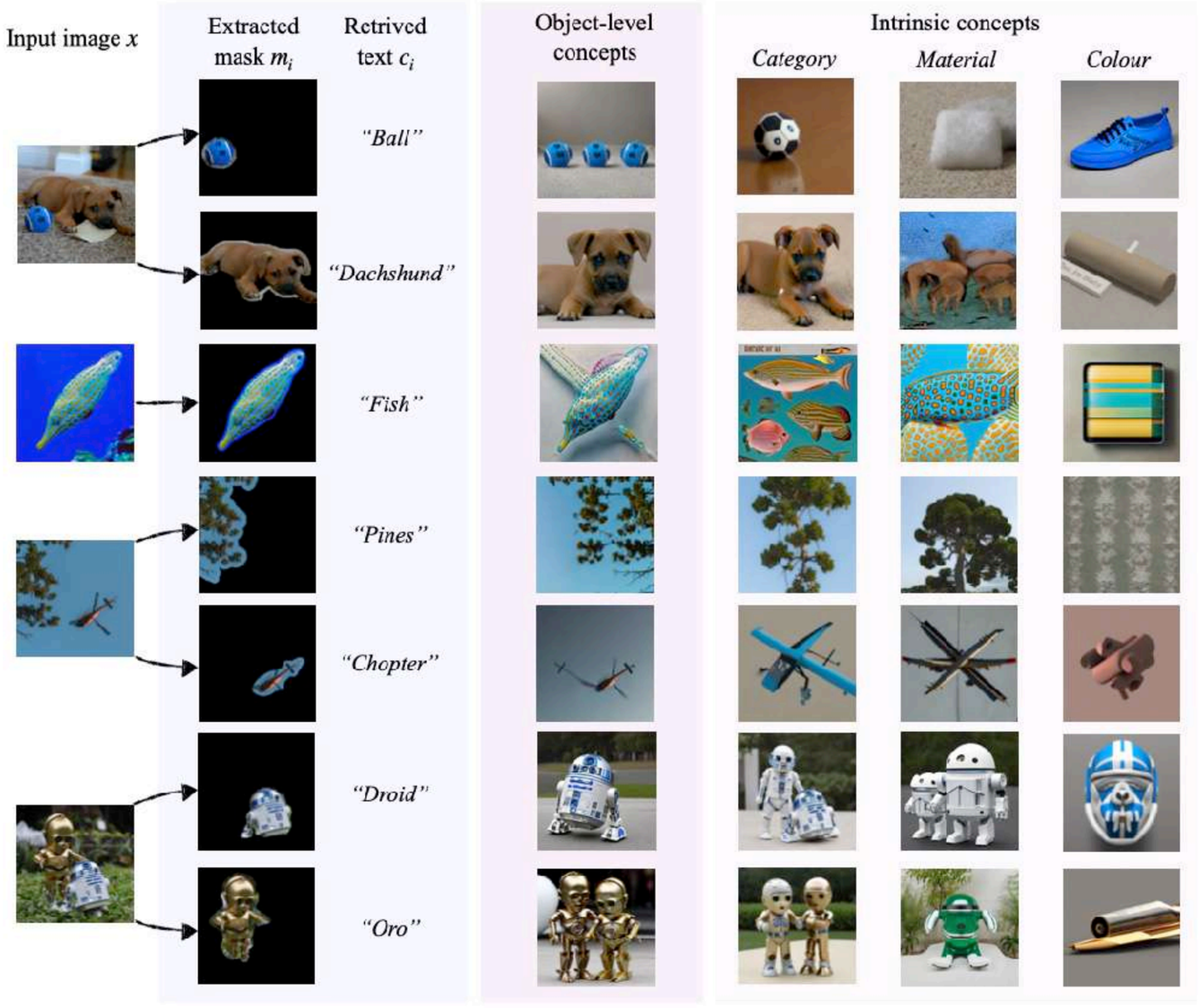
# Intrinsic Concept Extraction: Results

## ICE qualitative results (1/2)



# Intrinsic Concept Extraction: Results

## ICE qualitative results (2/2)



# Intrinsic Concept Extraction: Results

## ICE quantitative results

Table 2. Performance of ICE and relevant works on Unsupervised Concept Extraction (UCE) benchmarks using CLIP [25] encoder.

| Method | $SIM^I$ (%) | $SIM^C$ (%) | $ACC^1$ (%) | $ACC^3$ (%) |
|---|---|---|---|---|
| Break-A-Scene [1] | 0.627 | 0.773 | 0.174 | 0.282 |
| ConceptExpress[11] | 0.689 | 0.784 | 0.263 | 0.385 |
| **ICE (Ours)** | **0.738** | **0.822** | **0.325** | **0.518** |

Table 3. Performance of ICE and relevant works on Unsupervised Concept Extraction (UCE) benchmarks using DINO [7] encoder.

| Method | $SIM^I$ (%) | $SIM^C$ (%) | $ACC^1$ (%) | $ACC^3$ (%) |
|---|---|---|---|---|
| Break-A-Scene [1] | 0.254 | 0.510 | 0.202 | 0.315 |
| ConceptExpress [11] | 0.319 | 0.568 | 0.324 | 0.470 |
| **ICE (Ours)** | **0.677** | **0.755** | **0.476** | **0.638** |

## ICE ablation study

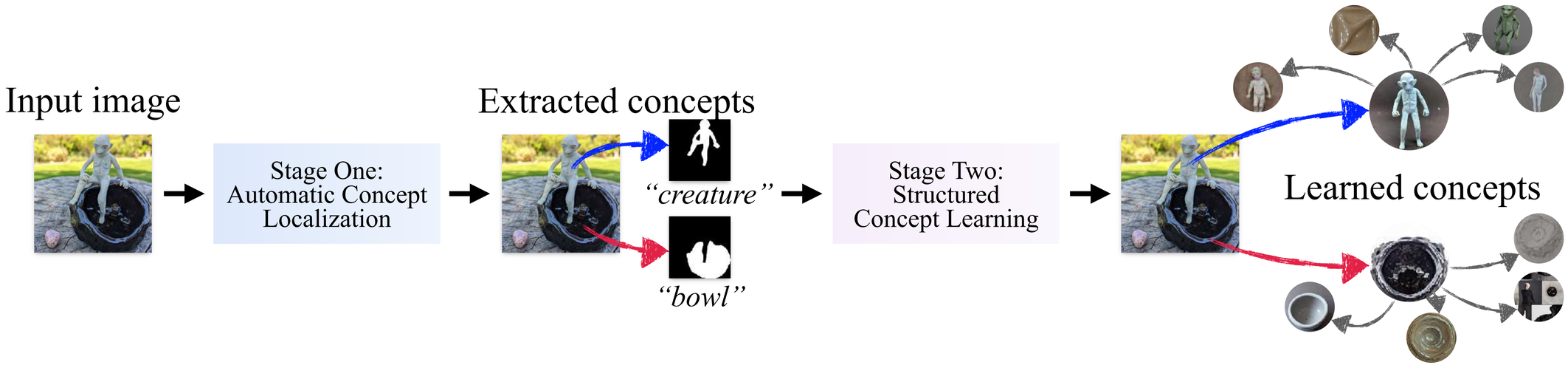
Table 4. Ablation study on **ICE** model components on the *D1* dataset.

| | Model's components | | | CLIP [25] encoder | | | | DINO [7] encoder | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Stage One's mask | Stage Two's token init. | Stage Two learning | $SIM^I$ (%) | $SIM^C$ (%) | $ACC^1$ (%) | $ACC^3$ (%) | $SIM^I$ (%) | $SIM^C$ (%) | $ACC^1$ (%) | $ACC^3$ (%) |
| ConceptExpress | ✗ | ✗ | ✗ | 0.689 | 0.784 | 0.263 | 0.385 | 0.319 | 0.568 | 0.324 | 0.470 |
| ICE *w.* mask | ✓ | ✗ | ✗ | 0.710 | 0.781 | 0.307 | 0.456 | 0.493 | 0.601 | 0.395 | 0.604 |
| ICE *w.o* Stage Two | ✓ | ✓ | ✗ | 0.726 | 0.807 | 0.301 | 0.452 | 0.501 | 0.621 | 0.411 | 0.604 |
| ICE *w.o* text init. | ✓ | ✗ | ✓ | 0.722 | 0.814 | 0.320 | 0.475 | 0.548 | 0.643 | 0.449 | 0.627 |
| **ICE (Ours)** | ✓ | ✓ | ✓ | **0.738** | **0.822** | **0.325** | **0.518** | **0.677** | **0.755** | **0.476** | **0.638** |

# Intrinsic Concept Extraction: Conclusion

**In this work:**

1. We introduce a novel task called Intrinsic Concept Extraction from a single image.
2. We propose ICE framework to automatically and systematically extract concepts.



# Summary & List of publications

(A) *PartCo: Part-Level Correspondence Priors Enhance Category Discovery*
Fernando Julio Cendra, Kai Han
*In submission, arXiv:2509.22769*

(B) *PromptCCD: Learning Gaussian Mixture Prompt Pool for Continual Category Discovery*
Fernando Julio Cendra, Bingchen Zhao, Kai Han
Published: *ECCV'24*


(C) *ICE: Intrinsic Concept Extraction from a Single Image via Diffusion Models*
Fernando Julio Cendra, Kai Han
Published: *CVPR'25 Highlight, Top 3% submissions*

# Limitation & Future works

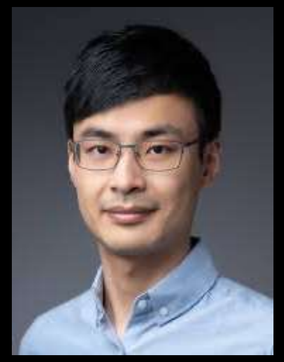
There is still a lot to be done to understand our open-world:

1. How to effectively process dynamic information from the physical world
2. To find / model dynamic representation that can mimic our open-world
3. Make use of these dynamic representations e.g, by learning motion and interaction (human, objects & scenes) to better understand our world

# — Thank you —

Prof. Kai Han

Prof. Guanbin Li



Prof. Lequan Yu

Prof. Binzheng Zhang

# Q&A