

APPENDIX

Anonymous authors

Paper under double-blind review

1 DETAILS OF DATASETS

The table 1 shows the statistical details of the datasets we choose.

Dataset	#Class	#Train	#Valid	#Test
ImageNet ¹	1000	128,1167	50,000	-
Food-101	101	60,600	15,150	25,250
CIFAR-100	100	40,000	10,000	10,000
CIFAR-10	10	40,000	10,000	10,000
CUB-200	200	4,794	1,200	5,794
Flowers	102	4,093	1,633	2,463

Table 1: Statistical details of datasets. “#Class” means the number of classifications. “#Train”, “#Valid”, and “#Test” denote the instance numbers of each dataset respectively.

2 CHOICE OF DIFFERENT PROMPTS

In this section, we discuss the choice of different prompts in concept discovery and compare the performance of concepts discovered with different prompts on the CUB dataset. From the result we can observe that different prompts provide similar performance, which is because the large language model is not sensitive to the prompts and give similar concepts.

3 PERFORMANCE OF DIFFERENT SIMILARITY THRESHOLD

In this section, we show the performance of our CDL model with different similarity threshold on the CUB dataset. From the results we can observe that the threshold of 0.9 can achieve the best performance.

4 UNSUPERVISED CLASSIFICATION RESULT WITH DIFFERENT BACKBONES

In this section we compare the unsupervised classification result of our fine-tuned CLIP and previous method (VDES) on different backbones. The comparison with “ViT-L/14” backbone is shown in Sec 4.1. Here we show the comparison with “ViT-B/32” backbone.

¹For ImageNet, we randomly sample 50 images per class as the valid dataset and use the original valid dataset as the test dataset

Prompts\#Concepts	200	400
What are useful visual features for distinguishing a {category name} in a photo?	83.2	83.4
What visual features do you use to recognize a {category name} in a photo?	83.0	83.3
What are the identifying features of a {category name} in a photo?	82.9	83.3

Table 2: Classification Performance of concepts generated by different prompts on the CUB dataset.

#Concepts	200	400
Threshold = 0.8	81.8	82.5
Threshold = 0.85	82.3	82.7
Threshold = 0.9	83.2	83.4
Threshold = 0.95	82.9	83.1

Table 3: The performance of CDLwith different threshold on the CUB dataset.

	ImageNet	Food-101	CIFAR-100	CIFAR-10	CUB-200	Flowers-102
CLIP + Name	58.5	79.3	63.5	89.0	52.0	65.9
CLIP + Name w/ Concept	63.0	83.6	64.7	90.3	52.6	66.1
CLIP + Concept	16.2	2.5	22.8	59.4	3.2	4.6
CLIP + Name w/ Random Concept	61.2	80.4	63.3	90.1	52.6	66.3
CDL + Concept	62.7	82.0	65.2	90.7	53.9	67.4

Table 4: The unsupervised classification results of the original and our fine-tuned CLIP model with different prompts. “Name” corresponds to the simple prompt “A photo of a class name”. “Name w/ Concept” denotes the prompts in the previous work (Menon & Vondrick, 2022), which are like “A photo of a class name, which has “concept”. “Concept” corresponds to the pure concept. “Name w/ Random Concept” means that we replace the correct concept with random concepts. The large gap between “Name w/ Concept” and “Concept” and the small gap between “Name w/ Random Concept” and “Name w/ Concept” mean that the class names instead of the descriptive features in the prompts make the main contribution to the decision of the CLIP model. “CDL + Concept” means the prediction of our fine-tuned CLIP model with class-agnostic concepts.

5 HUMAN EVALUATION DETAILS

We hire workers on <https://www.mturk.com> to conduct human evaluation. In order to make sure the correctness of human annotation, for one data point we ask three human workers to annotate. For the factuality and groundability metric, we randomly sample 10 classes from each dataset and annotate the factuality and groundability of the top-3 concepts of each class. In order to calculate the factuality and groundability, we select 10 images for each concept to annotate. Therefore, we annotate 10,800 data points in total for those two task. For the visual discriminability and classname containing, we conduct annotation on selected 400 concepts of LaBo and our method on the CUB dataset. Hence we annotate 1,600 data points for those two task. We pay the human workers \$0.05 each data point. The total cost of human annotation is \$1,860. **In the annotation, we randomly shuffle the order of instances to remove possible biases.**

In order to validate the effectiveness of our human evaluation, we calculate the pairwise annotator agreement score following previous work Yang et al. (2023). The average pairwise annotator agreement proportion on all datasets is 69.2%, which is comparable with the 69.8% proportion in the previous work.

We conduct Students’ T-test to evaluate the statistical significance of the human evaluation results. We set the threshold of p-value to be 0.05 following previous works. When p-value is lower than 0.05, the null hypothesis is rejected and our method performs significantly better than the baseline method. From the results we can observe that both our concept learning and concept discovery method significantly outperform the baseline methods regarding the intervention, factuality and groundability metrics.

We show some examples about the interface of our human annotation. In the annotation platform, the workers can see an image and is asked to select whether the given concept describes the image.

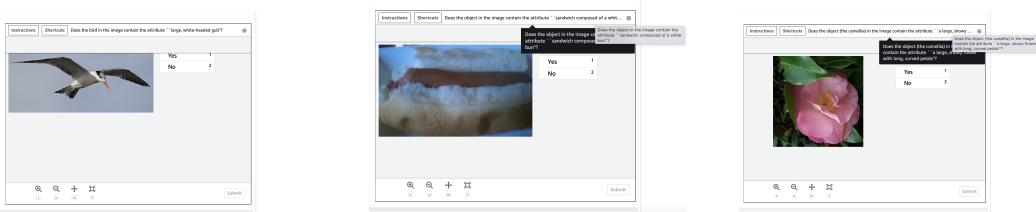


Figure 1: Examples of the annotation interface.

Dataset	Method	Intervention		Factuality		Groundability	
		p-value	significance	p-value	significance	p-value	significance
ImageNet	CLIP + CDL v.s. CLIP + LaBo	2.6e-22	✓	4.3e-29	✓	0.34	×
	CDL + CDL v.s. CLIP + CDL	5.6e-2	×	5.2e-4	✓	8.4e-2	×
Food-101	CLIP + CDL v.s. CLIP + LaBo	1.3e-106	✓	5.5e-12	✓	5.3e-3	✓
	CDL + CDL v.s. CLIP + CDL	8.5e-5	✓	3.4e-3	✓	0.80	×
CIFAR-100	CLIP + CDL v.s. CLIP + LaBo	9.2e-45	✓	1.1e-50	✓	1.8e-6	✓
	CDL + CDL v.s. CLIP + CDL	1.3e-3	✓	0.82	×	0.62	×
CIFAR-10	CLIP + CDL v.s. CLIP + LaBo	1.3e-2	×	0.14	×	7.0e-2	×
	CDL + CDL v.s. CLIP + CDL	2.7e-29	✓	7.8e-5	✓	5.6e-2	×
CUB-200	CLIP + CDL v.s. CLIP + LaBo	8.9e-20	✓	2.8e-15	✓	0.73	×
	CDL + CDL v.s. CLIP + CDL	2.0e-9	✓	1.9e-2	✓	7.2e-5	✓
Flowers-102	CLIP + CDL v.s. CLIP + LaBo	1.5e-39	✓	1.2e-23	✓	1.8e-5	✓
	CDL + CDL v.s. CLIP + CDL	4.2e-3	✓	2.4e-4	✓	0.25	×

Table 5: The statistical significance of the human evaluation results.

6 EXAMPLES OF MUTUAL INFORMATION BASED CONCEPT SELECTION

In this section we showcase some examples of the concepts selected by our Mutual Information based method. From the examples we can see that our method can effectively select visually discriminative concepts and exclude non-visual ones.

7 EXAMPLES OF CONCEPT-BASED MULTI-MODAL RECOGNITION

In this section we show some examples of different concept-based image classification methods. From the examples in Figure 2 we can observe that previous works suffer from class-conditional and non-visual concepts, while our method can learn interpretable concept-class map based on class-agnostic concepts.

Category	Concept Selected	Concept Excluded
Giant Panda	black patches around eyes large, round head black fur on ears	a rare animal popular in zoo
Black-footed Albatross	black and white a long, hooked bill long, narrow wings	found in North America dive to depths of over 30 meters
Grey Whale	long, curved mouth dark grey or black white patches on the skin	large marine mammal long-distance migration

Table 6: The examples of selected and excluded concepts by our Mutual Information based concept selection method




<p>great pyrenees dog</p> 	<p>VDES predicts great pyrenees dog with these concepts:</p> <p>Great pyrenees dog, which has bushy tail.</p> <p>Great pyrenees dog, which has thick coat of fur.</p> <p>Great pyrenees dog, which is a large, white, fluffy dog.</p> <p>Great pyrenees dog, which has big, round eyes.</p>	<p>After class names removal, VDES predicts cardigan welsh corgi with these concepts:</p> <p>large, pointy ears</p> <p>short-legged dog</p> <p>a large, white, fluffy dog</p> <p>thick fur</p>	<p>LaBo predicts great pyrenees dog with these concepts:</p> <p>wise and regal dog</p> <p>gentle giant, known for being calm and patient</p> <p>known for its thick fur</p> <p>comes from the pyrenees mountain range</p>	<p>CDL (Ours) predicts great pyrenees dog with these concepts:</p> <p>white and fluffy</p> <p>thick mane around the neck</p> <p>thick coat of fur</p> <p>white, cream, or biscuit-colored</p>
<p>paella</p> 	<p>VDES predicts paella with these concepts:</p> <p>Paella, which is made of metal or ceramic</p> <p>Paella, which has has rice, seafood, and vegetables</p> <p>Paella, which is a brightly-colored rice-based dish.</p>	<p>After class names removal, VDES predicts ceviche with these concepts:</p> <p>onions, peppers, and herbs</p> <p>garnished with avocado, lime, or cilantro</p> <p>brightly-colored rice-based dish</p>	<p>LaBo predicts paella with these concepts:</p> <p>made with either fresh or frozen seafood</p> <p>popular Spanish dish</p> <p>popular choice for large gatherings</p>	<p>CDL (Ours) predicts paella with these concepts:</p> <p>rice, seafood, vegetables, and meats</p> <p>cooked in broth</p> <p>bright-colored seafood rice served with lemon wedges</p>
<p>black-footed albatross</p> 	<p>VDES predicts black-footed albatross with these concepts:</p> <p>Black-footed Albatross, which is a black and white bird.</p> <p>Black-footed Albatross, which has black wingtips.</p> <p>Black-footed Albatross, which can be seen gliding over the ocean.</p> <p>Black-footed Albatross, which is a large, long-winged, seabird.</p>	<p>After class names removal, VDES predicts sooty albatross with these concepts:</p> <p>black or dark grey plumage</p> <p>long, black legs</p> <p>webbed feet</p> <p>a white band around its neck</p>	<p>LaBo predicts black-footed albatross with these concepts:</p> <p>most abundant albatross species</p> <p>one of the largest albatrosses</p> <p>help protect black-footed albatrosses by supporting organizations</p> <p>only albatross species that has completely black legs and feet</p>	<p>CDL (Ours) predicts black-footed albatross with these concepts:</p> <p>long, narrow wings</p> <p>a long, hooked bill</p> <p>webbed feet</p> <p>black or grey body</p>

Figure 2: Examples of how different models conduct image classification based on the concepts. Correct predictions and concepts are in green, while wrong concepts and non-visual concepts are in red. Though VDES (Menon & Vondrick, 2022) and LaBo (Yang et al., 2023) can both classify the image correctly and the concepts are mostly correlated with the class names (highlighted in orange). After the removal of class name in VDES, we observe that VDES classifies this image as ring tailed lemus and correlate the image with irrelevant concepts. Our proposed method (CDL) can predict *giant panda* correctly based on the class-agnostic concepts.

REFERENCES

- Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023.