GLASS: A Differentiable Geometric Alignment Layer on Manifolds and Graphs via Learned Slices and Soft Optimal Transport

Nizar Benbouchta
Université Paris-Cité
nizarbenbouchta@gmail.com

Abstract

We introduce **GLASS** (Geometric Learned Alignment via Soft Slices), a lightweight, fully differentiable layer for aligning non-Euclidean representations on compact manifolds and graphs. GLASS learns intrinsic one-dimensional projections, performs entropic one-dimensional optimal transport (a smooth soft sorting), and lifts the soft plan back to the ambient geometry to produce a geometry-aware loss and a reusable coupling. We prove equivariance to isometries and automorphisms, approximation guarantees to manifold W_2 and a diffusion-feature surrogate on graphs up to curvature, truncation, and entropy terms, uniform generalization bounds, and end-to-end differentiability. On synthetic spheres and stochastic block model graphs, GLASS achieves compelling accuracy and efficiency with near-linear runtime via banded Sinkhorn, and outperforms naive baselines when gate widths are tied or learned relative to capacity. The layer is plug-and-play and directly applicable as a geometry-aware alignment head inside foundation-model pipelines operating over manifolds and graphs. Code to reproduce all experiments in this paper are available at: `https://github.com/Nizben/glass`

1 Introduction

Modern representation learning increasingly targets non-Euclidean domains: embeddings on spheres or Lie groups, and data living on graphs or meshes. Aligning such representations across views, modalities, or timesteps is central to pretraining and adaptation, yet common objectives like cosine similarity or mean-squared error ignore intrinsic geometry and do not expose an explicit transport plan that can be reused downstream.

We propose **GLASS**, a simple, fast, theory-backed *geometric alignment layer*. First, GLASS learns geometry-aware one-dimensional scores on the manifold or graph. Second, it aligns these scores by solving entropic optimal transport in one dimension, yielding a smooth soft coupling. Third, it lifts this plan to the ambient non-Euclidean space to compute a geometry-aware loss. The result is a differentiable objective and a reusable coupling that preserves symmetry structure and scales near-linearly in practice.

Positioning and relevance. Unlike Sliced Wasserstein (SW), which averages costs of many random projections and discards the plan, GLASS *learns* intrinsic projections adapted to the data geometry and *keeps* a coupling useful for supervision or consistency. Unlike Gromov–Wasserstein (GW) and Fused GW (FGW), which optimize a quadratic objective over pairwise structures with higher per-iteration cost and delicate differentiation, GLASS uses a low-cost ordering device (learned 1D scores) and a soft plan lifted back to manifold or diffusion-feature distances. This makes GLASS a practical *alignment head* for foundation models whose intermediate features live on non-Euclidean spaces.

Contributions.

- A general alignment layer for compact manifolds and graphs with learned intrinsic projections, soft one-dimensional OT, and a lifted non-Euclidean loss that returns both a scalar objective and a coupling.
- Theoretical statements covering symmetry preservation, approximation to W_2 on manifolds and a diffusion-feature surrogate on graphs, uniform generalization, differentiability, and near-linear complexity via banded kernels.
- Synthetic evaluations on S^2 and SBMs validating capacity–coverage trade-offs, diffusion-scale behavior, and empirical near-linear scaling, supporting GLASS as a geometry-aware head for non-Euclidean foundation-model stacks.

2 Related Work

Optimal transport in machine learning. OT provides a principled way to align distributions via couplings, with entropic regularization enabling scalable, differentiable computations [2, 7]. Beyond distances, OT-based layers and losses have been integrated into deep nets as differentiable modules for matching, sorting, and assignment.

Sliced and generalized sliced Wasserstein. SW projects to $\mathbb{R}$ along many directions, solves 1D OT, and averages costs [4, 9]. While efficient, SW typically uses random linear projections and reports only a scalar, discarding the coupling. GLASS differs by *learning intrinsic projections* (log maps on manifolds, diffusion features on graphs), *retaining a plan*, and *lifting* to ambient non-Euclidean costs.

Gromov–Wasserstein and Fused GW. GW aligns structures by matching pairwise distances, and FGW augments with features [5, 8, 10]. These methods can achieve strong structural alignment but are heavier per iteration and can be brittle under gradients. GLASS offers a complementary regime that trades exact structure matching for stability and speed by combining learned orderings with soft (e.g., [3, 6]) 1D OT and a geometric lift.

Diffusion geometry for graphs. Diffusion maps embed graphs using heat kernels and provide a multi-scale structural representation [1]. GLASS builds intrinsic projections from Chebyshev-filtered diffusion features and defines lifted losses directly in that spectral space, with the diffusion time τ controlling the structural scale.

Non-Euclidean architectures and FM alignment. Equivariant layers on manifolds and GNNs on graphs motivate geometry-aware objectives that respect symmetries. GLASS is designed as an FM-compatible alignment head that preserves isometries/automorphisms and exposes a reusable coupling for supervision and consistency constraints.

3 Preliminaries

We consider either a compact Riemannian manifold $(\mathcal{M}, g)$ with geodesic distance $d_{\mathcal{M}}$ or a graph $\mathcal{G}$ with Laplacian L and diffusion features. Given empirical measures $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}$, our goal is to construct a differentiable coupling $P_\gamma \in \mathbb{R}^{n\times n}$ and a geometry-aware loss.

Entropic one-dimensional optimal transport. For real-valued scores $s, t \in \mathbb{R}^n$ with quadratic cost $C_{ij} = (s_i - t_j)^2$, the entropic OT problem is

$$P_\gamma \;=\; \arg\min_{P\in\mathcal{U}} \langle P, C\rangle \;+\; \gamma \sum_{ij} P_{ij} \log P_{ij}, \qquad \mathcal{U} = \{\, P \geq 0,\ P\mathbf{1} = \tfrac{1}{n}\mathbf{1},\ P^\top\mathbf{1} = \tfrac{1}{n}\mathbf{1}\,\}. \quad (1)$$

Sinkhorn scaling yields a unique, smooth solution and stable Jacobian-vector products for backpropagation.

4 The GLASS Layer

GLASS composes three steps: learned intrinsic projections, soft one-dimensional alignment, and a lifted ambient loss.

Learned intrinsic projections. On manifolds, pick anchors $a_k \in \mathcal{M}$ (within injectivity radius) and tangent directions $u_k \in T_{a_k}\mathcal{M}$, and define

$$\varphi(x;\theta) = \sum_{k=1}^{K} \alpha_k \, \langle u_k, \mathrm{Log}_{a_k}(x) \rangle \, w_k(x), \quad \alpha \in \Delta_K, \; w_k \geq 0, \; \sum_k w_k = 1. \tag{2}$$

On graphs, with diffusion features $\psi_\tau(v) = (e^{-\tau\lambda_\ell} u_\ell(v))_{\ell=1}^{p}$, set

$$\varphi(v;\theta) = \langle w, \psi_\tau(v) \rangle, \qquad w \in \mathbb{R}^p. \tag{3}$$

Soft one-dimensional alignment. Compute $s_i = \varphi(x_i;\theta)$ and $t_j = \varphi(y_j;\theta)$ and obtain $P_\gamma(s,t)$ via entropic OT in one dimension.

Lifted ambient loss. Lift the soft plan to a geometry-aware objective:

$$\mathcal{L}_{\text{geo}}(\theta) = \frac{1}{n} \sum_{ij} P_{\gamma,ij} \, d_{\mathcal{M}}(x_i, y_j)^2 \qquad \text{(manifolds)}, \tag{4}$$

$$\mathcal{L}_{\text{spec}}(\theta) = \frac{1}{n} \sum_{ij} P_{\gamma,ij} \, \|\psi_\tau^X(i) - \psi_\tau^Y(j)\|^2 \qquad \text{(graphs)}. \tag{5}$$

Gradients flow through projections, Sinkhorn, and distances, enabling end-to-end training.

Complexity via banded kernels. Sorting induces near-diagonal structure in 1D. Truncating the Gaussian kernel $K = \exp(-C/\gamma)$ to a band of radius $r = \sqrt{-\gamma \log \varepsilon}$ yields $\mathcal{O}(\mathrm{nnz})$ Sinkhorn updates with $\mathrm{nnz} = \mathcal{O}(n \, \mathrm{band\_width})$, delivering near-linear empirical time while controlling bias.

5 GLASS as a Geometry-Aware Alignment Head

We view GLASS as a head attached to a backbone that outputs non-Euclidean features. Let f_θ be the backbone producing sets $X = \{x_i\}$ and $Y = \{y_j\}$ on a manifold or graph domain from two related inputs (e.g., views or timesteps). The GLASS head returns a coupling P_γ and a lifted loss:

$$(\mathcal{L}_{\text{geo}}, P_\gamma) = \text{GLASS}(X, Y; \theta_{\text{head}}), \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \, \mathcal{L}_{\text{geo}} \;\text{ or }\; \lambda \, \mathcal{L}_{\text{spec}}. \tag{6}$$

Gradients propagate through P_γ (via Sinkhorn) into both the head and backbone:

$$\frac{\partial \mathcal{L}_{\text{geo}}}{\partial \theta} = \sum_{ij} \frac{1}{n} \left(\frac{\partial P_{\gamma,ij}}{\partial \theta} \, d^2(x_i, y_j) + P_{\gamma,ij} \, \frac{\partial d^2(x_i, y_j)}{\partial \theta} \right).$$

An annealing schedule $\gamma_t = \max(\gamma_{\min}, \gamma_0 \alpha^t)$ hardens the plan during training; the band radius can follow $r_t = \sqrt{-\gamma_t \log \varepsilon}$ to keep truncation error stable. For multi-view settings $\{X^{(v)}\}_{v=1}^{V}$, one can sum GLASS losses across pairs:

$$\mathcal{L}_{\text{MV}} = \sum_{v<w} \lambda_{vw} \, \mathcal{L}_{\text{geo}}\Big(X^{(v)}, X^{(w)}\Big),$$

which enforces view consistency while respecting symmetries. Complexity per step is $\mathcal{O}(nK)$ (projections on manifolds) or $\mathcal{O}(np)$ (graphs) plus $\mathcal{O}(\mathrm{nnz})$ for banded Sinkhorn.

6 Theoretical Properties

We summarize core statements; detailed sketches appear in the Appendix.

Symmetry preservation. If anchors and gates relabel under isometries $g \in \mathrm{Isom}(\mathcal{M})$ as $a_k \mapsto g a_k$ and $u_k \mapsto D g_{a_k} u_k$, then $\varphi(gx; \theta^g) = \varphi(x; \theta)$ and the lifted loss is invariant. On graphs, if diffusion features transform orthogonally within eigenspaces under automorphisms, then the lifted spectral loss is invariant.

Approximation on manifolds. On compact regions covered by normal charts with bounded curvature and Lipschitz projections, there exist anchors and parameters such that

$$\mathcal{L}_{\text{geo}}(\theta^\star) \;\le\; W_2^2(\mu,\nu) \;+\; \varepsilon \;+\; C_\kappa r^4 \;+\; C_\gamma \gamma,$$

where r bounds chart radii and γ is the entropic temperature.

Approximation on graphs. Let d_{diff} denote diffusion distance at time τ and η the truncation error from using p eigenpairs or Chebyshev filters. Then for suitable (p, τ, w) and small γ,

Let D_X, D_Y be the pairwise diffusion distance matrices at time τ for graphs G_X, G_Y. We write

$$\mathrm{GW}^2_{\text{diff}}(G_X, G_Y) := \min_{P \in \mathcal{U}} \left\| D_X - P D_Y P^\top \right\|_F^2,$$

a GW-like objective where within-space distances are diffusion distances.

$$\mathcal{L}_{\text{spec}}(\theta^\star) \;\le\; \mathrm{GW}^2_{\text{diff}}(\mathcal{G}_X, \mathcal{G}_Y) \;+\; \varepsilon \;+\; C_{\text{trunc}}\eta \;+\; C_\gamma \gamma.$$

Lipschitz dependence and generalization. On bounded sets, the map $C \mapsto P_\gamma(C)$ is Lipschitz, yielding a uniform bound

$$\sup_\theta \left| \widehat{\mathcal{L}}_n(\theta) - \mathcal{L}(\theta) \right| \;\le\; c_1 \, \mathfrak{R}_n(\mathcal{F}) + c_2 \sqrt{\tfrac{\log(1/\delta)}{n}},$$

with constants depending on geometric and entropic smoothness.

Banded kernel bias. Truncating K at radius r introduces a controlled plan perturbation and a bounded lifted-loss gap that decays with $e^{-r^2/\gamma}$.

7 Synthetic Experiments

We evaluate GLASS on S^2 and on stochastic block model (SBM) graphs. All runs use banded Sinkhorn; reported metrics are final lifted losses. All experiments run on an NVIDIA T4 16GB GPU.

7.1 S^2 Capacity and Gate Width

Table 1: S^2 GLASS final loss vs. anchors K at $\gamma_0 = 0.1$. Tied gates help at small K.

K	Baseline	Tied	Δ (Tied–Base)
8	2.7914	**2.6594**	**-0.1320**
16	2.8307	**2.8201**	**-0.0106**
32	2.8320	**2.8312**	**-0.0008**

7.2 S^2 Chart Radius Ablation

Table 2: S^2 GLASS final loss vs. chart radius (degrees).

Radius (deg)	10	20	30	40
Loss	2.8319	2.8308	2.8259	**2.8112**

7.3 S^2 Runtime Scaling

See **Figure 1**

7.4 SBM Graphs Diffusion Scale

See **Figure 2**

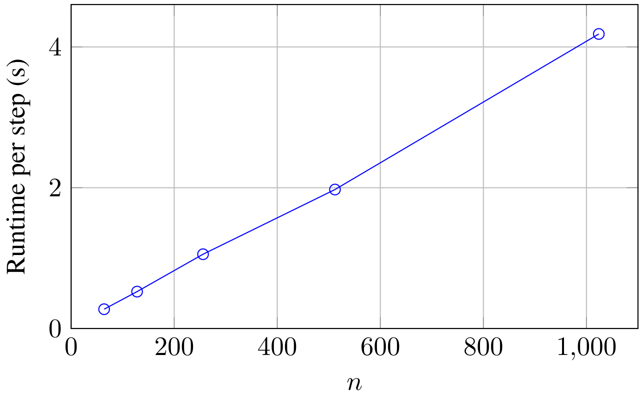

Figure 1: S^2 banded Sinkhorn shows near-linear scaling (power exponent ≈ 0.98).

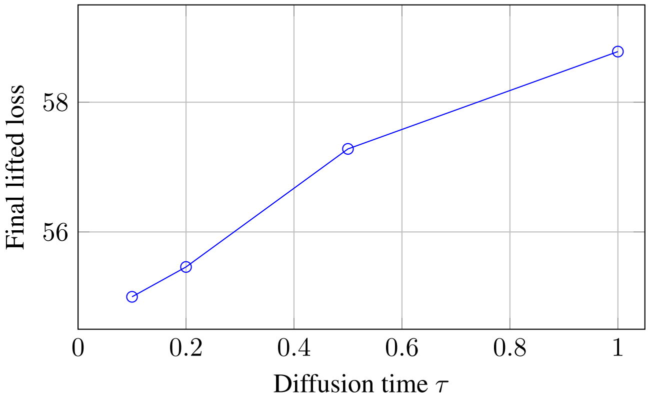

Figure 2: SBM graphs: smaller τ preserves more structural detail and yields lower loss.

8 Limitations

GLASS provides approximation-style guarantees rather than exact optimality. On manifolds, assumptions include chart coverage and moderate curvature; on graphs, results depend on diffusion truncation error and the choice of diffusion time. Entropic bias trades approximation for smooth gradients; annealing mitigates bias at the cost of variance. Computing all ambient distances can be heavy for large n, but minibatching and blocking alleviate this.

9 Conclusion

We introduced **GLASS**, a differentiable geometric alignment layer that learns intrinsic one-dimensional projections, solves entropic one-dimensional optimal transport to obtain a soft coupling P_γ, and *lifts* that plan to non-Euclidean distances to define the training objective. In contrast to objectives that only compare pooled statistics or discard transport structure, GLASS returns both a scalar loss and an explicit coupling that can be reused for supervision, warping, or consistency constraints inside foundation-model pipelines operating on manifolds and graphs. Our theory formalizes why this works: symmetry preservation under isometries and automorphisms, approximation guarantees to manifold W_2 and a diffusion-feature surrogate on graphs (up to curvature, truncation, and entropy terms), and stability via entropic smoothing. Empirically, banded Sinkhorn yields near-linear scaling, and simple capacity controls (anchors, gate widths, diffusion scale) match the qualitative predictions of the analysis.

Practical takeaways. For practitioners, GLASS is a drop-in alignment head that interfaces with a backbone through $(\mathcal{L}_{\text{geo}}, P_\gamma)$ or $(\mathcal{L}_{\text{spec}}, P_\gamma)$. We find that (i) annealing γ stabilizes early training while converging to crisper plans, (ii) choosing the band radius $r = \sqrt{-\gamma \log \varepsilon}$ keeps bias controlled and time nearly linear, and (iii) tying or learning gate widths against anchor count improves coverage on curved domains. These heuristics make the layer an attractive replacement for cosine/MSE when features live on non-Euclidean spaces.

Outlook. The present work opens several avenues: learning projection families beyond linear functionals in normal coordinates and diffusion features; principled selection of diffusion time and spectral truncation for graphs; adaptive banding schemes with explicit bias control; extensions to non-uniform marginals and partial matching; and integrating GLASS with equivariant backbones on S^2, Lie groups (e.g., $\mathrm{SE}(3)$), and large-scale graph encoders. Beyond synthetic settings, we plan to evaluate GLASS on panoramic vision, protein structures, and multi-view scene graphs, where the reusable coupling can act as a supervision signal or a curriculum for alignment-heavy pretraining. Overall, GLASS provides a simple, theory-grounded, and scalable mechanism to bring *geometry-aware alignment* into modern foundation-model stacks for non-Euclidean data.

References

[1] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[3] Marco Cuturi, Obay Teboul, Olivier Vert, and Jean-Philippe Vert. Differentiable sorting and ranking via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Soheil Kolouri, Philip E Pope, Charles E Martin, and Gustavo K Rohde. Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

[6] Gonzalo Mena, David Belanger, Scott Mallis, et al. Learning latent permutations with Gumbel-Sinkhorn networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[7] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[8] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov–wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning (ICML)*, pages 2664–2672, 2016.

[9] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 435–446, 2011.

[10] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 12(10):196, 2019.

Appendix

A Proof Sketches for Theoretical Statements

A.1 Symmetry preservation. On manifolds, assume anchors and directions transform under $g \in \mathrm{Isom}(\mathcal{M})$ by $a_k \mapsto ga_k$ and $u_k \mapsto Dg_{a_k}u_k$, with gates relabeled accordingly. In normal coordinates, $\mathrm{Log}_{ga_k}(gx) = Dg_{a_k}\mathrm{Log}_{a_k}(x)$, so $\langle u_k, \mathrm{Log}_{a_k}(x)\rangle = \langle Dg_{a_k}u_k, \mathrm{Log}_{ga_k}(gx)\rangle$. Hence $\varphi(gx; \theta^g) = \varphi(x; \theta)$, C is invariant, and so are P_γ and the lifted loss. Graphs follow from orthogonal invariance of diffusion features within eigenspaces under automorphisms.

A.2 Manifold approximation bound. Cover supports by normal charts of radius $r < \mathrm{inj}(\mathcal{M})$. In a chart, $d^2_{\mathcal{M}}(x, y) = \|x - y\|^2 + \mathcal{O}(\kappa r^4)$. Choose K anchors and linear functionals so that the resulting score orders approximate the support pattern of an ε-optimal W_2 plan. In 1D with distinct scores, $P_\gamma \to P_{\text{perm}}$ as $\gamma \downarrow 0$; thus the induced coupling cost is within $\mathcal{O}(\gamma)$ of the 1D optimum. Summing over charts and accounting for gate mixing gives

$$\mathcal{L}_{\text{geo}} \le W_2^2 + \varepsilon + C_\kappa r^4 + C_\gamma \gamma.$$

A.3 Graph surrogate consistency. Let d_{diff} be diffusion distance at time τ, and let $\eta(p, \tau)$ be the error from p-term eigen truncation or Chebyshev filtering. The lifted objective equals the expected squared Euclidean distance in the diffusion embedding under P_γ. For suitable (p, τ, w) and small γ, this matches a diffusion-distance GW-like surrogate within $\varepsilon + C_{\mathrm{trunc}}\eta + C_\gamma\gamma$.

A.4 Lipschitz dependence and generalization. Entropic OT with uniform marginals is strongly convex in P, and the Sinkhorn map is contractive in dual variables, yielding a local Lipschitz bound $\|P_\gamma(C) - P_\gamma(C')\|_1 \le L_\gamma \|C - C'\|_\infty$, with L_γ polynomial in $1/\gamma$ for fixed n. Composing with Lipschitz projections yields standard uniform generalization bounds via Rademacher complexity.

A.5 Banded kernel bias and feasibility. Truncate $K = \exp(-C/\gamma)$ at radius r so off-band entries are at most $\varepsilon = e^{-r^2/\gamma}$. Running Sinkhorn with dual scalings preserves marginals. One obtains $\|P_\gamma^{\text{band}} - P_\gamma\|_1 \le c_1 n\varepsilon + c_2 e^{-r^2/\gamma}$; the lifted-loss gap is bounded by this perturbation times $\max_{ij} d^2(x_i, y_j)$, so bias decays exponentially in r^2/γ.