Towards Compositionality in Concept Learning

Adam Stein [1] **Aaditya Naik** [1] **Yinjun Wu** [2] **Mayur Naik** [1] **Eric Wong** [1]

Abstract

Concept-based interpretability methods offer a lens into the internals of foundation models by decomposing their embeddings into high-level concepts. These concept representations are most useful when they are *compositional*, meaning that the individual concepts compose to explain the full sample. We show that existing unsupervised concept extraction methods find concepts which are not compositional. To automatically discover compositional concept representations, we identify two salient properties of such representations, and propose Compositional Concept Extraction (CCE) for finding concepts which obey these properties. We evaluate CCE on five different datasets over image and text data. Our evaluation shows that CCE finds more compositional concept representations than baselines and yields better accuracy on four downstream classification tasks. [1]

1. Introduction

Foundation models continue to enable impressive performance gains across a variety of domains, tasks, and data modalities (Srivastava et al., 2023). However, their black-box nature severely limits the ability to debug, monitor, control, and trust them (Turpin et al., 2024; Tamkin et al., 2023; Schaeffer et al., 2024).

Concept-based explanations (Kim et al., 2018; Zhou et al., 2018) are a promising approach that seeks to explain a model's behavior using individual concepts such as object attributes (e.g. `striped`) or linguistic sentiment (e.g. `happiness`). Decomposing a model's learned representation can derive these concepts. For instance, a model's embedding of a dog image may decompose into the sum of concept vectors representing its fur, snout, and tail.

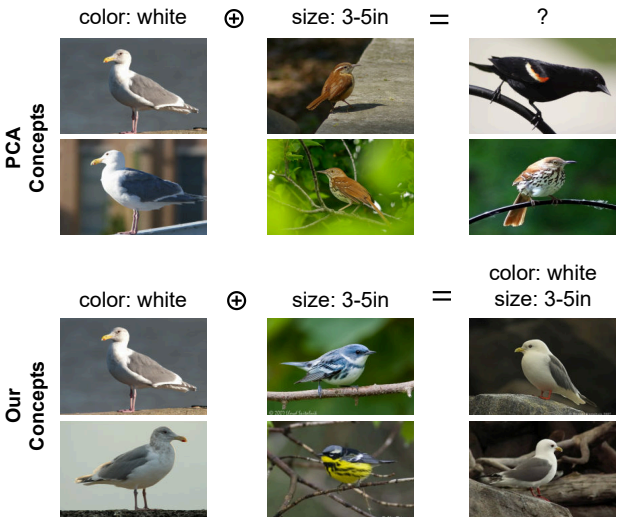


Figure 1. We illustrate the issue of concept compositionality with respect to concepts extracted from the embeddings of the CLIP model over the CUB dataset. Specifically, we visualize the concepts `white birds` and `small birds` learned by PCA (Zou et al., 2023a) and CCE along with their compositions. We show the top two images that best represent each concept. Ideally, composing the `white birds` and `small birds` concepts should result in a concept representing small white birds. This is not the case with the concepts learned by PCA. On the other hand, the concepts extracted by CCE are composable, as shown by the images of small white birds that best represent the resulting concept.

Existing works based on methods such as PCA (Zou et al., 2023a) or KMeans (Ghorbani et al., 2019) extract such concept vectors reasonably well for basic concepts. For instance, Figure 1 shows images from the CUB (Wah et al., 2011) dataset containing concepts extracted by PCA from the CLIP (Radford et al., 2021) model. These techniques are able to correctly extract the representations of concepts like `white birds` and `small birds`, however, composing them by adding their representations does not yield the representation of the concept of `small white birds`.

The *compositionality* of concepts is vital for several use cases. First, model predictions can be explained by combining concepts (Abid et al., 2022). Compositional concepts also allow for editing fine-grained model behavior, like im-

[1]Department of Computer and Information Science, University of Pennsylvania, Pennsylvania, USA [2]School of Computer Science, Peking University, Beijing, China. Correspondence to: Adam Stein <steinad@seas.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Code and data are available at https://github.com/adaminsky/compositional_concepts.

proving the truthfulness of an LLM without compromising other behaviors (Zou et al., 2023a). Models can also be trained to compose basic concepts for new tasks, e.g. using concepts for beak shapes, wing colors, and environments to classify bird species (Yuksekgonul et al., 2023).

In this paper, we study the unsupervised extraction of compositional concepts. Existing work does not directly evaluate the compositionality of extracted concepts, but rather focuses on the individual concept representations. We therefore evaluate the compositionality of concepts extracted by existing unsupervised approaches.

For this purpose, we first validate the compositionality of ground-truth representations of concepts in controlled settings. We observe that concepts can be grouped into *attributes*, where each attribute consists of concepts over some common property, such as the color of objects or the shape of objects. Concepts from different attributes (e.g. `blue` and `cube`) can be composed, while those from the same attribute (e.g. `red` and `green`) cannot. We also observe that the concepts from different attributes are roughly orthogonal, while those from the same attribute are not. We prove in a generalized setting that these properties are crucial for the compositionality of concepts. Since existing approaches do not enforce these properties, they often extract non-composable concept representations.

To extract compositional concepts in an unsupervised manner, we propose Compositional Concept Extraction (CCE). Our key insight is to search for entire subspaces of concepts at once instead of individual concepts, allowing CCE to enforce the aforementioned properties of compositional concepts. We show that CCE recovers the representation of known compositional concepts better than existing approaches, can discover compositional concepts in existing image and text datasets, and the discovered concepts improve downstream classification accuracy.

We thus summarize the contributions of our paper:

- We study concept-based explanations of foundation models from the lens of compositionality—a property desirable for many use-cases. We observe that concept representations extracted by state-of-the-art methods fail to compose, and set out to remedy this problem.
- We validate that models can in fact represent concepts compositionally in embedding space. We identify two salient properties of compositional concept representations that existing methods fail to satisfy.
- We prove in a generalized setting that the identified properties are necessary for compositionality. We present a novel method called Compositional Concept Extraction (CCE) that guarantees to yield concept representations that satisfy these properties by construction.
- We demonstrate that CCE extracts more compositional concepts than baselines on vision and language datasets, and they improve downstream performance.

2. Concepts and Compositionality

Concept Representations. In machine learning, concepts are symbols that are assigned some human-interpretable meaning, often used to explain predictions made by models.

A concept extractor E extracts concepts from the intermediate representation of some pretrained model M over a dataset D. $E(M, D)$ thus yields a set of *concept vectors* representing the concepts $C = \{c_1, \ldots, c_i\}$. Concept vectors are denoted as $R(c)$, where $R : \mathbb{C} \to \mathbb{R}^d$ is the concept representation function, $\mathbb{C}$ is the set of all possible concepts, and $\mathbb{R}^d$ is an embedding space in some dimension d. The set of extracted concepts C can be grouped into mutually exclusive *attributes* $A_1, \ldots A_k$ each containing concepts about some common property such that $C = \bigcup_{i=1}^{k} A_i$.

To measure the presence (or degree of expression) of a concept in a sample's embedding, we borrow the following definition of concept score from (Yeh et al., 2020).

Definition 2.1. (Concept Score) For a concept $c \in \mathbb{C}$ and concept representation function $R : \mathbb{C} \to \mathbb{R}^d$, a sample embedding $z \in \mathbb{R}^d$ has *concept score* $s(z, c) = S_{\cos}(z, R(c))$ where $S_{\cos}$ is the *cosine similarity* function.

Existing work makes use of concept scores to quantify the presence of concepts on a per-sample basis. This has uses in several applications, such as creating concept bottleneck models where a sample's embedding is converted to concept scores used for classification (Yuksekgonul et al., 2023), and sorting samples by a concept (Kim et al., 2018).

Compositionality. Following work on compositional representations (Andreas, 2019) and pretrained embeddings (Trager et al., 2023), we define the compositionality of concept representations.

Definition 2.2. (Compositional Concept Representations) For concepts $c_i, c_j \in \mathbb{C}$, the concept representation $R : \mathbb{C} \to \mathbb{R}^d$ is compositional if for some $w_{c_i}, w_{c_j} \in \mathbb{R}^+$,

$$R(c_i \cup c_j) = w_{c_i} R(c_i) + w_{c_j} R(c_j).$$

In other words, the representation of the composition of concepts corresponds to the weighted sum of the individual concept vectors in the embedding space.

Furthermore, concept scores for the concepts satisfying Definition 2.2 also behave compositionally, since each concept score quantifies the presence of that concept in a sample.

Lemma 2.3. *For compositional concepts $c_i, c_j \in \mathbb{C}$, the concept score of their composition $c_k = c_i \cup c_j$ over a sample embedding $z \in \mathbb{R}^d$ is the composition of the concept scores of c_i and c_j, weighted by $w_{c_i}, w_{c_j} \in \mathbb{R}^+$:*

$$s(z, c_k) = w_{c_i} s(z, c_i) + w_{c_j} s(z, c_j).$$

Since concept scores are used for several downstream tasks discussed above, this property about the compositionality of concept scores can simplify such tasks and improve the overall performance on them.

Besides finding compositional concepts, we also want to explain embeddings based on the concepts which compose it. Prior work also performs a decomposition into a sum of concept representations (Zhou et al., 2018), but we modify the definition of such a decomposition so that a sample embedding is composed of only the concept representations that are truly present for the sample.

Definition 2.4. (Concept-based Decomposition) Consider a sample that is associated with a set of concepts $C \subseteq \mathbb{C}$, such that each attribute $A_i \subseteq C$ contains exactly one concept. A concept representation $R : \mathbb{C} \to \mathbb{R}^d$ decomposes that sample's embedding $z_i \in \mathbb{R}^d$ if it can be expressed as the weighted sum of the sample's associated concepts:

$$z_i = \sum_{c \in C} \lambda_{i,c} R(c), \text{ such that } \lambda_{i,j} > 0.$$

As an example, consider the CLEVR dataset (Johnson et al., 2017) consisting of images of objects of different shapes and colors. A concept extractor for a vision model may extract the set of concepts $C_{\text{CLEVR}} = \{\{\texttt{red}\}, \{\texttt{blue}\}, \{\texttt{cube}\}, \{\texttt{sphere}\}\}$. C_{CLEVR} can also be grouped into attributes $A_1 = \{\{\texttt{red}\}, \{\texttt{blue}\}\}$ and $A_2 = \{\{\texttt{cube}\}, \{\texttt{sphere}\}\}$ containing color and shape concepts respectively. As such, a composite concept like $\{\texttt{red, sphere}\}$ can be represented as the weighted sum of $R(\{\texttt{red}\})$ and $R(\{\texttt{sphere}\})$.

3. Evaluating Concept Compositionality

In this section, we validate the compositionality of ground-truth concept representations and evaluate the same for concepts extracted using existing approaches. We first discuss our controlled setting and show that concept representations from the CLIP model are compositional. We then evaluate the compositionality of concepts extracted by existing approaches. Finally, we outline the necessary properties of compositional concept representations.

3.1. Setup

In order to validate the compositionality of ground-truth concepts, we focus on concepts extracted from subsets of the CLEVR (Johnson et al., 2017), CUB (Wah et al., 2011), and Truth (Azaria & Mitchell, 2023) datasets, all of which have labelled attributes with compositional structure.

We follow a setup similar to (Lewis et al., 2022) for the synthetic CLEVR (Johnson et al., 2017) dataset and consider images with single objects labelled as one of three shapes (sphere, cube, or cylinder) and one of three colors (red, green, or blue). We also consider a subset of the CUB dataset consisting of bird images labelled as one of three colors and one of three sizes. Finally, we consider a subset of the Truth (Zou et al., 2023b) dataset consisting of facts relating to one of three topics and labelled true or false.

3.2. Ground-Truth Concept Compositionality

We evaluate the compositionality of ground-truth concept representations learned by the CLIP model over each labelled dataset. Since these representations are not provided, for each concept, we consider the mean of the model's embeddings for samples belonging to that concept as a surrogate of its true representation (Zou et al., 2023a).

For example, for the CLEVR dataset, we extract the ground-truth representation of the `red` concept by calculating the mean of all sample embeddings of images with red objects. We similarly extract the ground-truth representations for the other two color concepts, the three shape concepts, and composite concepts like $\{\texttt{red, sphere}\}$, for a total of 15 concepts. We repeat this process for each dataset.

As stated in Lemma 2.3, the concept score for a composite of two concepts is the weighted sum of the concept scores of each concept. This implies that a linear model should be able to predict the concept score for a composed concept given the concept scores for each of the base concepts. We thus train a linear model to predict the presence or absence of a composed concept given its base concepts. We measure the average precision of the model for each composed concept, and report the mean average precision (MAP) score in Table 2a for each dataset. We see that in all cases, the ground truth (GT) concepts have high MAP (up to 0.971 for CLEVR) when predicting concept compositions from their components, meaning the ground-truth concept representations are reasonably compositional.

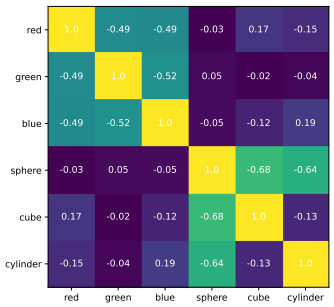
3.3. Compositionality Issues with Existing Methods

We next study the compositionality of concept representations discovered by existing unsupervised concept extraction methods. We train a linear model similar to the one described in Section 3.2, but with concepts extracted by baseline methods instead of the ground truths. From the MAP results in Table 2a we see that all the baselines have significantly lower compositionality than the ground-truth.

This is the case even for techniques that extract the concepts reasonably well, i.e. where the extracted concepts are able to discriminate between positive and negative samples of that concept. For each dataset and concept extraction method, we calculate the ROC-AUC score to measure the ability of the extracted concept to perform such a discrimination. We provide the full ROC-AUC results in Appendix E.6.

Method	CLEVR	CUB-sub	Truth-sub
GT	1.000 ± 0.000	0.808 ± 0.000	0.625 ± 0.000
PCA	0.981 ± 0.000	0.663 ± 0.000	0.467 ± 0.000
ACE	0.834 ± 0.029	0.651 ± 0.011	0.551 ± 0.017
DictLearn	0.891 ± 0.005	0.650 ± 0.010	0.533 ± 0.006
SemiNMF	0.780 ± 0.029	0.629 ± 0.029	0.525 ± 0.050
CT	0.575 ± 0.039	0.510 ± 0.003	0.428 ± 0.055
Random	0.568 ± 0.087	0.445 ± 0.079	0.461 ± 0.034
CCE	1.000 ± 0.000	0.648 ± 0.008	0.545 ± 0.004

(a) MAP score of predicting concept compositions.



(b) Cosine similarities between CLEVR concepts.

Figure 2. Compositionality of ground-truth concepts compared with concepts extracted by existing approaches and CCE. Figure 2a shows that the ground-truth concepts (GT) are quite compositional, but existing methods are not. Figure 2b shows the cosine similarities between pairs of ground-truth concepts for the CLEVR dataset. The darker blue cells represent concepts that are orthogonal, while the lighter yellow ones represent non-orthogonal ones. We observe that concepts tend to be more orthogonal if they belong to different attributes.

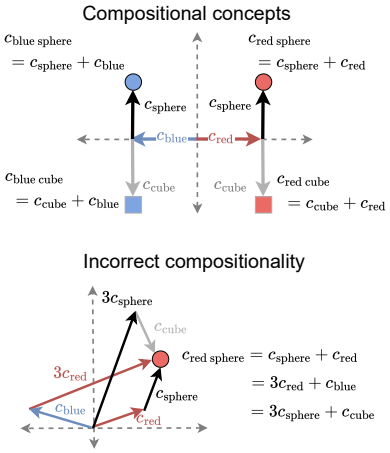


Figure 3. Illustration of concepts on a dataset of cubes and spheres that are either red or blue. The concepts on the top are compositional while those on the bottom are not. Even though the concepts on the bottom can perfectly represent the four samples, they still fail to compose properly. For instance, the composition of the `red` and `blue` concepts can form the {`red, sphere`} concept even though the `blue` concept is not present in a red sphere.

In the case of NMF, despite this score averaging as high as 0.907 for the CLEVR dataset, the extracted concepts are not compositional. This implies that finding concept representations simply based on their ability to discriminate positive and negative samples of a concept does not mean that those representations will compose as expected.

We further demonstrate this point with a toy illustration in Figure 3. This figure depicts four perfectly composed concepts at the top, and four incorrectly composed concepts at the bottom, even though each concept is perfectly discriminative of the samples with the concept. Therefore, we must ensure that we explicitly extract compositional concepts.

3.4. Desired Properties of Compositional Concepts

To extract compositional concepts, we must first identify characteristics of such concepts. Since the ground-truth concepts were compositional, we investigate the salient characteristics of those concepts.

Consider the ground-truth concepts for the CLEVR dataset. In order to understand the relationship between different ground-truth concepts and their compositionality, we center the sample embeddings and visualize cosine similarities between pairs of these concepts in Figure 2b. We observe that the ground-truth representations of color concepts are roughly orthogonal (cosine similarity near 0) to those of shape concepts. In contrast, the representations of concepts within the same attribute, such as the `red` and `blue` concepts, are non-orthogonal. Furthermore, the orthogonal concepts are also those that can compose to form new concepts, since they lie in different attributes. For instance, the `red` and `sphere` concepts are orthogonal, and can compose to form the {`red, sphere`} concept, while the `red` concept can't compose with the `blue` concept.

We visualize the same for the CUB-sub and Truth-sub datasets in Appendix C, and empirically observe the following trend over all three datasets: concept representations from different attributes are roughly orthogonal while those from the same attribute are non-orthogonal. Also, the orthogonal concepts tend to be compositional, while the non-orthogonal ones can't be composed.

Orthogonality is a generally helpful property for several use cases, such as disentangling concepts in embedding space (Chen et al., 2020). Some approaches therefore try to enforce orthogonality on the concepts being extracted. Table 1 summarizes existing unsupervised approaches for concept extraction and whether the method enforces any orthogonality constraints (Ortho.) between concepts of different attributes and allows for non-orthogonality between those

Table 1. Properties enforced by unsupervised concept extraction.

Method	Example	Ortho.	Corr.
PCA	RepE (Zou et al., 2023a)	✓	✗
KMeans	ACE (Ghorbani et al., 2019)	✗	✓
Dictionary-Learning	TransformerVis (Yun et al., 2021)	✗	✓
NMF	CRAFT (Fel et al., 2023)	✗	✓
Custom	Concept Tf (Rigotti et al., 2022)	✗	✓
Custom	CCE (Ours)	✓	✓

of the same attribute (Corr.). We see that these approaches allow for only one of the two, but not both.

We now formally prove that the observed properties regarding concept compositionality hold in a generalized setting.

Theorem 3.1. *For some dataset, consider two attributes A and A' where A has l concepts $c_1, \dots c_l$ and A' has l' concepts $c'_1, \dots c'_{l'}$. Assuming that for each compositional concept $c = \{c_i, c'_j\}$, its representation $v_{i,j}$, follows a spherical normal distribution with zero mean and unit covariance, i.e. $v_{i,j} \sim N(\mathbf{0}, \mathbf{I}^d)$, the following statements are true with high probability for a large dimension d:*

- *There exists $c_1, c_2 \in A$ and $c'_1, c'_2 \in A'$ such that the representations of these base concepts are non orthogonal.*
- *For all $c_1 \in A$ and $c_2 \in A'$, the representations of c_1 and c_2 are orthogonal with high probability.*

We show the proof in Appendix B. The takeaway from this result is that compositional concepts will be roughly orthogonal, while concepts of the same attribute may not be orthogonal. In addition, we show in Corollary B.4 that given concepts which follow the consequent of the above theorem, that the concepts will have compositional concept representations, meaning the representations of composite concepts consist of a sum of their component base concept representations, as defined in Definition 2.2. We leverage this to design an unsupervised concept extraction method which can find compositional concepts when they exist.

4. Compositional Concept Extraction (CCE)

To achieve this orthogonality property between concepts, we propose CCE, summarized in Algorithm 1 and visualized in Figure 4. As the outer loop of the algorithm suggests, once we find concepts for an attribute in a subspace P, we remove that subspace using orthogonal rejection and find concepts in a new subspace. This enforces orthogonality between the discovered subspaces, thus respecting the orthogonality property described in Section 3.

To discover concepts within each attribute, we employ a two-step process consisting of LearnSubspace and LearnConcepts, as illustrated in Figure 4. The LearnSubspace step, shown on the left, is given a clustering of the data (in terms of centroids V) and optimizes a subspace, defined by $P \in \mathbb{R}^{d \times S}$, so that the data in this subspace (ZP) becomes well clustered according to the fixed centroids V. In the next step, LearnConcepts, shown on the right, we identify concepts by performing spherical K-Means clustering on ZP, the data within subspace P.

Algorithm 1 Compositional Concept Extraction

Input: embeddings Z, num. attr. M, concepts per attr. K, subspace dimension S
Initialize concepts $C = \{\}$
for $m = 1 \dots M$ **do**
  Initialize $P \in \mathbb{R}^{d \times S}$ such that $P^T P = I$.
  Initialize K concepts $V = \{v_1, \dots, v_K\}$.
  repeat
    $P = \text{LearnSubspace}(P, Z, V)$
    $V = \text{LearnConcepts}(ZP, K)$
  until Converged
  $C = C \cup V$
  $Z = Z - ZPP^T$
end for
Return C

This clustering process is performed within a learned subspace and the subspace is learned according to the learned clustering. Therefore, we jointly learn the subspace P and the clustering centroids V. Specifically, for LearnSubspace, we employ the Silhouette score (Rousseeuw, 1987) to quantify how well clustered the projected data ZP is given a cluster assignment L determined by the centroids from spherical K-Means clustering. The Silhouette score measures the ratio of average within cluster distance to average between cluster distance. Since the Silhouette score is differentiable, once we fix a clustering L from LearnConcepts, we perform a step of gradient ascent in LearnSubspace to increase the Silhouette score. Thus, we solve the following optimization problem by iteratively fixing P to learn L (with LearnConcepts) and then fixing L to learn P by a gradient step (with LearnSubspace) until convergence:

$$\arg\max_{P,L} \text{Sil}(ZP, L).$$

We further observe that simply maximizing the above objective leads to overfitting issues since projecting the learned cluster centroids from LearnConcepts back to the original space may not necessarily correspond to cluster centroids in the original space. Therefore, in the LearnSubspace step we additionally try to match the cluster centroids learned within the subspace and projected out to the original space to the centroids of the clusters in the original space. This is integrated into the above full objective function as a regularization term, i.e.:

$$\arg\max_{P,L} \left(\text{Sil}(ZP, L) + \sum\nolimits_k S_{\cos}\left(C_k P^T, \hat{C}_k\right) \right),$$

where C_k represents the clustering centroids in the subspace

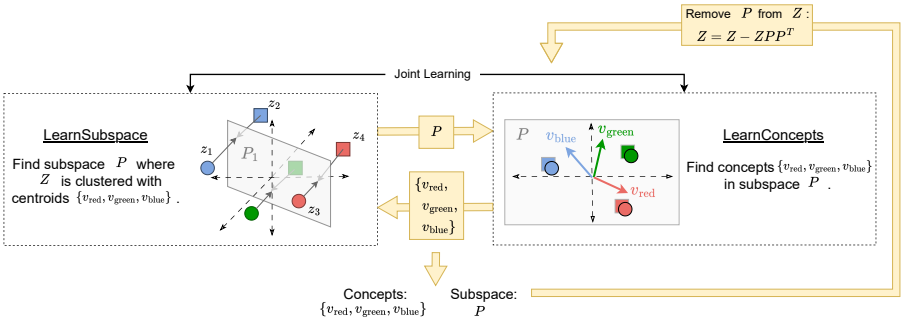


Figure 4. Finding color concepts in one iteration of CCE, which can be proceeded by finding other concepts, such as shapes.

P while $\hat{C}_k = \frac{1}{\sum_i \mathbb{1}[L_i=k]} \sum_i \mathbb{1}[L_i = k] Z_i$ represents the clustering centroids in the original space.

5. Experiments

5.1. Experimental Setup

Datasets and Models. We evaluate using five datasets across vision and language settings: CLEVR (Johnson et al., 2017) (vision), CUB (Wah et al., 2011) (vision), HAM10000 (Tschandl et al., 2018) (vision), Truth (Zou et al., 2023b) (language), and News (Mitchell, 1999) (language). We perform experiments on both *controlled* and *full* settings. In the controlled setting, we follow the same configuration as Section 3.1 for the CLEVR, CUB and Truth datasets. Further information on our datasets is included in Appendix F. The *full setting* considers all samples from the CUB, Ham, Truth, and News datasets.

For the image datasets, we obtain sample representations from the CLIP model (Radford et al., 2021) while for the NLP dataset, this is achieved with Llama-2 13B Chat model (Touvron et al., 2023). We also perform ablation studies on the choices of different models in Appendix E.8.

Baseline Methods. Since the concept representations are learned by CCE in an unsupervised manner, we therefore primarily compare CCE against the following state-of-the-art unsupervised concept extraction methods, i.e., PCA (Zou et al., 2023a), NMF (Fel et al., 2023), ACE (KMeans) (Ghorbani et al., 2019), and Dictionary Learning (Bricken et al., 2023; Yun et al., 2021). In addition, we include a Random baseline where we randomly initialize concept vectors from a normal distribution of mean zero and variance one.

Recent studies like Concept Transformer (Rigotti et al., 2022) explore how to jointly learn concept representations and perform training of downstream classification tasks with learned concept representations. Hence, we treat Concept Transformer (Concept Tf) (Rigotti et al., 2022) as another baseline. Note that Concept Tf can optionally incorporate concept labels as additional supervisions, which are not considered in our experiments for fair comparison.

Experiment Design. We aim to answer these questions regarding the quality of the learned concept representations:

- **RQ1** In the controlled setting with known compositional ground-truth concept representations, does CCE compose concepts more effectively than baselines?
- **RQ2** In the full setting where the ground-truth concepts are typically unknown, can CCE successfully discover new and meaningful compositional concepts?
- **RQ3** In both controlled and full settings, how can the learned compositional concept representations impact downstream performance?

To address **RQ1**, we evaluate the compositionality score (Andreas, 2019) on the concept representations extracted by CCE and the baselines, which is defined as follows:

Definition 5.1. (Compositionality Score) Given a dataset D consisting of embeddings $z \in \mathbb{R}^d$, their associated ground-truth concepts $C \subset \mathbb{C}$, and a concept representation function $R : \mathbb{C} \to \mathbb{R}^d$ obtained from a concept extractor, the compositionality score is the following:

$$\min_{\Lambda \geq 0} \frac{1}{|D|} \sum_{(z,C) \in D} \left\| z - \sum_{i=1}^{|C|} \Lambda_{z,i} R(C_i) \right\|$$

Intuitively speaking, for a sample embedding z, this metric quantifies how much z can be reconstructed by composing a list of concept representation $R(c_i)$'s that correspond to the i_{th} ground-truth concepts of z. Each $R(c_i)$ is weighted by a coefficient $\Lambda_{z,i}$, which is determined by optimizing the above formula with respect to all $\Lambda_{z,i}$.

In addition, for each ground-truth concept, we also report the cosine similarity between the learned concept representation $R(c_i)$ and the corresponding ground-truth representation.

To study **RQ2** for the full setting, we primarily perform qualitative studies to identify whether CCE is capable of discovering reasonable compositional concepts. Specifically, for each learned concept representation, we assign a name to the concept by inspecting the ten images with the top concept score. Then for each pair of the learned concepts, we first identify those samples with the highest concept scores. Then, we sum the two concept representations, and find

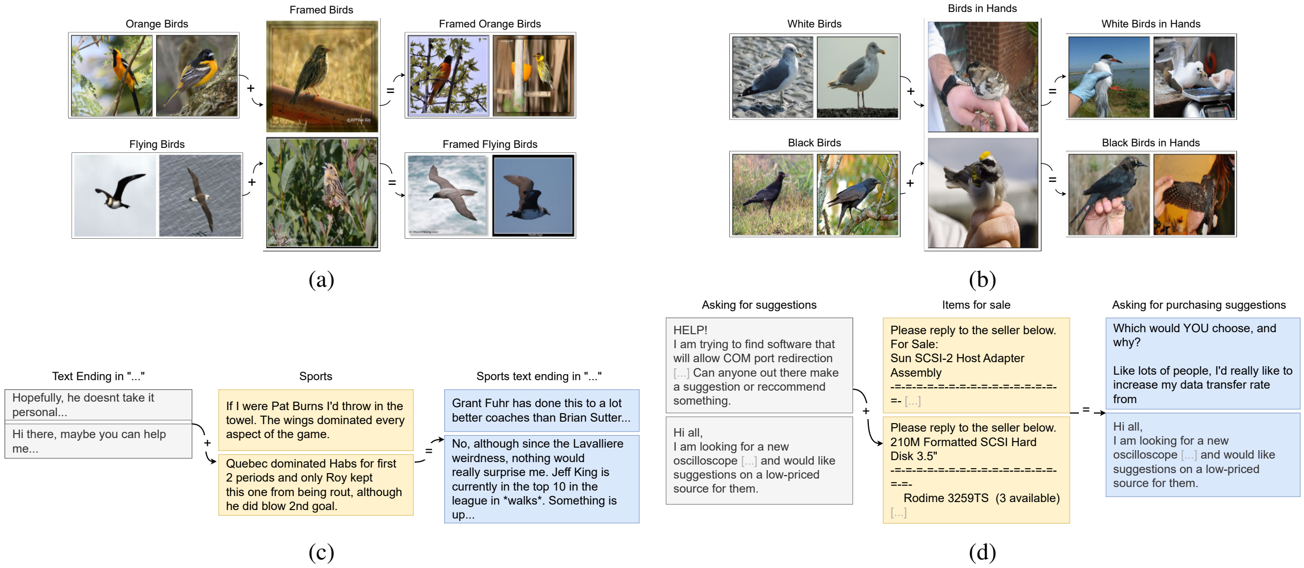


Figure 5. Examples of compositional concepts identified by CCE. Figures 5a and 5b are from the CUB dataset while Figures 5c and 5d are from the News dataset. These figures suggest that CCE can not only discover new meaningful concepts outside the ground-truth concepts, such as the `Birds in Hands` concept in Figure 5b, but also compose these concepts correctly, e.g. `White Birds` + `Birds in Hands` = `White Birds in Hands`.

Table 2. Compositionality Scores (lower is better).

	CLEVR	CUB-sub	Truth-sub
GT	3.162 ± 0.000	0.462 ± 0.000	3.743 ± 0.000
PCA	3.684 ± 0.000	0.472 ± 0.000	3.975 ± 0.000
ACE	3.474 ± 0.134	0.496 ± 0.007	3.727 ± 0.032
DictLearn	3.367 ± 0.016	0.498 ± 0.002	3.708 ± 0.007
SemiNMF	3.716 ± 0.053	0.495 ± 0.004	3.781 ± 0.074
CT	4.929 ± 0.002	0.545 ± 0.000	4.348 ± 0.000
Random	4.925 ± 0.000	0.545 ± 0.000	4.348 ± 0.000
CCE	3.163 ± 0.000	0.459 ± 0.004	3.689 ± 0.002

the samples with largest concept score for this aggregated representation. By investigating these examples, we visually examine whether the composition is reasonable or not.

Lastly, we answer **RQ3** by evaluating the downstream classification performance with the learned concept representations. Specifically, we follow Yuksekgonul et al. (2023) to learn a linear classifier by predicting class labels with the concept scores of a sample. We further report the performance of training a linear classifier on sample embeddings without involving any concepts, denoted by "No concept".

5.2. Experimental Results

Compositionality in Controlled Settings. We first evaluate the compositionality scores on the CLEVR, CUB-sub, and Truth-sub datasets and report them in Table 2. In all cases, CCE obtains the best score compared to the baselines, indicating the advantage of CCE in discovering compositional concepts. Moreover, CCE's scores are comparable to those of the ground-truth concept representations. This shows that the concepts learned by CCE almost align with the ground-truth concept representations.

This is further supported by the results in Table 3. This table summarizes the cosine similarities between the ground-truth concept representations and the ones learned by the baselines and CCE. Again, the concepts learned by CCE are the closest to the ground truths. Note that some baselines like Dictlearn also produce highly accurate concept representations. However, as Table 2 shows, their compositions fail to be consistent with the ground truths.

Compositionality in Real Data Settings. To address **RQ2**, we perform some qualitative studies on compositional concepts discovered by CCE on the CUB and News dataset, which are visualized in Figure 5. As shown in this figure, CCE is capable of identifying reasonable concepts, such as `White Birds`, `Framed Birds` and `Text Ending in ``...''`. Some of these concepts are even beyond the ground-truth concept labels that are provided by the dataset itself. For example, CCE identifies the "Birds in Hands" concept which is not labeled in the CUB dataset. But its top activated samples are images with a bird in someone's hand (see Figure 5b). Furthermore, the composition of those learned concepts is also representative of the properties of each concept. For example, in Figure 5c, the composition of the concept `Text Ending in ``...''` and `Sports` represents sentences about "sports" ending in "...".

Downstream Performance Analysis. For **RQ3**, we studied the impact of the extracted compositional concepts on

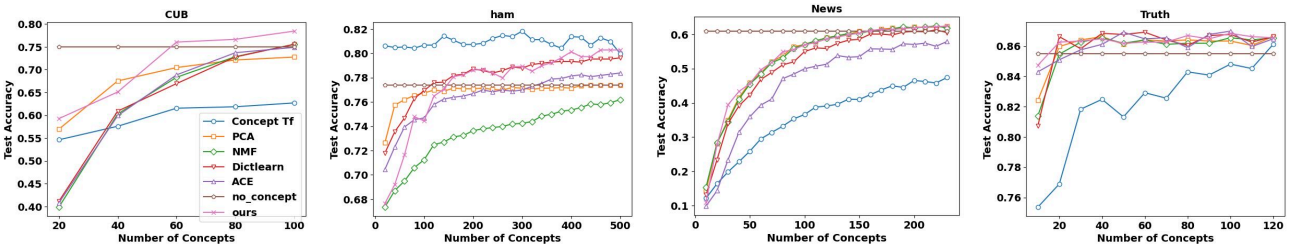


Figure 6. Downstream classification accuracy on the full setting.

Table 3. The average cosine similarity between individual learned concept representations and the ground truth (higher is better).

	CLEVR	CUB-sub	Truth-sub
PCA	0.580 ± 0.000	0.503 ± 0.000	0.459 ± 0.000
ACE	0.728 ± 0.009	0.719 ± 0.016	0.648 ± 0.007
DictLearn	0.745 ± 0.003	0.661 ± 0.010	0.686 ± 0.007
SemiNMF	0.732 ± 0.014	0.696 ± 0.002	0.673 ± 0.052
CT	0.044 ± 0.009	0.066 ± 0.001	0.019 ± 0.002
Random	0.059 ± 0.003	0.043 ± 0.011	0.024 ± 0.001
CCE	0.992 ± 0.000	0.770 ± 0.001	0.804 ± 0.001

downstream performance across all datasets in the full setting. Throughout the experiments, we observe that the total number of concepts is a crucial factor in determining the performance. Therefore, we also vary this number and report the performance numbers accordingly for all datasets and methods in Figure 6. As this figure suggests, across all the datasets, despite the poor performance with a small number of concepts, CCE gradually gains performance with an increasing number of concepts, eventually outperforming all the unsupervised baseline methods.

Also, it is worth noting that CCE outperforms Concept Tf most times and is on par with it in the worst case (see the experimental results on the ham dataset with 500 concepts). This thus indicates the performance advantage of CCE even in the absence of supervision from downstream tasks.

Furthermore, CCE discovers concept representations by performing a series of linear transformations on top of the sample embeddings. But by comparing against "No concept" where sample embeddings are directly used for downstream tasks, CCE can even outperform it by a large margin on CUB and Ham dataset. This implies that the concept representations extracted by CCE might be more relevant to the downstream classification tasks than the raw embeddings.

6. Related Work

Concept-based Interpretability. Concept-based interpretability encompasses the building of models using human-interpretable concepts (Koh et al., 2020; Espinosa Zarlenga et al., 2022; Yuksekgonul et al., 2023) and extracting such concepts post-hoc from models (Kim et al., 2018; Zhou et al., 2018). In either case, how do we choose which concepts to use? Some existing work specifies concepts using human supervision to select and provide their labels (Kim et al., 2018), large-scale concept annotation datasets (Bau et al., 2017), general knowledge bases (Yuksekgonul et al., 2023), and large language models (Yang et al., 2023). Another line of work uses regularization (Wong et al., 2021), or other inductive biases (Rigotti et al., 2022) to learn concepts during standard supervised training of a model. Finally, there is work which leverages unsupervised methods to automatically discover concepts (Ghorbani et al., 2019; Fel et al., 2023; Yun et al., 2021; Bricken et al., 2023) which is the approach taken in this paper. Unlike existing unsupervised concept learning methods which focus on properties such as faithfulness (Ghorbani et al., 2019) or human-meaningfulness (Fel et al., 2023), we focus specifically on compositionality.

Compositionality in Foundation Models. Since the observation of compositional word vectors by Mikolov et al. (2013) there has been interest in finding and utilizing compositional behavior of deep learning models. Existing work has leveraged insights from psychology and cognitive science to find concepts learned by generative models (Frankland & Greene, 2020; Lake, 2014). Compositionality has been used to uncover and mitigate bias in word embeddings (Bolukbasi et al., 2016), edit classifier behavior (Santurkar et al., 2021), and recently to monitor and control the behavior of foundational language (Todd et al., 2023; Zou et al., 2023a) and vision models (Wang et al., 2023; Kwon et al., 2023). To the best of our knowledge, we are the first to evaluate compositionality of concept representations learned by unsupervised approaches and to propose a method to improve compositionality of discovered concepts.

Compositional and Disentangled Representations. In representation learning, there is considerable effort to encourage *disentangled* representations (Bengio et al., 2013; Higgins et al., 2016; Wang et al., 2022). While disentanglement concerns how to distinguish separate concepts in embedding space, compositionality concerns what happens when separate concepts get combined. Existing work has shown that disentanglement and compositionality do not have to be correlated (Xu et al., 2022). Unlike representation learning, we start with a pretrained model and try to uncover the compositional concepts it learned.

Structures beyond compositionality. This paper focuses on compositionality in concept-based interpretability, but other important structures include subpopulation, relational, and causal structures. Group, or subpopulation, structure has been used as a way to interpret datasets with existing work on automatically finding such structure (Blei et al., 2001) and explaining models with respect to this structure (Havaldar et al., 2023). In addition, existing work has developed methods to steer explanations to respect group structures (Stein et al., 2023). Relational structures have also been studied as a lens into understanding the behavior of pretrained models (Todd et al., 2024; Lovering & Pavlick, 2022; Hill et al., 2018). Beyond group and relational structures, recent work proposes a method to identify known causal structures in pretrained LLMs (Wu et al., 2023).

7. Limitations

We study the case where concepts compose compositionally, but concepts may also be non-compositional. For instance, the concepts of `hot` and `dog` do not compose to form the meaning of `hot dog` (Zhai, 1997). In addition, we supposed a flat concept structure, which does not distinguish between "(small blue) car" and "small (blue car)". We leave the study of such non-compositional and hierarchical concepts to future work.

Another limitation of unsupervised concept extraction is that discovered concept vectors are not associated with any name. We assign names to the concept through manual inspection of samples with a high concept score, but this can require significant effort with large numbers of concepts.

8. Conclusion

In this paper, we studied concept-based explanations of foundation models from the lens of compositionality. We validated that the ground-truth concepts extracted from these models are compositional while the existing unsupervised concept extraction methods usually fail to guarantee compositionality. To address this issue, we first identified two salient properties for compositional concept representations and designed a novel concept extraction method called CCE that respects these properties by design. Through extensive experiments across vision and language datasets, we demonstrated that CCE not only learns compositional concepts but also enhances downstream performance.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grand No. DGE-2236662, the Google Research Fellowship, and "The Fundamental Research Funds for the Central Universities, Peking University".

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Hello GPT-4o. URL https://openai.com/index/hello-gpt-4o/.

Abid, A., Yuksekgonul, M., and Zou, J. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, pp. 66–88. PMLR, 2022.

Andreas, J. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019.

Azaria, A. and Mitchell, T. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*, 2023.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *Advances in neural information processing systems*, 14, 2001.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022.

Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Frankland, S. M. and Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, 71:273–303, 2020.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Havaldar, S., Stein, A., Wong, E., and Ungar, L. H. Topex: Topic-based explanations for model comparison. In Maughan, K., Liu, R., and Burns, T. F. (eds.), *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=AidIUjh__t.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

Hill, F., Santoro, A., Barrett, D., Morcos, A., and Lillicrap, T. Learning to make analogies by contrasting abstract relational structure. In *International Conference on Learning Representations*, 2018.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

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 *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

Lake, B. M. *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn*. PhD thesis, Massachusetts Institute of Technology, 2014.

Lewis, M., Nayak, N. V., Yu, P., Yu, Q., Merullo, J., Bach, S. H., and Pavlick, E. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022.

Lovering, C. and Pavlick, E. Unit testing for concepts in neural networks. *Transactions of the Association for Computational Linguistics*, 10:1193–1208, 2022. doi: 10.1162/tacl_a_00514. URL https://aclanthology.org/2022.tacl-1.69.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Mitchell, T. Twenty Newsgroups. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C5C323.

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.

Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., and Scotton, P. Attention-based interpretability with concept transformers. In *International conference on learning representations*, 2022.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale

Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Santurkar, S., Tsipras, D., Elango, M., Bau, D., Torralba, A., and Madry, A. Editing a classifier by rewriting its prediction rules. *Advances in Neural Information Processing Systems*, 34:23359–23373, 2021.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Stein, A., Wu, Y., Wong, E., and Naik, M. Rectifying group irregularities in explanations for distribution shift. *arXiv preprint arXiv:2305.16308*, 2023.

Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., and Ganguli, D. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.

Todd, E., Li, M., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Trager, M., Perera, P., Zancato, L., Achille, A., Bhatia, P., and Soatto, S. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15395–15404, 2023.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2024.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, X., Chen, H., Tang, S., Wu, Z., and Zhu, W. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*, 2022.

Wang, Z., Gui, L., Negrea, J., and Veitch, V. Concept algebra for (score-based) text-controlled generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wegner, S.-A. Lecture notes on high-dimensional data. *arXiv preprint arXiv:2101.05841*, 2021.

Wong, E., Santurkar, S., and Madry, A. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*, pp. 11205–11216. PMLR, 2021.

Wu, Z., Geiger, A., Icard, T., Potts, C., and Goodman, N. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36, 2023.

Xu, Z., Niethammer, M., and Raffel, C. A. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and 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.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yun, Z., Chen, Y., Olshausen, B., and Lecun, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 1–10, 2021.

Zhai, C. Exploiting context to identify lexical atoms–a statistical view of linguistic context. *arXiv preprint cmp-lg/9701001*, 1997.

Zhou, B., Sun, Y., Bau, D., and Torralba, A. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023b.

A. Proof of Lemma 2.3

Proof. Let $z \in \mathbb{R}^d$ be a sample embedding, $R : \mathbb{C} \to \mathbb{R}^d$ be a compositional concept representation function, and $c_i, c_j \in \mathbb{C}$ be two compositional concepts which compose as $c_k = c_i \cup c_j$. From Definition 5.1, the concept scores for c_i and c_j are the following:

$$s(z, c_i) = S_{\cos}(z, R(c_i))$$
$$s(z, c_j) = S_{\cos}(z, R(c_j)).$$

The concept score for the composition c_k can then be written as:

$$\begin{aligned}
s(z, c_k) &= s(z, c_i \cup c_j) \\
&= S_{\cos}(z, R(c_i \cup c_j)) \\
&= S_{\cos}(z, w_{c_i} R(c_i) + w_{c_j} R(c_j)) && \text{(since } R \text{ is compositional)} \\
&= \frac{z \cdot (w_{c_i} R(c_i) + w_{c_j} R(c_j))}{\|z\| \|w_{c_i} R(c_i) + w_{c_j} R(c_j)\|} && \text{(definition of cosine similarity)} \\
&= \frac{z \cdot w_{c_i} R(c_i)}{\|z\| \|R(c_k)\|} + \frac{z \cdot w_{c_j} R(c_j)}{\|z\| \|R(c_k)\|} \\
&= \frac{(w_{c_i} \|R(c_i)\|) z \cdot R(c_i)}{\|R(c_k)\| \|z\| \|R(c_i)\|} + \frac{(w_{c_j} \|R(c_j)\|) z \cdot R(c_j)}{\|R(c_k)\| \|z\| \|R(c_j)\|} \\
&= \frac{w_{c_i} \|R(c_i)\|}{\|R(c_k)\|} S_{\cos}(z, R(c_i)) + \frac{w_{c_j} \|R(c_j)\|}{\|R(c_k)\|} S_{\cos}(z, R(c_j)) && \text{(definition of cosine similarity)}
\end{aligned}$$

□

B. Proof of Theorem 3.1

Lemma B.1 (curse of dimensionality)**.** *(Wegner, 2021) For a pair of vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *randomly sampled from* $N(0, \mathbf{I}^d)$*,* $\mathbf{x}$ *and* $\mathbf{y}$ *are orthogonal with high probability for large enough* d*. Mathematically speaking, for a fixed small constant,* ϵ*, the following inequality holds:*

$$\mathbb{P}\left[|\langle \frac{\mathbf{x}}{|\mathbf{x}|}, \frac{\mathbf{y}}{|\mathbf{y}|} \rangle| \leq \epsilon \right] \geq 1 - \frac{M_1}{\sqrt{d}\epsilon} - \frac{M_2}{\sqrt{d}},$$

where $M_1 = 2$ *and* $M_2 = 7$

Lemma B.2 (Gaussian Annulus Theorem)**.** *(Wegner, 2021) For a vector* v *randomly sampled from* $N(0, \mathbf{I}^d)$*,* $\|v\|$ *is approaching* $\sqrt{d}$ *with high probability for large enough* d*. Mathematically speaking, the following inequality holds:*

$$\mathbb{P}\left[|\|\mathbf{x}\| - \sqrt{d}| \leq \epsilon \right] \geq 2 \exp(-M_3 \epsilon^2),$$

in which $M_3 = \frac{1}{16}$

Based on the above two lemmas, for any two randomly sampled vectors $\mathbf{x}$ and $\mathbf{y}$ from $N(0, \mathbf{I}^d)$, the following equality holds with high probability:

$$\langle \mathbf{x}, \mathbf{y} \rangle = o(d) \tag{1}$$

Lemma B.3. *As defined in Theorem 3.1, for a composite concept* $c = \{c_i, c'_j\}$*, its representation is denoted by* $v_{i,j}$*, then the representation of the base concept* c_i *belonging to attribute* A *is:*

$$v_i = \frac{1}{l'} \sum_{j=1}^{l'} v_{i,j}.$$

Similarly, the representation of the base concept $c'_j \in A'$ is:

$$v'_j = \frac{1}{l}\sum_{i=1}^{l} v_{i,j}.$$

Proof. v_i could be derived by calculating the mean of the representations of all samples with concept c_i in the attribute A. Since those samples may have different concepts in the attribute A', then the composite concept among these samples could be $\{c_i, c'_1\}, \{c_i, c'_2\}, \ldots, \{c_i, c'_l\}$. Therefore, v_i is derived by:

$$v_i = \frac{1}{N}\sum_{x \text{ with concept } c_i \text{ in attribute A}} x = \frac{1}{N}\sum_{j=1}^{l'}\sum_{x \text{ with concept } \{c_i,\, c'_j\}} x,$$

in which N represents the number of samples with concept c_i in attribute A. By further assuming that there is a large enough number of samples for each composite concept, this implies that the number of each composite concept is roughly the same, i.e., around N/l'. Then the above formula could be transformed to:

$$v_i = \frac{1}{N}\sum_{j=1}^{l'}\sum_{x \text{ with concept } \{c_i,\, c'_j\}} x = \frac{1}{N}\sum_{j=1}^{l'}\frac{N}{l'}v_{i,j} = \frac{1}{l'}\sum_{j=1}^{l'} v_{i,j}.$$

The last step in the above formula leverages the fact that $v_{i,j}$ is calculated by the mean of all samples belonging to composite concept $\{c_i, c'_j\}$.

We can further illustrate this with one concrete example from the CLEVR dataset. By reusing the running example from Section 3, we assume that there are three colors {red, green, blue} and three shapes {sphere, cube, cylinder} in the CLEVR dataset. By following the notations of Theorem 3.1, the representation of a composite concept, say, $\{c_{\text{red}}, c_{\text{sphere}}\}$, is represented by $v_{\text{red,sphere}}$. Then the representation of the base concept sphere should be the mean of all samples belonging to this base concept. This can be derived by the mean of the samples belonging to the concept $\{c_{\text{red}}, c_{\text{sphere}}\}$, the ones belonging to $\{c_{\text{green}}, c_{\text{sphere}}\}$ and the ones belonging to $\{c_{\text{blue}}, c_{\text{sphere}}\}$. Therefore, the representation of c_{sphere} is denoted by:

$$v_{\text{sphere}} = \frac{1}{3}[v_{\text{red,sphere}} + v_{\text{green,sphere}} + v_{\text{blue,sphere}}].$$

□

We next present the formal proof of Theorem 3.1:

Proof. We split our proof into two parts. The first part is for proving "For the base concepts belonging to the same attribute, there exists at least one pair of non-orthogonal concepts." while the second part is for proving "For any pair of base concepts from two different attributes, they are orthogonal with high probability."

Part 1: There exists $c_1, c_2 \in A$ and $c'_1, c'_2 \in A'$ such that the representations of these base concepts are non orthogonal.

According to Lemma B.3, the concept representation for the base concept c_i (denoted by $\hat{v}_i$) is:

$$\hat{v}_i = \frac{1}{l'}\sum_{j=1}^{l'} v_{i,j}, \tag{2}$$

which sums over all concepts in A'.

Since we also want to perform centering operations over the entire dataset, then this suggests that we need to leverage the mean of all concepts, i.e.,:

$$\mu = \frac{1}{ll'}\sum_{i,j} v_{i,j}. \tag{3}$$

Then after the centering operation, $\hat{v}_i$ is transformed into:

$$v_i = \frac{\hat{v}_i - \mu}{\sigma}. \tag{4}$$

In the formula above, we use σ to represent the standard deviation vector calculated over the entire dataset.

Then let us fix i and sum up all v_i over all i, which yields:

$$\sum_{i=1}^{l} v_i = \sum_{i=1}^{l} \frac{\hat{v}_i - \mu}{\sigma} = \sum_{i=1}^{l} \frac{\hat{v}_i}{\sigma} - \frac{l\mu}{\sigma} \tag{5}$$

Then by integrating Equation 2 and Equation 3 into the above formula, we get:

$$\begin{aligned}
\sum_{i=1}^{l} v_i &= \sum_{i=1}^{l} \frac{\hat{v}_i}{\sigma} - \frac{l\mu}{\sigma} \\
&= \frac{1}{\sigma} \sum_{i=1}^{l} \frac{1}{l'} \sum_{j=1}^{l'} v_{i,j} - \frac{l\mu}{\sigma} \\
&= \frac{1}{\sigma l'} \sum_{i=1}^{l} \sum_{j=1}^{l'} v_{i,j} - \frac{1}{\sigma l'} \sum_{i=1}^{l} \sum_{j=1}^{l'} v_{i,j} \\
&= 0
\end{aligned}$$

We can equivalently show that $\sum_{j=1}^{l'} v'_j = 0$.

Therefore, the concept representations v_i within the attribute A are linearly dependent and the representations v'_i within the attribute A' are linearly dependent, meaning there exist concepts c_i and c_j such that $\langle v_i, v_j \rangle \neq 0$, and concepts c'_k and c'_m such that $\langle v'_k, v'_m \rangle \neq 0$.

Part 2: For all $c_1 \in A$ and $c_2 \in A'$, the representations of c_1 and c_2 are orthogonal with high probability.

To prove that all concept representations from A are orthogonal to all concept representations from A' , we will show that the dot product between these two representations is zero. Let $c_i \in A$ and $c'_j \in A'$ and v_i, v'_j are the concept representations for c_i and c'_j respectively. We can expand the dot product as follows:

$$\langle v_i, v'_j \rangle = \left\langle \frac{\hat{v}_i}{\sigma} - \frac{\mu}{\sigma}, \frac{\hat{v}'_j}{\sigma} - \frac{\mu}{\sigma} \right\rangle$$

Then by integrating Equation 2 and Equation 3 into the above formula, we can expand the above into the following:

$$\langle v_i, v'_j \rangle = \frac{1}{\sigma^2} \left\langle \frac{1}{l'} \sum_{j=1}^{l'} v_{i,j} - \mu, \frac{1}{l} \sum_{i=1}^{l} v_{i,j} - \mu \right\rangle$$

We note that for arbitrary pairs of $v_{i,j}$ and $v_{i',j'}$ with $i \neq i'$ or $j \neq j'$, since they are two different random vectors sampled from a spherical normal distribution $N(\mathbf{0}, \mathbf{I}^d)$, their dot product is $o(d)$ according to Equation 1. Therefore, through some

linear algebraic operations, the above formula could be reformulated as follows:

$$\begin{aligned}
\langle v_i, v'_j \rangle &= \frac{1}{\sigma^2} \left\langle \frac{1}{l'} \sum_{s=1}^{l'} v_{i,s} - \mu, \frac{1}{l} \sum_{t=1}^{l} v_{t,j} - \mu \right\rangle \\
&= \frac{1}{\sigma^2} \left\langle \frac{1}{l'} \sum_{s=1}^{l'} v_{i,s} - \frac{1}{ll'} \sum_{t,s} v_{t,s}, \frac{1}{l} \sum_{t=1}^{l} v_{t,j} - \frac{1}{ll'} \sum_{t,s} v_{t,s} \right\rangle \\
&= \frac{1}{\sigma^2 ll'} \left\langle \sum_{s=1}^{l'} v_{i,s} - \frac{1}{l} \sum_{t,s} v_{t,s}, \sum_{t=1}^{l} v_{t,j} - \frac{1}{l'} \sum_{t,s} v_{t,s} \right\rangle \\
&= \frac{1}{\sigma^2 ll'} \left[\sum_{s=1}^{l'} v_{i,s} \sum_{t=1}^{l} v_{t,j} - \frac{1}{l'} \sum_{s=1}^{l'} v_{i,s} \sum_{t,s} v_{t,s} - \frac{1}{l} \sum_{t=1}^{l} v_{t,j} \sum_{t,s} v_{t,s} + \frac{1}{ll'} \sum_{t,s} v_{t,s} \sum_{t,s} v_{t,s} \right] \\
&= \frac{1}{\sigma^2 ll'} \left[\|v_{i,j}\|^2 - \frac{1}{l'} \sum_{s=1}^{l'} \|v_{i,s}\|^2 - \frac{1}{l} \sum_{t=1}^{l} \|v_{t,j}\|^2 + \frac{1}{ll'} \sum_{t,s} \|v_{t,s}\|^2 \right] + o(d)
\end{aligned}$$

in which $o(d)$ is derived by applying Equation 1 to all the cross terms of the form $\langle v_{i,j}, v_{i',j'} \rangle$ where at least one pair of i, i' and j, j' are different.

We can further simplify this expression using Lemma B.2 which says that for each vector x randomly sampled from $N(\mathbf{0}, \mathbf{I}^d)$, its norm is bounded by $[\sqrt{d} - \epsilon, \sqrt{d} + \epsilon]$ with high probability, which applies to each $v_{i,j}$. Therefore, we can bound the above equation by:

$$\begin{aligned}
\langle v_i, v'_j \rangle &\leq \frac{1}{\sigma^2 ll'} \left[(\sqrt{d} + \epsilon)^2 - \frac{1}{l'} l' (\sqrt{d} - \epsilon)^2 - \frac{1}{l} l (\sqrt{d} - \epsilon)^2 + \frac{1}{ll'} ll' (\sqrt{d} + \epsilon)^2 \right] o(d) \\
&= \frac{8\sqrt{d}\epsilon}{\sigma^2 ll'} + o(d)
\end{aligned}$$

Similarly, we can prove that $\langle v_i, v'_j \rangle \geq -\frac{8\sqrt{d}\epsilon}{\sigma^2 ll'} + o(d)$, so we can conclude that

$$|\langle v_i, v'_j \rangle| = o(d) \tag{6}$$

Our goal is to get a bound on the cosine similarity of v_i and v'_j to show that it is zero. The cosine similarity is written $S_{\cos}(v_i, v'_j) = \frac{\langle v_i, v'_j \rangle}{\|v_i\| \|v'_j\|}$, so we have a bound on the numerator, but we now want a bound on the terms in the denominator. We can compute the norm of v_i and v'_j and follow the same derivation as above by leveraging Equation 1, which results in:

$$\begin{aligned}
\|v_i\|_2^2 &= \langle v_i, v_i \rangle \\
&= \frac{1}{\sigma^2 l'^2} \left\langle \sum_{s=1}^{l'} v_{i,s} - \frac{1}{l} \sum_{t,s} v_{t,s}, \sum_{s=1}^{l'} v_{i,s} - \frac{1}{l} \sum_{t,s} v_{t,s} \right\rangle \\
&= \frac{1}{\sigma^2 l'^2} \left[\sum_{s=1}^{l'} v_{i,s} \sum_{s=1}^{l'} v_{i,s} - 2\frac{1}{l} \sum_{s=1}^{l'} v_{i,s} \sum_{t,s} v_{t,s} + \frac{1}{l^2} \sum_{t,s} v_{t,s} \sum_{t,s} v_{t,s} \right] \\
&= \frac{1}{\sigma^2 l'^2} \left[\sum_{s=1}^{l'} \|v_{i,s}\|^2 - \frac{2}{l} \sum_{s=1}^{l'} \|v_{i,s}\|^2 + \frac{1}{l^2} \sum_{t,s} \|v_{t,s}\|^2 \right] + o(d)
\end{aligned}$$

Similarly, we can get the following:

$$\|v'_j\|_2^2 = \frac{1}{\sigma^2 l^2} \left[\sum_{t=1}^{l} \|v_{t,j}\|^2 - \frac{2}{l'} \sum_{t=1}^{l'} \|v_{t,j}\|^2 + \frac{1}{l'^2} \sum_{t,s} \|v_{t,s}\|^2 \right] + o(d)$$

By Lemma B.2, the norm of each $v_{i,j}$ is bounded by $\sqrt{d}-\epsilon$ and $\sqrt{d}+\epsilon$ with high probability, so the above formula can be bounded by:

$$\frac{1}{\sigma^2 ll'}((l-1)d-(2l+6)\sqrt{d}\epsilon+(l-1)\epsilon^2)+o(d) \leq \|v_i\|_2^2 \leq \frac{1}{\sigma^2 ll'}((l-1)d+(2l+6)\sqrt{d}\epsilon+(l-1)\epsilon^2)+o(d),$$

Therefore,

$$\|v_i\|_2^2 = O(d) \tag{7}$$

and we can equivalently show that $\|v'_j\| = O(d)$.

As a consequence, we can now calculate the cosine similarity between v_i and v'_j:

$$S_{\text{cos}}(v_i, v'_j) = \frac{\langle v_i, v'_j\rangle}{\|v_i\| \cdot \|v'_j\|} = \frac{o(d)}{O(d)} = o(1),$$

which means that this converges to zero as desired. □

Corollary B.4. *Given Theorem 3.1, for the representation of the composite concepts $v_{i,j}$, it can be (approximately) decomposed into the linear combinations of the representations of the base concepts (after the centering operation), v_i, v_j but is orthogonal to the representations of other base concepts with high probability. In other words, compositionality holds with high probability.*

Proof. To prove this, let us consider the cosine similarity between $v_{i,j}$ and v_t.

According to Equation 2, we first compute the inner product between these two vectors, i.e.:

$$\langle v_{i,j}, v_t\rangle = \frac{1}{l'}\sum_{n=1}^{l'}\langle v_{i,j}, v_{t,n}\rangle, \tag{8}$$

Depending on whether $t=i$ or not, there are two different cases.

Case 1: $t \neq i$ Note that according to Lemma B.2, since $v_{i,j}$ and $v_{t,n}$ are twowvectors randomly sampled from the spherical normal distribution, their inner product is $o(d)$. Therefore, the above inner product between $v_{i,j}$ and v_t becomes:

$$\langle v_{i,j}, \mu_t\rangle = o(d).$$

Also note that according to Equation 4, $v_t = \frac{\hat{v_t}-\mu}{\sigma}$, we thus need to leverage this equation to derive the inner product between $v_{i,j}$ and v_t. Furthermore, according to equation 3, μ is the mean of all the representations of the composite concepts, which are all randomly sampled from a spherical normal distribution. Therefore, μ is approaching 0 with high probability and thus the following equation holds with high probability:

$$\langle v_{i,j}, v_t\rangle = \langle v_{i,j}, \frac{\hat{v_t}-\mu}{\sigma}\rangle = \langle v_{i,j}, \frac{\hat{v_t}}{\sigma}\rangle = o(d), t \neq i.$$

In addition, according to Lemma B.2 and Equation 7, the norms of $v_{i,j}$ and v_t are both $O(\sqrt{d})$. Therefore, the cosine similarity between $v_{i,j}$ and v_t :

$$\text{cosine}(v_{i,j}, v_t) = \frac{\langle v_{i,j}, v_t\rangle}{\|v_{i,j}\| \cdot \|v_t\|} = \frac{o(d)}{\|v_{i,j}\| \cdot \|v_t\|} = \frac{o(d)}{O(d)} = o(1).$$

Intuitively speaking, this indicates that for the representation of a composite concept $v_{i,j}$, it is not correlated with the representation of a base concept that does not appear in this composite concept with high probability. For example, this could mean that the representation of the composite concept $\{c_{\text{red}}, c_{\text{sphere}}\}$ is not correlated to the representation of the concept c_{blue}, which is intuitively true.

Case 2: $t = i$ In Equation 8, according to Lemma B.2, the inner product between $v_{i,j}$ and most $v_{t,m}$ is $o(d)$ except when $j = m$. Therefore, Equation 8 becomes:

$$\langle v_{i,j}, v_t \rangle = \|v_{i,j}\|_2^2 + o(d),$$

Then according to Lemma B.2, since $\|v_{i,j}\|$ is approaching $\sqrt{d}$, then the above formula is transformed to:

$$\langle v_{i,j}, v_t \rangle = O(d),$$

Then according to Lemma B.2 and Equation 7, the norms of $v_{i,j}$ and v_t are both $O(\sqrt{d})$. Therefore, the cosine similarity between $v_{i,j}$ and v_t is:

$$\text{cosine}(v_{i,j}, v_t) = \frac{\langle v_{i,j}, v_t \rangle}{\|v_{i,j}\| \cdot \|v_t\|} = \frac{O(d)}{\|v_{i,j}\| \cdot \|v_t\|} = \frac{O(d)}{O(d)} = O(1),$$

which is thus a nonzero value.

As indicated by the above analysis, we can conclude that each $v_{i,j}$ is only correlated to the representation of the base concepts v_i, and v'_j. Since the representations of those base concepts are from different attributes, thus orthogonal to each other, then we can regard them as the basis vectors in the vector space, which can then be linearly combined to approximately reconstruct $v_{i,j}$, i.e.:

$$v_{i,j} = \text{cosine}(v_{i,j}, v_i)v_i + \text{cosine}(v_{i,j}, v'_j)v'_j$$

This thus matches the definition of the compositionality (see Definition 2.2).

□

Theorem B.5. *For some dataset, consider two attributes A and A' where we have l concepts for A, $c_1, \ldots, c_l$, and l' concepts for A', $c'_1, \ldots, c'_{l'}$. Define normalized concept representations $v_1, \ldots, v_l$ and $v'_1, \ldots, v'_{l'}$ for the concepts in A and A' such that v_i is orthogonal to v'_j for all i and j and for v_i and samples x and x' such that x has concept c_i and x' does not, then $S_{cos}(x, v_i) > S_{cos}(x', v_i)$. Then the concept representations are compositional.*

Proof. Let v_i be the concept representation for c_i and v'_j be the concept representation for c'_j. We are given that for any two samples x and x' with and without concept c_i respectively, $S_{\cos}(x, v_i) > S_{\cos}(x', v_i)$ and similarly for any two samples x and x' with and without concept c'_j respectively, $S_{\cos}(x, v'_j) > S_{\cos}(x', v'_j)$. We will show that a concept representation for $c_{i,j}$, the composition of concept c_i and c'_j, exists and is represented by $v_{i,j} = v_i + v'_j$.

Let $v_{i,j} = v_i + v'_j$. We will show that this concept can perfectly rank samples with the concept $c_{i,j}$. Since v_i and v'_j result in perfect rankings, for all x, x' such that x has c_i and x' does not, $S_{\cos}(x, v_i) - S_{\cos}(x', v_i) > 0$. Similarly, for any x, x' such that x has c'_j and x' does not, $S_{\cos}(x, v'_j) - S_{\cos}(x', v'_j) > 0$.

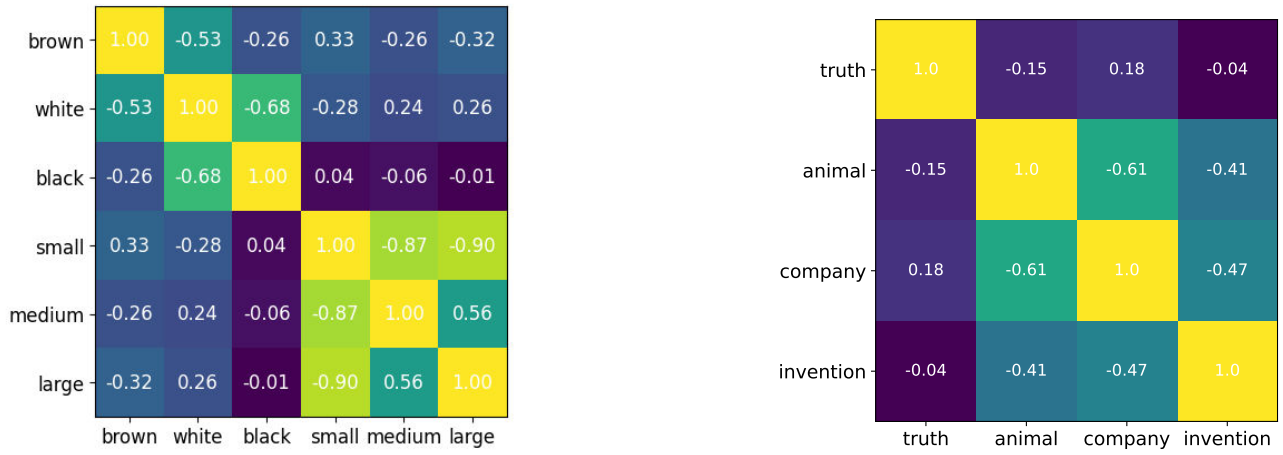
Now let, x, x' be such that x has concept $c_{i,j}$ and x' does not. We can write the following:

$$\begin{aligned} S_{\cos}(x, v_i + v'_j) &= \frac{\langle x, v_i + v'_j \rangle}{\|x\|\|v_i + v'_j\|} \\ &= \frac{\langle x, v_i \rangle + \langle x, v'_j \rangle}{\|x\|\sqrt{2}} && \text{Since } \langle v_i, v'_j \rangle = 0, \langle v_i, v_i \rangle = 1, \text{ and } \langle v'_j, v'_j \rangle = 1 \\ &= \frac{1}{\sqrt{2}}(S_{\cos}(x, v_i) + S_{\cos}(x, v'_j)) \end{aligned}$$

Therefore, we can now show that the concept score for the composed concept is larger for x than x':

$$\begin{aligned} S_{\cos}(x, v_i + v'_j) - S_{\cos}(x', v_i + v'_j) &= \frac{1}{\sqrt{2}}(S_{\cos}(x, v_i) + S_{\cos}(x, v'_j)) - \frac{1}{\sqrt{2}}(S_{\cos}(x', v_i) + S_{\cos}(x', v'_j)) \\ &= \frac{1}{\sqrt{2}}\left((S_{\cos}(x, v_i) - S_{\cos}(x', v_i)) + (S_{\cos}(x, v'_j) - S_{\cos}(x', v'_j))\right) \\ &> 0. \end{aligned}$$

□



(a) CUB-sub

(b) Truth-sub

Figure 7. Compositionality of Ground-Truth Concepts for the CUB-sub and Truth-sub datasets.

C. Compositionality of Ground-Truth Concepts

The cosine similarities between concepts is shown for the CUB-sub and Truth-sub datasets in Figure 7. We see similar findings as in Figure 2b.

D. Qualitative Examples

We provide additional qualitative results for the CUB dataset in Figure 8 and the ImageNet (Russakovsky et al., 2015) validation set in Figure 9. The concepts are named by manually looking at the top 20 images for each concept and coming up with a short description which is as specific as possible to the images while being general enough to apply to each image.

As an alternative to manual concept labelling, we also experimented with using a vision-text language model to automatically name concepts from their top 20 examples. We used GPT-4o (gpt) to get concept labels. For each concept, we produce a single image containing the top 20 samples for the concept and we pass the image to GPT-4o with the following prompt:

```
You are given 20 images representing a single concept and your task is to label
the name of the concept from just the 20 images. First, output a detailed caption
for each image. Then output a concept name which is specific to the images but
summarizes what is common among all of them. For example, for images of red cars
in different environments and positions, the concept name could be 'Red cars'.
Output the name of the concept after 'Concept Name:'.
```

The labels for the additional CUB examples in Figure 8 are the following where each line labels a row of the figure:

```
Hummingbirds, Birds, Hummingbirds
Black Birds, Birds in Natural Habitats, Black Birds
Wrens, Birds with food in their beaks, Wrens
Seagulls, Birds with food in their beaks, Birds with fish in their beaks
```

Similarly, the labels from GPT-4o for Figure 9 are the following:

```
Dogs, Sleeping in various environments, Sleeping Dogs
Reptiles and Amphibians in Natural Habitats, Pairs of Dogs, Pairs of Animals
Wild Animals, Pairs of Dogs, Animals in Pairs
Waterfront Structures and Transportation,
    Outdoor Activities and Wildlife,
    British Heritage and Infrastructure
```

```
Tools and Objects in Close-Up,
    Laboratory and Scientific Equipment,
    Vintage and Everyday Objects
```

E. Additional quantitative results

E.1. Runtime analysis

Table 4. Runtimes in seconds

Dataset	PCA	ACE	DictLearn	SemiNMF	CT	CCE
CLEVR	0.10 ± 0.12	0.02 ± 0.00	28.65 ± 0.29	8.13 ± 1.03	63.66 ± 0.73	190.98 ± 2.38
CUB-sub	0.11 ± 0.15	0.03 ± 0.01	14.38 ± 0.15	3.99 ± 0.09	6.89 ± 0.15	112.73 ± 2.67
CUB	0.84 ± 0.06	0.46 ± 0.03	51.53 ± 1.51	25.85 ± 0.22	495.49 ± 10.81	207.17 ± 0.70
Truth-sub	0.16 ± 0.03	0.06 ± 0.02	43.36 ± 4.35	29.83 ± 0.62	165.06 ± 1.21	316.45 ± 2.63
Truth	1.10 ± 0.16	2.64 ± 0.09	88.81 ± 6.54	194.67 ± 10.18	712.16 ± 7.70	1574.88 ± 17.68
HAM	1.89 ± 0.03	2.97 ± 0.03	367.67 ± 8.71	165.80 ± 2.22	693.73 ± 1.88	7460.52 ± 47.95
News	3.28 ± 0.72	25.75 ± 2.39	241.75 ± 38.70	934.69 ± 117.66	431.78 ± 7.11	7947.31 ± 70.64

E.2. Downstream performance error bars

We include error bars for the downstream performance results using the greatest number of concepts in Table 5.

Table 5. Error bars of the downstream performance (%). Three decimal places are given when necessary to show non-zero standard deviation.

Method	CUB	Truth	HAM	News
PCA	72.71 ± 0.01	87.137 ± 0.000	77.42 ± 0.01	62.029 ± 0.001
ACE	74.99 ± 0.06	87.161 ± 0.001	78.67 ± 0.12	57.019 ± 0.004
DictLearn	75.33 ± 0.07	87.500 ± 0.002	79.65 ± 0.01	61.015 ± 0.002
SemiNMF	75.81 ± 0.11	87.355 ± 0.001	76.30 ± 0.03	$\mathbf{62.215 \pm 0.002}$
CT	65.60 ± 0.12	84.520 ± 0.004	72.71 ± 0.06	47.207 ± 0.007
CCE	$\mathbf{76.49 \pm 0.47}$	$\mathbf{87.888 \pm 0.001}$	$\mathbf{80.05 \pm 0.01}$	61.670 ± 0.003

E.3. Ablation on regularization in CCE

To see the impact of the regularization step in the LearnSubspace step of CCE, we performan an additional ablation on the CLEVR dataset. We compare CCE without this regularization step to the full implementation of CCEin Table 6, and we see that regularization improves all three metrics.

E.4. Ablation on clustering loss function

We perform an ablation on the use of the Silhouette score as our clustering loss. Instead of Silhouette we experiment with the cross entropy loss based on the technique from Caron et al. (2018), but our results in Table 7 show that the Silhouette results in better compositionality.

Table 6. Regularization ablation on CLEVR.

Method	MAP	Comp. Score	Mean Cosine
CCE	$\mathbf{1.00 \pm 0.00}$	$\mathbf{3.41 \pm 0.18}$	$\mathbf{0.99 \pm 0.00}$
CCE-NoReg	0.97 ± 0.03	3.81 ± 0.21	0.78 ± 0.09

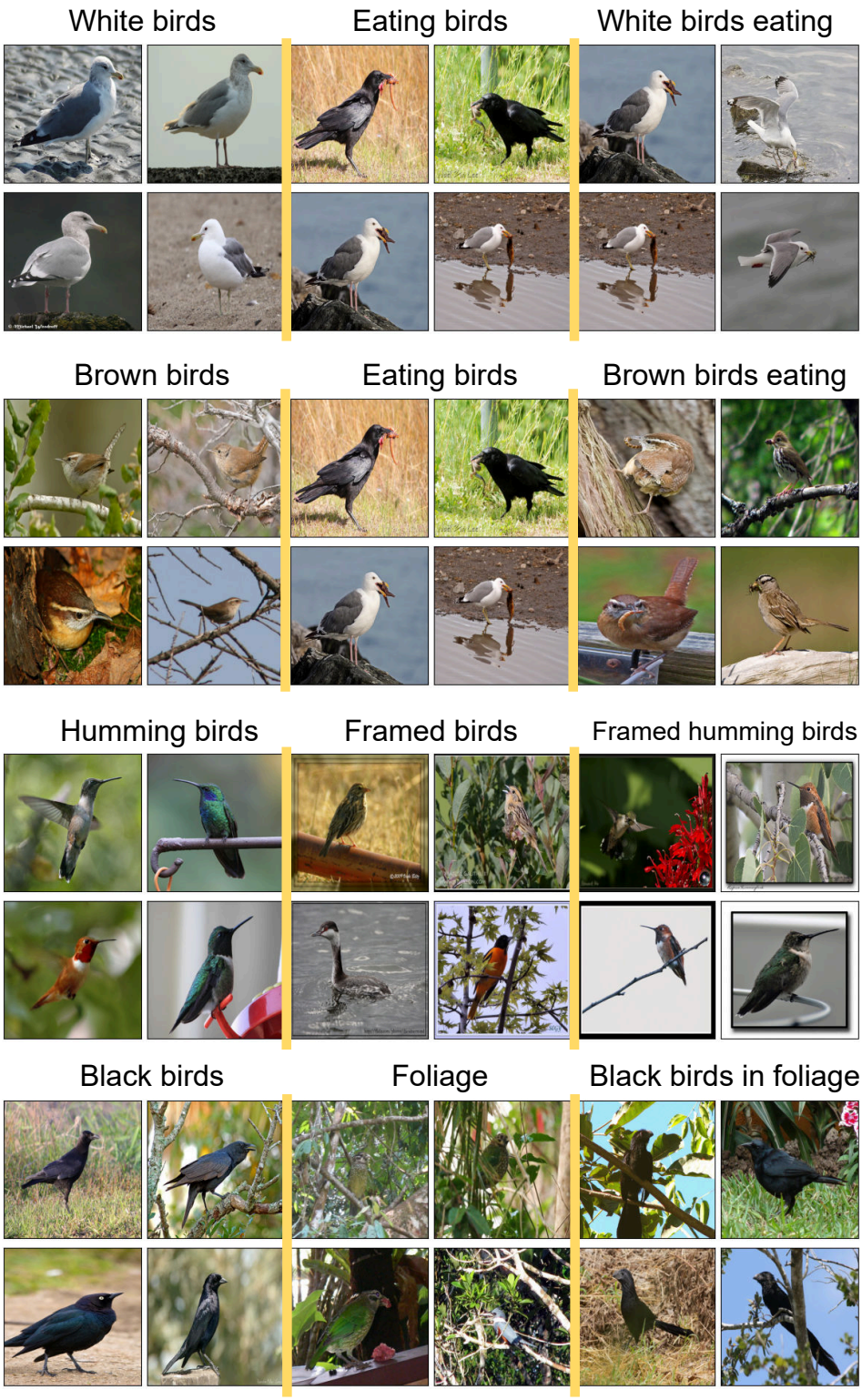


Figure 8. Additional CUB qualitative examples.

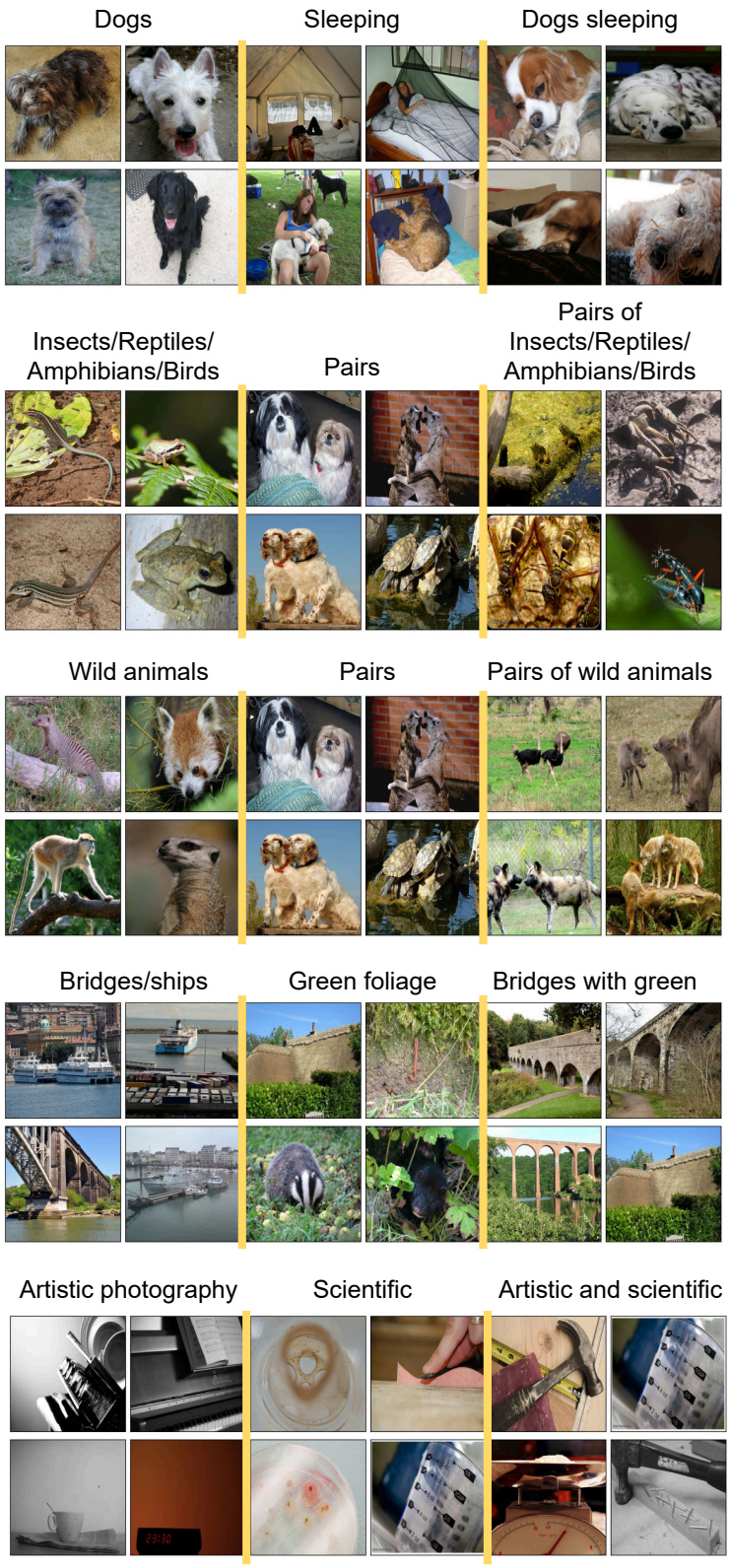


Figure 9. ImageNet qualitative examples.

Table 7. Loss function ablation on CLEVR.

Dataset	Loss	MAP	Comp. Score	Mean Cosine
CLEVR	Silhouette	**1.00** $\pm$ **0.00**	**3.41** $\pm$ **0.18**	**0.99** $\pm$ **0.00**
CLEVR	Cross Entropy	0.94 $\pm$ 0.08	3.44 $\pm$ 0.14	0.89 $\pm$ 0.10
Truth-sub	Silhouette	**0.56** $\pm$ **0.02**	**3.68** $\pm$ **0.01**	**0.81** $\pm$ **0.01**
Truth-sub	Cross Entropy	0.50 $\pm$ 0.04	3.94 $\pm$ 0.04	0.75 $\pm$ 0.02
CUB-sub	Silhouette	**0.65** $\pm$ **0.01**	**0.48** $\pm$ **0.00**	**0.77** $\pm$ **0.01**
CUB-sub	Cross Entropy	0.62 $\pm$ 0.04	0.49 $\pm$ 0.00	0.76 $\pm$ 0.01

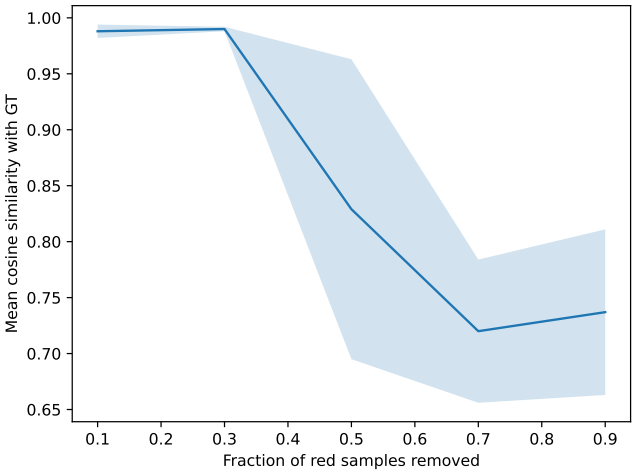


Figure 10. Cosine similarity between discovered "red" concept and the ground-truth "red" concept after removing a certain fraction of the red samples in the training set. As the attribute imbalance becomes larger, meaning there are less red samples than other colored samples, CCE performs worse at finding the true red concept.

E.5. Ablation on attribute imbalance

We perform an ablation experiment on the effect of attribute imbalance by testing CCE's ability to recover the ground truth concepts on the CLEVR dataset after removing different fractions of samples labeled with the "red" concept. The results are shown in Figure 10 where we see that removing more red samples, which creates a greater imbalance, decreases the average cosine similarity of the discovered concepts with the ground truth.

E.6. ROC-AUC Scores between Concept Representations and Ground-Truth

The maximum ROC-AUC between the concept score and the true label for the ground-truth concepts is presented in Table 8 for CLEVR, Table 9 for CUB-sub, and Table 10 for Truth-sub.

Table 8. Max AUC score CLEVR v/s GT

Concepts	CCE	ACE	ACE	PCA	DictLearn	SemiNMF
red	1.000	0.765	0.728	0.985	0.757	0.793
green	1.000	0.771	0.711	0.996	0.797	0.818
blue	1.000	0.753	0.745	0.972	0.782	0.836
sphere	1.000	1.000	0.736	1.000	1.000	1.000
cube	1.000	0.998	0.742	0.971	0.994	0.999
cylinder	1.000	0.998	0.831	0.977	0.992	0.998
(red and sphere) object	0.987	0.993	0.911	0.950	0.978	0.983
(red and cube) object	0.923	0.999	1.000	0.965	0.983	0.999
(red and cylinder) object	0.899	0.940	0.932	0.964	0.998	0.943
(green and sphere) object	0.858	0.991	0.870	0.863	0.980	0.986
(green and cube) object	0.878	1.000	1.000	0.877	0.951	1.000
(green and cylinder) object	0.936	0.916	0.960	0.969	1.000	0.994
(blue and sphere) object	0.952	0.996	1.000	0.834	0.940	0.997
(blue and cube) object	0.878	1.000	1.000	0.973	0.842	0.978
(blue and cylinder) object	0.923	0.992	1.000	0.990	0.995	0.995

Table 9. ROC AUC of baseline methods on recovering the labeled concepts.

Method	Brown	White	Black	Small	Medium	Large
GT	0.984	0.999	0.998	1.000	0.923	0.847
PCA	0.881	0.985	0.931	0.997	0.886	0.677
ACE	0.895	0.785	0.677	0.726	0.584	0.678
DictLearn	0.849	0.645	0.650	0.702	0.519	0.551
SemiNMF	0.086	0.164	0.099	0.116	0.066	0.168
CT	0.923	0.837	0.887	0.926	0.754	0.736
Random	0.867	0.933	0.855	0.888	0.849	0.723
CCE	0.894	0.834	0.710	0.743	0.656	0.661

E.7. The analysis of the cosine similarity score between learned concept representations and ground-truth

We further break down the results reported in Table 3 average cosine similarity between the learned concept representation and the ground-truth concept representations.

E.8. Ablation studies on other pretrained models

Recall that in the experiment section, we primarily focus on discovering concepts from pretrained CLIP model. In this section, we study with different choices of pretrained models, can we obtain similar results as that in Section 5?

To answer this question, we leverage vision transformer (ViT), another widely used pretrained vision model, to repeat the

Table 10. ROC AUC of baseline methods on recovering the labeled concepts.

Method	Truth	Animal	Company	Invention
GT	0.91	1.00	1.00	1.00
PCA	0.829	0.917	0.832	0.863
ACE	0.777	0.999	0.941	0.795
DictLearn	0.353	0.734	0.627	0.539
SemiNMF	0.759	0.708	0.629	0.521
CCE	0.91	1.00	0.96	0.78

Table 11. Max AUC score CLEVR v/s GT ViT

Concepts	CCE	ACE	PCA	DictLearn	SemiNMF
red	1.000	0.735	0.945	0.710	0.712
green	1.000	0.711	0.922	0.716	0.680
blue	1.000	0.642	0.995	0.704	0.629
sphere	1.000	0.610	1.000	1.000	1.000
cube	1.000	0.735	0.970	0.999	1.000
cylinder	1.000	0.695	1.000	1.000	1.000
(red and sphere) object	0.972	1.000	0.980	0.997	0.991
(red and cube) object	0.884	0.720	0.881	0.992	0.967
(red and cylinder) object	0.933	0.837	0.962	0.998	1.000
(green and sphere) object	0.904	1.000	0.923	0.998	0.985
(green and cube) object	0.913	0.731	0.886	0.920	0.937
(green and cylinder) object	0.895	0.660	0.866	0.988	0.939
(blue and sphere) object	0.939	0.844	0.970	0.954	0.949
(blue and cube) object	0.825	0.770	0.905	0.838	0.851
(blue and cylinder) object	0.854	0.766	0.842	0.913	0.875

experiments on CLEVR dataset. The results are summarized in Table 12-13. The results from these tables maintain the same trends as the one shown in Section 5.

Table 12. ViT results on CLEVR

Method	MAP	Comp. Score	Mean Cosine
GT	1.00 ± 0.00	3.69 ± 0.00	1.00 ± 0.00
PCA	0.90 ± 0.00	4.33 ± 0.00	0.64 ± 0.00
ACE	0.70 ± 0.05	4.36 ± 0.11	0.67 ± 0.00
DictLearn	0.80 ± 0.04	3.98 ± 0.06	0.70 ± 0.01
SemiNMF	0.76 ± 0.01	4.29 ± 0.02	0.67 ± 0.00
CT	0.58 ± 0.05	6.26 ± 0.00	0.04 ± 0.01
Random	0.64 ± 0.03	6.26 ± 0.00	0.05 ± 0.00
CCE	1.00 ± 0.00	3.87 ± 0.25	1.00 ± 0.00

F. Dataset Details

We provide the details for all datasets in Table 15.

G. Hyperparameters

The hyperparameters of all experiments are given in Table 16.

Table 13. ResNet-50 results on CLEVR

Method	MAP	Comp. Score	Mean Cosine
GT	0.95 ± 0.00	1.77 ± 0.00	1.00 ± 0.00
PCA	0.90 ± 0.00	2.08 ± 0.00	0.58 ± 0.00
ACE	0.77 ± 0.04	1.92 ± 0.02	0.68 ± 0.01
DictLearn	0.71 ± 0.08	1.95 ± 0.11	0.68 ± 0.01
SemiNMF	0.64 ± 0.00	2.01 ± 0.01	0.69 ± 0.00
CT	0.63 ± 0.08	2.83 ± 0.00	0.03 ± 0.00
Random	0.57 ± 0.03	2.83 ± 0.00	0.03 ± 0.00
CCE	0.92 ± 0.01	1.78 ± 0.01	0.96 ± 0.04

Table 14. Cosine similarity of baseline methods for recovering the labeled concepts.

Method	Truth	Animal	Company	Invention
PCA	0.367	0.139	0.688	0.583
ACE	0.244	0.956	0.733	0.642
DictLearn	0.760	0.988	0.917	0.879
SemiNMF	0.824	0.898	0.931	0.725
CCE	0.90	0.94	0.85	0.64

Table 15. Dataset details for all experiments

Dataset	Total Samples	Number of GT Concepts	Modality
CLEVR	1001	6	Image
CUB	11788	NA	Image
CUB-sub	261	6	Image
Truth	4127	NA	Text
Truth-sub	1125	4	Text
HAM	10015	NA	Image
News	18846	NA	Text

Table 16. Hyperparameters

Dataset	K	M	learning rate
CLEVR	3	3	0.001
CUB	20	5	0.001
CUB-sub	5	4	0.1
Truth	12	10	0.001
Truth-sub	[4, 2, 3]	3	0.001
HAM	20	25	0.02
News	15	30	0.001