On Relation-Aware Slicing in Cross-Domain Alignment

Dhruv Sarkar[1] **Aprameyo Chakrabartty**[1] **Anish Chakrabarty**[2] **Swagatam Das**[2]
[1]Indian Institute of Technology, Kharagpur [2]Indian Statistical Institute, Kolkata

Abstract

The Sliced Gromov-Wasserstein (SGW) distance, aiming to relieve the computational cost of solving a non-convex quadratic program that is the Gromov-Wasserstein distance, utilizes projecting directions sampled uniformly from unit hyperspheres. This slicing mechanism incurs unnecessary computational costs due to uninformative directions, which also affects the representative power of the distance. However, finding a more appropriate distribution over the projecting directions (*slicing distribution*) is often an optimization problem in itself that comes with its own computational cost. In addition, with more intricate distributions, the sampling itself may be expensive. As a remedy, we propose an optimization-free slicing distribution that provides fast sampling for the Monte Carlo approximation. We do so by introducing the Relation-Aware Projecting Direction (RAPD), effectively capturing the pairwise association of each of two pairs of random vectors —each following their ambient law —in the statistical sense. This enables us to derive the Relation-Aware Slicing Distribution (RASD), a location-scale law corresponding to sampled RAPDs. Finally, we introduce the RASGW distance and its variants, e.g., IWRASGW (Importance Weighted RASGW), which overcome the shortcomings experienced by SGW. We theoretically analyze its properties and substantiate its empirical prowess using extensive experiments on various alignment tasks.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Optimal Transport (OT) theory serves as a cornerstone of contemporary machine learning. A notable outcome of OT theory is the Wasserstein distance (WD), widely used in various learning applications [Khamis et al., 2024]. Metrizing weak convergence in a compact metric space, WD offers a strong measure of discrepancy between probability distributions of the same dimensionality. However, it is the associated computational complexity that depreciates its appeal. Given n replicates from a d-variate distribution, calculating the empirical WD incurs a space complexity of $\mathcal{O}(n^2)$. Moreover, the associated rate of convergence cannot escape the curse of dimensionality ($\mathcal{O}(n^{-\frac{1}{d}})$) if the underlying distributions have arbitrary smoothness. Among the several remedies proposed [Cuturi, 2013, Altschuler et al., 2017, Guo et al., 2020], it is the Sliced Wasserstein Distance (SWD) [Bonneel et al., 2015] that achieves linear space complexity without sacrificing statistical or topological properties [Nadjahi et al., 2020]. The underlying trick involves linearly projecting observations onto $\mathbb{R}$, which renders the remaining transportation merely sorting and matching the data points. The projected distributions incur a sample complexity of $\mathcal{O}(n^{-\frac{1}{2}})$ under WD, making SWD impervious to the dimensionality curse.

The correspondence problem becomes particularly challenging when the two underlying distributions originate from distinct and often incomparable metric spaces. In such cases, the notion of mass transportation breaks down, as there is no common geometric medium through which the two spaces can be meaningfully compared. Invoking a L^p-relaxation to the well-known Gromov-Hausdorff metric, the Gromov-Wasserstein distance (GW) [Mémoli, 2011] overcomes this issue. Instead of calculating the average cost of transportation, it compares the geometry of the underlying spaces. Precisely, it is the expected departure from isometry between them. While theoretically rich, computing the GW distance runs into roadblocks more demanding than WD. Typically, it requires solving a quadratic assignment problem (QAP) that is in general NP-hard to compute. To circumvent this, one may approximately calculate a linear lower bound as proposed by Chowd-

hury and Mémoli [2019]. So far, the most reliable solution to the problem remains Entropic regularization [Scetbon et al., 2022, Rioux et al., 2024]. It follows earlier estimation techniques using regularized versions of entropy, which can be solved more efficiently by repeatedly employing Sinkhorn projections [Solomon et al., 2016, Peyré et al., 2016]. However, among the many methods inspired by similar techniques in OT, the most attractive in terms of ease of implementation and theoretical elegance is still *slicing*. The Sliced GW (SGW) [Vayer et al., 2019] again uses Radon transforms to project the measures to univariate distributions, followed by computing the GW-distance between them. Taking the expectation over all slices, it emerges as a proxy for GW.

Regardless, the slicing has its drawbacks. Firstly, as also observed in WD, the complexity of the function evaluations corresponding to the projections may increase exponentially with the data dimension (e.g., LIPO for Max-slicing [Nietert et al., 2022]). Moreover, the outer expectation over the slices typically becomes intractable [Vayer et al., 2019] and has to be replaced by a Monte-Carlo (MC) approximate. The approximation is done by sampling uniformly from the unit hypersphere ($\mathbb{S}^{d-1}$). However, this can be highly computationally inefficient since randomly sampled directions may include highly *uninformative* ones. As GW encapsulates the similarity between pairwise distances from the underlying spaces, the geometric structure must remain intact relative to one another after projections. This is imperative to a proxy distance reflecting the same properties as GW. Clearly, there exist projecting directions that perturb pairwise distances (see Section 4) and, as a result, distort the alignment of cross-domain observations. We only focus on the directions $\Theta \subset \mathbb{S}^{d-1}$ that maintain the geometric association between the two spaces.

We prioritize the concentration of such informative *relation-aware* projecting directions while designing the slicing distribution. They are defined as the directions that capture the intra-pair distance of each pair sampled from the respective input measures. More specifically, we take the unit bisecting vector of the displacement vectors of each pair (see Figure 1). Based on what we call the *Relation-aware Slicing Distribution* (RASD), we introduce a novel and efficient slicing technique into cross-domain alignment metrics such as GW. We show that sampling from the same is also quite simple, as it involves no additional optimization and only relies on input samples from both input distributions, which are assumed to be easy to sample from in most machine learning (ML) tasks.

2 RELATED WORKS

To better appreciate our propositions, let us first go through the ever-expanding list of slicing techniques addressing diverse objectives, such as minimizing slicing complexity or projection variability. Notably, GSW [Kolouri et al., 2019] and ASW [Chen et al., 2020] extend the approach to non-linear projections to represent complex associations between distributions. In search of the most efficient slice in terms of information, Deshpande et al. [2019] and Nguyen et al. [2021] propose Max-SW and DSW distances, respectively. Most of these methods involve precise optimizations for finding a class of suitable directions. For example, the HSW [Nguyen et al., 2023] undergoes a selection of *bottleneck* slices that linearly combine to simulate projections. HHSW [Nguyen and Ho, 2024] follows a similar combining mechanism, however, non-linearly. At the end of parameter-free techniques, EBSW [Nguyen and Ho, 2023] relies on slicing distributions whose density is proportional to an energy function of the projected univariate WD.

We emphasize that all such methods originate to cater to the OT problem, out of which only the vanilla slicing [Bonneel et al., 2015] (similarly, Max-SW) is adopted for cross-domain alignment to date. The first obstacle in doing so is the involvement of distributions from disparate spaces, which immediately renders a projection suitable for one measure useless for the other. Even recent methods such as Piening and Beinert [2025] resort to the fused GW and its lower bounds, succumbing to this issue. Secondly, the alignment problem being fundamentally different from transportation makes the notion of *informativeness* of a projecting direction incompatible in the case of GW. Keeping in mind the cross-domain problem, where preserving the geometry is key, we redefine what it means to be 'informative' based on relation-awareness. The work that relates to our approach the best is RPSW [Nguyen et al., 2024], which facilitates Monte-Carlo sampling by introducing normalized differences between replicates from the two distributions while choosing slices. For a detailed discussion on the adoption of existing slicing techniques in the purview of the cross-domain problem, we refer the reader to Appendix B.

Contributions. In summary, our contributions are as follows.

- We introduce a novel optimization-free slicing mechanism based on what we call the *relation aware* projecting directions (RAPD). This involves randomly perturbing the bisecting unit vector of the displacement between sample pairs from different input measures. Based on RAPD, we define the relation-aware slicing distribution (RASD), from which drawing ran-

dom projecting sample directions is both easy and efficient.

- Based on RA slices, we develop two new variants of Gromov-Wasserstein distances that preserve metric topological and statistical properties of GW besides overcoming computational drawbacks. The first is termed *relation aware sliced Gromov-Wasserstein* (RASGW), which replaces uniform sampling in SGW by RASD. The second variant, *importance-weighted relation aware sliced GW* (IWRASGW), is defined as the expectation of a weighted ensemble of several randomly projected distances using RAPDs. We explore the theoretical attributes of both RASGW and IWRASGW, including topological, statistical, and computational properties. More specifically, we assess the metric properties, their relation to other SGW variants, incurred sample complexity, and computational complexities when utilizing Monte Carlo techniques.
- Comparing the performance of RASGW and IWRASGW with other slicing mechanisms poses new challenges. This is primarily due to the absence of existing benchmarks in cross-domain alignment tasks. In a first, we adapt several OT projections, such as SW, Max-SW, DSW, RPSW, and EBSW in the GW framework. We empirically show that in generative tasks involving unalike spaces, RASGW and IWRASGW surpass the competition on diverse datasets. In the process, we modify GW-GANs [Bunne et al., 2019] and GW-Autoencoders [Nakagawa et al., 2023] using our metrics, which reveal that our metric is indeed superior in capturing the relational aspects between datasets at minimal computational overhead.

We defer the proofs of key results and additional experimental details to the Supplementary material. All codes for our experiments, along with execution instructions, are placed in the repository[1].

3 BACKGROUND

In this section, we review the preliminary concepts needed to develop our framework, namely the Gromov-Wasserstein distance, the sliced Gromov-Wasserstein distance, and also elucidate the idea of relational discrepancy.

Notations. We denote by $\mathcal{P}(\mathcal{X})$ the set of all probability measures supported on a space $\mathcal{X}$. For $p \geq 1$, $\mathcal{P}_p(\mathcal{X})$ denotes the set of all probability that have finite p-moments. Given $d \geq 2$, we denote $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d \mid ||\theta||_2^2 = 1\}$ and $\mathcal{U}(\mathbb{S}^{d-1})$ as the unit hyper-sphere and the uniform law supported on it. For any two sequences a_n and b_n, the notation $a_n = \mathcal{O}(b_n)$ means that $a_n \leq Cb_n$ for all $n \geq 1$, where C is some universal constant. Moreover, we use the notation $v//v'$ to denote that the two vectors v and v' are parallel.

Gromov-Wasserstein Distance. Let $(\mathcal{X}, c_X, \mu)$ and $(\mathcal{Y}, c_Y, \nu)$ be two metric measure (mm) spaces, where c_X and c_Y are the endowed metrics. For simplicity, we consider without loss of generality that both $\mathcal{X}$ and $\mathcal{Y}$ are Polish. We sparingly refer to the mm spaces as $\mathscr{X}$ and $\mathscr{Y}$, respectively. Let $\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi(\cdot \times \mathcal{Y}) = \mu, \pi(\mathcal{X} \times \cdot) = \nu\}$ be the set of all couplings between μ and ν. Then the p-GW distance [Mémoli, 2011] between the two mm spaces is defined as follows.

$$\mathrm{GW}_p^p(\mu, \nu) = \inf_{\pi \in \Pi} \mathbb{E}_{\pi \otimes \pi}[c_p(\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}')], \tag{1}$$

where $c_p(\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}') = |c_X(\mathbf{x}, \mathbf{x}') - c_Y(\mathbf{y}, \mathbf{y}')|^p$ denotes the realized distortion given samples $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \sim \mu \otimes \nu$.

Sliced Gromov-Wasserstein Distance. It is essentially defined as the average of several GW distances between μ and ν once they are projected onto the space of univariate measures [Vayer et al., 2019]. In the population regime, $\mathrm{SGW}_p^p(\mu, \nu) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[\mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu)]$, where $\theta\sharp\mu$ is the push-forward of the measure μ through the function $f_\theta : \mathbb{R}^d \to \mathbb{R}$. Specifically, one employs the linear projection $f_\theta(x) = \theta^\top x$. However, the expectation in the definition of the SGW distance is intractable to compute. Therefore, the Monte Carlo scheme is employed to approximate the value: $\widehat{\mathrm{SGW}}_p^p(\mu, \nu; M) = \frac{1}{M}\sum_{l=1}^M \mathrm{GW}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)$, where $\theta_1, \ldots, \theta_M \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$ and are referred to as *projecting directions*. The pushfoward measures $\theta_1\sharp\mu, \ldots, \theta_M\sharp\mu$ are called projections of μ (similarly for ν). The number of Monte Carlo samples M is often referred to as the number of projections. When μ and ν are discrete measures that have at most n supports (sample problem), the 'best achievable' time complexity and memory complexity of the SGW are $\mathcal{O}(Mn \log n + Mdn)$ and $\mathcal{O}(Md + Mn)$ respectively [Vayer et al., 2019]. Note that it is only conditionally achievable, and we elaborate on a sufficient condition for it in the Appendix D. Evidently, to get a fair estimate, M should be large enough compared to d. Employing instead the accumulation operation max, one may define the max-sliced GW (Max-SGW) distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, given as $\text{Max-SGW}_p(\mu, \nu) = \max_{\theta \in \mathbb{S}^{d-1}} \mathrm{GW}_p(\theta\sharp\mu, \theta\sharp\nu)$. Since most slices from either space turn out non-informative, obtaining a suitable coupling becomes a challenge in Max-SGW.

On Relational Awareness. In cross-domain alignment, pairwise distances between observations is key to defining any measure of discrepancy. Preserving

[1] https://anonymous.4open.science/r/Relation-Aware-Slicing-in-Cross-Domain-Alignment-DB42/README.md

this information is crucial for ensuring that the intrinsic structure of one population (μ) is maintained during translation, matching or transform sampling [Hur et al., 2024] from the other (ν). Given samples $\{(\mathbf{x}_i, \mathbf{y}_j)\}^{m,n} \sim \mu \otimes \nu$, we refer to the distribution of pairwise distances $c_X(\mathbf{x}_i, \mathbf{x}_{i'})$, $i, i' = 1, \cdots, m$ (similarly for $c_X(\mathbf{y}_j, \mathbf{y}_{j'})$) as *relational information*. Collectively, they characterize the relational structure (local geometry) of the ambient space. The extent of departure from perfect alignment (i.e. isometry) $c_p = |c_X(\mathbf{x}_i, \mathbf{x}_{i'}) - c_Y(\mathbf{y}_j, \mathbf{y}_{j'})|^p$ embodies the association between such structures. Unlike the case $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^{d'}$, it is quite difficult to characterize the concentration of sample distortions between arbitrary spaces, e.g., networks. In general, given an optimal coupling, we refer to the same as *relational discrepancy*.

In this context, we identify a crucial property a projection (f_θ) must possess during slicing. Ideally, it should preserve the relational information of one space with respect to the other. As such, any perturbation due to projection must reflect symmetrically on pairwise distances from both spaces. We call this the *relational awareness* of a slice θ. Additionally, in Appendix C we provide further background on OT and GW. Appendix B is dedicated to establishing some baselines extending existing slicing methods in OT to GW, besides SGW and Max-SGW.

4 PROPOSED METHOD

We first introduce RAPD as candidate directions having relational awareness, which in turn invokes RASD in Section 4.1. We then discuss the Relation-Aware Sliced Gromov Wasserstein variants in Section 4.2.

Motivation. Note that, while a dimensionality reduction of a d-variate distribution inevitably leads to a loss of information, the relational information can still be recovered. However, given two participating distributions μ, ν, slicing directions also must ensure that relational discrepancy is kept intact to the greatest extent. On the other hand, for $\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$, we have $\mathbb{P}\big(\theta^T(X - X') \leq \epsilon\big) \geq 1 - e^{-d\epsilon^2}$, given any $X, X' \sim \mu$ [Kolouri et al., 2019]. As such, slicing directions chosen to preserve relational information in one space may always distort the same corresponding to the other domain. The degradation of relational discrepancy, as a result, occurs exponentially with d if slices are chosen uniformly at random.

Searching for a remedy, we identify only a handful of effective directions that become relation-aware. To illustrate our construction, let us consider a simple example. Assume both spaces to be $\mathbb{R}^d$, endowed with the Euclidean norm. As depicted in Figure 1(a), a set

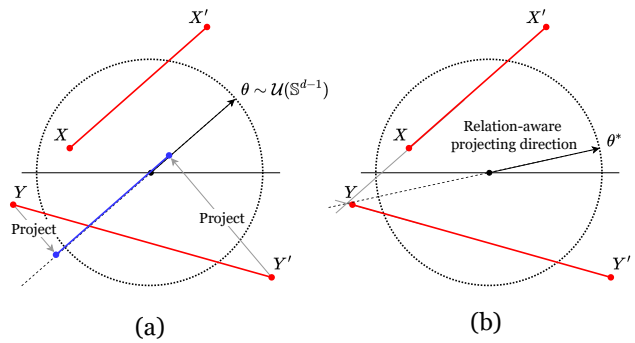


Figure 1: (a) For any $\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$, drawn uniformly randomly, it is feasible to observe $X - X' // \theta$, where $X, X' \sim \mu$. While this preserves the distance $||X - X'||$ post-projection, slicing $Y - Y'$ along the same direction (marked in blue) distorts the relational information. (b) In the absence of specific scaling for each pair (X, X') (or (Y, Y')), θ^* is the only direction that has the same effect on both pairwise distances during slicing.

of random observations $(X, Y), (X', Y') \sim \mu \otimes \nu$ may experience unequal perturbation in intra-pair distances after projection based on an arbitrary θ. Those are rendered uninformative. While rescaling is a remedy for such unwanted amplifications, it immediately makes the slicing mechanism dependent on parameters. A parameter-free approach demands a θ that leaves a symmetric impression on $X - X'$ and $Y - Y'$.

To that end, let us consider a bisecting direction θ^* at an angle α_1 with $\overrightarrow{XX'}$ and α_2 with $\overrightarrow{YY'}$. Thus,

$$
\begin{aligned}
c_p^{\frac{1}{p}} &= \Big| ||\theta^\top (X - X')|| - ||\theta^\top (Y - Y')|| \Big| \\
&= \big| ||X - X'|| \cos\alpha_1 - ||Y - Y'|| \cos\alpha_2 \big| .
\end{aligned}
$$

As such, we observe distorted relational discrepancies given an arbitrary choice of (α_1, α_2), which may lead to a suboptimal coupling π different from the intended. Only by selecting $\alpha_1 = \alpha_2 = \alpha$ do we keep the underlying optimization unscathed. Hence, the corresponding $\theta^\star$ is a candidate for relation-aware slices. The only impact on the proxy sliced GW value is a multiplicative factor of the form $\times \cos\alpha$. Notably, our construction results in a unique distortion $|c_X - c_Y|/\sqrt{2}$ if $\overrightarrow{XX'} \perp \overrightarrow{YY'}$, which has 0 probability of appearing in a pool of random slices (e.g., in SGW). While Max slicing can realize an optimal direction theoretically in SW, its maximization objective makes little sense in the alignment problem where relation preservation is more important, and the direction that maximizes the value of the metric need not be relation preserving.

4.1 Relation-Aware Projecting Direction

Let us now statistically characterize the directions we intuitively constructed.

Definition 4.1 (Intra-Relational Path)**.** Given $p \geq 1$ and dimension $d \geq 1$, the Intra-relational path (IRP) between $X, X' \sim \mu \in \mathcal{P}(\mathbb{R}^d)$ is defined as the transformation $Z_{X,X'} = X - X'$.

The distribution of $Z_{X,X'}$ can be written as $\mu * (-)\sharp\mu$, where $*$ denotes the convolution operator, and $(-)\sharp\mu$ denotes the pushforward measure of μ through the function $f(x) = -x$ [Nguyen et al., 2024]. While the density (corresponding to the distribution) is intractable in general, it is simple to sample from. Specifically, one needs to draw $X, X' \sim \mu$, followed by setting $Z_{X,X'} = X - X'$. We may readily define a corresponding projecting direction by normalizing it. That is, $\bar{Z}_{X,X'} = \frac{Z_{X,X'}}{\|Z_{X,X'}\|_2}$. This is the trivial slice that preserves relational information in $\mathscr{X}$ exactly.

Remark 4.2. Note that there lies a chance of $\bar{Z}_{X,X'}$ being undefined in a sample problem. It can be primarily solved by following a without-replacement sampling scheme while drawing the dataset. Given an existing dataset, sacrificing repeated samples leads to information loss. Nguyen et al. [2024] suggest considering a small additive constant c, which makes $Z_{X,X'} + c$ a valid intra-relational path.

Since we require a direction that takes into account the relational structure of both distributions, we consider

$$Z_{X,X',Y,Y'} = \frac{\bar{Z}_{X,X'} + \bar{Z}_{Y,Y'}}{\|\bar{Z}_{X,X'} + \bar{Z}_{Y,Y'}\|_2}, \text{ and}$$
$$Z'_{X,X',Y,Y'} = \frac{\bar{Z}_{X,X'} - \bar{Z}_{Y,Y'}}{\|\bar{Z}_{X,X'} - \bar{Z}_{Y,Y'}\|_2}.$$

Taking both of the above directions into account allows us to ensure that at least one of them is the bisecting vector of the acute angle, which further implies $\alpha \leq \pi/4$. Drawing such a direction θ^* with high probability requires defining a distribution that has a high central tendency towards it, with decaying tails. Thus, we add a continuous random perturbation, drawn from a suitable location-scale distribution (see Lehmann and Casella [1998]) around the normalized random path. This additionally helps cover the entire support $\mathbb{S}^{d-1}$ in representative samples. Based on this idea, a RAPD is defined as follows.

Definition 4.3 (Relation-Aware Projecting Direction)**.** Given a location-scale distribution σ_κ on $\mathbb{S}^{d-1}$, exactly characterized by $\kappa > 0$, the relation aware projecting direction (RAPD) between $(X, Y), (X', Y') \sim \mu \otimes \nu$ such that $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ is defined as

$$\theta | X, X', Y, Y', \kappa \sim D_\kappa(X, X', Y, Y'),$$

where $D_\kappa = \frac{1}{2}\sigma_\kappa(\cdot; Z_{X,X',Y,Y'}) + \frac{1}{2}\sigma_\kappa(\cdot; Z'_{X,X',Y,Y'})$.

Observe that, simulation from D_κ essentially means tossing a fair coin to select either $\sigma_\kappa(\cdot; Z)$ or $\sigma_\kappa(\cdot; Z')$. Typically, we choose σ_κ as the vMF distribution, whose density follows the form $\text{vMF}(\theta; \epsilon, \kappa) \propto \exp(\kappa\epsilon^\top\theta)$ Jupp and Mardia [1979] and PS, given as $\text{PS}(\theta; \epsilon, \kappa) \propto (1 + \epsilon^\top\theta)^\kappa$ De Cao and Aziz [2020]. In Definition 4.3, the reason for choosing the location-scale family is to guarantee that the projecting direction concentrates around the prescribed direction. For further details justifying the choice and a detailed discussion on the two distributions, we refer the reader to Appendix C.

Relation-Aware Slicing Distribution. From Definition 4.3, we can obtain the slicing distribution of the RAPD by marginalizing out X, Y. In particular, we define the RASD as follows.

Definition 4.4. Given $0 < \kappa < \infty$, the relation-aware slicing distribution based on $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $d \geq 1$ is defined as

$$\sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa) = \int D_\kappa(X, X', Y, Y') d(\mu \otimes \nu)^{\otimes 2}.$$

We can sample from $\sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa)$ easily by sampling $X, X' \sim \mu$ and $Y, Y' \sim \nu$ independently, then $\theta \sim D_\kappa(\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}')$, where $\mathbf{x}$ is the realization of X.

4.2 Relation-Aware Sliced GW

We now propose two sliced GW variants that rely on RASDs.

Definition 4.5 (Relation-Aware SGW)**.** Given $0 < \kappa < \infty$, $p \geq 1$, and $d \geq 1$, the relation-aware sliced Gromov-Wasserstein (RASGW) distance between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$\text{RASGW}_p^p(\mu, \nu; \sigma_\kappa) = \mathbb{E}_{\theta \sim \sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa)}[\text{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu)]$$
$$= \mathbb{E}_{X,X' \sim \mu, Y,Y' \sim \nu} \mathbb{E}_{\theta \sim D_\kappa(\theta; X, X', Y, Y')}[\text{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu)],$$

where σ_{RA} is as in Definition 4.4.

To diversify the slices according to multiple views of the dataset, we can generalize the weights of RAPDs, namely $\theta_1, \cdots, \theta_L \sim \sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa)$ by their corresponding projected distances $\text{GW}_p^p(\theta_1\sharp\mu, \theta_1\sharp\nu)), \cdots, \text{GW}_p^p(\theta_L\sharp\mu, \theta_L\sharp\nu))$. This takes inspiration from the importance sampling estimation of EBSGW as in Equation (5) (see Appendix B). The GW variant so constructed is called the Importance-Weighted RASGW (IWRASGW) distance and is given as follows.

Definition 4.6. Let $f : [0, \infty) \to (0, \infty)$ be an increasing function. Then, given $0 < \kappa < \infty$ and $L \geq 1$, the

importance weighted relation aware projection sliced Gromov-Wasserstein (IWRASGW) between two probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$\text{IWRASGW}_p^p(\mu,\nu;\sigma_\kappa,L,f) = \mathbb{E}_{\theta_1,\ldots,\theta_L\sim\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_\kappa)} \left[\sum_{l=1}^{L} \text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\frac{w_l(f;\mu,\nu)}{\sum_{j=1}^{L} w_j(f;\mu,\nu)}\right],$$

where σ_{RA} is defined as in Definition 4.4, and $w_l(f;\mu,\nu) := f(\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))$, $d \geq 1$.

The intuition behind utilizing L appropriately weighted random projecting directions is to prioritize the relational discrepancy corresponding to a pair of observations from $(\mu\otimes\nu)^{\otimes 2}$ that has greater contribution towards finding an optimal coupling.

Metric Properties. Since GW metrizes the equivalence class of isomorphic mm spaces [Mémoli, 2011], any proposed sliced variant must ideally emulate the same. Towards the same, we first investigate the metricity of RASGW and IWRASGW.

Theorem 4.7. *Given an increasing function $f : [0,\infty) \to (0,\infty)$ and $0 < \kappa < \infty$, $RASGW_p(\cdot,\cdot;\sigma_\kappa)$ and $IWRASGW_p(\cdot,\cdot;\sigma_\kappa,L)$ form semi-metrics on $\mathcal{P}(\mathcal{X})$ for any $p \geq 1$, $L \geq 1$. Namely, RASGW and IWRASGW satisfy non-negativity, symmetry, and identity of isometric isomorphism. The RASGW satisfies the 'quasi'-triangle inequality given as,*

$$RASGW_p(\mu_1,\mu_2;\sigma_\kappa) \leq RASGW_p(\mu_1,\mu_3;\sigma_\kappa,\mu_1,\mu_2) + RASGW_p(\mu_3,\mu_2;\sigma_\kappa,\mu_1,\mu_2),$$

where

$$RASGW_p^p(\mu_1,\mu_3;\sigma_\kappa,\mu_1,\mu_2) = \mathbb{E}_{\theta\sim\sigma_{RA}(\theta;\mu_1,\mu_2,\sigma_\kappa)}[GW_p^p(\theta\sharp\mu_1,\theta\sharp\mu_3)]$$

and a similar definition holds for $RASGW_p^p(\mu_3,\mu_2;\sigma_\kappa,\mu_1,\mu_2)$, for all $\mu_i \in \mathcal{P}(\mathcal{X})$, $i = 1,2,3$ such that $\mathcal{X} \subset \mathbb{R}^d$ Polish.

The proof of Theorem 4.7 is given in Appendix A.1. Observe that both RASGW and IWRASGW can be immediately extended to serve distinct dimensional measures by introducing a *padding* operation first. For example, if $\nu \in \mathcal{P}(\mathcal{Y})$, given $\mathcal{Y} \subset \mathbb{R}^{d'}$, $d' < d$; we rather calculate the RASGW distance between $\mu \in \mathcal{P}(\mathcal{X})$ and $\Delta\sharp\nu$, where $\Delta : \mathcal{Y} \to \mathcal{X}$ is a padding operator. Vayer et al. [2019], in the context of SGW, use the 'uplifting' padding given as $\Delta(y) := (y_1,\cdots,y_{d'},\underbrace{0,\cdots,0}_{d-d'})$.

The properties in Theorem 4.7 still hold under such adjustments.

Proposition 4.8 (Metric Hierarchy)**.** *For any $p \geq 1$, $L \geq 1$, increasing function $f : [0,\infty) \to (0,\infty)$, and $0 < \kappa < \infty$, we have the following relationships:*

(i) $RASGW_p(\mu,\nu;\sigma_\kappa) \leq IWRASGW_p(\mu,\nu;\sigma_\kappa,L,f) \leq Max\text{-}SGW_p(\mu,\nu)$,

(ii) $RASGW_p(\mu,\nu;\sigma_\kappa) \xrightarrow{\kappa\to 0} SGW_p(\mu,\nu)$,

(iii) $IWRASGW_p(\mu,\nu;\sigma_\kappa,L,f) \xrightarrow{L\to\infty} EBSGW_p(\mu,\nu;f)$.

Proposition 4.8 positions RASGW and IWRASGW within the SGW hierarchy, showing they approach slicing (SGW) and the energy-based slicing (EBSGW) respectively, theoretically justifying their adaptiveness. The proof of Proposition 4.8 is given in Appendix A.2.

Statistical Properties. Given that sliced GW variants innately benefit tasks such as shape matching or surface correspondence [Solomon et al., 2016], it is crucial to study the one-sided sample complexity of RASGW and IWRASGW.

Proposition 4.9. *Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mu$, such that μ is fully supported on $\mathcal{X} \subset \mathbb{R}^d$. We denote the empirical measure by $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$. Then, for any $p > 1$, $L \geq 1$ and $0 < \kappa < \infty$, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}[RASGW_p(\mu_n,\mu;\sigma_\kappa)] \leq \mathbb{E}[IWRASGW_p(\mu_n,\mu;\sigma_\kappa,L,f)] \leq C\sqrt{\frac{(d+1)\log(n+1)}{n}},$$

where the outer expectation is taken with respect to $X_1, X_2, \ldots, X_n$.

Proposition 4.9 proves sublinear sample complexity ($\mathcal{O}(n^{-1/2})$), at par with the conventional SGW [Nadjahi et al., 2019]. This positions RASGW —already computationally superior —as a better choice for a suitable loss in cross-domain generative and shape-matching tasks, compared to SGW and Max-SGW. The proof of Proposition 4.9 is given in Appendix A.3.

Computational Properties. The upcoming section is dedicated to computational details regarding our distances.

Monte Carlo Estimation. In an empirical setup, we resort to Monte Carlo simulations to estimate RASGW. This is also crucial due to the optimization that underlies the outer expectation. In particular, we sample $\theta_1,\ldots,\theta_M \sim \sigma_{\text{RA}}(\theta;\mu,\nu,\kappa)$ (as described in Section 4.1) to construct the Monte Carlo estimate of RASGW:

$$\widehat{\text{RASGW}}_p^p(\mu,\nu;\sigma_\kappa,M) = \frac{1}{M}\sum_{l=1}^{M}\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu).$$

We refer the reader to Algorithm 3- 4 in Appendix D for a detailed discussion on the computation of RASGW and IWRASGW. The immediate concern related to

such an approximation is its realized error. To that end, given that the sampling of RA slices incurs finite variability, we ensure that the approximation error is of order $\mathcal{O}(M^{-1/2})$.

Proposition 4.10. *For any $p \geq 1$, $0 < \kappa < \infty$, dimension $d \geq 1$, and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, we have:*

$$\left|\mathbb{E}[\widehat{RASGW}_p^p(\mu,\nu;\sigma_\kappa)] - RASGW_p^p(\mu,\nu;\sigma_\kappa)\right| \leq \frac{1}{\sqrt{M}} Var_{\theta\sim\sigma_{RA}(\theta;\mu,\nu,\sigma_\kappa)}\left[GW_p^p\left(\theta\sharp\mu,\theta\sharp\nu\right)\right]^{\frac{1}{2}}.$$

The proof of Proposition 4.10 is given in Appendix A.4. Since the stochasticity of σ_{RA} is completely user-defined, it is easy to control the $\text{Var}(\cdot)$ and the error bound as a result. Asymptotically, the estimation adheres to the strong law, given that the underlying measures have finite second moments. As such, we may always observe a tighter bound than SGW, which shares a similar rate in terms of M.

Similarly, for IWRASGW, we draw H MC samples each corresponding to the L random projecting directions, namely $\theta_{11}, \ldots, \theta_{HL} \sim \sigma_{\text{RA}}(\theta;\mu,\nu,\kappa)$. This forms the MC estimate of IWRASGW as follows:

$$\widehat{\text{IWRASGW}}_p^p(\mu,\nu;\sigma_\kappa,L,H) = \frac{1}{H}\sum_{h=1}^{H}\left[\sum_{l=1}^{L}\text{GW}_p^p(\theta_{hl}\sharp\mu,\theta_{hl}\sharp\nu)\frac{w_{hl}(f;\mu,\nu)}{\sum_{j=1}^{L}w_{hj}(f;\mu,\nu)}\right]. \tag{2}$$

To simplify the parameter selections, we often set $H = 1$ in the experiments. This leaves L as the only tunable parameter (number of projections). In contrast to RASGW, the error rate of IWRASGW with respect to H is non-trivial due to the importance weights.

Computational Complexities. When μ and ν are discrete measures that have at most n supports, sampling from them only cost $\mathcal{O}(n)$ in terms of time and space. Hence, sampling M random paths between μ and ν cost $\mathcal{O}(Mdn)$ in time and memory. Moreover, sampling from the vMF and the PS distribution costs $\mathcal{O}(Md)$ (see Algorithm 1 in De Cao and Aziz [2020]) in time and memory. Now, it is only feasible to incur time and space complexities for the ensuing RASGW as $\mathcal{O}(Mn\log n + Mdn)$ and $\mathcal{O}(Md + Mn)$ respectively *if* the associated GW distance in $\mathbb{R}$ (post-projection) is solved optimally by *identity* or *anti-identity* [Vayer et al., 2019]. While this may not occur almost everywhere, both Vayer et al. [2019] and Beinert et al. [2023] remark that in numerical simulations, it "often" appears to be true. Our experiments corroborate the claim empirically. In Appendix D, we present a sufficient condition on slices such that the ensuing sliced GW proxies are optimally solved, achieving the above complexity. This is crucial given that the map $\theta \mapsto \text{GW}_p^p(\theta\sharp\cdot,\theta\sharp\cdot)$ is Lipschitz continuous and any dependence on θ in the optimal plan in 1D can be traced back to the ambient spaces. Future work may look into the high-density regions of σ_{RM} that contribute to satisfying the sufficient condition. Anyway, the involvement of data dimensionalities in the overall complexity stems solely from the projections, making RASGW scale linearly with the same. This complements the complexities achieved by SGW, given that it results in an optimal plan. In the case of IWRASGW, the complexities remain the same.

Gradient Estimation. In applications (see Section 5), where RASGW and IWRASGW are used as a risk to estimate some parameters of interest, i.e., μ_ϕ with $\phi \in \Phi$, we might want to estimate the gradient of RASGW and IWRASGW with respect to ϕ. For RASGW, we have

$$\nabla_\phi \text{RASGW}_p^p(\mu_\phi,\nu;\sigma_\kappa) = \nabla_\phi \mathbb{E}_{X,X'\sim\mu_\phi;Y,Y'\sim\nu} \mathbb{E}_{\theta\sim\sigma_\kappa(\theta;D_\kappa(X,X',Y,Y')}[\text{GW}_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)].$$

Since vMF and PS are reparameterizable [De Cao and Aziz, 2020], we may do the same for $\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_\kappa)$ given μ_ϕ is reparameterizable (e.g., $\mu_\phi := f_\phi\sharp\varepsilon$ for ε being a fixed noise, or as discussed in Kingma and Welling [2013]). A common example of this may be a generative model, such as a GAN, where f_ϕ is a transformation induced by a neural network. With parameterized sampling, we can sample $\theta_{1,\phi}, \ldots, \theta_{M,\phi} \sim \sigma_{\text{RA}}(\theta;\mu_\phi,\nu,\sigma_\kappa)$, then form an unbiased gradient estimator as follows:

$$\nabla_\phi \text{RASGW}_p^p(\mu_\phi,\nu;\sigma_\kappa) \approx \frac{1}{M}\sum_{l=1}^{M}\nabla_\phi \text{GW}_p^p(\theta_{l,\phi}\sharp\mu_\phi,\theta_{l,\phi}\sharp\nu).$$

For IWRASGW, we sample $\theta_{11,\phi}, \ldots, \theta_{HL,\phi} \sim \sigma_{\text{RA}}(\theta;\mu_\phi,\nu,\kappa)$ to form a similar gradient estimator, the expression of which can be found in Appendix E. In this context, we mention that the requirement of gradient computation renders fast GW approximation methods, such as Scetbon et al. [2022], unsuitable to be adapted as a loss in cross-domain generative models (see Appendix B for details).

Along with the other OT-inspired baselines (see Appendix B, we introduce the **Random Path Sliced Gromov-Wasserstein (RPSGW)**. In Nguyen et al. [2024], the authors proposed a variant of Sliced-Wasserstein known as Random Path Sliced Wasserstein (RPSW). It introduces the random-path projecting direction (RPD), constructed as a normalized difference between two random vectors sampled from the input measures, with a random perturbation to ensure continuity. This creates a new slicing distribution, the random-path slicing distribution (RPSD), which is optimization-free and efficient to sample. It is not immediately clear how to adapt RPSW to the SGW setting, as RPSW constructs its projecting directions based on just two points (one taken from each distribution), whereas in GW, we sample a pair of points

Table 1: GW-2 distance and computational times for 2D→3D and 3D→2D generations in 4-point experiment based on 10 repetitions.

Method	Step 10000 (3D to 2D)		Step 10000 (2D to 3D)	
	$GW_2(\downarrow)$	Time(s)(↓)	$GW_2(\downarrow)$	Time(s)(↓)
SGW	122.54±11.94	42.30±0.50	2.23±0.62	41.90±0.43
Max-SGW	385.28±8.98	213.45±2.23	2.25±0.25	210.98±2.76
DSGW	102.95±3.59	247.68±5.65	1.64±0.40	242.39±4.87
EBSGW	85.54±5.51	**40.09±0.88**	1.79±0.70	**41.39±0.63**
RPSGW	39.73±3.54	52.90±0.63	1.34±0.35	55.65±0.89
IWRPSGW	37.44±3.23	53.12±1.09	1.31±0.45	57.84±1.21
RASGW	**18.80±2.46**	51.89±4.09	1.28±0.54	59.03±3.54
IWRASGW	22.45±3.18	54.45±3.41	**1.15±0.32**	56.54±5.49

belonging to each distribution. We can extend it to the GW setting by taking the quartet (X, X', Y, Y') followed by constructing a direction by taking a cross pair. However, this is quite arbitrary, and it is not exactly clear how this serves any purpose in our alignment problem. The upcoming experimental studies show that our metric outperforms this by quite a margin.

5 SIMULATION RESULTS

Gromov-Wasserstein GAN. In contrast to the vanilla GAN [Goodfellow et al., 2014], Bunne et al. [2019] proposed a variant that operates between spaces of different dimensions by capturing their geometric relations using GW loss instead. Given a lower-dimensional input noise $Z \sim \nu$ (typically a Gaussian), the underlying optimization over a class of generators $\mathcal{G}$ becomes: $G^* = \text{argmin}_{G\in\mathcal{G}}\ \text{GW}_2^2(\mu, G\sharp\nu)$, where μ is the target distribution. Observe that here, the transformation G executes an uplifting operation. Vayer et al. [2019] adopts this framework by replacing GW with SGW as the loss function for training G. Our first experiment involves checking the efficacy of RASGW and IWRASGW in such cross-domain generations. We consider both cases, i.e., μ (target) and ν (source), of similar and dissimilar dimensionalities.

I. Learning Across Identical Spaces. First, we consider both mm spaces to be similar, as in the case of a typical GAN. Specifically, we assess the model's ability to recover 2D Gaussian mixtures, a common task for mode recovery [Che et al., 2016, Metz et al., 2016]. For this synthetic experiment, we follow the setup of Bunne et al. [2019], where both the generator and adversary are feed-forward networks with ReLU activations. The immediate consequence of plugging in RASGW as the loss is a significant improvement in convergence speed without compromising generation quality (see Section E.2.2).

II. Learning Across Dimensionalities. We consider a mixture of bivariate Gaussians as the source distribution for a truly cross-domain experiment. The target is constructed as a tri-variate Gaussian mixture, each component being drawn around a vertex of a hypercube in $\mathbb{R}^3$ as its center. Here, we also use ReLU-activated generators. For a detailed discussion, we refer the reader to Section E.2.3.

Results. Based on our setting (see Section E.2.1 of Appendix), following Bunne et al. [2019], RASGW and IWRASGW bear superior outcomes. For the 2D source mixture distribution of 4 Gaussians (4-point experiment), Table 1 shows that the average GW-2 distance between the target and the generated 3D distributions is minimally realized using our models. The corresponding computational times are also comparable to SGW and other optimization-free variants, such as EBSGW, RPSGW and IWRPSGW. Figure 6 (see Section E.2.4) compares the same for the generation of the base law. The efficient bidirectional generation hints at achieving lower relational discrepancy. Figure 7 and Figure 8 (see Appendix) illustrate the propagation of errors across iterations. Furthermore, RASGW and IWRASGW demonstrate significantly lower computational costs than DSGW and Max-SGW.

Gromov-Wasserstein Autoencoder. We select autoencoders for the second generative experiment, which inherently considers unalike spaces in the form of the input space and the latent representation. Specifically, we modify the GW Autoencoder (GWAE) [Nakagawa et al., 2023] to reconstruct input image datasets (CIFAR-10 [Krizhevsky et al., 2009] and Omniglot [Liu et al., 2015]). Instead of likelihood maximization, GWAE models learn representations by matching the geometric structure between the latent and data spaces. The training objective of GWAE between the metric measure spaces $(\mathcal{X}, d_X, \mu)$ (input) and $(\mathcal{Z}, d_Z, \pi_\theta)$ (latent) is given as $\min_\theta \text{GW}_p^p(\mu, \pi_\theta)$, where π_θ is the trainable latent prior. Nakagawa et al. [2023] recast the GW objective into a main GW component $\mathcal{L}_{gw}$ with three regularizations: a reconstruction loss $\mathcal{L}_w$, a joint dual loss $\mathcal{L}_d$, and an entropy regularization $\mathcal{L}_h$. We replace the objective using RASGW and IWRASGW and perform extensive experiments against competitors SGW and the other sliced variants for image reconstruction.

Results. We follow the architecture of Nakagawa et al. [2023] as discussed in Section E.3.1 of the Appendix. For all the experiments, we followed early stopping with a patience of 10 epochs, ensuring that training terminated if the validation performance did not improve for 10 consecutive epochs. In CIFAR-10, based on FID scores [Heusel et al., 2017], RASGW and IWRASGW perform the best and the second best, respectively, whereas for Omniglot, IWRASGW performs the best with RASGW following closely in second. On PSNR scores, IWRASGW performs best, followed by SGW in CIFAR-10 (Table 2). The corresponding runtime

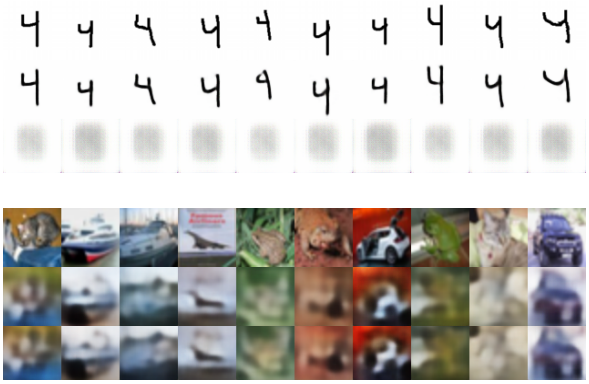


Figure 2: **(Top)** Target (*top*), reconstructed Omniglot images (*middle*) and outputs (*bottom*) at initialization using IWRASGW. **(Bottom)** Input (*top*), reconstructed images using RASGW at 40th epoch (*middle*) and reconstructed images using SGW at 48th epoch (*bottom*) for CIFAR-10.

Table 2: Results on CIFAR-10 (C) and Omniglot (O) datasets. FID scores (↓), PSNR values (in dB, ↑), total training time (seconds, ↓), and epochs (↓) are reported for each method. Columns C and O refer to CIFAR-10 and Omniglot, respectively.

Method	FID (↓)		PSNR (↑)		Total (↓)		Epochs(↓)	
	C	O	C	O	C	O	C	O
SGW	72.87	20.95	22.65	**22.13**	22459	16932	48	70
Max-SGW	88.97	22.27	22.38	21.94	51345	42975	**27**	63
DSGW	83.45	21.66	22.41	22.10	59876	47585	30	67
EBSGW	78.50	38.99	22.48	20.53	28765	15533	52	55
RPSGW	80.80	29.88	22.55	21.33	24356	**12321**	43	**45**
RASGW	**70.46**	21.48	22.70	20.56	23613	16731	40	62
IWRASGW	71.99	**20.25**	**22.73**	21.87	**19879**	18972	35	76

attests to the efficacy of our models. For a discussion on the evaluation metrics, refer to Section E.3.3 in the Appendix. The qualitative results on reconstruction and plots of change in PSNR and loss values are placed in Section E.4 (CIFAR-10) and E.5 (Omniglot).

6 CONCLUSION AND FUTURE WORK

We introduce a novel approach for developing informative projecting directions for SGW. Based on this, we derive the concept of relation-aware slicing distributions along with two optimization-free SGW variants, namely RASGW and IWRASGW, that preserve topological (metric) properties and improve computational cost. It improves over uniform and Max slicing based on informative projections that reflect in cross-domain generative tasks over distinct spaces. Future research may integrate the WD into our framework to develop a fused metric, better serving attributed network spaces.

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The z-gromov-wasserstein distance, 2025.

Robert Beinert, Cosmas Heiss, and Gabriele Steidl. On assignment problems related to gromov–wasserstein distances on the real line. *SIAM Journal on Imaging Sciences*, 16(2):1028–1032, 2023.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015.

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *International Conference on Machine Learning*, pages 851–861, 2019.

Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

Xiongjie Chen, Yongxin Yang, and Yunpeng Li. Augmented sliced wasserstein distances. *arXiv preprint arXiv:2006.08812*, 2020.

Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Nicola De Cao and Wilker Aziz. The power spherical distribution. *arXiv preprint arXiv:2006.04437*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Wenshuo Guo, Nhat Ho, and Michael Jordan. Fast algorithms for computational optimal transport and wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. PMLR, 2020.

Sigurdur Helgason. The radon transform on r n. In *Integral Geometry and Radon Transforms*, pages 1–62. Springer, 2011.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

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.

YoonHaeng Hur, Wenxuan Guo, and Tengyuan Liang. Reversible gromov–monge sampler for simulation-based inference. *SIAM Journal on Mathematics of Data Science*, 6(2):283–310, 2024.

Peter E Jupp and Kanti V Mardia. Maximum likelihood estimators for the matrix von Mises-Fisher and bingham distributions. *The Annals of Statistics*, 7 (3):599–606, 1979.

Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338, 2015.

Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 250–260, 2019.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

Nao Nakagawa, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Gromov-wasserstein autoencoders. In *The Eleventh International Conference on Learning Representations*, 2023.

Khai Nguyen and Nhat Ho. Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 2023.

Khai Nguyen and Nhat Ho. Hierarchical hybrid sliced wasserstein: A scalable metric for heterogeneous

joint distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=XwrMd1njqq.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Minh Nguyen, and Nhat Ho. Hierarchical sliced wasserstein distance. In *The Eleventh International Conference on Learning Representations*, 2023.

Khai Nguyen, Shujian Zhang, Tam Le, and Nhat Ho. Sliced wasserstein with random-path projecting directions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

Sloan Nietert, Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 2022.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

Moritz Piening and Robert Beinert. A novel sliced fused gromov-wasserstein distance. *arXiv preprint arXiv:2508.02364*, 2025.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Entropic gromov-wasserstein distances: Stability and algorithms. *Journal of Machine Learning Research*, 25(363):1–52, 2024.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Suvrit Sra. Directional statistics in machine learning: a brief review. *arXiv preprint arXiv:1605.00316*, 2016.

Nico M Temme. *Special functions: An introduction to the classical functions of mathematical physics*. John Wiley & Sons, 2011.

Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced gromov-wasserstein. *arXiv preprint arXiv:1905.10124*, 2019.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K Sriperumbudur. Gromov–wasserstein distances: Entropic regularization, duality and sample complexity. *The Annals of Statistics*, 52(4):1616–1645, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

On Relation-Aware Slicing in Cross-Domain Alignment Supplementary Material

First, we present skipped proofs in the main text in Appendix A. We then discuss some related works in B. We provide the algorithms in Appendix D. After that, we provide some additional experimental details in Appendix E. Finally, we provide details on the computational infrastructure in Appendix F. For codes for our experiments, we refer the reader to the repository https://anonymous.4open.science/r/Relation-Aware-Slicing-in-Cross-Domain-Alignment-DB42/README.md.

A TECHNICAL PROOFS

A.1 Proof of Theorem 4.7

Non-negativity. Since the Gromov-Wasserstein distance is non-negative, we have $\mathrm{GW}_p(\theta\sharp\mu, \theta\sharp\nu) \geq 0$ for all $\theta \in \mathbb{S}^{d-1}$. Therefore, $\mathbb{E}_{\theta\sim\sigma_{\mathrm{RA}}(\theta;\mu,\nu,\sigma_\kappa)}[\mathrm{GW}_p(\theta\sharp\mu, \theta\sharp\nu)] \geq 0$ which leads to $\mathrm{RASGW}_p(\mu,\nu;\sigma_\kappa) \geq 0$. Similarly, we have $\mathbb{E}_{\theta_1,\ldots,\theta_L\sim\sigma_{\mathrm{RA}}(\theta;\mu,\nu,\sigma_\kappa)}\left[\sum_{l=1}^L \mathrm{GW}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)\frac{f(\mathrm{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))}{\sum_{j=1}^L f(\mathrm{GW}_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))}\right] \geq 0$ which implies $\mathrm{IWRASGW}_p(\mu,\nu;\sigma_\kappa, L) \geq 0$

Symmetry. From the definition of RASGW and $f_{D_\kappa(x,x',y,y')}(\theta)$ as the density function of $D_\kappa(x,x',y,y')$ we have

$$
\begin{aligned}
&\mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{D_\kappa(x,x',y,y')}(\theta)\\
=&\frac{1}{2}\mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(Z_{x,x',y,y'})}(\theta) + \frac{1}{2}\mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(Z'_{x,x',y,y'})}(\theta)\\
=&\frac{1}{2}\mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(Z_{y,y',x,x'})}(\theta) + \frac{1}{2}\mathrm{GW}_p^p(-\theta\sharp\mu, -\theta\sharp\nu) f_{\sigma_\kappa(Z'_{y,y',x,x'})}(-\theta).
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{D_\kappa(x,x',y,y')}(\theta) d\theta\\
=&\frac{1}{2}\int_{\mathbb{S}^{d-1}} \mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(Z_{y,y',x,x'})}(\theta) d\theta + \frac{1}{2}\int_{\mathbb{S}^{d-1}} \mathrm{GW}_p^p(-\theta\sharp\mu, -\theta\sharp\nu) f_{\sigma_\kappa(Z'_{y,y',x,x'})}(-\theta) d\theta\\
=&\frac{1}{2}\int_{\mathbb{S}^{d-1}} \mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(Z_{y,y',x,x'})}(\theta) d\theta + \frac{1}{2}\int_{\mathbb{S}^{d-1}} \mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(Z'_{y,y',x,x'})}(\theta) d\theta\\
=&\int_{\mathbb{S}^{d-1}} \mathrm{GW}_p^p(\theta\sharp\nu, \theta\sharp\mu) f_{D_\kappa(y,y',x,x')}(\theta) d\theta,
\end{aligned}
$$

where we use the reflection property of $\mathbb{S}^{d-1}$, i.e.,

$$
\begin{aligned}
\mathrm{GW}_p^p(\theta\sharp\mu, \theta\sharp\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int |\theta^\top(x-y)|^p d\pi(x,y)\\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int |-\theta^\top(x-y)|^p d\pi(x,y) = \mathrm{GW}_p^p(-\theta\sharp\mu, -\theta\sharp\nu),
\end{aligned}
$$

and

$$
\begin{aligned}
f_{\sigma_\kappa(\theta;Z_{x,x',y,y'})}(\theta) &= f_{\sigma_\kappa(\theta;Z_{y,y',x,x'})}(\theta)\\
f_{\sigma_\kappa(\theta;Z'_{x,x',y,y'})}(\theta) &= f_{\sigma_\kappa(\theta;Z'_{y,y',x,x'})}(-\theta)
\end{aligned}
$$

which holds for both the vMF density $f_{\sigma_\kappa(\theta;\epsilon)}(\theta) \propto \exp\left(\kappa\epsilon^\top\theta\right)$ and the PS density $f_{\sigma_\kappa(\theta;\epsilon)}(\theta) \propto \left(1+\epsilon^\top\theta\right)^\kappa$. After integrating over the measures, we get,

$$\text{RASGW}_p^p(\mu,\nu;D_\kappa) = \text{RASGW}_p^p(\nu,\mu;D_\kappa)$$

Similarly, we have $\text{IWRASGW}_p^p(\mu,\nu;\sigma_\kappa,L,f) = \text{IWRASGW}_p^p(\nu,\mu;\sigma_\kappa,L,f)$.

Existence of Isometric Isomorphism. We need to show that $\text{RASGW}_p(\mu,\nu;\sigma_k) = 0$ if and only if $\exists$ an isometric isomorphism between μ and ν. For the forward direction, since $\text{GW}_p(\theta\sharp\mu,\theta\sharp\nu) = 0$ when μ and ν have an isometric isomorphism between them, we obtain directly $\text{RASGW}_p(\mu,\nu;\sigma_k) = 0$. For the reverse direction, we use the same proof technique in Vayer et al. [2019]. If $\text{RASGW}_p(\mu,\nu;\sigma_k) = 0$, we have $\int_{\mathbb{S}^{d-1}} \text{GW}_p\left(\theta\sharp\mu,\theta\sharp\nu\right) \mathrm{d}\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_k) = 0$. Hence, we have $\text{GW}_p(\theta\sharp\mu,\theta\sharp\nu) = 0$ for $\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_k)$-almost surely $\theta \in \mathbb{S}^{d-1}$. Since $\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_k)$ is continuous due to the continuity of σ_k, we have $\text{GW}_p(\theta\sharp\mu,\theta\sharp\nu) = 0$ for all $\theta \in \mathbb{S}^{d-1}$. Therefore, the measures are isometrically isomorphic. This follows from the proof of Theorem 10 in Vayer et al. [2019].

For IWRASGW, when μ, ν, are isometrically isomorphic we have $\text{GW}_p(\theta_l\sharp\mu,\theta_l\sharp\nu) = 0$ for any $\theta_1,\ldots,\theta_L \in \mathbb{S}^{d-1}$. Therefore, we have $\sum_{l=1}^L \text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\frac{f(\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))}{\sum_{j=1}^L \text{GW}_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu)} = 0$ for any $\theta_1,\ldots,\theta_L \in \mathbb{S}^{d-1}$ which implies

$$\text{IWRASGW}_p(\mu,\nu;\sigma_\kappa,L) = \mathbb{E}\left[\sum_{l=1}^L \text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\frac{f(\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))}{\sum_{j=1}^L \text{GW}_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu)}\right] = 0.$$

In the reverse direction, when $\text{IWRASGW}_p(\mu,\nu;\sigma_\kappa,L) = 0$, it means that we have $\sum_{l=1}^L \text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\frac{f(\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))}{\sum_{j=1}^L \text{GW}_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu)} = 0$ for any $\theta_1,\ldots,\theta_L \in \mathbb{S}^{d-1}$. Since $f(\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)) > 0$ for any θ_j, it implies that $\text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu) = 0$ for all $\theta_l \in \mathbb{S}^{d-1}$. With similar arguments to the proof of RASGW, we obtain μ, ν are isometrically isomorphic, which completes the proof.

Quasi-Triangle Inequality. Given three probability measures $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_p(\mathbb{R}^d)$ we have:

$$\begin{aligned}
&\text{RASGW}_p(\mu_1,\mu_2;\sigma_\kappa) = \left(\mathbb{E}_{\theta\sim\sigma_{\text{RA}}(\theta;\mu_1,\mu_2,\sigma_\kappa)}[\text{GW}_p^p(\theta\sharp\mu_1,\theta\sharp\mu_2)]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{\theta\sim\sigma_{\text{RA}}(\theta;\mu_1,\mu_2,\sigma_\kappa)}[(\text{GW}_p(\theta\sharp\mu_1,\theta\sharp\mu_3) + \text{GW}_p(\theta\sharp\mu_3,\theta\sharp\mu_2))^p]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{\theta\sim\sigma_{\text{RA}}(\theta;\mu_1,\mu_2,\sigma_\kappa)}[\text{GW}_p^p(\theta\sharp\mu_1,\theta\sharp\mu_3)]\right)^{\frac{1}{p}} + \left(\mathbb{E}_{\theta\sim\sigma_{\text{RA}}(\theta;\mu_1,\mu_2,\sigma_\kappa)}[\text{GW}_p^p(\theta\sharp\mu_3,\theta\sharp\mu_2)]\right)^{\frac{1}{p}} \\
&= \text{RASGW}_p(\mu_1,\mu_3;\sigma_\kappa,\mu_1,\mu_2) + \text{RASGW}_p(\mu_3,\mu_2;\sigma_\kappa,\mu_1,\mu_2),
\end{aligned}$$

where the first inequality is due to the triangle inequality of Wasserstein distance, and the second inequality is due to the Minkowski inequality. Since each pair (μ_i,μ_j) uses a different slicing distribution, it is perhaps important to note that the expectation is first taken over $\sigma_{\text{RA}}(\cdot;\mu_1,\mu_2,\sigma_\kappa)$ and then integrated. We complete the proof here.

A.2 Proof of Proposition 4.8

(i) To prove that $\text{RASGW}_p(\mu,\nu;\sigma_\kappa) \leq \text{IWRASGW}_p(\mu,\nu;\sigma_\kappa,L)$, we introduce the following lemma which had been proved in Nguyen and Ho [2023]. Here, we provide the proof for completeness.

Lemma A.1. *For any $L \geq 1$, $0 \leq a_1 \leq a_2 \leq \ldots \leq a_L$ and $0 < b_1 \leq b_2 \leq \ldots \leq b_L$, we have:*

$$\frac{1}{L}(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) \leq \sum_{i=1}^L a_i b_i. \tag{3}$$

Proof. We prove Lemma A.1 via an induction argument. For $L = 1$, it is clear that $a_i b_i = a_i b_i$. Now, we assume that the inequality holds for L i.e., $\frac{1}{L}(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) \leq \sum_{i=1}^L a_i b_i$ or $(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) \leq L\sum_{i=1}^L a_i b_i$. Now,

we want to show that the inequality holds for $L+1$ i.e., $(\sum_{i=1}^{L+1} a_i)(\sum_{i=1}^{L} b_i) \leq (L+1)\sum_{i=1}^{L+1} a_i b_i$. First, we have:

$$\begin{aligned}(\sum_{i=1}^{L+1} a_i)(\sum_{i=1}^{L+1} b_i) &= (\sum_{i=1}^{L} a_i)(\sum_{i=1}^{L} b_i) + (\sum_{i=1}^{L} a_i)b_{L+1} + (\sum_{i=1}^{L} b_i)a_{L+1} + a_{L+1}b_{L+1} \\ &\leq L\sum_{i=1}^{L} a_i b_i + (\sum_{i=1}^{L} a_i)b_{L+1} + (\sum_{i=1}^{L} b_i)a_{L+1} + a_{L+1}b_{L+1}.\end{aligned}$$

By rearrangement inequality, we have $a_{L+1}b_{L+1} + a_i b_i \geq a_{L+1}b_i + b_{L+1}a_i$ for all $1 \leq i \leq L$. By taking the sum of these inequalities over i from 1 to L, we obtain:

$$(\sum_{i=1}^{L} a_i)b_{L+1} + (\sum_{i=1}^{L} b_i)a_{L+1} \leq \sum_{i=1}^{L} a_i b_i + La_{L+1}b_{L+1}.$$

Therefore, we have

$$\begin{aligned}(\sum_{i=1}^{L+1} a_i)(\sum_{i=1}^{L+1} b_i) &\leq L\sum_{i=1}^{L} a_i b_i + (\sum_{i=1}^{L} a_i)b_{L+1} + (\sum_{i=1}^{L} b_i)a_{L+1} + a_{L+1}b_{L+1} \\ &\leq L\sum_{i=1}^{L} a_i b_i + \sum_{i=1}^{L} a_i b_i + La_{L+1}b_{L+1} + a_{L+1}b_{L+1} \\ &= (L+1)(\sum_{i=1}^{L+1} a_i b_i),\end{aligned}$$

which completes the proof. □

From Lemma A.1, with $a_i = \mathrm{GW}_p^p(\theta_i\sharp\mu, \theta_j\sharp\nu)$ and $b_i = f(\mathrm{GW}_p^p(\theta_i\sharp\mu, \theta_j\sharp\nu))$, we have:

$$\frac{1}{L}\sum_{i=1}^{l} \mathrm{GW}_p^p(\theta_i\sharp\mu, \theta_i\sharp\nu) \leq \sum_{i=1}^{L} \mathrm{GW}_p^p(\theta_i\sharp\mu, \theta_i\sharp\nu)\frac{f(\mathrm{GW}_p^p(\theta_i\sharp\mu, \theta_i\sharp\nu))}{\sum_{j=1}^{L} f(\mathrm{GW}_p^p(\theta_j\sharp\mu, \theta_j\sharp\nu))}.$$

Taking the expectation with respect to $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \sigma_{\mathrm{RA}}(\theta; \mu, \nu, \sigma_\kappa)$, we obtain $\mathrm{RASGW}_p(\mu, \nu; \sigma_\kappa) \leq \mathrm{IWRASGW}_p(\mu, \nu; \sigma_\kappa, L)$.

Now to show that $\mathrm{IWRASGW}_p(\mu, \nu; \sigma_\kappa, L) \leq \text{Max-SGW}(\mu, \nu)$, we have $\theta^\star = \mathrm{argmax}_{\theta\in\mathbb{S}^{d-1}} \mathrm{GW}_p(\theta\sharp\mu, \theta\sharp\nu)$ since $\mathbb{S}^{d-1}$ is compact and the function $\theta \to \mathrm{GW}_p(\theta\sharp\mu, \theta\sharp\nu)$ is continuous. From the definition of the IWRASGW, for any $L \geq 1, \sigma_\kappa \in \mathcal{P}(\mathbb{S}^{d-1})$ we have:

$$\begin{aligned}&\mathrm{IWRASGW}_p(\mu, \nu; \sigma_\kappa, L) \\ &= \left(\mathbb{E}_{\theta_1,\ldots,\theta_L\sim\sigma_{\mathrm{RA}}(\theta;\mu,\nu,\sigma_\kappa)}\left[\sum_{l=1}^{L} \mathrm{GW}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)\frac{f(\mathrm{GW}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu))}{\sum_{j=1}^{L} f(\mathrm{GW}_p^p(\theta_j\sharp\mu, \theta_j\sharp\nu))}\right]\right)^{\frac{1}{p}} \\ &\leq \left(\mathbb{E}_{\theta\sim\sigma_{\mathrm{RA}}(\theta;\mu,\nu,\sigma_k)}\left[\mathrm{GW}_p^p\left(\theta^\star\sharp\mu, \theta^\star\sharp\nu\right)\right]\right)^{\frac{1}{p}} = \mathrm{GW}_p^p\left(\theta^\star\sharp\mu, \theta^\star\sharp\nu\right) = \text{Max-SGW}_p(\mu, \nu).\end{aligned}$$

(ii) We first recall the density of the von Mises-Fisher distribution and the Power spherical distribution. In particular, we have $vMF(\theta; \epsilon, \kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}\exp(\kappa\epsilon^\top\theta)$ with I_s denotes the modified Bessel function of the first kind, and the PS distribution $PS(\theta; \epsilon, \kappa) = \left(2^{d+\kappa-1}\pi^{(d-1)/2}\frac{\Gamma((d-1)/2+\kappa)}{\Gamma(d+\kappa-1)}\right)^{-1}(1+\epsilon^\top\theta)^\kappa$. We have:

$$\begin{aligned}&\lim_{\kappa\to 0}\frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}\exp(\kappa\epsilon^\top\theta) \to C_1, \\ &\lim_{\kappa\to 0}\left(2^{d+\kappa-1}\pi^{(d-1)/2}\frac{\Gamma((d-1)/2+\kappa)}{\Gamma(d+\kappa-1)}\right)^{-1}(1+\epsilon^\top\theta)^\kappa \to C_2\end{aligned}$$

for some constant C_1 and C_2 which do not depend on θ, hence, the vMF distribution and the PS distribution converge to the uniform distribution when $\kappa \to 0$. Therefore, we have $vMF(\theta;\epsilon,\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$ and $PS(\theta;\epsilon,\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$. Therefore, we have $\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$.

Now, we need to show that $\text{GW}_p^p(\theta\sharp\mu,\theta\sharp\nu)$ is bounded and continuous with respective to θ. For boundedness, it is sufficient to note that the GW distance is a bounded distance.

For the continuity, let $(\theta_t)_{t\geq 1}$ be a sequence on $\mathbb{S}^{d-1}$ which converges to $\theta \in \mathbb{S}^{d-1}$ i.e., $\|\theta_t - \theta\| \to 0$ as $t \to \infty$, and a arbitrary measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$. Then we have:

$$\begin{aligned}
\text{GW}_p(\theta\sharp\mu,\theta_t\sharp\mu) &= \left(\inf_{\pi\in\Pi(\mu,\mu)}\int_{\mathbb{R}^d} ||\theta_t^\top x - \theta_t^\top x'| - |\theta^\top y - \theta^\top y'||^p d\pi(x,y)d\pi(x',y')\right)^{1/p} \\
&\leq \left(\int_{\mathbb{R}^d} ||\theta_t^\top x - \theta_t^\top x'| - |\theta^\top x - \theta^\top x'||^p d\pi(x,x)d\pi(x',x')\right)^{1/p} \\
&\leq \left(\int_{\mathbb{R}^d} \|x-x'\|^p \mu(dx)\mu(dx')\right)^{1/p} \|\theta_t - \theta\| \to 0 \quad \text{as } t \to \infty,
\end{aligned}$$

where $\left(\int_{\mathbb{R}^d} \|x\|^p \mu(dx)\right)^{1/p} < \infty$ since $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, and the second inequality is due to the Cauchy-Schwartz inequality.

Using the triangle inequality, we have:

$$\begin{aligned}
&\left|\text{GW}_p(\theta_t\sharp\mu,\theta_t\sharp\nu) - \text{GW}_p(\theta\sharp\mu,\theta\sharp\nu)\right| \\
&\leq \left|\text{GW}_p(\theta_t\sharp\mu,\theta_t\sharp\nu) - \text{GW}_p(\theta\sharp\mu,\theta_t\sharp\nu)\right| + \left|\text{GW}_p(\theta\sharp\mu,\theta_t\sharp\nu) - \text{GW}_p(\theta\sharp\mu,\theta\sharp\nu)\right| \\
&\leq \text{GW}_p(\theta\sharp\mu,\theta_t\sharp\mu) + \text{GW}_p(\theta\sharp\nu,\theta_t\sharp\nu) \to 0 \text{ as } t \to \infty,
\end{aligned}$$

hence, $\text{GW}_p(\theta_t\sharp\mu,\theta_t\sharp\nu) \to \text{GW}_p(\theta\sharp\mu,\theta\sharp\nu)$ as $t \to \infty$, which complete the proof of continuity.

From the boundedness, continuity, and $\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$, we have

$$\begin{aligned}
\text{RASGW}_p^p(\mu,\nu;\sigma_\kappa) &= \mathbb{E}_{\theta\sim\sigma_{\text{RA}}(\theta;\mu,\nu,\sigma_\kappa)}[\text{GW}_p(\theta\sharp\mu,\theta\sharp\nu)] \\
&\to \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[\text{GW}_p(\theta\sharp\mu,\theta\sharp\nu)] = \text{SGW}_p^p(\mu,\nu).
\end{aligned}$$

Applying the continuous mapping theorem for $x \to x^{1/p}$, we obtain $\lim_{\kappa\to 0}\text{RASGW}_p(\mu,\nu;\sigma_\kappa) \to \text{SGW}_p(\mu,\nu)$.

(iii) Since we have proved that $\text{GW}_p^p(\theta\sharp\mu,\theta\sharp\nu)$ is bounded and continuous with respective to θ, we can show that $\sum_{i=1}^L \text{GW}_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)\frac{f(\text{GW}_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))}{\sum_{j=1}^L f(\text{GW}_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))}$ are bounded and continuous with respect to $\theta_1,\ldots,\theta_L$. As $L \to \infty$, we have

$$\sum_{i=1}^L \text{GW}_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)\frac{f(\text{GW}_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))}{\sum_{j=1}^L f(\text{GW}_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))} \to \mathbb{E}_{\gamma\sim\sigma_{\mu,\nu,f}(\gamma)}[\gamma\sharp\mu,\gamma\sharp\nu] = \text{EBSGW}_p^p(\mu,\nu;f).$$

Applying the continuous mapping theorem for $x \to x^{1/p}$, we obtain $\lim_{L\to\infty}\text{IWRASGW}_p(\mu,\nu;\sigma_\kappa,L,f) \to \text{EBSGW}_p(\mu,\nu;f)$.

A.3 Proof of Proposition 4.9

The proof of this result follows from the proof of Proposition 3 in Nguyen and Ho [2023]. We assume that μ has a compact support $\mathcal{X} \subset \mathbb{R}^d$.

From Proposition 4.8, we have

$$\begin{aligned}
\mathbb{E}[\text{RASGW}_p(\mu_n,\mu;\sigma_\kappa)] &\leq \mathbb{E}[\text{IWRASGW}_p(\mu_n,\mu;\sigma_\kappa,L)] \\
&\leq \mathbb{E}\left[\text{Max-SGW}_p(\mu_n,\mu)\right] \leq c\mathbb{E}\left[\text{Max-SW}_p(\mu_n,\mu)\right],
\end{aligned}$$

for any $\sigma_\kappa \in \mathcal{P}(\mathbb{S}^{d-1})$ and $L \geq 1$. Further, the last inequality follows from Zhang et al. [2024].

Therefore, the proposition follows as long as we can demonstrate that

$$\mathbb{E}[\text{Max-SW}_p(\mu_n, \mu)] \leq C' \sqrt{(d+1)\log n/n}$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ with $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mu$, and $C' > 0$ is some universal constant which satisfies $C = cC'$ and the outer expectation is taken with respect to $X_1, \ldots, X_n$.

Using the closed-form of one-dimensional Wasserstein distance, we have:

$$\text{Max-SW}_p(\mu_n, \mu) = \max_{\theta \in \mathbb{S}^{d-1}} \int_0^1 |F_{n,\theta}^{-1}(z) - F_\theta^{-1}(z)|^p dz,$$

where $F_{n,\theta}$ and F_θ as the cumulative distributions of $\theta\sharp\mu_n$ and $\theta\sharp\mu$. Since $\text{GW}_p((t\theta)\sharp\mu, (t\theta)\sharp\nu) = t\text{GW}_p(\theta\sharp\mu, \theta\sharp\nu)$ for $t > 0$. We can rewrite Max-SW as:

$$\begin{aligned}
\text{Max-SW}_p^p(\mu_n, \mu) &= \max_{\theta \in \mathbb{R}^d : \|\theta\| = 1} \int_0^1 |F_{n,\theta}^{-1}(z) - F_\theta^{-1}(z)|^p dz \\
&\leq \text{diam}(\mathcal{X}) \max_{x \in \mathbb{R}, \theta \in \mathbb{R}^d : \|\theta\| \leq 1} |F_{n,\theta}(x) - F_\theta(x)|^p \\
&= \text{diam}(\mathcal{X}) \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|,
\end{aligned}$$

where $\mathcal{A}$ is the set of half-spaces $\{z \in \mathbb{R}^d : \theta^\top z \leq x\}$ for all $\theta \in \mathbb{R}^d$ such that $\|\theta\| \leq 1$. From VC inequality (Theorem 12.5 in Devroye et al. [2013]), we have

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\mu_n(B) - \mu(A)| > t\right) \leq 8S(\mathcal{A}, n)e^{-nt^2/32}.$$

with $S(\mathcal{A}, n)$ is the growth function. From the Sauer Lemma (Proposition 4.18 in Wainwright [2019]), the growth function is upper bounded by $(n+1)^{VC(\mathcal{A})}$. Moreover, we can get $VC(\mathcal{A}) = d+1$ from Example 4.21 in Wainwright [2019].

Let $8S(\mathcal{A}, n)e^{-nt^2/32} \leq \delta$, we have $t^2 \geq \frac{32}{n}\log\left(\frac{8S(\mathcal{A},n)}{\delta}\right)$. Therefore, we obtain

$$\mathbb{P}\left(\sup_{A \in \mathcal{B}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{32}{n}\log\left(\frac{8S(\mathcal{A}, n)}{\delta}\right)}\right) \geq 1 - \delta,$$

Using the Jensen inequality and the tail sum expectation for non-negative random variable, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right] \\
&\leq \sqrt{\mathbb{E}\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right]^2} = \sqrt{\int_0^\infty \mathbb{P}\left(\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right)^2 > t\right) dt} \\
&= \sqrt{\int_0^u \mathbb{P}\left(\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right)^2 > t\right) dt + \int_u^\infty \mathbb{P}\left(\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right)^2 > t\right) dt} \\
&\leq \sqrt{\int_0^u 1 dt + \int_u^\infty 8S(\mathcal{A}, n)e^{-nt/32} dt} = \sqrt{u + 256S(\mathcal{A}, n)\frac{e^{-nu/32}}{n}}.
\end{aligned}$$

Since the inequality holds for any u, we search for the best u that makes the inequality tight. Let $f(u) = u + 256S(\mathcal{A}, n)\frac{e^{-nu/32}}{n}$, we have $f'(u) = 1 + 8S(\mathcal{A}, n)e^{-nu/32}$. Setting $f'(u) = 0$, we obtain the minima $u^\star = \frac{32\log(8S(\mathcal{A},n))}{n}$. Plugging $u^\star$ in the inequality, we obtain:

$$\mathbb{E}\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right] \leq \sqrt{\frac{32\log(8S(\mathcal{A}, n))}{n} + 32} \leq C'\sqrt{\frac{(d+1)\log(n+1)}{n}},$$

by using Sauer Lemma i.e., $S(\mathcal{A}, n) \leq (n+1)^{VC(\mathcal{A})} \leq (n+1)^{d+1}$. Putting the above results together leads to

$$\mathbb{E}[\text{Max-SW}_p(\mu_n, \mu)] \leq C' \sqrt{(d+1)\log n / n},$$

where $C' > 0$ is some universal constant. As a consequence, we obtain the conclusion of the proof.

A.4 Proof of Proposition 4.10

For any $p \geq 1$, $d \geq 1$, $\sigma_\kappa \in \mathcal{P}(\mathbb{S}^{d-1})$, and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, using Jensen's inequality, we have:

$$\begin{aligned}
&|\mathbb{E}[\widehat{\text{RASGW}}_p^p(\mu, \nu; \sigma_\kappa)] - \text{RASGW}_p^p(\mu, \nu; \sigma_\kappa)| \\
&\leq \mathbb{E}|\widehat{\text{RASGW}}_p^p(\mu, \nu; \sigma_\kappa) - \text{RASGW}_p^p(\mu, \nu; \sigma_\kappa)| \\
&\leq \left(\mathbb{E}|\widehat{\text{RASGW}}_p^p(\mu, \nu; \sigma_\kappa) - \text{RASGW}_p^p(\mu, \nu; \sigma_\kappa)|^2\right)^{\frac{1}{2}} \\
&= \left(\mathbb{E}\left(\frac{1}{M}\sum_{l=1}^{M} \text{GW}_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu) - \mathbb{E}_{\theta \sim \sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa)}\left[\text{GW}_p^p\left(\theta \sharp \mu, \theta \sharp \nu\right)\right]\right)^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}[\frac{1}{M}\sum_{l=1}^{M} \text{GW}_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu)] &= \frac{1}{M}\sum_{l=1}^{M} \mathbb{E}[\text{GW}_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu)] \\
&= \frac{1}{M}\sum_{l=1}^{M} \widehat{\text{RASGW}}_p^p(\mu, \nu; \sigma_\kappa) = \widehat{\text{RASGW}}_p^p(\mu, \nu; \sigma_\kappa),
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}|\widehat{\text{RASGW}}_p^p(\mu, \nu; \sigma_\kappa) - \text{RASGW}_p^p(\mu, \nu; \sigma_\kappa)| &\leq \left(Var_{\theta \sim \sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa)}\left[\frac{1}{M}\sum_{l=1}^{M} \text{GW}_p^p\left(\theta_l \sharp \mu, \theta_l \sharp \nu\right)\right]\right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{M}} Var\left[\text{GW}_p^p\left(\theta \sharp \mu, \theta \sharp \nu\right)\right]^{\frac{1}{2}}.
\end{aligned}$$

Now, since $\theta_1, \ldots, \theta_M \overset{i.i.d}{\sim} \sigma_{\text{RA}}(\theta; \mu, \nu, \sigma_\kappa)$, this completes the proof.

B ADDITIONAL BASELINES

In this section, we present two baselines constructed based on similar concepts in the Sliced-Wasserstein literature. We adapt them to the Sliced Gromov-Wasserstein setup to provide a clearer exposition of the advantages of our distance. These baselines operate in addition to SGW and Max-SGW, which already exist. The first baseline is based on a similar premise of identifying an appropriate slicing distribution, but uses a computationally expensive optimization procedure to obtain it. On the other hand, the second baseline is optimization-free, but its main computational bottleneck is sampling from the proposed distribution.

Distributional Sliced Gromov-Wasserstein. To choose an improved slicing distribution, Nguyen et al. [2021] introduced the distributional sliced Wasserstein (DSW) distance. Adapting this to our context, we define the distributional sliced Gromov-Wasserstein (DSGW) distance, which is defined between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ as:

$$\text{DSGW}_p^p(\mu, \nu) = \max_{\psi \in \Psi} \mathbb{E}_{\theta \sim \sigma_\psi(\theta)}[\text{GW}_p^p(\theta \sharp \mu, \theta \sharp \nu)], \tag{4}$$

where $\sigma_\psi(\theta) \in \mathcal{P}(\mathbb{S}^{d-1})$, for example, an implicit distribution [Nguyen et al., 2021], the von Mises-Fisher [Jupp and Mardia, 1979] (vMF) distribution, and the Power Spherical (PS) [Cao et al., 2018] distribution with unknown

location parameter $\sigma_\psi(\theta) := (\text{PS})\text{vMF}(\theta|\epsilon,\kappa)$, with $\psi = \epsilon$. By performing $T \geq 1$ iterations of (projected) stochastic (sub)-gradient ascent to estimate the parameter $\hat{\psi}_T$, Monte Carlo samples $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \sigma_{\hat{\psi}_T}(\theta)$ are used to approximate DSGW. The time and space complexities of DSGW are at best $\mathcal{O}(LTn\log n + LTdn)$ and $\mathcal{O}(Ld + Ln)$, respectively, excluding the complexities of sampling from $\sigma_{\hat{\psi}_T}(\theta)$. As the concentration parameter $\kappa \to \infty$, both the vMF and PS distributions converge to the Dirac distribution, giving the max sliced Gromov-Wasserstein distance [Deshpande et al., 2019]. The Max-SGW involves performing $T \geq 1$ iterations of (projected) (sub)-gradient ascent to find the approximate "max" projection direction $\hat{\theta}_T$. The estimated value of Max-SGW is then set as $\text{GW}_p(\hat{\theta}_T\sharp\mu, \hat{\theta}_T\sharp\nu)$. Notably, the optimization problem is non-convex [Nietert et al., 2022], implying that the global optimum $\theta^\star$ cannot be achieved. Therefore, Max-SGW approximations do not constitute a metric, even as $T \to \infty$. The time and space complexities of Max-SGW are again at best $\mathcal{O}(Tn\log n + Tdn)$ and $\mathcal{O}(d+n)$, respectively.

Energy-driven Sliced Gromov-Wasserstein. In practical applications, the computational effort of optimization often surpasses that of sampling from a fixed slicing distribution. To circumvent costly optimization processes, recent work [Nguyen and Ho, 2023] introduces a novel approach that leverages an energy-driven slicing distribution, bypassing optimization altogether. While this was originally developed within the sliced Wasserstein framework, we have adapted it for our context. The energy-driven sliced Gromov-Wasserstein (EBSGW) distance between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as follows:

$$\text{EBSGW}_p^p(\mu,\nu;f) = \mathbb{E}_{\theta\sim\sigma_{\mu,\nu}(\theta;f,p)}\left[\text{GW}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right], \tag{5}$$

where $f : [0,\infty) \to (0,\infty)$ is an increasing energy function such as $f(x) = e^x$, and $\sigma_{\mu,\nu}(\theta;f,p) \propto f(\text{GW}_p^p(\theta\sharp\mu,\theta\sharp\nu))$. EBSGW can be estimated using importance sampling with the uniform proposal distribution $\sigma_0 = \mathcal{U}(\mathbb{S}^{d-1})$. For $\theta_1,\ldots,\theta_L \overset{i.i.d}{\sim} \sigma_0(\theta)$, the estimate is given by:

$$\widehat{\text{EBSGW}}_p^p(\mu,\nu;f,L) = \sum_{l=1}^{L} \text{GW}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\hat{w}_{\mu,\nu,\sigma_0,f,p}(\theta_l), \tag{6}$$

where $\hat{w}_{\mu,\nu,\sigma_0,f,p}(\theta_l) = \frac{w_l(f;\mu,\nu)/\sigma_0(\theta)}{\sum_{l'=1}^{L} w_{l'}(f;\mu,\nu)/\sigma_0(\theta)}$ are the normalized importance weights, and $\frac{w_l(f;\mu,\nu)}{\sigma_0(\theta)}$ represents the importance weight function. The computational complexity in time and space for calculating EBSGW is at best $\mathcal{O}(Ln\log n + Ldn)$ and $\mathcal{O}(Ld + Ln)$, respectively. For a detailed discussion regarding the complexity and its relation to optimality, see Section D. However, this estimation introduces bias, meaning $\mathbb{E}[\widehat{\text{EBSGW}}_p^p(\mu,\nu;f,L)] \neq \text{EBSGW}_p^p(\mu,\nu;f)$. Beyond importance sampling, Markov Chain Monte Carlo (MCMC) methods can also approximate EBSGW, although they carry a significant computational burden [Nguyen and Ho, 2023].

In the context of competing methods, we mention the approximately linear algorithm to estimate GW (between measures in their ambient spaces) proposed by Scetbon et al. [2022]. Given that it employs low rank approximations to replace complexity-heavy computations underlying GW, e.g., Sinkhorn scaling, it is inherently hostile to gradient computations (non-smooth, not differentiable), and hence unsuitable to be used as an objective function in generative models (e.g., GWGAN). In the absence of a dedicated nested optimization scheme, it also becomes numerically unstable (similar to the vanilla GW [Bunne et al., 2019])during training when used as a loss.

C MORE PRELIMINARIES

We briefly revisit the concept of the von Mises-Fisher distribution.

Definition C.1. The von Mises–Fisher distribution (vMF) represents a probability distribution that spans the unit sphere $\mathbb{S}^{d-1}$, characterized by the following density function [Jupp and Mardia, 1979]:

$$f(x|\epsilon,\kappa) := C_d(\kappa)\exp(\kappa\epsilon^\top x), \tag{7}$$

where $\kappa \geq 0$ serves as the concentration parameter, $\epsilon \in \mathbb{S}^{d-1}$ is designated as the location vector, and $C_d(\kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$ functions as the normalization constant. The term I_v refers to the modified Bessel function of the first kind with order v [Temme, 2011].

Algorithm 1 Sampling from a mixture of vMF distributions

Require: The number of vMF components k, location $\{\epsilon_i\}_{i=1}^k$, concentration $\{\kappa_i\}_{i=1}^k$, mixture weights $\{\alpha_i\}_{i=1}^k$, dimension d, unit vector $e_1 = (1, 0, .., 0)$.

Sample index $i \sim \text{Categorical}(\alpha_1, ..., \alpha_k)$
Sample $v \sim \mathcal{U}(\mathbb{S}^{d-2})$
$b \leftarrow \frac{-2\kappa_i + \sqrt{4\kappa_i^2 + (d-1)^2}}{d-1}$, $a \leftarrow \frac{(d-1) + 2\kappa_i + \sqrt{4\kappa_i^2 + (d-1)^2}}{4}$, $m \leftarrow \frac{4ab}{(1+b)} - (d-1)\ln(d-1)$
repeat
  Sample $\psi \sim \text{Beta}\left(\frac{1}{2}(d-1), \frac{1}{2}(d-1)\right)$
  $\omega \leftarrow h(\psi, \kappa_i) = \frac{1-(1+b)\psi}{1-(1-b)\psi}$
  $t \leftarrow \frac{2ab}{1-(1-b)\psi}$
  Sample $u \sim \mathcal{U}(0, 1)$
until $(d-1)\ln(t) - t + m \geq \ln(u)$
$h_1 \leftarrow (\omega, \sqrt{1-\omega^2} v^\top)^\top$
$\epsilon' \leftarrow e_1 - \epsilon_i$
$u = \frac{\epsilon'}{\|\epsilon'\|}$
$U = \mathbb{I} - 2uu^\top$
return Uh_1

The vMF distribution centers around the mode ϵ, and its density diminishes as x moves further from ϵ. As $\kappa \to 0$, vMF transitions to a uniform distribution, while for $\kappa \to \infty$, it converges to a Dirac distribution located at ϵ [Sra, 2016]. Details of the sampling procedure with a mixture of vMF distributions are placed in Algorithm 1.

We proceed to introduce the power spherical distribution [De Cao and Aziz, 2020]:

Definition C.2. The power spherical distribution (PS) describes a probability distribution on the unit sphere $\mathbb{S}^{d-1}$ with the following density function:

$$f(x|\epsilon, \kappa) = C(\kappa, d)(1 + \epsilon^\top x)^\kappa, \tag{8}$$

here, $\kappa \geq 0$ serves as the concentration parameter, $\epsilon \in \mathbb{S}^{d-1}$ is the location vector, and the normalization constant $C_d(\kappa)$ is defined as $\left\{2^{d-1+\kappa}\pi^{\frac{d-1}{2}}\frac{\Gamma(\frac{d-1}{2}+\kappa)}{\Gamma(d-1+\kappa)}\right\}^{-1}$.

When $\kappa \to 0$, the power spherical distribution tends to the uniform distribution because, for $\kappa \in (0, 1]$, $f(x|\epsilon, \kappa)$ remains uniformly bounded. Moreover, for all x not equal to $-\epsilon$

$$\lim_{\kappa \to 0} f(x|\epsilon, \kappa) = C(0, d),$$

indicating it matches the uniform distribution's density on $\mathbb{S}^{d-1}$. Thus, by invoking the Lebesgue dominated convergence theorem, for any bounded function g on $\mathbb{S}^{d-1}$, it holds that

$$\lim_{\kappa \to 0} \int_{\mathbb{S}^{d-1}} g(x) f(x|\epsilon, \kappa) dx = \int_{\mathbb{S}^{d-1}} g(x) \mathcal{U}(\mathbb{S}^{d-1}) dx,$$

where $\mathcal{U}(\mathbb{S}^{d-1})$ denotes the uniform distribution on $\mathbb{S}^{d-1}$. This indicates that the density f of the power spherical distribution converges to that of the uniform distribution as $\kappa \to 0$. Conversely, as $\kappa \to \infty$, the function f at its mode ϵ increases unboundedly, and the following conclusion holds for any uniformly bounded function g on $\mathbb{S}^{d-1}$.

$$\lim_{\kappa \to \infty} \int_{\mathbb{S}^{d-1}} g(x) f(x|\epsilon, \kappa) dx = \int_{\mathbb{S}^{d-1}} g(x) \delta_\epsilon(x) dx.$$

It is true since for $x = \epsilon$, and $\kappa \to \infty$ we have

$$f(\epsilon|\epsilon, \kappa) = 2^\kappa \frac{\Gamma(d-1+\kappa)}{\Gamma(\frac{d-1}{2}+\kappa)} \pi^{\frac{1-d}{2}} 2^{1-d-\kappa} = \pi^{\frac{1-d}{2}} 2^{1-d} \frac{\Gamma(d-1+\kappa)}{\Gamma(\frac{d-1}{2}+\kappa)} \to \infty.$$

Algorithm 2 Sampling from power spherical distribution

Require: location parameter ϵ, concentration κ, dimension d, unit vector $e_1 = (1, 0, .., 0)$.

1: Sample $z \sim \text{Beta}(\frac{(d-1)}{2} + \kappa, \frac{(d-1)}{2})$
2: Sample $v \sim \mathcal{U}(\mathbb{S}^{d-2})$
3: $w \leftarrow 2z - 1$
4: $h_1 \leftarrow (\omega, \sqrt{1-\omega^2} v^\top)^\top$
5: $\epsilon' \leftarrow e_1 - \epsilon$
6: $u = \frac{\epsilon'}{\|\epsilon'\|}$
7: $U = \mathbb{I} - 2uu^\top$
8: **return** Uh_1

Consequently, as $\kappa \to \infty$, the density of the power spherical distribution with a location vector ϵ approaches the Dirac delta measure at ϵ.

Power Spherical Distribution Sampling Procedure. We revisit the sampling technique for the power spherical (PS) distribution, as outlined in Algorithm 2 in De Cao and Aziz [2020]. A key distinction between sampling from the PS and the von Mises-Fisher (vMF) distribution is that PS does not require rejection sampling, unlike the vMF's approach as detailed in Algorithm 1. Thus, PS sampling is more efficient than vMF sampling. Moreover, parameter gradient estimation for PS density is simpler compared to vMF since the entire PS sampling algorithm is differentiable, including sampling from the Beta distribution using the implicit reparameterization trick [Figurnov et al., 2018].

Radon Transform. The classical Radon transform, symbolized by $\mathcal{R}$, converts a function $I \in L^1(\mathbb{R}^d)$— where

$$L^1(\mathbb{R}^d) = \{I : \mathbb{R}^d \to \mathbb{R} \mid \int_{\mathbb{R}^d} |I(x)| dx < \infty\}$$

— into an infinite collection of its integrals over the hyperplanes of $\mathbb{R}^d$, as defined by

$$\mathcal{R}I(t, \theta) = \int_{\mathbb{R}^d} I(x)\delta(t - \langle x, \theta \rangle) dx, \tag{9}$$

for $(t, \theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$. Here, $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ represents the unit sphere in d dimensions, $\delta(\cdot)$ is the Dirac delta function, and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner-product. The mapping $\mathcal{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d-1})$ is established. A typical hyperplane is described as:

$$H(t, \theta) = \{x \in \mathbb{R}^d \mid \langle x, \theta \rangle = t\}, \tag{10}$$

which can be viewed as a level set for the function $g : \mathbb{R}^d \times \mathbb{S}^{d-1} \to \mathbb{R}$ defined by $g(x, \theta) = \langle x, \theta \rangle$. For any fixed θ, the integrals over hyperplanes perpendicular to θ yield a continuous function $\mathcal{R}I(\cdot, \theta) : \mathbb{R} \to \mathbb{R}$, acting as a projection (or slice) of I. The Radon transform is a linear bijective mapping [Helgason, 2011], with its inverse $\mathcal{R}^{-1}$ given by:

$$\begin{aligned} I(x) &= \mathcal{R}^{-1}\big(\mathcal{R}I(t, \theta)\big) \\ &= \int_{\mathbb{S}^{d-1}} (\mathcal{R}I(\langle x, \theta \rangle, \theta) * \eta(\langle x, \theta \rangle)) d\theta \end{aligned} \tag{11}$$

where $\eta(\cdot)$ stands for a one-dimensional high-pass filter whose Fourier transform is $\mathcal{F}\eta(\omega) = c|\omega|^{d-1}$, as justified by the Fourier slice theorem [Helgason, 2011] (see supplementary materials), and '$*$' signifies the convolution operation. This formulation of the inverse Radon transform is known as the filtered back-projection technique, extensively used in the reconstruction of images in the field of biomedical imaging. Conceptually, each one-dimensional projection (or slice) $\mathcal{R}I(\cdot, \theta)$ is first filtered using a high-pass filter and then back-projected into $\mathbb{R}^d$ along $H(\cdot, \theta)$ to reconstruct an approximation of I. The final image I is rebuilt by combining all such back-projected approximations. In practical scenarios, obtaining an infinite number of projections is impossible, thus, the filtered back-projection method replaces integral calculations with a finite sum of projections, akin to a Monte Carlo estimation.

Sliced Wasserstein Distance. The sliced p-Wasserstein distance approach computes a series of one-dimensional perspectives of a multidimensional probability distribution using linear projections, facilitated by the Radon transform. The distance between two distributions is then determined by evaluating the p-Wasserstein distance of these one-dimensional representations, i.e., their marginal distributions. Formally, the sliced p-Wasserstein distance between I_μ and I_ν is expressed as:

$$SW_p(I_\mu, I_\nu) = \left(\int_{\mathbb{S}^{d-1}} W_p^p\big(\mathcal{R}I_\mu(.,\theta), \mathcal{R}I_\nu(.,\theta)\big)\, d\theta \right)^{\frac{1}{p}}. \tag{12}$$

This serves as a true distance function, complying with positive-definiteness, symmetry, and the triangle inequality conditions [Bonnotte, 2013, Kolouri et al., 2016]. The SW distance is computed through integration over the unit sphere in $\mathbb{R}^d$. In practice, this is approximated using a Monte Carlo method that samples $\{\theta_l\}$ uniformly from $\mathbb{S}^{d-1}$ and then replaces the integral with an average over these samples:

$$SW_p(I_\mu, I_\nu) \approx \left(\frac{1}{L} \sum_{l=1}^{L} W_p^p\big(\mathcal{R}I_\mu(\cdot,\theta_l), \mathcal{R}I_\nu(\cdot,\theta_l)\big) \right)^{\frac{1}{p}} \tag{13}$$

The sliced p-Wasserstein distance is beneficial in practice: if $\mathcal{R}I_\mu(\cdot,\theta_l)$ and $\mathcal{R}I_\nu(\cdot,\theta_l)$ can be determined for each sampled θ_l, then the SW distance is derived from solving several one-dimensional optimal transport problems with closed-form solutions. This is particularly advantageous when estimating a high-dimensional PDF I, since one-dimensional kernel density estimation, necessary for PDF slices, simplifies the process. However, as dimensionality increases, more projections are needed to accurately estimate I from $\mathcal{R}I(\cdot,\theta)$. Essentially, a two-dimensional distribution requiring L projections demands approximately $\mathcal{O}(L^{d-1})$ projections for a similarly smooth d-dimensional distribution for $d \geq 2$. For clarification, consider $I_\mu = \mathcal{N}(0, I_d)$ and $I_\nu = \mathcal{N}(x_0, I_d)$, $x_0 \in \mathbb{R}^d$, as two multivariate Gaussian densities with identity covariance matrices. Their projections result in one-dimensional Gaussian distributions $\mathcal{R}I_\mu(\cdot,\theta) = \mathcal{N}(0,1)$ and $\mathcal{R}I_\mu(\cdot,\theta) = \mathcal{N}(\langle\theta, x_0\rangle, 1)$. Here, $W_2(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta))$ achieves its peak when $\theta = \frac{x_0}{\|x_0\|_2}$ and measures zero for θ orthogonal to x_0. Randomly chosen vectors from the unit sphere tend to be nearly orthogonal in high dimensions. More precisely, the inequality $Pr(|\langle\theta, \frac{x_0}{\|x_0\|_2}\rangle| < \epsilon) > 1 - e^{(-d\epsilon^2)}$ indicates that in high dimensions, most sampled θs will be nearly orthogonal to x_0, leading to $W_2(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)) \approx 0$ with high probability. To counteract this, one can skip uniform sampling of the unit sphere, selecting θs that contain significant distinguishing information between I_μ and I_ν. This approach is used by Deshpande et al. [2018], who first computes a linear discriminant subspace and then assesses the empirical SW distance by choosing the θs as the subspace's discriminant components. A similar but less empirical method involves the maximum sliced p-Wasserstein (max-SW) distance, an alternative OT metric, defined as:

$$\text{max-}SW(I_\mu, I_\nu) = \max_{\theta \in \mathbb{S}^{d-1}} W_p\big(\mathcal{R}I_\mu(\cdot,\theta), \mathcal{R}I_\nu(\cdot,\theta)\big). \tag{14}$$

D ALGORITHMS

Computational algorithms of RASGW and IWRASGW. We present the pseudo-codes for computing RASGW and IWRASGW with Monte Carlo estimation in Algorithm 3 and Algorithm 4.

Algorithm 3 Computational algorithm of RASGW

Require: Probability measures μ and ν, $p \geq 1$, the number of projections L, $0 < \kappa < \infty$

1: **for** $l = 1$ to L **do**
2: Sample $X, X' \sim \mu$, $Y, Y' \sim \nu$
3: Sample $\theta_l \sim D_\kappa(X, X', Y, Y')$
4: Compute $v_l = \mathrm{GW}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
5: **end for**
6: Compute $\widehat{\mathrm{RASGW}}_p(\mu, \nu; L, \sigma_\kappa) = \left(\frac{1}{L}\sum_{l=1}^{L} v_l\right)^{\frac{1}{p}}$
7: **return** $\widehat{\mathrm{RASGW}}_p(\mu, \nu; L, \sigma_\kappa)$

Algorithm 4 Computational algorithm of the IWRASGW

Require: Probability measures μ and ν, $p \geq 1$, the number of projections L, $0 < \kappa < \infty$, and the energy function f.

1: **for** $l = 1$ to L **do**
2: Sample $X, X' \sim \mu$, $Y, Y' \sim \nu$
3: Sample $\theta_l \sim D_\kappa(X, X', Y, Y')$
4: Compute $v_l = \mathrm{GW}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
5: Compute $w_l = f(\mathrm{GW}_p(\theta_l \sharp \mu, \theta_l \sharp \nu))$
6: **end for**
7: Compute $\widehat{\mathrm{IWRASGW}}_p(\mu, \nu; \sigma_\kappa, L, f) = \left(\sum_{l=1}^{L} v_l \frac{w_l}{\sum_{i=1}^{L} w_i} \right)^{\frac{1}{p}}$
8: **return** $\widehat{\mathrm{IWRASGW}}_p(\mu, \nu; \sigma_\kappa, L, f)$

Computation of GW in 1D. Let us consider input distributions $\mu \in \mathcal{P}_4(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_4(\mathbb{R}^{d'})$ such that the projections $\mu_\theta \coloneqq \theta_\# \mu$ (similarly define $(\Delta \sharp \nu)_\theta$), given that $\theta \sim \sigma_{\mathrm{RA}}$ preserve finiteness in moments. The equality is understood setwise. Observe that, calculating $\mathrm{RASGW}_2^2(\mu, \nu)$ essentially underlies computing $\mathrm{GW}_2^2(\mu_\theta, (\Delta \sharp \nu)_\theta)$, given replicates of θ. Now, if both μ and ν are centered (without loss of generality), the following decomposition holds due to Zhang et al. [2024]

$$\mathrm{GW}_2^2(\mu_\theta, (\Delta \sharp \nu)_\theta) = S_1(\mu_\theta, (\Delta \sharp \nu)_\theta) + S_2(\mu_\theta, (\Delta \sharp \nu)_\theta),$$

such that

$$S_1(\mu_\theta, (\Delta \sharp \nu)_\theta) = \int \big(\theta(x) - \theta(x')\big)^4 d\mu^{\otimes 2} + \int \big(\theta(y) - \theta(y')\big)^4 d(\Delta \sharp \nu)^{\otimes 2} - 4 \int \theta(x)^2 \theta(y)^2 d\mu \otimes (\Delta \sharp \nu) \tag{15}$$

$$S_2(\mu_\theta, (\Delta \sharp \nu)_\theta) = \inf_{a \in [0.5W_-, 0.5W_+]} 32a^2 + \inf_{\pi \in \Pi(\mu_\theta, (\Delta \sharp \nu)_\theta)} \int (-4x^2y^2 - 32axy) d\pi, \tag{16}$$

where $W_- = \inf_{\pi \in \Pi(\mu, \Delta \sharp \nu)} \int \theta(x)\theta(y)\, d\pi(x, y)$, $W_+ = \sup_{\pi \in \Pi(\mu, \Delta \sharp \nu)} \int \theta(x)\theta(y)\, d\pi(x, y)$, and we write $\theta(\cdot) \coloneqq \langle \theta, \cdot \rangle$. Since S_1 is independent of the coupling π, solving the RASGW problem boils down to solving a constrained OT (as given in (16)) once we have identified an optimal a^*. As such, there may indeed be linear solvers for sliced GW variants. However, the question arises whether they exist uniformly corresponding to all optimal a^*'s in the feasible set, as Beinert et al. [2023] construct cases where the *cyclic* permutation attains a lower GW cost compared to *identity* and *anti-identity* permulations, rendering the latter two sub-optimal in 1D. Here, note that the boundaries (involving W_- and W_+) rather depend on sliced OT costs. In this line, Zhang et al. [2024] shows that in an empirical setup, solving the GW_2^2 distance, i.e.,

$$\frac{1}{n^2} \min_\sigma \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \left| \theta(x)_i - \theta(x)_j \right|^2 - \left| \theta(y)_{\sigma(i)} - \theta(y)_{\sigma(j)} \right| \right|^2, \tag{17}$$

optimally using $\sigma =$ identity or anti-identity permutation is possible if and only if $a^* \in \{\frac{W_-}{2}, \frac{W_+}{2}\}$. Here, $\theta(x)_i$ and $\theta(y)_i$ correspond to the ith order statistic in a pool of n projected samples from μ and $\Delta \sharp \nu$, respectively. As such, the counterexample Beinert et al. [2023] provides belong to the cases other than the boundaries of a. The following result summarizes the discussion as a sufficient condition on the slices θ such that the computational complexities, as mentioned is Section 4.2, are achieved.

Proposition D.1 (Zhang et al. [2024]). *Given n independent samples each from $\mu \in \mathcal{P}_4(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_4(\mathbb{R}^{d'})$, $d' < d$; solving the problem (17) incurs time and space complexities as $\mathcal{O}(n \log n + dn)$ and $\mathcal{O}(d + n)$ respectively, if $\theta \sim \sigma_{RA}$ is such that the optima in (16) is achieved within $\{\frac{W_-}{2}, \frac{W_+}{2}\}$.*

The proposition recasts the sufficiency in finding an appropriate permutation as a sufficient criteria over finding the slice that imposes the same alignment. Given that σ_{RA} can be arbitrarily reparametrized without increasing sampling complexity, finding such a sufficient direction to achieve the desired computational complexity in the corresponding RASGW is feasible.

An immediate future direction may look into the concentration of such useful directions among σ_{RA}. To this end, an encouraging observation is that the optimal θ corresponding to the coupling $\pi \in \Pi(\mu_\theta, (\Delta \sharp \nu)_\theta)$ (corresponding

Algorithm 5 Relation-Aware Sliced Gromov-Wasserstein between discrete measures in an Euclidean setup

Require: $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i} \in \mathcal{P}(\mathbb{R}^d)$ and $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i} \in \mathcal{P}(\mathbb{R}^{d'})$, given $d' < d$; number of Monte Carlo samples M; location-scale distribution σ_κ on $\mathbb{S}^{d-1}$, $0 < \kappa < \infty$

$y_i \leftarrow \Delta(y_i)$, $\forall i$
for $l = 1, \ldots, M$ **do**
  Sample $\theta_l \sim \sigma_{\mathrm{RA}}(\theta; \mu, \nu, \sigma_\kappa)$ {Definition 4.4}
  Sort $(\langle x_i, \theta_l\rangle)_i$ and $(\langle y_j, \theta_l\rangle)_j$ in increasing order
  Solve $\min_{\sigma\in S_n} \sum_{i,j} -\big(\theta_l(x)_i - \theta_l(x)_j\big)^2\big(\theta_l(y)_{\sigma(i)} - \theta_l(y)_{\sigma(j)}\big)^2$, where S_n is the set of all permutations of $\{1, \ldots, n\}$ {solution: $\sigma_{\theta_l} \in$ Id or Anti-Id}
end for
return $\widehat{\mathrm{RASGW}}_2^2(\mu, \nu; \sigma_\kappa, M) = \frac{1}{n^2 M}\sum_{l=1}^{M}\sum_{i,j=1}^{n}\Big((\langle x_i - x_j, \theta_l\rangle)^2 - (\langle y_{\sigma_{\theta_l}(i)} - y_{\sigma_{\theta_l}(j)}, \theta_l\rangle)^2\Big)^2$

to W_- or W_+) can be traced back to its domain due to (i) continuity of the objective under GW in couplings [Mémoli, 2011, Bauer et al., 2025], (ii) the continuity of the map $\theta \mapsto \mathrm{GW}_p(\theta\sharp\cdot, \theta\sharp\cdot)$. The following result formalizes (ii).

Lemma D.2. *Given any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the map $f : \theta \mapsto GW_p^p(\mu_\theta, \nu_\theta)$ is Lipschitz on $\mathbb{S}^{d-1}$.*

Proof of Lemma D.2. Let $\theta, \theta' \in \mathbb{S}^{d-1}$. Due to the triangle inequality of GW [Mémoli, 2011],

$$\begin{aligned}\big|\mathrm{GW}_p(\mu_\theta, \nu_\theta) - \mathrm{GW}_p(\mu_{\theta'}, \nu_{\theta'})\big| &\le \mathrm{GW}_p(\mu_\theta, \mu_{\theta'}) + \mathrm{GW}_p(\nu_{\theta'}, \nu_\theta)\\ &\lesssim \mathrm{W}_p(\mu_\theta, \mu_{\theta'}) + \mathrm{W}_p(\nu_{\theta'}, \nu_\theta).\end{aligned}$$

Moreover,

$$\mathrm{W}_p^p(\mu_\theta, \mu_{\theta'}) = \inf_{X,Y\sim\mu} \mathbb{E}\big|\theta(X) - \theta'(Y)\big|^p \le \big\|\theta - \theta'\big\|^p \mathbb{E}_\mu\|X\|^p.$$

As such,

$$\big|\mathrm{GW}_p(\mu_\theta, \nu_\theta) - \mathrm{GW}_p(\mu_{\theta'}, \nu_{\theta'})\big| \lesssim \big\|\theta - \theta'\big\| \left[\sqrt[p]{\vee_{\mu,\nu}\mathbb{E}\|X\|^p}\right].$$

□

Since the continuity holds for arbitrary slices, the lemma extends to RASGW as well. The proof of the same involves an additional usage of Jensen's inequality. However, we point out that Proposition D.1 may not hold for arbitrary p (instead of 2) and supports other than Euclidean, since the decomposition need not hold in general.

In light of the discussion, Algorithm 3 can be understood as Algorithm 5.

E ADDITIONAL EXPERIMENTAL DETAILS

The gradient calculation of other sliced GW variants follows similarly to RASGW. For example, given samples $\theta_{11,\phi}, \ldots, \theta_{HL,\phi} \sim \sigma_{\mathrm{RA}}(\theta; \mu_\phi, \nu, \kappa)$, we have

$$\nabla_\phi \mathrm{IWRASGW}_p^p(\mu_\phi, \nu; \sigma_\kappa, L, H) = \frac{1}{H}\sum_{h=1}^{H}\left[\nabla_\phi \sum_{l=1}^{L} \mathrm{GW}_p^p(\theta_{hl,\phi}\sharp\mu_\phi, \theta_{hl,\phi}\sharp\nu)\frac{w_{hl,\phi}(f; \mu_\phi, \nu)}{\sum_{j=1}^{L} w_{hj,\phi}(f; \mu_\phi, \nu)}\right].$$

E.1 Ablation Study

We observe that for a given value of κ, increasing the number of projection replicates $\theta_1, \ldots, \theta_M \sim \sigma_{\mathrm{RA}}$ (in approximating RASGW using MC estimates) indeed results in better alignment (i.e., declining overall GW_2 values). This is expected due to Proposition 4.10, as the realized bias shrinks gradually. However, that comes at the cost of increasing time required. The following table (Table 4) justifies our choice of $M = 500$ as the

corresponding alignment is satisfactory without incurring elevated time (e.g., at $M = 10000$). The ablation of L in an empirical IWRASGW setup follows a similar protocol. For κ, heuristically, the value should not be considered too large since it signifies the power of the density $\mathrm{PS}(\theta; \epsilon, \kappa) \propto (1 + \epsilon^T \theta)^\kappa$. Observe that based on the definition of Power Spherical (PS) distributions (Definition C.2), SGW can be shown to be a special case of RASGW in the limit ($\kappa \to 0$). On the other end of the limit (in $\kappa \to \infty$), the sampled direction according to PS becomes degenerated at the location, even in the high-dimensional regime: $d \to \infty$ such that $\frac{d}{\kappa} \to c$, for some $c > 0$. Two facts emerge from this. First, RASGW (and IWRASGW), based on κ, enables a trade-off in the degree of relation-awareness in slicing directions. While the increasing concentration of informative directions in high dimensions makes it harder for the fact to reflect in generated outputs, the very observation that RASGW (and IWRASGW) achieves improved perceptibility implies that $\kappa = 0$ is not an optimal choice. Second, this is a reminder that increasing κ drastically, even along d, does not guarantee enhanced performance. This is also reflected in low dimensions (Table 3), as improvement in corresponding GW values stagnates after a point. The degeneration is quite prominent, looking at Figure 1 of De Cao and Aziz [2020] (e.g., for $d = 3$). Here, we mention that the wide basin of values we conduct our ablation study over follows the protocol of De Cao and Aziz [2020] (see Figure 2(b)). It is also not straightforward to think that the value of κ would naturally be significantly higher for high-dimensional image datasets, since they typically have low intrinsic dimensionality ($\simeq 26$ for CIFAR-10 according to Pope et al. [2021]).

Our ablations (Table 3) suggest that at $\kappa = 50$, the corresponding alignment improves significantly over smaller values and does not improve that much upon further increment. Overall, the simulations reflect that the proposed semi-metric is not too sensitive to the choice of M, but needs careful tuning when it comes to κ. The inference is anticipated as κ is instrumental in choosing an appropriate relation-aware projection.

Table 3: Ablation study of Hyperparameter κ (when $M = 500$) for RASGW in 3D to 2D experiment for 10000 steps

κ	$\mathrm{GW}_2(\downarrow)$	Time (s)($\downarrow$)
1	113.43±2.46	**48.156±2.84**
5	100.10±4.02	49.98±2.45
10	67.08±3.91	50.81±2.76
50	<u>18.80</u>±4.64	51.89±4.09
100	20.41±4.17	50.09±2.23
500	19.16±5.50	49.75±2.98
1000	20.65±5.21	49.34±2.43
5000	21.04±4.188	<u>48.90±3.01</u>
10000	**18.10±2.46**	51.21±3.19
50000	20.51±2.13	50.21±2.89

Table 4: Ablation study of Hyperparameter M (when $\kappa = 50$) for RASGW in 3D to 2D experiment for 10000 steps

M	$\mathrm{GW}_2(\downarrow)$	Time (s)($\downarrow$)
1	90.42±16.11	**49.33±2.84**
5	40.23±5.50	49.90±2.45
10	36.77±5.30	<u>49.81±3.21</u>
50	29.05±5.98	50.12±3.09
100	23.83±5.95	50.54±2.23
500	18.80±4.64	51.89±4.09
1000	<u>18.32±5.21</u>	57.49±2.43
5000	19.67±4.188	115.66±3.01
10000	**17.38±2.46**	185.78±11.09

E.2 GWGAN

E.2.1 Architecture

Bunne et al. [2019] proposed an architecture consisting of two primary components: the Generator (G) and the Adversary (A), which operate within an adversarial learning framework to generate and evaluate data distributions. The Generator is designed as a four-layer *Multilayer Perceptron (MLP)* that processes a 256-dimensional input and produces a 2-dimensional output. Each hidden layer incorporates ReLU activation functions to introduce non-linearity. The Adversary is implemented as a five-layer *MLP* that processes inputs based on their dimensionality. It accepts either a 3-dimensional or 2-dimensional input, applies ReLU activations at each layer, and outputs either a 3-dimensional or 2-dimensional result, depending on the processing path. To enhance training stability, they utilized Xavier Normal Initialization for the generator to mitigate vanishing gradients, while the adversary employed Orthogonal Initialization to maintain a well-conditioned weight space.

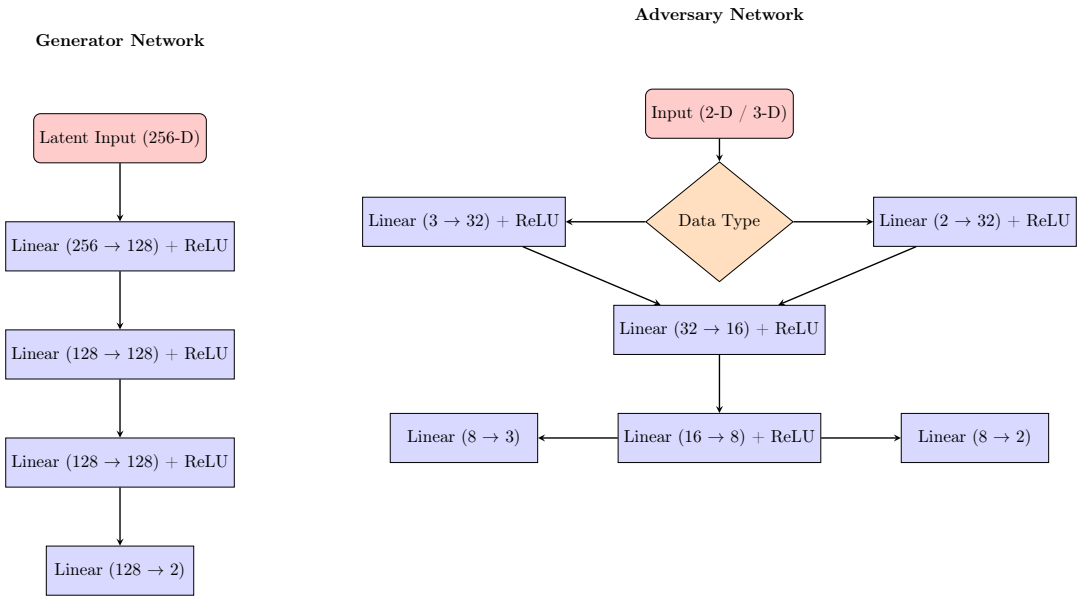


Figure 3: Flowchart of the Generator and Adversary networks.

E.2.2 For Identical Spaces:

For the experiment of generation of 2D gaussian distribution from a 2D distribution, the Adversary neural network was used along with l_1-Regularization to ensure fast convergence. This experiment was performed for 10,000 iterations with RASGW as the loss function during training.

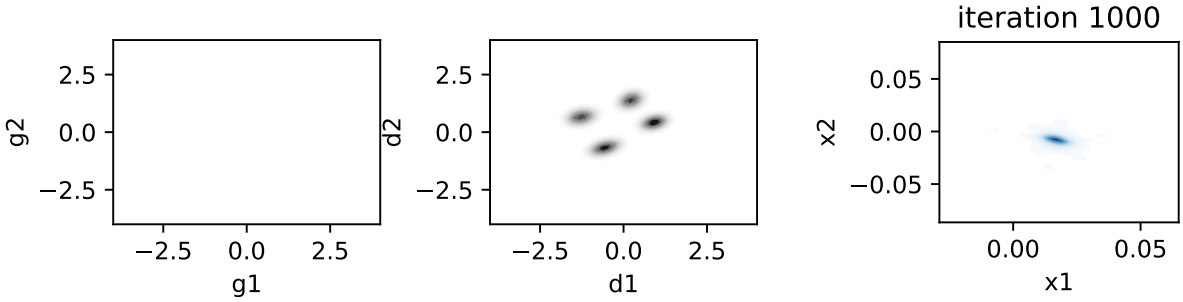


Figure 4: At iteration 1000, using RASGW as the loss function in training: **Left:** Feature space of the Generator, **Middle:** Feature space of the Adversary, **Right:** Generated image. All plots are done using 1000 samples.

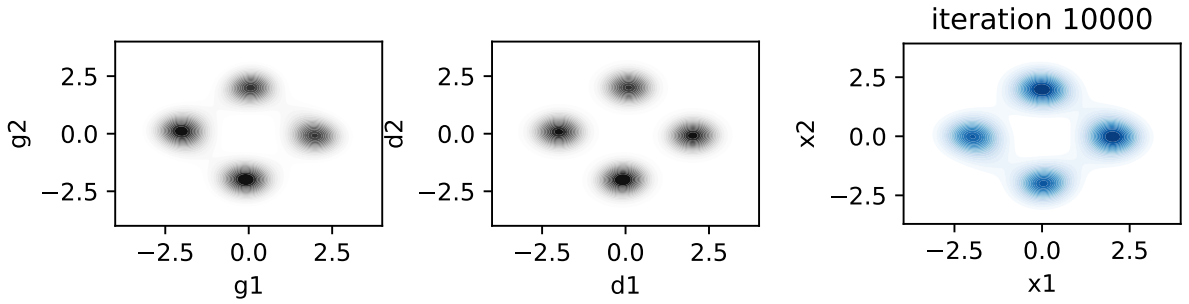


Figure 5: At iteration 10000, using RASGW as the loss function in training: **Left:** Feature space of the Generator, **Middle:** Feature space of the Adversary, **Right:** Generated image. All plots are done using 1000 samples.

Table 5: Gromov Wasserstein-2 distance and computational times across iterations in generation of 2D distribution from 3D distribution

Distance	Step 100		Step 500		Step 1000		Step 5000		Step 10000	
	$GW_2(\downarrow)$	Time (s)(↓)	$GW_2(\downarrow)$	Time (s)(↓)	$GW_2(\downarrow)$	Time (s)(↓)	$GW_2(\downarrow)$	Time (s)(↓)	$GW_2(\downarrow)$	Time (s)(↓)
SGW	142.30	**1.26**	115.64	3.22	103.29	5.53	104.91	22.34	111.32	42.82
Max-SGW	399.92	2.88	390.43	10.61	390.09	20.29	382.56	95.47	379.07	180.87
DSGW	120.25	3.07	113.71	12.04	100.23	23.62	106.12	111.33	115.54	220.90
EBSGW	195.09	1.30	100.23	**3.21**	96.36	**5.46**	82.74	**21.65**	82.20	**41.84**
RPSGW	111.93	1.29	48.85	3.25	**39.47**	5.63	40.58	22.36	49.16	43.20
IWRPSGW	136.22	1.34	**45.34**	3.55	43.79	6.15	38.78	23.47	41.08	47.97
RASGW	122.98	1.32	58.96	3.53	39.97	6.17	**30.08**	25.30	**18.21**	50.08
IWRASGW	**111.03**	1.76	55.82	**3.21**	40.46	**5.46**	45.43	**21.65**	23.10	46.04

E.2.3 For Incomparable Spaces

For the experiment on generating a 2D Gaussian distribution from a 3D Gaussian distribution or the other way around (a 3D Gaussian distribution from a 2D Gaussian distribution), the adversary neural network was not used following Bunne et al. [2019]. This experiment was also performed for 10,000 iterations with different sliced variants as the loss function during training.

E.2.4 The 4-point Experiment

The Gaussian 4 experiment generates a dataset with 4 clusters. In the 3D case, the clusters are centered at the points (1, 0, 0), (-1, 0, 0), (0, 1, 0), and (0, -1, 0). Each point is sampled from a Gaussian distribution with small noise around these centers, scaled by a factor of 2. This experiment produces 4 distinct clusters in 3D space. It is also possible to use this experiment in 2D by setting the Z-axis coordinate to 0 for each cluster center, resulting in clusters lying in the XY-plane. The points are randomly assigned to one of these clusters, with the corresponding label indicating which center the point was sampled from. Thus, both 3D and 2D versions of the experiment are possible. It was performed for 10,000 iterations.

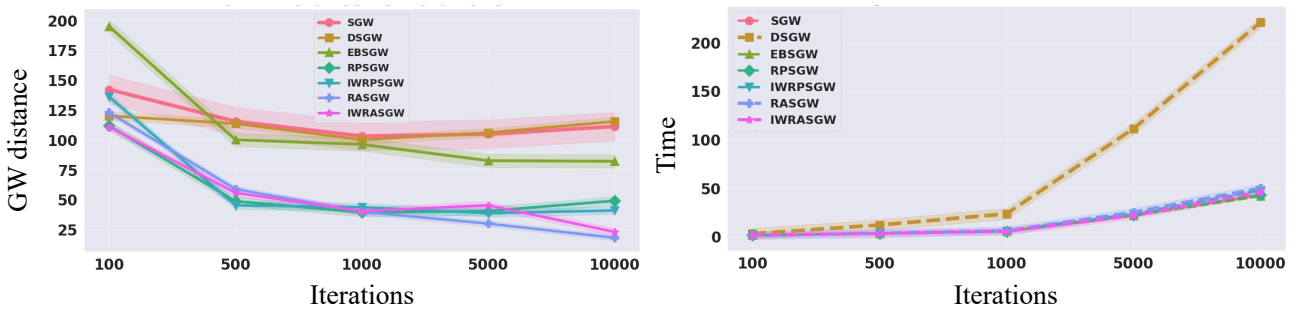


Figure 6: (*left*) GW distances and (*right*) computation time for 4-point 3D→2D generations.

Table 6: Gromov Wasserstein-2 distance and computational times across iterations in generation of 3D distribution from 2D distribution

Distance	Step 100		Step 500		Step 1000		Step 5000		Step 10000	
	$GW_2(\downarrow)$	Time (s)($\downarrow$)	$GW_2(\downarrow)$	Time (s)($\downarrow$)	$GW_2(\downarrow)$	Time (s)($\downarrow$)	$GW_2(\downarrow)$	Time (s)($\downarrow$)	$GW_2(\downarrow)$	Time (s)($\downarrow$)
SGW	20.35	1.35	3.65	**3.07**	2.10	**5.23**	1.67	**21.25**	2.59	**41.98**
Max-SGW	14.80	2.77	4.39	10.38	3.60	19.71	2.09	94.08	2.50	186.66
DSGW	13.01	3.33	3.34	12.34	3.14	23.52	3.23	111.70	1.56	222.39
EBSGW	23.10	**1.33**	4.66	3.27	3.46	5.38	3.72	21.86	1.82	42.33
RPSGW	23.04	1.39	**3.25**	3.62	3.29	6.26	2.66	26.62	1.60	52.51
IWRPSGW	23.42	1.37	5.47	3.50	**2.00**	6.13	1.35	26.69	1.47	52.04
RASGW	**12.31**	1.75	4.67	4.64	2.71	8.67	1.63	39.38	1.26	79.92
IWRASGW	29.12	1.80	5.86	4.96	2.99	9.04	**0.73**	41.00	**1.03**	78.31

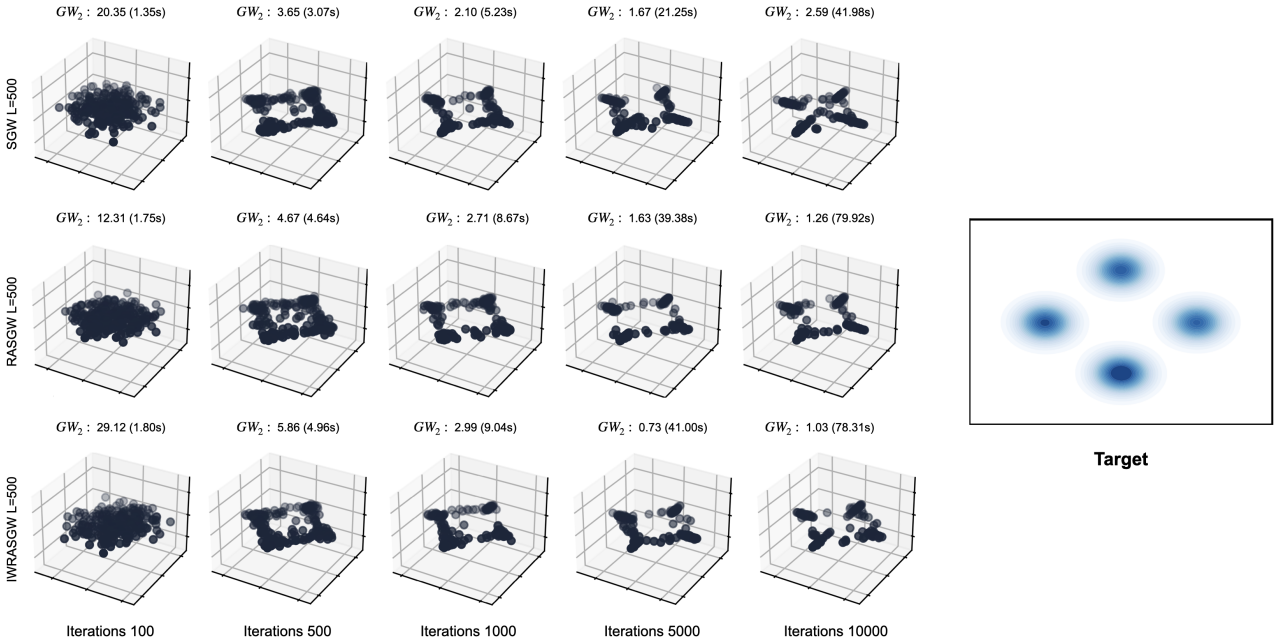


Figure 7: Generation of 3D distribution from 2D distribution for the 4-point experiment

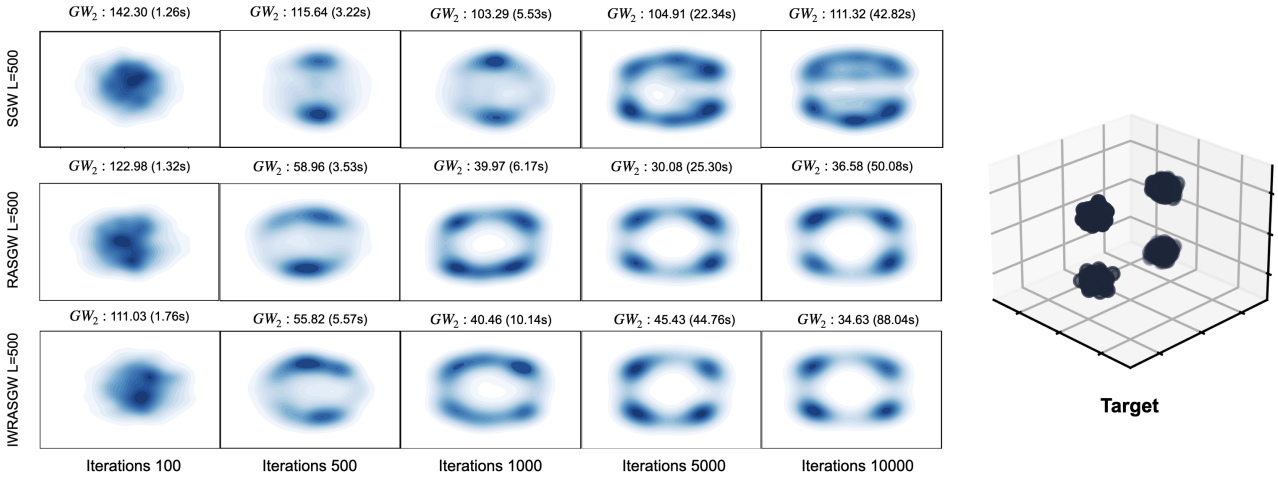
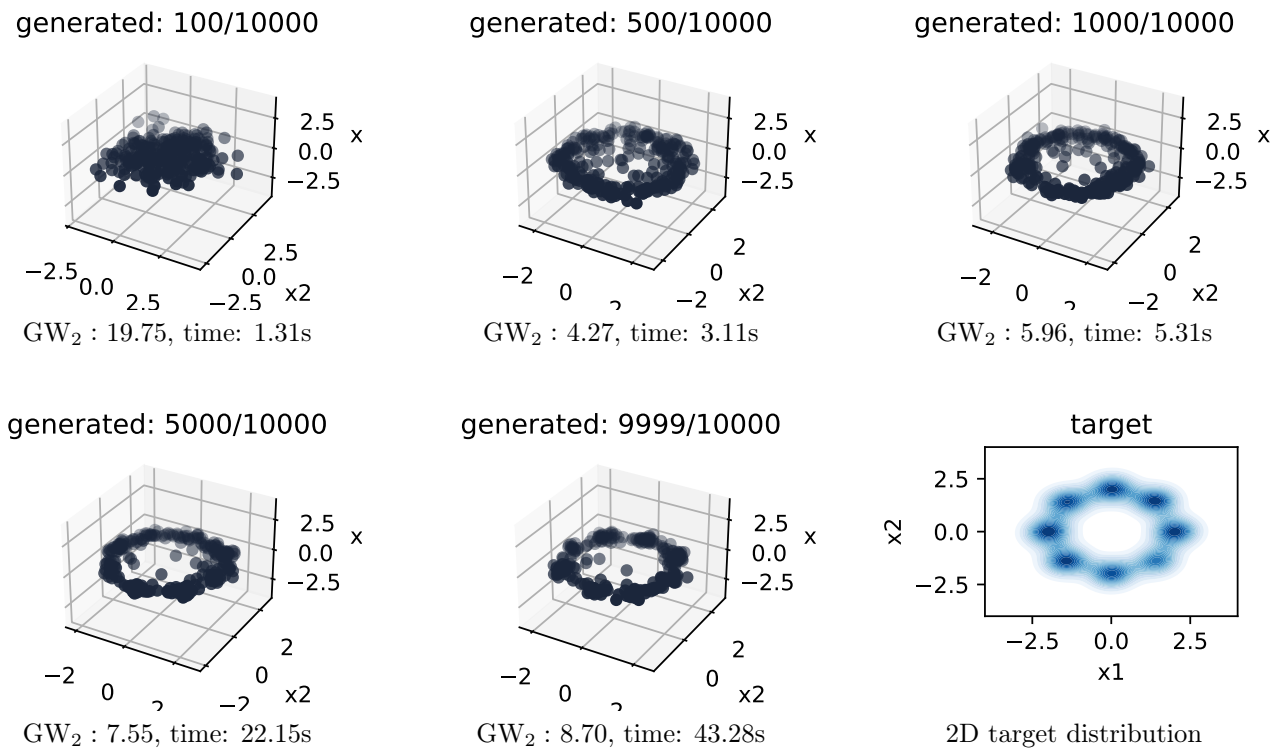


Figure 8: Generation of 2D distribution from 3D distribution for 4-point experiment

E.2.5 The 8-point Experiment

The Gaussian 8 experiment generates a dataset with 8 clusters arranged symmetrically around the origin. The centers of the clusters are at the points: (1, 0), (-1, 0), (0, 1), (0, -1), and four diagonal points: $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}), (\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), (\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$. Each point is sampled from a Gaussian distribution with small noise around these centers, scaled by a factor of 2. This experiment generates 8 distinct clusters, and the points are randomly assigned to these clusters, with each cluster having its own corresponding label. Now, we aim to generate a 3D distribution for this target 2D distribution. It was performed for 10,000 iterations.



GW_2 : 19.75, time: 1.31s | GW_2 : 4.27, time: 3.11s | GW_2 : 5.96, time: 5.31s

GW_2 : 7.55, time: 22.15s | GW_2 : 8.70, time: 43.28s | 2D target distribution

Figure 9: Generation of 3D distribution from 2D distribution for the 8-point experiment using SGW

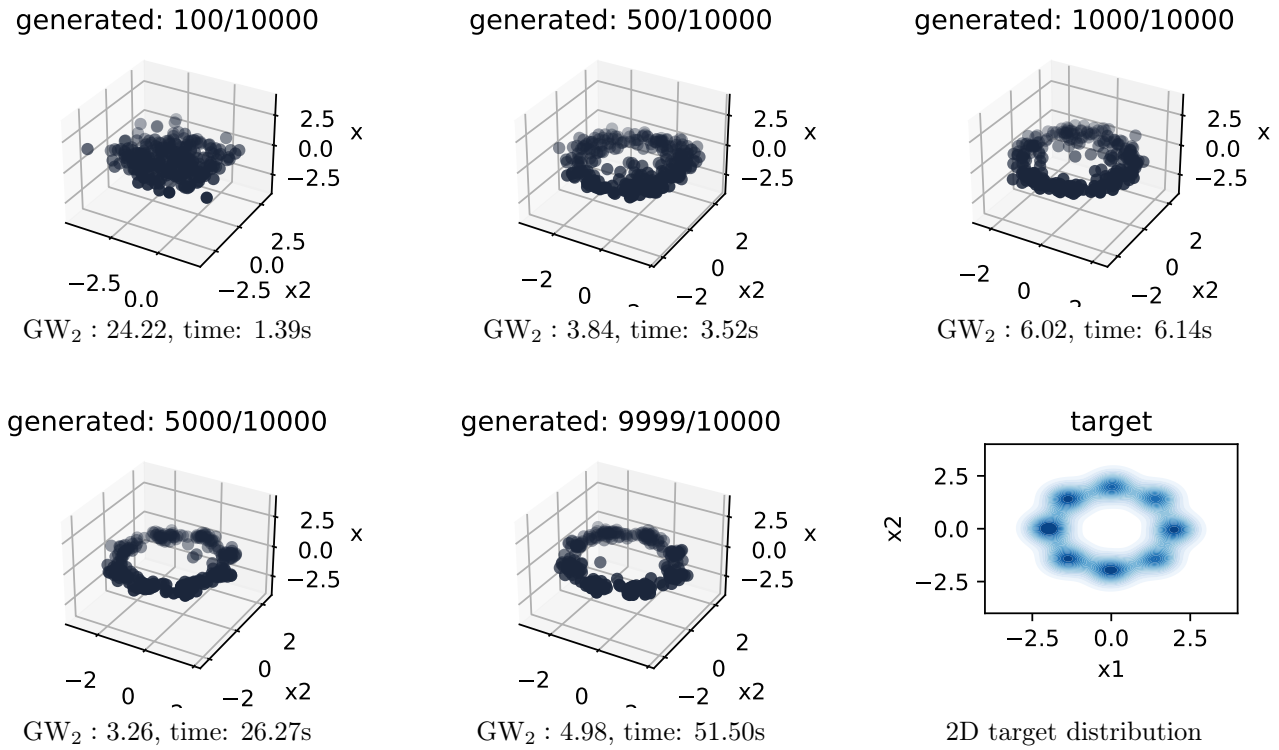


Figure 10: Generation of 3D distribution from 2D distribution for the 8-point experiment using RASGW

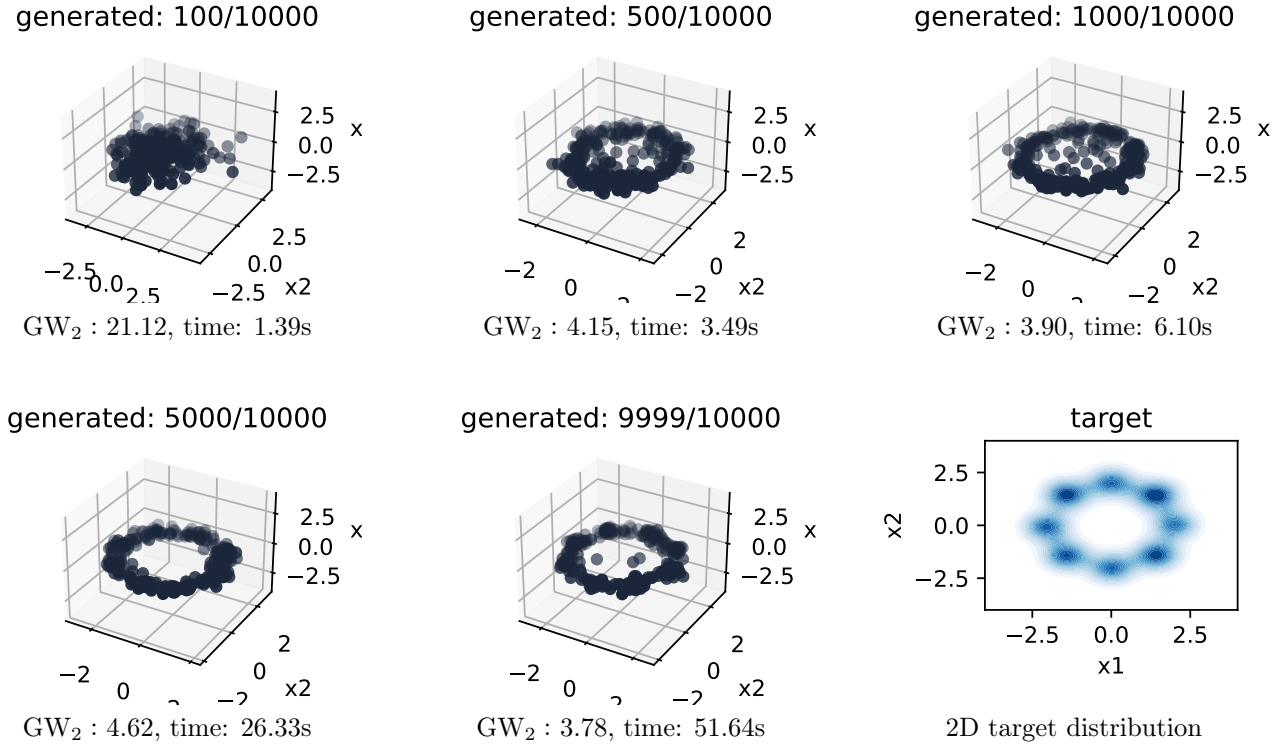


Figure 11: Generation of 3D distribution from 2D distribution for the 8-point experiment using IWRASGW

E.3 GWAE

E.3.1 Architecture

The encoder and decoder networks in the GWAE models, as proposed by Nakagawa et al. [2023], follow a convolutional and fully connected architecture. For the encoder network, the input is a 64×64 RGB image with 3 channels. The encoder consists of several convolutional layers with increasing feature map sizes and decreasing spatial dimensions, followed by fully connected layers that output the latent variables μ and σ^2. The decoder network takes these latent variables as input, first expanding the latent space through fully connected layers and then using deconvolutional layers to progressively upsample the feature maps until the image is reconstructed. The encoder and decoder both use SiLU Hendrycks and Gimpel [2023] activations, except for the final layer of the decoder, which uses a sigmoid activation function to ensure that the output is in the range $[0, 1]$.

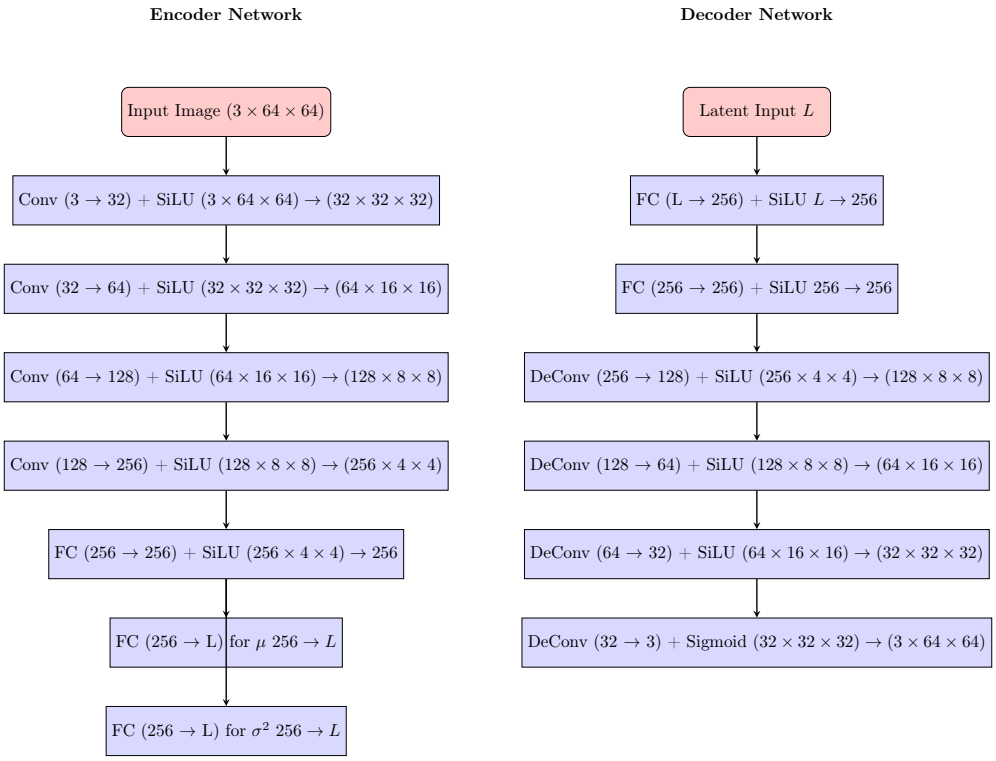


Figure 12: The architecture used in the GWAE models has been designed for 64×64 RGB images. In the experiments, the input size is set to $(\text{Channels}, \text{Height}, \text{Width}) = (3, 64, 64)$. The terms "FC" and "Conv" refer to fully-connected (linear) layers and convolutional layers, respectively.

E.3.2 Datasets

Omniglot. The Omniglot [Lake et al., 2015] dataset contains 1,623 handwritten characters from 50 different alphabets, with a total of 80,000 binary-valued images. Each image is a 105×105 pixel representation of a character, with 20 samples per character. The dataset is divided into a training set of 1,100 characters and a test set of 523 characters, making it a challenging benchmark for few-shot learning models.
CIFAR-10. The CIFAR [Krizhevsky et al., 2009] dataset contains 60,000 32x32 color images across 10 classes in the CIFAR-10 version and 100 classes in the CIFAR-100 version. In CIFAR-10, the images are labeled into categories such as airplane, automobile, bird, cat, dog, and other objects, with each class containing 6,000 images. The dataset is split into 50,000 training images and 10,000 test images. In CIFAR-100, the images are divided into 100 classes, with 600 images per class, grouped into 20 superclasses. It has the same training-test split as CIFAR-10, with 50,000 training and 10,000 test images. The images are 3-channel color images, with each image measuring 32x32 pixels.

E.3.3 Evaluation metrics

FID score. Fréchet Inception Distance (FID) [Heusel et al., 2017] is commonly used as a metric to evaluate the quality of images generated by generative models. The FID score is defined as the squared 2-Wasserstein distance between the feature distributions of real and generated images. It is assumed that the features of both real and generated images follow multivariate Gaussian distributions, with the real images having mean μ_r and covariance matrix Σ_r, and the generated images having mean μ_g and covariance matrix Σ_g.

The FID score is expressed as:

$$\text{FID} = W_2^2\left(N(\mu_r, \Sigma_r), N(\mu_g, \Sigma_g)\right) = \|\mu_r - \mu_g\|_2^2 + \text{tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}\right),$$

where W_2^2 denotes the squared Wasserstein-2 distance between the Gaussian distributions of real and generated features. The FID score quantifies the discrepancy between these distributions, and lower values are associated with better generation performance, indicating that the generated images are closer to the real images in terms of feature distribution.

As stated in Heusel et al. [2017], we follow Nakagawa et al. [2023] in using the features for the computation of FID, which are extracted from the final pooling layer of the Inception-v3 model that has been pre-trained on the ImageNet dataset [Deng et al., 2009].

Peak Signal-to-Noise Ratio. The Peak Signal-to-Noise Ratio (PSNR) is used for evaluating the quality of image reconstruction. The PSNR is defined as

$$\text{PSNR} = 20\log_{10}(\text{MAX}) - 10\log_{10}(\text{MSE}),$$

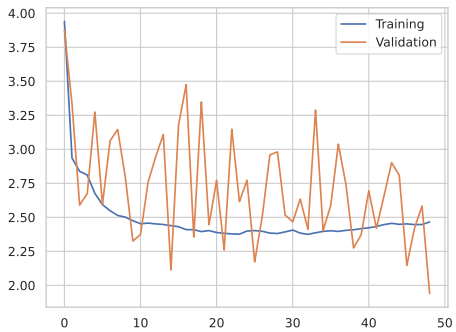
where MAX denotes the maximum possible pixel value, and MSE represents the Mean Squared Error (MSE), which is calculated as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_i - \hat{I}_i)^2,$$

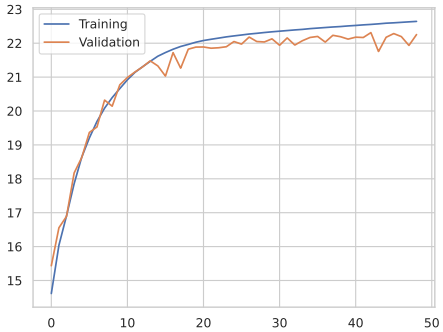
where I_i and $\hat{I}_i$ represent the pixel values of the original and reconstructed images, respectively, and N is the total number of pixels. In all the experiments, the value of MAX is set to 1, as the images in the dataset are scaled within the range [0, 1].

E.4 Results for CIFAR-10

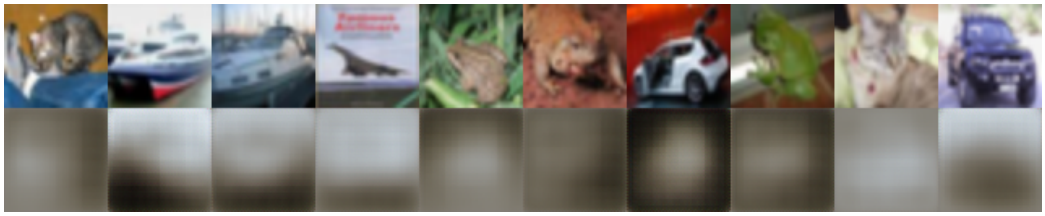
E.4.1 SGW



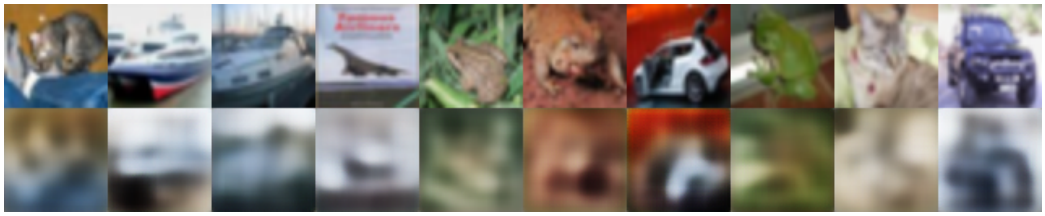
a. SGW Loss



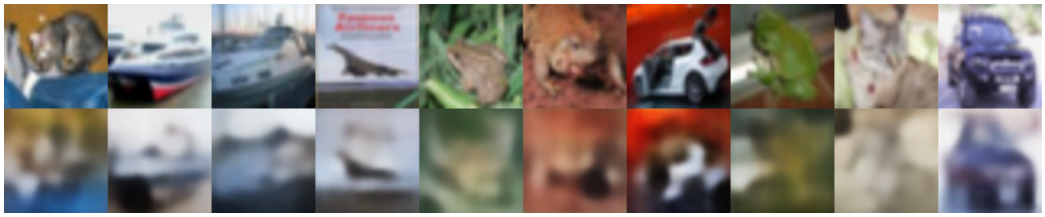
b. PSNR (in dB)



c. Epoch 1



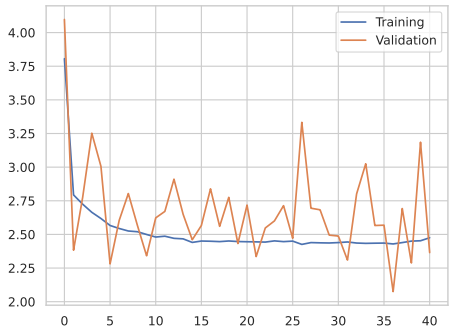
d. Epoch 10



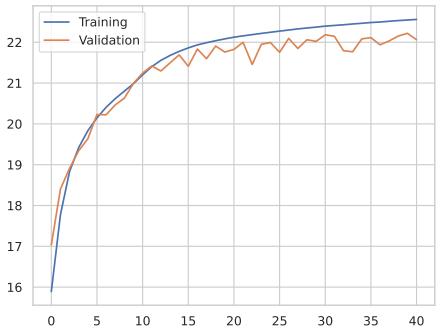
e. Last(48-th) iteration

Figure 13: a. Variation of SGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**72.87, **PSNR :**22.65, **Total time (in sec):** 22459

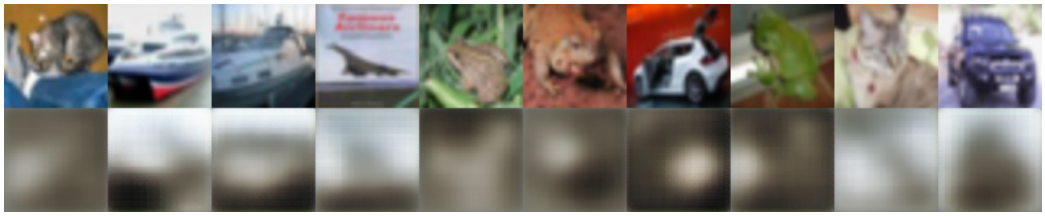
E.4.2 RASGW



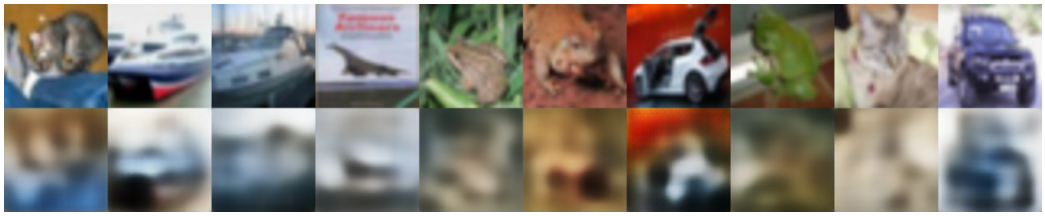
a. RASGW Loss



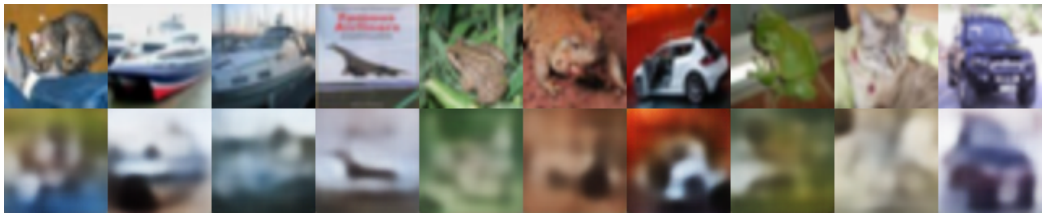
b. PSNR (in dB)



c. Epoch 1



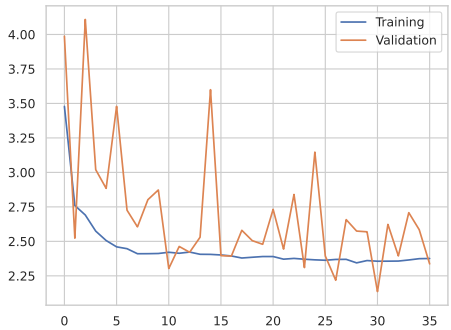
d. Epoch 10



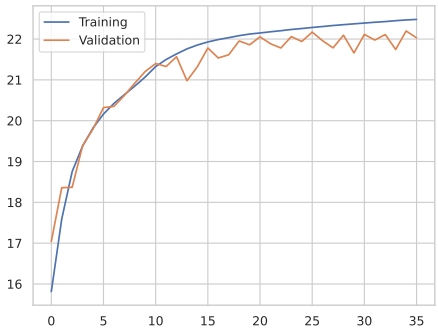
e. Last(40-th) iteration

Figure 14: a. Variation of RASGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**70.46, **PSNR :**22.70, **Total time (in sec):** 23613

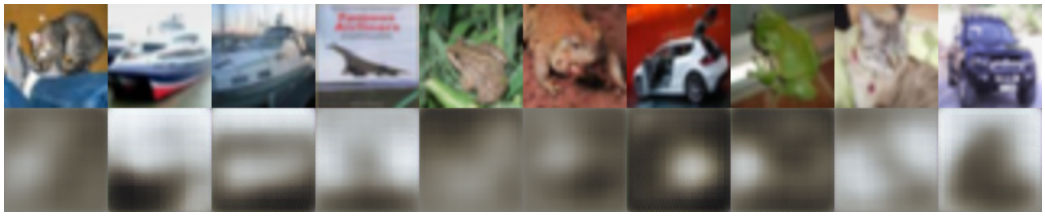
E.4.3 IWRASGW



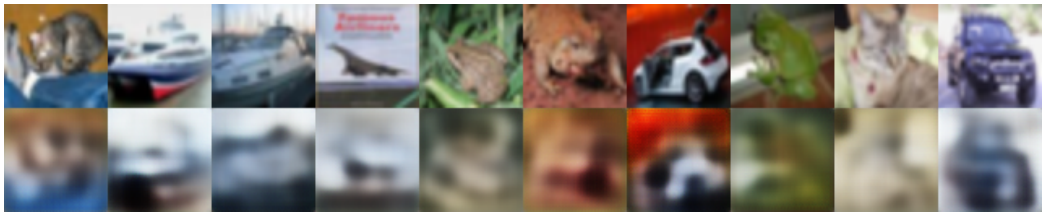
a. IWRASGW Loss



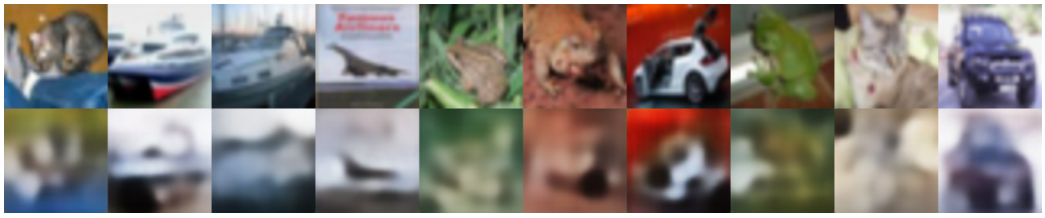
b. PSNR (in dB)



c. Epoch 1



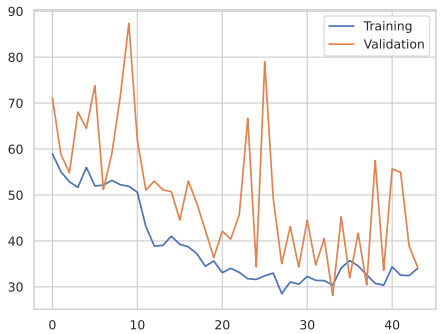
d. Epoch 10



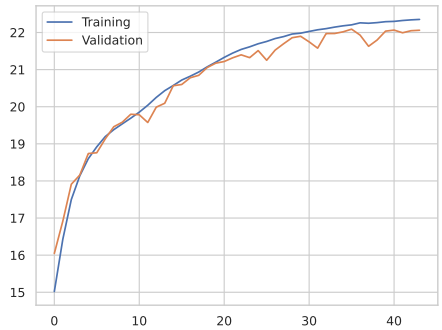
e. Last(35-th) iteration

Figure 15: a. Variation of RASGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**71.99, **PSNR :**22.73, **Total time (in sec):** 19879

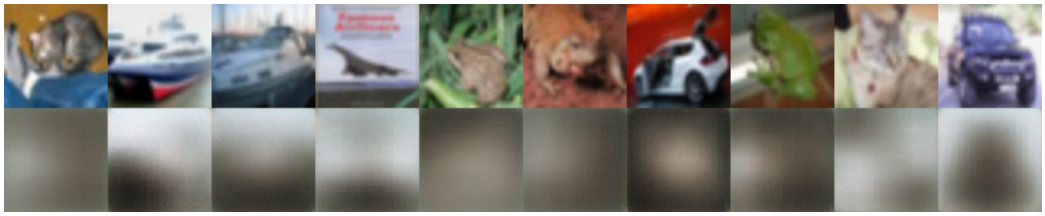
E.4.4 RPSGW



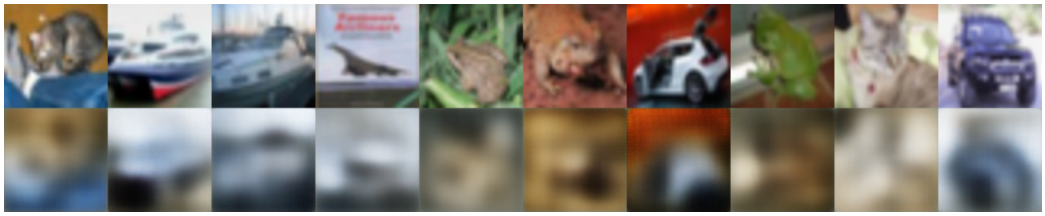
a. RPSGW Loss



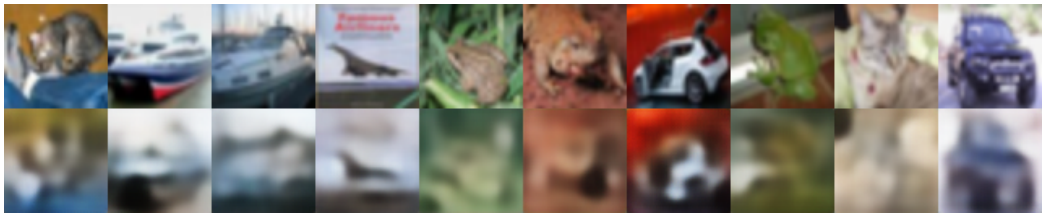
b. PSNR (in dB)



c. Epoch 1



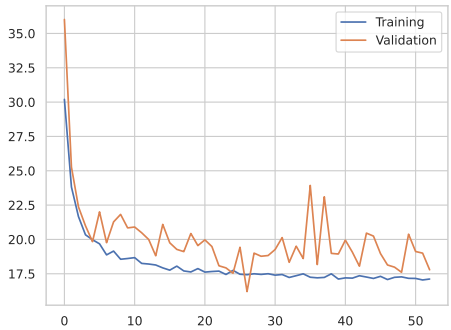
d. Epoch 10



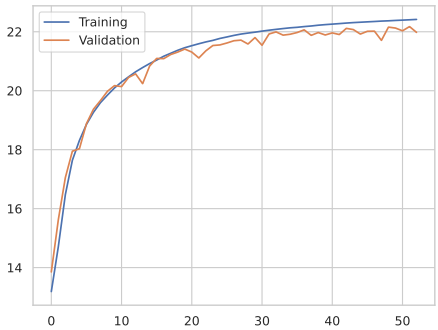
e. Last(43-th) iteration

Figure 16: a. Variation of RPSGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**80.80, **PSNR :**22.55, **Total time (in sec):** 24356

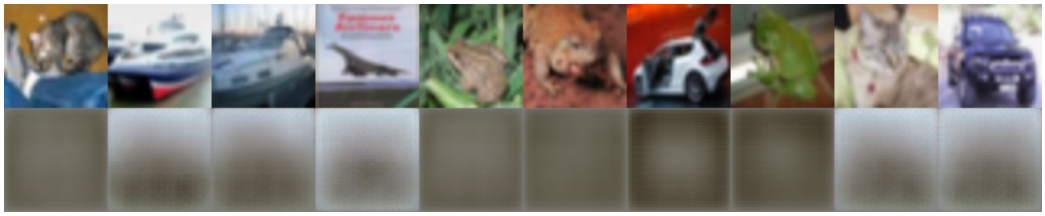
E.4.5 EBSGW



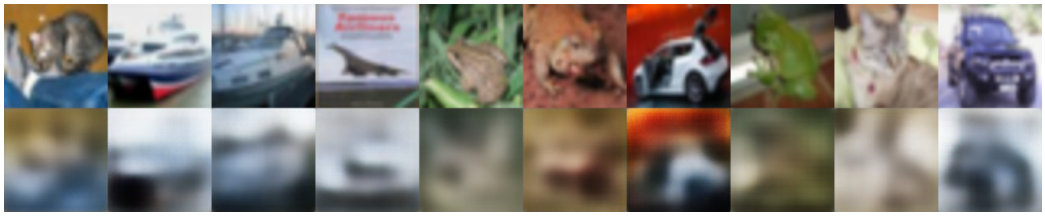
a. EBSGW Loss



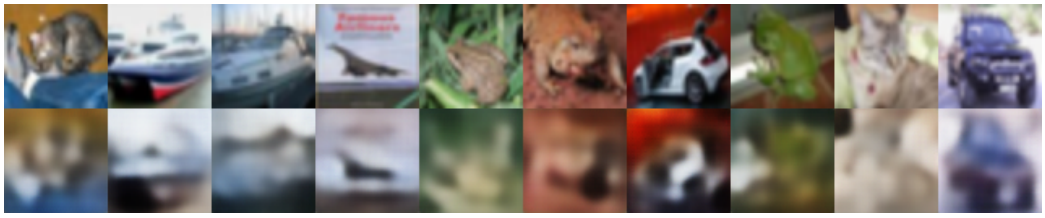
b. PSNR (in dB)



c. Epoch 1



d. Epoch 10

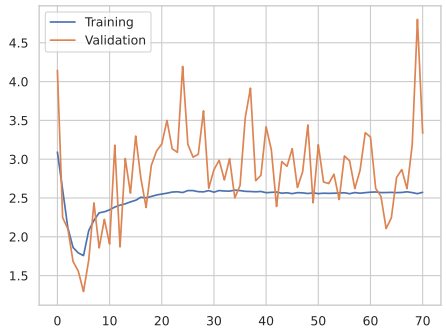


e. Last(52-th) iteration

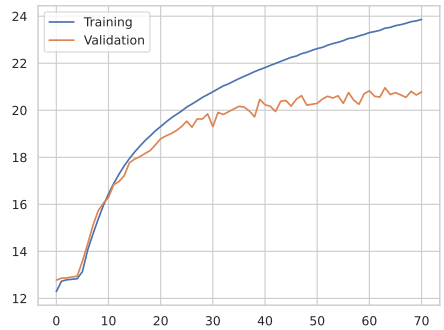
Figure 17: a. Variation of EBSGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**78.50, **PSNR :**22.48, **Total time (in sec):** 28765

E.5 Results on Omniglot

E.5.1 SGW

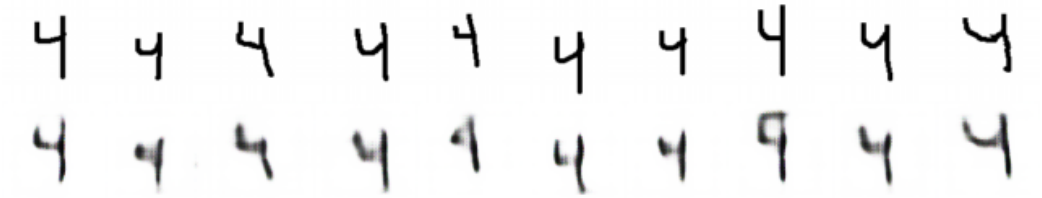


a. SGW Loss

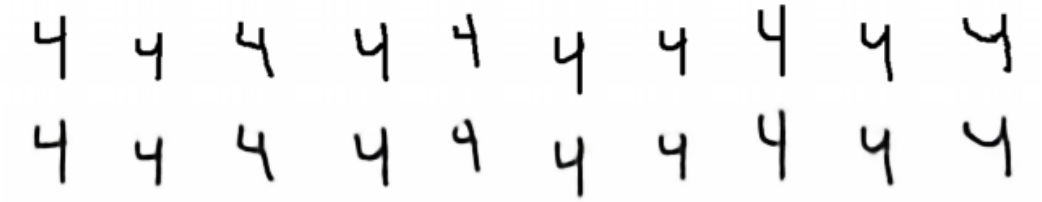


b. PSNR (in dB)

c. Epoch 1



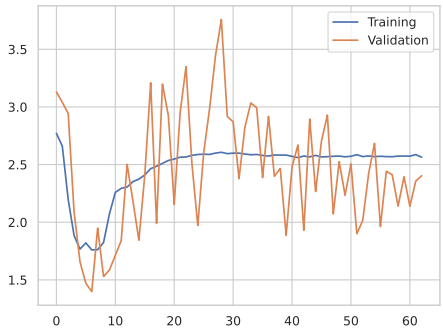
d. Epoch 10



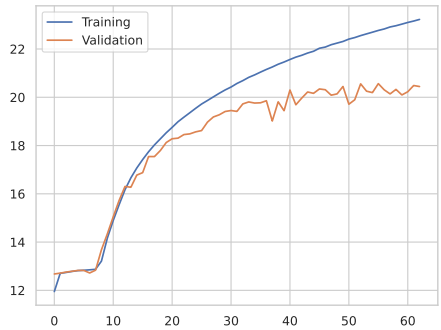
e. Last(70-th) iteration

Figure 18: a. Variation of SGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**20.95, **PSNR :**22.13, **Total time (in sec):** 16932

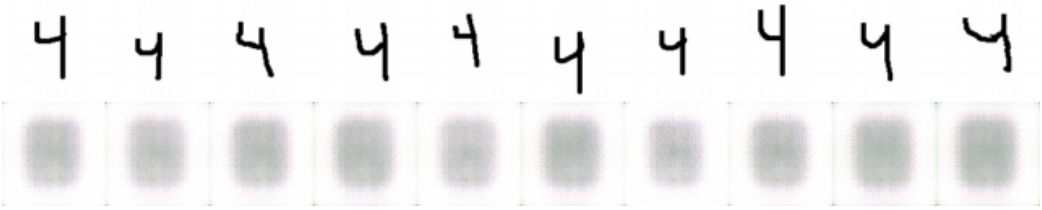
E.5.2 RASGW



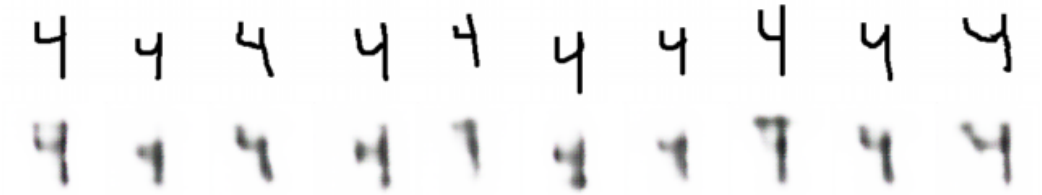
a. RASGW Loss



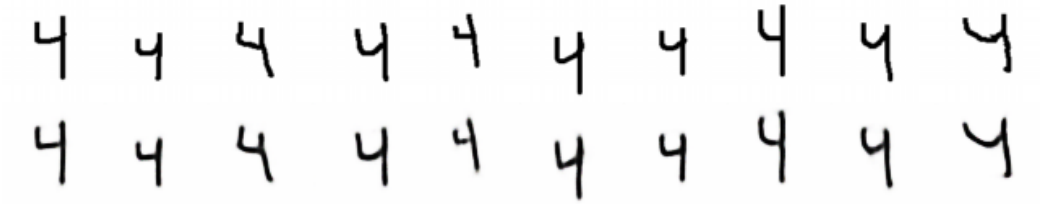
b. PSNR (in dB)



c. Epoch 1



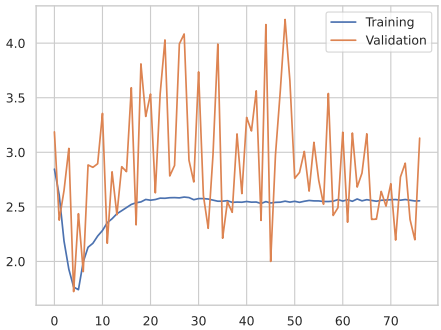
d. Epoch 10



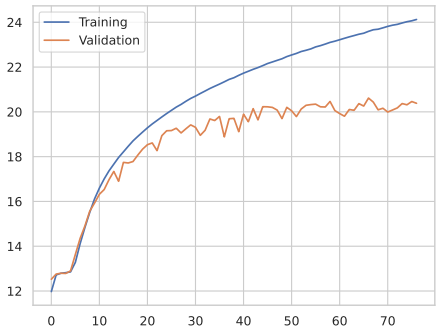
e. Last(62-nd) iteration

Figure 19: a. Variation of RASGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**21.48, **PSNR :**20.56, **Total time (in sec):** 16731

E.5.3 IWRASGW

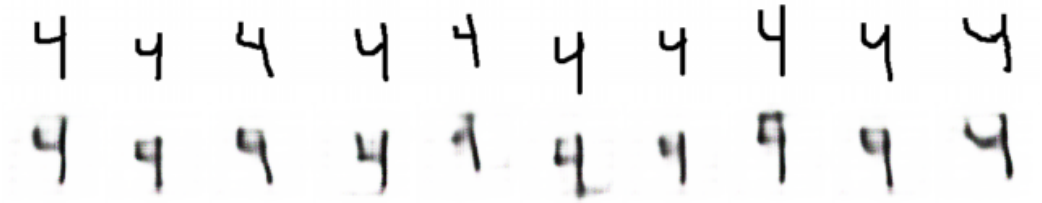


a. RASGW Loss

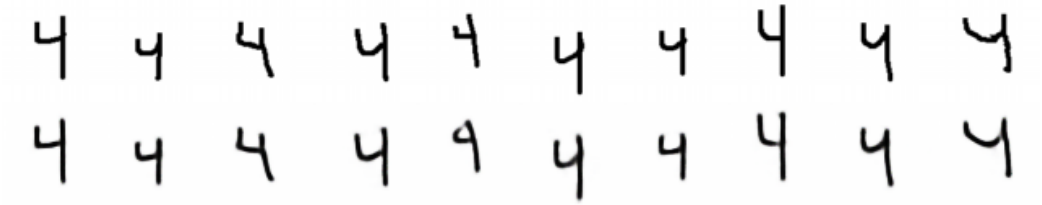


b. PSNR (in dB)

c. Epoch 1



d. Epoch 10



e. Last(76-th) iteration

Figure 20: a. Variation of IWRASGW Loss with epochs, b. Variation of PSNR with epochs, c. Reconstruction in Epoch 1, d. Reconstruction in Epoch 10, e. Reconstruction in Last Epoch, **FID :**20.25, **PSNR :**21.87, **Total time (in sec):** 18972

Remark E.1. To demonstrate the fact that instilling relation-awareness in slicing directions indeed improves performance over uniformly sampled counterparts (SGW), we highlight key takeaways from the above experiments. As shown in the results for the CIFAR-10 experiment (Appendix E.4), the standard SGW approach required 48 epochs to converge, whereas IWRASGW converged in only 35 epochs. Remarkably, this 27% reduction in training iterations is accompanied by an improvement in FID score from 72.87 (SGW) to 71.99 (IWRASGW). The convergence efficiency also directly translates into superior wall-clock performance, addressing concerns about whether the wasted Monte Carlo samples in SGW affect practical resource usage. Despite computing the Relation-Aware Projecting Directions (RAPD) in the process, the total training time for IWRASGW (19,879 seconds) was notably lower than that of SGW (22,459 seconds). The results are rather comparable in Omniglot, which employs a GWAE backbone. Nonetheless, the more noticeable empirical evidence of superiority in performance comes in the 4-point and 8-point (Figures 9 - 11) experiments. For example, while SGW achieves a GW benchmark

of 104.91 in 22.34 seconds, RASGW reaches 30.08 in approximately the same time (25.30 seconds) after 5000 iterations (see Figure 8). This goes on to infer that the directions that enable achieving a much better alignment following slicing are more efficiently sampled via RASD compared to uniform sampling (SGW).

F COMPUTATIONAL INFRASTRUCTURE

All the experiments of GWGAN (Gromov-Wasserstein Generative Adversarial Network) were performed on a single NVIDIA RTX 3090 GPU (24 GB), and all the experiments of GWAE (Gromov-Wasserstein Auto-Encoder) were performed on a single NVIDIA GeForce GTX Titan Xp (12 GB).