Supplementary Material

This supplementary material is organized as follows: in Section A we provide proofs for all the statements of the paper and we discuss some connections with mathematical representation theory; in Section B we give details on the datasets and prompts used for our experiments; in Section C we present some additional experimental results and qualitative examples.

A. Proofs

Lemma 2. *1) A collection of vectors $r(\mathcal{Z})$ is linearly factored if and only if the vector difference $\boldsymbol{u}_z - \boldsymbol{u}_{z'}$ does not depend on the components that $z, z' \in \mathcal{Z}$ share in common. 2) If $|\mathcal{Z}_i| = n_i$, then the dimension of Span$(r(\mathcal{Z}))$ is at most $1 + \sum_{i=1}^{k}(n_i - 1)$.*

Proof. (1) If the vectors are linearly factored, then clearly the vector differences $\boldsymbol{u}_z - \boldsymbol{u}_{z'}$ do not depend on the components that z, z' share in common since the corresponding vectors cancel out. For the converse, fix $z = (z_1, \ldots, z_k) \in \mathcal{Z}$ arbitrarily and choose any k vectors $\boldsymbol{u}_{z_1}, \ldots, \boldsymbol{u}_{z_k}$ such that $\boldsymbol{u}_z = \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}$. Now for any $z'_i \in \mathcal{Z}_i$ and any $i = 1, \ldots, k$, define

$$\begin{aligned} &\boldsymbol{u}_{z'_i} := \boldsymbol{u}_{z_i} + \boldsymbol{u}_{z'} - \boldsymbol{u}_z, \\ &\qquad \text{where } z' = (z_1, \ldots, z'_i, \ldots, z_k). \end{aligned}$$

If $z'' = (z'_1, \ldots, z'_k)$, it now holds that

$$\begin{aligned} \boldsymbol{u}_{z''} &= \boldsymbol{u}_{z''} - \boldsymbol{u}_{(z_1, z'_2, \ldots, z'_k)} \\ &\quad + (\boldsymbol{u}_{(z_1, z'_2, \ldots, z'_k)} - \boldsymbol{u}_{(z_1, z_2, \ldots, z'_k)}) \\ &\quad + \ldots + (\boldsymbol{u}_{(z_1, z_2, \ldots, z'_k)} - \boldsymbol{u}_z) + \boldsymbol{u}_z \\ &= (\boldsymbol{u}_{z'_1} - \boldsymbol{u}_{z_1}) + \ldots + (\boldsymbol{u}_{z'_k} - \boldsymbol{u}_{z_k}) + \boldsymbol{u}_z \\ &= \boldsymbol{u}_{z'_1} + \ldots + \boldsymbol{u}_{z'_k}. \end{aligned}$$

(2) We have that

$$\begin{aligned} \sum_{z \in \mathcal{Z}} \gamma_z \boldsymbol{u}_z &= \sum_{z \in \mathcal{Z}} \gamma_z (\boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}) \\ &= \sum_{z \in \mathcal{Z}} \gamma_z (\bar{\boldsymbol{u}}_{\mathcal{Z}_1} + \ldots + \bar{\boldsymbol{u}}_{\mathcal{Z}_k} + \tilde{\boldsymbol{u}}_{z_1} + \ldots \tilde{\boldsymbol{u}}_{z_k}), \\ &= \sum_{z \in \mathcal{Z}} \gamma_z (\boldsymbol{u}_0 + \tilde{\boldsymbol{u}}_{z_1} + \ldots \tilde{\boldsymbol{u}}_{z_k}), \end{aligned} \tag{14}$$

where $\bar{\boldsymbol{u}}_{\mathcal{Z}_1} := \frac{1}{n_i} \sum_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{z_i}$ and $\tilde{\boldsymbol{u}}_{z_i} := \boldsymbol{u}_{z_i} - \boldsymbol{u}_{\mathcal{Z}_i}$. Since $\sum_{z_i \in \mathcal{Z}_i} \tilde{\boldsymbol{u}}_{z_i} = 0$, equation 14 shows that any linear combination of the vectors $\boldsymbol{u}_z, z \in \mathcal{Z}$ can be written as a linear combination of $1 + \sum_{i=1}^{k}(n_i - 1)$ vectors. □

Lemma 3 (Centered decomposition)**.** *If a collection of vectors $r(\mathcal{Z})$ is linearly factored, then there exist unique vectors $\boldsymbol{u}_0 \in V$ and $\boldsymbol{u}_{z_i} \in V$ for all $z_i \in \mathcal{Z}_i$ $(i = 1, \ldots, k)$ such that $\sum_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{z_i} = 0$ for all i and*

$$\boldsymbol{u}_z = \boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}, \tag{2}$$

for all $z = (z_1, \ldots, z_k)$.

Proof. Following the proof of part 2 of the previous Lemma, it is enough to let $\boldsymbol{u}_0 := \bar{\boldsymbol{u}}_{\mathcal{Z}_1} + \ldots + \bar{\boldsymbol{u}}_{\mathcal{Z}_k}$ where $\bar{\boldsymbol{u}}_{\mathcal{Z}_1} := \frac{1}{n_i} \sum_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{z_i}$, and then re-center the remaining vectors accordingly. For the uniqueness, we note that equation 2 implies that the vectors $\boldsymbol{u}_0, \boldsymbol{u}_{z_i}, z_i \in \mathcal{Z}_i$ satisfy

$$\boldsymbol{u}_0 = \frac{1}{N} \sum_{z \in \mathcal{Z}} \boldsymbol{u}_z, \quad \boldsymbol{u}_{z_i} = \frac{n_i}{N} \sum_{\substack{z' = (z'_1, \ldots, z'_k) \\ z'_i = z_i}} \boldsymbol{u}_{z'} - \boldsymbol{u}_0. \tag{15}$$

where $N = n_1 \ldots n_k$. In particular, equation 15 shows that $\boldsymbol{u}_0, \boldsymbol{u}_{z_i}, z_i \in \mathcal{Z}_i$ are uniquely determined by the original vectors $\boldsymbol{u}_z$. □

In the previous proof, we considered a map associating each $\boldsymbol{u}_z, z \in \mathcal{Z}$ with the vectors given by

$$\boldsymbol{u}_0 = \frac{1}{N} \sum_{z \in \mathcal{Z}} \boldsymbol{u}_z, \quad \boldsymbol{u}_{z_i} = \frac{n_i}{N} \sum_{\substack{z' = (z'_1, \ldots, z'_k) \\ z'_i = z_i}} \boldsymbol{u}_{z'} - \boldsymbol{u}_0. \tag{16}$$

It is easy to see that if we define $\tilde{\boldsymbol{u}}_z = \boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}$ then applying equation 16 with the new vectors $\tilde{\boldsymbol{u}}_z$ instead of $\boldsymbol{u}_z$ yields the same components $\boldsymbol{u}_{z_i}$. Thus, this map can be seen as a projection onto a linearly factored set of vectors. Note that the component vectors satisfy $\sum_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{z_i} = 0$. The following result considers a slightly more general setting in which these components vectors satisfy $\sum \alpha_{z_i} v_{z_i} = 0$ for some weights α_i that sum to 1.

Proposition 4. *Let α_{z_i} $z_i \in \mathcal{Z}_i$ be arbitrary positive weights such that $\sum_{z_i \in \mathcal{Z}_i} \alpha_{z_i} = 1$, and define $\beta_z := \prod_i \alpha_{z_i}$ for all $z = (z_1, \ldots, z_k)$. Then, for any norm $\|\cdot\|$ induced by an inner product on V, we have that*

$$\begin{aligned} &\arg\min_{\tilde{\boldsymbol{u}}_z} \sum_{z \in \mathcal{Z}} \beta_z \|\boldsymbol{u}_z - \tilde{\boldsymbol{u}}_z\|^2, \\ &\qquad s.t. \ \{\tilde{\boldsymbol{u}}_z\} \text{ is linearly factored,} \end{aligned} \tag{3}$$

is given by $\tilde{\boldsymbol{u}}_z = \boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}$ where

$$\boldsymbol{u}_0 := \sum_{z} \beta_z \boldsymbol{u}_z, \quad \boldsymbol{u}_{z_i} := \frac{1}{\alpha_{z_i}} \sum_{\substack{z' = (z'_1, \ldots, z'_k) \\ z'_i = z_i}} \beta_z \boldsymbol{u}_{z'} - \boldsymbol{u}_0. \tag{4}$$

Proof. Without loss of generality, we may assume that $\sum_z \beta_z \boldsymbol{u}_{z_i} = \sum \alpha_{z_i} \boldsymbol{u}_{z_i} = 0$. Imposing that the derivative of equation 3 with respect to $\boldsymbol{u}_0$ is zero leads to

$$\begin{aligned}&\sum_{z \in \mathcal{Z}} \beta_z(\boldsymbol{u}_z - (\boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k})) \\ &= \sum_{z \in \mathcal{Z}} \beta_z(\boldsymbol{u}_z - \boldsymbol{u}_0) = 0,\end{aligned} \tag{17}$$

which implies $\boldsymbol{u}_0 = \sum_z \beta_z \boldsymbol{u}_z$. Similarly, differentiating with respect to $\boldsymbol{u}_{z_i}$ we have

$$\begin{aligned}&\sum_{\substack{z'=(z'_1,\ldots,z'_k)\\ z'_i=z_i}} \beta_{z'}(\boldsymbol{u}_{z'} - (\boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k})) \\ &= \sum_{\substack{z'=(z'_1,\ldots,z'_k)\\ z'_i=z_i}} \beta_{z'}(\boldsymbol{u}_{z'} - \boldsymbol{u}_0 - \boldsymbol{u}_{z_i}) = 0\end{aligned} \tag{18}$$

which implies that

$$\sum_{\substack{z'=(z'_1,\ldots,z'_k)\\ z'_i=z_i}} \beta_z \boldsymbol{u}_z = \alpha_{z_i}(\boldsymbol{u}_0 + \boldsymbol{u}_{z_i}), \tag{19}$$

so $\boldsymbol{u}_{z_i}$ is as in equation 4. □

Proposition 6. *Let $r(\mathcal{Z})$ be a set of linearly factored vectors of maximal dimension. Then r is compositional for some disentangled action of $G = \mathfrak{S}_{n_1} \times \ldots \times \mathfrak{S}_{n_k}$ on V. Conversely, if r is compositional for a disentangled action of G, then the vectors $r(\mathcal{Z})$ are linearly factored.*

Proof. Let $r(\mathcal{Z})$ be a set of linearly factored vectors of maximal dimension. If $W := Span(\boldsymbol{u}_z, z \in \mathcal{Z})$, then we write $V = W \oplus W'$, and define a linear action of G on $\mathbb{R}^d$ by associating each group element $g = (g_1, \ldots, g_k)$ with an invertible linear transformation so that each g_i determines a permutation of the vectors $\boldsymbol{u}_{z_i}$, while fixing other terms and W'. This describes a disentangled action of G, where $V = W' \oplus \langle \boldsymbol{u}_0 \rangle \oplus V_{\mathcal{Z}_1} \oplus \ldots \oplus V_{\mathcal{Z}_k}$ (to be consistent with the original definition, we can set $V_1 = W' \oplus \langle \boldsymbol{u}_0 \rangle \oplus V_{\mathcal{Z}_1}$ and $V_i = V_{\mathcal{Z}_i}$ for $i \geq 2$).

For the converse, let $\rho : G \to GL(V)$ be any linear action of G on V (a group representation). Writing $G_{\hat{i}} = \mathfrak{S}_1 \times \ldots \times \{e\} \times \ldots \times \mathfrak{S}_k$ (with the identity at the i-th component), we define

$$\begin{aligned}V_0 &:= \{\boldsymbol{u} \in V \colon g \cdot \boldsymbol{u} = \boldsymbol{u}, \ \forall g \in G\}, \\ \tilde{V}_i &:= \{\boldsymbol{u} \in V \colon g \cdot \boldsymbol{u} = \boldsymbol{u}, \ \forall g \in G_{\hat{i}}\}.\end{aligned} \tag{20}$$

Since G acts linearly, these are vector spaces. We also define the linear maps

$$\begin{aligned}\pi_0 &: \boldsymbol{u} \mapsto \frac{1}{|G|} \sum_{g \in G} g \cdot \boldsymbol{u}, \\ \tilde{\pi}_i &: \boldsymbol{u} \mapsto \frac{1}{|G_{\hat{i}}|} \sum_{g \in G_{\hat{i}}} g \cdot \boldsymbol{u}.\end{aligned} \tag{21}$$

These are linear projections onto V_0 and $\tilde{V}_i$, respectively, since they map onto these spaces and they fix them. We now define $\pi_i := \tilde{\pi}_i - \pi_0$ and $V_i := Im(\pi_i)$. Since $\tilde{V}_i \cap \tilde{V}_j = V_0$ for $i \neq j$, we have that $V_i \cap V_j = \{0\}$ for $i \neq j$. In general, we now have that $V_0 \oplus V_1 \oplus \ldots \oplus V_k \subset V$; if the action ρ is disentangled, however, then

$$V = V_0 \oplus V_1 \oplus \ldots \oplus V_k. \tag{22}$$

Thus, for any $v \in V$, we have $v = \pi_0(v) + \pi_1(v) + \ldots + \pi_k(v)$. Now assume that $r : \mathcal{Z} \to V$ is a compositional embedding, so $g \cdot r(z) = r(g \cdot z)$. We observe that $\boldsymbol{u}_{z_i} = \pi_i(\boldsymbol{u}_z)$ is fixed by $\mathfrak{S}_j$ for $j \neq i$, and thus depends only on z_i. In fact, the expressions for π_0, π_i applied to $\boldsymbol{u}_z$ are exactly the projection maps from equation 16. Thus, we can write $\boldsymbol{u}_z = \boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}$, which means that $r(\mathcal{Z})$ are linearly factored. □

Proposition 7. *In the setting described above, and assuming that $Span(\boldsymbol{v}_y, y \in \mathcal{Y}) = \mathbb{R}^d$, the embedding $z \mapsto \boldsymbol{u}_{x(z)}$ of $\mathcal{Z}$ is linearly factored in the sense of Definition 1 if and only if there exists functions $q_0, \ldots, q_k$ such that*

$$p(x(z), y) = q_0(y) q_1(z_1, y) \ldots q_k(z_k, y), \tag{7}$$

for all $z = (z_1, \ldots, z_k) \in \mathcal{Z}$ and $y \in \mathcal{Y}$.

Proof. Assume that equation 7 holds, and let $g_0(y) := \log(q_0(y))$ and $g_i(z_i, y) := \log(q_i(z, y))$. For all $z \in \mathcal{Z}$, we can write

$$\begin{aligned}\log p(x(z), y) &= g_0(y) + g_1(z_1, y) + \ldots + g_k(z_k, y) \\ &= \bar{g}_0(y) + \bar{g}_1(z_1, y) + \ldots + \bar{g}_k(z_k, y), \\ &s.t. \sum_{z_i \in \mathcal{Z}_i} \bar{g}_i(z_i, y) = 0, \quad i = 1, \ldots, k,\end{aligned} \tag{23}$$

where $\bar{g}_0(y) := g_0(y) + \sum_{j=1}^{k} \frac{1}{n_j} \sum_{z_j \in \mathcal{Z}_j} g_j(z_j, y)$ and $\bar{g}_i(z_i, y) := g(z_i, y) - \frac{1}{n_i} \sum_{z'_i \in \mathcal{Z}_i} g_i(z'_i, y)$. It is easy to verify the following identities for $i = 1, \ldots, k$:

$$\begin{aligned}\bar{g}_0(y) &= \frac{1}{N} \sum_{z \in \mathcal{Z}} \log p(x(z), y) = \frac{1}{N} \sum_{z \in \mathcal{Z}} \boldsymbol{u}_{x(z)}^\top \boldsymbol{v}_y + c_0 \\ &= \boldsymbol{u}_0^\top \boldsymbol{v}_y + c_0 \\ \bar{g}_i(z_i, y) &= \frac{n_i}{N} \sum_{\substack{z'=(z'_1,\ldots,z'_k)\\ z'_i=z_i}} \log p(x(z), y) - \bar{g}_0(y) \\ &= \frac{n_i}{N} \sum_{\substack{z'=(z'_1,\ldots,z'_k)\\ z'_i=z_i}} \boldsymbol{u}_{x(z')}^\top \boldsymbol{v}_y - \boldsymbol{u}_0^\top \boldsymbol{v}_y = \boldsymbol{u}_{z_i}^\top \boldsymbol{v}_y,\end{aligned} \tag{24}$$

where we used the expression for $\log p(x, y)$ from equation 6 and the definition of the terms $\boldsymbol{u}_0, \boldsymbol{u}_{z_i}$ from equation 16. If we now define $\tilde{\boldsymbol{u}}_{x(z)} := \boldsymbol{u}_0 + \boldsymbol{u}_{z_1} + \ldots + \boldsymbol{u}_{z_k}$, then it follows from equation 24 that $\tilde{\boldsymbol{u}}_{x(z)}^\top \boldsymbol{v}_y = \boldsymbol{u}_{x(z)}^\top \boldsymbol{v}_y (= \log p(x(z), y)) - c_0)$ for all $z \in \mathcal{Z}$, $y \in \mathcal{Y}$. Since by hypothesis $Span(\boldsymbol{v}_y, y \in \mathcal{Y}) = \mathbb{R}^d$, we conclude that $\tilde{\boldsymbol{u}}_{x(z)} = \boldsymbol{u}_{x(z)}$. Conversely, it is clear that if all $\boldsymbol{u}_{x(z)}$ decompose as in equation 2, then $p(x(z), y)$ has a factored form as in equation 7 for all $y \in \mathcal{Y}$. □

Corollary 8. *Under the assumptions of Proposition 7, an embedding $z \mapsto \boldsymbol{u}_{x(z)}$ of $\mathcal{Z}$ is linearly factored if only if the factors z_i are conditionally independent given any image y.*

Proof. This follows immediately from the factored form of equation 7. More precisely, the statement means that

$$\tilde{p}(z \,|\, y) = \tilde{p}(z_1 \,|\, y) \dots \tilde{p}(z_k \,|\, y), \tag{25}$$

where $\tilde{p}(z \,|\, y) := \frac{1}{Z_y} p(x(z) \,|\, y)$, $\tilde{p}(z_i \,|\, y) := \frac{1}{Z_y} \sum_{z_{k \neq i}} p(x(z) \,|\, y)$ and $Z_y := \sum_z p(x(z) \,|\, y)$. We observe that equation 25 implies equation 7, since we can write

$$p(x(z), y) = Z_y p(y) \tilde{p}(z_1|y) \dots \tilde{p}(z_k|y), \tag{26}$$

which has the desired factored form. Conversely, equation 7 means that

$$\tilde{p}(z \,|\, y) = \frac{q_0(y) Z_1 \dots Z_k}{p(y) Z_y} \tilde{q}_1(z_1, y) \dots \tilde{q}_k(z_k, y), \tag{27}$$

where $Z_i = \sum_{z_i \in \mathcal{Z}_i)} q_i(z_i, y)$ and $\tilde{q}_i(z_i, y) = \frac{1}{Z_i} q(z_i, y)$. Since $\sum_{z \in \mathcal{Z}} \tilde{p}(z \,|\, y) = 1$, we deduce that the y-dependent constant on the right of equation 27 is equal to 1, and $\tilde{q}_i(z, y) = \tilde{p}(z_i|y)$. □

Proposition 9 (Relaxed feasibility of linear factorizations)**.** *1) If $y \in \mathcal{Y}$ is such that $p(x(z), y)$ is mode-disentangled, then one can replace the embedding vectors $\boldsymbol{u}_{x(z)}$ with their linearly factored approximations $\tilde{\boldsymbol{u}}_{x(z)}$ from Proposition 4 (for any choice of weights) and obtain the same prediction for z given y; 2) If $p(x(z), y)$ is order-disentangled for all images y sampled from a distribution with full support over the unit sphere, then the vectors $\boldsymbol{u}_{x(z)}$ are necessarily linearly factored.*

Proof. (1) Assume that $p(x(z), y)$ is mode-disentangled. Then we have that

$$\begin{aligned}
&\arg\max_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{(z_i, z_{-i})}^\top \boldsymbol{v}_y \\
&= \arg\max_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{(z_i, z'_{-i})}^\top \boldsymbol{v}_y \\
&= \arg\max_{z_i \in \mathcal{Z}_i} \sum_{\substack{z' = (z'_1, \dots, z'_k) \\ z'_i = z_i}} \boldsymbol{u}_{z'}^\top \boldsymbol{v}_y \\
&= \arg\max_{z_i \in \mathcal{Z}_i} \boldsymbol{u}_{z_i}^\top \boldsymbol{v}_y
\end{aligned} \tag{28}$$

where $\boldsymbol{u}_{z_i}$ is as in equation 16, or as in the weighted version from equation 4. This implies that we can perform inference using the linearly factored approximations $\tilde{\boldsymbol{u}}_{x(z)}$ instead of the original vectors.
2) We will use the notation $z = (z_i, z_j, z_{-\{i,j\}})$ where $z_{-\{i,j\}} := (z_1, \dots, z_{i-1}, z_{i+1}, \dots z_{j-1}, z_{j+1}, \dots, z_k)$. If $p(x(z), y)$ is order-disentangled for y, then for any $z_i, z'_i \in \mathcal{Z}_i$ and $z_j, z'_j \in \mathcal{Z}_j$

$$\begin{aligned}
&(\boldsymbol{u}_{(z'_i, z_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z_i, z_j, z_{-\{i,j\}})})^\top \boldsymbol{u}_y \geq 0 \\
&\Leftrightarrow (\boldsymbol{u}_{(z'_i, z'_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z_i, z'_j, z_{-\{i,j\}})})^\top \boldsymbol{u}_y \geq 0,
\end{aligned} \tag{29}$$

and similarly

$$\begin{aligned}
&(\boldsymbol{u}_{(z_i, z'_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z_i, z_j, z_{-\{i,j\}})})^\top \boldsymbol{u}_y \geq 0 \\
&\Leftrightarrow (\boldsymbol{u}_{(z'_i, z'_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z'_i, z_j, z_{-\{i,j\}})})^\top \boldsymbol{u}_y \geq 0.
\end{aligned} \tag{30}$$

If these relations hold for any vector $\boldsymbol{u}_y$, then it means that

$$\begin{aligned}
&\boldsymbol{u}_{(z'_i, z'_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z_i, z'_j, z_{-\{i,j\}})} \\
&\qquad = \lambda(\boldsymbol{u}_{(z'_i, z_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z_i, z_j, z_{-\{i,j\}})}) \\
&\boldsymbol{u}_{(z'_i, z'_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z'_i, z_j, z_{-\{i,j\}})} \\
&\qquad = \mu(\boldsymbol{u}_{(z_i, z'_j, z_{-\{i,j\}})} - \boldsymbol{u}_{(z_i, z_j, z_{-\{i,j\}})})
\end{aligned} \tag{31}$$

for some positive scalars $\lambda, \mu \in \mathbb{R}$. It follows from Lemma 11 below that either all four points in equation 31 are aligned, or $\lambda = \mu = 1$. However, we can exclude that all four points are aligned for otherwise the largest between $p(x(z_i, z_j, z_{-\{i,j\}}), y)$ and $p(x(z'_i, z_j, z_{-\{i,j\}}), y)$ would determine the largest among $p(x(z_i, z_j, z_{-\{i,j\}}), y)$ and $p(x(z_i, z'_j, z_{-\{i,j\}}), y)$, *i.e.*, the factors $\mathcal{Z}_i, \mathcal{Z}_j$ would not be distinct. (Technically, we can assume in our definition of "factors" that all possible rankings of values of $\mathcal{Z}_i$ are possible for any choice of z_{-i}). Thus, $\lambda = \mu = 1$ in equation 31 for all z_i, z'_i, z_j, z'_j. This implies that $\boldsymbol{u}_{(z_i, z_{-i})} - \boldsymbol{u}_{(z'_i, z_{-i})}$ does not depend on z_{-i}, which in turn means that the vectors $\boldsymbol{u}_z$ are linearly factored, since $\boldsymbol{u}_z - \boldsymbol{u}_{z'}$ does not depend on components that z, z' have in common. □

Lemma 11. *If $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^d$ are such that*

$$\boldsymbol{p} - \boldsymbol{q} = \lambda(\boldsymbol{r} - \boldsymbol{s}), \quad \boldsymbol{p} - \boldsymbol{r} = \mu(\boldsymbol{q} - \boldsymbol{s}), \tag{32}$$

*for some scalars $\lambda, \mu \in \mathbb{R}$, then either $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}, \boldsymbol{s}$ lie on the same affine line (*i.e.*, all pairwise differences are scalar multiples of each other) or $\lambda = \mu = 1$.*

Proof. Substituting $\boldsymbol{p} = \boldsymbol{q} + \lambda(\boldsymbol{r} - \boldsymbol{s})$ in the second equality in equation 32 yields

$$(1 - \mu)\boldsymbol{q} + (\lambda - 1)\boldsymbol{r} + (\mu - \lambda)\boldsymbol{s} = 0. \tag{33}$$

If $\mu \neq 1$ or $\nu \neq 1$, then this shows that $\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{s}$ are aligned (note that coefficients sum to 1). Using the relation for $\boldsymbol{p}$, we conclude that either $\mu = \nu = 1$ or all four points are aligned. □

We conclude this section by elaborating on the connection with mathematical representation theory. This discussion is not necessary for understanding the paper, but we believe that the symmetry-based viewpoint introduced in [26] is a useful framework for studying disentanglement and compositionality in machine learning. For convenience to the reader, we include here a minimal set of definitions and basic results from representation theory, focusing on the representation of finite groups. More details can be found, for example, in [21].

A *representation* of a group G is a homomorphism $\rho : G \to GL(V)$, where V is a finite-dimensional vector space (typically over the complex numbers, but we can focus on the the real setting here). Often the map ρ is omitted and the representation is identified with V. It also common to say that V is a "G-module" or a "G-representation." Given two G-representations V, W, a *homomorphism of representations* is a linear map $\varphi : V \to W$ that is G-equivariant:

$$\varphi(g \cdot v) = g \cdot \varphi(v), \quad \forall g \in G,\, v \in V. \tag{34}$$

A *subrepresentation* (or *submodule*) of a G-representation V is a vector subspace $H \subset V$ such that is G-invariant:

$$g(h) \in H, \qquad \forall g \in G,\, h \in H. \tag{35}$$

If $\varphi : V \to W$ is a homorphism of representations, then the kernel and image of φ are subrepresentations of V and W, respectively. A G-representation of V is *irreducible* if it has no proper subrepresentations, *i.e.*, if its only subrepresentations are $\{0\}$ and itself.

Example 12 (Trivial representation)**.** Let G be any group and let $V = \mathbb{R}$ be a one-dimensional vector space. Then the map $\rho : G \to GL(V)$ that every element of G with the identity on V is an irreducible representation, called the *trivial representation*.

Example 13 (Permutation representation)**.** Let $V = \mathbb{R}^n$ and consider the representation $\rho : \mathfrak{S}_n \to GL(V)$ that permutes coordinates. This is not an irreducible representation since the one-dimensional subspace $V_0 = \langle (1, \dots, 1) \rangle$ is a subrepresentation (a "copy" of the trivial representation). In fact, we have that $V = V_0 \oplus V_1$ where $V_1 = \{v \colon v_1 + \ldots + v_n = 0\}$. One can show that V_1 is irreducible, and it is called the *standard representation* of $\mathfrak{S}_n$.

The next statements imply that, for finite groups, irreducible representations can always be used as "building blocks" for describing arbitrary representations. The irreducible components of a representation are (nearly) uniquely determined; moreover, there are only finitely many irreducible representations of a group up to isomorphism.

Proposition 14 (Corollary 1.6, [21])**.** *If G is a finite group, any G-representation can be decomposed as a direct sum of irreducible representations.*

Proposition 15 (Proposition 1.8, [21])**.** *Let V be a G-representation, and consider its decomposition into irreducible representations:*

$$V = V_1^{\oplus a_1} \oplus \ldots \oplus V_k^{\oplus a_k}. \tag{36}$$

Then the spaces $V_i^{\oplus a_i}$ are uniquely determined. The irreducible representations V_i are determined up to isomoprhism.

Proposition 16 (Corollary 2.18, [21])**.** *Every finite group only has a finite set of irreducible representations, up to isomorphism.*

For example, the irreducible representations of a symmetric group $\mathfrak{S}_n$ are in one-to-one correspondence with the (unordered) partitions of n elements. See [21, Chapter 4] for an explicit description.

We now return to our factored set $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_k$. We consider the vector space $\langle \mathcal{Z} \rangle = Span(\boldsymbol{e}_z \colon z \in \mathcal{Z})$, spanned by independent basis vectors associated with elements of $\mathcal{Z}$. We can identify $\langle \mathcal{Z} \rangle$ with the space $\mathbb{R}^{n_1} \otimes \ldots \otimes \mathbb{R}^{n_k}$. As $\mathfrak{S}_i$-modules, $\mathbb{R}^{n_i} \cong V_{0,n_i} \oplus V_{1,n_i}$ where V_{0,n_i} is a trivial representation and V_{1,n_i} is the standard representation for $\mathfrak{S}_i$. We thus have that

$$\begin{aligned} \langle \mathcal{Z} \rangle &\cong \bigotimes_{i=1}^{k} (V_{0,n_i} \oplus V_{1,n_i}) \\ &\cong \bigoplus_{\epsilon_i \in \{0,1\}} V_{\epsilon_1,n_1} \otimes \ldots \otimes V_{\epsilon_k,n_k} \\ &\cong \bigoplus_{\epsilon \in \{0,1\}^k} V_\epsilon, \end{aligned} \tag{37}$$

with $V_\epsilon := V_{\epsilon_1,n_1} \otimes \ldots \otimes V_{\epsilon_k,n_k}$. This is a decomposition of $\langle \mathcal{Z} \rangle$ into irreducible G-representations (see [21, Exercise 2.36]). We can describe the projection π_ϵ onto V_ϵ explicitly

$$\pi_\epsilon = \pi_{\epsilon_1,n_1} \otimes \ldots \otimes \pi_{\epsilon_k,n_k}, \tag{38}$$

where $\pi_{\epsilon,n_i} : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ are given by

$$\begin{aligned} \pi_{0,n_i}(\boldsymbol{u}) &:= \frac{1}{|\mathfrak{S}_i|} \sum_{g \in \mathfrak{S}_i} g \cdot \boldsymbol{u}, \\ \pi_{1,n_i}(\boldsymbol{u}) &:= \boldsymbol{u} - \pi_{0,n_i}(\boldsymbol{u}). \end{aligned} \tag{39}$$

A data embedding $r : \mathcal{Z} \to \mathbb{R}^d$ can be uniquely associated with a linear map $\langle r \rangle : \langle \mathcal{Z} \rangle \to \mathbb{R}^d$ or can equivalently be viewed as a tensor in $[r] \in \mathbb{R}^{n_1} \otimes \ldots \otimes \mathbb{R}^{n_k} \otimes \mathbb{R}^d$. The image of $\langle r \rangle$ is a G-module in $\mathbb{R}^d$ and its decomposition will contain a subset of the irreducible components in equation 37. The notion of disentangled representation given in [26] means that the only irreducible components that contribute to the image of r are the representations V_ϵ such that $\epsilon_i = 1$ for at most one index i. Equivalently, we require that the projection of the image of r onto the “entangled components” is zero, *i.e.*, $\pi_\epsilon(\boldsymbol{u}_z) = 0$ whenever $|\{i\colon \epsilon_i = 1\}| > 1$. An intuitive way to understand this notion is in terms of the tensor $[r] \in \mathbb{R}^{n_1}\otimes\ldots\otimes\mathbb{R}^{n_k}\otimes\mathbb{R}^d$: we require that each of the d “slices” $\mathbb{R}^{n_1}\otimes\ldots\otimes\mathbb{R}^{n_k}$ can be obtained by summing “one-dimensional slices” of the form $\mathbf{1}\otimes\ldots\otimes\boldsymbol{u}_i\otimes\ldots\otimes\mathbf{1}$ (similar to summing vectors into a tensor by “array broadcasting”). In fact, this observation leads to the following characterization of linear factorization in terms of tensor-rank.

Proposition 17. *A tensor* $[r] \in \mathbb{R}^{n_1}\otimes\ldots\otimes\mathbb{R}^{n_k}\otimes\mathbb{R}^d$ *corresponds to a linearly factored representation if and only if all* $(\mathbb{R}^{n_1}\otimes\ldots\otimes\mathbb{R}^{n_k})$*-slices of* $\exp([r])$ *have tensor-rank one, where* $\exp([r])$ *is obtained from* $[r]$ *by exponentiating element-wise. This is true if and only if for all* $\varphi \in (\mathbb{R}^d)^*$ $\exp(\varphi([r]))$ *has tensor-rank one.*

Proof sketch. The first claim follows from the previous discussion and the fact that $\exp\left(\sum_i \mathbf{1}\otimes\ldots\otimes\boldsymbol{u}_i\otimes\ldots\otimes\mathbf{1}\right) = \exp(\boldsymbol{u}_1)\otimes\ldots\otimes\exp(\boldsymbol{u}_k)$. For the second statement, we note $\exp(t)$ having rank-one is a linear condition on a tensor t. □

For categorical distributions of multiple variables, the distribution tensor having rank equal to one corresponds to statistical independence of variables, so the result above can be seen as an algebraic reformulation of Proposition 7 in the main body of the paper. We also note that that other probabilistic conditions could be considered by allowing for more irreducible components in from equation 37 to appear in the image of r. This is similar to the log-linear representations of multivariate data. In fact, it is possible to express any conditional independence assumption on $p(\mathcal{Z}|\mathcal{Y})$ in terms of linear-algebraic conditions on the data representation r.

B. Experimental Details

Datasets. The MIT-states dataset [28] contains images of 245 objects modified by 115 adjectives, for a total of 28175 classes. The test set has size 12995. The UTZappos dataset [50] contains images of 12 shoe types with 16 fine-grained states. The test set has size 2914. Note that in both of these datasets only a small portion of all possible attribute-object pairs actually occurs in the test set. However, in our experiments we assume that we do not have access to this information. We also mention that prior works that have used these datasets such as [35, 38] have differentiatied between the performance on label pairs that were seen in training and those that were not. Since this distinction is not relevant in a zero-shot setting, we simply report accuracy on objects, attributes, and attribute-object pairs. In the Waterbird dataset [44] labels are “waterbird/landbird” and spurious attributes are “water background/land background.” There are 5794 test samples divided in four unbalanced groups. On the CelebA [33], labels are “not blond/blond” and spurious attributes are “male/female.” There are a total 19962 test samples with unbalanced groups. The DeepFashion2 dataset [23] with the captions provided in PerVL [14] contains 1700 images from 100 unique fashion items. Following [14] val/test splitting, we retrieve 50 of these concepts selected for testing. We use 5 randomly chosen images per fashion item as per-concept supporting images, and use a test set with 221 images containing all 50 concepts and their captions (see [14] for more details). Final results are obtained by averaging the Mean Reciprocal Rank metric over 5 random seeds.

<table>
<tr><th colspan="2">Class Prompts</th></tr>
<tr><td colspan="2">This is a picture of a landbird.
This is a picture of a waterbird.</td></tr>
<tr><th colspan="2">Spurious Prompts</th></tr>
<tr><td>This is a land background.</td><td>This is a picture of a forest.</td></tr>
<tr><td>This is a picture of a moutain.</td><td>This is a picture of a wood.</td></tr>
<tr><td>This is a water background.</td><td>This is a picture of an ocean.</td></tr>
<tr><td>This is a picture of a beach.</td><td>This is a picture of a port.</td></tr>
</table>

Table 4: Prompts for Waterbird dataset [44] from [11].

Prompts. For MIT-States and UTZappos, we use the prompt “image of a $[a][o]$,” “image of a $[a]$ object,” and “image of a $[o]$,” as explained in the main body of the paper. Here $[a]$ and $[o]$ are the lower-case original class labels.[3] For our experiments on debiasing on the Waterbirds and CelebA datasets we use the same prompts and spurious attributes used in [11]. These are shown in Tables 4 and 5. To compute debiased prompts we simply prepend all spurious prompts to each class prompts and then average the spurious prompts to obtain debiased class prompts (note that spurious prompts are “balanced” in their bias); this simpler but conceptually similar to the “Orth-Proj” approach used in in [11] that computes an orthogonal projection in the orthogonal complement of the linear space spanned by the spurious prompts. We do not make use of the “positive pairs” of prompts that are used in that work for regularization of the projection map.

C. Additional Results and Discussions

Quantifying linear compositionality. Given a set of vectors $\boldsymbol{u}_z$, $z \in \mathcal{Z}$ in $\mathbb{R}^d$, we can measure how close the vectors

[3] In the case of objects for UTZappos, we perform a simple split ‘Boots.Mid-Calf’ → “boots mid-calf”

Class Prompts
A photo of a celebrity with dark hair. A photo of a celebrity with blond hair.
Spurious Prompts
A photo of a male. A photo of a male celebrity. A photo of a man. A photo of a female. A photo of a female celebrity. A photo of a woman.

Table 5: Prompts for CelebA dataset [33] from [11].

	IW	**RW**	**Avg**
MIT-States [28]	0.23 ± 0.05	0.43 ± 0.06	0.78 ± 0.13
UT Zappos [50]	0.16 ± 0.04	0.51 ± 0.05	0.58 ± 0.18

Table 6: Quantifying compositionality using a trained encoder.

	IW	**RW**	**Avg**
MIT-States [28]	0.04 ± 0.02	0.16 ± 0.02	0.10 ± 0.03
UT Zappos [50]	0.10 ± 0.02	0.22 ± 0.04	0.14 ± 0.05

Table 7: Quantifying compositionality using a randomly initialized encoder.

are to being linearly factored by using

$$\begin{aligned} D(\boldsymbol{u}_z, z \in \mathcal{Z}) &:= \min_{\tilde{\boldsymbol{u}}_z} \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} \|\boldsymbol{u}_z - \tilde{\boldsymbol{u}}_z\|^2, \\ &s.t.\ \{\tilde{\boldsymbol{u}}_z\} \text{ is linearly factored.} \end{aligned} \tag{40}$$

The optimal vectors $\tilde{\boldsymbol{u}}_z$ here are the ideal word approximations given by Proposition 4. In Table 6, we report this quantity for embeddings of objects-attributes in the datasets MIT-States [28] and UT Zappos [50] (IW column). For comparison, we also include the average squared distance between the original embeddings and the average of the individual object and attribute embddings based on “real words” (RW column), and the average squared distance between pairs of the original embedding vectors (Avg). In Table 7, we report the same quantities but using embeddings obtained from a *randomly initialized* encoder. These results suggest that embeddings at initialization are already compositional. We discuss this point further in the next paragraph.

Visualized embeddings. We present more examples of projected embeddings of composite strings. In Figure 4, we consider again the four manually constructed examples from Figure 2 in the main body of the paper: “a photo of a {red, blue, pink} × {car, house}”; “a photo of a {big, small} × {cat, dog} × {eating, drinking}”; “{a photo of a, a picture of a} × {place, object, person}”; “king, queen, man, woman, boy, girl.” The top row of Figure 4 is the same as the top row from Figure 2. In the bottom row of 4, we visualize the embeddings of the same strings using a randomly initialized text encoder. In the first three examples, the factored structure is also *syntactic*, *i.e*., it is based on the string structure. In these cases, the embeddings remain roughly linearly factored even with random encoder. In the last case, however, linearly factored structures are not visible anymore, since the strings in this example contain no repeated substrings. Note also that in third case, the factor corresponding to {a photo of a, a picture of a} is no longer “squashed” since these two strings not considered similar by the randomly initialized encoder.

We show other examples of this effect in Figure 5. Here each pair of plots shows projections of the same strings using a trained encoder (left figure) and a randomly initialized encoder (right figure). As one might expect, for strings corresponding to capital-country relation (first row), the approximate symmetries that can be seen in the embbedings from the trained encoder are no longer present when using the random encoder. The strings in the second row, however, have a synctatic factored structure. In this case, we visually observe strong symmetries in the embeddings from the trained encoder as well as from the random encoder.

In Figure 6, we consider 2D projections of embeddings of factored strings that include idioms such as “cold shoulder,” “big apple”, “black friday,” “hot pepper.” We compare these embeddings with those of similar factored strings in which meanings of words are more conventional and uniform. In both cases, we quantify the amount of linear compositionality both visually and using the squared residual as in equation 40. The results confirm the natural intuition that linear compositionality is measurably weaker when strong contextual effects between words are present.

Other notions of probabilistic disentanglement. Proposition 7 shows that linear factorization of embeddings corresponds to conditional independence of factors z_i given the image y. One might also consider a different sort of probabilistic disentanglement in which conditionals are reversed:

$$p(y|z = (z_1, \ldots, z_k)) = p(y|z_1) \ldots p(y|z_k) q_0(y). \tag{41}$$

This can be viewed as a sort of “causal disentanglement” (similar to the notion used in [48]). It follows from Corollary 8 that linearly factored embeddings mean that

$$p(y|z) = p(y|z_1) \ldots p(y|z_k) p(y)^{1-k} \frac{p(z_1) \ldots p(z_k)}{p(z_1, \ldots, z_k)}. \tag{42}$$

Thus, conditional independence has the same form as equation 41 up to the factor $\frac{p(z_1)\ldots p(z_k)}{p(z_1,\ldots,z_k)}$ (pointwise mutual information) that does not depend on y. If factors are globally

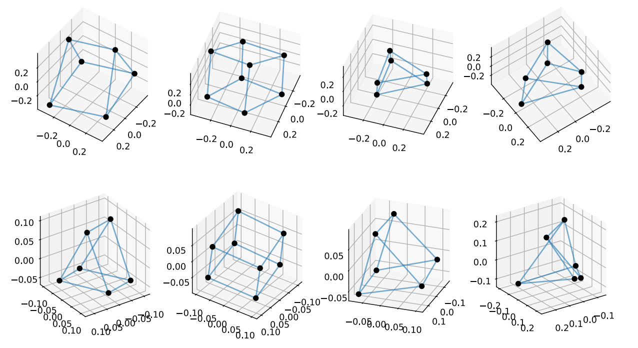

Figure 4: Projected embeddings of manually constructed strings associated with factored concepts, as described in Section 5 in the main body of the paper. *Top:* trained encoder (same as in Figure 2). *Bottom:* visualization of the embeddings for the same strings using a randomly initialized encoder. Even without semantic information, the embeddings in the first three examples are still roughly linearly factored.

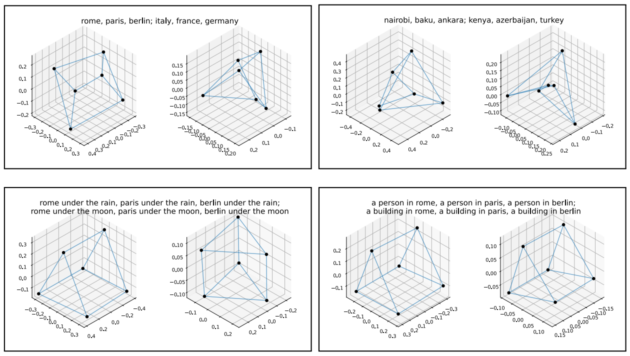

Figure 5: Comparison between projected embeddings using a trained encoder (left figure in each pair) and using a randomly encoder (right figure in each pair). Both encoders lead to symmetric structures when the strings have a factored syntax (bottom row), while only the trained encoder shows these approximate structures when the factorization is semantic (top row).

independent, then equation 42 and equation 41 are equivalent. It is also worth noting that equation 41 does not determine the marginal distribution $p(z = (z_1, \ldots, z_k))$. In general, linear factorization of the embeddings can be seen as a relaxed version of causal disentanglement.

Normalization. Embedding vectors for CLIP are typically normalized, however ideal word vectors are *never* normalized. While this may appear strange, we note that the norm of the embeddings does not carry a probabilistic meaning: we can replace the embeddings $\boldsymbol{u}, \boldsymbol{v}$ from the two modalities with $\boldsymbol{T}\boldsymbol{u}$ and $\boldsymbol{T}^{-1}\boldsymbol{v}$ for any invertible linear transformation $\boldsymbol{T}$ of $\mathbb{R}^d$ without changing the probability model on $\mathcal{X} \times \mathcal{Y}$. In general, ideal word manipulations require starting from normalized embeddings for consistency between modalities, but then normalization is never applied again (in fact, the inner product structure on the embedding space is not used). This explains our modification to the AvgIm+Text approach in Section 5 in the paper.

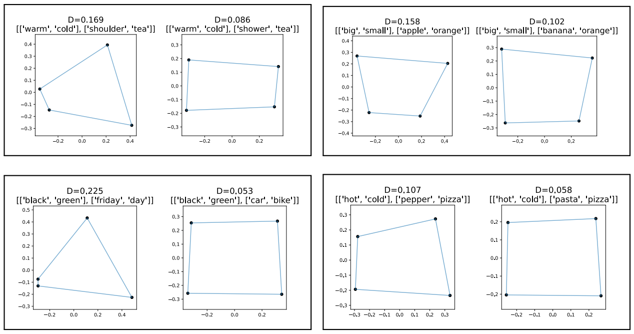

Figure 6: Comparison between projected embeddings for factored strings with and without idioms that have non-compositional meaning (left and right in the subfigures, respectively). We can qualitatively and quantitatively see that idioms lead to weaker compositionality.

Visualizations using SD. We present a few additional visualizations of ideal words using Stable Diffusion. In Figure 7, we consider the same ideal word approximation as in Figure 3 in the main body of the paper and observe the effect of scaling the ideal word corresponding to "green." That is, we consider $\boldsymbol{u}_0 + \boldsymbol{u}_{\text{house}} + \gamma \cdot \boldsymbol{u}_{\text{green}}$ for different γ. In the top row, we compute $\boldsymbol{u}_{\text{green}}$ using the standard "balanced" computation for ideal words (uniform α_i in Proposition 4). In the bottom row, we use weights $\alpha_{\text{house}} = 1$ and $\alpha_{\text{obj}} = 0$ otherwise. This implies that the IW corresponding to $\boldsymbol{u}_{\text{green}}$ is determined by how "green" composes with "house." Amplifying $\boldsymbol{u}_{\text{green}}$ now increases the "greenhouse-ness" of the generated image.

In Figure 8, we consider the problem of *transferring* ideal words. That is, we consider a different (*i.e.*, totally disjoint) set of objects and colors compared to the ones used for Figure 3 in the paper and compute the corresponding ideal words, that we write as $\boldsymbol{u}_{\text{color}'\,\text{object}'} \approx \boldsymbol{u}'_0 + \boldsymbol{u}'_{\text{color}'} + \boldsymbol{u}'_{\text{obj}'}$. We then investigate whether families of ideal words computed independently can be "mixed," combining ideal words for colors from the first collection and ideal words for objects from the second one, and vice-versa. Figure 8 shows that this is possible, at least in our restricted setting. In the first row, we show examples of four new objects with different colors computed by adding associated ideal words ({white, pink, orange, black} × {chair, wallet, shirt, pen}). In the next two rows, we use the ideal words

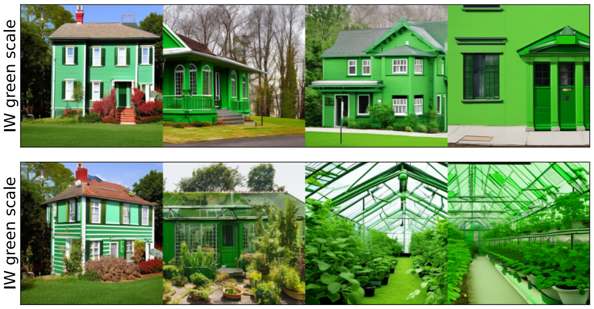

Figure 7: Scaling the ideal word $\boldsymbol{u}_{\text{green}}$ a by factor $\gamma = .5, 1, 1.5, 2$, respectively. *Top:* $\boldsymbol{u}_{\text{green}}$ is computed using all objects as contexts. *Bottom:* $\boldsymbol{u}_{\text{green}}$ is computed only "house" as context.

for objects with the ideal words for colors obtained previously; in the last two rows, we use the ideal words for the new colors together with the ideal words for the objects obtained previously. To obtain all of these images, we simply used $\boldsymbol{u}_{\text{color}'\,\text{object}} \approx (\boldsymbol{u}_0 + \boldsymbol{u}'_0)/2 + \boldsymbol{u}'_{\text{color}'} + 2 \cdot \boldsymbol{u}_{\text{obj}}$ (we found that amplifying the ideal words for objects helps ensure that objects are more centered). Analyzing the limits of this sort of transferability is left for future work.

Finally, in Figure 9 we generate images with ideal words while also using a third "context" factor, in addition to the ones corresponding to color and object (for those we use the same colors and objects as in Figure 3). Here we see that linear compositionality is effective using simple contexts such as {on the beach, on a street} (first two rows), however using more complex contexts such as {underwater, in a volcano} (third and fourth row) it fails to produce good results.

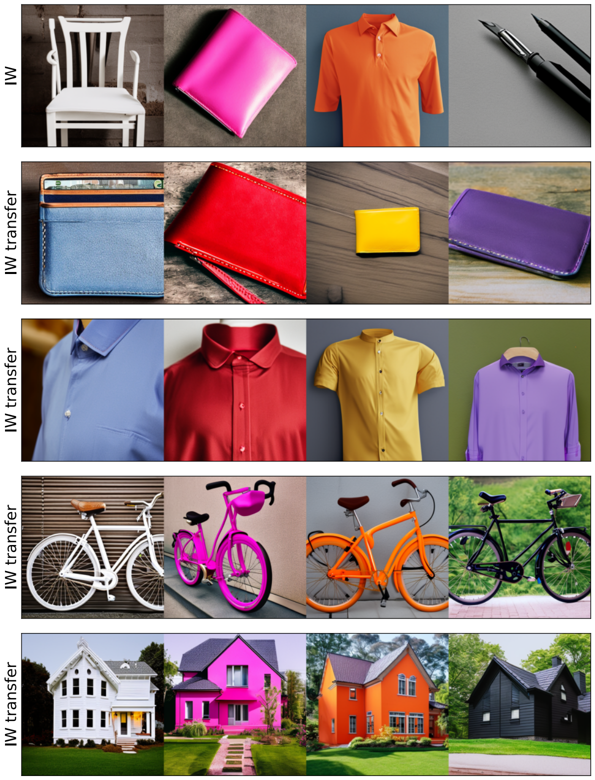

Figure 8: Transferring ideal words. *Top row:* Images generated ideal words for a different set of colors and objects compared to the ones used Figure 3. *Second and third rows:* images generated by adding new ideal words for objects with the previous ideal words for colors; *Fourth and fifth rows:* images generated by adding new ideal words for colors with the previous ideal words for objects.

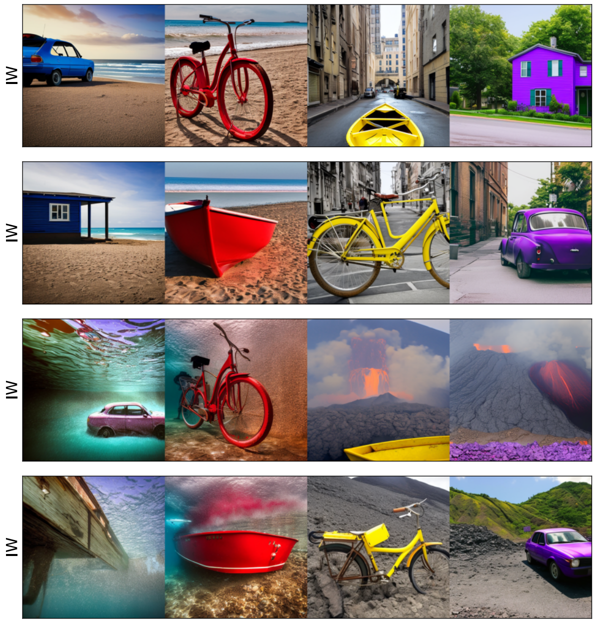

Figure 9: Images generated using ideal words with three factors: color, object, context. *First two rows:* using context factor {on the beach, on a street}; *Second two rows:* using context factor {underwater, in a volcano}.