THE SHAPE OF FLOWS: TOPOLOGICAL REGULARIZATION OF VECTOR FIELDS IN FLOW MATCHING FOR BETTER DIVERSITY AND FIDELITY.

Anonymous authors
Paper under double-blind review

ABSTRACT

flow matching models have become the de facto paradigm for the generation of continuous data. Despite impressive results, conditional sampling from these models implies a fidelity/diversity tradeoff, which greatly impacts their applicability. To address this issue, we propose the use of persistent homology to identify relevant topological features both at the pixel level and at the sample level, from which we derive two training objectives. On the one hand, we use filtrations on cubical complexes to align the topological persistence of pixel velocity sublevel sets with that of the ground truths; on the other hand, we employ filtrations on Vietoris-Rips complexes to ensure the topology of the generated distribution approximates that of the training set. We demonstrate these approaches improve both the diversity and fidelity of generations, showing the promise of persistent homology and geometric invariants to improve the quality of flow matching models.

1 INTRODUCTION

Topological Data Analysis (TDA) and its symbiosis with deep learning has a long and very successful history, from differentiable topological layers that allow for the optimization of shape-aware neural networks (Hofer et al., 2017; Carrière et al., 2020) to the more recent emergence of Topological Deep Learning, which extends traditional message-passing (originally limited to nodes and edges in graph neural networks) to higher-order structures like simplicial and cell complexes, marking a "new frontier" in relational learning (Hajij et al., 2022; Papamarkou et al., 2024).

One of the most powerful techniques in TDA is Persistent Homology (PH), which transforms complex, high-dimensional data into a simplified 'topological signature", that describes both its global topology and local geometry. It has been successfully applied to molecular design and protein folding (Qiu & Wei, 2023), image segmentation (Clough et al., 2020), signal processing (Xu et al., 2021) and point clouds via learned filtrations (Nishikawa et al., 2023).

PH can be particularly beneficial in domains that encode geometrical information, such as protein folding and image/video generation. State-of-the-art models for continuous data generation are predominantly diffusion or flow matching models (Esser et al., 2024; Black Forest Labs, 2026; OpenAI, 2024; Chefer et al., 2026; Wan et al., 2025; HaCohen et al., 2026; Seedance et al., 2026), which work by learning probability paths that move samples from a source distribution p_0 to generate data samples from a target data distribution p_1. diffusion models typically parametrize the score field (or the noise to be removed) along a probability path, whereas flow matching models parametrize the velocity field that transports samples directly along that path. But in both cases, models are trained via pixel level Bregman Divergences such as Mean Squared Error, which don't necessarily capture global topological features.

While current generative models achieve impressive results, they often sacrifice diversity for quality, specially when generating conditionally on an input. This is due to the fact that conditional generation in these models generally relies on Classifier Free Guidance (Ho & Salimans, 2022), which simultaneously trains the model with and without guidance. In the conditioned case, guidance information is usually added via trainable cross-attention layers. There is, however, ample evidence that this method implies a quality/diversity tradeoff and can lead to mode collapse (Chung et al., 2024; Li et al., 2025; Ifriqi et al., 2025; Pavasovic et al., 2025), an issue that becomes more acute

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for higher guidance scales (Patel et al., 2023). This ultimately means the data distribution sampled from the model p_1^θ is far from the actual data distribution p_1, which hampers the applicability of these models.

Similar to how language models are refined from a pretrained base via preference alignment and other post-training methods, a growing body of work attempts the same with diffusion and flow matching models, using regularization objectives (Wang & He, 2025; Chandran et al., 2026) or, more commonly, RL (Black et al., 2024; Liu et al., 2026) to push generations towards human preference, aesthetic quality or higher diversity when sampling conditionally.

We follow this trend, asking the following question: can we use the topological information present in high dimensional continuous data (such as images or videos) to improve simultaneously the diversity and fidelity of the generations of flow matching and diffusion models? To the best of our knowledge, this is the first work to apply differentiable persistent homology features to post-train flow matching models. Our contributions can be summarized as follow:

1. We adapt a cubical complex filtration based loss to improve the fidelity of each predicted velocity field.
2. We introduce a completely novel loss based on filtrations on Vietoris-Rips complexes to improve the diversity of predicted batches.
3. We show topological-finetuning for a few steps measurably improves the diversity and fidelity of our models.

This paper is divided as follows. Section 3 contains a brief summary of the persistent homology and flow matching theory. Section 4 describes two approaches to improve generations via persistent homology: either force the topology of predicted velocity fields to match that of ground truths while ignoring the noise, improving fidelity, or consider a training batch as a point cloud, forcing the overall topology of all predicted samples to match that of the ground truth batch, retaining the diversity present in the training dataset. Finally, experimental setup and results are presented in Section 5.

2 Related Work

Post-training and regularization of diffusion models. Numerous works in recent years deal with improving specific aspects of generative models after pretraining. The most popular approach is RL-based post-training, where denoising is understood as a multi-step Markov decision process, which enables policy-gradient optimization against a reward. Methods span early RL fine-tuning techniques such as DPO,DDPO and DPOK (Wallace et al., 2024; Black et al., 2024; Fan et al., 2023), and, most recently, GRPO-style group-relative methods (Liu et al., 2026; Xue et al., 2025). However, optimizing with respect to a scalar reward tends to concentrate the sampling distribution around a narrow set of high-reward outputs. Some solutions exist to this issue, including GFlowNet-based fine-tuning, which trains the sampler so that its terminal distribution matches a target distribution proportional to reward (Liu et al., 2025). A different approach, more in line with what we intend in this work, attempts to regularize the model's internal representations or spectral structure at training time via differentiable objectives, improving fidelity and diversity without an explicit reward (Wang & He, 2025; Chandran et al., 2026).

Topological objectives for diffusion models. The most relevant work is *TopoDiffusionNet* (Gupta et al., 2025), which uses persistent homology to force the model to generate images with a desired $H0$ or $H1$ topology (for example, *a line dividing the plane in three distinct regions*). They achieve this by adding a topological term to the regular diffusion loss, which represents the distance between the persistence diagrams of the cubical complexes of the predicted image and target. This builds on previous work using cubical complex filtrations to move predictions towards a topological prior (Clough et al., 2020; Hu et al., 2019). The use of Vietoris-Rips filtrations also appears in the literature. Moor et al. (2020); Hofer et al. (2020) define VR-based objectives to match the topologies of the data and autoencoder latent space, and to ensure beta-connectedness respectively. Another interesting, although orthogonal, idea appears in Hu et al. (2024): adding the topological information directly to the network via a trained topology encoder.

3 BACKGROUND

3.1 PERSISTENT HOMOLOGY

Let $X \subset \mathbb{R}^d$ be a point-cloud. How could we recover the underlying "shape" of the data? The idea behind persistent homology is to substitute each point $x \in X$ with a closed ball $B(x, r)$ of radius r. If two balls intersect, then we draw an edge (i.e.: a 1-dim simplex) between them. If three balls intersect, we draw a triangle (2-dim simplex) etc. The result is a *simplicial complex*, which evolves as we gradually increase r. This results in a sequence of simplicial complexes that approximates the shape of the data at different resolutions, known as *Čech filtration*, see Figure 1 (left).

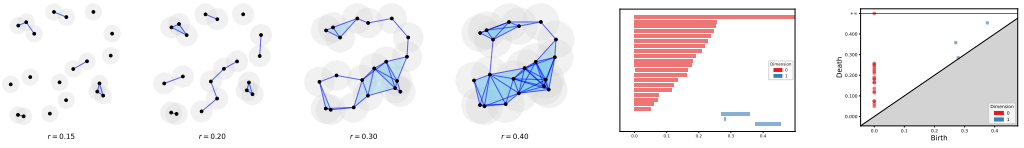


Figure 1: Čech complex for various values of r (**left**). Persistence barcode and diagram encoding the topology of a point-cloud X (**right**).

For each value of r, we compute the homology groups of the simplicial complex to find connected components (H_0), loops (H_1), and voids (H_2). This is equivalent to computing the homology groups of the union of balls, since it can be shown to be homotopy equivalent to the *Čech complex* (Nerve Theorem). However, as computing *Čech filtrations* is computationally costly, in the following section we will work with the *Vietoris-Rips filtration*, on which we add a k-simplex between points $x_1, ..., x_k$ for radius r if and only if the distance between all pairs is smaller or equal than r. This filtration can be considered an approximation of the *Čech filtration* up to a scale factor, and can be easily computed from a distance matrix. We use the `ripser` (Bauer, 2021) library to that end.

Persistent homology tracks the 'life cycle' of components, loops and voids as we increase r. A feature is said to be born at radius r_{birth} when a new hole or component is formed. A feature is said to die at radius r_{death} when it is filled in or merged with another component. The persistence of a feature is then defined as its lifespan: $r_{\text{death}} - r_{\text{birth}}$. If we draw these pairs on $\mathbb{R}$ and $\mathbb{R}^2$, we obtain respectively the persistence barcode and diagram of X, see Figure 1 (right). Topological features with short lifespans (close to the diagonal) can in most cases be discarded as noise, whereas larger lifespans point to descriptive features of our data.

Given two persistence diagrams obtained from point-clouds X_1 and X_2, we can compare them using a Wasserstein or Sinkhorn distance (Entropy-regularized Optimal Transport, (Cuturi, 2013)), with a slight modification. In standard Optimal Transport, we compute the cost of moving mass from one distribution to another. However, points in a persistence diagram can disappear into the diagonal, so we must account for this by allowing matchings of points with their projection onto the diagonal. More details on the exact formulation of these metrics in Appendix A.

While computing PH on Vietoris-Rips filtrations is powerful, for certain types of high dimensional data, such as images, it is preferable to use so-called *filtrations of Cubical Complexes*.

A cubical complex is a collection of hypercubes. Given a matrix of shape $m \times n$ representing grayscale pixel intensities, we treat each pixel as a 0-dimensional vertex (or a 2-dimensional square cell) in a grid. To build the filtration, we define a threshold value r and consider the sublevel set C_r consisting of all cells whose associated values (pixel intensities) are less than or equal to r. As r increases, new vertices, edges, and squares are added to the complex based on the pixel intensities. Higher-dimensional cells are assigned the maximum value of its constituent lower-dimensional faces. This sequence of nested complexes, $C_{r_1} \subseteq C_{r_2} \subseteq ...$, allows us to track the birth and death of topological features, such as connected components and holes, directly on the image grid. We will also employ filtrations on cubical complexes on Section 4.

3.2 FLOW MATCHING

Let $\psi_t : \mathbb{R}^d \to \mathbb{R}^d$, $\psi_t \in C^r(\mathbb{R}^d, \mathbb{R}^d)$ be a C^r diffeomorphism, i.e.: a continuously (r times) differentiable and invertible map such that $\psi_t^{-1} \in C^r(\mathbb{R}^d, \mathbb{R}^d)$.

Now let $\psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ such that $\psi(t,x) \mapsto \psi_t(x)$. The function ψ is called a C^r *flow*. Given a random variable $X_0 \sim p_0$, a *flow model* is a continuous time Markov process $(X_t)_{0 \leq t \leq 1}$ such that $X_t = \psi_t(X_0),\ X_0 \sim p_0$. We can also express the above Markov process via an ODE:

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)),\ \ \psi_0(x) = x,$$

where u is a $C^r([0,1] \times \mathbb{R}^d, \mathbb{R}^d)$ velocity field $u : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, $u(t,x) \mapsto u_t(x)$. Existence and uniqueness of solutions to this ODE can be shown up to time $t = 1$ (almost everywhere) as long as we assume global Lipschitzness or integrability of u (see (Lipman et al., 2024)). Such a flow defines a time dependent probability path $(p_t)_{0 \leq t \leq 1}$, $X_t = \psi_t(X_0) \sim p_t$ for $X_0 \sim p_0$. Conversely, given a probability path p_t, we say u_t generates p_t if $X_t = \psi_t(X_0) \sim p_t$.

Under this setting, training a *flow matching model* amounts to

1. building a probability path $(p_t)_{0 \leq t \leq 1}$ generated by a flow u_t such that p_1 aligns with our target distribution and,
2. finding a parametrization u_t^θ of u_t such that it minimizes the loss

$$L_{FM}(\theta) = \mathbb{E}_{t \sim U(0,1), X_t \sim p_t}\left[\|u_t(X_t) - u_t^\theta(X_t)\|^2\right],$$

where instead of the L2-norm we could use any Bregman divergence, see (Lipman et al., 2024; Holderrieth et al., 2024). The issue here is that u_t is not tractable. Luckily, it suffices to optimize with respect to the conditional flow. Given a r.v. $Z \sim p_1$, the conditional loss is defined as,

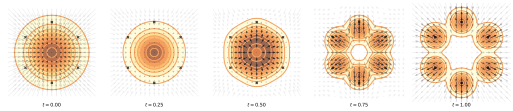
$$L_{CFM}(\theta) = \mathbb{E}_{t, Z X_t \sim p_{t|Z}(\cdot|Z)}\left[\|u_t(X_t \mid Z) - u_t^\theta(X_t)\|^2\right],$$

where $u_t(x \mid Z = z)$ is a conditional vector field generating the conditional probability path $p_{t|1}(\cdot \mid z)$. It can be proven that under mild assumptions, and if $u_t(x \mid z)$ is conditionally integrable and generates $p_{t|1}(\cdot \mid z)$, then the marginal velocity field u_t generates the marginal probability path p_t. Further, optimizing θ for $\mathcal{L}_{CFM}$ is equivalent to optimizing for $\mathcal{L}_{FM}$, given that the gradients of both losses coincide (refer to (Lipman et al., 2024; Holderrieth et al., 2024) for proofs of both facts).

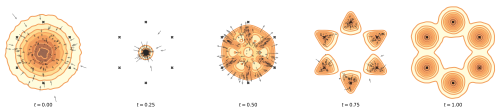
4 TOPOLOGICAL OBJECTIVES

Generative quality in flow matching and diffusion is usually measured via Fréchet Inception Distance (FID, (Heusel et al., 2017)), which encodes real and generated samples using an Inception-v3 network, and fits a multivariate gaussian to each set to compare both distributions. Intrinsically, FID measures both diversity (distance between variances) and fidelity (distance between means). However, ample research points to the limitations of FID optimization (see (Jayasumana et al., 2024; Stein et al., 2023)), namely: poor feature representations and lack of correlation with human perceived realism. More fundamentally, these limitations reflect the broader challenge of defining what it means for a generative model to faithfully approximate the underlying target distribution. While the answer is likely multifaceted, we propose a new direction that, to the best of our knowledge, remains unexplored in the literature: approximating the target data manifold using the topological signature of the target vector field (or x-prediction) via persistent homology. The key idea can be seen in Figure 2.

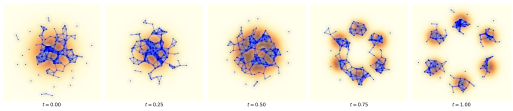
On what follows we show that (1) it is possible to define topological objectives that approximate the topology of the marginal vector field via the persistent homology of the conditional vector fields and (2) optimizing these objective improves the diversity and fidelity of generative models after a few steps of finetuning as measured by FID, precision, recall, and the coverage and density metrics proposed by (Naeem et al., 2020). We will define the objectives in velocity space, but notice the same applies on image space (for x-prediction models).



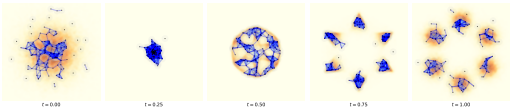
(a) Position space: $p_t(x)$ and velocity $u_t(x)$



(b) Velocity space: density of $u_t(x_t)$ and acceleration



(c) Vietoris-Rips on $x_t \sim p_t$ ($n = 128$, $r = 0.45$)



(d) VR on $u_t(x_t)$, velocity space ($n = 128$, $r = 0.45$)

Figure 2: The idea behind the proposed topological objectives is approximating the target data manifold via its topological signature. This works both on image space (x/ϵ-prediction models) and on velocity space (u-prediction models). Figure (a) represents the push-forward measure on a distribution under CondOT paths, together with the marginal vector field. Figure (c) approximates the target densities p_t as the simplicial complex of a randomly sampled point-cloud (in practice we employ the full filtration, but for simplicity here we draw only $r = 0.45$). In velocity space (b) we show the push-forward of p_t under the map $x \mapsto u_t(x)$, where $u_t = \mathbb{E}[x_1 - x_0 \mid x_t = x]$ is the marginal vector field, estimated by a smoothed histogram of 150000 samples $u_t(x_t)$, $x_t \sim p_t$. Figure (d) approximates this distribution from a randomly sampled point-cloud.

4.1 Targeting 'Fidelity': pixel-level cubical complex objective

Let $u_t^\theta(X_t)$ be the output from a flow matching model, and let $u_t(X_t \mid Z)$ for $Z \sim p_1 = p_{data}$ be the target velocity field on X_t given target image Z. As explained in the previous section, training a flow matching model θ amounts to differentiating through the distance between $u_t^\theta(X_t)$ and $u_t(X_t \mid Z)$. Assume θ is a minimizer (although imperfect) of the flow matching loss.

First, we focus on the topological features at the pixel level. Consider outputs $u_t^\theta(X_t)[i] \in \mathbb{R}^{c \times h \times w}$, where we denote by $u_t^\theta(X_t)[i]$ the i-th sample in a batch of predictions. For each individual channel, we build a filtration on the cubical complex on the $H \times W$ grid, which captures topological features of the pixel intensities: as we move the intensity threshold, the velocity sublevel sets grow, adding more faces to the hypercubes of the complex. This produces a persistence diagram that contains holes H_0 and loops H_1 in the velocity space, and which we denote $PD(u_t^\theta(X_t)[i, c])$. We do the same for the ground truth velocities to obtain $PD(u_t(X_t \mid Z)[i, c])$.

Now we identify which topological features in $PD(u_t^\theta(X_t)[i, c])$ correspond to a feature in $PD(u_t(X_t \mid Z)[i, c])$. This is done via a Wasserstein/Earth mover's distance (see Appendix A), which finds the lowest cost matching between points in $PD(u_t^\theta(X_t)[i, c])$ and $PD(u_t(X_t \mid Z)[i, c])$ while allowing points to be matched to the diagonal. The result is what we define as the set of *critical point matchings*, where we call a pair of pixels $(a_1, b_1), (a_2, b_2)$ a pair of critical points of PD, if as r grows in the filtration an edge appearing between them causes the birth or death of a topological feature in the PD. Thus the set of critical point matchings is:

$$Cr = Cr(u_t^\theta(X_t)[i, c], u_t(X_t \mid Z)[i, c]) = \{(x, y, v, w) | (x, y) \text{ is a critical point of } u_t^\theta(X_t), \\ (v, w) \text{ is a critical point of } u_t(X_t \mid Z), \text{Wasserstein maps } (x, y) \to (v, w)\}.$$

That is, we match topological features, and then select the corresponding pixels where these features were born, and where they died. With this we can define the fidelity loss as:

$$\mathcal{L}_{fid}(\theta) = \mathbb{E}_{t, Z, X_t}\left[\sum_{i,c} \sum_{(x,y,v,w) \in Cr} \text{dist}(u_t^\theta(X_t)[i, c, x, y], u_t(X_t \mid Z)[i, c, v, w])\right].$$

This is a common approach in TDA for grid data (see, for example, (Gupta et al., 2025)), and can be directly applied as is on pixel-space diffusion and flow matching models.

Perhaps the fact that (x, y) does not need to be equal to (v, w) might seem unintuitive, but it gains meaning when examined from the light of cubical complexes and pixel sublevel sets: consider a topological feature of the target, for example a loop that corresponds to the eye contour on a face.

All pixels along it have similar brightness. As we increase the brightness threshold, the feature is born when we add all contour pixels. Now consider the prediction, which is more noisy. The eye contour appears at a higher threshold, because some pixels closing it are too bright. So we push the brightness of these pixels down, thus making the eye contour in the prediction more clearly defined.

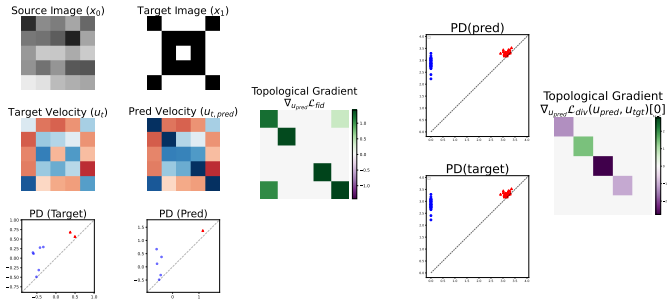


Figure 3: Proof of concept for fidelity and diversity losses to illustrate how gradients flow from critical points in the PDs to pixels. (**Left**) The 'predicted velocity" here is just a noisy modification of the target velocity, where we add gaussian noise to the diagonal to showcase how gradients $\nabla\mathcal{L}_{fid}(u_{pred}, u_{tgt})$ flow from critical points to the error in the diagonal (right-side). Below, the persistence diagrams $PD(u_{pred})$ and $PD(u_{tgt})$ on the cubical complexes. (**Right**) Similar set up with a batch of 32 images. We show $\nabla_{u_{pred}}\mathcal{L}_{div}(u_{pred}, u_{tgt})[0]$, together with the persistence diagrams of the filtrations on the Vietoris-Rips complex of the batch.

This objective is differentiable (see a simple example of how gradients flow through critical points on Figure 3), and can be interpreted as a version of MSE where we don't sample pixels uniformly, but from a distribution skewed towards topologically relevant pixels. In practice, to compute critical points we must compute a cubical complex filtration per index in the batch and per channel. Due to the computational overhead, when working with Representation Autoencoders (RAEs, (Zheng et al., 2025)), we further select specific channels to compute the loss over, making gradients zero on the rest. More information on Appendix E.

4.2 Targeting 'diversity': distribution-level Vietoris-Rips objective

Let $Z \in \mathbb{R}^{bs\times c\times h\times w}$ be a batch of samples from the target data distribution p_1. We know that if we sample $X_t \sim p_t(\cdot \mid Z)$ and transport it following the ODE $d/dtX_t = u_t(X_t)$ we can recover the data distribution p_1. We aim at training the flow u_t^θ so that we can recover the target distribution following trajectories consistent with $d/dtX_t = u_t^\theta(X_t)$. The idea now is to consider the whole batch as a point-cloud that approximates the probability distribution at time t, $X_t[0], ..., X_t[bs] \sim p_t(\cdot \mid Z)$, or in other words, a point-cloud that contains some information about the data manifold across the probability path. See Figure 2.

Consider the derivative at each point, given by $u_t(X_t \mid Z)[0], ..., u_t(X_t \mid Z)[bs]$, which contains information about the direction in which we have to transport each pixel to approximate the target distribution. As bs grows, this can be considered an approximation of the flow $u_t(x \mid Z)$, $x \sim p_t(\cdot \mid Z)$, that transports the distribution, $p_t(\cdot \mid Z)$ towards p_1; in other words, after marginalization over Z, the push-forward of p_t under the map $x \mapsto u_t(x)$. It is therefore possible to re-define the problem of optimizing u_t^θ as matching the topologies of $u_t^\theta(X_t)[0], ..., u_t^\theta(X_t)[bs]$ and $u_t(X_t \mid Z)[0], ..., u_t(X_t \mid Z)[bs]$.

Let $X_t \sim p_t(\cdot \mid Z)$. We can now define the Vietoris-Rips ('diversity') loss as

$$\mathcal{L}_{\text{div}}(\theta) = \mathbb{E}_{t,Z,X_t}\Bigg[\sum_{(i,j,k)\in C(X_t|Z)} \lambda_t \underbrace{\left|\text{dist}\big(u_t^\theta(X_t)[i], u_t^\theta(X_t)[j]\big) - birth/death(u_t(X_t \mid Z)[k])\right|}_{\text{topological accuracy}}$$
$$\cdot \underbrace{\text{dist}\big(u_t^\theta(X_t)[i] - u_t^\theta(X_t)[j], u_t(X_t \mid Z)[i] - u_t(X_t \mid Z)[j]\big)}_{\text{geometric alignment}}\Bigg].$$

For a constant λ_t which can also be made time dependent (although in practice we leave this parameter constant, as a simple scale control for the loss). $C(X_t \mid Z) = C(PD(u_t^\theta(X)), PD(u_t(X \mid Z))$ is the set of critical point matchings between $PD(u_t^\theta(X_t)$ and $PD(u_t(X_t \mid Z))$, which we define as the set of tuples of indices i, j, k, ℓ such that:

1. At radius r, the connection of samples i and j via an edge produced the *birth* or *death* of feature $f \in PD(u_t^\theta(X))$,
2. Wasserstein distance maps $f \to g \in PD(u_t(X \mid Z))$,
3. k, ℓ are the indices of the corresponding sample in $u_t(X \mid Z)$ responsible for the *birth* resp. *death* of g. Refer to Appendix B.1 for more details.

Finally, we write birth/death$(PD(u_t(X_t \mid Z))[k])$ as a shorthand for birth/death$(PD(u_t(X_t \mid Z)[k], u_t(X_t \mid Z)[k]))$ which returns the *birth* or *death* time of the corresponding feature on $PD(u_t(X_t \mid Z))$. Notice this is in fact a distance, i.e.: the radius in the filtration at which an edge appeared that caused the death or birth of that feature.

In essence, the topological accuracy term of the loss can be interpreted as: *move $u_t^\theta(X_t \mid Z)[i]$ and $u_t^\theta(X_t \mid Z)[j]$ closer or further apart so that the feature they give birth/death to appears at the right time in the persistence diagram* (that is, the time of birth/death time of the corresponding ground truth feature).

The issue here is that if we backpropagate through the topological accuracy term only, this will move samples i and j across the straight line that connects them, which might not be a meaningful direction in the space of velocities. To deal with this issue, we add the geometric alignment term, which forces them to move across the right direction. We could thus define the topological loss as the *sum* of the topological acc. and geometric alignment components; however, we observed empirically the product to work slightly better in practice. Further discussion on this topic in Appendix **??**. As distance metric we use both euclidean (for SD-VAE latents based models) and cosine distance (for RAE latents based models). More details on the impact of different distance metrics in Appendix B and C.

With the diversity loss defined in this way, we make the following crucial claim: *optimizing the conditional diversity loss leads to approximating the topology of the marginal vector field*. Refer to Appendix B.1 for a full proof.

5 Experiments

5.1 Experiments in pixel space.

Post-training. For our initial proof-of-concept, we utilized a class-conditioned U-Net flow matching model pretrained for 50k steps on the MNIST dataset. We fine-tuned this baseline for an additional 10k steps, comparing standard MSE against our proposed topological objectives (We write Diversity and Fidelity as a shorthand for the Vietoris-Rips filtration based loss and the cubical complex loss respectively). Results can be seen in Table 5.1. We observe that models fine-tuned with the Diversity loss consistently outperform both the MSE baseline and Fidelity-only models.

Pretraining. We also experiment with pretraining 50k steps from scratch on MNIST with different loss combinations to study whether topological objectives could be used beyond post-training. Results can be found in AppendixD. Topological losses don't improve much over MSE. This is to be expected, as in the beginning of training there is not much geometrical information to guide optimization. In fact, persistence diagrams of the Vietoris-Rips complexes are so different that all points get mapped to the diagonal, resulting on 0.0 loss and no gradients. We leave further research in pretraining with topological priors as future work.

5.2 Experiments in RAE space

Finetuning with topological losses in Text-to-Image with RAE latents. For these experiments, we evaluate topological losses in text-to-image synthesis on the MSCOCO dataset. We represent images in a latent space using the RAE approach (Zheng et al., 2025), where representations inherit the semantic topology of a pretrained DINOv2 encoder (Oquab et al., 2024). The RAE preserves a

Table 1: Performance of models finetuned with various combinations of MSE and topological losses. We present results for the class-conditioned flow matching baseline on MNIST (pixel space) and the text-to-image MMDiT on MSCOCO using RAE latents.

Models	**FID** ↓	**Density** ≈ 1	**Coverage** ↑	**Precision** ↑	**Recall** ↑	**ClipScore** ↓
Class-to-Image (MNIST)						
Baseline	4.890	0.735	0.862	0.776	0.833	-
MSE	4.077	0.818	0.888	0.798	0.823	-
MSE + Fidelity	4.257	0.807	0.882	0.803	0.813	-
MSE + Diversity	3.802	0.818	0.896	0.812	**0.831**	-
Fidelity	4.921	0.697	0.848	0.758	0.829	-
Diversity	**3.697**	**0.859**	**0.897**	**0.820**	0.815	-
Text-to-Image with RAE						
Baseline	18.901	**0.983**	0.876	0.777	0.516	27.561
MSE	15.798	1.091	0.915	0.804	0.557	**27.070**
MSE + Fidelity	16.649	1.202	0.923	**0.826**	0.523	27.336
MSE + Diversity	**13.779**	1.084	0.929	0.793	**0.584**	27.305
MSE + Fid + Div	14.703	1.146	**0.931**	0.807	0.566	27.418
Diversity	15.465	1.094	0.919	0.807	0.550	27.238

manifold where euclidean/cosine distances correlate with conceptual similarity, providing a stable foundation for topological optimization. Our architecture is based on a combination of the Multi-Modal diffusion Transformer (MMDiT) (Esser et al., 2024) and the diffusion Transformer with Decoupled Head (DiT-DH) introduced in (Zheng et al., 2025), adapted specifically for text-conditional generation. The base model was pretrained on the CC3M dataset (Sharma et al., 2018) for 240k steps. Its performance on the MSCOCO2014 validation set is presented in the baseline row of Table 5.1. Subsequently, we fine-tuned this baseline for 50k steps using various combinations of MSE, Diversity, and Fidelity losses. All hyperparameters detailed in Appendix D.

The quantitative results are detailed in Table 5.1. Notably, the integration of Diversity loss with the standard MSE objective yields a significant reduction in FID. Regarding distributional metrics, the baseline model already exhibited high Density; however, topological losses further increased this value. As observed in the MNIST experiments, Density values exceeding 1.0 indicate a degree of sample over-concentration, a characteristic effect of enhanced structural regularization (Chung et al., 2024). Our results further confirm a distinct performance trade-off: the Fidelity loss primarily optimizes Precision, reflecting improved sample quality, while the Diversity loss significantly enhances Recall, indicating broader manifold coverage.

Finetuning with topological losses in Class-to-Image with RAE latents. We then take a pretrained state-of-the-art model trained on the ImageNet dataset (Representation Autoencoders, Zheng et al. (2025)) and fine-tune for a few further steps using different objectives. Configuration and hyperparameters detailed in Appendix D. Besides the regular evaluation metrics, we evaluate the topological of the generated data manifold. To do that, we sample 50k images, compute their topological signature (PD) via Vietoris-Rips filtration, and compute the distance to the real (test set) data manifold. More details in Appendix A.4. Results can be seen on Table 2. For comparison we add to recent regularization methods: Spectral regularization (Chandran et al., 2026) and Diffuse and Disperse (Wang & He, 2025). For ablation, we also provide results for training with MSE+Div loss, detaching from the gradients either the geometrical alignment or the topological accuracy terms.

5.3 Experiments in VAE space

Finetuning with topological losses in Text-to-Image with VAE latents. Finally, we evaluated the topological losses within the VAE-based latent space of Stable diffusion 2.1 (Rombach et al., 2022). Fine-tuning for 10k steps with a constant learning rate of 10^{-6} did not yield performance gains, revealing two critical compatibility factors. First, the VAE latent space is regularized toward a Gaussian prior, which can conflate semantic features compared to the unconstrained RAE manifold.

Model	FID ↓	Precision ↑	Recall ↑	Bottleneck ↓		Wasserstein ↓	
				H_0	H_1	H_0	H_1
RAE Baseline	9.452	0.794	0.566	0.066	0.024	17.378	4.105
+ cfg 1.5	8.967	0.850	0.478	**0.046**	0.024	7.986	2.951
50k MSE finetune	9.101	0.727	0.660	0.079	0.033	28.638	5.031
+ cfg 1.5	7.367	0.832	0.546	0.058	0.033	**6.820**	4.014
10k MSE + Div ft	8.598	0.775	0.633	0.058	0.020	13.697	3.677
+ cfg 1.5	**7.203**	0.859	0.559	**0.046**	**0.019**	8.700	**2.418**
10k Div geo detach	9.610	0.663	**0.714**	0.076	**0.019**	20.333	3.521
10k Div topo detach	15.117	0.697	0.664	0.080	0.040	28.256	5.058
Spectral	8.699	0.775	0.627	-	-	-	-
+ cfg 1.5	8.454	0.870	0.485	0.049	0.027	10.449	3.658
Disperse	9.364	0.796	0.586	-	-	-	-
+ cfg 1.5	8.764	**0.874**	0.469	0.057	0.035	11.339	4.150

Table 2: Topological evaluation of the generated data manifold.

Second, we hypothesize that our topological constraints are better suited for the straight-line velocity fields of flow matching rather than the stochastic trajectories of the diffusion process used in SD 2.1. Future work will focus on re-formulating these topological objectives to accommodate the manifold geometry of VAEs and the non-linear sampling dynamics inherent to standard diffusion models.

5.4 Computational Performance

Computing persistence diagrams implies a computational overhead with respect to MSE. For comparison, in RAE experiments the Div losses require around 1.28x the backpropagation time of MSE, with full pipeline taking 1.37x. For MNIST experiments, the difference in performance grows (for more details and numbers, see Appendix E). While this overhead is not negligible, we still believe our method to be useful due to the following reasons: (1) Regularization requires few steps, with just 5-10k steps already showing results. (2) Scaling with respect to model size, as PDs computation is independent of number of model parameters. (3) Scaling with respect to input resolution, as PD's computation scales linearly exactly as MSE. (4) Potential for further computational improvements. For example, we observe that we can leverage the Johnson-Lindenstrauss lemma, which states that a high dimensional point-cloud can be projected onto a low dimensional space with distances between points being nearly preserved via a random orthogonal projection, to reduce computational cost of Vietoris-Rips filtrations (2.08x speedup) without loosing much performance. More details on Appendix B.3.

6 Conclusion & Limitations

In this work we asked the following question: can we use topological invariants to improve simultaneously the diversity and fidelity of the generations in flow matching and diffusion models? We presented two different methods to target this problem: either follow a classical approach in TDA for images, computing the filtration on the cubical complex of the velocity fields; or consider each batch of predicted samples as an approximation of the data topology, and try to match it with the ground truth (approximated) topology. We achieve this using filtrations on the Vietoris-Rips complexes of the flattened batches. We demonstrated that topological invariants can guide model finetuning, and showed that the Vietoris-Rips loss improves both diversity and fidelity over MSE and the cubical complexes loss. We would like to point out that this paper lacks large scale pre-training, as the goal is not to compare against state-of-the-art methods trained with MSE over millions of steps. Rather, we intend to show that regularizing the predicted velocity fields using persistent homology for a few steps yields a better topological approximation of the target, which translates into better diversity and fidelity. We hope this work contributes to improving generative modeling with flow matching.

7 AI USE STATEMENT

AI was used in this work for the following. Literature review; implementing parts of the training pipeline (RAE paper repository for the experiments in 5.2 was adapted to our training pipeline via Claude, with human verification); verifying the implementation of state of the art methods in section 5.2. In particular, both the spectral and Diffuse&Disperse regularization methods we compare against did not release code. As we believed comparison valuable, we implemented them with the help of AI, verifying step by step that the produced code followed the statements in each respective paper.

AI was not used for writing or styling. AI was not used for the mathematical derivations in Appendix B, except as a verifier. In particular, errors in the proof of Claim B.8 were found by LLM review. Corrections were made without AI assistance.

REFERENCES

Ulrich Bauer. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *J. Appl. Comput. Topol.*, 5(3):391–423, 2021. ISSN 2367-1726. doi: 10.1007/s41468-021-00071-5. URL https://doi.org/10.1007/s41468-021-00071-5.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Black Forest Labs. Flux 3 - real world models: Towards multimodal flow models as the backbone of visual intelligence. https://bfl.ai/blog/flux-3, July 2026.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, 2015.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pp. 2786–2796. PMLR, 2020.

Satish Chandran, Nicolas Roque dos Santos, Yunshu Wu, Greg Ver Steeg, and Evangelos Papalexakis. Spectral regularization for diffusion models. *arXiv preprint arXiv:2603.02447*, 2026.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

James R Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8766–8778, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Vin De Silva and Gunnar E Carlsson. Topological estimation using witness complexes. In *PBG*, pp. 157–166, 2004.

Mehdi Esmaeilzadeh, Alexia Jolicoeur-Martineau, Chirag Vashist, and Ke Li. One pass is not enough: Recursive latent refinement for generative models. *arXiv preprint arXiv:2605.15309*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in neural information processing systems*, 36:79858–79885, 2023.

Rickard Brüel Gabrielsson, Bradley J Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1553–1563. PMLR, 2020.

Florian Graf, Paolo Pellizzoni, Martin Uray, Stefan Huber, and Roland Kwitt. The flood complex: Large-scale persistent homology on millions of points. *arXiv preprint arXiv:2509.22432*, 2025.

Leonidas J Guibas and Steve Y Oudot. Reconstruction using witness complexes. *Discrete & computational geometry*, 40(3):325–356, 2008.

Saumya Gupta, Dimitris Samaras, and Chao Chen. TopoDiffusionNet: A Topology-aware Diffusion Model, March 2025. URL http://arxiv.org/abs/2410.16646. arXiv:2410.16646 [cs].

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Mustafa Hajij, Ghada Zamzmi, Theodore Papamarkou, Nina Miolane, Aldo Guzmán-Sáenz, Karthikeyan Natesan Ramamurthy, Tolga Birdal, Tamal K Dey, Soham Mukherjee, Shreyas N Samaga, et al. Topological deep learning: Going beyond graph data. *arXiv preprint arXiv:2206.00606*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Yasuaki Hiraoka, Yusuke Imoto, Shu Kanazawa, and Enhao Liu. Curse of dimensionality on persistence diagrams. *arXiv preprint arXiv:2404.18194*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. *Advances in neural information processing systems*, 30, 2017.

Christoph Hofer, Florian Graf, Marc Niethammer, and Roland Kwitt. Topologically densified distributions. In *International conference on machine learning*, pp. 4304–4313. PMLR, 2020.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky TQ Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. *arXiv preprint arXiv:2410.20587*, 2024.

Jiangbei Hu, Ben Fei, Baixin Xu, Fei Hou, Weidong Yang, Shengfa Wang, Na Lei, Chen Qian, and Ying He. Topology-Aware Latent Diffusion for 3D Shape Generation, January 2024. URL http://arxiv.org/abs/2401.17603. arXiv:2401.17603 [cs].

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32, 2019.

Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. In *NeurIPS 2025-Thirty-ninth Conference on Neural Information Processing Systems*, 2025.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9307–9315, 2024.

Shizuo Kaji, Takeki Sudo, and Kazushi Ahara. Cubical ripser: Software for computing persistent homology of image and volume data. *arXiv preprint arXiv:2005.12692*, 2020.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Xiang Li, Rongrong Wang, and Qing Qu. Towards understanding the mechanisms of classifier-free guidance. *arXiv preprint arXiv:2505.19210*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Zhen Liu, Tim Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed gflownets. In *International Conference on Learning Representations*, volume 2025, pp. 94806–94841, 2025.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International conference on machine learning*, pp. 7045–7054. PMLR, 2020.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020. URL https://arxiv.org/abs/2002.09797.

Naoki Nishikawa, Yuichi Ike, and Kenji Yamanishi. Adaptive topological feature via persistent homology: Filtration learning for point clouds. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 9131–9143. Curran Associates, Inc., 2023.

OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. Technical Report. Updates to the Sora framework focus on geometric consistency and physical law adherence.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Theodore Papamarkou, Tolga Birdal, Michael Bronstein, Gunnar Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, et al. Position: Topological deep learning is the new frontier for relational learning. *Proceedings of machine learning research*, 235:39529, 2024.

Dogyun Park and Suhyun Kim. Probabilistic precision and recall towards reliable evaluation of generative models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20099–20109, 2023.

Niket Patel, Luis Salamanca, and Luis Barba. Bridging the gap: Addressing discrepancies in diffusion model training for classifier-free guidance. *arXiv preprint arXiv:2311.00938*, 2023.

Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. Classifier-free guidance: From high-dimensional analysis to generalized guidance forms. *arXiv preprint arXiv:2502.07849*, 2025.

Yuchi Qiu and Guo-Wei Wei. Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models. *Briefings in Bioinformatics*, 24(5):bbad289, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

Xiaoqi Xu, Nicolas Drougard, and Raphaëlle N Roy. Topological data analysis as a new tool for eeg processing. *Frontiers in Neuroscience*, 15:761703, 2021.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025. URL https://arxiv.org/abs/2510.11690.

A TOPOLOGICAL DETAILS

Using a topological objective on velocity fields to optimize FM models requires two things: first, a computationally efficient and differentiable way of computing topological features and second, a differentiable metric for the comparison of these topological features.

A.1 COMPUTING TOPOLOGICAL FEATURES

As explained in Section 4, in this work we compute topological features computing the persistent homology of both cubical complexes on model predictions, and Vietoris-Rips complexes on batches of predictions. These are computed using the `cripser` (Kaji et al., 2020) and `ripser` (Bauer, 2021) libraries respectively. But there is a problem: persistent homology computation is non-differentiable. While there are some libraries that provide differentiable PH computations (for example `TopologyLayer` (Gabrielsson et al., 2020)), we found them to be too slow for our use-case.

What we do instead is keep track of critical indexes (i.e.: those that produce the birth or death of a component in the persistence diagram). In the fidelity loss, critical indexes correspond to pixels. In the diversity loss, critical indexes correspond to samples. In any case, we can backpropagate through pixels and samples, weighting their contribution by their topological relevance.

A.2 COMPARING TOPOLOGICAL FEATURES

Given two persistence diagrams, X_1 and X_2, they could simply be compared as pointsets in $\mathbb{R}^2$, by looking for *pairings* between the points in X_1 and X_2, and taking as distance between X_1 and X_2 the smallest among all pairings. But we need to take into account that PD_1 and PD_2 could have a different number of points. In most cases, this is due to noise, that is, features that are very close to the diagonal and have thus very short lifespans. We deal with this problem by allowing points to be matched to their closest projection in the diagonal. In this way, noise disappears into the diagonal and conservation of mass between distributions is maintained.

Definition A.1. The **bottleneck distance** $d_B(\cdot,\cdot)$ between two subsets X_1 and X_2 of $\mathbb{R}^2$ is defined as:

$$d_B(X_1, X_2) = \inf_{\gamma: X_1 \to X_2} \sup_{x \in X_1} \|x - \gamma(x)\|_\infty,$$

where γ ranges over bijections between points in X_1 and X_2, allowing matches to the diagonal.

Given that the bottleneck distance is computationally costly to compute, and not differentiable, it is frequent in TDA to rely on earth mover's alternatives.

Definition A.2. Let (M, d) be a metric space. The **Wasserstein** p-distance between two probability measures μ, η on M with finite p-moments is defined as

$$W_p(\mu, \eta) = \inf_{\gamma \in \Gamma(\mu,\eta)} (\mathbb{E}_{(x,y)\sim\gamma} d(x, y)^p)^{1/p}$$

where Γ ranges over all couplings of μ, η.

Or even better, the *Sinkhorn* distance, which is differentiable and can be computed in $O(n^2)$.

Definition A.3. Given two discrete probability distributions a and b, and a cost matrix C (where C_{ij} represents the cost to move mass from point i to point j), the **Sinkhorn distance** (entropic-regularized Optimal Transport cost) $W_\epsilon(a, b)$ is defined as:

$$W_\epsilon(a, b) = \min_{P \in \Pi(a,b)} \sum_{ij} P_{ij} C_{ij} + \epsilon \sum_{ij} P_{ij} (\log P_{ij} - 1).$$

Where $\Pi(a, b)$ is the set of all transport plans (matrices) that match the marginals a and b.

As mentioned above, to find a matching between features in two persistence diagrams and compute the distance, we need to allow for points in a diagram to be "destroyed" by being matched to the diagonal. We do this by simply matching a point with the diagonal whenever the cost of moving it

to its closest projection onto the diagonal is smaller than moving it to its closest match. In this case, we label it as a noisy feature.

On this work we use the Wasserstein implementation in the `gudhi` package to compute matchings between persistence diagrams.

A.3 Details on the final formulation

While the Sinkhorn distance provides a direct and differentiable comparison between target and predicted topological features, in practice we find this approach to not work. The Sinkhorn distance diffuses the signal over pixels (cubical complex filtration loss) and over samples (Vietoris-Rips filtration loss), providing information that is too noisy at the pixel level.

Instead of backpropagating through the difference between persistence diagrams, we use the Wasserstein distance (non backpropagatable) to find a matching of topological features between the predicted and target velocity fields. For each pair of matched features $f \in PD(u_t^\theta(X))$, $g \in PD(u_t(X \mid Z))$, we identify the indices of the samples that caused the *birth* or *death* of those features. In the case of the cubical complex loss, these correspond to pixels. In the case of the Vietoris-Rips filtration loss, these correspond to batch samples. We then backpropagate through the distance between these indices, forcing it to match the birth or death time of the ground truth topological feature. See Definition B.1 for a formal definition of Critical Indices.

A.4 Evaluation metrics.

Generative quality in flow matching and diffusion is usually measured via Fréchet Inception Distance (FID, (Heusel et al., 2017)), which encodes real and generated samples using a Inception-v3 network, and fits a multivariate gaussian to each set to compare both distributions:

$$FID((\mu_r, \Sigma_r), (\mu_g, \Sigma_g)) = \|\mu_r - \mu_g\|_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}.$$

Intrinsically, FID measures both diversity (distance between variances) and fidelity (distance between means).

However, fidelity and diversity are more accurately captured by the notions of precision and recall, which measure how many generated samples fall inside the real data manifold, and how many real samples are covered by the generated data manifold respectively. Here, each manifold is approximated by the union of adaptive k-nearest-neighbor neighborhoods in feature space. In addition to these, we evaluate coverage and density as proposed by (Naeem et al., 2020).

Coverage measures the fraction of real data points whose neighborhoods contain at least one generated sample, or how much of the real data distribution is covered by p_1 (diversity). For n real samples,

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\exists j \,:\, x_{gen}[j] \in S(x_{real}[i], r_k(x_{real}[i]))\}}.$$

where $\mathbb{1}$ is the indicator random variable, and S is the sphere centered in $x_{real}[i]$ with radius r_k (the distance to the k-th nearest neighbor). On the other hand, density measures how many real data spheres contain a generated point, that is, how well generated samples populate high-probability regions of the data manifold (fidelity). For m generated samples:

$$\text{Density} = \frac{1}{km} \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{\{x_{gen}[j] \in S(x_{real}[i], r_k(x_{real}[i]))\}}.$$

Evaluating topological quality. There is increasing evidence that FID evaluation is saturated ((Kynkäänniemi et al., 2022; Esmaeilzadeh et al., 2026)), and that it does not correlate with human perceived quality ((Stein et al., 2023; Jayasumana et al., 2024)). This is an issue, given that the quality of generative methods is evaluated against the baseline provided by FID. While it is out of the scope of this paper to propose alternative or more varied evaluation methods, we stress the need to complement FID evaluation with a variety of manifold coverage metrics such as precision and recall, as has been suggested many times in the literature (see (Stein et al., 2023)). We point

out, however, that precision and recall may not be enough, as they are well known to be sensitive to outliers ((Park & Kim, 2023)). To advance towards a more robust evaluation of the generated data manifold, we propose the use of persistent homology to compare its topological signature with that of the real data manifold.

We do this in the following manner. As for precision and recall evaluation, we generate a large set of samples (50k for Imagenet evaluation) to compare against the test set. We then compute embedding vectors using a visual encoder (we choose Inception-v3 to allow for a fair comparison with FID metrics). We then compute the persistence diagrams of the generated and real images, separate H_0 and H_1 features (connected components and loops) and compute their distance using the bottleneck or the Wasserstein distances as explained in the previous section.

In practice, this approach is prohibitive, as computing the persistence diagram on a set of thousands of points is very costly. The solution is to compute the PD on a set of landmarks, a subset of the original pointcloud that preserves its topology ((De Silva & Carlsson, 2004)). It is a common approach to obtain the landmark points via furthest point sampling ((Guibas & Oudot, 2008)). However, this method may oversample distant regions, which we fix by iteratively replacing each selected landmark with the medoid of its Voronoi cell, reducing the bias of greedy furthest point sampling toward boundary points. We set the number of landmarks to 1000.

Results can be seen in Table 2. If we focus on pure FID, we observe that a finetuning with MSE plus the Vietoris-Rips diversity loss obtains the best results, arguably because it significantly improves recall without harming precision greatly. When using CFG, the FID score actually goes up, because the precision improvement doesn't compensate for the recall loss. The most interesting insights come, however, when we look into the topological losses. The best H_0 distance is achieved by the baseline with CFG, and we can directly map this to the fact that this model has by far the highest precision, which means that for every connected component in the real data manifold there's an equivalent connected component in the generated data. However, we can observe worse values for the first homology group, indicating deficient manifold approximation not visible in pure FID measuring.

B Diversity Loss

Let $X_t \sim p_t(\cdot \mid Z)$. The diversity loss is formally defined as

$$\mathcal{L}_{\text{div}}(\theta) = \mathbb{E}_{t,Z,X_t}\Bigg[\sum_{(i,j,k)\in C(X|Z)} \lambda_t \underbrace{\left|\text{dist}\big(u_t^\theta(X_t)[i], u_t^\theta(X_t)[j]\big) - birth/death(u_t(X_t \mid Z)[k])\right|}_{\text{topological accuracy}} \cdot \underbrace{\text{dist}\big(u_t^\theta(X_t)[i] - u_t^\theta(X_t)[j], u_t(X_t \mid Z)[i] - u_t(X_t \mid Z)[j]\big)}_{\text{geometric alignment}}\Bigg].$$

For simplicity of notation, we will write

$$\mathcal{L}_{\text{topo}}^{i,j,k} = \mathcal{L}_{\text{topo}}^{i,j,k,x_t,z} := \left|\text{dist}\big(u_t^\theta(X_t)[i], u_t^\theta(X_t)[j]\big) - birth/death(u_t(X_t \mid Z)[k])\right|,$$

and

$$\mathcal{L}_{\text{geo}}^{i,j} = \mathcal{L}_{\text{geo}}^{i,j,x_t,z} := \text{dist}\big(u_t^\theta(X_t)[i] - u_t^\theta(X_t)[j], u_t(X_t \mid Z)[i] - u_t(X_t \mid Z)[j]\big).$$

On what follows, we will also denote the *marginal* components of the untractable objective as:

$$\mathcal{L}_{\text{g-topo}}^{i,j,k} := \left|\text{dist}\big(u_t^\theta(X_t)[i], u_t^\theta(X_t)[j]\big) - birth/death(u_t(X_t)[k])\right|,$$

and

$$\mathcal{L}_{\text{g-geo}}^{i,j} := \text{dist}\big(u_t^\theta(X_t)[i] - u_t^\theta(X_t)[j], u_t(X_t)[i] - u_t(X_t)[j]\big).$$

It therefore follows that the marginal version of the Vietoris-Rips loss is:

$$\mathcal{L}_{\text{g-div}}(\theta) = \mathbb{E}_{t,X_t}\Bigg[\sum_{i,j,k\in C(X_t)} \lambda_t \mathcal{L}_{\text{g-geo}}^{i,j} \cdot \mathcal{L}_{\text{g-topo}}^{i,j,k}\Bigg]$$

We define the set of critical points, $C(X \mid Z)$ as:

Definition B.1. The set of critical point matchings

$$C(X|Z) = C(PD(u_t^\theta(X)), PD(u_t(X \mid Z)) \\ = C^b(PD(u_t^\theta(X_t)), PD(u_t(X_t \mid Z))) \cup C^d(PD(u_t^\theta(X_t)), PD(u_t(X_t \mid Z))),$$

is the set of tuples of indices i, j, k such that:

1. At radius r, the connection of samples i and j via an edge produced the *birth* or *death* of feature $f \in PD(u_t^\theta(X))$,
2. Wasserstein distance maps $f \to g \in PD(u_t(X \mid Z))$,
3. k is the index of the corresponding sample in $u_t(X \mid Z)$ responsible for the *birth* of g, resp. *death* of g.

We denote the set of critical points that caused the *birth* of a feature by $C^b(PD(u_t^\theta(X_t))$, and the set of critical points that caused the *death* of a feature by $C^d(PD(u_t^\theta(X_t))$.

Similarly, the set of critical point matchings of the marginal vector field will be denoted by

$$C(X) = C(PD(u_t^\theta(X)), PD(u_t(X)) = C^b(u_t^\theta(X_t), u_t(X_t)) \cup C^d(u_t^\theta(X_t), u_t(X_t)).$$

We compute the set of critical point matchings as follows. For a fixed homology dimension (0 or 1) we identify which samples produce the birth or death (respectively) of that feature in both PDs. When studying connected components (0-th homology group, H_0), all features are born at distance 0, and their death is produced by the merge of two components. When looking into loops (H_1), their birth can occur by the combination of two features as well. Using a Wasserstein/earth mover's distance we can map features in both persistence diagrams, and thus obtain tuples $((i, j), k) \in C^b(u_t^\theta(X_t), u_t(X_t \mid Z))$ (or C^d) such that the combination of batch samples $u_t^\theta(X_t \mid Z)[i]$ and $u_t^\theta(X_t \mid Z)[j]$ produces the birth (or death) of a feature in $PD(u_t^\theta(X_t \mid Z))$ whose matching feature (according to Wasserstein/earth mover's distance) is $PD(u_t(X_t))[k]$. Notice that k is likely to be i or j, but does not need to be.

As mentioned in the main body, backpropagating through the conditional diversity loss $\mathcal{L}_{\text{div}}(\theta) = \mathbb{E}_{t,Z,X_t}\left[\sum_{(i,j,k)\in C(X_t|Z)} \lambda_t \mathcal{L}_{\text{geo}}^{i,j} \cdot \mathcal{L}_{\text{topo}}^{i,j,k}\right]$ essentially amounts to *moving* $u_t^\theta(X_t \mid Z)[i]$ *and* $u_t^\theta(X_t \mid Z)[j]$ *closer or further apart so that their position in the persistence diagram matches that of their corresponding feature in the ground truth, while ensuring they move along the correct direction*.

That is, optimizing this objective leads to a better approximation of the topology of the conditional vector field, as captured by its persistence diagram. But our goal, when training under the topological objective, is to generate data that aligns with the global topology of the target data manifold, be it directly via x-prediction or through the proxy of the velocity field (as in our experiments). On the following subsection, we will study under which conditions does the following hold:

> *Hypothesis*: Optimizing the conditional diversity loss leads to approximating the topology of the marginal vector field.

B.1 Approximating the topology of the marginal velocity field.

In a way akin to how the global vector field that moves the initial distribution p_0 to the target distribution p_1 is learned by optimizing over the conditional loss over batches (see Section 3.2), we'd hope that matching the topologies of predicted and ground truth batches results in a parametrization $u_t^\theta(x)$ of the marginal vector field $u_t(x)$. Moreover, we'd hope that optimizing the conditional topology loss allows us to learn the topology of the marginal vector field. Ideally, optimizing the conditional diversity loss should lead to optimization of the (untractable) marginal diversity loss:

$$\mathcal{L}_{\text{g-div}}(\theta) = \mathbb{E}_{t,X_t}\left[\sum_{(i,j,k)\in C(X_t)} \lambda_t \mathcal{L}_{\text{g-geo}}^{(i,j)} \mathcal{L}_{\text{g-topo}}^{b(i,j,k)}\right].$$

Before we explore whether the claim above holds, let us take a look at the topological accuracy component of the loss[1]:

Claim B.2. *Fix a time t and indices i, j, k. Then*

$$\mathbb{E}_{z,x_t}\left[\mathcal{L}^{i,j,k}_{topo}\right] \geq \mathbb{E}_{x_t}\left[\mathcal{L}^{i,j,k}_{g\text{-}topo}\right]$$

if we choose cosine distance as a metric to compute the persistence diagram and assuming condOT probability paths. In particular, optimizing the conditional topology loss leads to optimizing the marginal topology loss.

Proof. We will first consider the case that we compute the Vietoris-Rips filtration using cosine distances. To simplify notation once more, let us denote by $u^{\theta}_t(x_t)[i] = u^{\theta,i}_t(x)$. Assuming normalized outputs, by the law of total expectation and Jensen's inequality we have:

$$\begin{aligned}
\mathbb{E}_{z,x}[\mathcal{L}^{i,j,k}_{\text{topo}}] &= \mathbb{E}_{z,x}\left[\left|(1 - u^{\theta,i}_t(x) \cdot u^{\theta,j}_t(x)) - birth/death(u^k_t(x \mid z))\right|\right] \\
&= \int\int \left|(1 - u^{\theta,i}_t(x) \cdot u^{\theta,j}_t(x)) - birth/death(u^k_t(x \mid z))\right| \; p(z,x)dxdz \\
&= \int\int \left|(1 - u^{\theta,i}_t(x) \cdot u^{\theta,j}_t(x)) - birth/death(u^k_t(x \mid z))\right| \; p(z \mid x)p(x)dzdx \\
&= \mathbb{E}_x\mathbb{E}_{z|x}\left[\left|(1 - u^{\theta,i}_t(x) \cdot u^{\theta,j}_t(x)) - birth/death(u^k_t(x \mid z))\right|\right] \\
&\geq \mathbb{E}_x\left|(1 - u^{\theta,i}_t(x) \cdot u^{\theta,j}_t(x)) - \mathbb{E}_{z|x}\left[birth/death(u^k_t(x \mid z))\right]\right|.
\end{aligned}$$

Notice the first term of the difference in the absolute value corresponds to the first term of the difference in the marginal topology loss. Let us focus on the second term and show that it is possible to marginalize $\mathbb{E}_{z|x}\left[death/birth(PD((u^k_t(x \mid z))\right]$. Essentially, the value $death/birth(PD(u^k_t(x)))$ represents the time (radius) in the Vietoris-Rips filtration when the feature $g \in PD(u_t(x \mid z))$ which corresponds to the k-th vector, $u^k_t(x \mid z)$ is merged to another feature $h \in PD(u_t(x \mid z))$, which happens because there's an edge between $u^k_t(x \mid z)$ and some other element $u^\ell_t(x \mid z)$. The other possibility is that $death/birth(PD(u^k_t(x))) = 0$, which happens for the birth of all H_0 features, in which case we are done. Thus, we can write $death/birth(PD(u^k_t(x))) = 1 - u^k_t(x \mid z)u^\ell_t(x \mid z)$. With this in mind:

$$\begin{aligned}
\mathbb{E}_{z|x}\left[birth/death(u^k_t(x \mid z))\right] &= 1 - \mathbb{E}_{z|x}\left[u^k_t(x \mid z)u^\ell_t(x \mid z)\right] \\
&= 1 - \mathbb{E}_{z|x}\left[u^k_t(x \mid z)\right]\mathbb{E}_{z|x}\left[u^\ell_t(x \mid z)\right] + Cov_{z|x}(u^k_t(x \mid z), u^\ell_t(x \mid z)).
\end{aligned}$$

Where $\mathbb{E}_{z|x}\left[u^k_t(x \mid z)\right] = u^k_t(x)$ by the definition of marginal vector field (see (Lipman et al., 2024)):

$$u^k_t(x) = \int u^k_t(x \mid z)p_{1|t}(z \mid x)dz = \mathbb{E}_{z|x}[u^k_t(x_t \mid z)].$$

On what follows, we will write $p(z \mid x) = p_{1|t}(z \mid x_t)$ when it is clear from context that $x_1 = z$, $x_t = x$. Let us focus now on the covariance term. Following the assumption that we are using condOT probability paths, we have: $u^k_t(x \mid z) = z^k - x^k_0$, and similarly $u^k_t(x) = \mathbb{E}[z^k] - x^k_0$. Therefore

$$\begin{aligned}
Cov_{z|x}(u^k_t(x \mid z), u^\ell_t(x \mid z)) &= \mathbb{E}_{z|x}\left[\left(u^k_t(x \mid z) - \mathbb{E}_{z|x}[u^k_t(x \mid z]\right)\left(u^\ell_t(x \mid z) - \mathbb{E}_{z|x}[u^\ell_t(x \mid z]\right)\right] \\
&= \mathbb{E}_{z|x}\left[\left(u^k_t(x \mid z) - u^k_t(x)\right)\left(u^\ell_t(x \mid z) - u^\ell_t(x)\right)\right] \\
&= \mathbb{E}_{z|x}\left[\left(z^k - x^k_0 - (\mathbb{E}[z^k] - x^k_0)\right)\left(z^\ell - x^\ell_0 - (\mathbb{E}[z^\ell] - x^\ell_0)\right)\right] \\
&= \mathbb{E}_{z|x}\left[\left(z^k - x\mathbb{E}[z^k]\right)\left(z^\ell - \mathbb{E}[z^\ell]\right)\right] = Cov_{z|x}(z^k, z^\ell) = 0.
\end{aligned}$$

Where the last equality holds, crucially, because batches are sampled independently from out dataset, i.e.: z^k and z^ℓ are independent. Which means $\mathbb{E}_{z|x}\left[birth/death(u^k_t(x \mid z))\right] = 1 - u^k_t(x)u^\ell_t(x) = birth/death(u^k_t(x))$. And therefore

$$\mathbb{E}_{z,x}[\mathcal{L}^{i,j,k}_{\text{topo}}] \geq \mathbb{E}_x\left[\left|(1 - u^{\theta,i}_t(x) \cdot u^{\theta,j}_t(x)) - birth/death(u^k_t(x))\right|\right] = \mathbb{E}_x[\mathcal{L}^{i,j,k}_{\text{g-topo}}].$$

□

[1] As a note, all the following derivations are assuming we train our models for v-prediction, but can be easily adapted for the x-prediction setup.

Corollary B.3. *In fact,*

$$\mathbb{E}_x[\mathcal{L}^{i,j,k}_{g\text{-}topo}] + C \geq \mathbb{E}_{z,x}[\mathcal{L}^{i,j,k}_{topo}] \geq \mathbb{E}_x[\mathcal{L}^{i,j,k}_{g\text{-}topo}]$$

Proof. From the point where we obtained the lower bound by Jensen, notice that also, using the triangle inequality:

$$\begin{aligned}
\mathbb{E}_{z,x}[\mathcal{L}^{i,j,k}_{\text{topo}}] &= \mathbb{E}_x\mathbb{E}_{z|x}\Big[\big|(1 - u_t^{\theta,i}(x)\, u_t^{\theta,j}(x)) - birth/death\big(u_t^k(x \mid z)\big)\big|\Big] \\
&= \mathbb{E}_x\mathbb{E}_{z|x}\Big[\big|(1 - u_t^{\theta,i}(x)\, u_t^{\theta,j}(x)) - \mathbb{E}_{z'|x}\big[birth/death\big(u_t^k(x \mid z')\big)\big] \\
&\qquad + \mathbb{E}_{z'|x}\big[birth/death\big(u_t^k(x \mid z')\big)\big] - birth/death\big(u_t^k(x \mid z)\big)\big|\Big] \\
&\leq \mathbb{E}_x\Big|(1 - u_t^{\theta,i}(x)\, u_t^{\theta,j}(x)) - \mathbb{E}_{z|x}\big[birth/death\big(u_t^k(x \mid z)\big)\big]\Big| \\
&\quad + \underbrace{\mathbb{E}_x\mathbb{E}_{z|x}\Big|birth/death\big(u_t^k(x \mid z)\big) - \mathbb{E}_{z'|x}\big[birth/death\big(u_t^k(x \mid z')\big)\big]\Big|}_{=:\, C \text{ (independent of } \theta)}.
\end{aligned}$$

□

Corollary B.4. *Using a simple Vietoris-Rips loss without geometric alignment component, $\mathcal{L} = \mathcal{L}_{topo}$ as above, and using cosine distance,* ***optimizing over the conditional vector field leads to approximating the topology of the marginal vector field****.*

However, notice that the same cannot be proven in the case that we use the l2-norm to compute distances in our Vietoris-Rips filtration:

Remark B.5. Fix a time t and indices i, j, k. If we choose the l2-norm as metric to compute the persistence diagram, it **may not hold in general** that

$$\mathbb{E}_{z,x_t}\big[\mathcal{L}^{(i,j,k)}_{\text{topo}}\big] \geq \mathbb{E}_{x_t}[\mathcal{L}^{(i,j,k)}_{\text{g-topo}}].$$

Proof. The expression simplifies when we consider that by Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}_{z,x_t}\big[\|\|(u_t^\theta(x_t)[i] - u_t^\theta(x_t)[j])\|^2 - death/birth(PD((u_t(x_t \mid z))[k]\|^2\big] \\
&\geq \mathbb{E}_{x_t}\big[\|\mathbb{E}_{z|x_t}\big[\|u_t^\theta(x_t)[i] - u_t^\theta(x_t)[j])\|^2 - death/birth(PD((u_t(x_t \mid z))[k]\big]\|^2\big] \\
&= \mathbb{E}_{x_t}\big[\|\|u_t^\theta(x_t)[i] - u_t^\theta(x_t)[j])\|^2 - \mathbb{E}_{z|x_t}\big[death/birth(PD((u_t(x_t \mid z))[k]\big]\|^2\big]
\end{aligned}$$

As in the case above, we can write $death/birth(PD(u_t(x))[k]) = \|u_t^k(x \mid z) - u_t^\ell(x \mid z)\|^2$. And, as we have shown in Claim B.7, this can be marginalized (plus covariance terms ≥ 0). Therefore

$$\begin{aligned}
\mathbb{E}_{z|x}\big[death/birth(PD((u_t(x_t \mid z))[k]\big] &= \mathbb{E}_{z|x}\big[\|u_t^k(x \mid z) - u_t^\ell(x \mid z)\|^2\big] = \|u_t^k(x) - u_t^\ell(x)\|^2 \\
&\quad + Tr(Cov_{z|x}(u_t^k(x \mid z)) + Tr(Cov_{z|x}(u_t^\ell(x \mid z)) \\
&\quad - 2Tr(Cov_{z|x}(u_t^k, u_t^\ell))) \\
&= \|u_t^k(x) - u_t^\ell(x)\|^2 + Tr(Cov_{z|x}(u_t^k(x \mid z) - u_t^\ell(x \mid z))).
\end{aligned}$$

However, $Tr(Cov_{z|x}(u_t^k(x \mid z) - u_t^\ell(x \mid z))) \geq 0$, and in fact it is unlikely to hold that the variance of $u_t^k(x \mid z) - u_t^\ell(x \mid z)$ is zero. Even if this term is a constant with respect to gradients, the outer l2 norm means that when computing the topological accuracy in euclidean distances we are pushing the predicted outs i, j to have the target birth/death time, *corrected* by the variance over targets z of vector differences for that x_t. The effect of this is not immediately clear to us, but it may be the reason why in experiments (see Section 5) the diversity loss doesn't work for SD-VAE latents, as well as why in pixel space we found euclidean distances to underperform with respect to cosine distances. Another reason may be the fact that when using euclidean distances, high dimensional noise destroys meaningful information in the PD, see (Hiraoka et al., 2024). □

Consider now the sets of critical point matchings. Let us see

vspace4pt

Claim B.6. *Fix a time t and indices (i, j, k). Let S_{ijk} denote the set of fields with a critical point matching at (i, j, k), and let us write*

$$a(x) := \mathbb{1}_{\{u^\theta(x) \in S_{ijk}\}}, \qquad b(x) := \mathbb{1}_{\{u(x) \in S_{ijk}\}}, \qquad b(x, z) := \mathbb{1}_{\{u(x|z) \in S_{ijk}\}},$$

so that $\mathbb{1}_{\{(i,j,k) \in C(x)\}} = a(x)\, b(x)$ and $\mathbb{1}_{\{(i,j,k) \in C(x|z)\}} = a(x)\, b(x, z)$. Let

$$\pi(x) := \mathbb{P}_{z \sim p_t(\cdot|x)}\big(u(x \mid z) \in S_{ijk}\big) = \int b(x, z)\, p_t(z \mid x)\, dz$$

be the posterior probability that a conditional field is critical at (i, j, k). Then

$$\mathbb{E}_{z \sim p_1}\big[\mathbb{P}_{x \sim p_t(\cdot|z)}\big((i, j, k) \in C(x \mid z)\big)\big] = \mathbb{E}_{x \sim p_t}\big[a(x)\, \pi(x)\big].$$

And therefore

$$\mathbb{E}_{z \sim p_1}\big[\mathbb{P}_{x \sim p_t(\cdot|z)}\big((i, j, k) \in C(x \mid z)\big)\big] = \mathbb{P}_{x \sim p_t}\big((i, j, k) \in C(x)\big) + \mathbb{E}_{x \sim p_t}\big[a(x)\, \big(\pi(x) - b(x)\big)\big].$$

Proof. As established by (Lipman et al., 2022), flow matching models employ probability paths p_t such that $p_t(x) = \int p_t(x \mid z)\, p_1(z)\, dz$, where p_1 is the data distribution. For $t \in [0, 1)$, $p_t(\cdot \mid z)$ is a density, and we assume $(x, z) \mapsto b(x, z)$ is jointly measurable. Wherever $p_t(x) > 0$, define the posterior $p_t(z \mid x) = p_t(x \mid z)\, p_1(z)/p_t(x)$, so that $p_t(x \mid z)\, p_1(z) = p_t(x)\, p_t(z \mid x)$. Since u^θ does not depend on z, neither does a. Hence

$$\begin{aligned}
\mathbb{E}_{z \sim p_1}\big[\mathbb{P}_{x \sim p_t(\cdot|z)}\big((i, j, k) \in C(x \mid z)\big)\big] &= \int p_1(z) \bigg(\int \mathbb{1}_{\{(i,j,k) \in C(x|z)\}}\, p_t(x \mid z)\, dx \bigg) dz \\
&= \int p_1(z) \bigg(\int a(x)\, b(x, z)\, p_t(x \mid z)\, dx \bigg) dz \\
&= \int \int a(x)\, b(x, z)\, p_t(x \mid z)\, p_1(z)\, dz\, dx \\
&= \int \int a(x)\, b(x, z)\, p_t(x)\, p_t(z \mid x)\, dz\, dx \\
&= \int p_t(x)\, a(x) \bigg(\int b(x, z)\, p_t(z \mid x)\, dz \bigg) dx \\
&= \int p_t(x)\, a(x)\, \pi(x)\, dx \\
&= \int p_t(x)\, a(x)\, b(x)\, dx + \int p_t(x)\, a(x)\, \big(\pi(x) - b(x)\big)\, dx \\
&= \int \mathbb{1}_{\{(i,j,k) \in C(x)\}}\, p_t(x)\, dx + \mathbb{E}_{x \sim p_t}\big[a(x)\, \big(\pi(x) - b(x)\big)\big] \\
&= \mathbb{P}_{x \sim p_t}\big((i, j, k) \in C(x)\big) + \mathbb{E}_{x \sim p_t}\big[a(x)\, \big(\pi(x) - b(x)\big)\big].
\end{aligned}$$

The second step uses $\mathbb{1}_{\{(i,j,k) \in C(x|z)\}} = a(x)\, b(x, z)$, and the third is Tonelli's theorem (which we can use, because the integrand is nonnegative). For the fourth we apply Bayes' rule $p_t(x \mid z)\, p_1(z) = p_t(x)\, p_t(z \mid x)$. In the before-last we add and substract $\int p_t(x)\, a(x)\, b(x)\, dx$ and then conclude using $\mathbb{1}_{\{(i,j,k) \in C(x)\}} = a(x)\, b(x)$. □

That is, when marginalizing over z, we are recovering the critical point set of the marginal vector field. The intuition behind this is the following: for i, j, k fixed, and given that batch samples are selected randomly, the above probability is equal to the amount of Wasserstein matchings between features in $PD(u_t^\theta(X_t))$ and $PD(u_t(X))$ (i.e.: features that are not mapped to the diagonal). But all $X_t \sim p_t$ can be recovered by marginalizing the conditional probability distribution over z.

Let us now turn our attention to the geometrical alignment component of the loss to see whether something similar holds:

Claim B.7. *Let, as before, t and i, j, k fixed. Then*

$$\begin{aligned}
&\mathbb{E}_{z, x_t}\big[\|(u_t^\theta(x_t)[i] - u_t^\theta(x_t)[j]) - (u_t(x_t \mid z)[i] - u_t(x_t \mid z)[j])\|^2\big] \\
&\qquad \geq \mathbb{E}[\|(u_t^\theta(x_t)[i] - u_t^\theta(x_t)[j]) - (u_t(x_t)[i] - u_t(x_t)[j])\|^2].
\end{aligned}$$

That is, if we choose l2 norm as distance,

$$\mathbb{E}_{z,x_t}\big[\mathcal{L}_{geo}^{(i,j)}\big] \geq \mathbb{E}_{x_t}[\mathcal{L}_{g\text{-}geo}^{(i,j)}].$$

Proof. To simplify notation once more, let us denote by $u_t^\theta(x_t)[i] = u_t^{\theta,i}(x)$. First, let us expand the second term of the inequality:

$$\begin{aligned}
&\mathbb{E}_x[\|(u_t^\theta(x_t)[i] - u_t^\theta(x_t)[j]) - (u_t(x_t)[i] - u_t(x_t)[j])\|^2] \\
&= \mathbb{E}_x\big[\|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))\|^2 - 2(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))^\top(u_t^i(x) - u_t^j(x)) + \|u_t^i(x) - u_t^j(x)\|^2\big] \\
&= \|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))\|^2 - 2\mathbb{E}_x\big[u_t^{\theta,i}(x)^\top u_t^i(x)\big] - 2\mathbb{E}_x\big[u_t^{\theta,i}(x)^\top u_t^j(x)\big] - 2\mathbb{E}_x\big[u_t^{\theta,j}(x)^\top u_t^i(x)\big] \\
&- 2\mathbb{E}_x\big[u_t^{\theta,j}(x)^\top u_t^j(x)\big] + \mathbb{E}_x\big[\|u_t^i(x) - u_t^j(x)\|^2\big]
\end{aligned}$$

On the other hand:

$$\begin{aligned}
&\mathbb{E}_{z,x}\big[\|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x)) - (u_t^i(x \mid z) - u_t^j(x \mid z))\|^2\big] \\
&= \mathbb{E}_{z,x}\big[\|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))\|^2 - 2(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))^\top(u_t^i(x \mid z) - u_t^j(x \mid z)) \\
&+ \|u_t^i(x \mid z) - u_t^j(x \mid z)\|^2\big] \\
&= \|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))\|^2 - 2\mathbb{E}_{z,x}\big[u_t^{\theta,i}(x)^\top u_t^i(x \mid z)\big] - 2\mathbb{E}_{z,x}\big[u_t^{\theta,i}(x)^\top u_t^j(x \mid z)\big] \\
&- 2\mathbb{E}_{z,x}\big[u_t^{\theta,j}(x)^\top u_t^i(x \mid z)\big] \\
&- 2\mathbb{E}_{z,x}\big[u_t^{\theta,j}(x)^\top u_t^j(x \mid z)\big] + \mathbb{E}_{z,x}\big[\|u_t^i(x \mid z) - u_t^j(x \mid z)\|^2\big]
\end{aligned}$$

Component by component we can apply the following argument (see (Lipman et al., 2024)):

$$\begin{aligned}
\mathbb{E}_{z,x}\big[u_t^{\theta,i}(x)^\top u_t^j(x \mid z)\big] &= \int\int u_t^{\theta,i}(x)^\top u_t^j(x \mid z) p_t(x \mid z) p_1(z) dz dx \\
&= \int\int p_t(x) u_t^{\theta,i}(x)^\top u_t^j(x \mid z) \frac{p_t(x \mid z) p_1(z)}{p_t(x)} dz dx \\
&= \int p_t(x) u_t^{\theta,i}(x)^\top \int u_t^j(x \mid z) \frac{p_t(x \mid z) p_1(z)}{p_t(x)} dz dx \\
&= \int p_t(x) u_t^{\theta,i}(x)^\top u_t^j(x) dx = \mathbb{E}_x[u_t^{\theta,i}(x) u_t^j(x)]
\end{aligned}$$

And similarly $\mathbb{E}_{z,x}\big[u_t^{\theta,i}(x)^\top u_t^i(x \mid z)\big] = \mathbb{E}_x[u_t^{\theta,i}(x)^\top u_t^i(x)]$, $\mathbb{E}_{z,x}\big[u_t^{\theta,i}(x)^\top u_t^j(x \mid z)\big] = \mathbb{E}_x[u_t^{\theta,i}(x)^\top u_t^j(x)]$ and $\mathbb{E}_{z,x}\big[u_t^{\theta,j}(x)^\top u_t^j(x \mid z)\big] = \mathbb{E}_x[u_t^{\theta,j}(x)^\top u_t^j(x)]$.

That gives us all crossed components of the claimed inequality. And we also have $\mathbb{E}_{z,x}\big[\|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))\|^2\big] = \mathbb{E}_x\big[\|(u_t^{\theta,i}(x) - u_t^{\theta,j}(x))\|^2\big]$, which gives us the first term. Finally,

$$\begin{aligned}
\mathbb{E}_{z,x}\big[\|u_t^i(x \mid z) - u_t^j(x \mid z)\|^2\big] &= \mathbb{E}_{z,x}\big[\|u_t^i(x \mid z)\|^2\big] - 2\mathbb{E}_{z,x}\big[u_t^i(x \mid z)^\top u_t^j(x \mid z)\big] + \mathbb{E}_{z,x}\big[\|u_t^j(x \mid z)\|^2\big] \\
&= \|\mathbb{E}_{x,z|x}[u_t^i(x \mid z)]\|^2 + \|\mathbb{E}_{x,z|x}[u_t^j(x \mid z)]\|^2 \\
&+ \mathbb{E}_{x,z|x}[u_t^i(x \mid z)]\mathbb{E}_{x,z|x}[u_t^j(x \mid z)] \\
&+ \mathbb{E}_x[Tr(Cov_{z|x}(u_t^i(x \mid z)) + Tr(Cov_{z|x}(u_t^j(x \mid z)) \\
&- 2Tr(Cov_{z|x}(u_t^i, u_t^j))] \\
&= \mathbb{E}_x[\|u_t^i(x)\|^2] + \mathbb{E}_x[\|u_t^j(x)\|^2] - 2\mathbb{E}_x[u_t^{i\top}(x) u_t^j(x)] \\
&+ \mathbb{E}_x[Tr(Cov_{z|x}(u_t^i(x \mid z)) + Tr(Cov_{z|x}(u_t^j(x \mid z)) \\
&- 2Tr(Cov_{z|x}(u_t^i, u_t^j)))]
\end{aligned}$$

Where we are using the law of total expectation:

$$\begin{aligned}
\mathbb{E}_{z,x}\big[\|u_t^i(x \mid z)\|^2\big] &= \int\int \|u_t^i(x \mid z)\|^2 p(z,x) dx dz = \int\int \|u_t^i(x \mid z)\|^2 p(z \mid x) p(x) dz dx \\
&= \mathbb{E}_x \mathbb{E}_{z|x}\big[\|u_t^i(x \mid z)\|^2\big] = \mathbb{E}_x[\|\mathbb{E}_{z|x}[u_t^i(x \mid z)]\|^2 + Tr(Cov_{z|x}(u_t^i(x \mid z)))\big] \\
&= \mathbb{E}_x[\|u_t^i(x)\|^2] + \mathbb{E}_x[Tr(Cov_{z|x}(u_t^i(x \mid z))].
\end{aligned}$$

Overall, we are obtaining all terms and expressions of $\mathcal{L}^{(i,j)}_{\text{g-geo}}$, plus the covariance terms. Consider then that, by Cauchy-Schwarz we have $Tr(Cov_z(u^i_t(x \mid z)) + Tr(Cov_z(u^j_t(x \mid z)) \geq 2|Tr(Cov_z(u^i_t, u^j_t))|$. Or, to make the argument simpler:

$$\mathbb{E}_x[Tr(Cov_{z|x}(u^i_t(x \mid z)) + Tr(Cov_{z|x}(u^j_t(x \mid z)) - 2Tr(Cov_{z|x}(u^i_t, u^j_t)))] \\ = \mathbb{E}_x[Tr(Cov(u^i_t - u^j_t))] \geq 0,$$

because the covariance matrix is positive semi-definite. This allows us to conclude, for some $C \geq 0$:

$$\mathbb{E}_z\big[\mathcal{L}^{(i,j)}_{\text{geo}}\big] = \mathcal{L}^{(i,j)}_{\text{g-geo}} + C \implies \mathbb{E}_z\big[\mathcal{L}^{(i,j)}_{\text{geo}}\big] \geq \mathcal{L}^{(i,j)}_{\text{g-geo}}.$$

□

Which means that as we minimize $\mathbb{E}_z\big[\mathcal{L}^{(i,j)}_{\text{geo}}\big]$ we also minimize $\mathcal{L}^{(i,j)}_{\text{g-geo}}$. It is not difficult to prove an analogous result here if we consider cosine distance.

Notice there is, apriori, no reason why we couldn't define the diversity loss as simply the sum of the geometrical and topological terms, instead of the product. We settle for the product loss, because we want geometrical alignment gradients to be scaled by the topological accuracy and vice versa, and because we found it to empirically work slightly better, but would like to mention here that the sum is also valid choice of loss, which offers besides strong theoretical guarantees. We will outline the following proof for this version of the loss, namely:

$$\mathcal{L}_{\tilde{div}}(\theta) = \mathbb{E}_{t,Z,X_t}\Bigg[\sum_{(i,j,k)\in C(X|Z)} \Big(\underbrace{(\text{dist}(u^\theta_t(X_t)[i], u^\theta_t(X_t)[j]) - birth/death(u_t(X_t \mid Z)[k])}_{\text{topological accuracy}} \\ + \underbrace{\text{dist}\big(u^\theta_t(X_t)[i] - u^\theta_t(X_t)[j]), (u_t(X_t \mid Z)[i] - u_t(X_t \mid Z)[j])\big)}_{\text{geometric alignment}} \Big)\Bigg].$$

Under this setting, let us prove:

Claim B.8. *Assume we use cosine distances to compute the persistence diagram and CondOT paths as probability paths, and let c be the constant of Claim B.7. For each t and (i,j,k), let $\bar{q}_{ijk} = \mathbb{P}_{x_t}\big((i,j,k) \in C(x_t)\big)$. Then, for every θ,*

$$\mathcal{L}_{g\text{-}\tilde{div}}(\theta) \leq \mathcal{L}_{\tilde{div}}(\theta) + (4-c)^+ \mathbb{E}_t \mathbb{E}_{x_t}\big[|C(x_t)|\big].$$

where $(c)^+ = \max(c, 0)$. That is, up to a constant independent of θ, $\mathcal{L}_{\tilde{div}}$ is an upper bound on the marginal *loss $\mathcal{L}_{g\text{-}\tilde{div}}$: any θ with $\mathcal{L}_{\tilde{div}}(\theta) \leq \delta$ satisfies $\mathcal{L}_{g\text{-}\tilde{div}}(\theta) \leq \delta + (4-c)^+ N$.*

Proof. This proof is reminiscent of the the proof of the marginal flow matching loss (see (Lipman et al., 2024)), but we need to keep in mind that we are summing over critical points. By definition of the conditional diversity loss, we have:

$$\mathcal{L}_{\tilde{div}}(\theta) = \mathbb{E}_{t,z,x_t}\Bigg[\sum_{(i,j,k)\in C(x_t|z)} \mathcal{L}^{(i,j)}_{\text{geo}} + \mathcal{L}^{(i,j,k)}_{\text{topo}}\Bigg] = \mathbb{E}_{t,z,x_t}\Bigg[\sum_{i,j,k} \mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}(\mathcal{L}^{(i,j)}_{\text{geo}} + \mathcal{L}^{(i,j,k)}_{\text{topo}})\Bigg] \\ = \sum_{i,j,k}\Bigg(\mathbb{E}_{t,z,x_t}\Big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{(i,j)}_{\text{geo}}\Big] + \mathbb{E}_{t,z,x_t}\Big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{(i,j)}_{\text{topo}}\Big]\Bigg)$$

After applying the law of total expectation to obtain

$$\mathcal{L}_{\tilde{div}}(\theta) = \sum_{i,j,k}\Bigg(\mathbb{E}_{t,x_t}\mathbb{E}_{z|x_t}\Big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{(i,j)}_{\text{geo}}\Big] + \mathbb{E}_{t,x_t}\mathbb{E}_{z|x_t}\Big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{(i,j)}_{\text{topo}}\Big]\Bigg),$$

let us now consider each component separately. First, notice that for $\mathbb{E}_{z|x}\big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{i,j,k}_{\text{topo}}\big]$,

$\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}} = 0 \implies \mathcal{L}^{i,j,k}_{\text{topo}} = 0.$ Therefore: $\mathbb{E}_{z|x}\big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{i,j,k}_{\text{topo}}\big] = \mathbb{E}_{z|x}\big[\mathcal{L}^{i,j,k}_{\text{topo}}\big]$.

And we may apply Claim B.2 to conclude $\mathbb{E}_{z,x_t}[\mathcal{L}^{i,j,k}_{\text{topo}}] \geq \mathbb{E}_{x_t}[\mathcal{L}^{i,j,k}_{\text{g-topo}}]$. Here we write $x = x_t$ when it is clear from context that time is fixed. On the other hand:
$\mathbb{E}_{z|x}\big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{i,j}_{\text{geo}}\big] = \mathbb{E}_{z|x}\big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\big]\mathbb{E}_{z|x}\big[\mathcal{L}^{i,j}_{\text{geo}}\big] + Cov_{z|x}\big(\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}, \mathcal{L}^{i,j}_{\text{geo}}\big)$.
Let us denote for simplicity of notation

$$I = I(x_t \mid z) = \mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}, \qquad p = \mathbb{P}(I=1), \qquad D = D(x_t, z) = \mathcal{L}^{i,j,x_t,z}_{\text{geo}}.$$

Since I is Bernouilli we have

$$\mathbb{E}_{z|x}[D] = p\mathbb{E}_{z|x}[D \mid I=1] + (1-p)\mathbb{E}_{z|x}[D \mid I=0]. \quad \text{Also,}$$
$$\mathbb{E}_{z|x}[ID] = \mathbb{E}_{z|x}[D\mathbb{1}_{\{I=1\}}] = p\mathbb{E}_{z|x}[D \mid I=1]$$

Which means that

$$\begin{aligned} Cov_{z|x}(I,D) &= p\mathbb{E}_{z|x}[D \mid I=1] - p(p\mathbb{E}_{z|x}[D \mid I=1] + (1-p)\mathbb{E}_{z|x}[D \mid I=0]) \\ &= p(1-p)(\mathbb{E}_{z|x}[D \mid I=1] - \mathbb{E}_{z|x}[D \mid I=0]), \end{aligned}$$

where we can easily bound $0 \leq p(1-p) \leq 1/4$ by convexity, and given that $D \in [0,2]$ (cosine distance) this gives a hard bound of $\big|Cov_{z|x}(I,D)\big| \leq 1/2$, which corresponds to the worst case scenario $p = 1/2$.

Consider now

$$\begin{aligned} \mathbb{E}_x\mathbb{E}_{z|x}\big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\mathcal{L}^{i,j}_{\text{geo}}\big] &= \mathbb{E}_x\big[p\mathbb{E}_{z|x}[D]\big] + \mathbb{E}_x\big[Cov_{z|x}(I,D)\big] \\ &= \mathbb{E}_x[p]\mathbb{E}_x\mathbb{E}_{z|x}[D] + Cov_x[p, \mathbb{E}_{z|x}[D]] + \mathbb{E}_x\big[Cov_{z|x}(I,D)\big] \\ &= \mathbb{E}_x[p]\mathbb{E}_x\mathbb{E}_{z|x}[D] + Cov_{x,z}[I,D] \end{aligned}$$

where the last step holds by the law of total covariance. Using the same Bernoulli identity as above, with $\bar{p} = \mathbb{E}_x[p(x)]$, we have $\mathrm{Cov}_{x,z}(I,D) = \bar{p}(1-\bar{p})\big(\mathbb{E}_{x,z}[D \mid I=1] - \mathbb{E}_{x,z}[D \mid I=0]\big)$ and thus $|Cov_{x,z}(I,D)| \leq 1/2$. Using Claim B.6 and Claim B.7, and writing $\bar{q} = \mathbb{E}_x[\mathbb{1}_{\{i,j,k\in C(x)\}}]$ we have:

$$\begin{aligned} \mathbb{E}_x\mathbb{E}_{z|x}\big[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\,\mathcal{L}^{i,j}_{\text{geo}}\big] &= \mathbb{E}_{x,z}[I]\,\mathbb{E}_{x,z}[D] + \mathrm{Cov}_{x,z}(I,D) \\ &= \bar{q}\,\mathbb{E}_x\big[\mathcal{L}^{i,j}_{\text{g-geo}} + c\big] + \mathrm{Cov}_{x,z}(I,D) \\ &= \bar{q}\,\mathbb{E}_x\big[\mathcal{L}^{i,j}_{\text{g-geo}}\big] + c\,\bar{q} + \mathrm{Cov}_{x,z}(I,D) \\ &= \mathbb{E}_x\big[\mathbb{1}_{\{i,j\in C(x)\}}\mathcal{L}^{i,j}_{\text{g-geo}}\big] + c\,\bar{q} + \underbrace{\mathrm{Cov}_{x,z}(I,D) - \mathrm{Cov}_x\big(\mathbb{1}_{\{i,j\in C(x)\}}, \mathcal{L}^{i,j}_{\text{g-geo}}\big)}_{=:\,\alpha_{ijk}}. \end{aligned}$$

Write $J = \mathbb{1}_{\{i,j\in C(x)\}}$. Then, using the same argument as above:

$$\mathrm{Cov}_{x,z}(I,D) = \bar{q}(1-\bar{q})\big(\mathbb{E}[D \mid I=1] - \mathbb{E}[D \mid I=0]\big) \geq -2\,\bar{q}(1-\bar{q}),$$
$$\mathrm{Cov}_x\big(J, \mathcal{L}^{i,j}_{\text{g-geo}}\big) = \bar{q}(1-\bar{q})\big(\mathbb{E}[\mathcal{L}^{i,j}_{\text{g-geo}} \mid J=1] - \mathbb{E}[\mathcal{L}^{i,j}_{\text{g-geo}} \mid J=0]\big) \leq 2\,\bar{q}(1-\bar{q}),$$

which implies, since $4\bar{q}^2 \geq 0$:

$$c\,\bar{q} + \varepsilon_{ijk} \geq \bar{q}\,(c - 4 + 4\bar{q}) \geq -(4-c)^+\,\bar{q},$$

where $(c)^+ = \max(c,0)$. Afterwards we only need to sum over (i,j,k) and use $\sum_{i,j,k}\bar{q}_{ijk} = \mathbb{E}_x\big[|C(x)|\big]$ (the number of critical points), to reduce the final term. With that we arrive at

$$\begin{aligned} \mathcal{L}_{\text{div}}(\theta) &= \sum_{i,j,k}\left(\mathbb{E}_{t,x_t}\mathbb{E}_{z|x_t}\left[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\,\mathcal{L}^{i,j}_{\text{geo}}\right] + \mathbb{E}_{t,x_t}\mathbb{E}_{z|x_t}\left[\mathbb{1}_{\{(i,j,k)\in C(x_t|z)\}}\,\mathcal{L}^{i,j}_{\text{topo}}\right]\right) \\ &= \mathbb{E}_t\left[\sum_{i,j,k}\left(\mathbb{E}_{x_t}\left[\mathbb{1}_{\{(i,j,k)\in C(x_t)\}}\,\mathcal{L}^{i,j}_{\text{g-geo}}\right] + c\,\bar{q}_{ijk} + \alpha_{ijk} + \mathbb{E}_{x_t}\left[\mathbb{1}_{\{(i,j,k)\in C(x_t)\}}\,\mathcal{L}^{i,j}_{\text{g-topo}}\right]\right)\right] \\ &= \mathbb{E}_t\left[\sum_{i,j,k}\mathbb{E}_{x_t}\left[\mathbb{1}_{\{(i,j,k)\in C(x_t)\}}\left(\mathcal{L}^{i,j}_{\text{g-geo}} + \mathcal{L}^{i,j}_{\text{g-topo}}\right)\right]\right] + \mathbb{E}_t\left[\sum_{i,j,k}\left(c\,\bar{q}_{ijk} + \alpha_{ijk}\right)\right] \\ &= \mathcal{L}_{\text{g-div}}(\theta) + \mathbb{E}_t\left[\sum_{i,j,k}\left(c\,\bar{q}_{ijk} + \alpha_{ijk}\right)\right] \geq \mathcal{L}_{\text{g-div}}(\theta) - (4-c)^+\,\mathbb{E}_t\left[\sum_{i,j,k}\bar{q}_{ijk}\right] \\ &= \mathcal{L}_{\text{g-div}}(\theta) - (4-c)^+\,\mathbb{E}_t\mathbb{E}_{x_t}\big[|C(x_t)|\big]. \end{aligned}$$

□

That is, we can't guarantee that a minimizer of the conditional loss will be a minimizer of the marginal loss, but we can guarantee that optimizing the conditional loss approximates the topology of the marginal vector field. If a stronger theoretical guarantee is needed for convergence under optimization, we recommend relying on Corollary B.4 and defining the loss with the geometrical alignment component detached.

A similar result can perhaps be derived for the product version of the loss. We have not yet been able to arrive to it, given that there are many complicated cross-terms and covariances, but given the experimental evidence we consider it safe to optimize using the product version of the loss.

B.2 Differentiability of $\mathcal{L}_{div}(\theta)$

Let us make a note about how gradients flow through $\mathcal{L}_{div}$ and $\mathcal{L}_{\tilde{div}}$, given that persistence diagram computations are not differentiable in `ripser`. This implies: (1) no gradients flow through birth$(u_t(X_t \mid Z)[k]$, and (2) we do not act on the persistence diagram itself, but on a set of critical indices C. Therefore, gradients flow normally through $u_t^\theta(X_t)[i]$ and $u_t^\theta(X_t)[j]$, for a set of i, j defined by the persistence diagram. In this way, everything remains differentiable and gradients correctly modify the pixels of $u_t^\theta(X_t)[i]$ and $u_t^\theta(X_t)[j]$.

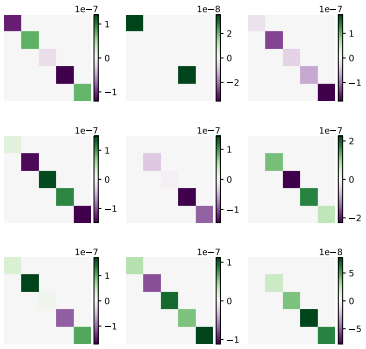


Figure 4: Similar set up as in Figure 3 with a batch of 9 images. We show $\nabla_{u_{pred}} \mathcal{L}_{div}(u_{pred}[i], u_{tgt}[i])$ for each sample.

There is another possibility, though: compute gradients directly on the difference between $PD(u_t^\theta(X_t))$ and $PD(u_t(X_t \mid Z)$, via a Sinkhorn loss, which is differentiable, or on the mean squared error over the persistence landscape.

Persistence landscapes are computed by mapping persistence diagrams PD_1 and PD_2 to a metric space where we can simply measure the distance between $f(PD_1)$ and $f(PD_2)$. This can be achieved via persistence landscapes (Bubenik, 2015), which transform the birth-death coordinates of a persistence diagram into a sequence of continuous, piecewise-linear functions by associating each point (b, d) with a function:

$$f_{(b,d)}(r) = \begin{cases} r - b, & \text{if } t \in [b, \frac{b+d}{2}] \\ d - r, & \text{if } t \in (\frac{b+d}{2}, d] \\ 0, & \text{otherwise.} \end{cases}$$

We tried to backpropagate through the distance between persistence diagrams directly in our experiments, both using Sinkhorn distances and euclidean distances on the persistence landscape, but found that they consistently underperformed when compared to $\mathcal{L}_{div}$ as described above. The issue seemed to be that gradients diffused over samples, providing a noisy signal to the network. We still believe backpropagating directly through the persistence diagrams can be fruitful approach, but more research is needed in this direction.

B.3 Computational optimization of the diversity loss

There is evidence that shows persistent homology works best in low dimensional settings, but is sensitive to the curse of dimensionality. This is proved, for example, in (Hiraoka et al., 2024).

The issue is that Vietoris-Rips complex computation rests solely on pairwise distances. In high dimension, noise produces large artifacts in the persistence diagram, rendering it unusable. We take this obstruction seriously, as in our Vietoris-Rips complex loss the batch size (size of the pointcloud) is always going to be much smaller than the total dimension for most use cases in Deep Learning. This is why we work with Representation Autoencoders: we can compute the filtration projecting onto the sphere, i.e.: using cosine distances, which allows the loss to capture geometrical information even in high dimensions and small batch sizes, as we demonstrated. See Section 5.1.

This works because cosine distances are meaningful in the RAE (Dinov2) latent space. However, when moving to VAE latent diffusion and flow matching models, cosine distances no longer keep that much information, which results in our approach not working. On the other hand, as mentioned in Section 5, computing the persistence diagram implies a costly computational overhead, that in some cases might be a concern.

To deal with this issues, we tried three different approaches:

1. **Sparse Random Projection.** We leverage the Johnson-Lindenstrauss lemma, which states that a high dimensional pointcloud can be projected onto a low dimensional space with distances between points being nearly preserved via a random orthogonal projection. We assume nearly preserved distances are good enough to capture the topological information of a space
2. **PCA Projection.** Following the advise in (Hiraoka et al., 2024), we project to a lower dimensional space using normalized PCA, and compute the Vietoris-Rips filtration there using cosine distances. In order for this to work, we precomputed a projection matrix on a sample of 10k images on the COCO dataset, and projected to 128 dimensions.

However, in our experiments we found that while the diversity loss over PCA projections roughly matches the performance of MSE, it greatly underperformed with respect to the one computed in the full vector space. The one computed with respect to the Johnson-Lindenstrauss projection fared better, see Table 3. We also include for comparison the Div loss where we use Spectral distances to compute the Vietoris-Rips filtration, which also leads to faster computation with performance gains.

	MSE	MSE_Div	JL proj	PCA proj	Spectral	Div	MSE_Fid
Performance	**4.36it/s**	1.54it/s	3.17it/s	2.92it/s	2.46it/s	1.52it/s	1.05s/it
FID	13.25	7.39	8.70	13.83	9.19	13.02	11.31

Table 3: Performance accounts for full pipeline performance in it/s. FID evaluation is reported after 10k steps of finetuning on the RAE Baseline. *JL proj* refers to approach 1. of computing the diversity loss via a sparse random projection, and *Spectral* refers to approach 3. All *JL proj*, *PCA proj* and *Spectral* correspond to training with MSE + Div loss.

C The effect of the choice of distance metric for the computation of filtrations

As we have seen, the choice of metric space greatly impacts persistent homology computation. While l2 looks like a sensible choice, in fact (see (Hiraoka et al., 2024)) it is unreliable due to the appearance of high dimensional noise. We tried to move around this using PCA projections using a precomputed matrix before the metric computation, but this approached performed worse than using cosine distance both in pixel space and RAE space. This is why we settle for cosine distance in all our experiments.

We could, however, experiment with different metrics. As we saw in Section 5, our experiments fail in SD-VAE latent space. We also conducted experiments on Cifar pixel space, and gains where limited when using cosine distance, which shows specially suited for RAE latents. Spectral metrics such as gaussian distances are suggested by (Hiraoka et al., 2024) as a way to avoid the curse of dimensionality in PD computation. We also experimented with Fermat distances, but leave further steps in this area as future work.

Experiment	FID	Precision	Recall	Density	Coverage	CLIP
BASE mse	15.798	0.804	0.557	1.091	0.915	27.070
Div only, $\lambda_t = 100$	15.465	0.807	0.550	1.094	0.919	27.238
Div only, $\lambda_t = 1000$	15.833	0.810	0.547	1.149	0.918	27.244
Div only, $\lambda_t = 10000$	15.854	0.805	0.548	1.129	0.918	27.237

Table 4: Small ablation on Div loss scale.

D Details on the experiments

We detail here all hyperparameters used in our experiments.

D.1 Experiments in pixel-space

Experiments with MNIST. As model we employ a UNet, with $lr = 1e-5$. We train for 50k steps and finetune for 10k. We fix the $\lambda_t = 100$ for the Div loss if using cosine distances, and $\lambda_t = 0.001$ if using euclidean. For the Fid loss we use $\lambda_t = 0.01$.

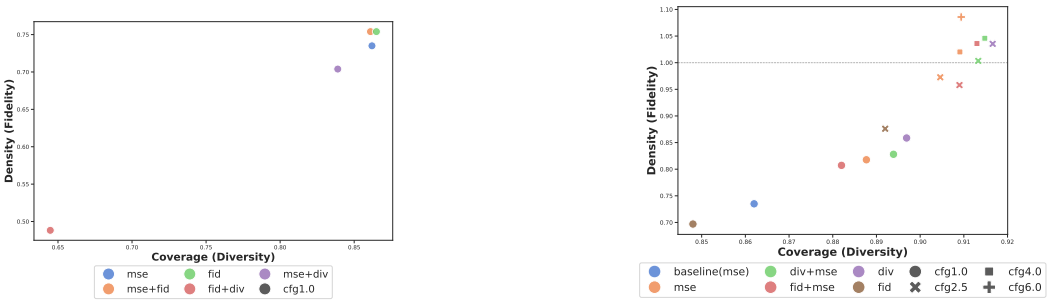


Figure 5: **Left:** Training 50k steps from scratch on the MNIST dataset. Metrics are computed sampling 10k synthetic images using cfg=1.0. **Right:** Model evaluation on the MNIST dataset. Baseline corresponds to 50k training steps with MSE. The other points are 10k further steps of finetuning with different losses. Metrics are computed sampling 10k synthetic images with different guidance scales.

D.2 Experiments in RAE space

Fine-tuning with MSCOCO. The base model was pretrained on the CC3M dataset (Sharma et al., 2018) with 2M samples for 240k steps. Its performance on the MSCOCO2014 validation set is presented in the baseline row of Table 5.1. Subsequently, we fine-tuned this baseline for 50k steps using various combinations of MSE, Diversity, and Fidelity losses. To maintain consistency, all fine-tuning iterations employed a constant learning rate of 10^{-6} and identical hyperparameters. For evaluation, we generated a synthetic dataset of 10k image-prompt pairs using a CFG scale of 3.0 and 50 sampling steps using Euler sampler. The λ_t hyperparameter is fixed to $\lambda_t = 100$ for the Div loss, and to $\lambda = 0.01$ for the Fid loss. The only criterion used is to keep its scale close to MSE. See Table 4 for a small ablation.

Fine-tuning with ImageNet. We take as base the pre-trained Flow Matching model released in the Representation Autoencoder paper (Zheng et al., 2025). We fine-tune for 10-50k further steps, using a linear (CondOT) probability path (ICPlan) as in the original work. We use a $lr = 1e-5$, gradient accumulation of 1, 4 GPUs and a local batch size of 64. For evaluation we sample 50k images with various CFG scales, and compare against ImageNet test set. Topological loss scales are fixed too $\lambda_t = 100$ for the Div loss, $\lambda_t = 0.01$ for the Fid loss.

Qualitative Analysis on ImageNet. We perform a qualitative analysis on several images with Imagenet. Pretraining for 200k steps with SiT/XL-2 architecture, using an effective batch size of 1024. We continue with MSE for a further 5k steps, and as well fine-tune with Vietoris Rips filtration loss + MSE to obtain two 205k steps models (MSE vs div + MSE). FiD evaluation does not show



(a) Standard MSE flow matching samples (L_{MSE})



(b) Diversity loss-enhanced flow matching samples ($L_{MSE} + L_{DIV}$)

Figure 6: Qualitative comparison of generative samples after 5k steps finetuning with MSE vs MSE + DIV after 200k steps of MSE pre-training with imagenet. The top row shows baseline flow matching results, while the bottom row displays samples from our proposed diversity-regularized variant.

improvement (baseline FiD after 200k steps is bad, around 30). Coverage increases very lightly. Still, we provide some qualitative samples for reference, see Figure 6.

E Computational performance

Computing persistence diagrams implies a computational overhead with respect to MSE. While we tried to keep computations as efficient as possible, this overhead is not negligible. In the Vietoris-Rips loss we only need to compute persistent homology once per batch, for which we use the very efficient `ripser` package which runs on `C++`. But in the case of the cubical complex loss, there is a computation per sample in the batch and per channel, which makes it more costly. For small MNIST models, computing gradients for Fidelity and Diversity losses takes roughly 27x and 6x the GPU time of MSE, respectively. For bigger models, the disadvantage of topological losses is smaller, as most of the computation time rests not on the PH computation but on the gradient propagation through the weights. For the RAE models we trained , Fidelity and Diversity losses required only 1.15x and 1.28x the time of MSE. When we apply PCA reduction to compute the Vietoris-Rips filtration, the cost of computing the Diversity loss is reduced considerably, to 1.03x of the MSE cost. Finally, notice the Fidelity loss implementation for RAE models obtains better numbers because we only compute cubical complexes on RAE channels. See Table E.

Models	MSE	Fidelity	Diversity	Diversity (PCA)
MNIST Models	**17.12 ± 0.0050**	466.97 ± 6.4075	119.38 ± 2.5032	-
Text-to-Image with RAE	**575.54 ± 0.0147**	663.99 ± 0.0317	743.90 ± 0.2465	594.33 ± 0.0465

Table 5: Average time in ms required to compute a backward pass through MSE and both topological losses with respect to the model parameters.

Despite recent and promising advances for more efficient persistent homology computation (Graf et al., 2025), which we still did not adapt to our method, we acknowledge there will always be a computational overhead with respect to raw mse, and that this a big limitation. Nonetheless, we still believe our method to be useful because of the following reasons:

1. **Regularization requires few steps**. These methods are not thought for large-scale pre-training vs mse. Experiments (see Section 5) show finetuning for as little as 5-10k steps already produces results. The method can be specially useful when mse training hits a plateau.
2. **Scaling with respect to model size**. Persistence diagram computation mainly depends on batch size and input resolution. Computational costs decompose linearly: $forward(ins) + pd_{computation}(outs) + backward(outs)$. As model size grows, the PD computation part of the equation becomes negligible. For reference, on MNIST/CIFAR experiments the div

loss cost is 2-6$\times$ of the mse cost (model size $35M$); on Coco experiments, it is $1.3\times$ of mse cost (model size, $875M$).

3. **Scaling with respect to input resolution**. As we increase resolution, cubical complex computation grows exponentially. However, that is not the case for the Vietoris-Rips filtration computation, as we only need to compute pairwise distances. The cost here grows exactly as in mse, see Table 6.
4. **Potential for further computational improvements**. We believe some of the computational overheads of our methods can be optimized. For example: the diversity loss computation shouldn't be greatly dependent on bs, as we only need to compute the VR filtration once per batch. However, since the `ripser` package does not return critical points (i.e.: the vertices that cause the birth or death of a feature), we are forced to find them via a distance matrix computation, which has cost $O(n^2)$ w.r.t. batch size. We are already studying how to deal with this by modifying `ripser`'s C++ source code. With respect to the cubical complex computation, subsampling could be greatly effective, given that we only need high level features. In any case, these are not fundamental limitations of our method. Further, there is new and exciting research in much more efficient persistent homology computational methods, as the cited (Graf et al., 2025).

Input Resolution	MSE	Fidelity Loss	Div Loss
16×16	9.4709 ± 0.0505	30.4900 ± 0.1627	20.1580 ± 0.1194
28×28	16.5038 ± 0.0071	69.5557 ± 0.5838	27.5492 ± 0.0542
64×64	71.4434 ± 0.0262	864.3511 ± 17.1775	82.7722 ± 0.0653
128×128	318.0614 ± 0.0428	16357.1667 ± 555.0040	330.7278 ± 0.0566

Table 6: Performance metrics across different input resolutions. Computed averaging backprop. time over 100 batches, batch size fixed at 8, model size $35M$.

Batch Size	MSE	FID Loss	Div Loss
8	16.5038 ± 0.0071	69.5557 ± 0.5838	27.5492 ± 0.0542
16	28.3166 ± 0.0256	133.3280 ± 1.2518	49.5385 ± 0.3393
32	54.4998 ± 0.0186	256.7658 ± 3.1695	96.1528 ± 0.6335
64	104.9235 ± 0.0181	518.0332 ± 4.7115	196.7253 ± 2.0524

Table 7: Scaling table with respect to batch size (computed averaging backprop. time over 100 batches, resolution fixed at 28×28, model size $35M$).