Not Only Text: Exploring Compositionality of Visual Representations in Vision-Language Models

Davide Berasi[1] Matteo Farina[2] Massimiliano Mancini[2]
Elisa Ricci[1,2] Nicola Strisciuglio[3]
[1]Fondazione Bruno Kessler [2]University of Trento [3]University of Twente

Abstract

Vision-Language Models (VLMs) learn a shared feature space for text and images, enabling the comparison of inputs of different modalities. While prior works demonstrated that VLMs organize natural language representations into regular structures encoding composite meanings, it remains unclear if compositional patterns also emerge in the visual embedding space. In this work, we investigate compositionality in the image domain, where the analysis of compositional properties is challenged by noise and sparsity of visual data. We address these problems and propose a framework, called Geodesically Decomposable Embeddings (GDE), that approximates image representations with geometry-aware compositional structures in the latent space. We demonstrate that visual embeddings of pre-trained VLMs exhibit a compositional arrangement, and evaluate the effectiveness of this property in the tasks of compositional classification and group robustness. GDE achieves stronger performance in compositional classification compared to its counterpart method that assumes linear geometry of the latent space. Notably, it is particularly effective for group robustness, where we achieve higher results than task-specific solutions. Our results indicate that VLMs can automatically develop a human-like form of compositional reasoning in the visual domain, making their underlying processes more interpretable. Code is available at https://github.com/BerasiDavide/vlm_image_compositionality.

1. Introduction

Compositionality is the principle by which cognitive and computational systems create meaning of a complex expression by combining the meaning of its (simpler) parts [50, 51]. Humans leverage compositionality instinctively, combining known elements to interpret novel situations. In machine intelligence, efforts were made to replicate this capability by developing models that imitate compositional processes, *e.g.*, solving complex tasks via sub-goals [9, 31, 49, 61], modeling objects as compositions of their parts [13, 14, 47, 60, 62], encoding concept hierachies [11, 17, 33, 55], explicitly learning compositional representations [1, 18, 38, 44], or architectures [20, 24, 36, 63, 68].

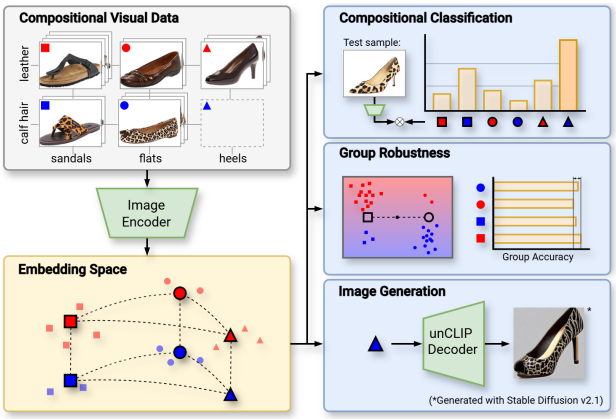

Figure 1. **Compositional structures in visual embedding space**. (*left*) Pre-trained VLM represents visual inputs of composite meanings in regular geometric shapes. The modularity of these structures enables the separation of the primitive components and the composition of unseen combinations. (*right*) We evaluate the usefulness of these properties in compositional classification, group robustness, and image generation.

With the rise of modern Vision-Language Models (VLMs) [26, 54, 74] jointly trained on large-scale image-text pairs, there has been growing interest in investigating whether these models exhibit instrinsic compositional behaviors [45, 52, 71]. In particular, Trager *et al.* [67] investigated latent compositional structures within the CLIP [54] text embedding space, demonstrating that composite concepts can be represented as linear combinations of embedding vectors corresponding to various factors. These vectors, called *ideal words*, can be used to compose new concepts in the embedding space. Their work focuses on finding compositional structures in the text embedding space of

Corresponding author: dberasi@fbk.eu.

CLIP, motivated by the fact that the structured and symbolic nature of language may facilitate the study of computational approaches to capture compositional meaning. However, cognitive studies show that language itself is used to describe and interpret the visual world and directly affects visual perception [6]. Hence, similar to text, human visual representations exhibit a compositional structure [19], made of simpler components systematically combined. Despite this connection, compositional properties of visual embeddings of VLMs have remained so far mostly unexplored.

To fill this gap, in this paper, we introduce GEODESICALLY DECOMPOSABLE EMBEDDINGS (GDE), a framework grounded in differential geometry and designed to investigate compositional structures of pre-trained embeddings within Riemannian manifolds (Fig.1). Visual embeddings exhibit unique challenges not present in compositional analysis of text embeddings, namely *data sparsity* in the compositional space and *noise and ambiguity* in images. Specifically, we deal with the sparsity of composite concepts, as certain combinations of elementary primitives may not appear in real image collections (*e.g.*, focusing on objects and attributes, "blue dog" images are unlikely to exist). Noise and ambiguity concern additional visual cues and information present in images, e.g. background, context, etc., that do not correspond to the composite concepts. We evaluated the compositional representations computed with the proposed approach in two relevant applications, namely compositional classification and group robustness (Fig. 1), considering publicly available datasets, showing that it better captures visual compositional structures than the alternatives (*e.g.*, [10]). GDE is particularly effective for group robustness, where we achieve better debiasing results than task-specific methods. Furthermore, we show that GDE can be successfully used in combination with state-of-the-art generative models to synthesize images of compositional concepts. Our contributions are:

i) We study compositional structures within visual embeddings for VLMs and demonstrate that the latent representations of visual signals also exhibit a degree of compositionality.
ii) We show that, unlike for text embeddings, linear structures are insufficient to (de)compose visual concepts; thus, the manifold geometry must be considered.
iii) We propose a framework that deals with the sparsity and noise of composite concepts in images, enabling the compositional analysis of visual embeddings.

2. Related Work

Compositionality in Vision. Compositionality is considered a cornerstone of perception [34], and compositional representations offer an effective tool to represent real-world phenomena [12]. The primary benefit of compositionality is the possibility of combining the representation of simpler concepts to understand and reason on complex ones, allowing for generalization to new unseen combinations of concepts [31, 44, 60]. In computer vision, early efforts focused on recognizing objects as composition of parts [13, 14, 47, 48] and evolved into architectures that can recognize and model objects in a compositional fashion [60], compositional generation [49, 64, 76], and interpretable representations [4, 63]. Compositionality has also lead to progress in various tasks, such as human-object interaction detection, model spatial/semantic relationships [21, 22, 28], and compositional zero-shot learning, where the goal is to recognize unseen compositions of training primitives [37, 41–44]. While these works focus on specific applications, in this paper we aim to study whether there exists an underlying compositional structure in the visual embeddings of VLMs.

Compositionality in VLMs. Modern Vision-Language Models (VLMs) like CLIP [54] are trained to extract meaningful representations from complex visual scenes guided by textual inputs without a priori imposing any form of compositionality. In this context, a natural question is: *Does compositional behavior emerge automatically in VLMs?*

Previous works already showed how VLMs are more suitable for tasks such as compositional zero-shot learning [40, 45, 52], and how their representations allow for cross-modal compositions, such as visual editing [5, 29, 75] and compositional retrieval [3, 23, 27, 59]. At the same time, works studied the challenges of VLMs in model compositional inputs, *e.g.*, at the level of word order, object-attribute bindings, spatial relationships and other compositional challenges [23, 65, 66, 70].

In this paper, we study the compositional structure in the visual embeddings extracted from VLMs. Close to our goal is [35], studying the compositional properties of the CLIP text encoder through compositional distributional semantics models in synthetic test scenarios. Similarly, [67] show that the textual embeddings of VLMs can be well approximated by linear compositions of smaller sets of ideal vectors. Motivated by the cross-modal alignment of VLMs, we investigate whether the embeddings of visual inputs exhibit an analogous compositional property. We achieve this by constructing a geometry-aware decomposition framework, following ideas similar to [46], where Principal Geodesic Analysis (PGA) [15] is applied to learn lower-dimensional submanifolds of the CLIP sphere that are associated to distinct parts-of-speech. To the best of our knowledge, this is the first work that investigates the emergence of compositional structures in the visual embeddings of VLMs.

3. Method

We propose a framework to analyze the compositional properties of image embeddings of neural encoders. We start by reviewing the fundamentals of the CLIP model along

with key concepts from differential geometry (Sec. 3.1). We then formalize the concept of geodesic decomposability (Sec. 3.2) and we discuss our methodology for dealing with visual inputs (Sec. 3.3).

3.1. Preliminaries

Contrastive Language-Image Pretraining (CLIP) consists of a pre-trained image encoder $\phi_{im} : \mathcal{X} \to \mathbb{R}^d$ and a text encoder $\phi_t : \mathcal{Y} \to \mathbb{R}^d$ that represent multi-modal text-visual inputs in a shared vision-language space. The latent representations of an image $x \in \mathcal{X}$ and text $y \in \mathcal{Y}$ are compared by cosine similarity, which is the scalar product $\mathbf{u}_x^\top \mathbf{u}_y$ of their normalized versions $\mathbf{u}_x = \phi_{im}(x)/||\phi_{im}(x)||$, $\mathbf{u}_y = \phi_t(y)/||\phi_t(y)||$. The weights of the encoders are trained to optimize a contrastive objective on a huge collection of paired image-text samples. Since the norm of CLIP embeddings does not carry any meaningful information, spherical geometry applies to their post-hoc analysis.

Riemannian Manifolds are geometric spaces where intrinsic distances can be measured. For a generic manifold $\mathcal{M} \subset \mathbb{R}^d$ with intrinsic distance $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to [0, \infty)$, we now recall the notions of *exponential map* and *intrinsic mean*. These tools permit operating with non-linear data, like the spherical normalized CLIP embeddings, while respecting their intrinsic shape. Let μ be a point on $\mathcal{M}$ and let $T_\mu\mathcal{M}$ be the tangent space in μ. The *exponential map* projects a tangent vector $\mathbf{v} \in T_\mu\mathcal{M}$ onto the manifold by moving along the geodesic segment it defines. Formally, if $\gamma_\mathbf{v} : [0, 1] \to \mathcal{M}$ is the unique geodesic path starting from $\gamma_\mathbf{v}(0) = \mu$ with initial velocity $\dot{\gamma}_\mathbf{v}(0) = \mathbf{v}$, then $\mathrm{Exp}_\mu(\mathbf{v}) := \gamma_\mathbf{v}(1)$. This function is locally invertible and its inverse is the logarithmic map $\mathrm{Log}_\mu = \mathrm{Exp}_\mu^{-1}$. The exponential and logarithmic maps send straight lines of the tangent plane into geodesic curves of the manifold, and vice-versa. Moreover, they approximately preserve distances between elements close to the point of tangency μ:

$$d_{\mathcal{M}}(\mathbf{u}, \mathbf{u}') \approx ||\,\mathrm{Log}_\mu(\mathbf{u}) - \mathrm{Log}_\mu(\mathbf{u}')||, \quad \mathbf{u}, \mathbf{u}' \in \mathcal{M} \quad (1)$$

Note that in (1) the equality holds if $\mathbf{u} = \mu$ or $\mathbf{u}' = \mu$. When applying the logarithmic map to a set of points $\{\mathbf{u}_i\}_{i=1}^N \subset \mathcal{M}$, the natural choice for the point of tangency μ is the *intrinsic mean*, *i.e.*, the element of $\mathcal{M}$ minimizing the average squared distance to the given points. In a more general definition, each point $\mathbf{u}_i$ $(i = 1, \ldots, N)$ is associated to a scalar weight w_i belonging to a probability-simplex vector Δ_N and the (weighted) intrinsic mean is:

$$\mu = \arg\min_{\mathbf{u} \in \mathcal{M}} \sum_{i=1}^{N} w_i \, d_{\mathcal{M}}(\mathbf{u}, \mathbf{u}_i)^2 \quad (2)$$

This distance-minimizing element μ guarantees that the images of the points through the logarithmic map are centered in the origin of the tangent space: $\sum_i w_i \,\mathrm{Log}_\mu(\mathbf{u}_i) = 0$.

3.2. Geodesically Decomposable Embeddings

We now formalize our proposed notion of compositional embeddings. We consider a set of composite meanings $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_s$, defined as the Cartesian product between finite lists of primitive concepts, and refer to the $\mathcal{Z}_i$ $(i = 1, \ldots, s)$ as the *dimensions* of $\mathcal{Z}$. For example, $\mathcal{Z} = \{\text{red, blue}\} \times \{\text{car, dress, flower}\}$ combines primitives from an attribute dimension and an object dimension.

We then consider an embedding map $\phi : \mathcal{Z} \to \mathcal{M}$ representing the composite concepts as points on a manifold $\mathcal{M} \subset \mathbb{R}^d$. Intuitively, the set $\phi(\mathcal{Z}) = \{\mathbf{u}_z \,|\, z \in \mathcal{Z}\}$ is compositional if it has a regular structure reflecting the composite nature of the inputs, *i.e.*, if one can *compose* primitive concepts within the geometric space to obtain embeddings of complex meanings. In this paper, we associate compositionality to the notion of *geodesic decomposability* which accounts for the intrinsic geometry of the manifold.

Definition 1 (Geodesically decomposable embeddings)**.** A set of embeddings $\phi(\mathcal{Z}) = \{\mathbf{u}_z \,|\, z \in \mathcal{Z}\} \subset \mathcal{M}$ with intrinsic mean μ is *geodesically decomposable* if there exist $\mathbf{v}_{z_i} \in T_\mu\mathcal{M}$ for all $z_i \in \mathcal{Z}_i$ $(i = 1, ..., s)$ such that

$$\mathbf{u}_z = \mathrm{Exp}_\mu(\mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s}) \qquad \forall z = (z_1, \ldots, z_s) \quad (3)$$

Note that in a decomposable set $\phi(\mathcal{Z})$ a new valid decomposition is obtained by adding the same tangent vector to all $\mathbf{v}_{z_i}$ and subtracting it from all $\mathbf{v}_{z_j}$, for any $i \neq j$. However, we can guarantee the uniqueness of the factorization by imposing a centering constraint.

Lemma 1. *Let $\phi(\mathcal{Z})$ be a geodesically decomposable set. Then there exist unique vectors $\mathbf{v}_{z_i} \in T_\mu\mathcal{M}$ for all $z_i \in \mathcal{Z}_i$ such that $\sum_{z_i \in \mathcal{Z}_i} \mathbf{v}_{z_i} = 0$ for all $i = 1, \ldots, s$ and Eq. (3) holds.*

For an intuitive interpretation, the intrinsic mean μ of a decomposable set can be seen as the *context* of the decomposition, and each unique direction $\mathbf{v}_{z_i}$ represents the meaning of the primitive concept z_i relative to μ. These "universal directions" are combined by addition on the tangent space $T_\mu\mathcal{M}$. The exponential map of the resulting tangent vector defines the geodesic segment on the manifold $\mathcal{M}$ from μ to the corresponding composite meaning (see Fig. 2).

Our notion of geodesic decomposability is general and applicable to manifolds of any shape. It generalizes that of [67], which is equivalent to ours in the special case $\mathcal{M} = \mathbb{R}^n$, where the intrinsic mean is the arithmetic mean, and the exponential and logarithmic maps behave like the identity function. Our manifold formalization agrees with the fact that lower-dimensional semantic subspaces in CLIP latent space are captured by submanifolds better than linear subspaces [46].

Best decomposable approximation. Decomposable sets live in a lower dimension subspace of their manifold $\mathcal{M}$.

The dimension of $Span(\{\mathbf{v}_{z_i}\}_{z_i \in \mathcal{Z}_i})$ is indeed at most $|\mathcal{Z}_i| - 1$ for all $i = 1, \dots, s$, implying the additive combinations of the primitive directions belong to a subspace of dimension at most $\sum_i (|\mathcal{Z}_i| - 1)$. This suggests that a generic set of embeddings $\{\mathbf{u}_z\}$ is unlikely to be perfectly decomposable. We thus search for its best decomposable approximation, that is the set $\{\tilde{\mathbf{u}}_z\}$ that minimizes the error

$$\sum_{z \in \mathcal{Z}} d_{\mathcal{M}}(\mathbf{u}_z, \tilde{\mathbf{u}}_z)^2 \tag{4}$$

In general, this is a hard problem to solve. Similarly to the standard solution to Principal Geodesic Analysis [15], we use Eq. (1) to approximate the objective in the "simpler" Euclidean space $T_\mu \mathcal{M}$, and rewrite Eq. (4) as:

$$\sum_{z \in \mathcal{Z}} || \operatorname{Log}_\mu(\mathbf{u}_z) - \operatorname{Log}_\mu(\tilde{\mathbf{u}}_z)||^2, \tag{5}$$

The solution to the approximate problem is obtained by computing vector means in $T_\mu \mathcal{M}$, as described in the next proposition. For a fixed primitive concept $z_i \in \mathcal{Z}_i$, let $\mathcal{Z}(z_i) = \{(z'_1, \dots, z'_r) \in \mathcal{Z} \,|\, z'_i = z_i\}$ denote the *slice* of $\mathcal{Z}$ containing all tuples with the i-th component equal to z_i.

Proposition 1. *Given a set $\phi(\mathcal{Z}) = \{\mathbf{u}_z \,|\, z \in \mathcal{Z}\} \subset \mathcal{M}$ with intrinsic mean μ, the minimization problem*

$$\begin{aligned} &\arg\min_{\{\tilde{\mathbf{u}}_z\}} \sum_{z \in \mathcal{Z}} || \operatorname{Log}_\mu(\mathbf{u}_z) - \operatorname{Log}_\mu(\tilde{\mathbf{u}}_z)||^2, \\ &\quad s.t.\ \{\tilde{\mathbf{u}}_z\} \textit{ is geodesically decomposable} \end{aligned} \tag{6}$$

is solved by $\tilde{\mathbf{u}}_z = \operatorname{Exp}_\mu(\mathbf{v}_{z_1} + \dots + \mathbf{v}_{z_r})$, where

$$\mathbf{v}_{z_i} = \frac{1}{|\mathcal{Z}(z_i)|} \sum_{z \in \mathcal{Z}(z_i)} \operatorname{Log}_\mu(\mathbf{u}_z) \tag{7}$$

Moreover, $\sum_{z_i \in \mathcal{Z}_i} \mathbf{v}_{z_i} = 0$ for all $i = 1, \dots, s$.

This result tells us that each vector $\mathbf{v}_{z_i}$ in the optimal decomposition is the tangent mean of all the input compositions including the primitive z_i. Moreover, the choice of the intrinsic mean as the point of tangency guarantees the uniqueness constraint is satisfied (see Appendix for details).

3.3. Decomposable Embeddings of Visual Inputs

Our framework holds for arbitrary manifolds and for any embedding map, hence being independent of the input modality. However, collections of natural visual data contain *noise* and are *sparse*. We account for these properties in our framework as presented in the following.

3.3.1 Removing *noise* from finite image sets

We refer to *noise* as information carried by images in addition to the composite concept of interest. For example, an image from the tuple $z = $ (red, car) likely contains non-negligible extra information, e.g. a driver, a road, or a blue sky in the background. This stems from the inherent ambiguity and non-uniqueness of visual signals. Most importantly, it is absent in text, for which it is easier to manually craft the string "a red car" ensuring no extra information.

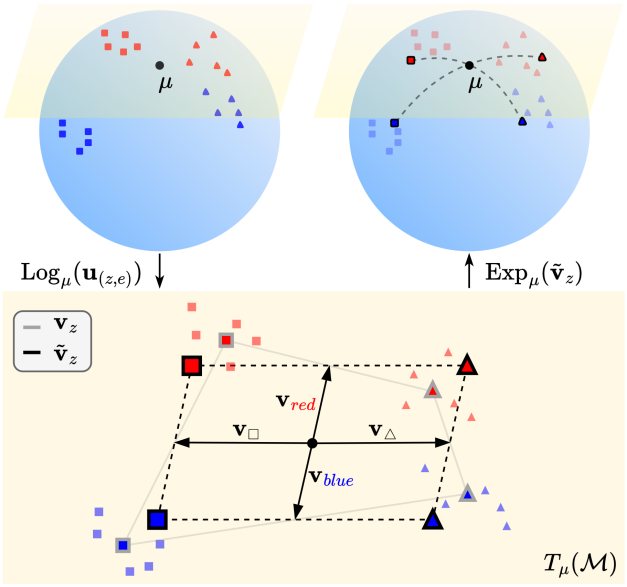

Figure 2. **Sketch of our decomposition method**. (*top-left*) Each concept in $\mathcal{Z} = \{red, blue\} \times \{\square, \triangle\}$ is represented by $k = 5$ embeddings on a manifold. (*bottom*) These are mapped in the tangent space where optimal primitive directions are computed as vector means and combined by addition. (*top-right*) The obtained compositions are mapped back to the manifold to obtain a decomposable approximation of the input embeddings.

Problem formulation. Since images contain noise in addition to represented concepts, we consider an input set $\phi(\mathcal{Z} \times \mathcal{E}) = \{\mathbf{u}_{(z,e)} \,|\, (z, e) \in \mathcal{Z} \times \mathcal{E}\}$ where each z is represented by $k = |\mathcal{E}|$ different image embeddings varying along the unknown noise dimension $\mathcal{E}$. Also, different images may contain different amounts of noise. For each fixed z, we model this aspect with a probability distribution $\{p_{(z,e)}\}_{e \in \mathcal{E}}$ describing how well the elements in $\{\mathbf{u}_{(z,e)}\}_{e \in \mathcal{E}}$ represent their label z. In this setting, we want the decomposable set $\{\tilde{\mathbf{u}}_z\}_{z \in \mathcal{Z}}$ minimizing the objective

$$\sum_{(z,e) \in \mathcal{Z} \times \mathcal{E}} p_{(z,e)} d_{\mathcal{M}}(\mathbf{u}_{(z,e)}, \tilde{\mathbf{u}}_z)^2, \tag{8}$$

where the importance given to the approximation error for each input embedding is weighted according to the noise distribution. The next result generalizes Proposition 1, which addresses the special case $k = 1$, and provides an easy-to-compute approximate solution to the problem.

Proposition 2. *Let $p_{(z,e)}$, $(z, e) \in \mathcal{Z} \times \mathcal{E}$, be non-negative scalars such that $\sum_{e \in \mathcal{E}} p_{(z,e)} = 1$ for each $z \in \mathcal{Z}$, and let*

$\phi(\mathcal{Z}\times\mathcal{E}) = \{\mathbf{u}_{(z,e)} \,|\, (z,e)\in\mathcal{Z}\times\mathcal{E}\} \subset \mathcal{M}$ *be a set of embeddings with weighted intrinsic mean* μ *w.r.t. the weights* $w_{(z,e)} = p_{(z,e)}/\sum_{(z,e)} p_{(z,e)}$. *The minimization problem:*

$$\begin{aligned}\arg\min_{\{\tilde{\mathbf{u}}_z\}} \sum_{(z,e)\in\mathcal{Z}\times\mathcal{E}} p_{(z,e)} ||\operatorname{Log}_\mu(\mathbf{u}_{(z,e)}) - \operatorname{Log}_\mu(\tilde{\mathbf{u}}_z)||^2,\\ s.t.\ \{\tilde{\mathbf{u}}_z\} \text{ is geodesically decomposable}\end{aligned} \tag{9}$$

is solved by $\tilde{\mathbf{u}}_z = \operatorname{Exp}_\mu(\mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s})$, *where*

$$\mathbf{v}_{z_i} = \frac{1}{|\mathcal{Z}(z_i)|}\sum_{z\in\mathcal{Z}(z_i)} \mathbf{v}_z, \quad \mathbf{v}_z = \sum_{e\in\mathcal{E}} p_{(z,e)} \operatorname{Log}_\mu(\mathbf{u}_{(z,e)}) \tag{10}$$

Moreover, $\sum_{z_i\in\mathcal{Z}_i} \mathbf{v}_{z_i} = 0$ *for all* $i = 1,\dots,s$.

Fig. 2 visualizes the decomposition procedure. Notice that, using the same notation of the proposition, the vectors $\mathbf{v}_z$ can be seen as a denoised tangent representation of the tuples in $\mathcal{Z}$ and the solution $\{\tilde{\mathbf{u}}_z\}_{z\in\mathcal{Z}}$ to the weighted optimization problem corresponds to the decomposable approximation given by Proposition 1 applied to the denoised embeddings $\{\mathbf{u}_z := \operatorname{Exp}_\mu(\mathbf{v}_z)\}_{z\in\mathcal{Z}}$. Indeed, these have intrinsic mean equal to the weighted intrinsic mean μ.

Lemma 2. *Using the notation of Proposition 2, the set* $\{\mathbf{u}_z := \operatorname{Exp}_\mu(\mathbf{v}_z)\}_{z\in\mathcal{Z}}$ *has intrinsic mean* μ.

3.3.2 Dealing with *sparsity* in finite image sets

The previously described setup assumes that every $z\in\mathcal{Z}$ is represented by $k > 0$ images. This requirement can be too restrictive in practice, because some combinations of primitives may not occur in real image collections. For example, if $\mathcal{Z} = \{\text{red, blue}\}\times\{\text{car, apple}\}$, there will probably be no pictures of a (blue, apple). We refer to the absence of composite concepts as *sparsity*. Once more, please note that sparsity is not an issue with text, since strings can be manually crafted for any $z\in\mathcal{Z}$.

Problem Formulation. In general, in a labeled image collection, only a subset $\mathcal{T}\subset\mathcal{Z}\times\mathcal{E}$ is available, and only a subgroup $\mathcal{Z}'\subset\mathcal{Z}$ of composite concepts is represented by at least one element in $\mathcal{T}$. In this scenario, we obtain a decomposable approximation of $\phi(\mathcal{T})$ by approximating the vector means in Eq. (10) with the mean of the available elements. The only requirement is that every primitive $z_i\in\mathcal{Z}_i$ $(i=1,\dots,s)$ appears in at least one tuple of $\mathcal{Z}'$. Precisely, we first compute the weighted intrinsic mean μ of $\phi(\mathcal{T})$ with weights $w_{(z,e)} = p_{(z,e)}/\sum_{(z,e)\in\mathcal{T}} p_{(z,e)}$, and then consider $\tilde{\mathbf{u}}_z = \operatorname{Exp}_\mu(\mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s})$, where:

$$\mathbf{v}_{z_i} = \frac{1}{|\mathcal{Z}'(z_i)|}\sum_{z\in\mathcal{Z}'(z_i)} \mathbf{v}_z, \quad \mathbf{v}_z = \sum_{\substack{e\in\mathcal{E}\ s.t.\\ (z,e)\in\mathcal{T}}} p_{(z,e)} \operatorname{Log}_\mu(\mathbf{u}_{(z,e)}) \tag{11}$$

Note that the obtained decomposable set contains vector representations of all the concepts in $\mathcal{Z}$, including the unseen elements of $\mathcal{Z}\setminus\mathcal{Z}'$. The formulation in (11) deals with all aspects mentioned so far: the manifold $\mathcal{M}$, noise, and sparsity. In the next section, we use it to evaluate the compositional structure of real visual embeddings.

Noise distribution. The described setup requires the noise scores $p_{(z,e)}$. Given a collection of visual inputs $\mathcal{T}$ representing each label $z\in\mathcal{Z}'$ with $k_z > 0$ elements, a simple choice is using uniform scores $p_{(z,e)} = 1/k_z$. Alternatively, we propose using the CLIP image-to-text distribution $p_{(z,e)} = \mathbb{P}((z,e)|y(z))$, where $y(z)$ is a text prompt for label $z\in\mathcal{Z}'$. This is the softmax of the scaled similarities

$$\mathbb{P}((z,e)|y(z)) = \frac{\exp(\mathbf{u}_{(z,e)}^\top \mathbf{u}_{y(z)}/t)}{\sum_e \exp(\mathbf{u}_{(z,e)}^\top \mathbf{u}_{y(z)}/t)} \tag{12}$$

The temperature parameter t is learned during training, but it can be tweaked to smooth or sharpen the distribution.

4. Experimental Validation

We carry out experiments to analyze the decomposable properties of visual embeddings of VLMs. When not specified differently, we use the pre-trained CLIP ViT-L/14 [54]. We also consider CLIP ResNet50 [54] and SigLIP [72]. All considered models are from the OpenCLIP repository [8]. We use images with attribute-object labels to represent sets of composite concepts of the form $\mathcal{Z} = \mathcal{Z}_{attr}\times\mathcal{Z}_{obj}$.

In this setup, we first assess the decomposable nature of small sets of embeddings inspecting their geometric arrangement according to Proposition 2 (Sec. 4.1). Then, we leverage the structured nature of the decomposed embeddings and experiment on the tasks of compositional classification (Sec. 4.2) and group robustness (Sec. 4.3). Finally, we visualize the approximate decomposable embeddings using a diffusion model (StableDiffusion v2.1 [57]) with the unCLIP technique [56] (Sec. 4.4).

Attribute-object decomposition. We usually deal with sparse collections of visual inputs $\mathcal{T}$ where only a subset $\mathcal{Z}'$ of labels present at least one image. Thus, we compute the embedding decomposition according to Eq. (11): the optimal vectors $\tilde{\mathbf{u}}_{(a,o)} = \operatorname{Exp}_\mu(\mathbf{v}_a + \mathbf{v}_o)$ are the combinations of the attribute directions $\mathbf{v}_a = \frac{1}{\mathcal{Z}'(a)}\sum_o \mathbf{v}_{(a,o)}$ and the object directions $\mathbf{v}_o = \frac{1}{\mathcal{Z}'(o)}\sum_a \mathbf{v}_{(a,o)}$, where the denoised representations $\mathbf{v}_{(a,o)}$, $(a,o)\in\mathcal{Z}'$, are the mean tangent vectors within pairs. For compositional classification and group robustness, we use the CLIP image-to-text probabilities as the noise distribution discussed in Sec. 3.3.2. We finetune the temperature parameter (see Appendix for details). In the other experiments, we utilize uniform scores.

Datasets. We represent composite concepts with images from the training sets of diverse compositional datasets. We test compositional classification on the typical benchmark

datasets UT-Zappos [69] and MIT-states [25] with the splits from [53]. UT-Zappos contains images of shoes centered on a white background all sharing the same orientation. There are 12 object classes referring to the footwear type and 16 attribute categories referring to the material. MIT-states is a collection of natural objects in different states. The dataset contains 115 attribute categories and 245 object categories, generating a large number of possible combinations.

We test group robustness on the Waterbirds and CelebA datasets with the splits in [58]. These contain objects with spuriously correlated attributes, making them suitable for debiasing tasks. Waterbirds contains images of two bird species $\mathcal{Z}_{obj}$={waterbird, landbird} on two types of background $\mathcal{Z}_{attr}$={land, water}. We use the version of CelebA from [58] that contains close-ups photos of celebrities labeled with hair-color $\mathcal{Z}_{obj}$ = {blonde, dark} and gender $\mathcal{Z}_{attr}$={male, female}. The data distribution over the four different groups is highly unbalanced in the train sets of these two datasets, implying spurious correlations.

4.1. Visualizing Compositional Embeddings

We evaluated the decomposability of the embeddings from a geometric perspective. We visualize lower-dimensional PCA projections of the tangent vectors $\{\mathbf{v}_{(a,o)}\}$, considering that the denoised representations $\mathbf{u}_{(a,o)} := \mathrm{Exp}_{\mu}(\mathbf{v}_{(a,o)})$ are geodesically decomposable if and only if their tangent directions are the vertices of a geometric shape with parallel faces. For example, decomposable sets of size $|\mathcal{Z}| = 2 \times 2$ and $|\mathcal{Z}| = 2 \times 3$ correspond to a parallelogram and a triangular prism, respectively.

In Fig. 3 we show the (first row) 2-D projection of image embeddings from the Waterbirds dataset, representing the four compositions of two attributes and two objects, and the (second row) 3-D projection of the two-by-three concepts in the set {leather, suede} × {boots ankle, boots knee high, shoes flats} of the UT-Zappos dataset. By increasing the number k of images per pair, the noise is successfully removed and the resulting representations define shapes with parallel faces, indicating approximate geodesical decomposability. This highlights the importance of the denoising step and demonstrates compositional regularities of visual embeddings.

4.2. Compositional Classification

We perform compositional classification on the UT-Zappos and MIT-states datasets using the decomposable approximation of the train data as classifiers. This task serves to evaluate the generalization capabilities towards novel compositions of objects and states. Specifically, we follow the standard generalized zero-shot evaluation protocol in both closed-world and open-world scenarios [41].

We compute decomposable embeddings on a subset $\mathcal{Z}' \subset \mathcal{Z}$ of *seen* pairs from the training set, while not all

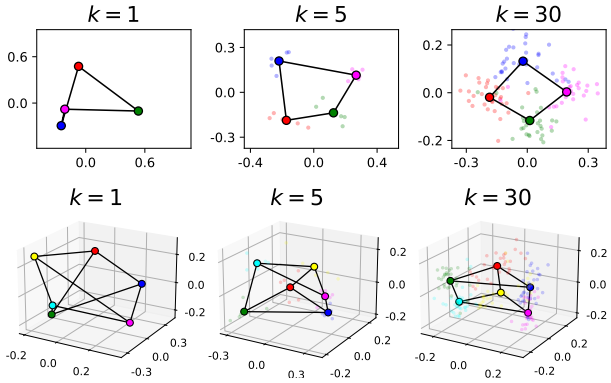

Figure 3. (*top*) 2-D projections of image embeddings representing the 2×2 composite labels of the Waterbirds dataset. (*bottom*) 3-D projections of image embeddings representing 2×3 composite labels with images from the UT-Zappos dataset. Denoised pair representations (marked with a black contour) are computed with $k = 1, 5, 30$ randomly selected images.

labels in the test set are in $\mathcal{Z}'$. In the closed-world setting, the set of target labels $\mathcal{Z}^{test} \subset \mathcal{Z}$ contains only the pairs appearing in the dataset, while in the open-world framework, no prior knowledge is assumed and all the attribute-object combinations in $\mathcal{Z}^{test} = \mathcal{Z}$ are considered. Both settings require generalizing the prior knowledge about the primitives to understand the *unseen* compositions in $\mathcal{Z}^{test} \setminus \mathcal{Z}'$. This operation is particularly challenging in the open-world scenario, where the more numerous novel compositions in the test set are a distraction for the predictor.

Our framework provides a straightforward solution to the complex problem of compositional classification. The geodesically decomposable set $\{\tilde{\mathbf{u}}_{(a,o)}\}$ computed with the full train data $\mathcal{T}$ represents all the pairs, including the unseen ones. Thus we classify an image x as $\arg\max_{(a,o)\in\mathcal{Z}^{test}} \tilde{\mathbf{u}}_{(a,o)}^{\top}\mathbf{u}_x$. We evaluate the prediction with the standard metrics [7, 53]: attribute accuracy (ATTR), object accuracy (OBJ), best seen accuracy (SEEN), best unseen accuracy (UNSEEN), best harmonic mean (HM) between the seen and unseen accuracy and area under the seen-unseen curve (AUC).

Baselines. Our primary goal is to examine if the Geodesically Decomposable Embeddings (GDE) approximating the train data contain semantically meaningful information about the composite concepts they represent. We evaluate the relative performance ρ (AUC ratio) obtained with decomposed embeddings w.r.t. the results achieved with the standard zero-shot baseline (CLIP) using the full-state embeddings (attribute-object labels $(a, o) \in \mathcal{Z}^{test}$ are represented by the text embedding of "An image of a {a} {o}").

We investigate the importance of complying with data geometry and compare with the Linearly Decomposable Embeddings (LDE) proposed in [67], which we compute

Dataset	Method	Closed-world							Open-world						
		Attr	Obj	Seen	Unseen	HM	AUC	ρ	Attr	Obj	Seen	Unseen	HM	AUC	ρ
UT-Zappos	CLIP [54]	*24.1*	*58.3*	*11.9*	*45.7*	*15.3*	*4.4*	-	*18.8*	*57.4*	*11.9*	*23.8*	*12.0*	*2.3*	-
	LDE (text)* [67]	24.1	58.8	11.9	45.7	14.1	4.0	92.4 %	**19.2**	57.2	11.9	20.0	11.1	1.9	83.2 %
	GDE (text)	**25.3**	**60.0**	**17.0**	**48.2**	**18.9**	**6.4**	**146.6 %**	18.7	**59.9**	**17.0**	**21.4**	**12.2**	**2.5**	**111.1 %**
	LDE (image)	13.9	52.6	5.6	32.1	6.6	0.9	21.1 %	9.8	48.0	5.6	14.9	2.3	0.2	8.9 %
	GDE (image)	**36.3**	**64.1**	**31.4**	**55.9**	**29.3**	**13.9**	**317.9 %**	**28.6**	**61.7**	**31.3**	**33.3**	**19.0**	**6.7**	**293.5 %**
MIT-states	CLIP [54]	*33.0*	*52.1*	*30.6*	*45.3*	*26.3*	*11.1*	-	*15.6*	*47.7*	*30.6*	*8.3*	*8.4*	*1.7*	-
	LDE (text)* [67]	30.6	51.2	24.7	43.0	21.9	8.2	73.4 %	21.1	**50.7**	24.7	**13.8**	11.9	2.5	148.1 %
	GDE (text)	**32.6**	**51.7**	**27.8**	**45.2**	**24.5**	**10.0**	**89.7 %**	**21.3**	49.9	**27.8**	13.0	**12.1**	**2.6**	**158.5 %**
	LDE (image)	15.3	30.5	15.0	20.9	11.1	2.0	18.4 %	11.0	34.8	15.0	5.6	4.6	0.4	27.1 %
	GDE (image)	**28.1**	**45.3**	**30.7**	**36.1**	**23.4**	**8.6**	**77.7 %**	**18.5**	**43.6**	**29.7**	**8.5**	**9.3**	**1.8**	**106.6 %**

Table 1. Compositional classification results on the UT-Zappos and MIT-states datasets. Highest values within modality are in **bold** and "*" indicates that the results of our implementation are shown.

Dataset	Method	CLIP, RN50							CLIP, ViT-L/ 14							SigLIP, ViT-SO400M/ 14						
		Attr	Obj	Seen	Unseen	HM	AUC	ρ	Attr	Obj	Seen	Unseen	HM	AUC	ρ	Attr	Obj	Seen	Unseen	HM	AUC	ρ
UT-Zappos	CLIP [54]	*24.4*	*40.5*	*4.8*	*41.9*	*6.7*	*1.5*	-	*24.1*	*58.3*	*11.9*	*45.7*	*15.3*	*4.4*	-	*52.5*	*74.4*	*44.9*	*68.1*	*39.2*	*24.6*	-
	LDE (image)	15.1	44.6	3.2	21.7	4.7	0.5	33.5 %	13.9	52.6	5.6	32.1	6.6	0.9	21.1 %	21.4	50.3	7.0	42.1	8.2	1.6	6.5 %
	GDE (image)	**28.2**	**56.1**	**24.0**	**43.7**	**23.7**	**8.6**	**578.5 %**	**36.3**	**64.1**	**31.4**	**55.9**	**29.3**	**13.9**	**317.9 %**	**48.1**	**72.4**	**42.5**	**68.7**	**41.3**	**24.7**	**100.4 %**
MIT-states	CLIP [54]	*26.6*	*42.3*	*23.5*	*35.2*	*19.4*	*6.2*	-	*33.0*	*52.1*	*30.6*	*45.3*	*26.3*	*11.1*	-	*45.9*	*61.2*	*43.8*	*58.1*	*39.7*	*22.2*	-
	LDE (image)	13.7	25.1	10.6	16.2	8.0	1.1	17.8 %	15.3	30.5	15.0	20.9	11.1	2.0	18.4 %	18.7	34.6	18.6	27.2	14.6	3.6	16.1 %
	GDE (image)	**20.8**	**34.4**	**18.9**	**25.1**	**14.2**	**3.4**	**54.9 %**	**28.1**	**45.3**	**30.7**	**36.1**	**23.4**	**8.6**	**77.7 %**	**32.3**	**50.3**	**36.8**	**40.8**	**27.3**	**11.9**	**53.6 %**

Table 2. Ablation on backbone architecture in compositional classification, closed-world scenario.

by setting $\mathcal{M} = \mathbb{R}^n$ in our method, for both text and image modalities. We indicate the modality by adding "(text)" or "(image)" next to method names. Decomposed text-embeddings are given by Proposition 1, as noise and sparsity belong only to visual data.

Results. Table 1 reports the results in the closed-word and open-world settings. In general, GDEs of visual data perform closely to the zero-shot full-state baseline, demonstrating they encode semantically meaningful information about the labels. Interestingly, on the UT-Zappos dataset, they improve the standard zero-shot approach by a large margin. We attribute this gap to the fact that in UT-Zappos numerous representations are used for the computation of each primitive direction on average ($\sim$1400 per attribute, $\sim$1900 per object). In contrast, the MIT-states dataset contains noisy annotations [2] and on average fewer representations to compute the primitives ($\sim$260 per attribute, $\sim$120 per object). The decomposition shows robustness to sparsity, as indicated by the good open-world unseen accuracy on the MIT-states datasets, for which seen pairs are less than 5% of the total. LDE for visual data performs much worse than GDE on both datasets and when ablating the VLM backbone (see Tab. 2). This indicates that image embeddings are not closely linearly decomposable, and highlights the importance of respecting the data geometry when dealing with the extra complexity given by noise and sparsity. This verifies also on other geometries (see Appendix).

4.3. Group Robustness

Pre-trained VLMs produce biased representations, leading to zero-shot classifiers not robust to group shifts [73]. Our framework offers a training-free method to compute unbiased embeddings. We evaluate it on the group robustness benchmark presented in [58], which requires classifying an image without leveraging spurious correlations. In this setting, a set of target classes $\mathcal{Z}_{obj}$ has spurious correlations with a set of attributes $\mathcal{Z}_{attr}$ due to the highly unbalanced data distribution over the *groups* in $\mathcal{G} = \mathcal{Z}_{attr} \times \mathcal{Z}_{obj}$. The goal is to obtain an object classifier that does not exploit spurious correlations, improving the average accuracy over all the groups (AVG) while keeping the (GAP) on the worst group accuracy (WG) small. We use the object embeddings $\tilde{\mathbf{u}}_o := \mathrm{Exp}_\mu(\mathbf{v}_o)$, $o \in \mathcal{Z}_{obj}$ computed with our method to evaluate the group robustness performance. Intuitively, these embed only object representations that are not correlated with attribute-related spurious features. We thus predict the object class of an image x as $\arg\max_{o \in \mathcal{Z}_{obj}} \tilde{\mathbf{u}}_o^\top \mathbf{u}_x$.

Baselines. In addition to the zero-shot CLIP and LDE method, we include two standard baselines that use labeled data, namely Empirical Risk Minimization (ERM) with linear probing [32] and ERM with feature adapters [16]. Furthermore, we compare with two recent methods improving the performance of VLMs, Deep Feature Reweighting (DFR) [30] and Contrastive Adapters (CA) [73], and with FairerCLIP [10] that performs debiasing of the frozen CLIP representations in the training-free setting like our method.

Results. Table 3 reports the results on the group robustness benchmarks. GDE considerably outperforms CLIP and LDE , with an increase of WG accuracy on the Waterbirds and CelebA datasets of about 42 and 21.8, respectively. This indicates that our method effectively decomposes the embeddings of object and attribute primitives, producing robust classifiers. Notably, it achieves state-of-the-art WG accuracy and smaller Gap compared to all other methods

Method	Waterbirds			CelebA		
	WG	Avg	Gap	WG	Avg	Gap
CLIP [54]	44.4	84.3	40.0	74.4	86.9	12.4
LDE (text) [67]	64.6	88.0	23.3	83.9	85.5	1.6
ERM Linear Probe [32]	65.4	97.7	32.3	30.4	**94.6**	64.2
ERM Adapter [16]	76.1	**97.8**	21.7	40.0	94.3	54.3
DFR (Subsample) [30]	58.8	95.9	37.1	78.7	91.8	13.1
DFR (Upsample) [30]	66.5	96.4	29.8	83.9	91.2	7.2
CA [73]	85.3	94.5	9.3	83.9	90.4	6.4
FairerCLIP [10]	86.0	92.2	6.1	85.2	87.8	2.5
GDE (image)	**86.4**	91.5	**5.0**	**87.5**	87.9	**0.4**

Table 3. Comparison of results on group robustness.

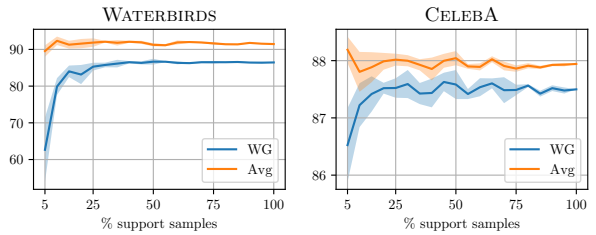

Figure 4. Data efficiency of GDE on the group robustness benchmark. Subsets of the full support set are sampled keeping the group ratios fixed. The shaded confidence band shows the standard deviation over five experiments.

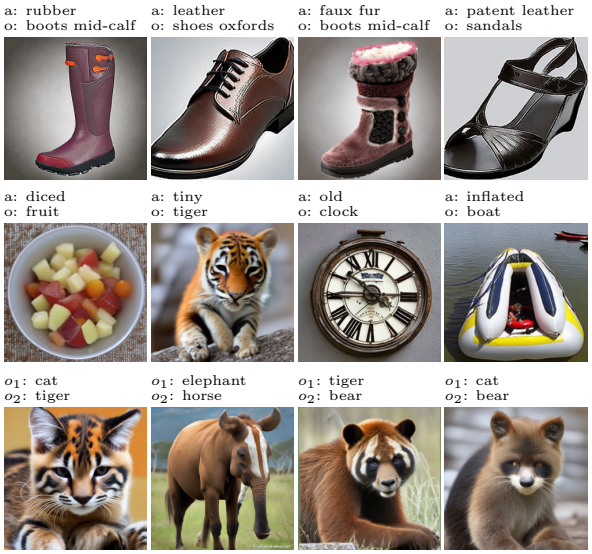

Figure 5. Attribute-object pairs generated using decomposed embeddings with StableDiffusion for the UT-Zappos (*first row*) and MIT-states (*second row*) datasets. The two leftmost labels are seen pairs, while the two right-most are unseen pairs. We also generate object-object pairs (*third row*) blending animal species.

that use labeled data, including the task-specific FairerCLIP. GDE is thus an effective training-free solution to compute unbiased embeddings. Furthermore, GDE demonstrates remarkable data-efficiency performance, achieving high results using limited amount of data (see Fig. 4). For example, when using 25% of the full train samples (randomly selected keeping group ratios fixed) the WG decreases less than 1% on both the Waterbirds and CelebA datasets.

4.4. Visualize Decomposable Approximations

We visualize the decomposed visual embeddings using a diffusion model implementing the unCLIP mechanism (StableDiffusion v2.1) [56, 57], trained to invert the CLIP image encoder by conditioning the generative process with the image embeddings. We invert the decomposable vectors $\tilde{\mathbf{u}}_{(a,o)}$ obtained in previous experiments. In this way, we can qualitatively examine the information they contain.

In Fig. 5 we show some generated images for object-attribute pairs where the attribute is not the most common state of the object (i.e. we avoid common pairs like "green broccoli" or "big elephant"), observing whether the generated image correctly represents the full label and not just the attribute/object. The generated images well represent both the object and the attribute of the label, with no difference in the quality of the outputs from seen (two leftmost columns) and unseen pairs (two rightmost columns). This emphasizes the generalization properties of our decomposable image embeddings, with potential to be applied in practical tasks like augmenting compositional sparse datasets.

The modularity of the decomposable structures allows representing the composition of two objects $o_1, o_2 \in \mathcal{Z}_{obj}$ as $\text{Exp}_\mu(\mathbf{v}_{o_1} + \mathbf{v}_{o_2})$. Inspired by [39], we experiment with blending different animal species (Fig. 5, third row). The generated images portray photorealistic creatures with features of the two input species. This further highlights the power and versatility of the proposed framework. More generated images are in the Appendix.

5. Conclusion

We investigated the emergence of compositional structures within the image latent space of vision-language models and demonstrated that visual embeddings also exhibit a degree of compositionality similar to that of textual representations. We proposed a training-free framework, Geodesically Decomposable Embeddings (GDE), designed to address the noisy and sparse nature of image data. GDE decomposes visual representations as a geometry-aware combination of optimal directions representing primitive concepts. We demonstrated that these composed representations encode complex concepts and are effective in several tasks, including compositional classification and group robustness. Notably, GDE presents more robust abilities to perform compositionality than existing approaches based on linear decomposition of latent spaces, contributing to higher results in group robustness than existing task-specific methods. We believe this work contributes to achieving better interpretability and controllability of modern VLMs.

Acknowledgements. This work was sponsored by the ERJU project, the EU Horizon project ELIAS (No. 101120237), Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007), and the FAIR - Future AI Research (PE00000013), funded by NextGeneration EU. The authors acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support.

References

[1] Jacob Andreas. Measuring compositionality in representation learning. In *ICLR*, 2019. 1

[2] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *NeurIPS*, 33:1462–1473, 2020. 7

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 2

[4] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *CVPR*, 2022. 2

[5] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023. 2

[6] Sarah Chabal and Viorica Marian. Speakers of different languages process the visual world differently. *Journal of Experimental Psychology: General*, 144(3):539–550, 2015. 2

[7] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68. Springer, 2016. 6

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 5

[9] Konrad Czechowski, Tomasz Odrzygóźdź, Marek Zbysiński, Michał Zawalski, Krzysztof Olejnik, Yuhuai Wu, Łukasz Kuciński, and Piotr Miłoś. Subgoal search for complex reasoning tasks. *NeurIPS*, 2021. 1

[10] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing zero-shot predictions of clip in rkhss. In *ICLR*, 2024. 2, 7, 8

[11] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023. 1

[12] Jacob Feldman. Probabilistic origins of compositional mental representations. *Psychological Review*, 131(3):599–624, 2024. 2

[13] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 1, 2

[14] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973. 1, 2

[15] P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TMI*, 23(8):995–1005, 2004. 2, 4

[16] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 7, 8

[17] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, pages 6840–6849, 2023. 1

[18] Yunye Gong, Srikrishna Karanam, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Peter C Doerschuk. Learning compositional visual concepts with mutual consistency. In *CVPR*, 2018. 1

[19] Alon Hafri, E. J. Green, and Chaz Firestone. Compositionality in visual perception. *Behavioral and Brain Sciences*, 46: e277, 2023. 2

[20] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bošnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. In *ICLR*, 2018. 1

[21] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 2

[22] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *ECCV*, 2022. 2

[23] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *NeurIPS*, 2023. 2

[24] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018. 1

[25] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 6

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 1

[27] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 2

[28] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, 2018. 2

[29] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 2

[30] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023. 7, 8

[31] Jayanth Koushik, Hiroaki Hayashi, and Devendra Singh Sachan. Compositional reasoning for visual question answering. In *ICML*, 2017. 1, 2

[32] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 7, 8

[33] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *NeurIPS*, 35: 30233–30249, 2022. 1

[34] Kevin J. Lande. Compositionality in perception: A framework. *WIREs Cognitive Science*, 15(6):e1691, 2024. 2

[35] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500, 2024. 2

[36] Xilai Li, Xi Song, and Tianfu Wu. Aognets: Compositional grammatical architectures for deep learning. In *CVPR*, 2019. 1

[37] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 2

[38] Renjie Liao, Alex Schwing, Richard Zemel, and Raquel Urtasun. Learning deep parsimonious representations. *NeurIPS*, 2016. 1

[39] Giorgio Longari, Lorenzo Olearo, Simone Melzi, Rafael Peñaloza, and Alessandro Raganato. How to blend concepts in diffusion models. *arXiv preprint arXiv:2407.14280*, 2024. 8

[40] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *CVPR*, 2023. 2

[41] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, pages 5222–5230, 2021. 2, 6

[42] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE TPAMI*, 46(3):1545–1560, 2022.

[43] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017.

[44] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 1, 2

[45] Nihal V. Nayak, Peilin Yu, and Stephen Bach. Learning to compose soft prompts for compositional zero-shot learning. In *ICLR*, 2023. 1, 2

[46] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. Parts of speech–grounded subspaces in vision-language models. *NeurIPS*, 36:2700–2724, 2023. 2, 3

[47] Bjorn Ommer and Joachim Buhmann. Learning the compositional nature of visual object categories for recognition. *IEEE TPAMI*, 32(3):501–516, 2009. 1, 2

[48] Bjorn Ommer and Joachim M Buhmann. Learning the compositional nature of visual objects. In *CVPR*, 2007. 2

[49] Dim P Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba. How to make a pizza: Learning a compositional layer-based gan model. In *CVPR*, 2019. 1, 2

[50] Barbara Partee et al. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360, 1995. 1

[51] Barbara H Partee. *Compositionality in formal semantics: Selected papers*. John Wiley & Sons, 2008. 1

[52] Pramuditha Perera, Matthew Trager, Luca Zancato, Alessandro Achille, and Stefano Soatto. Prompt algebra for task composition. *arXiv preprint arXiv:2306.00310*, 2023. 1, 2

[53] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019. 6

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 7, 8

[55] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *CVPR*, pages 27263–27272, 2024. 1

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5, 8

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 5, 8

[58] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020. 6, 7

[59] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023. 2

[60] Sascha Saralajew, Lars Holdijk, Maike Rees, Ebubekir Asan, and Thomas Villmann. Classification-by-components: Probabilistic modeling of reasoning over a set of components. *NeurIPS*, 2019. 1, 2

[61] Jürgen Schmidhuber. Towards compositional learning in dynamic networks. *Technical University of Munich (Technical Report FKI-129-90)*, 1990. 1

[62] Zhangzhang Si and Song-Chun Zhu. Learning and-or templates for object recognition and detection. *IEEE TPAMI*, 35 (9):2189–2205, 2013. 1

[63] Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D Scott Phoenix, and Dileep George. Teaching compositionality to cnns. In *CVPR*, 2017. 1, 2

[64] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *CVPR*, 2019. 2

[65] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 2

[66] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 2

[67] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *ICCV*, pages 15395–15404, 2023. 1, 2, 3, 6, 7, 8

[68] Jianyu Wang and Alan L Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *CVPR*, 2015. 1

[69] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, pages 192–199, 2014. 6

[70] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023. 2

[71] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn composable primitive concepts? *TMLR*, 2023. 1

[72] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 5

[73] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *NeurIPS*, 35:21682–21697, 2022. 7, 8

[74] Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. *Advances in Neural Information Processing Systems*, 34:18513–18526, 2021. 1

[75] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, 2023. 2

[76] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal. Modular generative adversarial networks. In *ECCV*, 2018. 2