A Appendix

A.1 Assumptions: generative process

Let the generator $g : \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} = \mathbb{S}^{N-1}$ and $\mathcal{X} \subseteq \mathbb{R}^N$. We assume that the marginal distribution $p(\mathbf{z})$ over latent variables $\mathbf{z} \in \mathcal{Z}$ is uniform:

$$p(\mathbf{z}) = \frac{1}{|\mathcal{Z}|}. \tag{6}$$

Further, we assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is a von Mises-Fisher distribution

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \quad \text{with} \quad C_p := \int e^{\kappa \boldsymbol{\eta}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{7}$$

where κ is a parameter controlled the width of the distribution and $\boldsymbol{\eta}$ is any vector on the hypersphere. Finally, we assume that during training one has access to samples from both of these distributions.

A.2 Assumptions: model

Let $f : \mathcal{X} \to \mathbb{S}_r^{N-1}$, where $\mathbb{S}_r^{N-1}$ denotes a hypersphere with radius r, be the model whose parameters are optimized using constrastive learning. We associate a conditional distribution $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through $h = f \circ g$ and

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \text{with} \quad C_q(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top \; h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{8}$$

where $C_q(\mathbf{z})$ is the partition function and $\tau > 0$.

A.3 Proofs for Sec. 3

We begin be recalling a result of Wang and Isola [19], where the authors show an asymptotic relation between the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ and two loss functions, the *alignment* loss $\mathcal{L}_{\mathsf{align}}$ and the *uniformity* loss $\mathcal{L}_{\mathsf{uniform}}$:

Proposition A (Asymptotics of $\mathcal{L}_{\mathsf{contr}}$, 19)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M \to \infty} \mathcal{L}_{\mathsf{contr}}(f; \tau, M) - \log M = \mathcal{L}_{\mathsf{align}}(f; \tau) + \mathcal{L}_{\mathsf{uniform}}(f; \tau), \tag{9}$$

where

$$\begin{aligned} \mathcal{L}_{\mathsf{align}}(f; \tau) &:= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(\tilde{\mathbf{z}}, \mathbf{z}) \sim p(\tilde{\mathbf{z}}, \mathbf{z})} \left[(f \circ g)(\mathbf{z})^\mathsf{T} (f \circ g)(\mathbf{z}) \right] \\ \mathcal{L}_{\mathsf{uniform}}(f; \tau) &:= \mathop{\mathbb{E}}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{(f \circ g)(\tilde{\mathbf{z}})^\mathsf{T} (f \circ g)(\mathbf{z})/\tau} \right] \right]. \end{aligned} \tag{10}$$

Proof. See Wang and Isola [19]. Note that they originally formulated the losses in terms of observations $\mathbf{x}$ and not in terms of the latent variables $\mathbf{z}$. However, this modified version simplifies notation in the following. □

Based on this result, we show that the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ asymptotically converges to the cross-entropy between the ground-truth conditional p and our assumed model conditional distribution q_h. This is notable, because given the correct model specification for q_h, the cross entropy is minimized iff $q_h = p$, i.e., the ground truth conditional distribution and the model distribution will match.

Theorem 1 ($\mathcal{L}_{\mathsf{contr}}$ converges to cross-entropy between latent distributions)**.** *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M \to \infty} \mathcal{L}_{\mathsf{contr}}(f; \tau, M) - \log M + \log |\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z} \sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z})) \right] \tag{11}$$

where H is the cross-entropy between the ground-truth conditional distribution over positive pairs p and the conditional distribution over the recovered latent space q_{h}, and $C_h(\tilde{\mathbf{z}}) \in \mathbb{R}^+$ is the partition function of q_{h} (see Appendix A.2):

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\tilde{\mathbf{z}})^{-1} e^{h(\tilde{\mathbf{z}})^\mathsf{T} h(\mathbf{z})/\tau} \quad \textit{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\mathsf{T} h(\mathbf{z})/\tau} \, \mathrm{d}\mathbf{z}. \tag{12}$$

Proof. The cross-entropy between the conditional distributions p and q_h is given by

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} [H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))] \tag{13}$$

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} [-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})] \right] \tag{14}$$

$$= \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[-\frac{1}{\tau} h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) + \log C_h(\mathbf{z}) \right] \tag{15}$$

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log C_h(\mathbf{z}) \right]. \tag{16}$$

Using the definition of C_h in Eq. (12) yields

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \int_{\mathcal{Z}} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}} \right]. \tag{17}$$

We assume a uniform marginal distribution $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$. We expand by $|\mathcal{Z}||\mathcal{Z}|^{-1}$ and estimate the integral by sampling from $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$, yielding

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log |\mathcal{Z}| \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \right] \right] \tag{18}$$

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \right] \right] + \log |\mathcal{Z}| \tag{19}$$

By inserting the definition $h = f \circ g$,

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[(f \circ g)(\tilde{\mathbf{z}})^\top (f \circ g)(\mathbf{z}) \right] \tag{20}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})\tau} \right] \right] + \log |\mathcal{Z}|, \tag{21}$$

we can identify the losses introduced in Proposition A,

$$=\mathcal{L}_{\mathsf{align}}(f;\tau) + \mathcal{L}_{\mathsf{uniform}}(f;\tau) + \log |\mathcal{Z}|, \tag{22}$$

which recovers the original alignment term and the uniformity term for maximimizing entropy by means of a von Mises-Fisher KDE up to the constant $\log |\mathcal{Z}|$. According to proposition A this equals

$$= \lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log |Z|, \tag{23}$$

which concludes the proof. □

Proposition 1 (Minimizers of the cross-entropy maintain the dot product). *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} \exp(\kappa \tilde{\mathbf{z}}^\top \mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau/\kappa}$.*[3] *Consider the model conditional distribution*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \quad \textit{with} \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \, \mathrm{d}\tilde{\mathbf{z}}, \tag{24}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h^* is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.

[3] Note that in practice this can be implemented as a learnable rescaling operation of the network f.

Proof. Note that $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h—in particular, for $h(\mathbf{z}) = \sqrt{\tau/\kappa}\mathbf{z}$. The global minimum of the cross-entropy between two distributions is reached if they match by value and have the same support. Thus, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}^*}(\tilde{\mathbf{z}}|\mathbf{z}). \tag{25}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using that h maps from a unit hypersphere onto one with radius $\sqrt{\tau/\kappa}$ yields

$$\begin{aligned} p(\mathbf{z}|\mathbf{z}) &= q_{\mathrm{h}^*}(\mathbf{z}|\mathbf{z}) & (26)\\ C_p \exp(\kappa \mathbf{z}^\top \mathbf{z}) &= C_h(\mathbf{z}) \exp(h^*(\mathbf{z})^\top h^*(\mathbf{z})/\tau) & (27)\\ C_p \exp(\kappa) &= C_h(\mathbf{z}) \exp(\kappa) & (28)\\ C_p &= C_h. & (29) \end{aligned}$$

As the normalization constants are identical we get for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$\exp(\kappa \mathbf{z}^\top \tilde{\mathbf{z}}) = \exp(h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}})) \Leftrightarrow \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}}). \tag{30}$$

□

Lemma 1 (Kernel Density Estimators in the limit of unlimited samples)**.** *For* $0 < \alpha, \kappa < \infty$*, we have*

$$p(\mathbf{z}) = \frac{1}{\alpha} \int p(\tilde{\mathbf{z}}) \exp(-\alpha \|\mathbf{z} - \tilde{\mathbf{z}}\|) \mathrm{d}\tilde{\mathbf{z}} \qquad \mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{R}^N \tag{31}$$

$$p(\mathbf{z}) = C(\kappa) \int p(\tilde{\mathbf{z}}) \exp(\kappa \mathbf{z}^\top \tilde{\mathbf{z}}) \mathrm{d}\tilde{\mathbf{z}} \qquad \mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{S}^{N-1}. \tag{32}$$

Proof. First, we consider $\mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{R}^N$. We can define the Kernel Density Estimator (KDE) for which $\lim_{n\to\infty} \hat{p}_n(\mathbf{z}) = p(\mathbf{z})$ [34]. We use this result to obtain

$$\begin{aligned} \lim_{n\to\infty} \hat{p}_n(\mathbf{z}) &= \lim_{n\to\infty} \frac{1}{\alpha} \sum_{i=1}^{n} \frac{1}{n} \exp(-\alpha \|\mathbf{z} - \mathbf{z}_i\|) \mathrm{d}\tilde{\mathbf{z}}, \quad \mathbf{z}_i \sim^{\text{iid}} p(\mathbf{z}) & (33)\\ &= \frac{1}{\alpha} \int_{\mathbb{R}^N} p(\mathbf{z}) \exp(-\alpha \|\mathbf{z} - \tilde{\mathbf{z}}\|)\, \mathrm{d}\tilde{\mathbf{z}} = p(\mathbf{z}). & (34) \end{aligned}$$

The result for $\mathbb{S}^{N-1}$ can be derived analogously. □

Lemma 2 (Matching conditionals imply bounded pushforward densities)**.** *Let p and q_{h} be conditional distributions as defined above. If they match, i.e. $p = q_{\mathrm{h}}$, and the marginal distribution is bounded, i.e. $p(\mathbf{z}) \leq p_{max} < \infty$, the pushforward density $p_{\#h}$ of $h(\mathbf{z}), \mathbf{z} \sim p(\mathbf{z})$ is bounded.*

Proof. We will show, that the KDE of the pushforward distribution $p_{\#h}$ is related to the normalization constant $C_h(\mathbf{z})$. Let $\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \|\mathbf{z} - \tilde{\mathbf{z}}\|$ or $\delta(\mathbf{z}, \tilde{\mathbf{z}}) = -\mathbf{z}^\top \tilde{\mathbf{z}}$ be the function used in the model conditional q_h and let C_δ be the appropriate normalization constant as defined in Lemma 1. For $p(\mathbf{z}) < \infty$, we get (cf. Lemma 1)

$$\begin{aligned} p_{\#h}(\mathbf{z}) &= C_\delta \int p(\tilde{\mathbf{z}}) \exp(-\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})))\, d\tilde{\mathbf{z}} & (35)\\ &\leq C_\delta \int p_{\max} \exp(-\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})))\, d\tilde{\mathbf{z}} & (36)\\ &= C_\delta\, p_{\max}\, C_h(\mathbf{z}). & (37) \end{aligned}$$

For matching distributions $q_h = p$, we have $C_h(\mathbf{z}) = C_p$ acccording to Proposition 3 and get

$$p_{\#h}(\mathbf{z}) \leq C_\delta\, p_{\max}\, C_p \tag{38}$$

Since both $p(\mathbf{z}) < \infty$ and $C_p < \infty$ by assumption, it follows that $p_{\#h}(\mathbf{z}) < \infty$, concluding the proof. □

Lemma 3. *Let h be as described in Proposition 1, such that it minimizes the cross-entropy. Then the empirical marginal distribution of the model (pushforward of p through h, denoted as $p_{\#h}$) is a uniform distribution, i.e. $p_{\#h}(\mathbf{z}) = const.$*

Proof. As per Proposition 1, the conditional distributions' normalization constants have the same value, i.e. $C_p = C_q$. However, as noted above (see Sec. A.2), the definition of C_q coincides with a von Mises-Fisher KDE of the empirical marginal distribution $p(h(\mathbf{z}))$ in the limit of infinite samples (Lemma. 1, 35). As this C_q equals $C_p = |\mathcal{Z}|^{-1}$, the empirical distribution $p(h(\mathbf{z}))$ is also a uniform distribution. □

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$. If $h : \mathcal{Z} \to \mathcal{Z}$ maintains the dot product up to a constant factor, i.e. $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an affine transformation.*

Proof. As h maintains the dot product up to a factor, this also holds true if one rotates the coordinate system by an arbitrary rotation matrix $\mathbf{R} \in \mathrm{SO}(N)$. Thus, we get

$$\forall \mathbf{R} \in \mathrm{SO}(N), \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}^\top \mathbf{R}^\top \mathbf{R} \tilde{\mathbf{z}} = h(\mathbf{R}\mathbf{z})^\top h(\mathbf{R}\tilde{\mathbf{z}}). \tag{39}$$

We consider the partial derivatives w.r.t. $\mathbf{z}$ and obtain:

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \tilde{\mathbf{z}} = \mathbf{R}\mathbf{J}_h^\top(\mathbf{R}\mathbf{z}) h(\mathbf{R}\tilde{\mathbf{z}}). \tag{40}$$

We can recover the initial dot product by multiplying both sides of the equation with $\mathbf{z}^\top$ to obtain

$$\begin{aligned}\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} &= \mathbf{z}^\top \mathbf{R}\mathbf{J}_h^\top(\mathbf{R}\mathbf{z}) h(\mathbf{R}\tilde{\mathbf{z}}) && (41)\\ &= h(\mathbf{R}\tilde{\mathbf{z}})^\top \mathbf{J}_h(\mathbf{R}\mathbf{z}) \mathbf{R}^\top \mathbf{z}. && (42)\end{aligned}$$

From here, we take the partial derivative on both sides, this time w.r.t. $\tilde{\mathbf{z}}$, yielding

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z} = [\mathbf{R}\mathbf{J}_h(\mathbf{R}\tilde{\mathbf{z}})\mathbf{J}_h^\top(\mathbf{R}\mathbf{z})\mathbf{R}^\top]\mathbf{z}. \tag{43}$$

Multiplying with $\mathbf{R}^\top$ from the left and defining $\mathbf{z}' := \mathbf{R}^\top \mathbf{z}$ gives

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}' = [\mathbf{J}_h(\mathbf{R}\tilde{\mathbf{z}})\mathbf{J}_h^\top(\mathbf{R}^2\mathbf{z}')]\mathbf{z}'. \tag{44}$$

We define a transform from $(\mathbf{R}, \mathbf{z}, \tilde{\mathbf{z}})$ to $(\mathbf{a}, \mathbf{b}, \mathbf{z}')$: First, we select $\mathbf{R}$ and $\mathbf{z}$ s.t. $\mathbf{z}' = \mathbf{R}^T \mathbf{z}$ and $\mathbf{b} = \mathbf{R}\mathbf{z} = \mathbf{R}^2\mathbf{z}'$. Then, we select $\tilde{\mathbf{z}}$ s.t. $\mathbf{a} = \mathbf{R}\tilde{\mathbf{z}}$. With this transform, we rewrite the aforementioned equation and obtain:

$$\forall \mathbf{a}, \mathbf{b}, \mathbf{z}' \in \mathcal{Z} : \kappa \mathbf{z}' = [\mathbf{J}_h(\mathbf{a})\mathbf{J}_h(\mathbf{b})^\top]\mathbf{z}', \tag{45}$$

which can only be satisfied iff

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{a})\mathbf{J}_h(\mathbf{b})^\top = \kappa \mathbf{I}. \tag{46}$$

By evaluating this expression for $\mathbf{a} = \mathbf{b}$ we get the

$$\forall \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{b})^\top = \kappa \mathbf{J}_h^{-1}(\mathbf{b}) \tag{47}$$

Inserting this property again in the previous expression yields

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{a})\kappa\mathbf{J}_h(\mathbf{b})^{-1} = \kappa \mathbf{I}, \tag{48}$$

and finally:

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{a}) = \mathbf{J}_h(\mathbf{b}). \tag{49}$$

□

Taking all of this together, we can now prove Theorem 2:

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$ and the ground-truth marginal be uniform and the conditional a vMF distribution (cf. Eq. 2). If the mixing function g is differentiable and invertible, and if f is differentiable and minimizes the CL loss (1), then for fixed $\tau > 0$ and $M \to \infty$ we find that $h = f \circ g$ is affine, i.e. we recover the latent sources up to affine transformations.*

Proof. As f minimzes the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ we can apply Theorem 1 to see that f also minimizes the cross-entropy between $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$. This means, we can apply Proposition 1 to show that the concatenation $h = f \circ g$ is an isometry with respect to the dot product. Finally, according to Proposition 2, h must then be a linear transformation on the hypersphere, i.e., a combination of permutations, sign flips and rotations. Thus, f recorvers the latent sources up to affine transformations, concluding the proof. □

A.4 Extension to (subspaces of) $\mathbb{R}^N$

Here, we show how one can generalize the theory above from $\mathcal{Z} = \mathbb{S}^{N-1}$ to $\mathcal{Z} \subseteq \mathbb{R}^N$. Under mild assumptions regarding the ground-truth conditional distribution p and the model distribution q_{h}, we prove that all minimizers of the cross-entropy between p and q_{h} are linear functions, if $\mathcal{Z} = \mathbb{R}^N$ or $\mathcal{Z}$ is a convex body. Note that the hypercube $[a_1, b_1] \times \ldots \times [a_N, b_N]$ is an example of such a convex body.

A.4.1 Assumptions

First, we restate the core assumptions for this proof. The main difference with the assumptions for the hyperspherical case above is that we assume different conditional distributions: instead of rotation-invariant von Mises-Fisher distributions, we use translation-invariant distributions (up to restrictions determined by the finite size of the space) of the exponential family.

Generative Process Let $g : \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} \subseteq \mathbb{R}^N$ and $\mathcal{X} \subseteq \mathbb{R}^N$, with $\mathcal{Z} = \mathbb{R}^N$ or $\mathcal{Z}$ being a convex body (e.g. a hypercube). Further, let the marginal distribution fulfill $\operatorname{supp} p(\mathbf{z}) = \mathcal{Z}$. We assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is an exponential distribution

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) e^{-\lambda\delta(\tilde{\mathbf{z}},\mathbf{z})} \quad \text{with} \quad C_p(\mathbf{z}) := \int e^{-\lambda\delta(\mathbf{z},\tilde{\mathbf{z}})} \,\mathrm{d}\tilde{\mathbf{z}}, \tag{50}$$

where δ is a semi-metric, and $\lambda > 0$ a parameter controlling the width of the distribution. We make no further assumptions on the marginal distribution $p(\mathbf{z})$ except that it must be compatible with the conditional distribution:

$$p(\mathbf{z}) = \int p(\mathbf{z}, \tilde{\mathbf{z}}) \,\mathrm{d}\tilde{\mathbf{z}} = \int p(\mathbf{z}|\tilde{\mathbf{z}}) p(\tilde{\mathbf{z}}) \,\mathrm{d}\tilde{\mathbf{z}}. \tag{51}$$

Finally, we assume that during training one has access to samples from both of these distributions.

Model Let $\mathcal{Z}'$ be a subset of $\mathbb{R}^N$ that is a convex body and let $f : \mathcal{X} \to \mathcal{Z}'$ be the model whose parameters are optimized. We associate a conditional distribution $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))} \quad \text{with} \quad C_q(\mathbf{z}) := \int e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))} \,\mathrm{d}\tilde{\mathbf{z}}, \tag{52}$$

where $C_q(\mathbf{z})$ is the partition function and δ is defined above.

A.4.2 Minimizing the cross-entropy

In a first step, we derive a property similar to Theorem 1, which suggests a practical method to find minimizers of the cross-entropy between the ground-truth p and model conditional q_{h}. For this, we suggest a modification of the contrastive loss Eq. (1).

Proposition A.4.2. *Let δ be a semi-metric. Consider the ground-truth conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}))$ and the model conditional distribution*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))), \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\mathrm{d}\tilde{\mathbf{z}}. \tag{53}$$

Then the cross-entropy between p *and* q_{h} *is upper bounded by*

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}[H(p(\cdot|\mathbf{z}), q_{\mathrm{h}}(\cdot|\mathbf{z})] \leq \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z}))}}[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))] + \tag{54}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}[\exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))]\right)\right] + \mathit{const.}, \tag{55}$$

which can be implemented by sampling data from the accessible distributions. This bound becomes tighter, the more uniform *the marginal distribution* p *becomes. Note, the similarity to the InfoNCE objective in Eq. (1), where the distance function* δ *corresponds to the dot-product.*

Proof.

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}[H(p(\cdot|\mathbf{z}), q_{\mathrm{h}}(\cdot|\mathbf{z})] = -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}[\log(q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}))]\right] \tag{56}$$

By inserting the definition of q_{h} one gets

$$= -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\log(C_h^{-1}(\mathbf{z})) - \delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right]\right] \tag{57}$$

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}[\log(C_h(\mathbf{z})) + \delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))]\right] \tag{58}$$

As $C_h(\mathbf{z})$ does not depend on $\tilde{\mathbf{z}}$ it can be moved out of the inner expectation value, yielding

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))] + \log(C_h(\mathbf{z}))\right], \tag{59}$$

which can be written as

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}[\log(C_h(\mathbf{z}))], \tag{60}$$

Inserting the definition of C_h gives

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\int \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right)\right]. \tag{61}$$

By using that p is a regular probability density and writing $p_{\text{max}} = \max_{\mathbf{z}} p(\mathbf{z})$, the second term can be upper bounded with

$$\leq \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\int \frac{p(\tilde{\mathbf{z}})}{p_{\text{max}}}\exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right)\right], \tag{62}$$

with equality if the marginal p is a uniform distribution over $\mathcal{Z}$. Finally, this can be simplified as

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}[\exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))]\right)\right] - \log p_{\text{max}}. \tag{63}$$

□

A.4.3 Cross-entropy minimizers are isometries

Now we show a version of Proposition 1, that is generalized from hyperspherical spaces to (subsets of) $\mathbb{R}^N$. We note that while providing an exact link between cross-entropy minimization and CL with marginals on $\mathbb{R}^N$ is beyond the scope of this work, this result can still serve as a stepping stone for generalizing our identifiability result on the hypersphere in future work.

Proposition 3 (Minimizers of the cross-entropy are isometries)**.** *Let δ be a semi-metric. Consider the conditional distributions of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z})\exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}))$ and*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z})\exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))), \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\mathrm{d}\tilde{\mathbf{z}}, \tag{64}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h^* is a minimizer of the cross-entropy $\mathcal{L}_{\mathsf{CE}} = \mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then h is an isometry, i.e. $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \lambda\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h^*(\mathbf{z}), h^*(\tilde{\mathbf{z}}))$. Note, that this does not depend on the choice of $\mathcal{Z}$ but just on the class of conditional distributions allowed.

Proof. Note that $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h, e.g. the identity. The global minimum of cross-entropy between two distributions is reached if they match by value and have the same support. Hence, if p is a regular density, q_{h} will be a regular density, i.e. q_{h} is continuous and has only finite values $0 \leq q_{\mathrm{h}} < \infty$. As the two distributions match, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{h^*}(\tilde{\mathbf{z}}|\mathbf{z}). \tag{65}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using the property $\delta(\mathbf{z}, \mathbf{z}) = 0$ yields

$$\begin{aligned} & p(\mathbf{z}|\mathbf{z}) = q_{\mathrm{h}*}(\mathbf{z}|\mathbf{z}) & (66)\\ \Leftrightarrow \quad & C_p^{-1}(\mathbf{z})\exp(-\lambda\delta(\mathbf{z}, \mathbf{z})) = C_h^{-1}(\mathbf{z})\exp(-\delta(h^*(\mathbf{z}), h^*(\mathbf{z}))) & (67)\\ \Leftrightarrow \quad & C_p(\mathbf{z}) = C_h(\mathbf{z}). & (68) \end{aligned}$$

As the normalization constants are identical, we obtain for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$\exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z})) = \exp(-\delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z}))) \Leftrightarrow \lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}) = \delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z})). \tag{69}$$

By introducing a new semi-metric $\delta' := \lambda^{-1}\delta$, we can write this as $\delta(\tilde{\mathbf{z}}, \mathbf{z}) = \delta'(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z}))$, which shows that h is an isometry. □

A.4.4 Bijectivity of h

We proceed by showing an important property that will be needed later in Sec. A.5: the bijectivity of h. For this, we show that if the conditional probability distribtion q_{h}, which is implicitly defined by h, is a regular density function (i.e. $0 \leq q_{\mathrm{h}} < \infty$), then h is bijective. First, we start with introducing some existing concepts:

Definition 1. Let $\mathcal{M}, \mathcal{N}$ be manifolds. A map $h : M \to N$ is *proper* if for every compact set $S \in \mathcal{N}$ its preimage $h^{-1}(S)$ is compact in $\mathcal{M}$.

Proposition 4. *Let $\mathcal{M}$ and $\mathcal{N}$ be simply connected, oriented, d-dimensional $\mathcal{C}^1$-submanifolds of $\mathbb{R}^D$ ($D \geq d$), without boundary. Let $h : M \mapsto N$ be a proper $\mathcal{C}^1$-map such that the Jacobian determinant $|\mathbf{J}_h|$ never vanishes. Then h is bijective.*

Proof. See Theorem 2.1 in [36]. □

Many important manifolds fit the conditions of this proposition, including $\mathbb{R}^N$. We now prove that under mild conditions, any h, such that q_{h} as defined above is a regular density function, is proper and has a non-vanishing Jacobian determinant.

Proposition 5. *Let $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map, such that q_{h} as defined abvove is a regular density function; this also means that $\sup q_{\mathrm{h}} < \infty$. Then h is proper and has a non-vanishing Jacobian determinant.*

Proof. Suppose that the Jacobian of h vanishes for some $\mathbf{z} \in \mathcal{M}$. Then the inverse of the determinant of the Jacobian goes to infinity at this point and so does the density of $h(\mathbf{z})$ according to the well-known transformation of probability densities. By assumption, q_{h} must be a regular density function and, therefore, cannot be a delta distribution.

The mapping h is proper if pre-images of compact spaces are compact. According to the Heine–Borel theorem, compact subsets of $\mathbb{R}^D$ are closed and bounded. Additionally, a continuous mapping between $\mathcal{M}$ and $\mathcal{N}$ is also closed, i.e. pre-images of closed subsets are also closed [37]. In addition, it is well-known that continuous functions on compact spaces are bounded[4], concluding the proof. □

[4] See e.g. https://proofwiki.org/wiki/Continuous_Function_on_Compact_Space_is_Bounded

Finally, this is sufficient to prove that the map h is bijective, if q_{h} is a regular density:

Theorem 3. *Let $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map, such that q_{h} as defined abvove is a regular density function, and with $\mathcal{M}$, $\mathcal{N}$ defined as above. Then h is bijective.*

Proof. According to Proposition 5, h is proper and has non-vanishing Jacobians. Then, according to Proposition 4, h is bijective. □

This general property also holds for the special case of maps between $\mathbb{R}^N$:

Corollary 1. *Let $h : \mathbb{R}^N \to \mathbb{R}^N$ be a differentiable map, such that q_{h} as defined abvove is a regular density function as defined as above. Then h is bijective.*

A.5 Cross-entropy minimization identifies the ground-truth factors

Before we continue, let us recall a Theorem by Mazur and Ulam [38]:

Theorem 4 (Mazur-Ulam)**.** *Every bijective isometry between real normed spaces is affine.*

Proof. See Mazur and Ulam [38] or Nica [39]. □

Furthermore, let us recall an extension of that Theorem for convex bodies by Mankiewicz [40]:

Theorem 5 (Mankiewicz)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces and let $\mathcal{V}$ be a convex body in $\mathcal{X}$ and $\mathcal{W}$ a convex body in $\mathcal{Y}$. Then every isometry of between $\mathcal{V}$ and $\mathcal{W}$ can be uniquely extended to an affine isometry between $\mathcal{X}$ and $\mathcal{Y}$.*

Proof. See Mankiewicz [40]. □

By combining the properties derived before we can show that h is an affine function:

Theorem 6. *Let $\mathcal{Z}, \mathcal{Z}'$ be a convex bodies in $\mathbb{R}^N$ or the entire $\mathbb{R}^N$. If h minimizes the cross-entropy between p and q_{h} as defined in (4) and if the mixing function $\mathbf{g}$ is differentiable and invertible, then we find that $h = f \circ g$ is affine, i.e. we recover the latent sources up to affine transformations.*

Proof. According to Proposition 3 h is an isometry and q_{h} is a regular probability density function. Then, according to Corollary 1 h is also bijective. Finally, Theorem 4 says that h is an affine transformation. □

Note, that this result can be seen as a first step towards a generalized version of Theorem 2, as it is valid for $\mathcal{Z} = \mathbb{R}^N$ and allows a larger variety of conditional distributions. A missing step is to derive a connection between minimizing $\mathcal{L}_{\mathsf{contr}}$ and minimizing the cross-entropy, just as Theorem 1 does for the hyperspherical case. This will be addressed in future work.

A.6 Extended experiments

A.6.1 Implementation

Inspired by the experimental evaluation by Wang and Isola [19], we use a convex combination of the $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ losses. Given the severe mismatch between DSprites and our theory's assumptions, we generalized the $\mathcal{L}_{\mathsf{align}}$ loss function for more flexibility, resulting in the loss function

$$\mathcal{L}_{\mathsf{AU}}(\alpha, p) := \alpha \widetilde{\mathcal{L}}_{\mathsf{uniform}}(p) + (1-\alpha)\widetilde{\mathcal{L}}_{\mathsf{align}}(p), \tag{70}$$

with the components

$$\widetilde{\mathcal{L}}_{\mathsf{align}}(p) := \mathop{\mathbb{E}}_{(\tilde{\mathbf{z}},\mathbf{z})\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[\|(f \circ g)(\mathbf{z}) - (f \circ g)(\mathbf{z})\|_p^p\right] \tag{71}$$

$$\widetilde{\mathcal{L}}_{\mathsf{uniform}}(p) := \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\mathbf{z}^-\sim p(\mathbf{z}^-)} \left[e^{-\|(f\circ g)(\mathbf{z})-(f\circ g)(\mathbf{z})\|_p^p}\right]\right]. \tag{72}$$

A.6.2 Disentanglement Datasets

Standard datasets for disentanglement, most of which have been compiled by Locatello et al. [41], are limited in that the data generating process is independent and identically distributed (*i.i.d.*). This is problematic for evaluating the identifiability capabilities of contrastive methods, which require *positive pairs* to enforce alignment in latent space. Recent work in disentanglement, which make assumptions about pairs of instances in order to perform nonlinear demixing in an unsupervised fashion [25, 31, 32], introduced datasets in order to overcome the cited issues. We leverage these datasets for evaluation on more complex inputs.

UNI [31] Pairs of images are combined such that only k factors change, where $k \in \mathcal{U}\{1, D-1\}$ and D denotes the number of ground-truth factors, where the next value for each of the k factors is sampled uniformly from the set of possible values.

LAP [32] For each ground-truth factor, the first value in the pair is chosen i.i.d. from the dataset and the second is chosen by weighting nearby factor values using Laplace distributed probabilities. If all factors remain constant (no transition), then the sample is rejected because the pair would not result in any temporal learning signal.

NAT [32] For a given image pair, the position and scale of the sprite objects are set using measured values from adjacent time points for natural objects in YouTube-VOS [42, 43]. The sprite shapes are simple, like dSprites [44], and fixed for a given pair. The sprite orientations are fixed for the pair and are sampled uniformly from the same distribution as was used for dSprites. A version discretized to the granularity of dSprites was contributed, we refer to the discrete version as **ND** and the continuous version as **NC**.

KITTI [32] The dataset is composed of pedestrian segmentation masks from an autonomous driving vision benchmark KITTI-MOTS [45], with natural shapes and continuous natural transitions.

A.6.3 Extended Disentanglement Results

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
Ada-GVAE (UNI)*	92.3	84.7	26.6	47.9	91.3	7.4
SlowVAE (UNI)*	90.4 (3.3)	81.35 (7.6)	35.7 (8.3)	52.1 (6.5)	87.6 (2.0)	5.1 (1.4)
$\mathcal{L}_{\mathsf{AU}}(0.9999, 1)$ (UNI)	82.5 (3.4)	62.9 (3.8)	7.8 (1.9)	22.2 (2.9)	98.8 (0.7)	4.9 (1.5)
$\mathcal{L}_{\mathsf{AU}}(0.9, 1)$ (UNI)	80.2 (0.5)	61.0 (0.3)	20.5 (0.2)	42.4 (0.4)	100.0 (0.0)	7.6 (0.1)
SlowVAE (LAP)*	100.0 (0.0)	98.32 (2.5)	27.8 (7.9)	65.3 (3.1)	97.0 (1.5)	6.1 (2.6)
$\mathcal{L}_{\mathsf{AU}}(0.999, 1)$ (LAP)	100.0 (0.0)	96.0 (3.6)	24.9 (2.5)	45.6 (0.8)	92.9 (0.1)	9.6 (2.5)
$\mathcal{L}_{\mathsf{AU}}(0.99, 1)$ (LAP)	100.0 (0.0)	94.5 (3.3)	17.4 (3.7)	54.0 (1.8)	94.7 (1.4)	7.0 (0.6)
$\mathcal{L}_{\mathsf{AU}}(0.9, 1)$ (LAP)	100.0 (0.0)	89.5 (1.7)	18.1 (4.6)	54.4 (0.5)	90.5 (0.6)	6.7 (0.7)

Table 2: **DSprites.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (rows highlighted with * are from [32]).

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP	MCC
β-VAE	78.1 (3.0)	60.6 (6.0)	4.6 (1.9)	10.3 (1.8)	87.8 (2.3)	2.1 (1.0)	41.7 (3.4)
SlowVAE	82.6 (2.2)	76.2 (4.8)	11.7 (5.0)	18.9 (5.5)	88.1 (3.6)	4.4 (2.3)	52.6 (4.1)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$	72.2 (6.3)	62.1 (4.8)	9.4 (2.2)	18.1 (2.5)	96.5 (0.6)	5.1 (0.5)	33.4 (5.1)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$	55.7 (5.1)	44.1 (4.3)	1.7 (1.0)	3.2 (0.5)	91.2 (3.0)	0.7 (0.4)	22.4 (4.1)

Table 3: **Discrete Natural Sprites.** Mean (s.d.) performance levels over 10 random seeds.

Model (Data)	MCC
β-VAE	42.6 (4.7)
SlowVAE (C)	49.1 (4.0)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$	41.7 (5.3)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$	25.8 (4.6)

Table 4: **Continuous Natural Sprites**. Mean (s.d.) over 10 random seeds.

Model (frame separation)	MCC
β-VAE	62.7 (7.1)
SlowVAE (Δ t=1)	66.1 (4.5)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$ (Δ t=1)	77.1 (1.7)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$ (Δ t=1)	64.8 (2.4)
SlowVAE (Δ t=5)	79.6 (5.8)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$ (Δ t=5)	79.6 (2.6)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$ (Δ t=5)	68.3 (2.6)

Table 5: **KITTI Masks**. Mean (s.d.) over 10 random seeds.