Pre-trained Vision-Language Models Learn Discoverable Visual Concepts

Yuan Zang[1], Tian Yun[1], Hao Tan[2], Trung Bui[2], and Chen Sun[1]

[1] Brown University
{yuan_zang,tian_yun,chensun}@brown.edu
[2] Adobe Research
{hatan,bui}@adobe.com

Abstract. Do vision-language models (VLMs) pre-trained to caption an image of a *durian* learn visual concepts such as *brown* (color) and *spiky* (texture) at the same time? We aim to answer this question as visual concepts learned "for free" would enable wide applications such as neuro-symbolic reasoning or human-interpretable object classification. We assume that the visual concepts, if captured by pre-trained VLMs, can be extracted by their vision-language interface with text-based concept prompts. We observe that recent works prompting VLMs with concepts often differ in their strategies to define and evaluate the visual concepts, leading to conflicting conclusions. We propose a new concept definition strategy based on two observations: First, certain concept prompts include shortcuts that recognize correct concepts for wrong reasons; Second, multimodal information (e.g. visual discriminativeness, and textual knowledge) should be leveraged when selecting the concepts. Our proposed concept discovery and learning (CDL) framework is thus designed to identify a diverse list of generic visual concepts (e.g. *spiky* as opposed to *spiky durian*), which are ranked and selected based on visual and language mutual information. We carefully design quantitative and human evaluations of the discovered concepts on six diverse visual recognition datasets, which confirm that pre-trained VLMs do learn visual concepts that provide accurate and thorough descriptions for the recognized objects. All code and models are publicly released.[3][4]

Keywords: Visual Concepts · Vision-Language Models

1 Introduction

Can a vision-language model (VLM) pre-trained on images of *California seagull* learn visual concepts [12–14, 28], such as *yellow legs* and *white belly*, to describe an image of a *Gentoo penguin*, which the model might not see during training? Visual concepts such as color, shape, and texture help models generalize

[3] Code is available at: https://github.com/brown-palm/Concept-Discovery-and-Learning
[4] Project homepage: https://conceptdiscovery.github.io

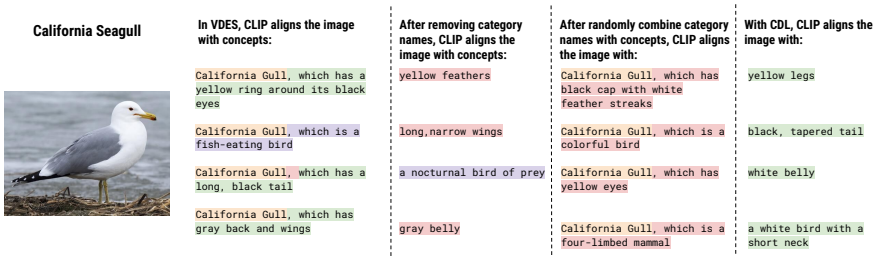

Fig. 1: The design of concept prompts plays a critical role on the understanding of whether VLMs learn visual concepts. We can observe that concept-augumented prompt can predict correct correct visual concepts (e.g., *yellow ring around black eyes*) when the prompt is associated with the category name (*California seagull*). When the category name is removed from the prompt (second column), the retrieved concepts are either non-visual or incorrect. We attribute this to the category name bias (third column), as the correct category can be retrieved by CLIP even when the paired descriptions are randomly shuffled and thus irrelevant. We propose a concept discovery and learning (CDL) framework and demonstrates that pre-trained VLMs can indeed learn visual concepts (e.g., last column). Correct predictions and concepts are in green, while wrong concepts are in red and non-visual concepts are in violet. The category names are in orange.

compositionally [7,9,22,29], and can be incorporated into neuro-symbolic frameworks [19,37] or offer concept-based explanations for classification decisions [10].

Our paper aims to investigate if VLMs, such as CLIP [25], learn visual concepts automatically when pre-trained on image and text pairs with contrastive learning objectives. We hypothesize that the visual concepts, if captured by the pre-trained VLMs, can be directly extracted by prompting their vision-language interface, without needing to probe their internal representations. Our research question can thus be formulated as discovering the visual concept prompts, and evaluating the quality of the extracted concepts. Intuitively, the desirable visual concepts should be *precise*, to faithfully reflect the visual concepts presented in an image; They should also be *thorough*, so that the concepts can provide most of the distinctive visual features for an object of interest.

Interestingly, several recent works [21,35,39] studying visual concepts in pre-trained VLMs reached different conclusions. We attribute these discrepancies to their different strategies for concept prompt selection and evaluation. For example, Yun et al. [39] observed that CLIP does not appear to recognize fine-grained visual attributes for birds [30], when a pre-defined list of visual concepts by bird experts was used. In contrast, Menon et al. [21] demonstrated that object prompts with concepts proposed by a large language model (LLM) appear to provide interpretable object classification, as the concept descriptions are nicely correlated with the recognized object categories. As illustrated in Figure 1, we observe that the design of concept prompts has a substantial impact on the "recognized" concepts: when certain *shortcuts* are included (e.g., the object category

California seagull) in the prompt, the retrieved concepts (e.g., *black eyes*) can be correct for wrong reasons, thus undermining their ability to answer our research question. We also observe that the visual concepts need to be selected based on both visual and language cues, as certain concepts returned by an LLM is correct, but cannot be recognized visually (e.g., *fish-eating*). As for evaluation, proxy benchmarks such as classification accuracy [10] are often used, which may not always positively correlate with the quality of the visual concepts [20].

In response to these observations, we propose a new concept discovery strategy to eliminate the shortcuts and utilize both visual and language information, in order to properly investigate the concepts in pre-trained VLMs. We first use a large and diverse image captioning dataset [26] as a source of objects to discover diverse and thorough visual concepts shared by multiple objects. We then rank and select the visual concepts based on multimodal information: Given a collection of images and their text descriptions, we prefer the visual concepts that can be both reliably recognized from the images (e.g., with a pre-trained VLM), and deemed as suitable descriptions based on the text descriptions (e.g., according to the prior knowledge encoded by an LLM). We compute this ranking with a mutual information formulation. When evaluating the concepts on a specific domain, we can use the same formulation to select a compact list of concepts. Finally, we propose a self-supervised method to adjust the final linear projection layer of pre-trained VLMs, which we show empirically can further improve the quality of the concepts without fine-tuning the encoder backbones. We name our overall framework as Concept Discovery and Learning (CDL).

We observe that as expected, when applying the concepts extracted by CDL for interpretable object recognition, our method consistently outperforms baselines in both full-shot and few-shot settings, across six visually diverse benchmarks. More importantly, we introduce a suite of quantitative and human evaluation protocols to measure the precision and thoroughness of concepts extracted by CDL. Across different evaluation protocols, we observe that the concepts extracted by CDL from CLIP are both precise and thorough, demonstrating that pre-trained vision-language models can indeed learn visual concepts via its vision-language interface. Our models and code will be publicly released.

2 Related Work

Vision-and-language models (VLMs) pretrained on unlabeled pairs of images and texts from the internet have shown great success on multi-modal benchmarks. Representations learned by these VLMs can be transferred to a wide range of tasks, such as visual question answering [2,16,17], and image and video captioning [18,34,40]. These pretrained VLMs can directly recognize complex concepts such as object categories in a zero-shot setting with text prompts [1,25]. However, it remains unclear whether VLMs recognize composite concepts by learning and reasoning over primitive concepts such as colors and shapes. Understanding the roles of these visual concepts is important since they can serve as clues

for interpretable fine-grained visual recognition [4, 8, 10], zero-shot or few-shot learning [21, 24], open-world object detection [36] and visual reasoning [19].

Previous studies have explored the capability of VLMs to capture and compose primitive concepts [38, 39] and bind visual concepts with objects [15, 31]. Yun et al. [39] demonstrate that VLMs do not capture composable primitive concepts by intervening a linear classifier learned from VLM-predicted concepts. Meanwhile, previous works [21, 24] show that concept-augmented prompts do provide improvements for VLM-based image recognition. In this paper, we aim to address this discrepency and understand pretrained VLMs' true capability of encoding interpretable visual concepts.

Previous research [10] have proposed the Concept Bottleneck Models (CBM) to decompose the end-to-end decision into concept-level reasoning. The concept bottlenecks can provide interpretable decision basis for complex recognition and reasoning, and are applied in various machine learning domains such as medical diagnosis [33]. Prior works [32, 35] have proposed to build the CBM with descriptive concepts to improve the recognition performance of VLMs.

3 Method

We first describe a framework of extracting visual concepts from pre-trained VLMs and its application for object recognition in Section 3.1. We show in Section 3.2 that the text prompts used to extract visual concepts may include shortcuts, and the concept activations may not correspond to how likely the visual concepts actually present in an image. We address this issue from two perspectives: We first propose a new framework in Sections 3.3 and 3.4 to discover the visual concepts that are category-generic and visually salient from a pre-trained VLM. We then propose a suite of quantitative and human evaluations on the interpretability, precision, thoroughness, and generalizability of the discovered visual concepts in Section 3.5. Together they help us understand if pre-trained VLMs learn to encode visual concepts.

3.1 Object Recognition via Visual Concepts

Vision-language models such as CLIP [25] jointly learn an image encoder $\mathcal{E}_\mathcal{I}$ and a text encoder $\mathcal{E}_\mathcal{T}$ to align images and texts in a shared embedding space. Thanks to the flexibility of the image-language interface, several recent works [21, 24, 32, 35, 39] attempted to construct semantically interpretable representations by projecting an encoded visual embedding with basis defined by encoded text embeddings. The text embeddings are obtained by encoding manually designed or automatically generated text "prompts" that are likely to correspond to visual concepts. Concretely, given an image I and a set of concept prompts $\mathbb{P}$ of N concepts, one can project the visual features $\mathcal{E}_\mathcal{I}(I)$ into the space of concepts as an n-dimensional concept activation vector $\mathbf{a} = [a_1, a_2, ..., a_N]$. Each concept activation is computed as $a_i = \text{sim}(\mathcal{E}_\mathcal{I}(I), \mathcal{E}_\mathcal{T}(p_i))$, where p_i is the i-th concept

prompt, and $\text{sim}(\cdot)$ is a similarity function, such as cosine similarity, that measures the similarity between the encoded visual and text embeddings. Our paper assumes that the visual concepts, if well learned by a VLM, can be extracted as concept activations $\mathbf{a}$ via the text prompts.

The concept activations can be utilized for multi-modal object recognition with a function $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ that predicts object categories, where M is the total number of categories. When f is a linear function $f(\mathbf{a}) = \mathbf{a} \cdot \mathcal{W}$, the learned $N \times M$ weight matrix $\mathcal{W}$ allows us to interpret how the visual concepts are utilized for object recognition. A higher positive weight w_{ij} indicates the i-th concept is deemed as important positive evidence for recognizing the j-th object category, and a near-zero weight indicates the concept is deemed as irrelevant. The linear classifier that maps concept activations to categorical predictions is referred to as Concept Bottleneck Models [10] (CBM), which have been adopted to study concept learning with VLMs in [32, 35, 39]. For zero-shot object recognition when $f(\cdot)$ cannot be learned from data, some works [21, 24] assume the visual concepts are category-specific (hence the object names are included in the prompts), and simplify $f(\cdot)$ to be a linear function $f_j(\mathbf{a}) = \frac{1}{|\mathcal{C}_j|} \sum_{a_i \in \mathcal{C}_j} a_i$, where $\mathcal{C}_j$ is the set of visual concepts for the j-th category, and $\mathcal{C}_j \cap \mathcal{C}_k = \emptyset$. We observe that this assumption does not hold for real-world datasets as many visual concepts are often shared by several objects.

3.2 Do Prompted Activations Correspond to Visual Concepts?

Large language models (LLMs) are often utilized as the knowledge source to propose the relevant visual concept prompts given an object category [21, 24, 32, 35]. One example is illustrated in Figure 1, where for California Gull, concept prompts such as "California Gull, which has a long, black tail" are proposed. As discussed in Section 3.1, the linear function $f(\mathbf{a})$ allows us to identify the most important visual concepts for object recognition by picking the highest weighted w_{ij} concepts i's for object j. One can then consider two proxy evaluations to measure whether the prompted concept activation a_i actually corresponds to the visual concept i that is present in an image: First, by measuring the object classification accuracy, and assuming the higher the accuracy is, the more precise the concept activations are. Second, by comparing the top ranked concepts for an object category and those identified by human experts, which allows us to understand not only the precision but also the thoroughness of the concepts.

We observe that these proxy evaluations, while intuitive, require careful design of concept prompting strategy in order to draw conclusions on whether the concepts are actually encoded by the pre-trained VLMs. The first issue we identify is the existence of certain shorts in the text prompts, leading to "false positive" conclusions. For example, Figure 1 illustrates the concepts with the highest activations according to CLIP. Although the first column appears to indicate that most of the selected concepts are semantically correlated with the input image of a California Gull, it remains unclear whether the concepts are retrieved because they are recognized by CLIP or the class name is utilized as

a shortcut. We perform a simple ablation to investigate this potential shortcut: In the second column of Figure 1, we observe that irrelevant concepts are selected when category names are removed from the prompts. In the third column, we observe that when we combine the category names with randomly shuffled descriptors, CLIP tends to align images with concepts that contain correct category names and wrong descriptors. This phenomenon indicates that the category names, instead of the visual concepts, are used to generate the concept activations. We reinforce these qualitative observations in Section 4.2, where we demonstrate consistently across six datasets that the zero-shot classification accuracy remains similar when LLM-generated concepts or randomly shuffled concepts are paired with category names, and that the accuracy drops significantly when category names are removed from the text prompts.

A second issue is the existence of sub-optimal text prompts, leading to "false-negative" conclusions. For example, the concepts marked with purple background in Figure 1 are not visually recognizable (e.g. "nocturnal bird"), and hence should not be considered as visual concepts. Similarly, we observe that the text prompts used by Yun et al. [39] are designed by birding experts, which may not be in a friendly format to serve as text prompts (e.g. "crested head pattern").

In order to address both issues, we propose to discover category-generic (hence no shortcuts) and visually discriminative concepts directly from a pre-trained VLM (hence the text prompts are more friendly to the VLM). We then propose a suite of evaluations in order to draw robust conclusions on the precision and thoroughness of the discovered concepts.

3.3 Visual Concept Discovery and Learning

We propose to discover visual concepts from a large pool of diverse objects, so that the discovered concepts would be generic and generalizable to different domains. Towards this goal, we adopt the Conceptual Captions 3M [26] dataset, which contains three million images and their captions. As illustrated in Figure 2, our overall goal is to identify the key objects and their associated visual concepts from the image captions, with the help of a large language model as the knowledge source. The candidate visual concepts are then checked for their visual discriminativeness, by verifying if a candidate visual concept can be reliably recognized from the corresponding image. Thus, we can select the concepts that can both be recognized by the VLM in the image and be deemed by the LLM as a relevant attribute for the object described in the caption.

As objects in captions often hold specific dependencies, such as serving as the subject or object of an action, we employ Dependency Parsing [5] along with a set of designed rules (see Appendix for details) to retrieve the words and phrases that potentially correspond to objects in the captions (e.g., "king penguin" from a sentence "a group of king penguins walking in the snow"). We then leverage the LLM as an external knowledge base to obtain visual concepts for recognizing the objects. We design a set of prompts (e.g., "What are useful visual features for distinguishing an {object} in a photo?") to query the LLM and retrieve relevant

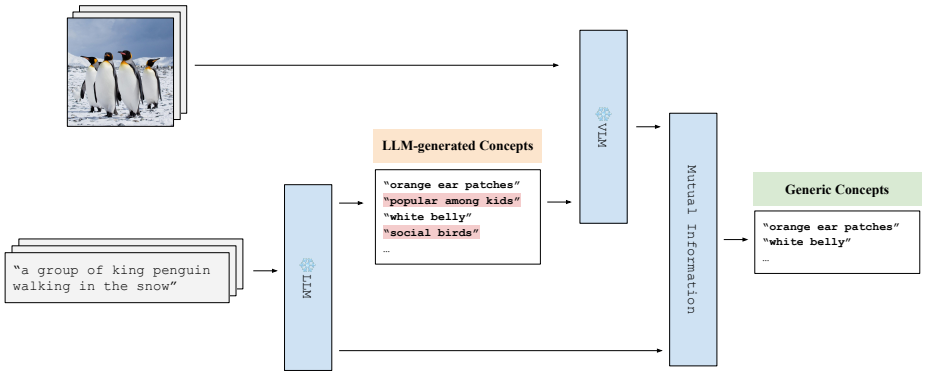

Fig. 2: Illustration of our proposed concept discovery method. Given image-caption pairs, we first identify objects from the captions and utilize a large language model to propose candidate concepts for the objects. The concepts are then ranked by the agreement between VLM knowledge (concept recognition from the image) and LLM knowledge (concept proposed based on the caption) based on mutual information.

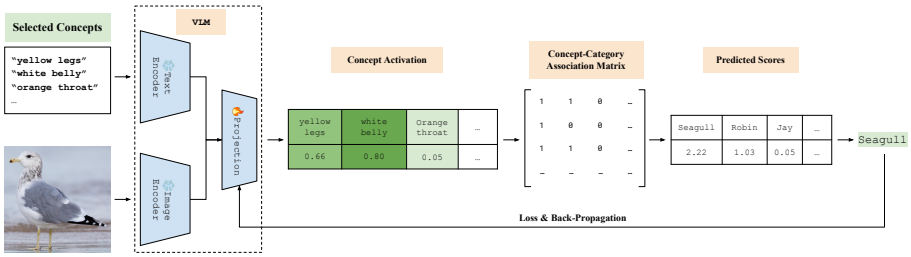

Fig. 3: Illustration of the concept-based object recognition framework and our proposed concept learning method. We map the concept activations $\mathbf{a}$ to categories with the concept-category association matrix $\mathcal{W}$. For object recognition, only $\mathcal{W}$ is optimized based on object classification supervision. For concept learning, we assume $\mathcal{W}$ is a binary matrix given by LLM knowledge, and learns to update $\mathbf{a}$ by fine-tuning the last layers of visual and text encoders in the VLM. We rely on the VLM recognized object labels as opposed to ground truth object labels, hence the process is self-supervised.

concepts for the object of interest. We take the union of the concepts discovered for all objects as the preliminary list of generic visual concepts.

We then propose to filter the concepts so that the selected ones are also visually discriminative. Intuitively, a visual concept should be ranked higher when the VLM can recognize it from the image if and only if when the same concept is proposed by the LLM based on the image caption. Otherwise, the concept either is not specific to the object of interest (i.e. can be recognized from the image but not proposed to describe any object mentioned in the caption), or likely to correspond to non-visual concepts (e.g. "popular among kids" and "social

birds" in Figure 2). Specifically, given a concept and a list of image-caption pairs, we define two variables X and Y, where X corresponds to the image-concept similarity as measured by a VLM, and the Y is a binary indicator on the caption-concept correspondence according to an LLM. A high X value x indicates that the VLM can recognize the concept from the image well, and a higher Y value y indicates the concept is deemed by the LLM as a relevant attribute for the object described in the caption. We adopt the Mutual Information (MI) to determine the agreement between the VLM and the LLM:

$$I(c) = \sum_{y \in Y_c} \sum_{x \in X_c} P_{x,y}(x,y) \log \frac{P_{x,y}(x,y)}{P_x(x) P_y(y)}, \tag{1}$$

which measures the amount of information gain of one variable by knowing another variable. A higher $I(c)$ means that X_c and Y_c are in agreement.

We assume that the vision-language pre-training learns powerful encoders for recognizing visual concepts, but visual concept recognition can be further improved by "re-aligning" the image-text interface. As illustrated in Figure 3, we keep the weights of pre-trained text and image encoders frozen and only fine-tune the last linear projection layers used to compute the concept activations $\mathbf{a}$. We re-purpose the concept-based object recognition framework $f(\mathbf{a}) = \mathbf{a} \cdot \mathcal{W}$: Instead of learning the $N \times M$ weight matrix $\mathcal{W}$, we construct a fixed concept-category association matrix $\mathcal{W}_{\text{LLM}}$ with binary weights, where each weight is obtained by querying the LLM with prompts like "Does the {category_name} usually have the attribute {concept}?". We then fine-tune only the linear projection layers of the VLM through the concept bottleneck, by asking it to justify its own prediction: Namely, the classification objective is obtained by asking the VLM to perform zero-shot object classification. We refer to our two-stage pipeline as the Concept Discovery and Learning (CDL) framework, it leverages knowledge already encoded in pre-trained LLMs and VLMs, and the whole framework is self-supervised as it requires no additional human annotations.

3.4 Visual Concept Applications

Our concept discovery and learning framework aims to obtain a list of generic and visually discriminative concepts. For object recognition applications, the downstream dataset usually focuses on a specific domain, and it is often desirable to obtain a compact and performant set of visual concepts (e.g. "four-wheeled" is useful for recognizing cars, but not so much for birds). Towards this goal, we first construct the concept-category association matrix $\mathcal{W}_{\text{LLM}}$ and perform concept learning only for the object categories in the target dataset, so that irrelevant concepts not used by any objects are automatically discarded.

The list of remaining visual concepts can be further optimized based on their usefulness and generalizability. We first try to identify the visual concepts that can be reliably recognized from the target dataset. We re-purpose $I(c)$ where Y_c is now obtained by looking up the concept association matrix $\mathcal{W}_{\text{LLM}}$ knowing the ground truth object label for an image. A higher $I(c)$ means the concept c is

useful to recognize a subset of object categories in the dataset (according to the LLM knowledge used to construct $\mathcal{W}_{\text{LLM}}$), and can be reliably recognized from the images when the concept is expected to appear.

We also expect the selected concepts to generalize to unseen but related object categories. We employ a heurstics where a visual concept is more likely to generalize if it is already used by many known object categories. The "generalizability" of a concept $G(c)$ can hence be estimated by the ratio of object categories that contain the concept c over the total number of object categories.

There exists a natural trade-off between the usefulness and generalizability of a visual concept, we hence compute the weighted average $\alpha \cdot I(c) + (1-\alpha) \cdot G(c)$ to rank the concepts and apply a fixed budget on the number of visual concepts to use for each downstream benchmark. We select α based on classification performance on the validation set (a lower α, namely more general concepts, is preferred whenever the accuracy remains high).

3.5 Visual Concept Evaluation

Finally, we propose a suite of evaluation protocols to measure the quality of VLM-discovered concepts in terms of *Interpretability*, *Precision*, *Thoroughness* and *Generalizability*.

Interpretability metric from [39] quantitatively measures how well a VLM learns concept-category associations. The discovered concepts are interpretable if they are associated with images in a human-understandable manner. For example, we expect the model to recognize an image of a "Giant Panda" according to the concepts "Rounded head", "Black eye patches" and "Black and white fur" instead of irrelevant concepts like "long wings". This metric is computed with an intervention approach. Given a trained concept-association matrix, we replace the concept activations with the binary ground-truth concept values to make predictions. High intervention accuracy indicates high interpretability of visual recognition since the concept-category association is consistent with human knowledge.

Precision is measured with human evaluation on how well the discovered concepts provide correct reasoning clues for visual recognition. As illustrated in Section 3.1, the contribution of a certain concept to the predicted category j can be measured by the dot product between the concept activation $\mathbf{a}$ and $\mathcal{W}_j$ from a CBM weight matrix $\mathcal{W}$. We expect the top-weighted concepts to be accurate depictions of the predicted object when the model prediction is correct. For each correctly-classified image, we ask human annotators to determine whether the top-k weighted concepts actually describe the image, and calculate the proportion as the precision score. The selection of the parameter k is contingent upon the scale of the concept space within each dataset. For datasets with large amounts of concepts, we can find more descriptive concepts for a certain category. Hence, we set a larger value for k. Conversely, datasets with relatively few concepts have a diminished pool of descriptive elements for individual categories. Consequently, a smaller value for k is deemed appropriate in such scenarios.

Thoroughness measures how well the discovered concepts cover important features to recognize the objects in downstream domains. Given an image correctly predicted for its category, we first leverage the LLM knowledge to obtain a complete list of concepts that are potentially related to the category. We then ask human annotators to select from the complete list of concepts that can be inferred visually from the image of interest. We call these the important visual concepts for an image. We then select the top-k weighted concepts and calculate the thoroughness score as the percentage of important visual concepts covered by the top-k weighted concepts.

Generalizability measures whether the discovered and learned concepts can benefit the recognition of unseen objects. We consider in-domain and cross-domain generalization of the discovered concepts. For the in-domain generalization, we first randomly split the category list of a given dataset into seen categories and unseen categories. We select from the discovered concepts separately for the seen and unseen categories (as in Section 3.4), we then perform concept learning only the examples of seen categories. These concepts are transferred to train concept-based classifiers for the unseen categories, and we report classification accuracy as the proxy to measure in-domain generalization. Cross-domain generalization is measured following the same procedure, but for a seen dataset and a different dataset.

4 Experiments

4.1 Experimental Setup

Datasets: We conduct experiments on several challenging fine-grained image classification datasets, including ImageNet [6], Food-101 [3], CIFAR-100 [11], CIFAR-10, CUB-200 [30] and Flowers-102 [23]. The statistics and split of the datasets are shown in the appendix.

Baselines: We compare with LaBo [35] and LM4CV [32], which are the state-of-the-art works on concept-based visual recognition. Following the setting in LM4CV, we control the bottleneck sizes (number of concepts) to be the same for the baselines and our model for fair comparison. Besides full-shot classification, we also compare with LaBo on few-shot settings, where we select concepts and train the model on limited data. It is infeasible for LM4CV to conduct few-shot classification since it requires the whole training set on concept generation.

Implementation Details: We use the same LLM `GPT-3-text-davinci-002` to obtain descriptors as previous works. We also use the same CLIP backbone (ViT-L-14) to compare with baseline models. Following [39], we use logistic regression to train the concept bottleneck models. We observe that the performance of CBM is robust to the choice of hyperparameters and use the default Scikit-learn hyperparameters for all datasets. For concept learning, we use the AdamW optimizer with 5e-4 learning rate and 1e-4 weight decay to fine-tune the CLIP model, and we use the validation loss to select checkpoints. For human evaluation experiments, we hire annotators from Amazon Mechanical Turk. For

Method	Prompt Design	ImageNet	Food-101	CIFAR-100	CIFAR-10	CUB-200	Flowers-102
VDES	Category Name	71.6	91.8	75.9	96.2	63.1	77.4
VDES	Name w/ LLM Concepts	75.0	**92.4**	77.7	**96.6**	63.5	78.9
VDES	Name w/ Random Concepts	70.1	91.6	75.4	95.0	62.1	78.7
VDES	LLM Concepts	22.1	3.6	30.9	70.7	5.3	7.0
CDL	CDL Concepts	**75.7**	91.5	**77.8**	96.5	**64.7**	**80.9**

Table 1: Zero-shot classification on six object recognition benchmarks. We observe that although augmenting category names with LLM-generated concepts improves classification accuracy (e.g. VDES [21]), the method still mainly relies on category names in the text prompts, and it remains unclear if the concepts included in the text prompts are properly recognized by pre-trained VLMs. When paired with randomly shuffled concept in the prompts, the accuracy drops only moderately; when the category names are removed, the accuracy drops significantly. Our proposed concept discovery and learning (CDL) framework achieves competitive zero-shot classification performance without using category names as a shortcut.

each example, we ask three annotators to annotate and use the majority vote to obtain the result. We conduct Students' T-test [27] to evaluate the statistical significance of the human evaluation results. The detailed design of human evaluation and the statistical significance results are shown in the appendix.

4.2 Concept-Augmented Prompts Are Category-Biased

As discussed in Section 3.2, we conduct ablation experiments to understand if the concepts used in the prompts of previous zero-shot classification methods lead to the improved classification accuracy. The results in Table 1 are consistent with Figure 1, both of which show that the concept-augmented text prompts do not offer conclusive evidence whether pre-trained VLMs learn to encode concepts. In contrast, our proposed CDL provides concept prompts independent of category names, and still achieves competitive zero-shot classification accuracy.

Besides category biases, the discovered concepts in previous works contain many non-visual concepts as illustrated in Figure 1. We conduct human evaluation to compare our discovered concepts with previous works by the proportion of non-visual concepts and concepts containing class names. For both baselines and our CDL, we randomly select 100 concepts for each dataset for human evaluation. The results are shown in Table 2. We can observe that CDL offers visual concepts that are more category-agnostic and visually discriminative.

4.3 Classification Performance with Visual Concepts

Following the settings in LaBo and LM4CV, we set the bottleneck size to be the same or 2 times of number of classes in a given dataset. The same ViT-L-14 CLIP model is used across all methods. To ensure a fair comparison, we conduct the classification using our concept discovery and selection methods, without

	%Category-agnostic	%Visual
LaBo	65.50	66.83
ML4CV	82.33	73.83
CDL	**92.33**	**85.50**

Table 2: Human evaluation results on the proportion of category-agnostic and visually-discriminative concepts. The results show that our concepts discovered from general corpora are less biased by category names and more visually discriminative.

	ImageNet		Food-101		CIFAR-100		CIFAR-10		CUB-200		Flowers-102	
#Concepts	1000	2000	101	202	100	200	10	20	200	400	102	204
LaBo	83.2	83.6	89.8	91.1	80.5	84.1	77.8	92.2	79.7	81.3	95.5	94.9
LM4CV	83.4	83.6	90.0	91.1	81.7	84.2	80.4	92.3	77.5	80.4	94.5	93.8
CDL	**83.6**	**84.0**	**94.2**	**94.9**	**83.8**	**85.8**	**82.1**	**93.5**	**83.2**	**83.4**	**96.3**	**95.7**

Table 3: Comparison with baselines on classification accuracy with different bottleneck sizes. All methods are based on ViT-L-14 CLIP checkpoint.

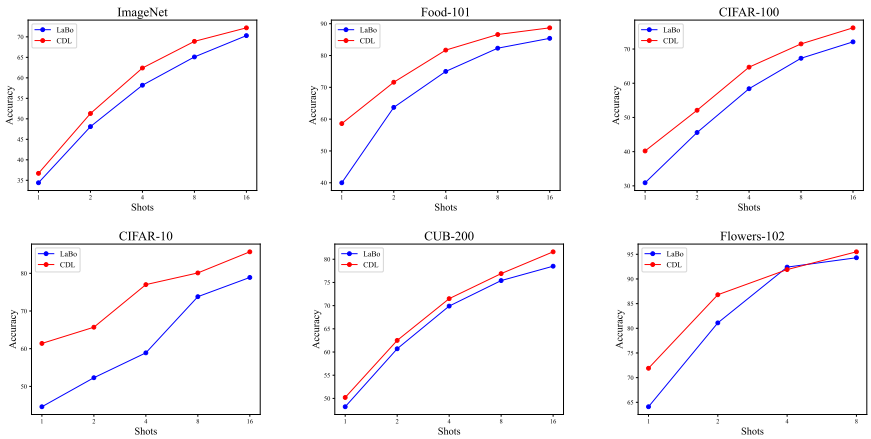

Fig. 4: Few-shot classification evaluation with LaBo and our proposed CDL.

concept learning. Table 3 shows that our method consistently outperforms the baseline methods on all datasets.

We then adopt the few-shot learning setting, where the concepts are selected only with the few training examples. Figure 4 shows that CDL consistently outperforms LaBo, especially when the number of training examples is smaller.

4.4 Evaluation of the Discovered Concepts

Figure 5 shows the evaluation results of the *Interpretability*, *Precision*, and *Thoroughness* of the baselines and our method. We can observe that despite having

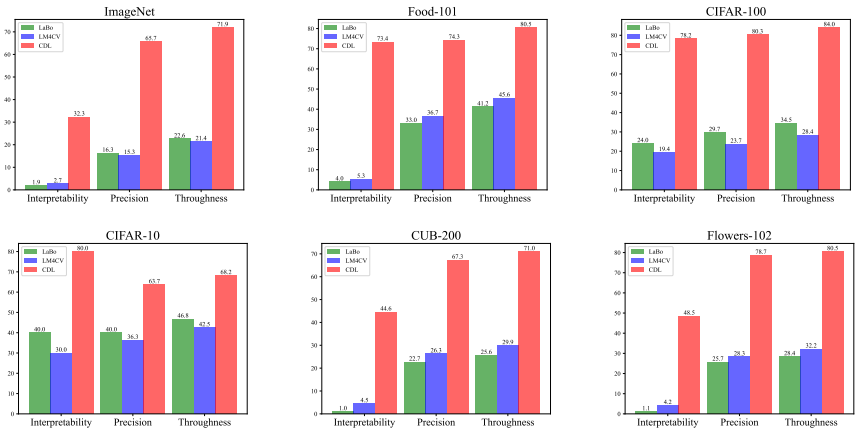

Fig. 5: Human evaluation of the discovered concepts for their interpretability, precision, and thoroughness.

	ΔClassification	ΔInterpretability	ΔPrecision	ΔThoroughness
LaBo	0.2	16.1	5.8	7.6
LM4CV	0.4	14.2	4.5	6.4
CDL	**1.0**	**18.3**	**7.9**	**10.4**

Table 4: In-domain generalization evaluation, where Δ denotes the improvement of fine-tuned CLIP compared to the original CLIP on the unseen categories.

	ΔClassification	ΔInterpretability	ΔPrecision	ΔThoroughness
LaBo	-1.5	0.1	-3.7	-4.5
LM4CV	-0.3	-0.6	-6.4	-8.2
CDL	**0.2**	**7.0**	**10.3**	**13.9**

Table 5: Cross-domain generalization evaluation from ImageNet to CUB-200.

high classification performance, the baseline models discover concepts that exhibit unsatisfactory interpretability, precision, and thoroughness. In contrast, our CDL framework provides significant improvements on all three metrics.

Tables 4 and 5 show the in-domain and cross-domain generalization results. , and Figure A3 shows the generalization of our discovered concepts under the few-shot setup. Our proposed CDL outperforms both baselines for its generalizability, especially for the cross-domain scenario when transferring from ImageNet to CUB-200. We observe that LaBo and LM4CV struggle with cross-domain generalization as they select completely different concepts for different datasets and few common patterns can be learned with their methods.

Finally, Table 6 shows the human evaluation results before and after concept learning on the ImageNet dataset. We can observe both our concept discovery

	Interpretability	Precision	Thoroughness
W/o Concept Learning	21.9	44.0	49.6
W Concept Learning	32.3	65.7	71.9

Table 6: Human evaluation of the discovered concepts before and after concept learning on the ImageNet dataset.

Fig. 6: Examples of the top weighted concepts correlated with the given images.

and concept learning can improve the quality of the extracted concepts compared with baselines.

5 Conclusion and Future Work

In this paper, we investigate the question of whether pre-trained Vision-Language Models (VLMs) can encode primitive visual concepts. We first illustrate that category-biased concepts extracted from specific datasets do not offer conclusive evidence of the concept learning capacity of VLMs. To resolve this issue, we design a novel framework to discover category-agnostic and visually discriminative concepts from a large image-caption dataset with the help of VLMs and LLMs.

To make use of the discovered concepts for downstream tasks, we propose a novel method to build concept bottlenecks with a compact and performant set of concepts for a specific domain. We also propose a self-supervised concept learning framework to re-align the concept knowledge in VLMs for the category classification in specific domains. To prove that VLMs do learn useful and interpretable concepts, we propose a suite of comprehensive protocols to evaluate the quality of the discovered concepts and perform exhaustive experiments including human evaluation. The experimental results demonstrate that VLMs do capture primitive concepts that can lead to effective, interpretable, and generalizable visual recognition.

While we illustrate that VLMs do learn discoverable concepts, it is still significant to understand what kind of concept and compositionality knowledge cannot be learned in the contrastive learning-based pre-training of VLMs. In future work, we plan to explore whether VLMs can capture the semantic and spatial relationships between concepts and utilize these relationships to perform complex multi-modal reasoning.

Acknowledgements: This work is in part supported by a gift from Adobe Research and a Richard B. Salomon award for Chen Sun. We thank helpful discussions with Professors Ellie Pavlick and Stephen Bach. We thank Yue Yang, An Yang, Yu Wang for providing the open-sourced code for the baselines.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 3
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 **3** (2023) 3
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014) 10
4. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5948–5955 (2023) 4
5. De Marneffe, M.C., MacCartney, B., Manning, C.D., et al.: Generating typed dependency parses from phrase structure parses. In: Lrec. vol. 6, pp. 449–454 (2006) 6
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009) 10
7. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1778–1785. IEEE (2009) 2

8. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in neural information processing systems **32** (2019) 4
9. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in Neural Information Processing Systems **36** (2024) 2
10. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning. pp. 5338–5348. PMLR (2020) 2, 3, 4, 5
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 10
12. Lake, B., Salakhutdinov, R., Gross, J., Tenenbaum, J.: One shot learning of simple visual concepts. In: Proceedings of the annual meeting of the cognitive science society. vol. 33 (2011) 1
13. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015) 1
14. Lee, S., Zhang, Y., Wu, S., Wu, J.: Language-informed visual concept learning. In: The Twelfth International Conference on Learning Representations (2023) 1
15. Lewis, M., Yu, Q., Merullo, J., Pavlick, E.: Does clip bind concepts? probing compositionality in large image models. arXiv preprint arXiv:2212.10537 (2022) 4
16. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 3
17. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020) 3
18. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019) 3
19. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In: International Conference on Learning Representations (2019) 2, 4
20. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A.: Do concept bottleneck models learn as intended? arXiv preprint arXiv:2105.04289 (2021) 3
21. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2022) 2, 4, 5, 11
22. Nagarajan, T., Grauman, K.: Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 169–185 (2018) 2
23. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008) 10
24. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15691–15701 (2023) 4, 5
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 3, 4
26. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) 3, 6
27. Student: The probable error of a mean. Biometrika pp. 1–25 (1908) 11
28. Sun, C., Gan, C., Nevatia, R.: Automatic concept discovery from parallel text and visual corpora. In: Proceedings of the IEEE international conference on computer vision. pp. 2596–2604 (2015) 1
29. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022) 2
30. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 2, 10
31. Yamada, Y., Tang, Y., Yildirim, I.: When are lemons purple? the concept association bias of clip. arXiv preprint arXiv:2212.12043 (2022) 4
32. Yan, A., Wang, Y., Zhong, Y., Dong, C., He, Z., Lu, Y., Wang, W.Y., Shang, J., McAuley, J.: Learning concise and descriptive attributes for visual recognition. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3067–3077. IEEE Computer Society (2023) 4, 5, 10
33. Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al.: Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023) 4
34. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10714–10726 (2023) 3
35. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: 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) 2, 4, 5, 10
36. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems **35**, 9125–9138 (2022) 4
37. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. Advances in neural information processing systems **31** (2018) 2
38. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2022) 4
39. Yun, T., Bhalla, U., Pavlick, E., Sun, C.: Do vision-language pretrained models learn composable primitive concepts? Transactions on Machine Learning Research (2023) 2, 4, 5, 6, 9, 10
40. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2021) 3

A Hyperparameters

We adjust hyperparameters according to the performance of the models on the validation dataset (For ImageNet, we randomly select 10% of the training set as the validation set). For α in Equation 2, we set it to 0.7 for ImageNet dataset, 0.8 for Food-101, CIFAR-100, CUB-200 and Flowers-102 datasets and 0.9 for CIFAR-10 dataset. According to Equation 2 in Section 3.4, a smaller α will increase the generalizability but decrease the discriminativeness of the selected concepts, and thus decrease classification performance. To achieve a trade-off between generalizability of the selected concepts and classification performance, we monitor the classification performance when gradually decreasing α. We pick the α right before classification performance drops significantly.

B Additional Implementation Details

B.1 Dependency Parsing Rules

Given a caption, we perform Dependency Parsing to extract the grammatical relations from the caption. We extract the nouns or phrases from the following dependencies as potential objects in an image caption:

- *nsubj* / *nsubjpass*: The *nsubj* (nominal subject) is the noun that performs the main action. For example, in the sentence "the horse is eating grass", "horse" is the nominal subject. The *nsubjpass* (passive nominal subject) is the noun that performs the main action in a sentence of passive voice. For example, in the sentence "The dog is led by a leash", "dog" is the passive nominal subject. We extract the nominal and passive nominal subject as a word potentially corresponding to an object in the image.
- *dobj* / *iobj*: The *dobj* (direct object) is the noun or noun phrase that receives the action of the verb. For example, in the sentence "the horse is eating grass", "grass" is the direct object. The *iobj* (indirect object) is the noun or noun phrase that indicates to or for whom the action of the verb is done. For example, in the sentence "The man gives the girl a flower", "girl" is the indirect object. We extract the direct and indirect object as a word potentially corresponding to an object in the image.
- *amod*: An *amod* (adjectival modifier) is an adjective that describes a noun (e.g. "black dog", "white bird"). We extract the amod and its object as a phrase potentially corresponding to an object in the image.
- *compound*: The *compound* label indicates the word is part of a compound phrase like "king penguin". Once select a word following above rules, we check whether it is part of a compound phrase. If so, we extract the whole phrase as the object.

B.2 Mutual Information Implementation

The Mutual Information (MI) measures the information gain of one variable by knowing another variable. Given two variable X and Y, it measures the

KL divergence between the joint distribution $P(X, Y)$ and the product of the marginal distribution $P(X)P(Y)$:

$$MI = \sum_{y \in Y} \sum_{x \in X} P_{x,y}(x, y) \log \frac{P_{x,y}(x, y)}{P_x(x) P_y(y)}. \tag{2}$$

The MI is high when X and Y are positively or negatively related (e.g. high x value indicates high y value and low y value indicates low y value or vice versa). Given a concept and an image-caption dataset, the X variable corresponds to the image-concept similarity as measured by a VLM, and the Y variable is a binary indicator on the caption-concept correspondence according to an LLM. Namely, a higher x indicates that VLM can recognize the concept from the image well, and a higher y indicates the concept is deemed by LLM as a relevant "attribute" for the object described in the caption. As such, we prefer concepts with higher MI since it indicates that if a concept can be recognized from the image, the same concept should be relevant for the paired caption, and vice versa. This helps filter out concepts that are non-visual (relevant according to LLM, but cannot be recognized visually by VLM, such as "fish-eating" and "endangered animal"), irrelevant to the objects of interest (can be recognized by VLM, but irrelevant according to LLM, such as concept "yellow-leg" for recognizing food). Figure A1 shows some examples for the MI calculation.

Since the image-concept similarity is continuous, we utilize K-nearest-neighbors-based entropy estimation toolkit in Scikit-learn to perform discretization and calculate the MI.

C Dataset Details

The table A1 shows the statistical details of the datasets we choose.

Dataset	#Class	#Train	#Valid	#Test
ImageNet	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	102	4,093	1,633	2,463

Table A1: Statistical details of datasets. "#Class" means the number of classifications. "#Train", "#Valid", and "#Test" denote the instance numbers of each dataset respectively. For ImageNet, we randomly select 10% of the training set as the validation set and regard the validation set as the test set.

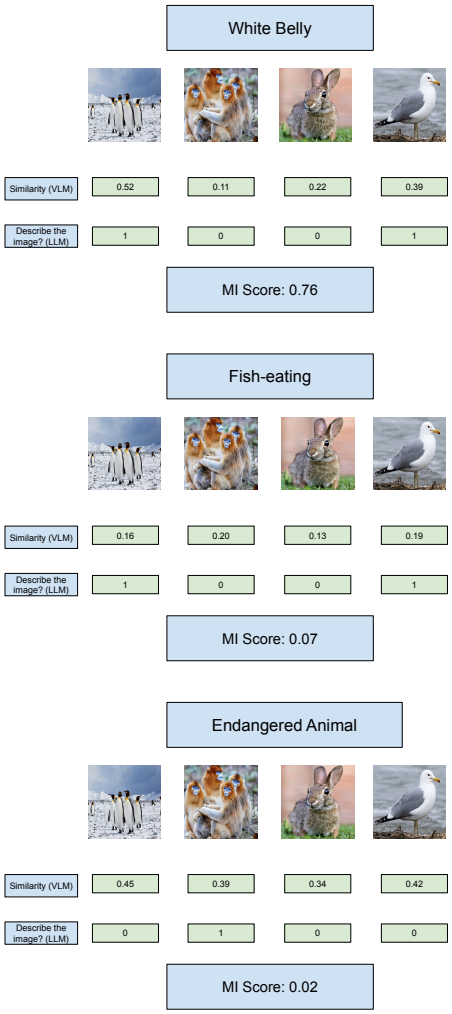

Fig. A1: Illustrations of Mutual Information calculation. We can observe that visually discriminative concepts such as "white belly" have high MI score because they can be reliably recognized from images that are supposed to contain the concept according to LLM, and vice versa. The non-visual concepts like "fish-eating" and "endangered animal" have low MI scores because VLMs cannot recognize these concepts from the images that are supposed to contain these concepts according to LLM.

D Human Evaluation Details

We hire workers on https://www.mturk.com to conduct human evaluation. To make our human evaluation more robust, for one data point we ask three human workers to annotate. For the precision and thoroughness metric, we randomly

Dataset	Method	Precision		Thoroughness		Category Agnostic		Visual	
		p-value	significance	p-value	significance	p-value	significance	p-value	significance
ImageNet	CDL v.s. LaBo	4.59e-65	✓	2.67e-62	✓	3.94e-11	✓	2.00e-3	✓
	CDL v.s. LM4CV	2.19e-68	✓	7.02e-66	✓	3.87e-01	✗	4.54e-2	✓
Food-101	CDL v.s. LaBo	7.38e-26	✓	8.73e-25	✓	1.32e-05	✓	1.30e-4	✓
	CDL v.s. LM4CV	1.59e-21	✓	4.97e-20	✓	9.31e-05	✓	2.78e-2	✓
CIFAR-100	CDL v.s. LaBo	2.85e-40	✓	3.44e-40	✓	8.59e-10	✓	7.26e-3	✓
	CDL v.s. LM4CV	1.27e-51	✓	4.99e-51	✓	1.56e-02	✓	1.78e-2	✓
CIFAR-10	CDL v.s. LaBo	4.28e-9	✓	9.58e-8	✓	5.35e-06	✓	2.36e-3	✓
	CDL v.s. LM4CV	5.28e-12	✓	1.39e-10	✓	7.90e-01	✗	2.27e-2	✓
CUB-200	CDL v.s. LaBo	1.31e-50	✓	4.54e-52	✓	3.57e-04	✓	6.18e-4	✓
	CDL v.s. LM4CV	5.12e-42	✓	2.95e-42	✓	5.09e-03	✓	1.40e-1	✗
Flowers-102	CDL v.s. LaBo	7.44e-45	✓	3.10e-42	✓	1.68e-09	✓	8.44e-3	✓
	CDL v.s. LM4CV	1.21e-40	✓	4.58e-37	✓	3.17e-02	✓	4.09e-2	✗

Table A2: The statistical significance of the human evaluation results. Smaller p-value indicates higher significance. The results show that out human evaluation results are statistically significant and our discovered concepts are consistently better than concepts in previous works on all datasets.

select 100 images on each dataset to evaluate. For CIFAR-10, CIFAR-100, Food-101, Flowers-102 we choose the top 3 concepts to annotate. For ImageNet and CUB-200 we choose the top 5 concepts to annotate. We annotate 2,200 data points for precision and around 3,000 data points for thoroughness. For the visual discriminability and category name containing, we utilize different methods to select 100 concepts for each dataset to evaluate and report the average score. Hence we annotate 1,800 data points for those two task. 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 LaBo. The average pairwise annotator agreement proportion on all datasets is 69.4%, which is comparable to the 69.8% proportion in LaBo.

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 out 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 precision and thoroughness metrics.

We show some examples of our human evaluation interface in Figure A2. The examples shown are for measuring precision of the concepts, and a similar interface is used to measure thoroughness, where the humans are asked to build a complete concept list for an image.

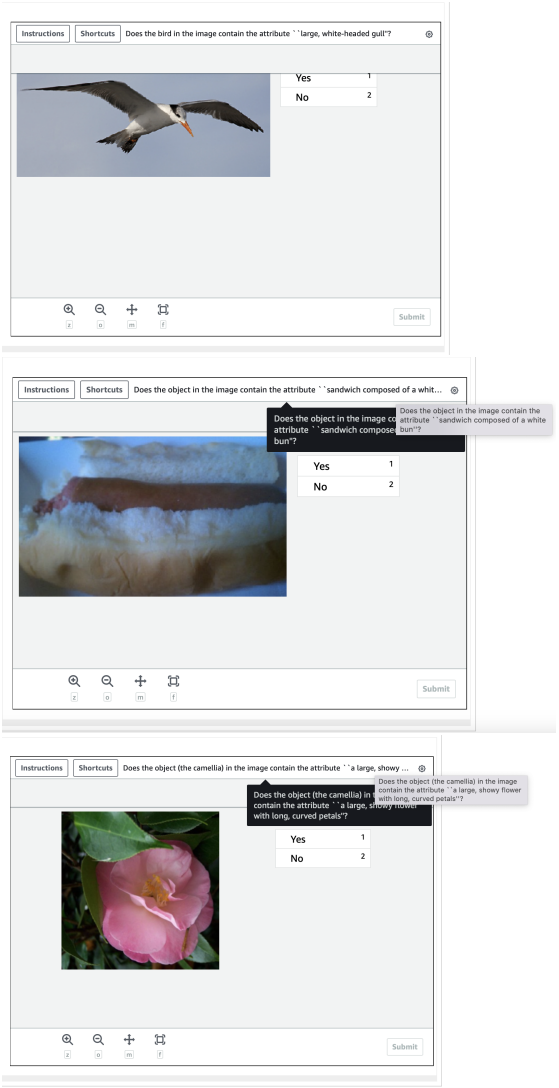

Fig. A2: Examples of the annotation interface.

E Ablation Study on Effectiveness of CDL Framework

In this section, we conduct ablation studies on the effect of different phases of our CDL framework. We compare the classification performance and interpretability of the concept-based image-recognition before and after concept learning to illustrate the effects of concept discovery and concept learning.

The results in Table A3 and A4 indicate that (1) it is the concept discovery that mainly contribute to the improvement of classification performance

	ImageNet		Food-101		CIFAR-100		CIFAR-10		CUB-200		Flowers-102	
#Concepts	1000	2000	101	202	100	200	10	20	200	400	102	204
W/o Concept Learning	83.6	84.0	94.2	94.9	83.8	85.8	82.1	93.5	83.2	83.4	96.3	95.7
W Concept Learning	83.8	83.9	94.4	94.9	83.6	85.3	96.1	96.5	82.5	82.1	96.6	96.2

Table A3: Comparison of classification performance with our discovered concepts before and after concept learning. The results show that it is the concept discovery that mainly contribute to the improvement of classification performance.

	ImageNet	Food-101	CIFAR-100	CIFAR-10	CUB-200	Flowers-102
W/o Concept Learning	21.9	55.7	62.8	60.0	25.5	27.6
W Concept Learning	32.3	73.4	78.2	80.0	44.6	48.5

Table A4: Comparison of interpretability of the discovered concepts before and after concept learning. The results of the intervention accuracy show that both concept discovery and learning parts provide significant improvement for the interpretability of the concepts.

and the concept learning would not affect the classification accuracy; (2) both concept discovery and learning parts provide significant improvement for the interpretability of the concepts according to the intervention accuracy resutls.

F Few-shot Performance about Generalization

To further measure the in-domain and cross-domain generalization of the discovered concepts, we perform few-shot classification with the fine-tuned CLIP model on the unseen categories. The results in Figure A3 show that our discovered concepts can enable much better few-shot classification performance, which suggests that the encoded knowledge of our discovered concepts are more generalizable to unseen categories and domains.

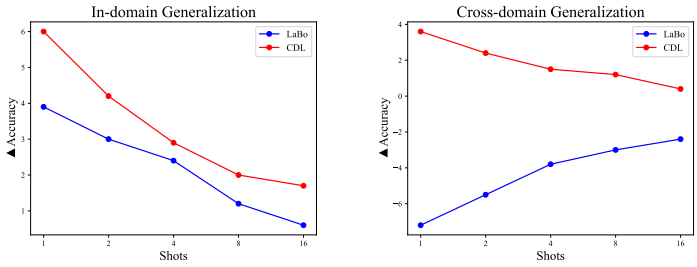

Fig. A3: Few-shot classification performance of in-domain and cross-domain generalization, where ▲ denotes the improvement of fine-tuned CLIP compared to the original CLIP. The comparison between our concepts and LaBo concepts suggest that our concepts can provide better generalization and lead to better performance for few-shot recognition on different categories and domains.