APPENDIX A: OUR APPROACH UNDER THE DEFINITION OF DISENTANGLEMENT

Our method follows from a prior definition of disentanglement (Shu et al., 2019), which decomposes disentanglement into two components, *restrictiveness* and *consistency*. Restrictiveness over a set of latent dimensions $z_I = z_{j:k}$ is met when changes to z_I correspond to only changes to the factors of variation s_I. Consistency is met when changes to the set of factors s_I is only controlled by changes to the latent dimensions z_I. This decomposition allows for statistically dependent factors of variation s_i to exist, which is often present in naturally occurring data. For a model to be fully disentangled requires every index of the model to be disentangled.

We approach consistency and restrictiveness from measuring manifold topology. We first assume the manifold X of a generative model is equipped with an atlas $g_j : Z_j \to X$, where Z_j is an open subset of $\mathbb{R}^n$ and n is the dimension of the manifold. Additionally, we have factors of variation $s_i : X \to \mathbb{R}^n$. We would like to drive only one of these maps s_i, while leaving the others $s_{\setminus i}$ fixed. Of course, we now have the composite maps $s_i \circ g_j : Z_j \to \mathbb{R}^n$, which can be thought of as a map from a small ball in $Z_j \in \mathbb{R}^n$ into $\mathbb{R}^n$. Our goal to find maps $\alpha_j : \mathbb{R}^n \to \mathbb{R}^n$ so that the map $s_i \circ g_j \circ \alpha_j$ has the form $(z_1, \ldots, z_n) \to (f_1(z_1), \ldots, f_n(z_n))$ and the i-th output depends only on the i-th input. Because $z_i \to f_i(z_i)$ is topologically distinct, we can use this to evaluate disentanglement.

We derive measures of failure on this diagonalization, and our idea is to study the submanifolds $(s_i \circ g_j)^{-1}(z_{\setminus i})$, in particular the persistent homology of these submanifolds, to generate an evaluation metric which is guiding us towards the disentangled situation. We believe that under a perfectly disentangled model, perturbing the value of z_i should not change the topology of the manifold $(s_i \circ g_j)^{-1}(z_i)$ (restrictiveness) or $(s_{\setminus i} \circ g_j)^{-1}(z_{\setminus i})$ (consistency). From this, we measure disentanglement through its decomposition of consistency and restrictiveness, by comparing the persistence barcodes of these submanifolds.

We also cluster latent dimensions, so our evaluation metric rewards disentangled clusters, and moreover rewards maximizing the number of disentangled clusters, which would be interpreted as products of tangled manifolds.

Assumptions. We make the following assumptions:

- *Assumption A.* Each map $z_i \to f_i(z_i)$ is topologically distinct.
- *Assumption B.* We can measure the persistent homology of the generated space.
- *Assumption C.* If a set of mappings $z_j \to f_j(z_j)$ are not topologically distinct, then we can treat their shared z_i or s_i dimension as the same dimension.
- *Assumption D.* In the supervised case, each s_i can be observed for each $x \in X$.

With our method, we can evaluate the degree to which a set of latent dimensions z_I corresponds to a single s_i. This is a stronger form of *restrictiveness* that disentanglement necessitates. In order to identify z_I, we cluster topologically similar latent dimensions. We *penalize intra-cluster variance*, which discourages having a set of latent dimensions z_I correspond to distinct factors of variation and which denotes higher restrictiveness.

Furthermore, we can evaluate the degree to which a single factor s_i is affected by different clusters of latent dimensions $\{z_I, z_J, \ldots\}$, which also control other factors of variation. Removing shared factors increases consistency, by increasing the distance between distinct clusters. As a result, distinct clusters cannot be similar if they are consistent. This corresponds to a stronger form of *consistency* that disentanglement necessitates. Thus, we *penalize extra-cluster similarity*, which encourages topological variation between clusters and which denotes higher consistency.

Additional note on other definitions of disentanglement. As noted in a foundational paper on disentanglement (Bengio et al., 2013), disentanglement constitutes a bijective mapping between factors of variation in the data to dimensions in the latent space Z, e.g. $\forall_i \mathbf{z}_i = e(g(\mathbf{z}_i))$. Using homology, we can determine whether this bijective mapping holds along different factors, by observing the topological similarity of their conditional submanifolds and measuring the extent to which they continuously deform into each other. Aligned with newer definitions of disentanglement (Higgins et al., 2018; Duan et al., 2020), our framing permits multiple valid factorizations, where different clusters of likely homeomorphic submanifolds conditioned on factors can compose alternate factorizations. Our supervised variant is meant to consider a target factorization corresponding to factors on the real

manifold. In this variant, we follow the existing definition of supervised disentanglement (Shu et al., 2019) that allows different subsets of dimensions to contribute to a target factor and for target factors to exhibit statistical dependence. Additionally, based on the motivating description of MIG-sup (Li et al., 2020), Mutual Information Gap (MIG) (Chen et al., 2018b) and MIG-sup (Li et al., 2020) correspond, respectively, to the disentanglement concepts of restrictiveness and consistency.

APPENDIX B: ADDITIONAL BACKGROUND

In addition to the background section of our paper, we would like to point out related work, some reiterated for ease of cohesively examining the related literature.

Disentanglement metrics. Existing disentanglement metrics depend on an external model such as an encoder or classifier to be applicable across datasets, or dataset-specific preprocessing. Several metrics train classifiers to detect separability of the data, generated by conditioning on different latent dimensions (Eastwood and Williams, 2018; Kim and Mnih, 2018; Karras et al., 2019). These are reliant on hyperparameters and the architecture of the classifiers. Recently, the mutual information gap MIG was proposed as an information-theoretic metric, yet relies on a readily available encoder in order to estimate latent entropy (Chen et al., 2018b). Many state-of-the-art GANs do not have an encoder readily available, and this has even been cited as a barrier to use (Karras et al., 2019). Finally, the perceptual path length was proposed to measure disentanglement without relying on an external model, but the method is specific to face datasets such as CelebA, as it crops out the background prior to evaluation (Karras et al., 2019). To address these limitations in a metric's applicability and scope, we propose a method that focuses on only using the generative model's decoder $g : \mathbb{Z} \to \mathbb{X}$ and can be applied across datasets. Additionally, because the utility of disentanglement is often with respect specific subsets of factors that are human-interpretable, we include a supervised variant of our metric that compares the real data manifold with the generated one. Finally, there is a difference between evaluating disentanglement and learning a disentangled representation, the latter of which requires constructing a valid loss function for learning and guaranteeing disentanglement, a process that requires at least weak supervision (F. Locatello et al., 2019).

The factors of a generating dataset such as dSprites, and their values, are provided with the dataset. Features in the dSprites dataset include shape, orientation, x-position, and y-position. An example image (a data point) is a heart rotated 90 degrees in the top right corner. The values of each feature (factor) are provided in this generating dataset and, in this case, discretized. In generating a submanifold, we would hold a factor, such as orientation, constant, while varying the others (sampling different values for the others) to create a subsample that we then use in the RLT procedure. In a non-toy dataset, such as CelebA, where the factors and their values are not known to high accuracy, we can only estimate a possible subset. In this case, we follow prior precedent and use the binary attributes provided in the dataset, such as wearing sunglasses or black hair color. This type of selection of factors and values are common in disentanglement literature; we do not introduce a novel protocol with the factor selection here.

For a generated manifold, we do not know the factors corresponding to the latent dimensions upfront. As a result, we hold latent dimensions constant and randomly sample values from the latent prior (spherical normal) within a dimension to hold constant while varying the values of others through random sampling. Each set of latent dimension values correspond to a point in the data manifold, which we use the corresponding generative model to generate. These points on the generative model's data manifold are then embedded using an ImageNet-pretrained VGG16 network as a feature extractor (these details are currently in Appendix G). These embeddings produce point clouds from which the persistence barcodes are computed and vectorized using the RLT procedure.

Geometry of deep generative models. Prior work has explored applying Riemannian geometry to deep generative models (Shao et al., 2018; Chen et al., 2018a; Rieck et al., 2018; Horak et al., 2020). One work approximates the geodesics of the latent manifold to visually inspect deep generative models as an alternative to linear interpolation (Chen et al., 2018a). Another work also explores computing geodesics efficiently and shows that style between interpolations can be transferred with the approach (Shao et al., 2018). Horak et al. (2020) use persistent homology for comparing GAN evaluation metrics FID, KID, IS, and the geometry score from (Khrulkov and Oseledets, 2018). One of the closest work to ours has explored the geometry, specifically the normalized margin and tangent space alignment, of latent spaces in disentangling VAE models (Shukla et al., 2018). This

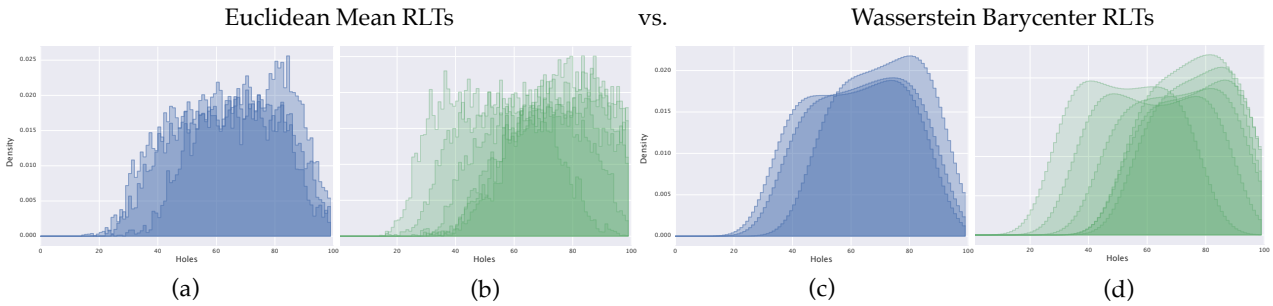

Figure 8: Comparison of each factor's topological signatures in *dSprites* derived from the Euclidean mean on the left (a) and (b) (Khrulkov and Oseledets, 2018) to our method of deriving them from taking the Wasserstein barycenters on the right (c) and (d). Notice visually that the Euclidean mean collapses geometric information, rendering the two topological signatures (a) and (b) more indistinct, while the Wasserstein barycenters exhibit much smoother distributions and construct distinct topological signatures for each factor, (c) and (d).

work is interesting in that it leverages the lower dimensionality of latent spaces to enable more computationally feasible calculations, such as singular value decomposition. However, they do not propose an disentanglement evaluation method and do not explore the learned data manifold or homology. Another recent study that relates closely to our work examines holonomy on disentangling data manifolds in synthetic manifolds and 3D objects Pfau et al. (2020). While they explore exciting differential geometry methods, our work examines the topology, specifically homology, a topological invariant, (and not to be confused with holonomy) of data manifolds, offer an evaluation metric (with unsupervised and supervised variants) based on this examination, and finally observe this property in practice on both toy and realistic datasets.

Persistent homology: barcodes and Wasserstein distance. Carlsson (2019) presents a survey of persistent homology and its applied uses. Specifically, there are multiple methods for vectorizing persistence barcodes, including persistence landscapes, persistence images, symmetric polynomials (Carlsson, 2019; Bubenik, 2015; Adcock et al., 2013). Additionally, Wasserstein distance defines a metric on barcode space, as detailed by Carlsson (2019):

$$W_p(B_1, B_2) = \inf_{\theta \in \mathcal{D}(B_1, B_2)} \Big(\sum_{I \in B_1'} \pi(I, \theta(I))^p \Big)^{\frac{1}{p}}$$

where $p > 0$ or $p = \infty$, B_1 and B_2 are two barcodes, $\mathcal{D}(B_1, B_2)$ denotes the set of all bijections $\theta : B_1 \to B_2$ for which $\pi(I, \theta(I)) \neq 0$ for only infinitely many $I \in B_1'$, π refers to the penalty function between barcodes. Thus, we use Wasserstein distance where $p = 2$, which underlies Wasserstein barycenters, on our barcodes, over prior work using Euclidean distance and Euclidean means (Khrulkov and Oseledets, 2018).

Geometry score implementation of persistent homology. The method for computing the relative living times (RLTs) originates from the Geometry Score paper (Khulkov et al. 2018) and we include the requested details on their method in the Appendix. Specifically, they use the Gudhi library and compute persistence intervals in a dimension by constructing witness complexes. The witness complex is computed for all filtration values at once to compute a persistence diagram, which summarizes the homology for all ϵ. Topological features are usually not computationally tractable in high dimensions, so we do not use high-dimensional features there — specifically, we are only using k=1 dimension, per the Geometry Score implementation.

APPENDIX C: WASSERSTEIN MEAN RLTS VS. EUCLIDEAN MEAN RLTS

We show visual comparisons of our method (Wasserstein Relative Living Times) against the prior method using the Euclidean mean to obtain the average distribution across Relative Living Times (Khrulkov and Oseledets, 2018) in Figure 8.

We empirically evaluate the use of Euclidean distance compared to W. distance and Euclidean mean compared to W. mean on the real *dSprites* dataset in Table 2. The table also indicates that the Wasserstein distance between Wasserstein barycenters is the most capable of differentiating similar and dissimilar topologies.

RLT Distance Metric	Wasserstein RLTs	Wasserstein Distance	Difference Ratio
Geometry Score	–	–	1.60x
(W. RLT)	✓	–	1.75x
(W. Distance)	–	✓	2.14x
Ours	✓	✓	**2.93x**

Table 2: **Wasserstein distance: empirical study and ablation**. We evaluate the ratio between the mean distance between homeomorphic RLTs and non-homeomorphic RLTs with an ablation of our proposed use of W. distance and W. RLTs on real images in *dSprites*, compared to the *Geometry Score* evaluation metric proposed in Khrulkov and Oseledets (2018). A higher ratio on known factors of variation indicates a distance metric on RLTs that is better at identifying similar topologies and distinguishing different topologies.

APPENDIX D: ADDITIONAL HOMEOMORPHIC AND NON-HOMEOMORPHIC WASSERSTEIN RELATIVE LIVING TIMES

An extended figure, from Figure 4 in the main text is shown here in Figure 9, illustrating that likely homeomorphic clusters of submanifolds conditioned on factors based on their W. RLTs look visually similar.

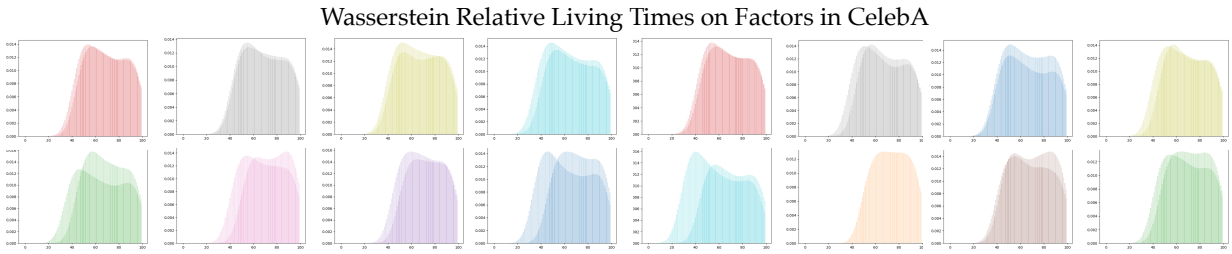

Figure 9: Additional Wasserstein RLTs from several more factors in the *CelebA* dataset. As before, factors whose conditional submanifolds are homeomorphic to each other are shown on top, and factors which are not homeomorphic to each other are shown below.

APPENDIX E: TOPOLOGICAL SIGNATURES OF DSPRITES

We show topological signatures for each factor from the real *dSprites* dataset in Figure 10. In the supervised variant, we discover that similar topological signatures in the generated manifold match the ones in the reals for corresponding latent interpretations that semantically adapt these factors.

APPENDIX F: TOPOLOGICAL SIMILARITY MATRIX OF *dSprites*

In Figure 11, we display the topological similarity matrix for the *dSprites* dataset, showing every value of each factor. A visible diagonal can be seen for each factor. Observe that the first three squares in the top left corner correspond to shape, the next six to scale, the next forty to orientation, the next thirty-six to x-position, and finally the last thirty-six to y-position. Note that this grid uses our method (W. distance) and is not spectrally coclustered.

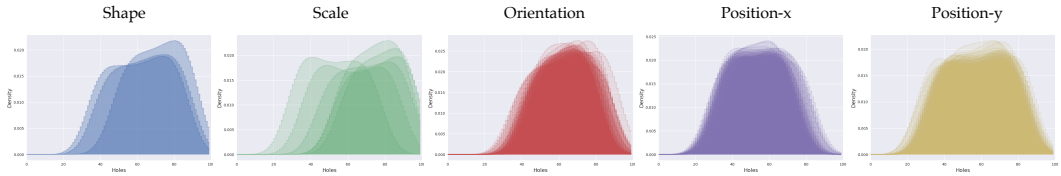

Figure 10: Topological signatures of each *dSprites* factor. Each graph illustrates an overlay of different topological signatures (W. RLTs), produced when holding a given value of that factor constant while varying others. For example, in the first graph, the three curves indicate three shapes: ellipse, heart, and square. Notably, these sets of topological signatures present distinguishing features in aggregate.

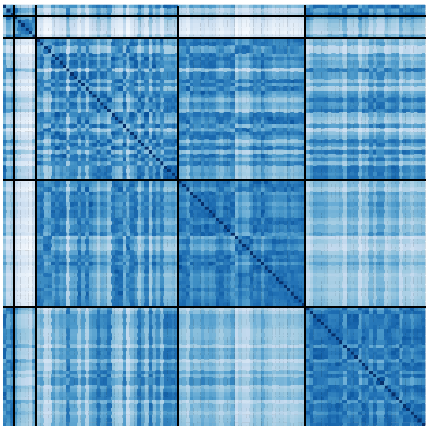

Figure 11: Topological similarity matrix on *dSprites* (reals), across values of every factor.

APPENDIX G: HYPERPARAMETERS AND EXPERIMENT DETAILS

For all models, we used open-sourced PyTorch implementations and model checkpoints that implement prior work using default hyperparameters. We use pretrained model checkpoints and do not tune them further. For VAE variants, these from the disentangling-vae library [1] to examine loss differences with network parity. One exception, in which we use TensorFlow instead of PyTorch, is the InfoGAN variants, where we could not reproduce results from any open-sourced PyTorch implementations, a known issue for InfoGAN (Higgins et al., 2017; Kim and Mnih, 2018). For these, we trained to the default number of epochs and hyperparameters based on the papers, because pretrained checkpoints were not available on these tasks. Additionally, we use default hyperparameters and functions for spectral co-clustering (scikit-learn, Pedregosa et al. (2011)) and Geometry Score implementations. The Geometry Score implementation used a gamma of $\frac{1}{128}$ and an n of 100. Following the suggestion in Khrulkov and Oseledets (2018), we also used a pretrained VGG16 (Simonyan and Zisserman, 2015) with the last 3 layers removed as a feature extractor to embed high-dimensional images into 64 feature dimensions. All of these hyperparameters were constant across all datasets, models, and experiments.

In addition, based on our preliminary experiments with MIG-sup, as indicated in the VAE results in Li et al. (2020), MIG and MIG-sup are highly correlated under VAE architectures. Over ten runs of VAE, β-VAE, and β-TCVAE, represented in Figure 12, we found that the two metrics were strongly correlated ($R^2 = 0.952$). We note that these results are meant to explore MIG-sup and demonstrate preliminary results in a contained study, for which further investigation should be pursued. As the metric is similar to MIG, it does not correspond particularly closely to either alternative supervised metric in Figure 7.

[1] https://github.com/YannDubs/disentangling-vae

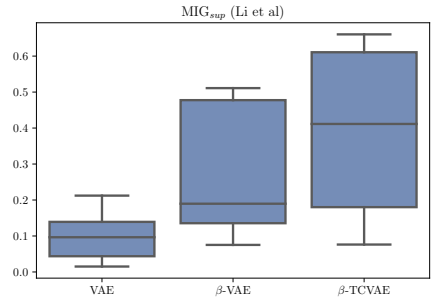

Figure 12: Supplementary comparison of MIG$_{sup}$ evaluations across various three models.

APPENDIX H: COMPUTATIONAL COMPLEXITY

Let n be the number of RLTs per latent dimension, L_0 be the number of RLT landmarks, N be the number of images sampled per latent dimension, D_z be the number of latent dimensions, D_s be the number of factors of variation, and B be the number of bins for the probability distribution histogram:

(a) Calculating the nD_z RLTs per D_z latent dimensions is O($nD_zD_sNL_0$) (Khrulkov and Oseledets, 2018).

(b) Calculating D_z W. barycenters is O(nD_zB^2m), with Maxiter m from Dognin et al. (2019).

(c) Calculating W. distances between all D_z barycenters is O($D_z^2B^2/\varepsilon^2$) by the Sinkhorn algorithm with tolerance ε (T. Lin et al., 2019).

(d) Spectral coclustering can be computed in O(D_z^2) (M. Vlachos et al., 2014), and we optimize over the number of biclusters, of which there are at most D_z, so this is O(D_z^3). Calculating the bicluster scores has the same runtime.

Treating ε and m as constants, and noting $B \leq L_0 \leq N$, then this is O($nD_zD_sNL_0 + D_z^2B^2 + D_z^3$). Note that many of these subprocedures can be substantially parallelized.

APPENDIX I: ALGORITHM

The following are additional algorithms of this paper:

Algorithm 2: Procedure for producing Wasserstein RLTs on real images: Given a dataset $\mathcal{D}$ with D_s factors of variation s and n_d possible (or sampled) values per factor $d \in D_s$, returns a W. RLT for each dimension. RLT is the RLT procedure from Khrulkov and Oseledets (2018).

for latent dimension d in $1 : D_s$ **do**
  for k in $1 : n_d$ **do**
    Set $x, s \leftarrow sample(\{x, s \in \mathcal{D} | s_d = k\})$
    Compute $e_z \leftarrow embedding(x)$ {e.g. using VGG16}
    Compute $rlt[d, k] \leftarrow RLT(e_z, \gamma = 1/128, L_0 = 64, n = 100)$
  end for
  Compute $WB[d] \leftarrow W.Barycenter(rlt[d, k])$
end for
return WB

Algorithm 3: Procedure for calculating μ: Given W. RLTs WB_1, WB_2, with D_1, D_2 dimensions respectively, finds a similarity matrix M with the μ-maximizing c clusters and returns M, c, μ.

for d_1 in $1 : D_1$ **do**
  for d_2 in $1 : D_2$ **do**
    Compute $M[d_1, d_2] \leftarrow W.Distance(WB_1[d_1], WB_2[d_2])$
  end for
end for
for c in $1 : D_2$ **do**
  Apply spectral coclustering to matrix M using c biclusters
  Compute μ based on in-group and out-group similarities
end for
Select c which minimizes total variance **if** unsupervised **else** real c
return M, c, μ

APPENDIX J: ETHICAL CONSIDERATIONS

This research can aid in alleviating bias in deep generative models and more generally, unsupervised learning. Disentanglement has been shown to help with potentially reducing bias or identifying sources of bias in the underlying data by observing the factors of variation. Those who will benefit from this research will be users of generative models, who wish to disentangle or evaluate the disentanglement of particular models for downstream use. This may include artists or photo editors who use generative models for image editing. For negative consequences, this research broadly advances research in deep generative models, which have been shown to have societal consequences when applied maliciously, e.g. mimicking a political figure in DeepFakes.