Do Vision-Language Pretrained Models Learn Primitive Concepts?

Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun

Brown University
{tian_yun, usha_bhalla, ellie_pavlick, chensun}@brown.edu

Abstract. Vision-language pretrained models have achieved impressive performance on multimodal reasoning and zero-shot recognition tasks. Many of these VL models are pretrained on unlabeled image and caption pairs from the internet. In this paper, we study whether the notion of primitive concepts, such as color and shape attributes, emerges automatically from these pretrained VL models. We propose to learn compositional derivations that map primitive concept activations into composite concepts, a task which we demonstrate to be straightforward given true primitive concept annotations. This compositional derivation learning (CompDL) framework allows us to quantitively measure the *usefulness* and *interpretability* of the learned derivations, by jointly considering the entire set of candidate primitive concepts. Our study reveals that state-of-the-art VL pretrained models learn primitive concepts that are highly useful as visual descriptors, as demonstrated by their strong performance on fine-grained visual recognition tasks, but those concepts struggle to provide interpretable compositional derivations, which highlights limitations of existing VL models. Code and models will be released.

1 Introduction

Vision-language (VL) models pretrained on raw images and text [42,32,47,37] have revolutionized deep learning in the last few years. They have demonstrated strong transfer learning performance on a wide range of tasks [42,43], even under the zero-shot setup, where text "prompts" [5] are used to specify a task instead of labeled data. Many of these VL models are pretrained on naturally labeled data, such as image and caption pairs from the internet, and thus have the potential to encode "commonsense knowledge" by processing huge quantities of multimodal data. These models often learn to label complex concepts (e.g., bohemian kingfisher) impressively well. However, it is not clear whether they do this by learning to reason over lower-level primitive concepts that humans naturally use to characterize these concepts (e.g., blue wing color, large bill shape, etc). In this paper, we ask whether pretrained VL models capture representations of such primitive concepts "for free" in the course of their pretraining. If they do so, it has important implications for the capacity of models to support compositional generalization, and for humans to interpret the reasoning procedures models undertake.

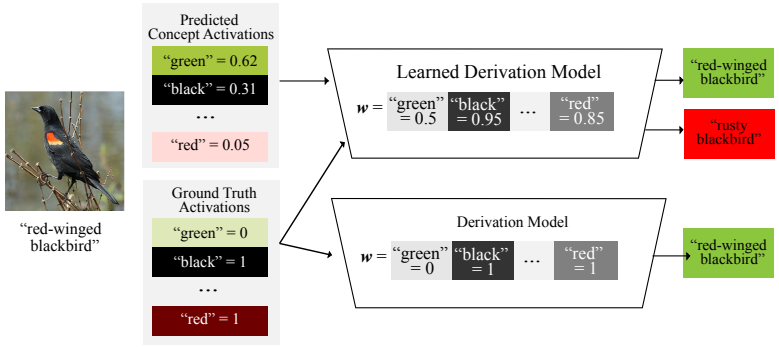

Fig. 1: Illustration of a learned derivation (top) and a groundtruth derivation (bottom), whose task is to map primitive concepts (e.g. colors of a bird) into composite concepts (e.g. *a red-winged blackbird*). While the derivation can be reliably obtained given true concepts, the task can be challenging with predicted concept activations, where the model makes correct predictions for the wrong reasons. We quantify the quality of the learned derivation by its usefulness, which is the empirical performance of composite concept classification, as well as its interpretability, which measures how well the learned model uncovers the groundtruth derivation

Why are we interested in the primitive concepts if a pretrained VL model can directly recognize composite concepts via "prompts"? For one reason, primitive concepts provide greater interpretability of and interaction with models [12,36]. They serve as interpretable "concept bottlenecks" [24], which allows for inspecting which concepts contribute the most to a prediction or having experts to correct the predicted concepts. As another example, primitive concepts allow users to construct a zero-shot classifier by specifying how a concept (e.g. *red apple*) can be derived from primitives (*red, apple*). Existing systems for these applications require manual annotation of the primitives to build corresponding classifiers [26]. Therefore, it would be particularly appealing if pretrained VL models are able to learn the primitives automatically from uncurated data.

In this paper, we propose a computational framework to measure quantitatively how well a VL model has learned primitive concepts. We assume that a composite concept can be derived from a collection of non-overlapping primitive concepts. The key motivation of our proposed framework is that this derivation can also be obtained in a data-driven fashion, by learning a *linear* classifier from the primitive concepts to the composite concepts, which we refer to as a derivation model. When the true primitive and composite concept annotations are available for a dataset, one can reliably learn a derivation model, which we validate experimentally. When only composite concept annotations are available, we ask a pretrained VL model to automatically annotate the primitive concepts via

its zero-shot prompts. Intuitively, if a pretrained VL model does learn primitive concepts, its concept annotations should allow us to learn the same derivation.

We quantify this with two metrics. First, we measure the *usefulness* of a learned derivation model for recognizing composite concepts, in terms of its classification accuracy. Second, we evaluate the *interpretability* of a learned derivation model by passing true primitive concept annotations as its inputs (see Figure 1 middle). We then compute the gap between this classification accuracy and the accuracy achieved by a ground truth derivation model with the same inputs (Figure 1 bottom): The gap is small when the learned derivation correctly maps the true primitive concept annotations to their corresponding composite concepts, as can be achieved by a ground truth derivation model. The derivation model learning framework serves as a proxy benchmark to measure how well pretrained VL models learn primitive concepts, where the *usefulness* and *interpretability* metrics jointly consider the entire set of primitive concepts annotated by a pretrained VL model on a (large) dataset. Our benchmark does not require exhaustive primitive concept labeling on the dataset, since we show ground truth derivations can be learned from a few examples.

We focus our study on three recent state-of-the-art VL models: (CLIP) [42], ViLT [22], and ALBEF [27]. They represent three categories of VL models: no cross-modal fusion (CLIP), early fusion (ViLT), and late fusion (ALBEF). To learn the derivations, we consider tasks where the target labels are composite concepts, such as in fine-grained visual recognition [52] (where the primitives are color, shape of the object parts) and object state recognition (where the primitives are objects and their attributes). Our study reveals both promises and limitations of VL pretrained models. For example, we find, encouragingly, that the primitive concepts recognized by VL models are highly useful for visual recognition tasks, such as few-shot fine-grained visual recognition [50] and zero-shot compositional learning [41]. However, we also observe that the learned derivations are *not* interpretable, indicating their inputs behave more as visual descriptors than actual primitive concept detectors.

To summarize, we make the following contributions: (1) We propose a framework to quantitatively measure how well VL pretrained models learn primitive concepts, via the derivation learning task; (2) We perform extensive quantitative and qualitative studies based on our proposed framework, using a range of recent VL pretrained models; (3) We demonstrate quantitatively that the "primitive concepts" from VL models are highly *useful*, to the extent that a simple linear classifier provides very competitive performance on fine-grained recognition and compositional generalization tasks. However, our analysis also shows that their learned derivations are not *interpretable*, indicating that the existing VL models do *not* learn interpretable primitive concepts. This highlights an important future direction for VL pretraining. Code and models will be released.

2 Related Works

Vision and language pretraining. The fast evolution of deep learning hardware enables researchers to train large deep neural networks on web-scale data [48,9,5]. One data source is vision and language pairs from the internet, such as images and captions [44], or videos and speech transcripts [37,42]. These vision-language datasets are usually not manually annotated. Representations learned by VL models can be transferred to a diverse range of tasks, such as image or video captioning [32,47] and visual question answering [32,4]. These pretrained models can directly recognize composite concepts, such as human actions [47] or object categories [42] with text prompts. To our knowledge, we make the first attempt to study if pretrained VL models learn primitive concepts.

Primitive concepts and their representation. Visual and linguistic concepts are highly compositional and can be represented with primitive concepts. For human activities, the primitives are humans, objects, and their interactions [18]. For objects, the primitives can be parts [14,11] and their attributes [40,25,26]. For sentences, the primitives are often taken to be words and their grammatical relations [46].

To study the relations between the representations of composite and primitive concepts, measurements on compositionality have been proposed [1,20,50]. One common technique is to rely on representation arithmetic, where the representation of a composite concept is expected to be reconstructed by its primitives (e.g. summation [1]). This measurement requires having a fixed set of primitive concepts and knowing the derivation from primitive to composite concepts, both of which might not hold for the composite concepts observed by a pretrained VL model. Instead, our paper assumes that the primitives can be selected *on-the-fly* based on different sets of composite concepts, and the derivation can be learned by a linear classifier.

Using and learning primitive concepts. Primitive concepts have wide applications in machine learning models, such as neural symbolic reasoning [55,36,3,2], few-shot [50] or zero-shot [40,26] visual recognition, building interpretable models [21,12,24], or simply as a visual descriptor [28]. Most of these frameworks require annotations of the primitives to build primitive classifiers before the primitive concepts can be incorporated into the overall framework. In this paper, we are interested in whether the primitive concepts can be recognized by a pretrained VL models via zero-shot prompts, a process that does not require any manual annotations of the concepts, or task-specific *prompt tuning* [53].

3 Method

This section describes our proposed framework to measure quality of the learned primitive concept predicted by VL pretrained models. We first introduce the two stages of our proposed framework, concept prompting and derivation learning. We name our proposed framework *Comp*ositional *D*erivation *L*earning, or *CompDL*. We then discuss our concept prompting setup and our two evaluation benchmarks. An illustration of the overall framework can be found in Figure 2.

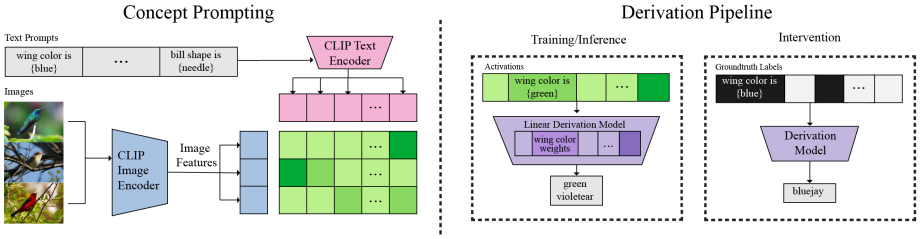

Fig. 2: Illustration of two-step framework, Compositional Derivation Learning (CompDL). On the left is our "Concept Prompting" module that generates concept activations from images, given a set of concepts. On the right is our "Derivation Learning" to train a linear derivation model and test it on predicted (inference) or ground truth activations (intervention)

3.1 Derivation Learning

We denote the process of inferring composite concepts (e.g. bird species) from primitive concepts (e.g. colors and shapes of a bird beak) as *derivation*. In this work, we assume that the primitive concepts are composable, and there exists a true derivation for any composite concept q, given an expressive set of candidate primitive concepts $\mathcal{C}_p$. As illustrated in Figure 1, a true derivation model can be represented in the form of a linear classifier, where the primitive concepts required for the derivation are assigned with $w_i = 1$ (or other positive weights), and the primitive concepts to be ignored are assigned with $w_j = 0$.

We are interested in *derivation learning*: Given a set of paired primitive concept activations $\mathbf{e}$ along with composite concept labels, can we learn a derivation model $\mathcal{H}$ that correctly maps the primitive concepts into the composite concepts, as achieved by the true derivation? Formally, we want to predict a composite concept q with a linear classifier $\mathcal{H}_q(\mathbf{e}) = \mathbf{w}_q^\intercal \cdot \mathbf{e}$, where each element e_i in $\mathbf{e}$ corresponds to how likely a concept p_i presents, and can be given by an oracle as groundtruth $\mathbf{e}^{\text{gt}}$, or predicted by a pretrained VL model $\mathbf{e}^{\text{pred}}$ (see Section 3.2). The classifier can be trained with a contrastive learning objective [42], which pushes the positive pairs of activated primitive concepts and the derivation of their corresponding composite concept to be more similar ($\mathcal{H}_q$ to be higher), and the negative pairs to be less similar. Given a fixed set of mutually-exclusive composite concepts, this objective is the same as Softmax cross-entropy.

The derivation model is regarded as a "proxy" to measure the quality of primitive concepts learned by VL models. We quantify the quality of the learned concepts with two metrics: *usefulness* and *interpretability*. First, the derivation model can be directly evaluated for its *usefulness* as a classifier, i.e. the classification accuracy when evaluated on an unseen test set. Second, we want to measure whether the derivation model is *interpretable*, meaning that the primitive concepts necessary to derive a composite concept are selected by the model and that the learned derivation is not "right for the wrong reason".

Quantifying usefulness. Intuitively, we claim the learned derivation model to be useful if it can correctly map the predicted primitives $\mathbf{e}^{\text{pred}}$ into composite concepts. A useful derivation model should sufficiently learn to compose these primitive concepts such that it would be able to derive unseen composite concepts correctly, even when deriving novel composite concepts. We use classification accuracy in composite concept recognition tasks to quantitatively measure the usefulness of a trained derivation model; however this is not the main contribution of the paper, but rather to validate that the linear classifier adopted by our derivation framework does not come at the cost of performance.

Quantifying interpretability. We assume (and validate experimentally in Section 4.2) that when the ground truth primitive concept activations $\mathbf{e}^{\text{gt}}$ and their corresponding composite concept q^{gt} are available, the derivation from primitive concepts to composite concepts can be reliably learned by a linear derivation model $\mathcal{H}_q^{\text{gt}}(\mathbf{e})$ for any composite concept q.

If pretrained VL models do learn primitive concepts well, their predicted primitives $\mathbf{e}^{\text{pred}}$ should learn a derivation model $\mathcal{H}_q^{\text{pred}}$ *similar* to $\mathcal{H}_q^{\text{gt}}$, based only on pairs of $\mathbf{e}^{\text{pred}}$ and q^{gt}: Namely, $\mathcal{H}_q^{\text{pred}}$ should mostly pick the true primitive concepts necessary to compose q, just like $\mathcal{H}_q^{\text{gt}}$ does. We consider this "similarity" to the true derivation as *interpretability* of a learned derivation model.

To quantitatively measure the interpretability of a derivation model $\mathcal{H}_q^{\text{pred}}$, we compare its performance with the corresponding ground truth derivation $\mathcal{H}_q^{\text{gt}}$. Rather than forcing $\mathcal{H}_q^{\text{pred}}$ to behave exactly as the true derivation for arbitrary inputs $\mathbf{e}$, we relax the constraint and only require it to behave similarly on true primitives $\mathbf{e}^{\text{gt}}$. Formally, we define the interpretability metric as: $\Delta = \texttt{Acc}\left(\mathcal{H}_q^{\text{gt}}(\mathbf{e}^{\text{gt}})\right) - \texttt{Acc}\left(\mathcal{H}_q^{\text{pred}}(\mathbf{e}^{\text{gt}})\right)$, where $\texttt{Acc}(\cdot)$ is the performance evaluation metric such as classification accuracy. Replacing the model inputs with ground truth concepts given by an oracle can be seen as a special case of *model intervention* [24]. An illustration of intervention can be found in Figure 2 (right).

Discussion. A possible alternative to quantitatively measure if VL pretrained models learn primitive concepts is to directly evaluate models' performance when predicting all primitive concepts. However, this evaluation requires exhaustive labeling of primitive concepts on a given image, which is often not available (e.g. an image of "red apple" in MIT-States [10] can also be "ripe" or "round"). Therefore, the derivation model learning and the *usefulness* and *interpretability* metrics serve as proxy benchmarks to measure how well pretrained VL models learn primitive concepts: they offer two aggregated metrics by jointly considering the whole set of primitive concepts, as opposed to measuring individual "performance" on each primitive concepts.

3.2 Concept Prompting

Given the vocabulary of all concepts $c \in \mathcal{C}$, primitive concepts as $p \in \mathcal{C}_p$, and composite concepts as $q \in \mathcal{C}_q$. We have $\mathcal{C}_p \subset \mathcal{C}$, $\mathcal{C}_q \subset \mathcal{C}$, and $\mathcal{C}_p \cap \mathcal{C}_q = \emptyset$. Prior works have worked on datasets with human-annotated images with primitive and composite concepts [52,17]. These primitive concepts can be defined by experts (e.g. the color and shape of body parts to recognize bird species), or can naturally

be inferred from human-annotated composite concepts (e.g. primitive concepts of attribute *ripe* and object *apple* from a label *ripe apple*).

A typical pretrained VL model consists of a text encoder $\mathcal{G}$ and an image encoder $\mathcal{F}$. Parts of the encoders can be shared to fuse vision-language signals. During VL pretraining, a score $s(x, t)$ is computed to measure the "compatibility" between an image x and a sentence t, such as with cosine similarity between the encoded $\mathcal{F}(x)$ and $\mathcal{G}(t)$. The score should be high for the positive pairs of images and their corresponding descriptions and low for the negative pairs.

To obtain $\mathbf{e}^{\text{pred}}$, we rely on the *concept prompting* technique which automatically extracts the concept activations from a pretrained VL model based on the score function $s(x, t)$. For a primitive concept p, we fit it into a template sentence $t(p)$ (for example, "`a photo of [concept]`"). A concept representation can be computed with the score function $e_p = s(x, t(p))$, as illustrated in Figure 2 (left). We can repeat the process for all primitive concepts to obtain the predicted primitive concept activations $\mathbf{e}^{\text{pred}} = [e_1, ..., e_N]$. In appendix, we study the impact of the choice of prompt template $t(\cdot)$.

We explore three types of VL models according to where cross-modal interactions take place: *no cross-modal fusion*, *early fusion*, and *late fusion*. Firstly, CLIP [42] is an example of no cross-modal fusion models. Such models learn $\mathcal{G}$ and $\mathcal{F}$ via contrastive learning objective without having an explicit architecture component to fuse visual and language signals. Secondly, some VL pretraind models [22] adopt early fusion on the low-level text and image features from $\mathcal{G}$ and $\mathcal{F}$. On the contrary, in late fusion VL pretrained models [27], text and image features are encoded separately through $\mathcal{G}$ and $\mathcal{F}$ and these high-level features are combined together via a fusion component.

3.3 Benchmarks

We consider two classification tasks whose labels consist of composite concepts:

Compositional zero-shot learning (CZSL) [17] aims to evaluate if a classifier trained to predict composite concepts generalizes to novel compositions, where each composite concept is a pair of object attribute (*red*) and category (*apple*). It provides a challenging setup for us to evaluate the *usefulness* and *interpretability* of a derivation model.

Given the set of composite concept labels $\mathcal{C}^s$ during training and $\mathcal{C}^t$ during evaluation, CZSL considers the scenario when $\mathcal{C}^s \neq \mathcal{C}^t$, but both share the same set of object attributes and categories. We consider the generalized CZSL setup, where $\mathcal{C}^s$ is a true subset of $\mathcal{C}^t$. We also consider both the "closed-world" version and the "open-world" version of generalized CZSL. In the closed-world setting, $\mathcal{C}^t$ is constraint to be a relatively small set of composite concepts. In the open-world setting, $\mathcal{C}^t$ contains every combination of object attribute and category, and is therefore much more challenging due to the large output space.

Fine-grained visual recognition. We evaluate whether a true derivation model could be learned in a few-shot or full-shot setting for fine-grained visual recognition tasks [52]. Our intuition is that fine-grained concepts could potentially be defined by the details given by primitive concepts, such as shape or color

of bird body parts. To compare with previous work, we also adopt a few-shot learning setup, in which the derivation model needs to discriminate n composite concepts by training on only k images from each composite concept.

4 Experiments

We first introduce our datasets, baselines, and implementation details. We then discuss our experimental results on two zero- and few-shot benchmarks.

4.1 Experimental Setup

Datasets. We conduct experiments on two image classification benchmarks where the target labels are composite concepts. The first dataset is MIT-States [10], which contains 53K images of 115 attributes and 245 objects. Each image is labeled with (object, attribute) tuples. We use the standard splits from [41]. The training split has 30K images of 1262 seen attribute-object compositions, the validation split has 10K images of 300 seen and 300 unseen compositions, and the test split has 13K images of 400 seen and unseen compositions. We report results on the validation and test splits.

The other dataset is Caltech-UCSD Birds-200-2011 (CUB) [52], which contains 11788 photographs of 200 mainly North American bird species. We use the standard training/testing split from [13]. Each image is also annotated with attributes corresponding to 28 different categories, such as throat color, wing shape, etc. There are 312 possible binary attributes in total. We adopt the attribute denoising setup from [24].

Metrics. For the MIT-States dataset, we follow the evaluation protocol from [41] whose goal is to mitigate models' bias on seen compositions. The evaluation protocol has four metrics: (1) Unseen-Seen Area Under the Curve (AUC), (2) best accuracy on data samples of seen compositions (*best seen*), (3) best accuracy on data samples of unseen compositions (*best unseen*), and (4) best harmonic mean (*best HM*) of seen and unseen accuracy (2, 3). For the CUB dataset, we follow the standard practice in n-way k-shot evaluation [45] and report mean accuracy over the 600 sampled tasks run for each setup. In each task, the n classes and k examples are chosen at random.

Implementation details. We choose three recently open-sourced VL models: CLIP [42], ViLT [22], and ALBEF [27] for this study. They represent three categories of VL models: no cross-modal fusion (CLIP), early fusion (ViLT), and late fusion (ALBEF). For CLIP, we use the pretrained CLIP with a ViT-B/32 [10] visual encoder and a transformer-based [51] text encoder. For ViLT, we use ViLT-B/32 pretrained with masked language modeling and image-text matching objectives. For ALBEF, we use ALBEF with a ViT-B/16 [10] visual encoder and a BERT [9] text encoder. This checkpoint is trained on Conceptual 12M dataset [6]. We follow the default setups to compute the concept activations.

For MIT-States, a contrastive objective is used to train the derivation models. We apply two *linear* projection layers on the primitive concept activations and

Table 1: Results of CZSL task on MIT-States in the *closed-world* setting. AUC (%) are computed using precision at k=1,2,3. Best seen and unseen accuracies, and best harmonic mean of the two are reported. Derivation models are learned from ground truth primitive concepts $\mathbf{e}^{\text{gt}}$ (first row), or from VL model concept prompting $\mathbf{e}^{\text{pred}}$. Usefulness is measured when the same predicted concepts $\mathbf{e}^{\text{pred}}$ is used for evaluation (rows in blue). Interpretability is measured as the performance gap from GT derivation when $\mathbf{e}^{\text{gt}}$ is provided as "intervention"

Method	Val AUC			Test AUC			Seen	Unseen	HM
Top k →	1	2	3	1	2	3			
GT Derivation	99.9	99.7	99.7	99.9	99.9	99.9	100.0	100.0	99.9
CLIP-Prim	8.2	17.0	24.1	6.9	15.6	22.8	34.0	27.9	20.4
CLIP-Interv(GT)	32.2	49.1	58.0	30.0	49.3	59.6	52.3	66.0	47.7
CLIP-Interv(GTX)	35.8	53.4	62.9	32.6	53.7	63.2	54.1	68.2	49.4
ViLT-Prim	4.7	11.1	16.9	3.8	9.7	15.2	25.4	20.8	15.1
ViLT-Interv(GT)	8.6	17.1	23.6	4.1	10.2	14.4	21.7	24.8	16.3
ViLT-Interv(GTX)	20.1	33.8	43.1	12.9	25.5	35.1	39.0	41.8	28.7
ALBEF-Prim	7.1	15.6	23.1	5.8	13.9	21.2	32.9	25.1	18.5
ALBEF-Interv(GT)	36.8	52.7	63.6	37.9	57.2	64.8	58.5	71.6	53.0
ALBEF-Interv(GTX)	38.7	56.5	68.3	41.1	61.8	70.6	61.3	73.0	55.8

the text embeddings of target composite concepts respectively, to embed them into a shared space. This approach works with both closed-world and open-world settings of CZSL, and maintains the linear derivation assumption. During training, a subset of composite concepts are supposed to be unknown, so we remove the unseen composite concepts (i.e. $|\mathcal{C}^s| = 1{,}262$ for training; $|\mathcal{C}^t| = 1{,}962$ for closed-world evaluation, and $|\mathcal{C}^t| = 28{,}175$ for open-world evaluation).

For CUB, a logistic regression model is used to train the derivation models since the number of target composite concepts is fixed. We used the default sklearn hyperparameters and $L2$ regularization of 1 for all experiments. We observed that the performances are robust against the choice of logistic regression hyperparameters. Due to annotation inconsistencies, we follow [24] and denoise the primitive concept annotations via majority voting.

We design two sets of prompts to extract attribute/object/pair activations from the VL models. For MIT States, e.g., *"this is {ripe}"*, *"this is {apple}"*, and *"this is {ripe apple}"* are used to produce the corresponding activations. For CUB, we compute attribute and class activations by *"a photo of bird whose {bill shape} is {needle}"* and *"the bird is {bohemian waxwing}"*. These two CUB prompts have different templates since we observe some bird attribute categories can only be understood with the context of birds (e.g. size and shape of birds), and thus we include the language prior "bird" in both prompts. We explore more prompt templates in Appendix and find our observations to be robust.

Table 2: n-way k-shot evaluation on CUB. Mean per-sample accuracy scores are reported. Derivation models are learned from ground truth primitive concepts $\mathbf{e}^{\text{gt}}$ (first row), or from VL model concept prompting $\mathbf{e}^{\text{pred}}$. Usefulness is measured when the same predicted concepts $\mathbf{e}^{\text{pred}}$ is used for evaluation (rows in blue). Interpretability is measured as the performance gap from GT derivation when $\mathbf{e}^{\text{gt}}$ is provided as "intervention"

Method	$n = 5$		$n = 10$		$n = 100$		$n = 200$		Full
$k \rightarrow$	1	5	1	5	1	5	1	5	Shot
GT Derivation	99.8	100.0	99.9	99.9	98.9	98.9	98.0	97.9	98.0
CLIP-Prim	76.5	87.3	60.9	78.0	21.0	43.0	13.6	33.2	52.6
CLIP-Interv(GT)	61.6	66.0	43.1	50.8	9.4	13.4	5.1	7.8	8.0
CLIP-Interv(GTX)	75.4	85.4	59.4	75.2	17.5	32.9	10.9	23.4	32.9
ViLT-Prim	70.0	83.4	54.0	71.3	14.8	28.8	8.9	19.7	40.1
ViLT-Interv(GT)	54.3	60.7	37.7	45.7	6.5	8.6	3.5	4.7	9.2
ViLT-Interv(GTX)	64.9	75.8	48.0	60.0	9.6	15.6	5.3	9.4	19.3
ALBEF-Prim	73.2	85.2	58.3	74.4	17.0	36.3	10.7	27.2	44.6
ALBEF-Interv(GT)	57.0	63.9	39.3	46.5	7.0	9.5	3.7	5.2	7.6
ALBEF-Interv(GTX)	69.9	80.0	52.4	66.3	12.0	24.6	7.0	16.9	30.2

4.2 Interpretability of Learned Derivations

We first explore if a derivation model learned from concept activations is interpretable. We approach this problem by inspecting if the linear classifier learns a true derivation for the composite concepts or it solves the task with some spurious correlations between the activations and the composite concepts. The true derivation is learned from ground truth primitives, while a learned derivation is learned from concept activations predicted by VL models.

When we intervene on the learned derivations in two ways. The first method is intervention with ground truth primitives (*Interv(GT)*), which means we replace the primitive concept activations with the binary ground truth primitives during inference. The comparison between *Interv(GT)* and a corresponding ground truth derivation can reflect how feasible it is to approximate a ground truth derivation from primitive concept activations. The second is intervention only on attributes activated by the ground truth primitives (*Interv(GTX)*). This method is similar to *Interv(GT)*, but all dimensions not activated by the ground truth labels are left as the activations predicted by the VL model. The comparison between *Interv(GTX)* and *Interv(GT)* shows the spurious correlations learned between non-activated primitives and composite concepts.

Table 1 and Table 2 summarize the results on MIT-States and CUB. First of all, the derivation models learned from ground truth primitive concept activations $\mathbf{e}^{\text{gt}}$ and their corresponding composite concept q_{gt} perform almost perfectly (*GT Derivation*). This validates our hypothesis that a true derivation can be

Table 3: Measuring interpretability Δ on MIT-States (Left) and CUB (Right). Δ is the performance gap between GT derivation and Interv(GT). The lower Δ is, the more interpretable the derivation model is

Model	Δ AUC	Seen	Unseen	HM
CLIP	69.9	47.7	34.0	52.2
ViLT	95.8	78.3	75.2	83.6
ALBEF	62.0	41.5	28.4	46.9

Model	$n \rightarrow$	Δ 5	100	200	FS
CLIP		34.0	85.5	90.1	90.0
ViLT		39.3	90.3	93.2	88.8
ALBEF		36.1	89.4	92.7	90.4

learned from $\mathbf{e}^{\text{gt}}$ and q_{gt}. We also observe two consistent patterns on both CZSL and few-shot learning tasks. First, with *Interv(GT)* intervention, all models perform worse than the ground truth derivation (Table 3), which generally reaches perfect performance. This implies that the learned derivation models are not interpretable. Second, by comparing *Interv(GTX)* intervention with *Interv(GT)*, all models increase in performance, indicating that the models utilize activations from the "background" concepts to predict the composite concepts.

When we compare *Prim* with *Interv(GT)*, we see different patterns on MIT-States and CUB. This indicates that the learned derivations of MIT-States are able to identify positive correlations between the composite concepts and their primitives, while the derivations of CUB are not able to identify such positive correlations sufficiently. We attribute this difference to the gap of the complexity in these two tasks, as a derivation model for MIT-States only needs to capture two primitives (an attribute and an object) for each composite concept, while a derivation model for CUB must capture multiple primitives for each bird class.

4.3 Qualitative Analysis of Derivations

In Figure 3, we carry out a qualitative analysis to inspect the top three primitive concepts assigned with highest weights by the linear derivation model.

We observe that, for MIT States, the top three most activated composite concepts are mostly relevant to the attribute-object pairs (e.g. *old book* and *open book* for a *thick book*), indicating that the model learns the correlations between composite concepts in training set and new composite concepts. We also notice that the predicted composite concepts of *ripe banana* all identify object *banana*, indicating that the model is able to identify objects correctly but that attribute classification is more challenging. Finally, the model sometimes labels the images with correct concepts (e.g. *open book* for *thick book*), but the labels strictly consider one single concept. This is a limitation and also a challenge of the annotations of compositional learning evaluation datasets.

For CUB, we observe that the top three most activated attributes of the shown bird classes are mostly but not always relevant to the corresponding bird classes. The derivation model seems to give the most importance to attributes related to the overall color of the bird, even if specific parts of birds have different color from the overall color of the bird (e.g. the bill color of the *Indigo Bunting*).

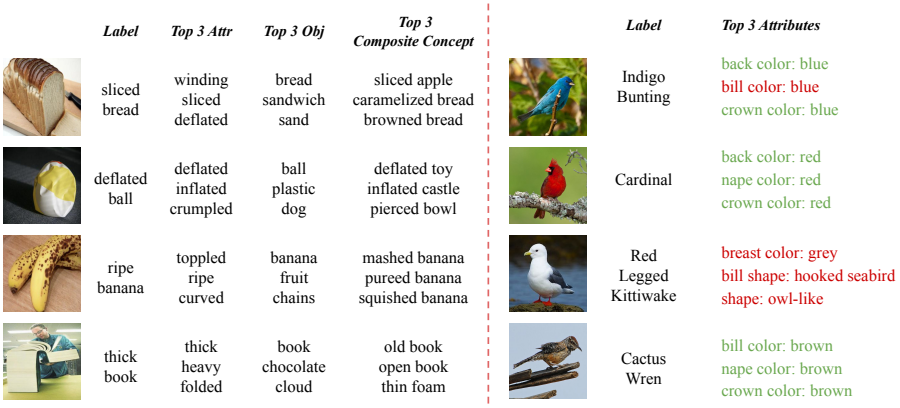

Fig. 3: Qualitative analysis on the unseen composite concepts in the test split of MIT-States (Left) and CUB (Right). Left: We show the most important primitive and composite concepts according to the weights of the learned derivation model. Right: We show the most important attributes for each bird class, given by the weights of the learned derivation model. Attributes in green are correct attributes for the corresponding bird class and attributes in red are incorrect. The images are randomly selected for illustration purposes

4.4 Usefulness of Derivations

We look into the classification performance based on the primitive concept activations. The primitive concepts are deemed useful if the derivation learned on top of them achieves good classification performance. To understand if VL models learn composite concepts better than primitive concepts, we follow the same concept prompting and derivation learning process but ask a VL model to directly predict composite concepts, a variant we name as *Comp*. We also try to combine *All* concepts, which measures whether primitive and composite concepts encode complementary information.

Table 4 shows the results on MIT-States in closed-world setting. We observe that the zero-shot CLIP already achieves on par or better performance than the previous approaches, which confirms the observation that CLIP learns composite concepts. We also observe that CLIP outperforms ALBEF, and ALBEF outperforms ViLT, which matches to the order of their pretraining data size. We see that derivations trained with primitive concepts always outperform those trained with composite concepts, indicating that for the set of primitive and composite concepts that appear in MIT-States, VL models appears to capture the composite concepts better than the primitive ones. Interestingly, combining the two sets of concepts together (e.g. CompDP-All), we observe very similar performance as the derivations trained with composite concepts. This demonstrates that the information encoded in the primitive and composite concepts is likely to be not complementary.

Table 4: Results of generalized CZSL on MIT-States in the closed-world setting for all CompDL variants. We observe that CLIP with CompDL performs the best among the selected VL models, and it outperforms previous methods

Method	Top k →	Val AUC 1	Val AUC 2	Val AUC 3	Test AUC 1	Test AUC 2	Test AUC 3	Seen	Unseen	HM
AOP [39]		2.5	6.2	10.1	1.6	4.7	7.6	14.3	17.4	9.9
LE+ [39]		3.0	7.6	12.2	2.0	5.6	9.4	15.0	20.1	10.7
TMN [41]		3.5	8.1	12.4	2.9	7.1	11.5	20.2	20.1	13.0
SymNet [31]		4.3	9.8	14.8	3.0	7.6	12.3	24.4	25.2	16.1
CompCos [34]		/	/	/	4.5	/	/	25.3	24.6	16.4
CGE_{ff} [38]		6.8	/	/	5.1	/	/	28.7	25.3	17.2
CLIP (zero-shot)		5.8	13.0	19.1	5.5	12.7	18.8	25.0	30.5	18.3
CLIP-Prim		8.2	17.0	24.1	6.9	15.6	22.8	34.0	27.9	20.4
CLIP-Comp		8.2	17.6	25.1	7.3	16.1	23.5	35.5	28.2	21.2
CLIP-All		7.8	16.2	23.3	6.1	14.3	21.2	32.8	25.5	19.1
ViLT-Prim		4.7	11.1	16.9	3.8	9.7	15.2	25.4	20.8	15.1
ViLT-Comp		5.5	12.6	18.8	4.5	11.0	16.8	28.4	22.4	16.5
ViLT-All		5.6	12.5	18.6	4.2	10.4	16.1	28.3	21.1	15.8
ALBEF-Prim		7.1	15.6	23.1	5.8	13.9	21.2	32.9	25.1	18.5
ALBEF-Comp		7.5	16.6	23.7	6.5	14.6	21.8	32.8	26.9	20.0
ALBEF-All		7.3	15.9	23.3	6.6	15.0	21.9	32.4	27.1	20.2

Table 5: Results of generalized CZSL on MIT-States in the open-world setting for CLIP withCompDL. We observe that CLIP with CompDL outperforms previous methods on AUC , but it does not necessarily generalize well to unseen composite concepts

Method	Top k →	Val AUC 1	Val AUC 2	Val AUC 3	Test AUC 1	Test AUC 2	Test AUC 3	Seen	Unseen	HM
CompCos [34]		/	/	/	1.6	/	/	25.4	10.0	8.9
CLIP-Prim		2.6	6.0	9.1	2.1	5.2	8.0	33.7	9.9	10.8
CLIP-Comp		2.8	6.3	9.4	2.4	5.8	8.6	33.6	10.9	11.8
CLIP-All		2.2	5.2	8.0	1.8	4.3	6.7	31.3	8.5	10.1

Table 5 shows the results on MIT-States in open-world setting. We mainly focus on CLIP due to its superior performance in the closed-world setting. With a linear derivation model trained on top of concept activations predicted by CLIP, all learned derivations outperform the previous state-of-the-art approach. However, most improvements come from better learning the seen concepts in the

Table 6: Results of 5-way k-shot learning on CUB. Mean accuracy across 600 sampled tasks for CompDL-All are reported for CLIP, ViLT, and ALBEF. CompDL provides competitive performance despite its simplicity

Method $n = 5$	1-shot	5-shot
FEAT [54]	73.3	85.8
DeepEMD [56]	75.7	88.7
RENet [19]	79.5	91.1
S2M2_R [35]	80.7	90.9
PEM_b-NCM [15]	80.8	91.5
CLIP-All	80.6	91.5
ViLT-All	66.6	76.8
ALBEF-All	72.6	83.5

training split but not necessarily from better generalizability to the unseen concepts, implying that the learned derivations are still distracted by the infeasible attribute-object pairs.

Table 6 shows the results on CUB in 5-way k-shot learning setting. We report the results on derivations learned by the combinations of primitive and composite concept activations (CompDP-All) due to its superior empirical performance. We observe that the learned derivation with CLIP again performs the best among the selected VL models, and it performs competitively against the state of the art in the 1-shot and 5-shot setting.

5 Conclusions

In this paper, we have proposed a framework for measuring how well vision-language pretrained models can learn primitive concepts, in terms of usefulness and interpretability. By conducting extensive experiments with recent VL pretrained models - CLIP, ViLT, and ALBEF - we observe that these models are able to learn primitive concepts that are useful for visual recognition tasks, but the learned derivations from primitive concepts to composite concepts are often not interpretable. Our study suggests that despite the strong performance of VL pretrained models for visual recognition tasks, more research is needed to improve their ability to better capture primitive concepts. We anticipate our framework serve as a meaningful step towards understanding the working mechanisms (and potentially biases) of VL pretrained models.

References

1. Andreas, J.: Measuring compositionality in representation learning. In: ICLR (2019)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705 (2016)
3. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: CVPR (2016)
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
6. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 3558–3568. Computer Vision Foundation / IEEE (2021)
7. Chen, Y., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: universal image-text representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX. Lecture Notes in Computer Science, vol. 12375, pp. 104–120. Springer (2020)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
12. Ghorbani, A., Wexler, J., Zou, J., Kim, B.: Towards automatic concept-based explanations. arXiv preprint arXiv:1902.03129 (2019)
13. Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C.D., Hodas, N.O.: Few-shot learning with metric-agnostic conditional embeddings. arXiv preprint arXiv:1802.04376 (2018)
14. Hoffman, D.D., Richards, W.A.: Parts of recognition. Cognition (1984)
15. Hu, Y., Gripon, V., Pateux, S.: Squeezing backbone feature distributions to the max for efficient few-shot learning. arXiv preprint arXiv:2110.09446 (2021)
16. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. CoRR **abs/2004.00849** (2020)
17. Isola, P., Lim, J.J., Adelson, E.H.: Discovering states and transformations in image collections. In: CVPR (2015)
18. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: CVPR (2020)
19. Kang, D., Kwon, H., Min, J., Cho, M.: Relational embedding for few-shot classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8822–8833 (2021)

20. Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., et al.: Measuring compositional generalization: A comprehensive method on realistic data. arXiv preprint arXiv:1912.09713 (2019)
21. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: ICML (2018)
22. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 5583–5594. PMLR (2021)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
24. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML (2020)
25. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
26. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
27. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021)
28. Li, L.J., Su, H., Lim, Y., Fei-Fei, L.: Object bank: An object-level image representation for high-level visual recognition. IJCV (2014)
29. Li, L.H., Yatskar, M., Yin, D., Hsieh, C., Chang, K.: What does BERT with vision look at? In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 5265–5275. Association for Computational Linguistics (2020)
30. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX. Lecture Notes in Computer Science, vol. 12375, pp. 121–137. Springer (2020)
31. Li, Y., Xu, Y., Mao, X., Lu, C.: Symmetry and group in attribute-object compositions. In: CVPR (2020)
32. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265 (2019)
33. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 13–23 (2019)
34. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: CVPR (2021)

35. Mangla, P., Kumari, N., Sinha, A., Singh, M., Krishnamurthy, B., Balasubramanian, V.N.: Charting the right manifold: Manifold mixup for few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2218–2227 (2020)
36. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. arXiv preprint arXiv:1904.12584 (2019)
37. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: CVPR (2020)
38. Naeem, M.F., Xian, Y., Tombari, F., Akata, Z.: Learning graph embeddings for compositional zero-shot learning. In: CVPR (2021)
39. Nagarajan, T., Grauman, K.: Attributes as operators: Factorizing unseen attribute-object compositions. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV (2018)
40. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (2011)
41. Purushwalkam, S., Nickel, M., Gupta, A., Ranzato, M.: Task-driven modular networks for zero-shot compositional learning. In: ICCV (2019)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
43. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092 (2021)
44. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
45. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. arXiv preprint arXiv:1703.05175 (2017)
46. Socher, R., Bauer, J., Manning, C.D., Ng, A.Y.: Parsing with compositional vector grammars. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 455–465 (2013)
47. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
48. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV (2017)
49. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5099–5110. Association for Computational Linguistics (2019)
50. Tokmakov, P., Wang, Y.X., Hebert, M.: Learning compositional representations for few-shot recognition. In: ICCV (2019)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
52. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)

53. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
54. Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8808–8817 (2020)
55. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.B.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. arXiv preprint arXiv:1810.02338 (2018)
56. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12203–12213 (2020)
57. 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: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 5579–5588. Computer Vision Foundation / IEEE (2021)

6 Appendix

6.1 Ablation Study of Prompts

We carry out an ablation study to explore the impact of prompts. Table A1 shows the prompts collected from CLIP. This set of 7 prompt templates is the best performing subset on ImageNet [8] over 80 prompt templates.

Different Prompt Templates We first explore how the *usefulness* and *interpretability* of learned derivations vary with different prompt templates. We perform concept prompting based on each prompt in Table A1 for both primitive and composite concepts. We then train $(7 \times 2 =)$ 14 derivation models in total. We use these models to investigate how robust our observations in the main paper are.

Tables A2 and A3 show the results for *usefulness* as measured on MIT-States in closed-world and open-world settings respectively. We observe that the aggregated results have slightly higher mean than what were reported in the main paper, and the standard deviations are small. This confirms that our observations are robust and the choice of prompts does not matter much for usefulness of the learned derivations. The gains over the best performing baselines are significant considering the low standard deviations in both settings.

Table A4 shows the results with or without intervention, a protocol we use to measure *interpretability* of the learned derivations. We observe that the results with intervention generally exhibit higher variances, indicating that the choice of prompts matters more for the interpretability of the learned derivations. However, the aggregated results are still far from those of the ground truth derivation.

Using More Prompts We further examine the impact of using more prompts. Following what CLIP does to combine prompts, we combine text embeddings from multiple prompts together. We vary the number of prompts to be 1, 4, or 7, and train their corresponding derivation models. In Table A5, we observe a positive correlation between the number of prompts used to generate the concept activations and the interpretability of the learned derivations. However, the impact on usefulness is marginal.

6.2 Additional Experimental Details

Concept Prompting with ViLT and ALBEF ViLT [22] is a single-stream vision-and-language (VL) pretrained model that simplies the visual embedding pipeline. Different from other VL models that use region features [33,29,49,7,30,57] or grid features [16], ViLT directly applies an linear projection layer on the patches of input images to encode these images. The model takes a concatenated sequence of a pair of text sequence and image sequence (of patches) as

Table A1: All seven prompt templates we use

Prompt ID	Prompt
1	itap of a {}
2	a bad photo of the {}
3	a origami {}
4	a photo of the large {}
5	a {} in a video game
6	art of the {}
7	a photo of the small {}

Table A2: Aggregated results (mean and standard deviation) over derivation models trained on seven prompt templates on MIT-States in closed-world setting

Template Source	Top k →	Val AUC 1	Val AUC 2	Val AUC 3	Test AUC 1	Test AUC 2	Test AUC 3	Seen	Unseen	HM
	CGE_{ff}	6.8	/	/	5.1	/	/	28.7	25.3	17.2
Single (main paper)	CLIP-Prim	8.2	17.0	24.1	6.9	15.6	22.8	34.0	27.9	20.4
	CLIP-Comp	8.2	17.6	25.1	7.3	16.1	23.5	35.5	28.2	21.2
Aggregated	CLIP-Prim	8.7 ± 0.2	18.1 ± 0.3	25.5 ± 0.2	7.1 ± 0.2	16.2 ± 0.1	23.7 ± 0.2	35.1 ± 0.5	27.7 ± 0.3	20.7 ± 0.4
	CLIP-Comp	8.3 ± 0.3	17.3 ± 0.5	24.9 ± 0.5	7.0 ± 0.2	15.9 ± 0.4	23.2 ± 0.5	34.3 ± 0.6	28.0 ± 0.8	20.6 ± 0.4

Table A3: Aggregated results (mean and standard deviation) over derivation models trained on seven prompt templates on MIT-States in open-world setting

Template Source	Top k →	Val AUC 1	Val AUC 2	Val AUC 3	Test AUC 1	Test AUC 2	Test AUC 3	Seen	Unseen	HM
	CompCos	-	-	-	1.6	-	-	25.4	10.0	8.9
Single (main paper)	CLIP-Prim	2.6	6.0	9.1	2.1	5.2	8.0	33.7	9.9	10.8
	CLIP-Comp	2.8	6.3	9.4	2.4	5.8	8.6	33.6	10.9	11.8
Aggregated	CLIP-Prim	2.9 ± 0.1	6.5 ± 0.2	9.7 ± 0.2	2.2 ± 0.1	5.3 ± 0.2	8.0 ± 0.2	34.0 ± 0.7	10.0 ± 0.3	11.3 ± 0.2
	CLIP-Comp	2.6 ± 0.2	6.1 ± 0.4	9.1 ± 0.5	2.2 ± 0.1	5.2 ± 0.3	7.9 ± 0.4	32.8 ± 0.9	10.1 ± 0.5	11.0 ± 0.3

input, and is trained with image-text matching (ITM), masked language modeling (MLM), and word patch alignment objectives.

ALBEF [27] is a dual-stream VL pretrained model that learns a text encoder, an image encoder, and a multimodal encoder. The text input and image input will be passed into their corresponding encoders, and the multimodal encoder has a cross-attention layer that fuses the encoded text input and encoded image input together. ALBEF is trained with image-text contrastive learning (ITC), image-text matching (ITM), and masked language modeling (MLM) objectives.

For ViLT and ALBEF, we introduce the concept prompting pipeline using primitive concepts as an example. We have a set of primitive concepts $\{p_1, ..., p_N\}$ and an image x. To obtain a concept activation $\mathbf{e}^{\text{pred}}$ of image x, we extract the ITM score for each concept p_i and treat each ITM score as e_i. We then repeat the

Table A4: Aggregated results (mean and standard deviation) over derivation models trained on seven prompt templates on MIT-States in closed-world setting. We perform intervention to measure interpretability

Template Source	Top k →	Val AUC 1	Val AUC 2	Val AUC 3	Test AUC 1	Test AUC 2	Test AUC 3	Seen	Unseen	HM
	GT Derivation	99.9	99.7	99.7	99.9	99.9	99.9	100	100	99.9
Single (main paper)	CLIP-Prim	8.2	17.0	24.1	6.9	15.6	22.8	34.0	27.9	20.4
	CLIP-Interv(GT)	32.2	49.1	58.0	30.0	49.3	59.6	52.3	66.0	47.7
	CLIP-Interv(GTX)	35.8	53.4	62.9	32.6	53.7	63.2	54.1	68.2	49.4
Aggregated	CLIP-Prim	8.7 ± 0.2	18.1 ± 0.3	25.5 ± 0.2	7.1 ± 0.2	16.2 ± 0.1	23.7 ± 0.2	35.1 ± 0.5	27.7 ± 0.3	20.7 ± 0.4
	CLIP-Interv(GT)	39.3 ± 3.9	58.6 ± 3.5	68.6 ± 3.4	39.9 ± 3.0	58.1 ± 3.1	68.4 ± 2.6	59.2 ± 2.4	72.7 ± 2.4	55.7 ± 2.6
	CLIP-Interv(GTX)	43.6 ± 4.3	64.0 ± 4.3	73.3 ± 4.2	44.0 ± 3.4	63.7 ± 3.5	73.3 ± 3.1	62.8 ± 3.2	75.2 ± 2.1	59.0 ± 2.9

Table A5: The impact of combining activations from multiple prompt templates on MIT-States in closed-world setting

Template Source	Top k →	Val AUC 1	Val AUC 2	Val AUC 3	Test AUC 1	Test AUC 2	Test AUC 3	Seen	Unseen	HM
	GT Derivation	99.9	99.7	99.7	99.9	99.9	99.9	100	100	99.9
CLIP w/ 1 prompt	CLIP-Prim	8.2	17.4	24.2	7.1	16.3	24.0	35.0	28.2	20.8
	CLIP-Interv(GT)	43.6	63.1	73.0	40.9	59.8	71.0	59.0	75.1	56.3
	CLIP-Interv(GTX)	47.2	68.4	77.3	45.4	64.4	75.2	63.4	76.9	60.1
CLIP w/ 4 prompts	CLIP-Prim	8.8	18.5	25.7	7.2	16.5	24.1	35.3	28.3	21.0
	CLIP-Interv(GT)	45.9	64.7	75.2	46.3	65.5	74.9	62.7	77.7	62.1
	CLIP-Interv(GTX)	49.7	70.4	77.4	50.1	69.8	78.3	66.2	79.5	65.2
CLIP w/ 7 prompts	CLIP-Prim	8.9	18.5	25.8	7.2	16.1	24.0	35.3	28.2	20.9
	CLIP-Interv(GT)	45.8	65.7	76.7	47.7	65.0	76.1	65.0	77.9	62.4
	CLIP-Interv(GTX)	51.5	72.1	80.7	51.1	70.1	80.2	67.3	79.8	65.4

process for all primitive concepts to obtain $\mathbf{e}^{\text{pred}} = [e_1, ..., e_N]$. We notice that computing concept activations for ViLT is significantly time-consuming due to its single-stream model input format - a concatenated sequence of text and image inputs. We cannot precompute any text or image features, and thus need to pass the concatenated sequences for each single ITM score. For dual-stream models (CLIP[42] and ALBEF), we directly cache the encoded images and encoded primitive concepts, and then compute the concept activations based on these precomputed features.

Metrics For MIT-States [10] dataset, we follow the evaluation protocol from [41] and report four metrics: Unseen-Seen Area Under the Curve (AUC), (2) best accuracy on data samples of seen compositions (*best seen*), (3) best accuracy on data samples of unseen compositions (*best unseen*), and (4) best harmonic mean (*best HM*) of seen accuracy (2) and unseen accuracy (3).

Since a derivation is trained only on seen composite concepts, the seen concepts can evaluate much better or worse than the unseen ones on validation and test splits. We follow [41] and apply calibration biases, which are scalars to be added to the predicted scores of unseen concepts. Specifically, with a larger bias, the accuracy of unseen concepts tends to increase, while that of seen concepts tends to decrease. We vary the values of the calibration bias, and compute a list of accuracy scores of seen concepts and another list for unseen concepts. *AUC* is the area under the curve of unseen accuracy and seen accuracy; *Best seen* and *best unseen* are the best accuracy in the lists of accuracy scores of seen and unseen concepts. *Best harmonic mean* is the highest harmonic mean computed by

$$HM = 2 \cdot \frac{ACC_{seen} \cdot ACC_{unseen}}{ACC_{seen} + ACC_{unseen}} \tag{1}$$

Baselines We compare our approach for MIT-States against several state-of-the-art baselines in CZSL. LabelEmbed+ (LE+) [39] maps pair embeddings, combinations of word embeddings of attribute and object, and image embeddings into a joint semantic space using two separate MLP. Attribute as Operators (AOP) [39] regards attributes as linear transformations and the transformed object word embeddings are taken as composition embeddings. Task-Modular Modular Networks (TMN) [41] trains a feature extraction model and a gating model which modifies the classifier based on the given attribute-object pairs. SymNet [31] utilizes the symmetry properties of attribute-based transformation to learn object embeddings. Compositional Cosine Logits (CompCos) [34] trains a linear layer as a composition function that projects a pair of attribute and object embeddings to a compositional space. Compositional Graph Embedding (CGE) [38] learns the dependency structure of attributes, objects, and their compositions using a Graph Convolutional Network (GCN) [23]. Since the proposed CGE is trained end-to-end and it finetunes its image encoder, we compare with the reported baseline, $\text{CompCos}_{\text{ff}}$, for MIT-States in closed-world setting, and with CompCos for MIT-States in open-world setting. Note that these models are task-specific supervised, whereas our approach only needs to learn derivation models and is capable of performing the task in a zero-shot fashion.

We compare our CUB results to multiple state-of-the-art baselines. FEAT [54] adapts instance embeddings to the target classification task with a set-to-set Transoformer function. DeepEMD [56] uses the Earth Mover's Distance as a metric to compute structural distances between dense image representations to determine image relevance during training. RENet [19] combines self-correlational representation and cross-correlational attention to learn relational embeddings. S2M2 [35] regularizes feature manifolds via Manifold Mixup by focusing on learning a general purpose representation robust to changes in data distribution. PEM_b-NCM [15] aims at processing feature vectors to be closer to Gaussian-like distributions to boost transfer learning accuracy.

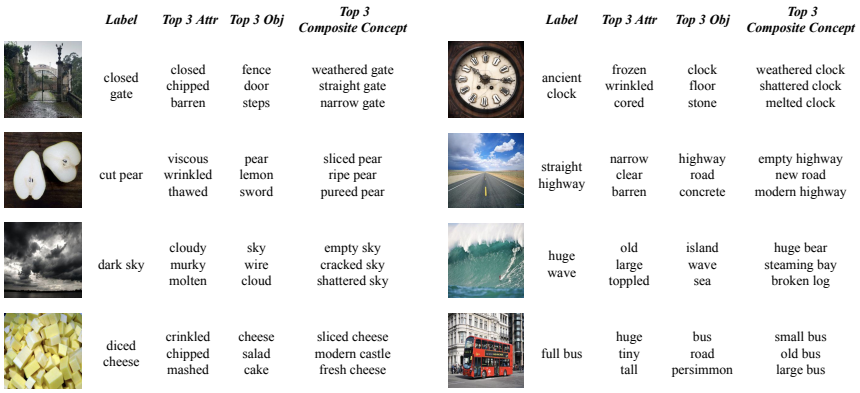

Fig. A1: Additional qualitative analysis on the unseen composite concepts in the test split of MIT-States. We show the ground truth label (first column), the most important primitives (top 3 attr, top 3 obj) and composite concept (last column) according to the weights of the learned derivations. The top concepts shown here are selected for the individual labels based on their corresponding derivation models. The images are randomly selected and for illustration purposes

Fig. A2: Additional qualitative analysis on the 5-way, 5-shot CUB task (each column corresponds to one task). We show the top 3 attributes the learned derivation model. Attributes are extracted from the weights of the linear model for each bird class. Attributes corresponding to activated ground truth concepts are highlighted in green. The images are randomly selected for illustration purposes

6.3 Additional Visualizations

We show additional visualizations on MIT-States and CUB in Figures A1 and A2 respectively.