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

## Supplementary Material

In this Supplementary Material, we provide additional details on Riemannian manifolds in Appendix A, we prove the theoretical results our framework builds upon in Appendix B, we describe extra information of our implementation in Appendix C and we present further experimental results in Appendix D.

## A. Details on Riemannian Manifold

We discuss some details of the tools we used in our framework to deal with the geometry of a data manifold $\mathcal{M}$. In the following, we focus on the spherical case $\mathcal{M} = \mathbb{S}^{d-1}$, which applies to the case with normalized embeddings.

### A.1. Closed form solutions

The exponential and logarithmic maps can be expressed in closed form on the unit sphere $\mathbb{S}^{d-1}$. For any point of tangency $\mu \in \mathbb{S}^{d-1}$, we have

$$\mathrm{Exp}_\mu(\mathbf{v}) = \cos(||\mathbf{v}||)\mu + \sin(||\mathbf{v}||)\frac{\mathbf{v}}{||\mathbf{v}||}, \quad \mathbf{v} \in T_\mu\mathbb{S}^{d-1} \tag{13}$$

and

$$\mathrm{Log}_\mu(\mathbf{u}) = \theta\frac{(I_d - \mu\mu^\top)(\mathbf{u}-\mu)}{||(I_d - \mu\mu^\top)(\mathbf{u}-\mu)||}, \quad \mathbf{u} \in \mathbb{S}^{d-1} \tag{14}$$

where $\theta = \arccos(\mathbf{u}^\top\mu)$ and $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix.[1]

### A.2. Intrinsic Mean

**Existence, uniqueness, and characterization.** The (weighted) intrinsic mean $\mu$ of a set of points $\{\mathbf{u}_i\}_{i=1}^N$, which is defined as the solution of a minimization problem, is not necessarily unique. For example, on $\mathbb{S}^2$ all the points on the equator minimize the average distance from the north and south poles. But, existence and uniqueness are guaranteed if the points live inside the same geodesic ball $\mathcal{B}_o(r) := \{\mathbf{u} \in \mathcal{M} \,|\, d_\mathcal{M}(o,\mathbf{u}) < r\}$ of radius $r > 0$ small enough [1]. Under the same condition, we also have that $\mu$ is the unique point on $\mathcal{M}$ centering the logarithmic map of the input points, *i.e.* such that $\sum_{i=1}^N w_i\ \mathrm{Log}_\mu(\mathbf{u}_i) = 0$. We will refer to this property as the *characterization of the intrinsic mean*. For the unit sphere $\mathbb{S}^{d-1}$, the closeness assumption is satisfied for any $r < \pi/2$. Note that we can expect this condition to be verified by the normalized embeddings of a neural encoder because of the cone effect [5].

[1] To be precise, the logarithmic map is defined on $\mathcal{M} \setminus \mathcal{C}_\mu$, where $\mathcal{C}_\mu$ is called the *cut locus* of $\mu$. We do not stress this detail because it is well known that $\mathcal{C}_\mu$ has measure zero on $\mathcal{M}$. On the unit sphere $\mathbb{S}^{d-1}$, the cut locus of any point $\mu$ is its antipode $-\mu$.

**Algorithm 1** Intrinsic mean

**Input:** $\mathbf{u}_1, \dots, \mathbf{u}_N \in \mathcal{M}, w_1, \dots, w_N \in \Delta_N, \mu_0 \in \mathcal{M}$
**Output:** the intrinsic mean $\mu \in \mathcal{M}$

**repeat**
  $\delta_\mu = \eta \sum_{i=1}^N w_i\ \mathrm{Log}_{\mu_j}(\mathbf{u}_i)$
  $\mu_{j+1} = \mathrm{Exp}_{\mu_j}(\delta_\mu)$
**until** $||\delta_\mu|| < \epsilon$

**Computation by gradient descent.** Computing the intrinsic mean $\mu$ of a weighted set of points requires minimizing the objective function

$$f(\mathbf{u}) = \frac{1}{2}\sum_{i=1}^N w_i\, d_\mathcal{M}(\mathbf{u},\mathbf{u}_i)^2, \quad \mathbf{u} \in \mathcal{M} \tag{15}$$

This can be done with a gradient descent algorithm [6]. Indeed, it can be shown that (15) has gradient

$$\nabla f(\mathbf{u}) = -\sum_{i=1}^N w_i\ \mathrm{Log}_\mathbf{u}(\mathbf{u}_i), \quad \mathbf{u} \in \mathcal{M} \tag{16}$$

Algorithm 1 shows the pseudocode for the gradient descent procedure. At each iteration, the new approximation $\mu_{j+1}$ is obtained by first moving in the opposite direction of the gradient and then mapping on the manifold with the exponential map centered in $\mu_j$. The cycle stops when the norm of the update is smaller than a fixed small value $\epsilon > 0$. Usually, the starting value $\mu_0 \in \mathcal{M}$ is chosen among the input points, which live on the manifold. Otherwise, in the special case $\mathcal{M} = \mathbb{S}^{d-1}$, a good choice is the normalized (weighted) arithmetic mean $\mu_0 = \sum_{i=1}^N w_i\,\mathbf{u}_i / \|\sum_{i=1}^N w_i\,\mathbf{u}_i\|$. The learning rate $\eta$ has to be carefully chosen to guarantee convergence. It has been shown that setting $\eta = 1$ is sufficient for spheres [2].

## B. Proofs

We provide the proof of the theoretical results stated in the methodology chapter. We omit the proof of Proposition 1 because it is the same of the more general Proposition 2 in the special case when $|\mathcal{E}| = 1$. In the following, we assume that a given composite concept $z \in \mathcal{Z}$ is the tuple $z = (z_1, \dots, z_s)$.

**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 $i = 1, \dots, s$ and*

$$\mathbf{u}_z = \mathrm{Exp}_\mu(\mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s}) \quad \forall z = (z_1, \dots, z_s) \tag{17}$$

*Proof.* Let $\phi(\mathcal{Z}) = \{\mathbf{u}_z\}$ be a geodesically decomposable set with tangent projections $\mathbf{v}_z = \mathrm{Log}_\mu(\mathbf{u}_z)$ decomposed as $\mathbf{v}_z = \mathbf{v}'_{z_1} + \cdots + \mathbf{v}'_{z_s}$. Indicating $\bar{\mathbf{v}}_{\mathcal{Z}_i} = \frac{1}{|\mathcal{Z}_i|}\sum_{z_i\in\mathcal{Z}_i} \mathbf{v}'_{z_i}$, we now show that the searched directions are $\mathbf{v}_{z_i} = \mathbf{v}'_{z_i} - \bar{\mathbf{v}}_{\mathcal{Z}_i}$ $(i = 1, \dots, s)$. The centering constrain $\sum_{z_i\in\mathcal{Z}_i} \mathbf{v}_{z_i} = 0$ immediately follows from the definition. Then, we observe that $\sum_i \bar{\mathbf{v}}_{\mathcal{Z}_i} = \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} \mathbf{v}_z = 0$ for the characterization of the intrinsic mean. This implies Eq. (17) is satisfied:

$$\begin{aligned}\mathbf{v}_z &= \mathbf{v}'_{z_1} + \cdots + \mathbf{v}'_{z_s} \\ &= (\bar{\mathbf{v}}_{\mathcal{Z}_1} + \cdots + \bar{\mathbf{v}}_{\mathcal{Z}_s}) + \mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s} \\ &= \mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s}\end{aligned} \tag{18}$$

To show uniqueness, we demonstrate the $\mathbf{v}_{z_i}$ are uniquely determined by the original vectors $\mathbf{v}_z$:

$$\begin{aligned}\mathbf{v}_{z_i} &= \mathbf{v}'_{z_i} - \bar{\mathbf{v}}_{\mathcal{Z}_i} \\ &= \mathbf{v}'_{z_i} + \sum_{j\neq i} \bar{\mathbf{v}}_{\mathcal{Z}_j} \\ &= \frac{1}{|\mathcal{Z}(z_i)|}\sum_{z\in\mathcal{Z}(z_i)} (\mathbf{v}'_{z_1} + \cdots + \mathbf{v}'_{z_s}) \\ &= \frac{1}{|\mathcal{Z}(z_i)|}\sum_{z\in\mathcal{Z}(z_i)} \mathbf{v}_z\end{aligned} \tag{19}$$

□

**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 $\mu$ w.r.t. the weights $w_{(z,e)} = p_{(z,e)} / \sum_{(z,e)} p_{(z,e)}$. The minimization problem:*

$$\begin{gathered}\arg\min_{\{\tilde{\mathbf{u}}_z\}} \sum_{(z,e)\in\mathcal{Z}\times\mathcal{E}} p_{(z,e)} ||\,\mathrm{Log}_\mu(\mathbf{u}_{(z,e)}) - \mathrm{Log}_\mu(\tilde{\mathbf{u}}_z)||^2, \\ s.t.\ \{\tilde{\mathbf{u}}_z\} \text{ is geodesically decomposable}\end{gathered} \tag{20}$$

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

$$\mathbf{v}_{z_i} = \frac{1}{|\mathcal{Z}(z_i)|}\sum_{z\in\mathcal{Z}(z_i)} \mathbf{v}_z, \quad \mathbf{v}_z = \sum_{e\in\mathcal{E}} p_{(z,e)}\,\mathrm{Log}_\mu(\mathbf{u}_{(z,e)}) \tag{21}$$

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

*Proof.* We start by observing that, if $\{\tilde{\mathbf{u}}_z\}$ is a geodesically decomposable set with intrinsic mean $\mu'$, then, following the proof of Lemma 1, we can write its tangent projection $\tilde{\mathbf{v}}_z = \mathrm{Log}_\mu(\tilde{\mathbf{u}}_z) \in T_\mu\mathcal{M}$ as $\tilde{\mathbf{v}}_z = \mathbf{v}_0 + \mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s}$ where $\sum_{z_i\in\mathcal{Z}_i} \mathbf{v}_{z_i} = 0$ and $\mathbf{v}_0 = \sum_i \bar{\mathbf{v}}_{\mathcal{Z}_i} = \frac{1}{|\mathcal{Z}|}\sum_z \tilde{\mathbf{v}}_z$. Note that $\mu' = \mu$ if and only if $\mathbf{v}_0 = 0$. Now, in the setting of the statement, we indicate $\mathbf{v}_{(z,e)} = \mathrm{Log}_\mu(\mathbf{u}_{(z,e)})$ and rephrase the objective in Eq. (20) as finding the vectors $\mathbf{v}_0$, $\mathbf{v}_{z_i} \in T_\mu\mathcal{M}$, $z_i \in \mathcal{Z}_i$ $(i = 1, \dots, s)$ minimizing

$$\frac{1}{2}\sum_{\substack{(z,e)\\ \in\mathcal{Z}\times\mathcal{E}}} p_{(z,e)} ||\mathbf{v}_{(z,e)} - (\mathbf{v}_0 + \mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s})||^2 \tag{22}$$

Observing that $\sum_{z\in\mathcal{Z}} \mathbf{v}_{z_i} = \sum_i \frac{|\mathcal{Z}|}{|\mathcal{Z}_i|}\sum_{z_i\in\mathcal{Z}_i} \mathbf{v}_{z_i} = 0$, the derivative of (22) with respect to $\mathbf{v}_0$ is

$$\begin{aligned}&\sum_{z\in\mathcal{Z}}\sum_{e\in\mathcal{E}} p_{(z,e)}(\mathbf{v}_{(z,e)} - (\mathbf{v}_0 + \mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s})) \\ &= \sum_{z\in\mathcal{Z}} (\mathbf{v}_z - \mathbf{v}_0),\end{aligned} \tag{23}$$

where $\mathbf{v}_z = \sum_{e\in\mathcal{E}} p_{(z,e)}\mathbf{v}_{(z,e)}$. Setting this equal to zero gives $\mathbf{v}_0 = \frac{1}{|\mathcal{Z}|}\sum_z \mathbf{v}_z = \sum_{(z,e)} w_{(z,e)}\mathbf{v}_{(z,e)} = 0$. Here the last equality follows from the characterization of the intrinsic mean and it implies the intrinsic mean of the solution is $\mu$. The derivative with respect to a fixed $\mathbf{v}_{z_i}$ is:

$$\begin{aligned}&\sum_{z\in\mathcal{Z}(z_i)}\sum_{e\in\mathcal{E}} p_{(z,e)}(\mathbf{v}_{(z,e)} - (\mathbf{v}_{z_1} + \cdots + \mathbf{v}_{z_s})) \\ &= \sum_{z\in\mathcal{Z}(z_i)} (\mathbf{v}_z - \mathbf{v}_{z_i})\end{aligned} \tag{24}$$

Setting this equal to zero gives $\mathbf{v}_{z_i} = \frac{1}{|\mathcal{Z}(z_i)|}\sum_{z\in\mathcal{Z}(z_i)} \mathbf{v}_z$. □

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

*Proof.* As observed in the proof of Proposition 2, we have $\frac{1}{|\mathcal{Z}|}\sum_z \mathbf{v}_z = 0$, implying the weighted mean $\mu$ is the intrinsic mean of $\{\mathbf{u}_z\}_{z\in\mathcal{Z}}$. □

# C. Experimental Details

## C.1. Closeness assumption

We numerically verify the closeness assumption, discussed in Appendix A.2, which guarantees the existence and uniqueness of the intrinsic mean. Given a set of points on $\mathbb{S}^{d-1}$, a good guess for the center $o \in \mathcal{M}$ of a small geodesic ball $\mathcal{B}_o(r)$ containing them is their normalized arithmetic mean $\mu_0$. So, for all the sets of embeddings used in our experiments, we verify their maximum intrinsic distance (*i.e.* angle) from $\mu_0$ is smaller than $\pi/2$. In Tab. 4 we show some statistics of the distances computed with the embeddings from the default model CLIP ViT-L\14 used in the experiments.

| | Image Embeddings | | | Text Embeddings | | |
|---|---|---|---|---|---|---|
| | avg | max | $r < \pi/2$ | avg | max | $r < \pi/2$ |
| UT-ZAPPOS | 0.49 | 1.0 | ✓ | 0.56 | 0.75 | ✓ |
| MIT-STATES | 0.78 | 1.4 | ✓ | 0.67 | 1.14 | ✓ |
| WATERBIRDS | 0.63 | 1.03 | ✓ | 0.41 | 0.48 | ✓ |
| CELEBA | 0.75 | 1.15 | ✓ | 0.4 | 0.43 | ✓ |

Table 4. Statistics of distances from embeddings to their normalized arithmetic mean. The closeness assumption is verified if all the embeddings are within a radius $r < \pi/2 \approx 1.57$.

### C.2. Noise distribution

**Temperature selection.** When performing compositional classification and group robustness, we use the image-to-text distribution $\mathbb{P}((z,e)|y(z))$ defined by the VLM as the noise distribution. For CLIP, this is given by the softmax activations described in the main paper and it depends on the temperature parameter $t \in (0, +\infty)$. For each dataset, we select the value for $t$ by performing a grid search on the validation set. We optimize the AUC metric for compositional classification and the WG accuracy for group robustness.

**SigLIP sigmoid probabilities.** Differently from the original CLIP, SigLIP uses a sigmoid-based loss processing every image-text pair independently and it defines the pair-specific probabilities

$$\mathbb{P}((z,e)|y(z)) = \frac{1}{1+\exp(-\mathbf{u}_{(z,e)}^{\top}\mathbf{u}_{y(z)}/t - b)} \tag{25}$$

When considering SigLIP embeddings, we use the noise distribution $p_{(z,e)} \propto \mathbb{P}((z,e)|y(z))$ proportional to the pair-specific sigmoid probabilities. We select the temperature parameter $t$ as described for the CLIP model while keeping the *logit bias* $b \in \mathbb{R}$ equal to the learned value ($b \approx -16.5$).

### C.3. Text prompts

For the UT-Zappos and MIT-states datasets, we consider the same text prompts used in [8]. Attribute-object pair $(a, o)$ is described by $y(a, o) =$ "an image of a $\{a\}$ $\{o\}$" where $\{a\}$, $\{o\}$ are the lower-case original category names. For UT-Zappos, every dot character is substituted with a space ("Synthetic Boots.Ankle" $\rightarrow$ "synthetic boots ankle"). We use these prompts both when decomposing text embeddings and when computing the image-to-text probabilities defining the noise distribution.

For the Waterbirds and CelebA datasets, we consider the text prompts defined in [3, 8]. These are obtained by representing each spurious attribute and each target class with the captions in Tabs. 5 and 6. Then, prepending the spurious prompts to the class prompts produces $k = 4$ and $k = 3$ textual descriptions for each composite group in the Waterbirds and CelebA datasets, respectively. We compute the image-to-text probabilities for the noise distribution using the decomposable text embeddings $\tilde{\mathbf{u}}_{y(z)}$, $z \in \mathcal{Z}$, given by Proposition 2 applied to the input embeddings. Note indeed that they can be written as $\{\mathbf{u}_{y(z,e)} \,|\, (z,e) \in \mathcal{Z} \times \mathcal{E}\}$, where $\mathcal{E}$ is a "prompt template" dimension.

| **Class prompt** | |
|---|---|
| This is a picture of a landbird. | |
| This is a picture of a waterbird. | |
| **Spurious attribute prompt** | |
| This is a land background. | This is a water background. |
| This is a picture of a forest. | This is a picture of a beach. |
| This is a picture of a moutain. | This is a picture of an ocean. |
| This is a picture of a wood. | This is a picture of a port. |

Table 5. The text prompts from [3] for the Waterbirds dataset.

| **Class prompt** | |
|---|---|
| A photo of a celebrity with dark hair. | |
| A photo of a celebrity with blond hair. | |
| **Spurious attribute prompt** | |
| A photo of a male. | A photo of a female. |
| A photo of a male celebrity. | A photo of a female celebrity. |
| A photo of a man. | A photo of a woman. |

Table 6. The text prompts from [3] for the CelebA dataset.

| DATASET | METHOD | ATTR | OBJ | SEEN | UNSEEN | HM | AUC | $\rho$ |
|---|---|---|---|---|---|---|---|---|
| UT-ZAPPOS | CLIP [7] | *24.1* | *58.3* | *11.9* | *45.7* | *15.3* | *4.4* | - |
| | $\text{GDE}_u$ (IMAGE) | 36.2 | 63.8 | 30.9 | 55.6 | 29.0 | 13.6 | 310.8 % |
| | GDE (IMAGE) | **36.3** | **64.1** | **31.4** | **55.9** | **29.3** | **13.9** | **317.9 %** |
| MIT-STATES | CLIP [7] | *33.0* | *52.1* | *30.6* | *45.3* | *26.3* | *11.1* | - |
| | $\text{GDE}_u$ (IMAGE) | 27.7 | 44.3 | 30.4 | 35.0 | 22.9 | 8.2 | 74.3 % |
| | GDE (IMAGE) | **28.1** | **45.3** | **30.7** | **36.1** | **23.4** | **8.6** | **77.7 %** |

Table 7. Results of ablating the use of CLIP scores as the noise distribution in compositional classification, closed-world scenario.

## D. Additional Results

### D.1. Ablation: noise distribution

Our decomposition method (GDE) computes the noise distribution using CLIP scores with a custom temperature parameter. In Tab. 7, we compare GDE against the decomposition obtained when using a uniform noise distribution ($\text{GDE}_u$) in the task of compositional classification. While the simpler $\text{GDE}_u$ performs well compared to the zero-shot baseline, leveraging the non-uniform noise distribution from the CLIP scores always improves performance.

### D.2. Decomposing hyperbolic representations

We investigate the compositional properties of visual representations on different geometries than the CLIP's hyper sphere. Specifically, we perform compositional classification of the pre-trained MERU ViT-L-16 [4] embeddings,

| Dataset | Method | Attr | Obj | Seen | Unseen | HM | AUC | $\rho$ |
|---|---|---|---|---|---|---|---|---|
| UT-Zappos | MERU [68] | *17.4* | *26.7* | *11.1* | *16.0* | *9.6* | *1.4* | - |
| | LDE (image) | 13.8 | 40.7 | 4.6 | 21.8 | 5.1 | 0.7 | 52.7 % |
| | GDE (image) | **22.9** | **49.7** | **15.2** | **40.3** | **16.0** | **4.7** | **340.4 %** |
| MIT-states | MERU [68] | *17.7* | *34.4* | *15.8* | *27.0* | *13.2* | *3.1* | - |
| | LDE (image) | 13.7 | 26.7 | 11.2 | 19.1 | 9.1 | 1.4 | 46.1 % |
| | GDE (image) | **18.9** | **34.2** | **18.5** | **25.3** | **13.7** | **3.3** | **107.0 %** |

Table 8. Compositional classification results of MERU's hyperbolic representations, closed-world scenario.

which are points on the *Lorentz model*:

$$\mathcal{L}^d = \{\mathbf{u} \in \mathbb{R}^{d+1} | \langle \mathbf{u}, \mathbf{u} \rangle_{\mathcal{L}} = -1/c\}. \tag{26}$$

Here $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ is the Lorentzian inner product and the parameter $c > 0$ is learned during pre-training. The exponential and logarithmic maps have a closed form solution for the hyperboloid $\mathcal{L}^d$ [4], enabling a simple application of the GDE framework also in this setting. Results in Tab. 8 show that, as observed for CLIP spherical embeddings, the GDEs of MERU's hyperbolic representations contain semantically meaningful information of the concepts they represent. Moreover, the significantly lower performance of LDE highlights the importance of GDE's geometry awareness also in this non-spherical setup.

### D.3. Runtime

A potential limitation of our proposed framework is the additional computational costs it requires for mapping embeddings to and from the tangent space. We now analyze the inference time of the decomposition procedure.

Suppose we compute a decomposable set for $M = |\mathcal{Z}|$ composite concepts using $N = |\mathcal{T}|$ visual embeddings on the sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$. Compared to LDE, GDE additionally computes $\text{Log}_\mu$ for the $N$ inputs and $\text{Exp}_\mu$ for the $M$ tangent compositions. The computational complexity of these operations is $\mathcal{O}(Nd)$ and $\mathcal{O}(Md)$, respectively. Note that the orthogonal projection in Eq. (14) can be rewritten as $(I_d - \mu\mu^\top)\mathbf{w} = \mathbf{w} - (\mu^\top \mathbf{w})\mu$, avoiding the explicit computation of the $d \times d$ matrix. Calculating $\mu$ with Algorithm 1 is $\mathcal{O}(Nd)$ per gradient step, keeping the extra compute linear in $N, M, d$. Tab. 9 reports the runtimes for GDE, LDE, $\mu, \text{Log}_\mu, \text{Exp}_\mu$ in our experiments (tolerance for $\mu$ is $\epsilon = 10^{-5}$). Both methods are fast on the relatively small datasets used for our analysis. GDE is significantly slower than LDE, with most of its extra runtime being spent on the computation of $\mu$. However, we argue that approximating $\mu$ with a smaller subset of $N' < N$ input embeddings could be sufficient and drastically improve efficiency when the number of inputs is large.

### D.4. Generated images

In Fig. 7 we show extra images generated using StabeDiffusion with the unCLIP module to invert composite embeddings. We include attribute-object pairs from all the datasets used in our experiments. Similarly to the animal-animal pairs shown in the main document, we identify other high-level categories within the MIT-states objects (items, environments and materials) and visualize animal-environment and item-material compositions.

Also, we observe that the modularity of the compositions allows to gain finer control on the composite embeddings. In Fig. 8, we invert embeddings of the form $\text{Exp}_\mu(\alpha \mathbf{v}_a + \mathbf{v}_o)$, where the attribute direction is scaled by a scalar $\alpha \in \mathbb{R}$. In the generated images, changing the value of $\alpha$ modifies the intensity of the attribute that results in a lower or strong appearance of it in the generated images. This experiment further demonstrates that the primitive vectors resulting from solving the proposed optimization problem are interpretable directions of the latent space.

Our initial goal of the generative experiments was to qualitative inspect the GDE compositions. However, the good quality of the results suggests that our framework could be useful for augmenting compositional datasets. To support this, we compute the average CLIP-score of 500 outputs (five generations of 100 random unseen concepts), to assess how a CLIP model perceives composite concepts in generated images. As a baseline, we consider the default text-to-image (T2I) version of the generative model.

| Dataset | $N$ | $M$ | LDE | GDE | $\mu$ | steps $\mu$ | $\text{Log}_\mu$ | $\text{Exp}_\mu$ |
|---|---|---|---|---|---|---|---|---|
| UT-Zappos | 22998 | 192 | 41±1 ms | 382±9 ms | 267±15 ms | 3 | 83±4 ms | 0±0 ms |
| MIT-states | 30338 | 28175 | 149±0 ms | 850±8 ms | 560±15 ms | 5 | 106±4 ms | 60±1 ms |
| Waterbirds | 4795 | 4 | 7±0 ms | 59±14 ms | 43±12 ms | 4 | 11±2 ms | 0±0 ms |
| CelebA | 162770 | 4 | 375±4 ms | 3812±35 ms | 2902±37 ms | 5 | 557±14 ms | 0±0 ms |

Table 9. Runtimes on a Titan Xp GPU, averaged over 5 runs.

| UT-Zappos | | MIT-states | |
|---|---|---|---|
| GDE: 0.68±0.06 | T2I: 0.62±0.10 | GDE: 0.58±0.08 | T2I: 0.55±0.10 |

Table 10. Average CLIP-score of SD generated images.

### D.5. Failure cases

We investigate failure cases in Stable Diffusion visualizations and noted that the decomposable embeddings may encode spurious correlations of the input data or produce ambiguous compositions. For instance, the generated images in Fig. 6 suggest that the 'inflated' and 'boat' primitive directions are respectively linked to 'round object' and 'water', and the 'tiger'+'horse' and 'dog'+'forest' compositions are respectively close to 'zebra' and 'bear'.


Figure 6. Failure instances in SD generations.

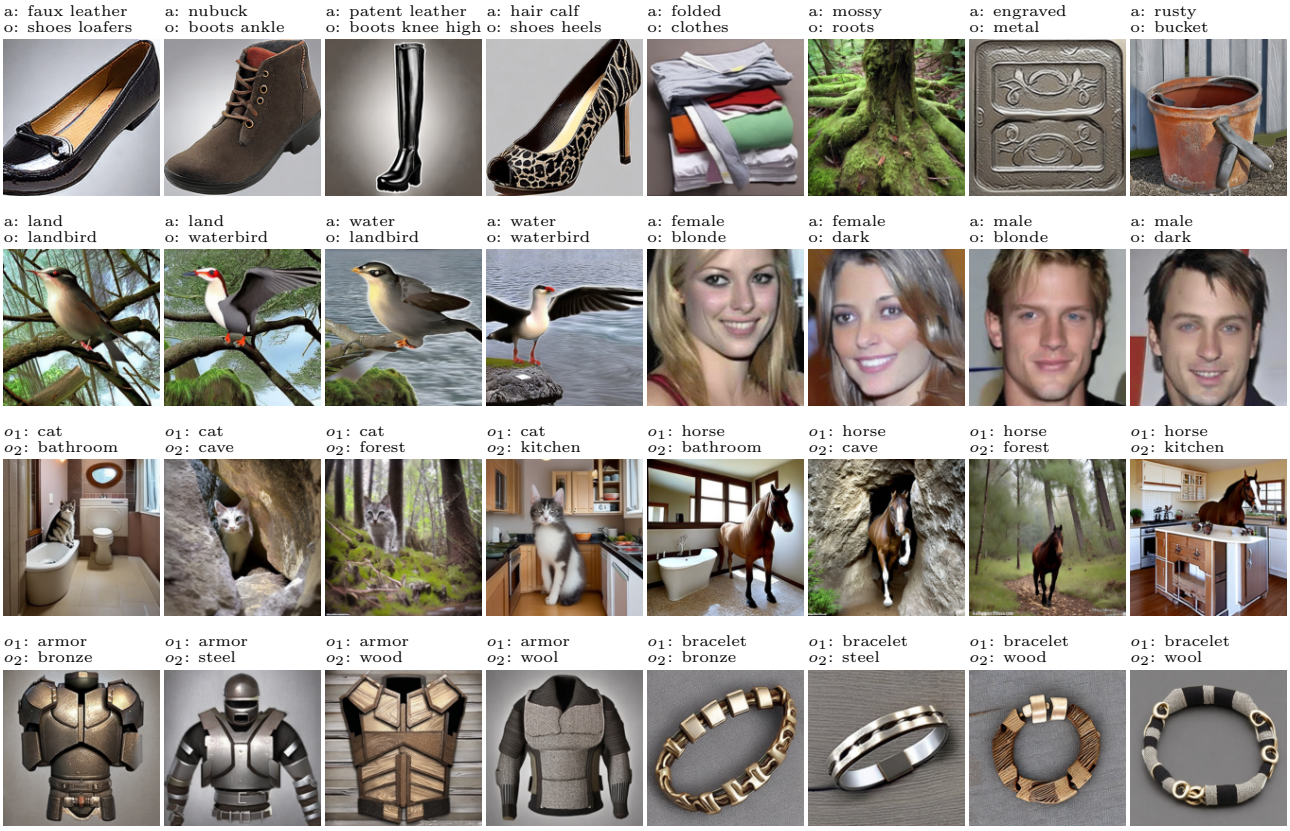


Figure 7. Additional generated images obtained by inverting the decomposable embeddings computed with GDE, using StableDiffusion with the unCLIP technique. We include attribute-object pairs from the UT-Zappos and MIT-states datasets (*first row*), and from the Waterbirds and CelebA datasets (*second row*). Similarly to the animal-animal pairs shown in the main document, we visualize animal-environment pairs (*third row*) and item-material pairs (*fourth row*) from the MIT-states objects.


Figure 8. Scaling attribute direction in attribute-object compositions.

## References


[1] Bijan Afsari. Riemannian $L^p$ center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011. 1

[2] Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3): 2230–2260, 2013. 1

[3] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 3

[4] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023. 3, 4

[5] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 35:17612–17625, 2022. 1

[6] Xavier Pennec. Probabilities and statistics on riemannian manifolds: Basic tools for geometric measurements. In *NSIP*, pages 194–198, 1999. 1

[7] 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. 3

[8] 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. 3