T-GROW: ROBUST AND SCALABLE TREE-BASED GROMOV-WASSERSTEIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparing structured and multi-modal data across different representation spaces remains a core challenge in machine learning, as metric distances between individual samples are not directly defined across domains. The Gromov–Wasserstein (GW) distance addresses this by seeking a geometry-preserving alignment between probability measures supported on different metric spaces. However, its underlying non-convex quadratic formulation becomes computationally prohibitive at scale. To alleviate this burden, scalable tree-based methods replace full pairwise geometry with simpler summaries of hierarchical structure. While relying solely on root-to-node path lengths enables fast alignment, this structural reduction discards rich pairwise relational geometry within the tree hierarchy. In this work, we introduce T-GROW, a robust and scalable tree-based GW discrepancy built on co-depth, defined as the length of the root path shared by two samples before their paths diverge. For any fixed hierarchy, we prove that solving GW over its leaf clusters under co-depth is exactly equivalent to solving GW over the original samples under the induced tree metric. This theoretical equivalence guarantees that T-GROW retains the complete pairwise geometry of the hierarchy without approximation loss, while shifting the quadratic optimization to a significantly compressed leaf resolution. Experimental evaluations show that T-GROW consistently improves structural matching over scalable GW baselines, exhibits strong robustness to structural noise, and scales efficiently to tens of millions of samples.

1 INTRODUCTION

Optimal transport and Wasserstein distances provide a powerful framework for comparing probability measures supported on a shared metric space. They have found broad applications across machine learning (Cuturi & Doucet, 2014; Arjovsky et al., 2017; Korotin et al., 2021; Haviv et al., 2025), computer vision (Rubner et al., 2000; Fiorucci et al., 2023; Liu et al., 2025; Yu et al., 2025), and natural language processing (Kusner et al., 2015; Xu et al., 2018; Lample et al., 2018). Within optimal transport, tree metrics offer a structured surrogate geometry on which transport can be computed particularly efficiently. This motivates Tree-Sliced Wasserstein, which averages transport over randomized tree representations (Le et al., 2019), and has inspired subsequent work on learned, robust, and richer tree-based transport constructions (Takezawa et al., 2021; Backurs et al., 2020; Le et al., 2024; Tran et al., 2025b;a; Düsterwald et al., 2025).

However, Wasserstein distance, including tree-based variants, still requires a ground cost that compares points across two supports living in the same metric space. This requirement becomes ill-defined when comparing structured and multi-modal data living in distinct spaces without canonical pointwise correspondence. The Gromov–Wasserstein (GW) distance (Mémoli, 2011; Peyré et al., 2016) overcomes this limitation by evaluating pairwise relations within each space rather than distances across the two spaces. Consequently, GW has become a central tool across machine learning and computer vision, with applications in areas such as graph and representation matching, geometric correspondence, and multi-modal alignment (Xu et al., 2019; Alvarez-Melis & Jaakkola, 2018; Solomon et al., 2016; Demetci et al., 2022; Schnaus et al., 2025). Despite its flexibility, the non-convex quadratic formulation of GW becomes computationally prohibitive for large-scale datasets. A broad literature addresses this scalability bottleneck through entropic regularization (Rioux et al., 2024), projection methods (Vayer et al., 2019), low-rank factorizations (Scetbon et al., 2022), and quantization (Chowdhury et al., 2021).

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

Leveraging tree structures, FlowAlign (Le et al., 2021) scales heterogeneous alignment by substituting full pairwise tree geometry with root-to-sample flow embeddings, reducing alignment to a significantly simpler transport problem. Nevertheless, root-flow reduction discards part of the pairwise hierarchical information within the tree hierarchy. While tracking how individual samples are positioned relative to the root, it fails to capture shared ancestry between sample pairs. Consequently, structural hierarchies with distinct pairwise tree geometries can yield identical root-flow representations.

To address this shortcoming, we introduce a robust and scalable tree-based GW discrepancy built on co-depth, defined as the length of the root path shared by two samples before their paths diverge. We prove that for any fixed hierarchy, solving GW over its leaf clusters under co-depth is mathematically equivalent to solving GW over the original samples under the induced tree metric. This sample-to-leaf reduction introduces no optimization gap, shifting the quadratic optimization from the original sample scale to a compressed leaf resolution without sacrificing relational fidelity.

Contributions. In this paper, we propose T-GROW, a tree-based Gromov–Wasserstein (GW) discrepancy that retains pairwise hierarchical structure while shifting second-order alignment to a resolution controlled by the hierarchy rather than sample size. Specifically, our contributions are as follows: i) we introduce co-depth to capture hierarchical relationships beyond standard root-flow alignment, where its diagonal recovers root-depth information while the full relation defines the underlying tree geometry; ii) we prove that the leaf-level GW problem under a fixed hierarchy is strictly equivalent to the sample-level problem with zero optimization gap, and by averaging over randomized hierarchies, we establish metric properties, approximation and computational guarantees, and a strict separation from root-flow alignment; and iii) we conduct extensive experiments demonstrating strong alignment and downstream performance with competitive scalability and exceptional robustness under both geometric perturbations and label corruption.

2 Tree-based Gromov-Wasserstein

2.1 Preliminaries and Motivation

We consider the comparison of probability measures supported on possibly different metric spaces. Let $\mathcal{X} = (X, d_{\mathcal{X}}, \mu)$ and $\mathcal{Y} = (Y, d_{\mathcal{Y}}, \nu)$ be two metric-measure spaces (mm-spaces), with discrete measures $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, where $a \in \Delta_n$ and $b \in \Delta_m$. Their admissible couplings are $\Pi(a, b) = \left\{\pi \in \mathbb{R}_+^{n \times m} : \pi \mathbf{1}_m = a,\ \pi^\top \mathbf{1}_n = b\right\}$. The Gromov-Wasserstein (GW) discrepancy compares the two spaces through their intra-domain pairwise geometries. For quadratic loss,

$$\mathrm{GW}_2^2(\mathcal{X}, \mathcal{Y}) = \min_{\pi \in \Pi(a,b)} \sum_{i,i',j,j'} \left| d_{\mathcal{X}}(x_i, x_{i'}) - d_{\mathcal{Y}}(y_j, y_{j'}) \right|^2 \pi_{ij} \pi_{i'j'}. \tag{1}$$

This formulation allows GW to compare measures across different ambient spaces and, through optimal coupling, to recover a soft correspondence between their supports. However, its second-order dependence on the coupling leads to a non-convex quadratic optimization and becomes computationally demanding as the support sizes increase (Mémoli, 2011; Peyré et al., 2016).

The second-order structure of Eq. 1 is central to the relational flexibility of GW, but it is also the main source of its computational cost. This has motivated scalable formulations that replace the original pairwise geometry with more structured representations. Tree-based approaches offer a particularly natural route, as they organize the support hierarchically and enable structural comparison through quantities derived from the tree. FlowAlign (Le et al., 2021) reduces tree geometry to root-to-point flows. Let $T_{\mathcal{X}}$ and $T_{\mathcal{Y}}$ be rooted trees associated with $\mathcal{X}$ and $\mathcal{Y}$, with roots $r_{\mathcal{X}}$ and $r_{\mathcal{Y}}$, respectively. For fixed roots, let $q_{\mathcal{X}}(x_i)$ and $q_{\mathcal{Y}}(y_j)$ denote the tree representatives associated with x_i and y_j, and define $\ell_i^{\mathcal{X}} = d_{T_{\mathcal{X}}}(r_{\mathcal{X}}, q_{\mathcal{X}}(x_i))$ and $\ell_j^{\mathcal{Y}} = d_{T_{\mathcal{Y}}}(r_{\mathcal{Y}}, q_{\mathcal{Y}}(y_j))$. The aligned-root FlowAlign objective is

$$\mathrm{FA}_2^2(\mu, \nu; r_{\mathcal{X}}, r_{\mathcal{Y}}) = \min_{\pi \in \Pi(a,b)} \sum_{i,j} \left| \ell_i^{\mathcal{X}} - \ell_j^{\mathcal{Y}} \right|^2 \pi_{ij}. \tag{2}$$

It is a one-dimensional Wasserstein problem between the weighted root-flow distributions and can be evaluated efficiently from the ordered flow lengths. The general FlowAlign formulation additionally optimizes over the roots, while DepthAlign extends the same principle across multiple tree depths.

Eq. 2 makes clear what is removed by this simplification. For two support points $x_i, x_{i'} \in \mathcal{X}$, let $q_{\mathcal{X}}(x_i)$ and $q_{\mathcal{X}}(x_{i'})$ denote their corresponding leaves in $T_{\mathcal{X}}$. Their induced tree distance can then be written as

$$d_{T_{\mathcal{X}}}(q_{\mathcal{X}}(x_i), q_{\mathcal{X}}(x_{i'})) = \ell_i^{\mathcal{X}} + \ell_{i'}^{\mathcal{X}} - 2d_{T_{\mathcal{X}}}(r_{\mathcal{X}}, \mathrm{LCA}(q_{\mathcal{X}}(x_i), q_{\mathcal{X}}(x_{i'})))\,, \tag{3}$$

and analogously for $T_{\mathcal{Y}}$. $\mathrm{LCA}(q_{\mathcal{X}}(x_i), q_{\mathcal{X}}(x_{i'}))$ denotes the lowest common ancestor of the corresponding tree representatives. The first two terms are determined by the individual root-flow lengths used in Eq. 2, whereas the last term depends jointly on the pair and measures the portion of their root paths that is shared before they separate.

This distinction exposes the structural limitation of the root-flow reduction. FlowAlign depends only on the weighted collections of root-flow lengths $\{(\ell_i^{\mathcal{X}}, a_i)\}_{i=1}^n$ and $\{(\ell_j^{\mathcal{Y}}, b_j)\}_{j=1}^m$. Therefore, rearranging points among subtrees without changing their root-flow lengths leaves FlowAlign discrepancy unchanged. Such a rearrangement can nevertheless change the lowest common ancestors of point pairs and hence alter the pairwise tree distances in Eq. 3.

This reveals the central trade-off. Collapsing a tree to unary root-depth statistics yields an efficient one-dimensional alignment, but removes the pair-dependent component that records shared hierarchical structure. The key issue is therefore whether the missing relational information can be retained without restoring the full point-level second-order comparison. For tree metrics, this possibility arises from the fact that pairwise relations are not independent but are coupled through the hierarchy of shared root paths. Section 2.2 develops this shared-path quantity as co-depth, which forms the relational basis of the subsequent construction.

2.2 Co-Depth Relational Geometry

The shared root paths provide a natural representation of the pairwise structure of a rooted tree. Let $T = (V, E, w, r)$ be a rooted weighted tree with edge-weight function $w : E \to \mathbb{R}_{\geq 0}$, where $w_e := w(e)$ denotes the weight of edge $e \in E$. For two nodes $u, v \in V$, their common root path terminates at their lowest common ancestor $\mathrm{LCA}(u, v)$. We encode the length of this shared path by the following quantity.

Definition 1 (Co-depth). *For any $u, v \in V$, their co-depth is defined as*

$$c_T(u, v) := d_T(r, \mathrm{LCA}(u, v))\,. \tag{4}$$

Co-depth has a standard interpretation in metric geometry (Gromov, 1987). For the tree metric d_T with base point r, $c_T(u, v) = (u \mid v)_r := \frac{1}{2}\left(d_T(r, u) + d_T(r, v) - d_T(u, v)\right)$, i.e., co-depth is precisely the Gromov product with respect to the root. For trees, this quantity equals exactly the length of the root path shared by u and v. Equivalently, the tree metric can be recovered from co-depth through

$$d_T(u, v) = c_T(u, u) + c_T(v, v) - 2c_T(u, v), \tag{5}$$

since $c_T(u, u) = d_T(r, u)$. Thus, the diagonal of co-depth stores unary root-depth information, whereas the off-diagonal entries encode shared ancestry.

The same geometry admits an explicit hierarchical representation. For each node $u \in V$, define the weighted root-path feature $\phi_T(u) \in \mathbb{R}^{|E|}$ by

$$[\phi_T(u)]_e := \sqrt{w_e}\,\mathbf{1}\{e \in P(r, u)\}\,, \qquad e \in E, \tag{6}$$

where $P(r, u)$ denotes the unique path from root r to u. For a set of nodes $\{u_1, \ldots, u_L\} \subseteq V$, define the co-depth matrix $K_T \in \mathbb{R}^{L \times L}$ by $[K_T]_{pq} := c_T(u_p, u_q)$. Let $A_T \in \{0, 1\}^{L \times |E|}$ be the root-path incidence matrix with $[A_T]_{pe} := \mathbf{1}\{e \in P(r, u_p)\}$, and let $W_T := \mathrm{diag}(\{w_e\}_{e \in E})$.

Proposition 1 (Hierarchical representation of co-depth). *For any $u, v \in V$,*

$$c_T(u, v) = \langle \phi_T(u), \phi_T(v) \rangle. \tag{7}$$

Consequently, $K_T = A_T W_T A_T^\top \succeq 0$.

Proposition 1 gives an explicit Gram representation of the rooted Gromov-product geometry induced by the tree. Together with Eq. 5, the co-depth kernel K_T and the induced tree metric contain the same leaf-level geometric information used in the reduced GW construction below. Further connections to classical metric-kernel theory are given in Appendix A.1.1.

2.3 TREE-INDUCED GROMOV–WASSERSTEIN

Having represented pairwise tree geometry through co-depth, we now compress each measure to the leaf resolution of its hierarchy. Let $T_\mathcal{X}$ and $T_\mathcal{Y}$ be fixed rooted weighted trees with nonempty leaf sets $\mathcal{L}_\mathcal{X} = \{u_1, \dots, u_{L_\mathcal{X}}\}$ and $\mathcal{L}_\mathcal{Y} = \{v_1, \dots, v_{L_\mathcal{Y}}\}$, where $L_\mathcal{X}$ and $L_\mathcal{Y}$ denote the corresponding numbers of leaves. Each support point is assigned to one leaf by maps $q_\mathcal{X} : X \to \mathcal{L}_\mathcal{X}$ and $q_\mathcal{Y} : Y \to \mathcal{L}_\mathcal{Y}$, inducing leaf clusters $C_p^\mathcal{X} := q_\mathcal{X}^{-1}(u_p)$ and $C_q^\mathcal{Y} := q_\mathcal{Y}^{-1}(v_q)$. Equivalently, the leaf representation identifies support points assigned to the same leaf, with $x_i \sim_{T_\mathcal{X}} x_{i'}$ if $q_\mathcal{X}(x_i) = q_\mathcal{X}(x_{i'})$, and analogously for $\mathcal{Y}$.

Each leaf inherits the total probability mass of the samples assigned to its cluster. The corresponding push-forward measures are

$$\bar{\mu} := (q_\mathcal{X})_\# \mu = \sum_{p=1}^{L_\mathcal{X}} \alpha_p \delta_{u_p}, \qquad \bar{\nu} := (q_\mathcal{Y})_\# \nu = \sum_{q=1}^{L_\mathcal{Y}} \beta_q \delta_{v_q}, \tag{8}$$

where $\alpha_p = \mu(C_p^\mathcal{X}) = \sum_{x_i \in C_p^\mathcal{X}} a_i$ and $\beta_q = \nu(C_q^\mathcal{Y}) = \sum_{y_j \in C_q^\mathcal{Y}} b_j$. Hence $\alpha \in \Delta_{L_\mathcal{X}}$ and $\beta \in \Delta_{L_\mathcal{Y}}$. The relational geometry on the leaf spaces is inherited from the two trees. Define the leaf co-depth matrices by $[K_\mathcal{X}]_{pp'} := c_{T_\mathcal{X}}(u_p, u_{p'})$ and $[K_\mathcal{Y}]_{qq'} := c_{T_\mathcal{Y}}(v_q, v_{q'})$, and let $h_\mathcal{X} := \mathrm{diag}(K_\mathcal{X}) \in \mathbb{R}^{L_\mathcal{X} \times 1}$ and $h_\mathcal{Y} := \mathrm{diag}(K_\mathcal{Y}) \in \mathbb{R}^{L_\mathcal{Y} \times 1}$ denote the corresponding column vectors of diagonal entries. By Eq. 5, the corresponding leaf-level distance matrices are

$$D_\mathcal{X} = h_\mathcal{X} \mathbf{1}_{L_\mathcal{X}}^\top + \mathbf{1}_{L_\mathcal{X}} h_\mathcal{X}^\top - 2K_\mathcal{X}, D_\mathcal{Y} = h_\mathcal{Y} \mathbf{1}_{L_\mathcal{Y}}^\top + \mathbf{1}_{L_\mathcal{Y}} h_\mathcal{Y}^\top - 2K_\mathcal{Y}. \tag{9}$$

Thus $[D_\mathcal{X}]_{pp'} = d_{T_\mathcal{X}}(u_p, u_{p'})$ and analogously for $\mathcal{Y}$. With positive path lengths between distinct leaves, these are valid metrics on the corresponding leaf spaces, allowing zero-length paths yields pseudometrics. A formal verification is given in Appendix A.1.3.

Definition 2 (Tree-induced Gromov–Wasserstein discrepancy)**.** *The leaf representations $\bar{\mathcal{X}}_T = (\mathcal{L}_\mathcal{X}, D_\mathcal{X}, \bar{\mu})$ and $\bar{\mathcal{Y}}_T = (\mathcal{L}_\mathcal{Y}, D_\mathcal{Y}, \bar{\nu})$ are finite metric-measure spaces when distinct leaves are separated by positive tree distance, and finite pseudometric-measure spaces otherwise. The tree-induced Gromov–Wasserstein discrepancy is defined as*

$$\mathrm{GW}_{T,2}^2(\mu, \nu) := \min_{P \in \Pi(\alpha, \beta)} \sum_{p,p'=1}^{L_\mathcal{X}} \sum_{q,q'=1}^{L_\mathcal{Y}} \left| [D_\mathcal{X}]_{pp'} - [D_\mathcal{Y}]_{qq'} \right|^2 P_{pq} P_{p'q'}. \tag{10}$$

$P \in \mathbb{R}_+^{L_\mathcal{X} \times L_\mathcal{Y}}$ transports mass between leaf clusters, so the second-order alignment is carried out at resolution $(L_\mathcal{X}, L_\mathcal{Y})$ rather than (n, m). The matrices $D_\mathcal{X}$ and $D_\mathcal{Y}$ are the canonical leaf-level geometries used by T-GROW, as they are exactly the tree metrics induced by the co-depth representations.

To determine whether passing from the original supports to the leaf spaces introduces an additional optimization approximation, associate each pair of original samples with the distance between their corresponding leaves

$$\widetilde{d}_\mathcal{X}(x_i, x_{i'}) := D_\mathcal{X}\big(q_\mathcal{X}(x_i), q_\mathcal{X}(x_{i'})\big), \widetilde{d}_\mathcal{Y}(y_j, y_{j'}) := D_\mathcal{Y}\big(q_\mathcal{Y}(y_j), q_\mathcal{Y}(y_{j'})\big). \tag{11}$$

These tree-induced distances are pseudometrics on the original supports. Samples assigned to the same leaf are indistinguishable at the chosen tree resolution.

Theorem 1 (Tree-quotient reduction)**.** *Let $q_X : X \to L_X$ and $q_Y : Y \to L_Y$ be finite partitions with induced masses α and β, and let D_X and D_Y be pairwise relations on the corresponding quotient supports. Define $\widetilde{d}_X(x_i, x_{i'}) = D_X(q_X(x_i), q_X(x_{i'}))$, and analogously for $\widetilde{d}_Y$. Then*

$$\begin{aligned} &\min_{P \in \Pi(\alpha, \beta)} \sum_{p,p',q,q'} \left| [D_X]_{pp'} - [D_Y]_{qq'} \right|^2 P_{pq} P_{p'q'} \\ &\quad = \min_{\pi \in \Pi(a,b)} \sum_{i,i',j,j'} \left| \widetilde{d}_X(x_i, x_{i'}) - \widetilde{d}_Y(y_j, y_{j'}) \right|^2 \pi_{ij} \pi_{i'j'}. \end{aligned} \tag{12}$$

The proof is provided in Appendix A.1.2. Theorem 1 shows that, once the quotient partitions and induced pairwise relations are fixed, replacing the $n \times m$ point-level coupling by a $L_\mathcal{X} \times L_\mathcal{Y}$ quotient-level coupling introduces no additional optimization gap, independently of the numerical solver. For

the tree-induced construction above, these relations are the co-depth-induced leaf metrics, hence, approximation relative to the GW problem arises from replacing $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ by their finite resolution induced counterparts, rather than from the quotient reduction itself. The metric properties of the resulting leaf mm-spaces are summarized in Appendix A.1.3.

2.4 T-GROW Construction and Analysis

A finite-depth hierarchy provides a data-adaptive representation of the underlying geometry at a prescribed resolution, but its realization depends on the particular hierarchical partition. To reduce this dependence, we evaluate the tree-induced GW discrepancy over multiple randomized hierarchies. Each support is recursively partitioned by farthest-point seeding followed by nearest-anchor assignment, with branching factor κ and maximum depth H. Each cluster is represented by an anchor, and parent–child edge lengths are the distances between the corresponding anchors. We use centroids for vector-valued inputs and medoids for general metric inputs. Further details are given in Appendix A.1.4.

For each slice s, let $T_{\mathcal{X}}^{(s)} = \mathcal{B}_{\kappa,H}(\mathcal{X};\xi_s), T_{\mathcal{Y}}^{(s)} = \mathcal{B}_{\kappa,H}(\mathcal{Y};\xi_s)$ denote the resulting randomized hierarchies. Their nonempty leaves determine masses $\alpha^{(s)}$ and $\beta^{(s)}$, while the co-depth construction of Section 2.2 induces the corresponding leaf-distance matrices $D_{\mathcal{X}}^{(s)}$ and $D_{\mathcal{Y}}^{(s)}$. The discrepancy associated with the s-th hierarchy is

$$d_s^2(\mathcal{X},\mathcal{Y}) = \min_{P\in\Pi(\alpha^{(s)},\beta^{(s)})} \sum_{p,p',q,q'} \left|[D_{\mathcal{X}}^{(s)}]_{pp'} - [D_{\mathcal{Y}}^{(s)}]_{qq'}\right|^2 P_{pq}P_{p'q'}. \tag{13}$$

Thus, each slice evaluates the canonical tree-induced GW discrepancy of Section 2.3 on a different randomized hierarchy. We define

$$\text{T-GROW}_{2,S}(\mathcal{X},\mathcal{Y}) := \left(\frac{1}{S}\sum_{s=1}^{S} d_s^2(\mathcal{X},\mathcal{Y})\right)^{1/2}. \tag{14}$$

When the same fixed family of randomized encoders is used across comparisons, each d_s is the GW distance between the corresponding encoded leaf mm-spaces, and their RMS aggregation defines a pseudometric on the encoded input domain. The formal statement is given in Appendix A.1.6.

Consider a fixed pair of leaf representations and suppress the slice index. Let $h_{\mathcal{X}}(p) = K_{\mathcal{X}}(p,p), h_{\mathcal{Y}}(q) = K_{\mathcal{Y}}(q,q)$ denote their root depths, and let $\mathcal{E}_T(P)$ denote the objective in Eq. 10 evaluated at a fixed coupling P.

Proposition 2 (Root-flow separation)**.** *If $P \in \Pi(\alpha,\beta)$ has zero aligned-root FlowAlign cost,*

$$\sum_{p,q} |h_{\mathcal{X}}(p) - h_{\mathcal{Y}}(q)|^2 P_{pq} = 0, \tag{15}$$

$$\mathcal{E}_T(P) = 4\sum_{p,p',q,q'} |K_{\mathcal{X}}(p,p') - K_{\mathcal{Y}}(q,q')|^2 P_{pq}P_{p'q'}. \tag{16}$$

Moreover, there exist rooted weighted tree mm-spaces for which $\mathrm{FA}_2 = 0$ *while* $\mathrm{GW}_{T,2} > 0$.

Proposition 2 isolates the pair-dependent information that is not determined by root-flow alignment. The diagonal entries of the co-depth matrices recover the unary root-depth statistics used by FlowAlign, whereas the off-diagonal entries encode shared ancestry between pairs of leaves. Consequently, even when the root-depth component is matched exactly, distinct pairwise tree geometries may remain. The proof and a strict-separation construction are given in Appendix A.1.6.

Theorem 1 shows that, once a sampled hierarchy and its induced leaf geometry are fixed, passing from the point-level coupling to the leaf-level coupling introduces no additional optimization gap. Therefore, relative to the original GW problem, the approximation arises from the finite-resolution replacement of the original intra-domain metrics by their tree-induced counterparts, rather than from the quotient reduction itself. Appendix A.1.7 quantifies this effect through a stability bound and decomposes the tree-metric distortion into within-leaf compression and inter-leaf geometric distortion.

Computational Complexity. Let n and m denote the numbers of samples in $\mathcal{X}$ and $\mathcal{Y}$, respectively, $N = \max\{n,m\}$, p the maximum ambient feature dimension of the two input spaces, S the number

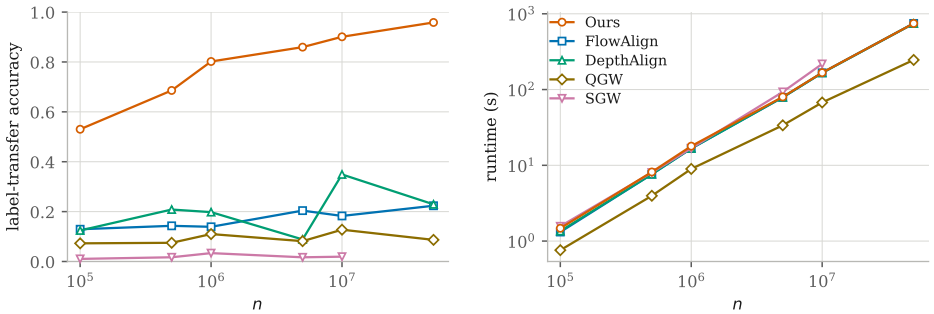
of sampled hierarchies, κ the branching factor, H the maximum tree depth, and $L \leq \min\{N, \kappa^H\}$ the largest number of realized nonempty leaves across slices. Constructing the S randomized hierarchies costs $\mathcal{O}(SH\kappa Np)$. Forming the co-depth-induced leaf geometries costs $\mathcal{O}(SHL^2)$. Solving Eq. 13 with I conditional-gradient iterations costs $\mathcal{O}(SIL^3)$ in the balanced case, giving a total complexity of $\mathcal{O}(SH\kappa Np + SHL^2 + SIL^3)$. Only the hierarchy-construction term depends directly on the original support size. Once L is fixed, the second-order comparison is confined to the leaf resolution and, by Theorem 1, introduces no additional optimization gap under the induced tree geometry. Appendix A.2 gives the full derivation.

3 Experiments

To rigorously evaluate T-GROW, we conduct experiments across synthetic setups and real-world applications. Specifically, we focus on three core objectives: (i) computational efficiency, evaluated through runtime scaling analysis; (ii) robustness, tested under controlled topological noise; and (iii) practical utility, demonstrated across applications to single-cell multi-omics and 3D computer vision. Throughout these experiments, we compare our approach with representative baselines, including unregularized GW (*GW*) (Peyré et al., 2016), Entropic GW (*EGW*) (Rioux et al., 2024), *FlowAlign* and *DepthAlign* (Le et al., 2021), Low-Rank GW (*LR-GW*) (Scetbon et al., 2022), Sliced GW (*SGW*) (Vayer et al., 2019), and Quantized GW (*QGW*) (Chowdhury et al., 2021).

3.1 Scalability

Large-Scale Synthetic Correspondence. To evaluate scalability and correspondence quality at large sample sizes, we construct a controlled synthetic benchmark inspired by the large-scale experiments of Scetbon et al. (2022). Two heterogeneous point clouds are independently sampled from the same latent hierarchy and embedded and perturbed in different feature spaces, with shared leaf-cluster identities defining the ground-truth correspondence. We vary the number of points per cloud as $n \in \{100\text{K}, 500\text{K}, 1\text{M}, 5\text{M}, 10\text{M}, 50\text{M}\}$. For each method, we use the estimated coupling to transfer source labels to the target and report label-transfer accuracy together with end-to-end runtime. All methods use a fixed configuration across scales, and the tree-based methods are evaluated on the same tree representations. GW and EGW are not included in this large-scale comparison because their dense pairwise computations become infeasible well before the sample sizes considered here. Full construction and experimental details are provided in Appendix A.3.



(a) accuracy (b) runtime

Figure 1: Large-scale synthetic correspondence. (a) Label-transfer accuracy and (b) Runtime as the number of points per cloud increases from 10^5 to 5×10^7.

As shown in Fig. 1, T-GROW consistently achieves the highest correspondence accuracy across all evaluated scales, with performance improving steadily as the sample size increases. Additional samples provide more reliable estimates of the underlying pairwise hierarchical relations, which T-GROW explicitly retains and exploits. In contrast, increasing the sample size cannot recover relational information discarded by root-flow summaries, so FlowAlign benefits much less from additional data. QGW achieves the lowest runtime, particularly at larger scales, but its correspondence accuracy remains substantially lower and shows no comparable improvement with increasing sam-

ple size. T-GROW maintains runtime scaling comparable to the lightweight tree-based baselines while achieving much stronger correspondence recovery. The results show a favorable accuracy–scalability trade-off, with T-GROW remaining practical at tens of millions of points while benefiting more from additional samples.

Real-World Large-Scale Scalability. To complement the synthetic scaling experiment, we further evaluate whether the recovered correspondences preserve semantic structure through large-scale segmentation transfer on S3DIS (Armeni et al., 2016), following the QGW protocol (Chowdhury et al., 2021). Given a labeled source point cloud and an unlabeled target point cloud, labels are transferred through the recovered correspondence. We report point-wise accuracy, mIoU, and runtime as mean $\pm$ standard deviation over three seeds. Full implementation details are provided in Appendix A.5.2. We evaluate two full-resolution scene pairs, the QGW-released Lobby pair (1.16M$\rightarrow$ 0.91M points) and an additional Lobby pair (1.16M$\rightarrow$ 2.05M points). For methods that operate on reduced representations, the full point clouds are first compressed to approximately 10^3 cells for correspondence estimation, while label transfer is evaluated on all target points. As shown in Table 1, T-GROW achieves the highest transfer accuracy and mIoU on both scene pairs while remaining computationally efficient at million-point scale.

	Released Lobby pair			*Additional Lobby pair*		
Method	Accuracy	mIoU	Runtime (s)	Accuracy	mIoU	Runtime (s)
FlowAlign	0.261 ± 0.048	0.076 ± 0.017	6.3 ± 0.1	0.180 ± 0.012	0.043 ± 0.005	9.4 ± 0.1
GW	0.283 ± 0.166	0.123 ± 0.067	22.1 ± 1.2	0.210 ± 0.146	0.071 ± 0.049	23.7 ± 4.7
DepthAlign	0.315 ± 0.023	0.086 ± 0.009	6.6 ± 0.3	0.196 ± 0.028	0.045 ± 0.004	10.3 ± 0.5
LR-GW	0.330 ± 0.115	0.085 ± 0.057	27.2 ± 10.8	0.229 ± 0.006	0.043 ± 0.005	28.6 ± 0.9
SGW	0.352 ± 0.024	0.096 ± 0.008	11.0 ± 0.0	0.243 ± 0.015	0.045 ± 0.006	15.2 ± 0.3
QGW	0.407 ± 0.023	0.170 ± 0.011	10.4 ± 0.5	0.191 ± 0.023	0.060 ± 0.004	8.3 ± 0.2
EGW	0.437 ± 0.010	0.171 ± 0.006	20.1 ± 0.4	0.228 ± 0.067	0.065 ± 0.021	21.5 ± 0.2
T-GROW	$\mathbf{0.462 \pm 0.034}$	$\mathbf{0.185 \pm 0.013}$	6.7 ± 0.1	$\mathbf{0.404 \pm 0.007}$	$\mathbf{0.121 \pm 0.001}$	7.9 ± 1.2

Table 1: Large-scale semantic segmentation transfer on two full-resolution S3DIS scene pairs.

3.2 Robustness

Robustness to Geometric Perturbations We evaluate robustness on a controlled synthetic matching task with known correspondences. Each problem contains paired source and target point-cloud distributions, where the targets are randomly permuted and methods must recover the correct pairing from structural discrepancies. We denote by n the number of sampled points per distribution and by ρ the perturbation strength. To probe robustness at different structural scales, we consider two complementary perturbations: a *local perturbation*, which rotates local substructures around branch centers and primarily alters fine-scale geometry, and a *global perturbation*, which rotates entire branches around the global origin and mainly changes their relative organization.

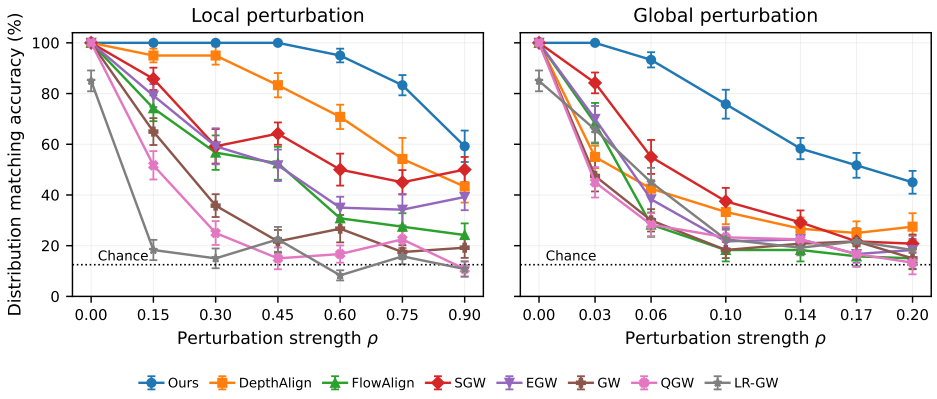


Figure 2: Robustness to local and global geometric perturbations. Accuracy is reported as mean $\pm$ std over 15 independent trials.

As shown in Fig. 2, our discrepancy degrades more gradually than the competing methods under both types of geometric mismatch. The local perturbation stresses fine-scale structure within branches, whereas the global perturbation changes the organization of whole branches. The consistent advantage across the two regimes indicates that the T-GROW discrepancy is not tied to a single preserved geometric scale but can exploit relational information distributed across multiple levels of the hierarchy. Full details are given in Appendix A.4.1.

Real-World Robustness to Label Corruption. We next move from synthetic perturbations to a real cross-modal problem in which no cross-domain cost is available. In the blind vision–language matching benchmark of Schnaus et al. (2025), the correspondence between class-level vision and language embeddings is withheld and recovered from within-modality relations alone, making it a stringent test of how much relational structure a discrepancy retains. We evaluate on CIFAR-100 (Krizhevsky et al., 2009) with CLIP vision embeddings (Radford et al., 2021) and all-mpnet-base-v2 language embeddings (Reimers & Gurevych, 2019). To progressively weaken the available structural signal, we corrupt a fraction ρ of the image labels before class-level averaging, thereby distorting the visual within-modality structure while preserving class sizes and the overall matching instance. Full details are given in Appendix A.4.2.

Table 2: Robustness to vision-side label noise on CIFAR-100 at $N{=}100$

Method	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	Runtime at $\rho = 0$ (s) $\downarrow$
SGW	0.014 ± 0.017	0.028 ± 0.019	0.044 ± 0.048	0.183 ± 0.004
DepthAlign	0.026 ± 0.019	0.030 ± 0.020	0.070 ± 0.043	1.716 ± 0.060
FlowAlign	0.032 ± 0.030	0.052 ± 0.035	0.088 ± 0.049	1.596 ± 0.079
LR-GW	0.178 ± 0.068	0.110 ± 0.072	0.148 ± 0.109	0.388 ± 0.043
EGW	0.250 ± 0.111	0.276 ± 0.085	0.192 ± 0.096	0.912 ± 0.049
QGW	0.294 ± 0.091	0.224 ± 0.127	0.170 ± 0.088	0.454 ± 0.039
GW	0.300 ± 0.146	0.120 ± 0.065	0.026 ± 0.021	3.277 ± 1.501
T-GROW	$\mathbf{0.656 \pm 0.096}$	$\mathbf{0.564 \pm 0.052}$	$\mathbf{0.396 \pm 0.132}$	2.340 ± 0.040

Table 2 reports results for $\rho \in \{0, 0.3, 0.5\}$ at $N = 100$. Ours achieves the highest matching accuracy across all corruption levels, including a substantial margin under moderate corruption. FlowAlign remains close to chance throughout, indicating that its root-flow summaries retain insufficient relational information for this blind matching task.

3.3 Cross-modal Alignment: Single-Cell Multi-Omics

We evaluate cross-modal alignment on SNARE-seq (Chen et al., 2019), which contains $1{,}047$ paired cells measured across two modalities. Following the data-processing setup of Scetbon et al. (2022), the two modalities are represented by 10- and 19-dimensional features after preprocessing. For each modality, we construct a k-nearest-neighbor graph with $k = 110$ using correlation-based neighborhoods and use shortest-path distances on the graph as the intra-modal geometry. We report tie-aware "fraction of samples closer than the true match" (FOSCTTM) error (Liu et al., 2019), which measures the fraction of samples in the other modality that are closer than its true match for each query cell, label accuracy, and runtime, with mean $\pm$ standard deviation over five runs. Complete configurations and the sensitivity analysis of LR-GW are provided in Appendix A.5.1.

Table 3: Cross-modal alignment on SNARE-seq.

Method	FOSCTTM $\downarrow$	Label Acc. $\uparrow$	Runtime (s) $\downarrow$
FlowAlign	0.5312 ± 0.0420	0.2821 ± 0.0640	1.49 ± 0.08
DepthAlign	0.4498 ± 0.0966	0.3885 ± 0.1475	0.66 ± 0.02
SGW	0.4207 ± 0.0322	0.3786 ± 0.0112	0.36 ± 0.01
QGW	0.3009 ± 0.2099	0.6819 ± 0.3639	0.41 ± 0.06
GW	0.2392 ± 0.0000	0.7011 ± 0.0000	7.97 ± 0.20
LR-GW	0.2260 ± 0.1038	0.7349 ± 0.2588	103.55 ± 14.61
EGW	0.1496 ± 0.0000	0.9379 ± 0.0000	138.26 ± 0.61
T-GROW	$\mathbf{0.1475 \pm 0.0012}$	$\mathbf{0.9381 \pm 0.0010}$	0.69 ± 0.03

As shown in Table 3, T-GROW achieves the best overall performance on SNARE-seq, yielding the lowest FOSCTTM (0.1475 ± 0.0012) and highest Label Accuracy (0.9381 ± 0.0010). While EGW is the closest baseline in alignment quality (0.1496 FOSCTTM, 0.9379 accuracy), it requires 138.26s. In contrast, T-GROW slightly surpasses EGW while executing in just 0.69s ($\sim 200\times$ speedup). Compared to fast tree-based baselines (e.g., FlowAlign, DepthAlign, SGW), T-GROW successfully bridges the gap between scalable surrogates and exact GW quality, offering an favorable trade-off between accuracy and efficiency on heterogeneous multi-omics data.

3.4 Blind Vision-Language Matching

We next evaluate on the blind vision-language matching benchmark of Schnaus et al. (2025), where class-level correspondences between CLIP vision (Radford et al., 2021) and all-mpnet-base-v2 language embeddings (Reimers & Gurevych, 2019) are recovered solely from within-modality geometry. On CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-100 (Russakovsky et al., 2015) ($N = 100$ classes), we report matching accuracy and runtime over five seeds (see Appendix A.4.2 for full results). As shown in Table 4, T-GROW achieves top matching accuracy on both datasets while remaining computationally fast. Notably, it outperforms the task-specific BlindMatch solver (FHG), which hits the $3600\,s$ timeout at $N = 100$. This confirms T-GROW's ability to uncover accurate cross-modal alignments purely from unlabelled relational structures.

Table 4: Blind vision–language matching on CIFAR-100 and ImageNet-100 at N=100

	CIFAR-100		ImageNet-100	
Method	Acc.	Runtime (s)	Acc.	Runtime (s)
SGW	0.014 ± 0.017	0.18 ± 0.00	0.046 ± 0.027	0.20 ± 0.02
DepthAlign	0.026 ± 0.019	1.72 ± 0.06	0.008 ± 0.016	1.69 ± 0.06
FlowAlign	0.032 ± 0.030	1.60 ± 0.08	0.060 ± 0.040	1.58 ± 0.08
LR-GW	0.178 ± 0.068	0.39 ± 0.04	0.054 ± 0.036	0.48 ± 0.03
QGW	0.294 ± 0.091	0.45 ± 0.04	0.076 ± 0.014	0.59 ± 0.07
EGW	0.250 ± 0.111	0.91 ± 0.05	0.084 ± 0.029	0.78 ± 0.09
GW	0.300 ± 0.146	3.28 ± 1.50	0.056 ± 0.022	2.54 ± 0.42
T-GROW	$\mathbf{0.656 \pm 0.096}$	2.34 ± 0.04	$\mathbf{0.094 \pm 0.016}$	2.99 ± 0.08
FHG (Schnaus et al., 2025)	0.033 ± 0.015	$3600^{\dagger}$	0.037 ± 0.015	$3600^{\dagger}$

4 Conclusion and Limitations

We introduced T-GROW, a scalable tree-based formulation of Gromov–Wasserstein discrepancy that preserves pairwise relational structure through co-depth. By encoding shared ancestry, co-depth incorporates root-to-node path lengths along its diagonal while recovering the induced tree metric. For a fixed hierarchy, we proved that the leaf-level GW problem is exactly equivalent to the sample-level problem under the induced tree geometry, this sample-to-leaf reduction introduces no additional optimization gap. This shifts the second-order optimization from the original support scale to a hierarchy-controlled leaf resolution, while retaining the pairwise geometry of the hierarchy. Across large-scale synthetic and real-world tasks, T-GROW achieves strong matching quality and robustness while maintaining scalable runtime.

Limitations and Future Work. The main limitation is that the tree geometry is only a finite-resolution approximation of the original GW distance. At a fixed resolution, T-GROW scales linearly with the number of samples because the second-order optimization is confined to the leaf space, while increasing the leaf resolution improves geometric fidelity at a higher computational cost. T-GROW also depends on the tree-construction scheme and structural hyperparameters, although averaging over multiple randomized hierarchies can reduce sensitivity to individual partitions. Extending T-GROW through adaptive or learned hierarchy construction, together with accelerated high-resolution solvers, is a promising direction for future work.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.) In this work, we used generative AI tools for polishing the writing. Additionally, we used generative AI tools for finding related work, and for editing and debugging software code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.) The research uses existing datasets and does not involve human-subject experiments, private data collection, or deployment of high-risk systems. We have reviewed the ICLR Code of Ethics and believe that the work conforms to its requirements.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The current submission does not yet provide an anonymized public code release. The experiments use publicly available datasets, and the appendix provides detailed dataset descriptions, experimental settings, and algorithmic specifications. We plan to release the implementation, scripts, and reproduction instructions upon acceptance. And we provide the full details of datasets, experimental setting and specific algorithm in the appendix.

REFERENCES

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1881–1890, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. *Proceedings of Machine Learning Research*, 70:214–223, 06–11 Aug 2017.

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1534–1543, 2016. doi: 10.1109/CVPR.2016.170.

Arturs Backurs, Yihe Dong, Piotr Indyk, Ilya Razenshteyn, and Tal Wagner. Scalable nearest neighbor search for optimal transport. In *International Conference on machine learning*, pp. 497–506. PMLR, 2020.

Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.

Samir Chowdhury, David Miller, and Tom Needham. Quantized gromov-wasserstein. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 811–827. Springer, 2021.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. *Proceedings of Machine Learning Research*, 32(2):685–693, 2014.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Scot: single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29(1):3–18, 2022.

Kira Michaela Düsterwald, Samo Hromadka, and Makoto Yamada. Fast unsupervised ground metric learning with tree-wasserstein distance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FBhKUXK7od.

Marco Fiorucci, Peter Naylor, and Makoto Yamada. Optimal transport for change detection on lidar point clouds. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 982–985, 2023.

Mikhael Gromov. Hyperbolic groups. In *Essays in group theory*, pp. 75–263. Springer, 1987.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=MRmI68k3gd.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14593–14605. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7a6a6127ff85640ec69691fb0f7cb1a2-Paper.pdf.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Tam Le, Nhat Ho, and Makoto Yamada. Flow-based alignment approaches for probability measures in different spaces. In *International conference on artificial intelligence and statistics*, pp. 3934–3942. PMLR, 2021.

Tam Le, Truyen Nguyen, and Kenji Fukumizu. Optimal transport for measures with noisy tree metric. In *International Conference on Artificial Intelligence and Statistics*, pp. 3115–3123. PMLR, 2024.

Haoyang Liu, Yijiang Li, Tiancheng Xing, Peiran Wang, Vibhu Dalal, Luwei Li, Jingrui He, and Haohan Wang. Dataset distillation via the wasserstein metric. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1205–1215, 2025. doi: 10.1109/ICCV51701.2025.00120.

Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 10:1–10:13. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.WABI.2019.10.

Facundo Mémoli. Gromov-wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

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

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24983–24992. IEEE, 2025.

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.

Yuki Takezawa, Ryoma Sato, and Makoto Yamada. Supervised tree-wasserstein distance. In *International Conference on Machine Learning*, pp. 10086–10095. PMLR, 2021.

Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Hoang V. Tran, Minh-Khoi Nguyen-Nhat, Huyen Trang Pham, Thanh Chu, Tam Le, and Tan Minh Nguyen. Distance-based tree-sliced wasserstein distance. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=OiQttMHwce.

Hoang V. Tran, Huyen Trang Pham, Tho Tran Huu, Minh-Khoi Nguyen-Nhat, Thanh Chu, Tam Le, and Tan Minh Nguyen. Tree-sliced wasserstein distance: A geometric perspective. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=StaRAs9n49.

Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. *Advances in Neural Information Processing Systems*, 32, 2019.

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled wasserstein learning for word embedding and topic modeling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/22fb0cee7e1f3bde58293de743871417-Paper.pdf.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pp. 6932–6941. PMLR, 2019.

Xi Yu, Xiang Gu, Zhihao Shi, and Jian Sun. Wasserstein style distribution analysis and transform for stylized image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17496–17505, 2025. doi: 10.1109/ICCV51701.2025.01625.

A Appendix

A.1 Theoretical proof

A.1.1 Co-Depth Representation

Connection to the Gromov product. For a metric space (V, d) with base point r, the Gromov product of $u, v \in V$ with respect to r is defined as

$$(u \mid v)_r := \frac{1}{2}\left(d(r,u) + d(r,v) - d(u,v)\right). \tag{17}$$

For a rooted weighted tree $T = (V, E, w, r)$, the unique-path structure gives

$$d_T(u,v) = d_T(r,u) + d_T(r,v) - 2d_T(r, \mathrm{LCA}(u,v)). \tag{18}$$

Combining Eq. 18 with Definition 1 yields

$$c_T(u,v) = d_T(r, \mathrm{LCA}(u,v)) = (u \mid v)_r. \tag{19}$$

Hence, co-depth is precisely the Gromov product of the tree metric with respect to the root. Equivalently,

$$d_T(u,v) = c_T(u,u) + c_T(v,v) - 2c_T(u,v), \tag{20}$$

since $c_T(u,u) = d_T(r,u)$.

Proof of Proposition 1. The paths $P(r,u)$ and $P(r,v)$ overlap exactly on $P(r, \mathrm{LCA}(u,v))$. Hence

$$\begin{aligned}\langle \phi_T(u), \phi_T(v) \rangle &= \sum_{e \in E} w_e \mathbf{1}\{e \in P(r,u)\}\mathbf{1}\{e \in P(r,v)\} \\ &= \sum_{e \in P(r, \mathrm{LCA}(u,v))} w_e \\ &= d_T(r, \mathrm{LCA}(u,v)) = c_T(u,v),\end{aligned} \tag{21}$$

which proves Eq. 7.

For the selected nodes $\{u_1, \ldots, u_L\}$,

$$\begin{aligned}[A_T W_T A_T^\top]_{pq} &= \sum_{e \in E} w_e \mathbf{1}\{e \in P(r,u_p)\}\mathbf{1}\{e \in P(r,u_q)\} \\ &= c_T(u_p, u_q) = [K_T]_{pq}.\end{aligned} \tag{22}$$

Therefore, $K_T = A_T W_T A_T^\top$. Since the edge weights are nonnegative, $W_T \succeq 0$, and consequently $K_T \succeq 0$. Thus, the co-depth matrix is the Gram matrix associated with the weighted root-path features ϕ_T. □

Connection to classical metric-kernel theory. The positive semidefiniteness of K_T also admits a classical metric-kernel interpretation. A weighted tree metric can be expressed as a nonnegative weighted sum of cut metrics and is therefore conditionally negative definite. For any conditionally negative definite metric d with $d(x,x) = 0$ and any fixed base point r, the base-point kernel

$$k_r(u,v) := \frac{1}{2}\left(d(r,u) + d(r,v) - d(u,v)\right) \tag{23}$$

is positive semidefinite. Since Eq. 19 shows that $c_T(u,v) = (u \mid v)_r = k_r(u,v)$ for the tree metric d_T, classical metric-kernel theory provides an alternative explanation for $K_T \succeq 0$.

Proposition 1 gives a more explicit tree-specific realization, rather than establishing positive semidefiniteness only abstractly, it identifies the feature coordinates with weighted root-path incidences and yields the constructive factorization $K_T = A_T W_T A_T^\top$.

A.1.2 TREE-QUOTIENT REDUCTION

Proof of Theorem 1. The argument uses only that the induced relations $\widetilde{d}_X$ and $\widetilde{d}_Y$ are constant on each pair of quotient cells; it does not depend on the particular form of D_X and D_Y.

Let $\pi \in \Pi(a, b)$ and aggregate its transported mass over pairs of quotient cells:

$$P_{pq} := \sum_{\substack{x_i \in C_p^{\mathcal{X}} \\ y_j \in C_q^{\mathcal{Y}}}} \pi_{ij}. \tag{24}$$

Summing the marginal constraints of π over each quotient cell gives

$$P\mathbf{1} = \alpha, \qquad P^\top \mathbf{1} = \beta,$$

and $P \in \Pi(\alpha, \beta)$.

By Eq. 11, $\widetilde{d}_{\mathcal{X}}(x_i, x_{i'}) = D_{\mathcal{X}}(p, p')$ for $x_i \in C_p^{\mathcal{X}}$ and $x_{i'} \in C_{p'}^{\mathcal{X}}$, and analogously $\widetilde{d}_{\mathcal{Y}}(y_j, y_{j'}) = D_{\mathcal{Y}}(q, q')$ for $y_j \in C_q^{\mathcal{Y}}$ and $y_{j'} \in C_{q'}^{\mathcal{Y}}$. Hence the point-level objective is block-constant with respect to the quotient partitions, and

$$\begin{aligned} &\sum_{i,i',j,j'} \left| \widetilde{d}_{\mathcal{X}}(x_i, x_{i'}) - \widetilde{d}_{\mathcal{Y}}(y_j, y_{j'}) \right|^2 \pi_{ij} \pi_{i'j'} \\ &\qquad = \sum_{p,p',q,q'} \left| D_{\mathcal{X}}(p, p') - D_{\mathcal{Y}}(q, q') \right|^2 P_{pq} P_{p'q'}. \end{aligned} \tag{25}$$

Conversely, after removing quotient cells with zero mass, let $P \in \Pi(\alpha, \beta)$. Define a point-level coupling by

$$\pi_{ij} = P_{pq} \frac{a_i}{\alpha_p} \frac{b_j}{\beta_q}, \qquad x_i \in C_p^{\mathcal{X}}, \quad y_j \in C_q^{\mathcal{Y}}. \tag{26}$$

Its row marginals satisfy $\sum_j \pi_{ij} = \frac{a_i}{\alpha_p} \sum_q P_{pq} = a_i$ and similarly $\sum_i \pi_{ij} = b_j$. Thus $\pi \in \Pi(a, b)$. Moreover, aggregating Eq. 26 over a pair of quotient cells recovers P_{pq}, and Eq. 25 shows that the lifted coupling attains exactly the same objective value.

Every feasible point-level coupling induces a feasible quotient-level coupling with the same cost, and every feasible quotient-level coupling admits a point-level lift with the same cost. Taking the infimum over the two feasible sets yields Eq. 12. □

This also shows that the exact quotient reduction is determined by the partition and the induced block-constant relations, rather than by a particular numerical solver.

A.1.3 VALIDITY AND METRIC STRUCTURE OF THE LEAF GEOMETRY

Lemma 1 (Validity of the leaf geometry)**.** *Assume that every path connecting two distinct leaves has positive total length. Then $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ in Eq. 9 are metrics on $\mathcal{L}_{\mathcal{X}}$ and $\mathcal{L}_{\mathcal{Y}}$, respectively. If zero-length paths between distinct leaves are allowed, they are pseudometrics.*

Proof. By Eq. 9, $D_{\mathcal{X}}(p, p')$ is exactly the weighted path distance between u_p and $u_{p'}$ in $T_{\mathcal{X}}$. Therefore,

$$D_{\mathcal{X}}(p, p') \geq 0, \qquad D_{\mathcal{X}}(p, p') = D_{\mathcal{X}}(p', p), \qquad D_{\mathcal{X}}(p, p) = 0.$$

For any three leaves $u_p, u_{p'}, u_{p''}$, the unique path from u_p to $u_{p''}$ cannot be longer than the walk obtained by concatenating the paths from u_p to $u_{p'}$ and from $u_{p'}$ to $u_{p''}$. Hence

$$D_{\mathcal{X}}(p, p'') \leq D_{\mathcal{X}}(p, p') + D_{\mathcal{X}}(p', p'').$$

Thus the triangle inequality holds.

Under the positive-path-length assumption, two distinct leaves have strictly positive distance, so

$$D_{\mathcal{X}}(p, p') = 0 \quad \Longleftrightarrow \quad p = p'.$$

Therefore $D_{\mathcal{X}}$ is a metric. The same argument applies to $D_{\mathcal{Y}}$. If distinct leaves can be connected by a zero-length path, nonnegativity, symmetry, and the triangle inequality remain valid, while identity of indiscernibles may fail. In that case the induced leaf geometry is a pseudometric. □

Corollary 1 (Metric structure at leaf resolution)**.** *Under the assumptions of Lemma 1, the leaf representations*

$$\bar{\mathcal{X}}_T = (\mathcal{L}_\mathcal{X}, D_\mathcal{X}, \bar{\mu}), \qquad \bar{\mathcal{Y}}_T = (\mathcal{L}_\mathcal{Y}, D_\mathcal{Y}, \bar{\nu})$$

are finite metric-measure spaces, and $\mathrm{GW}_{T,2}(\mu,\nu)$ *in Eq. 10 is Gromov–Wasserstein distance. Consequently, it defines a metric for the corresponding measure-preserving isometry classes. In particular,* $\mathrm{GW}_{T,2}(\mu,\nu) = 0$ *if and only if* $\bar{\mathcal{X}}_T$ *and* $\bar{\mathcal{Y}}_T$ *are measure-preserving isometric, and the triangle inequality holds on these equivalence classes.*

A.1.4 Hierarchical Tree Construction and T-GROW Algorithm

We adopt the clustering-based tree construction of Le et al. (2019). Given a weighted point cloud (X, a), branching factor κ, maximum depth H, and slice randomness ξ, the procedure starts from a root cluster containing the entire support. At each level, every nonempty cluster is partitioned into at most κ children. The first seed is selected according to ξ, and the remaining seeds are chosen sequentially by farthest-point selection. The samples are then assigned to their nearest anchors. Each resulting cluster C is represented by an anchor $\mathrm{anc}(C)$, and the length of a parent–child edge (C, C') is defined as $w(C, C') = d_\mathcal{X}(\mathrm{anc}(C), \mathrm{anc}(C'))$. The anchor depends on the representation of the input domain. For vector-valued data, we use the weighted centroid $\mathrm{anc}(C) = \frac{\sum_{x_i \in C} a_i x_i}{\sum_{x_i \in C} a_i}$, which is computationally efficient for the large-scale point-cloud experiments. For a general metric space, where an arithmetic mean need not be defined, we instead use the medoid $\mathrm{anc}(C) = \arg\min_{x \in C} \sum_{x' \in C} d_\mathcal{X}(x, x')$. The medoid is selected from the original support using pairwise distance evaluations. Thus, the hierarchy construction can be applied without requiring Euclidean coordinates.

For slice s, let $T_\mathcal{X}^{(s)} = \mathcal{B}_{\kappa,H}(\mathcal{X}; \xi_s)$ denote the hierarchy constructed for $\mathcal{X}$. Its nonempty leaves define the clusters $\{C_p^{\mathcal{X},(s)}\}_{p=1}^{L_\mathcal{X}^{(s)}}$, where $L_\mathcal{X}^{(s)} := \left|\mathcal{L}_\mathcal{X}^{(s)}\right| \le \min\{n, \kappa^H\}$ is the realized number of nonempty leaves, and the corresponding leaf masses are $\alpha_p^{(s)} = \sum_{x_i \in C_p^{\mathcal{X},(s)}} a_i$. The corresponding quantities for $\mathcal{Y}$ are defined analogously, with $L_\mathcal{Y}^{(s)} := \left|\mathcal{L}_\mathcal{Y}^{(s)}\right| \le \min\{m, \kappa^H\}$. We use $L_{\max} := \kappa^H$ to denote the nominal maximum number of leaves in a depth-H, branching-κ hierarchy.

Let $A_\mathcal{X}^{(s)}$ denote the root-path incidence matrix, $[A_\mathcal{X}^{(s)}]_{pe} = \mathbf{1}\{e \in P(r_\mathcal{X}, u_p)\}$, and let $W_\mathcal{X}^{(s)} = \mathrm{diag}\left(\{w_e^{(s)}\}_{e \in E(T_\mathcal{X}^{(s)})}\right)$ collect the corresponding edge lengths. By Proposition 1, the leaf co-depth matrix is

$$K_\mathcal{X}^{(s)} = A_\mathcal{X}^{(s)} W_\mathcal{X}^{(s)} (A_\mathcal{X}^{(s)})^\top. \tag{27}$$

Let

$$h_\mathcal{X}^{(s)} = \mathrm{diag}\left(K_\mathcal{X}^{(s)}\right).$$

The corresponding leaf-level tree metric is then recovered as

$$D_\mathcal{X}^{(s)} = h_\mathcal{X}^{(s)} \mathbf{1}^\top + \mathbf{1}(h_\mathcal{X}^{(s)})^\top - 2K_\mathcal{X}^{(s)}. \tag{28}$$

The same construction yields $\beta^{(s)}$, $K_\mathcal{Y}^{(s)}$, and $D_\mathcal{Y}^{(s)}$.

The s-th randomized hierarchy produces the leaf-level metric-measure representations $\left(\mathcal{L}_\mathcal{X}^{(s)}, D_\mathcal{X}^{(s)}, \alpha^{(s)}\right), \left(\mathcal{L}_\mathcal{Y}^{(s)}, D_\mathcal{Y}^{(s)}, \beta^{(s)}\right)$, which are precisely the encoded spaces compared by the reduced GW problem in Eq. 13.

By Proposition 1, the matrix computed in Algorithm 1 is equivalently $K = AWA^\top$. Computing each entry from the common root path requires at most H levels, giving $O(HL^2)$ time without forming a dense node-incidence product.

Algorithms 1 and 2 define the canonical T-GROW discrepancy through the optimal value of each GW problem at the reduced leaf-level. For experiments using the canonical soft-coupling realization of Eq. 13, we solve this problem with a conditional-gradient solver for unregularized quadratic GW. Other downstream tasks may use task-specific optimization while retaining the same tree-induced representation. The quotient result in Theorem 1 concerns the underlying reduced optimization problem and is independent of the numerical solver used to approximate its optimum.

Algorithm 1 Randomized Tree Encoder

Require: Weighted empirical measure $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$, branching factor κ, maximum depth H, random seed ξ

Ensure: Leaf representation $(\mathcal{L}, D, \alpha)$

1: Construct $T \leftarrow \mathcal{B}_{\kappa,H}(X; \xi)$ using the randomized hierarchical clustering procedure described above

2: Retain the nonempty leaves $\mathcal{L} = \{u_1, \ldots, u_L\}$, where

$$L \leq \min\{n, L_{\max}\}, \qquad L_{\max} = \kappa^H$$

3: Compute leaf masses

$$\alpha_p \leftarrow \sum_{x_i \in C_p} a_i$$

4: Compute the co-depth matrix

$$K_{pp'} \leftarrow c_T(u_p, u_{p'})$$

equivalently,

$$K = AWA^\top$$

5: Set $h \leftarrow \mathrm{diag}(K)$ and recover

$$D \leftarrow h\mathbf{1}^\top + \mathbf{1}h^\top - 2K$$

6: **return** $(\mathcal{L}, D, \alpha)$

Algorithm 2 Canonical T-GROW

Require: Weighted empirical measures $\mathcal{X} = (X, d_X, \mu)$ and $\mathcal{Y} = (Y, d_Y, \nu)$; branching factor κ; maximum depth H; number of slices S; fixed seeds $\{\xi_s\}_{s=1}^{S}$

Ensure: $\text{T-GROW}_{2,S}(\mathcal{X}, \mathcal{Y})$

1: **for** $s = 1, \ldots, S$ **do**

2: $(L_X^{(s)}, D_X^{(s)}, \alpha^{(s)}) \leftarrow \mathrm{TreeEncode}(X; \kappa, H, \xi_s)$

3: $(L_Y^{(s)}, D_Y^{(s)}, \beta^{(s)}) \leftarrow \mathrm{TreeEncode}(Y; \kappa, H, \xi_s)$

4: Define

$$E_s(P) := \sum_{p,p',q,q'} \left| [D_X^{(s)}]_{pp'} - [D_Y^{(s)}]_{qq'} \right|^2 P_{pq} P_{p'q'}.$$

5: $\widehat{P}_s \in \arg\min_{P \in \Pi(\alpha^{(s)}, \beta^{(s)})} \mathcal{E}_s(P)$

6: $d_s^2 \leftarrow \mathcal{E}_s(\widehat{P}_s)$

7: **end for**

8: $\text{T-GROW}_{2,S} \leftarrow \left(\frac{1}{S} \sum_{s=1}^{S} d_s^2 \right)^{1/2}$ **return** $\text{T-GROW}_{2,S}$

A.1.5 Coupling Extraction and Cross-Slice Aggregation

In addition to the sliced GW costs used in Eq. 14, Algorithm 2 yields an optimized leaf-level coupling $\widehat{P}_s \in \Pi(\alpha^{(s)}, \beta^{(s)})$ for each randomized tree slice s. Downstream correspondence tasks use these plans through a within-slice lift followed by cross-slice aggregation. Following the lifting construction of Theorem 1, the leaf-level plan $\widehat{P}_s$ induces the sample-level coupling

$$[\Pi_s]_{ij} = [\widehat{P}_s]_{pq} \frac{a_i b_j}{\alpha_p^{(s)} \beta_q^{(s)}}, \qquad p = q_{\mathcal{X}}^{(s)}(x_i), \quad q = q_{\mathcal{Y}}^{(s)}(y_j). \tag{29}$$

The lifted plan satisfies $\Pi_s \mathbf{1} = a$ and $\Pi_s^\top \mathbf{1} = b$, and hence $\Pi_s \in \Pi(a, b)$. When a soft point-level correspondence is required, we uniformly average the slice-specific lifted plans:

$$\overline{\Pi} = \frac{1}{S} \sum_{s=1}^{S} \Pi_s. \tag{30}$$

Since $\Pi(a, b)$ is convex, $\overline{\Pi} \in \Pi(a, b)$. The averaged plan is interpreted as an ensemble correspondence across the randomized hierarchies rather than as the optimizer of any single slice objective.

For large-scale label transfer, the dense matrix $\overline{\Pi} \in \mathbb{R}^{n \times m}$ is not explicitly materialized. Instead, the transported class scores are computed directly from the leaf-level plans. For a labeled source support, define the class profile of the source leaf p by

$$\widetilde{H}_{p,c}^{(s)} = \frac{1}{\alpha_p^{(s)}} \sum_{x_i \in C_p^{\mathcal{X},(s)}} a_i \, \mathbf{1}\{\ell_i = c\}. \tag{31}$$

The aggregated class score for target point y_j is

$$V_{j,c} = \frac{1}{S} \sum_{s=1}^{S} \frac{[(\widehat{P}_s)^\top \widetilde{H}^{(s)}]_{q_{\mathcal{Y}}^{(s)}(y_j), c}}{\beta^{(s)}_{q_{\mathcal{Y}}^{(s)}(y_j)}}, \qquad \widehat{\ell}_j = \arg\max_c V_{j,c}. \tag{32}$$

Thus, label transfer evaluates the class mass induced by the slice-specific couplings without constructing a dense point-level coupling. The averaging in Eq. 30 applies to soft couplings in $\Pi(a, b)$. For tasks whose feasible outputs are permutations, direct averaging is not admissible, and the cost-weighted consensus used for blind matching is described in Appendix A.4.2.

A.1.6 Metric Structure and Root-Flow Separation

Proposition 3 (Pseudometricity under shared slicing)**.** *Fix a collection of slice encoders $\{\Phi_s\}_{s=1}^S$, where each Φ_s maps an input mm-space to a leaf mm-space satisfying the conditions of Lemma 1. If the same encoders are used consistently across all pairwise comparisons, then* $\text{T-GROW}_{2,S}$ *in Eq. 14 is a pseudometric on the input domain.*

Proof. For every slice, $d_s(\mathcal{X}, \mathcal{Y}) = \mathrm{GW}_2\left(\Phi_s(\mathcal{X}), \Phi_s(\mathcal{Y})\right)$. Since GW_2 is a metric on measure-preserving isometry classes of metric-measure spaces, d_s is a metric on the image of Φ_s, and therefore a pseudometric on the original input domain. In particular, $d_s(\mathcal{X}, \mathcal{Z}) \le d_s(\mathcal{X}, \mathcal{Y}) + d_s(\mathcal{Y}, \mathcal{Z})$.

Applying Minkowski's inequality across the S slices gives

$$\begin{aligned} \text{T-GROW}_{2,S}(\mathcal{X}, \mathcal{Z}) &= \frac{1}{\sqrt{S}} \left\| \left(d_s(\mathcal{X}, \mathcal{Z})\right)_{s=1}^S \right\|_2 \\ &\le \frac{1}{\sqrt{S}} \left\| \left(d_s(\mathcal{X}, \mathcal{Y})\right)_{s=1}^S \right\|_2 + \frac{1}{\sqrt{S}} \left\| \left(d_s(\mathcal{Y}, \mathcal{Z})\right)_{s=1}^S \right\|_2 \\ &= \text{T-GROW}_{2,S}(\mathcal{X}, \mathcal{Y}) + \text{T-GROW}_{2,S}(\mathcal{Y}, \mathcal{Z}). \end{aligned} \tag{33}$$

Nonnegativity and symmetry hold slice-wise. Identity of indiscernibles may fail on the original input domain because distinct inputs can induce identical finite-resolution leaf representations. □

Proof of Proposition 2. Suppose that $P \in \Pi(\alpha, \beta)$ satisfies Eq. 15. Since every summand in that equation is nonnegative,

$$P_{pq} > 0 \quad \Longrightarrow \quad h_{\mathcal{X}}(p) = h_{\mathcal{Y}}(q).$$

Consequently, whenever $P_{pq}P_{p'q'} > 0$,

$$h_{\mathcal{X}}(p) = h_{\mathcal{Y}}(q), \qquad h_{\mathcal{X}}(p') = h_{\mathcal{Y}}(q').$$

Using $D_{\mathcal{X}}(p, p') = h_{\mathcal{X}}(p) + h_{\mathcal{X}}(p') - 2K_{\mathcal{X}}(p, p')$ and the analogous identity for $\mathcal{Y}$, the root-depth terms cancel

$$D_{\mathcal{X}}(p, p') - D_{\mathcal{Y}}(q, q') = -2\left[K_{\mathcal{X}}(p, p') - K_{\mathcal{Y}}(q, q')\right]$$

for every term with positive weight. Squaring and summing with weights $P_{pq}P_{p'q'}$ proves Eq. 16.

To establish strict separation, fix $L \geq 2$, $\rho > 0$, and $a \in (0, \rho)$, and assign the same positive masses to L leaves in both trees. Let $T_{\mathcal{X}}$ be a rooted star in which every leaf is connected directly to the root by an edge of length ρ. Let $T_{\mathcal{Y}}$ contain a common edge of length a from the root to an internal node, followed by L leaf edges of length $\rho - a$. Every leaf has root depth ρ in both trees, so the aligned-root FlowAlign discrepancy is zero.

For every pair of distinct leaves, $D_{\mathcal{X}}(p, p') = 2\rho, D_{\mathcal{Y}}(p, p') = 2(\rho - a)$. Hence the two leaf mm-spaces are not measure-preserving isometric. By the identity property of the Gromov–Wasserstein distance, $\mathrm{GW}_{T,2} > 0$, which establishes the strict separation. □

A.1.7 Approximation Stability and Distortion

For each slice s, let $\widetilde{d}^{(s)}_{\mathcal{X}}$ and $\widetilde{d}^{(s)}_{\mathcal{Y}}$ denote the pullbacks of $D^{(s)}_{\mathcal{X}}$ and $D^{(s)}_{\mathcal{Y}}$ to the original supports as in Eq. 11. Define the pointwise metric errors

$$\Delta^{(s)}_{\mathcal{X}}(i, i') := d_{\mathcal{X}}(x_i, x_{i'}) - \widetilde{d}^{(s)}_{\mathcal{X}}(x_i, x_{i'}), \tag{34}$$

$$\Delta^{(s)}_{\mathcal{Y}}(j, j') := d_{\mathcal{Y}}(y_j, y_{j'}) - \widetilde{d}^{(s)}_{\mathcal{Y}}(y_j, y_{j'}). \tag{35}$$

We first measure the distortion in the mass-weighted L^2 geometry naturally associated with the quadratic GW objective:

$$\delta^{(s)}_{\mathcal{X}} := \left[\sum_{i,i'} \left|\Delta^{(s)}_{\mathcal{X}}(i, i')\right|^2 a_i a_{i'}\right]^{1/2}, \delta^{(s)}_{\mathcal{Y}} := \left[\sum_{j,j'} \left|\Delta^{(s)}_{\mathcal{Y}}(j, j')\right|^2 b_j b_{j'}\right]^{1/2}. \tag{36}$$

For comparison, define also the corresponding worst-case distortions

$$\varepsilon^{(s)}_{\mathcal{X}} := \left\|d_{\mathcal{X}} - \widetilde{d}^{(s)}_{\mathcal{X}}\right\|_{\infty}, \varepsilon^{(s)}_{\mathcal{Y}} := \left\|d_{\mathcal{Y}} - \widetilde{d}^{(s)}_{\mathcal{Y}}\right\|_{\infty}. \tag{37}$$

Since a and b are probability vectors,

$$\delta^{(s)}_{\mathcal{X}} \leq \varepsilon^{(s)}_{\mathcal{X}}, \delta^{(s)}_{\mathcal{Y}} \leq \varepsilon^{(s)}_{\mathcal{Y}}.$$

Theorem 2 (Stability under tree approximation)**.** *For any fixed collection of S tree slices,*

$$\left|\mathrm{GW}_2(\mathcal{X}, \mathcal{Y}) - \text{T-GROW}_{2,S}(\mathcal{X}, \mathcal{Y})\right| \leq \left[\frac{1}{S}\sum_{s=1}^{S}\left(\delta^{(s)}_{\mathcal{X}} + \delta^{(s)}_{\mathcal{Y}}\right)^2\right]^{1/2}. \tag{38}$$

Consequently,

$$\left|\mathrm{GW}_2(\mathcal{X}, \mathcal{Y}) - \text{T-GROW}_{2,S}(\mathcal{X}, \mathcal{Y})\right| \leq \left[\frac{1}{S}\sum_{s=1}^{S}\left(\varepsilon^{(s)}_{\mathcal{X}} + \varepsilon^{(s)}_{\mathcal{Y}}\right)^2\right]^{1/2}. \tag{39}$$

Proof. For any $\pi \in \Pi(a, b)$, define

$$F(\pi) := \left[\sum_{i,i',j,j'} \left|d_{\mathcal{X}}(x_i, x_{i'}) - d_{\mathcal{Y}}(y_j, y_{j'})\right|^2 \pi_{ij}\pi_{i'j'}\right]^{1/2}, \tag{40}$$

$$F_s(\pi) := \left[\sum_{i,i',j,j'} \left|\widetilde{d}^{(s)}_{\mathcal{X}}(x_i, x_{i'}) - \widetilde{d}^{(s)}_{\mathcal{Y}}(y_j, y_{j'})\right|^2 \pi_{ij}\pi_{i'j'}\right]^{1/2}. \tag{41}$$

Let

$$E_s(i,i',j,j') := \Delta_{\mathcal{X}}^{(s)}(i,i') - \Delta_{\mathcal{Y}}^{(s)}(j,j').$$

By the reverse triangle inequality in the $L^2(\pi \otimes \pi)$ norm,

$$|F(\pi) - F_s(\pi)| \le \left[\sum_{i,i',j,j'} |E_s(i,i',j,j')|^2 \pi_{ij}\pi_{i'j'} \right]^{1/2}. \tag{42}$$

Applying Minkowski's inequality gives

$$\begin{aligned} &\left[\sum_{i,i',j,j'} |E_s(i,i',j,j')|^2 \pi_{ij}\pi_{i'j'} \right]^{1/2} \\ &\le \left[\sum_{i,i',j,j'} \left|\Delta_{\mathcal{X}}^{(s)}(i,i')\right|^2 \pi_{ij}\pi_{i'j'} \right]^{1/2} + \left[\sum_{i,i',j,j'} \left|\Delta_{\mathcal{Y}}^{(s)}(j,j')\right|^2 \pi_{ij}\pi_{i'j'} \right]^{1/2}. \end{aligned} \tag{43}$$

Because π has marginals a and b, the (i,i')-marginal of $\pi \otimes \pi$ is $a \otimes a$ and the (j,j')-marginal is $b \otimes b$. Hence

$$\sum_{i,i',j,j'} \left|\Delta_{\mathcal{X}}^{(s)}(i,i')\right|^2 \pi_{ij}\pi_{i'j'} = \sum_{i,i'} \left|\Delta_{\mathcal{X}}^{(s)}(i,i')\right|^2 a_i a_{i'}, \tag{44}$$

$$\sum_{i,i',j,j'} \left|\Delta_{\mathcal{Y}}^{(s)}(j,j')\right|^2 \pi_{ij}\pi_{i'j'} = \sum_{j,j'} \left|\Delta_{\mathcal{Y}}^{(s)}(j,j')\right|^2 b_j b_{j'}. \tag{45}$$

Therefore,

$$|F(\pi) - F_s(\pi)| \le \delta_{\mathcal{X}}^{(s)} + \delta_{\mathcal{Y}}^{(s)} \tag{46}$$

for every feasible coupling π.

By Theorem 1,

$$d_s(\mathcal{X},\mathcal{Y}) = \inf_{\pi \in \Pi(a,b)} F_s(\pi),$$

whereas

$$\mathrm{GW}_2(\mathcal{X},\mathcal{Y}) = \inf_{\pi \in \Pi(a,b)} F(\pi).$$

Since Eq. 46 holds uniformly over $\Pi(a,b)$, taking infima in both directions yields

$$|\mathrm{GW}_2(\mathcal{X},\mathcal{Y}) - d_s(\mathcal{X},\mathcal{Y})| \le \delta_{\mathcal{X}}^{(s)} + \delta_{\mathcal{Y}}^{(s)}. \tag{47}$$

Finally,

$$\mathrm{GW}_2(\mathcal{X},\mathcal{Y}) = \frac{1}{\sqrt{S}} \left\| (\mathrm{GW}_2(\mathcal{X},\mathcal{Y}))_{s=1}^S \right\|_2$$

and

$$\text{T-GROW}_{2,S}(\mathcal{X},\mathcal{Y}) = \frac{1}{\sqrt{S}} \left\| (d_s(\mathcal{X},\mathcal{Y}))_{s=1}^S \right\|_2.$$

The reverse triangle inequality in ℓ_2^S, together with Eq. 47, gives Eq. 38. The worst-case bound in Eq. 39 follows immediately from $\delta_{\mathcal{X}}^{(s)} \le \varepsilon_{\mathcal{X}}^{(s)}$ and $\delta_{\mathcal{Y}}^{(s)} \le \varepsilon_{\mathcal{Y}}^{(s)}$. □

We next relate the metric distortion to two geometrically interpretable sources of approximation. For each nonempty leaf $C_p^{\mathcal{X},(s)}$, choose a representative $z_p^{\mathcal{X},(s)} \in C_p^{\mathcal{X},(s)}$ and define

$$r_{\mathcal{X},2}^{(s)} := \left[\sum_i a_i d_{\mathcal{X}} \left(x_i, z_{q_{\mathcal{X}}^{(s)}(x_i)}^{\mathcal{X},(s)} \right)^2 \right]^{1/2}. \tag{48}$$

Define the mass-weighted inter-leaf distortion by

$$\eta^{(s)}_{\mathcal{X},2} := \left[\sum_{p,p'} \alpha^{(s)}_p \alpha^{(s)}_{p'} \left| d_{\mathcal{X}}\left(z^{\mathcal{X},(s)}_p, z^{\mathcal{X},(s)}_{p'}\right) - D^{(s)}_{\mathcal{X}}(p,p') \right|^2\right]^{1/2}. \tag{49}$$

The corresponding quantities for $\mathcal{Y}$ are defined analogously.

Proposition 4 (Weighted distortion decomposition)**.** *For every slice s,*

$$\delta^{(s)}_{\mathcal{X}} \le 2r^{(s)}_{\mathcal{X},2} + \eta^{(s)}_{\mathcal{X},2}, \tag{50}$$

and analogously for $\mathcal{Y}$.

Proof. Let $x_i \in C^{\mathcal{X},(s)}_p$ and $x_{i'} \in C^{\mathcal{X},(s)}_{p'}$. By the triangle inequality,

$$\begin{aligned} &\left| d_{\mathcal{X}}(x_i, x_{i'}) - D^{(s)}_{\mathcal{X}}(p,p') \right| \\ &\le d_{\mathcal{X}}\left(x_i, z^{\mathcal{X},(s)}_p\right) + d_{\mathcal{X}}\left(x_{i'}, z^{\mathcal{X},(s)}_{p'}\right) \\ &\quad + \left| d_{\mathcal{X}}\left(z^{\mathcal{X},(s)}_p, z^{\mathcal{X},(s)}_{p'}\right) - D^{(s)}_{\mathcal{X}}(p,p') \right|. \end{aligned} \tag{51}$$

Taking the $L^2(a \otimes a)$ norm of both sides and applying Minkowski's inequality gives

$$\begin{aligned} \delta^{(s)}_{\mathcal{X}} \le &\left[\sum_{i,i'} a_i a_{i'} d_{\mathcal{X}}\left(x_i, z^{\mathcal{X},(s)}_p\right)^2\right]^{1/2} \\ &+ \left[\sum_{i,i'} a_i a_{i'} d_{\mathcal{X}}\left(x_{i'}, z^{\mathcal{X},(s)}_{p'}\right)^2\right]^{1/2} + \eta^{(s)}_{\mathcal{X},2}. \end{aligned} \tag{52}$$

Since $\sum_i a_i = 1$, each of the first two terms equals $r^{(s)}_{\mathcal{X},2}$. Moreover, grouping the final term by leaf clusters converts the weights $a_i a_{i'}$ into $\alpha^{(s)}_p \alpha^{(s)}_{p'}$. Therefore,

$$\delta^{(s)}_{\mathcal{X}} \le 2r^{(s)}_{\mathcal{X},2} + \eta^{(s)}_{\mathcal{X},2}.$$

The proof for $\mathcal{Y}$ is identical. □

For completeness, a corresponding worst-case decomposition follows from the same pointwise argument. Define

$$r^{(s)}_{\mathcal{X},\infty} := \max_p \sup_{x_i \in C^{\mathcal{X},(s)}_p} d_{\mathcal{X}}\left(x_i, z^{\mathcal{X},(s)}_p\right), \tag{53}$$

$$\eta^{(s)}_{\mathcal{X},\infty} := \max_{p,p'} \left| d_{\mathcal{X}}\left(z^{\mathcal{X},(s)}_p, z^{\mathcal{X},(s)}_{p'}\right) - D^{(s)}_{\mathcal{X}}(p,p') \right|. \tag{54}$$

Then $\varepsilon^{(s)}_{\mathcal{X}} \le 2r^{(s)}_{\mathcal{X},\infty} + \eta^{(s)}_{\mathcal{X},\infty}$, and analogously for $\mathcal{Y}$. The two terms have distinct geometric interpretations. The within-leaf quantity $r^{(s)}_{\mathcal{X},2}$ measures the average mass-weighted error induced by representing all samples in a leaf through a single representative, whereas $\eta^{(s)}_{\mathcal{X},2}$ measures the discrepancy between the original geometry of leaf representatives and the tree metric induced by the hierarchy. Their worst-case counterparts provide uniform versions of the same two effects. Neither decomposition implies monotonic improvement under increasing tree depth. Refining a hierarchy can reduce within-leaf compression while simultaneously changing the geometry of the leaf representatives and the induced tree paths. Without additional assumptions on the underlying metric space and on the hierarchical partitioning procedure, no monotone decay of $\eta^{(s)}_{\mathcal{X},2}$ or $\eta^{(s)}_{\mathcal{X},\infty}$ with increasing depth is guaranteed.

A.2 Detailed Complexity Analysis

For the canonical soft-coupling realization of Eq. 13, we solve the reduced unregularized quadratic GW problem using the conditional-gradient method of Titouan et al. (2019). Each outer iteration linearizes the quadratic GW objective at the current coupling and solves the resulting linear optimal-transport subproblem. We denote the number of such outer iterations by I. Since the reduced GW objective is non-convex, the numerical solver provides an approximate solution; the exact quotient result of Theorem 1 concerns the underlying optimization problem and is independent of this numerical choice.

Let $N := \max\{n, m\}$ and $p := \max\{p_X, p_Y\}$, where p_X and p_Y denote the ambient feature dimensions of the two input spaces, and define $L := \max_s \left\{L_{\mathcal{X}}^{(s)}, L_{\mathcal{Y}}^{(s)}\right\} \leq \min\{N, \kappa^H\}$, where L is the maximum realized number of leaves across slices. Here κ is the branching factor, H the maximum tree depth, S the number of sampled hierarchies, and I the number of iterations of the reduced GW solver.

Complexity of the reduced-GW realization. For the direct farthest-point implementation considered in our experiments, each sample is compared with at most κ candidate centers at each level. Keeping the p-dimensional distance evaluations explicit, constructing S randomized hierarchies costs $\mathcal{O}(SH\kappa Np)$.

After the hierarchies are constructed, computing the pairwise co-depth matrices and the corresponding induced tree-distance matrices over at most L leaves costs $\mathcal{O}(SHL^2)$. For leaf spaces of sizes $L_{\mathcal{X}}$ and $L_{\mathcal{Y}}$, one quadratic GW iteration costs $\mathcal{O}\big(L_{\mathcal{X}}^2 L_{\mathcal{Y}} + L_{\mathcal{X}} L_{\mathcal{Y}}^2\big)$, which reduces to $\mathcal{O}(L^3)$ in the balanced case. Each iteration additionally solves a linear OT subproblem with the same leaf resolution. The quadratic GW arithmetic costs $\mathcal{O}(SIL^3)$ in the balanced case, in addition to the linear OT subproblems solved within the conditional-gradient iterations. Thus, the dominant second-order computation remains controlled entirely by the leaf resolution L rather than by the original support size N, here $L \leq \min\{N, \kappa^H\}$. When slices are processed sequentially, the memory requirement is $\mathcal{O}(Np + L^2)$. If point-level quantities are recovered from a leaf-level coupling, this can be done without explicitly materializing the full $n \times m$ coupling, with an additional $\mathcal{O}(N + L^2)$ cost under the aggregation–lifting representation of Theorem 1.

Comparison with FlowAlign. FlowAlign (Le et al., 2021) uses the same high-level clustering-based tree construction. Its reported complexity for this stage follows a different accounting convention from the direct implementation used here. For a consistent comparison, we account for both methods using the same direct farthest-point construction, yielding a shared hierarchy-construction cost of $\mathcal{O}(SH\kappa Np)$. After tree construction, the two methods differ in the representation they compare. Aligned-root FlowAlign operates on one-dimensional root-flow distributions at the original support resolution, with an additional cost of $\mathcal{O}(SN \log N)$. By contrast, T-GROW constructs the pairwise geometry of the leaf spaces and, in its canonical reduced-GW realization, performs the second-order optimization at leaf resolution, with additional cost $\mathcal{O}(SHL^2 + SIL^3)$.

The two post-construction costs depend on different scales. FlowAlign continues to depend on the original support size N, whereas the second-order computation of T-GROW is controlled by the hierarchy resolution L. For fixed (S, κ, H, I, p), we have $L \leq \kappa^H$, so once the tree resolution is fixed, the leaf-level second-order cost does not grow with N.

By Theorem 1, replacing the point-level coupling by the leaf-level coupling introduces no additional optimization gap under the induced tree geometry. Thus, the scalability of T-GROW arises from moving the second-order computation from the original support scale to a controlled leaf scale, rather than from discarding the pairwise relational structure itself.

A.3 Scalability

A.3.1 Large-Scale Synthetic Correspondence

Dataset Construction The large-scale synthetic benchmark is inspired by the scalability experiments of Scetbon et al. (2022), but is extended to include a known cross-space correspondence. This allows us to evaluate not only whether a method remains computationally feasible as the number of samples grows, but also whether the resulting alignment recovers meaningful structure. In particu-

lar, root-to-point distances provide only first-order information about a hierarchy: different pairwise organizations can induce the same collection of root depths. We construct a family in which these first-order statistics are deliberately non-discriminative, while the joint hierarchical organization remains informative. The benchmark is viewed as a controlled realization of the structural ambiguity identified by the theory.

We construct a balanced hierarchical tree with branching factor $\kappa = 4$ and depth $H = 3$, resulting in $L = \kappa^H = 64$ leaf clusters. Each leaf cluster is represented by a compact local point distribution, while the separation between cluster centers decreases with hierarchical depth, so that leaves sharing a deeper common ancestor are progressively closer in the latent hierarchy. Two point clouds, $X \subset \mathbb{R}^3$ and $Y \subset \mathbb{R}^6$, are then independently generated from the same underlying hierarchy. They share the same latent leaf identities and hierarchical mass structure, while their samples, embedding directions, rotations, and perturbations are generated independently.

The geometry below the top level is constructed to be homogeneous. Consequently, root-to-leaf depth carries only coarse information about the top-level branch and does not uniquely identify the 64 leaves. The finer correspondence is instead encoded in how probability mass is organized jointly across the hierarchy.

Evaluation Protocol Each sampled point is associated with the latent leaf from which it was generated. Because X and Y are sampled from the same abstract hierarchy corresponding leaves share the same latent identity, providing ground-truth correspondence at the cluster level. Given the two unlabeled point clouds, each method estimates the coupling between the source cloud X and the target cloud Y using only their observed within-domain structure. We transfer the source leaf labels to the target through the estimated coupling. For each target sample, class scores are computed directly from the slice-specific leaf couplings and averaged across randomized hierarchies according to Eq. 32, and the transferred label is the class with the largest aggregated score. The dense point-level coupling is never materialized. Label-transfer accuracy is the fraction of target samples whose transfer label agrees with their known latent leaf identity.

We evaluate $n \in \{100\text{K}, 500\text{K}, 1\text{M}, 5\text{M}, 10\text{M}, 50\text{M}\}$ points per cloud. Each method uses a single fixed configuration across all sample sizes, with no scale-specific tuning. For T-GROW, we use randomized hierarchies $\kappa = 4$, $H = 3$ and $S = 8$. FlowAlign and DepthAlign use the same tree realizations so that differences primarily reflect the alignment objective rather than tree construction. QGW uses 64 representatives, SGW uses 32 projections, and LR-GW uses rank 10 with 100 outer iterations.

Method	Accuracy ↑ Runtime (s) ↓					
	100K		500K		1M	
	Acc.	Time	Acc.	Time	Acc.	Time
SGW	0.010 ±0.004	1.57 ±0.09	0.017 ±0.012	8.01 ±0.92	0.034 ±0.015	16.78 ±0.74
QGW	0.073 ±0.020	0.76 ±0.03	0.075 ±0.065	3.96 ±0.78	0.110 ±0.126	8.96 ±0.22
DepthAlign	0.124 ±0.090	1.35 ±0.12	0.208 ±0.072	7.63 ±0.23	0.198 ±0.078	16.72 ±0.14
FlowAlign	0.129 ±0.053	1.32 ±0.12	0.143 ±0.041	7.62 ±0.23	0.139 ±0.031	16.67 ±0.24
Ours	**0.530** ±0.106	1.48 ±0.11	**0.686** ±0.129	8.16 ±0.23	**0.802** ±0.157	16.90 ±0.29

Method	Accuracy ↑ Runtime (s) ↓					
	5M		10M		50M	
	Acc.	Time	Acc.	Time	Acc.	Time
SGW	0.017 ±0.026	93.07 ±29.94	0.019 ±0.010	217.78 ±78.81	–	–
QGW	0.082 ±0.043	33.72 ±13.83	0.127 ±0.104	67.52 ±30.61	0.087 ±0.037	246.00 ±3.92
DepthAlign	0.088 ±0.094	79.11 ±23.65	0.349 ±0.228	165.92 ±63.25	0.229 ±0.036	744.36 ±18.67
FlowAlign	0.204 ±0.154	78.86 ±23.83	0.183 ±0.064	166.32 ±61.82	0.224 ±0.112	746.25 ±19.46
Ours	**0.859** ±0.158	79.99 ±21.41	**0.901** ±0.121	167.02 ±63.61	**0.958** ±0.072	746.13 ±18.02

Table 5: Large-scale synthetic correspondence results. We report label-transfer accuracy and runtime as mean ± std over three independent trials; "–" denotes runs exceeding the 3600s time upper bound or memory budget.

As shown in Table 5, Ours consistently achieves the highest correspondence accuracy across all evaluated scales while maintaining computational scaling comparable to the tree-based baselines. Its accuracy improves as n increases, as larger sample sizes produce more reliable estimates of the underlying hierarchical structure and reduce sampling noise in the induced correspondence. In contrast, the competing methods show much weaker gains from additional samples. The results demonstrate that our method can leverage increasing sample sizes to recover hierarchical correspondences more reliably while remaining scalable to tens of millions of points.

A.4 Robustness

A.4.1 Synthetic Dataset

Experimental setup We construct a controlled synthetic matching task with known source-target correspondences. Each matching instance contains eight source point-cloud distributions and eight corresponding target distributions. Every distribution follows the same broad two-level organization: four coarse branches, each containing multiple local geometric components. Different distributions share this hierarchical template but differ in the arrangement of their local geometric structure. This design creates explicit geometry at both within-branch and between-branch scales and therefore allows the two scales to be perturbed separately.

For a given perturbation level, the target distributions are generated from their corresponding sources and then randomly reordered. At $\rho = 0$, each target is therefore an unperturbed counterpart of its matched source up to permutation, providing a clean-reference condition. Approximate methods may nevertheless fall below perfect recovery because of their own representation or optimization constraints.

For each method, we compute a scalar discrepancy for every source-target pair, producing an 8×8 cost matrix C. We recover a one-to-one distribution-level matching by solving $\hat{\pi} = \arg\min_\pi \sum_{i=1}^{8} C_{i,\pi(i)}$, and report the fraction of correctly recovered source-target pairs.

We use n to denote the number of sampled points in each distribution and ρ to denote the perturbation strength. The main experiment uses $n = 128$ and 15 independent trials. All method settings are fixed across perturbation levels, and no method is retuned for individual values of ρ.

We introduce two complementary perturbations that act at different structural scales. All synthetic point clouds in this experiment lie in $\mathbb{R}^2$, so a planar rotation is specified by a single angle, $R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$

Local perturbation. Within each coarse branch, points are separated into two radial subgroups according to their distance from the branch center p_a. Each subgroup is rotated independently around the same branch center:

$$x_i' = R(\theta_{a,q})(x_i - p_a) + p_a, \theta_{a,q} = 1.9\rho z_{a,q}, z_{a,q} \sim \mathcal{N}(0,1).$$

This perturbation mainly modifies local within-branch geometry while leaving the coarse branch locations unchanged before the final jitter step. We evaluate $\rho \in \{0, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90\}$.

Global perturbation. All points belonging to the same coarse branch are rigidly rotated together around the global origin:

$$x_i' = R(\theta_a)x_i, \theta_a = 1.9\rho z_a, z_a \sim \mathcal{N}(0,1).$$

The rigid transformation preserves the internal geometry of each branch while changing the relative arrangement between branches. We evaluate $\rho \in \{0, 0.03, 0.06, 0.10, 0.14, 0.17, 0.20\}$. A smaller numerical range is used for the global perturbation because branch-level rotations induce stronger changes in the overall geometry. After either structured rotation, we add a small isotropic coordinate perturbation

$$\epsilon_i \sim \mathcal{N}(0, \sigma_\rho^2 I), \sigma_\rho = (0.010 + 0.025\rho)\rho,$$

followed by the same centering and scale normalization.

Full results analysis Tables 6 and 7 report the complete numerical results underlying the two robustness curves in the main text. Accuracy is reported as mean $\pm$ std over 15 independent trials.

Table 6: Local-perturbation distribution-level matching accuracy at $n = 128$. Runtime is reported as mean $\pm$ std (s) for constructing one complete 8×8 discrepancy matrix over 15 runs at $\rho = 0.15$.

Method	$\rho = 0.15$	0.30	0.45	0.60	0.75	0.90	Runtime
LR-GW	18.3±4.0	15.0±3.9	22.5±4.9	8.3±2.0	15.8±3.1	10.8±3.0	0.32±0.01
QGW	51.7±5.6	25.0±4.7	15.0±4.3	16.7±3.4	22.5±3.3	10.8±3.0	2.50±0.09
GW	65.0±5.3	35.8±4.5	21.7±4.5	26.7±5.4	17.5±2.7	19.2±4.0	0.85±0.02
FlowAlign	74.2±5.1	56.7±6.8	52.5±6.5	30.8±4.2	27.5±5.3	24.2±4.6	0.75±0.03
EGW	79.2±4.5	59.2±7.1	51.7±6.2	35.0±4.3	34.2±6.2	39.2±5.2	8.87±1.47
SGW	85.8±4.4	59.2±6.8	64.2±4.4	50.0±6.3	45.0±4.8	50.0±5.0	0.13±0.01
DepthAlign	95.0±2.7	95.0±3.6	83.3±4.8	70.8±4.8	54.2±8.3	43.3±6.3	0.16±0.01
Ours	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	**95.0±2.7**	**83.3±4.0**	**59.2±6.2**	1.54±0.07

Table 7: Global-perturbation distribution-level matching accuracy at $n = 128$. The runtime definition is identical to Table 6 and is reported as mean $\pm$ std over 15 runs at $\rho = 0.03$.

Method	$\rho = 0.03$	0.06	0.10	0.14	0.17	0.20	Runtime
QGW	45.0±6.0	28.3±4.5	23.3±4.2	22.5±5.1	16.7±5.1	13.3±4.6	2.50±0.08
GW	47.5±6.1	30.0±4.4	18.3±3.2	20.8±3.4	21.7±4.6	15.0±4.1	0.84±0.02
DepthAlign	55.0±4.4	42.5±3.2	33.3±4.8	26.7±4.4	25.0±4.6	27.5±5.3	0.15±0.01
LR-GW	65.8±5.1	45.0±5.7	22.5±4.4	19.2±3.0	21.7±5.2	18.3±5.6	0.32±0.01
FlowAlign	68.3±8.0	28.3±4.8	18.3±4.5	18.3±4.5	15.8±3.9	15.0±4.3	0.74±0.03
EGW	70.0±5.1	38.3±7.5	21.7±4.6	22.5±3.9	16.7±3.4	18.3±4.4	9.49±1.95
SGW	84.2±4.1	55.0±6.7	37.5±5.3	29.2±4.7	21.7±3.8	20.8±3.6	0.12±0.01
Ours	**100.0±0.0**	**93.3±3.0**	**75.8±5.7**	**58.3±4.2**	**51.7±4.9**	**45.0±4.5**	1.52±0.04

Runtime is measured as the wall-clock time required to construct the complete 8×8 source–target discrepancy matrix, while the final linear assignment is common to all methods and has negligible cost in comparison. Since varying the perturbation strength changes the geometry but not the problem size, runtime is primarily determined by n and the solver. For the two $n = 128$ settings, the reported runtimes remain similar for most methods, indicating that perturbation strength has only a limited effect on computational cost.

Across both perturbation axes, the proposed discrepancy degrades more gradually as ρ increases. The local perturbation stresses fine-scale within-branch geometry, while the global perturbation changes the between-branch organization. The consistent advantage across both settings is consistent with the use of multilevel relational information rather than dependence on a single preserved geometric scale.

A.4.2 Real-world Task: Blind Vision-Language Matching

Task formulation. We follow the blind vision–language matching setting introduced by Schnaus et al. (2025). Given N semantic classes, let $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_i\}_{i=1}^N$ denote their class-level vision and language representations, respectively. The cross-modal correspondence is withheld, and the goal is to recover it using only pairwise relations within the two modalities.

Class-level representations are obtained by averaging the embeddings associated with each semantic class. We then construct the within-modality distance matrices

$$\mathbf{X}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2, \qquad \mathbf{Y}_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|_2. \tag{55}$$

The blind correspondence problem can be written as the quadratic assignment problem

$$\pi^\star = \arg\min_{\pi \in \mathfrak{S}_N} \sum_{i,j=1}^{N} \left(\mathbf{X}_{ij} - \mathbf{Y}_{\pi(i)\pi(j)}\right)^2. \tag{56}$$

Matching accuracy is the fraction of semantic classes whose correspondence is correctly recovered. We use CIFAR-100 with CLIP (ViT-L/14@336) as the vision encoder and all-mpnet-base-v2 as the language encoder. Image embeddings belonging to the same semantic class are averaged to obtain a class-level visual representation, while the available language prompts are averaged to obtain the corresponding language representation.

We adapt the T-GROW construction in Algorithm 2 to correspondence recovery. The same randomized hierarchical construction and tree-induced relational representations are retained. However, instead of using the resulting structural comparison only to obtain a scalar discrepancy, the matching implementation retains the correspondence information and returns an estimated permutation $\hat{\pi}$ over the semantic classes. No ground-truth class correspondence or cross-modal cost matrix is provided during matching. The known class identities are used only after optimization to evaluate the recovered permutation by matching accuracy.

Each randomized slice returns a candidate permutation π_s together with its matching objective c_s. The slice-wise permutations are combined by the cost-weighted vote

$$V_{ij} = \sum_{s=1}^{S} w_s \, \mathbf{1}\{\pi_s(i) = j\}, \qquad w_s = \exp\left(-\frac{c_s - \min_{s'} c_{s'}}{\mathrm{std}(c)}\right). \tag{57}$$

The consensus permutation is obtained from

$$\pi_{\mathrm{cons}} = \arg\max_{\pi \in S_N} \sum_i V_{i,\pi(i)}, \tag{58}$$

using the Hungarian algorithm. The final correspondence is chosen as the better of π_{cons} and the lowest-cost individual slice under the matching objective.

Label Corruption To evaluate robustness to imperfect visual representations, we perturb image labels before computing the class-level vision embeddings. Let n_c denote the number of images assigned to class c. For a corruption ratio ρ, we randomly select

$$m_c = \lfloor \rho n_c \rfloor \tag{59}$$

images from each class. The selected images from all classes are pooled and their class assignments are randomly permuted within this pool. Consequently, the number of images assigned to each class is preserved exactly, while an increasing fraction of each class representation is contributed by images originating from other classes. The corruption is independently sampled for each random seed. The corrupted labels are applied before class-level averaging; all subsequent processing, including centering, distance construction, and correspondence recovery, remains unchanged. Varying ρ changes only the quality of the visual representation rather than the underlying matching instance.

We evaluate $\rho \in \{0, 0.1, \ldots, 0.5\}$ for $N = 100$. The setting $\rho = 0$ recovers the clean cross-modal matching problem and serves as an internal consistency check. Unlike the synthetic experiment, this does not imply identical source and target geometries, the vision and language modalities already exhibit intrinsic structural mismatch before any additional corruption is introduced.

Table 8 reports the complete results for $N = 100$. As the corruption ratio increases, the available cross-modal structural signal generally weakens, as reflected by the decreasing kernel correlation. T-GROW achieves the highest mean matching accuracy at every evaluated corruption level, from 0.656 in the clean setting to 0.396 at $\rho = 0.5$. These results indicate that T-GROW better preserves useful correspondence information as the vision-side relational structure is progressively degraded.

To quantify how the corruption changes the correspondence problem itself, we measure the remaining cross-modal structural signal using the Pearson correlation between the off-diagonal entries of the visual and language distance matrices

$$r(\rho) = \mathrm{Corr}\left(\mathrm{vec}(\mathbf{X}_{\mathrm{off}}), \mathrm{vec}(\mathbf{Y}_{\mathrm{off}})\right). \tag{60}$$

As shown in Fig. 3, the cross-modal correlation gradually decreases as ρ increases, indicating a progressive loss of shared pairwise structure between the two modalities. Notably, beyond approximately $\rho = 0.6$, the correlation drops into a substantially weaker regime, suggesting that much of the recoverable cross-modal structural signal has been destroyed. This also explains the sharp deterioration of matching performance at high corruption levels, the correspondence itself becomes increasingly difficult to identify from the available within-modality geometry. In contrast, at low and moderate corruption levels, sufficient structural signal remains, allowing differences between matching methods to be meaningfully assessed.

Table 8: Robustness to vision-side label corruption on CIFAR-100 at $N = 100$.Matching accuracy is reported as mean $\pm$ s.d. over five seeds. A fraction ρ of each class's images is reassigned across classes before class-level averaging, while class sizes are preserved exactly.

Low corruption			
Method	$\rho = 0$	$\rho = 0.1$	$\rho = 0.2$
SGW	0.014 ± 0.017	0.032 ± 0.033	0.038 ± 0.027
FlowAlign	0.032 ± 0.030	0.044 ± 0.051	0.072 ± 0.055
DepthAlign	0.026 ± 0.019	0.144 ± 0.239	0.070 ± 0.047
LR-GW	0.178 ± 0.068	0.180 ± 0.062	0.142 ± 0.073
QGW	0.294 ± 0.091	0.104 ± 0.088	0.180 ± 0.140
EGW	0.250 ± 0.111	0.366 ± 0.085	0.298 ± 0.108
GW	0.300 ± 0.146	0.188 ± 0.076	0.242 ± 0.140
T-GROW	$\mathbf{0.656 \pm 0.096}$	$\mathbf{0.422 \pm 0.141}$	$\mathbf{0.442 \pm 0.155}$
chance $(1/N)$	0.010	0.010	0.010

Moderate corruption			
Method	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$
SGW	0.028 ± 0.019	0.032 ± 0.032	0.044 ± 0.048
FlowAlign	0.052 ± 0.035	0.024 ± 0.008	0.088 ± 0.049
DepthAlign	0.030 ± 0.020	0.068 ± 0.044	0.070 ± 0.043
LR-GW	0.110 ± 0.072	0.112 ± 0.081	0.148 ± 0.109
QGW	0.224 ± 0.127	0.130 ± 0.095	0.170 ± 0.088
EGW	0.276 ± 0.085	0.226 ± 0.125	0.192 ± 0.096
GW	0.120 ± 0.065	0.084 ± 0.030	0.026 ± 0.021
T-GROW	$\mathbf{0.564 \pm 0.052}$	$\mathbf{0.370 \pm 0.111}$	$\mathbf{0.396 \pm 0.132}$
chance $(1/N)$	0.010	0.010	0.010

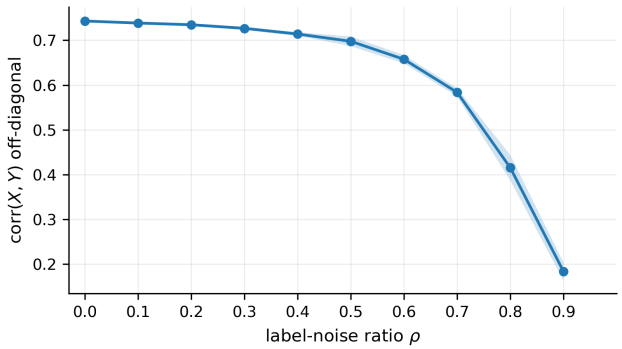


Figure 3: Cross-modal structural signal under label corruption. Pearson correlation between the off-diagonal entries of the visual and language distance matrices as the corruption ratio ρ increases.

A.5 Applications on Real-world Datasets

A.5.1 Cross-modal Alignment: Single-Cell Multi-Omics

We follow the SNARE-seq experiment of Scetbon et al. (2022). The dataset contains $n = m = 1047$ paired cells represented in 10- and 19-dimensional feature spaces after preprocessing. The known cell-wise correspondence is used only for evaluation. Following the reference protocol, a k-nearest-neighbor graph with $k = 110$ is independently constructed in each modality using correlation-based neighborhoods, and shortest-path distances are used to characterize the corresponding intra-modal geometry.

We evaluate alignment quality using the "fraction of samples closer than the true match" (FOSCTTM) error and label accuracy, together with runtime. Let x_i and y_i denote the true paired cells in the two modalities, and let d_{ij} denote the distance between x_i and the aligned representation of y_j. For the $X \to Y$ direction, we compute

$$\mathrm{FOSCTTM}_{X\to Y} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j\neq i}\left[\mathbb{I}(d_{ij} < d_{ii}) + \frac{1}{2}\mathbb{I}(d_{ij} = d_{ii})\right]}{m-1}, \tag{61}$$

and analogously for $Y \to X$. The reported FOSCTTM is the average of the two directions. Thus, FOSCTTM measures the fraction of samples in the opposite modality that are ranked closer to a query cell than its true counterpart, with lower values indicating better alignment. Stochastic methods are evaluated over five random seeds.

We further examine the sensitivity of LR-GW to its rank parameter and to k-means initialization. Using the official implementation and the SCOT kNN-graph geodesic geometry (Demetci et al., 2022) described above, we evaluate $r \in \{5, 10, 20, 50\}$ with five independent k-means initializations for each rank. In the main comparison, we report the mean performance at $r = 5$ over these five independent runs: 0.2260 ± 0.1038 FOSCTTM, 0.7349 ± 0.2588 label accuracy, and 103.55 ± 14.61 seconds runtime. The complete rank and initialization sensitivity results are summarized in Table 9.

Table 9: Sensitivity of LR-GW to rank and initialization on SNARE-seq. Single-run results are mean $\pm$ std over five k-means initializations. For multi-start, the solution with the lowest LR-GW objective among five initializations is retained.

Rank	Single-run FOSCTTM ↓	Multi-start FOSCTTM ↓	Multi-start time (s)
$r = 5$	0.226 ± 0.104	0.157	518
$r = 10$	0.398 ± 0.180	0.156	595
$r = 20$	0.309 ± 0.145	0.154	740
$r = 50$	0.252 ± 0.224	0.155	1313

LR-GW exhibits substantial sensitivity to k-means initialization on SNARE-seq. Favorable initializations can achieve FOSCTTM values of approximately 0.15, whereas less favorable runs can produce values around 0.4–0.65, resulting in substantial variation across single runs. The multi-start results reduce this variability by selecting, for each rank, the solution with the lowest LR-GW objective among five initializations. Across the tested ranks, multi-start LR-GW achieves FOSCTTM values between 0.153 and 0.157, consistent with the approximately 0.16 performance reported by Scetbon et al. (2022). The higher runtime of the multi-start setting results from performing five complete LR-GW optimizations rather than using a single initialization.

A.5.2 Semantic Segmentation Transfer on S3DIS

Data and evaluation protocol. We use the Stanford 3D Indoor Scene Dataset (S3DIS) (Armeni et al., 2016). Following the segmentation-transfer setting of QGW (Chowdhury et al., 2021), we consider a labeled source point cloud $X = \{x_i\}_{i=1}^{n_s}$ and an unlabeled target point cloud $Y = \{y_j\}_{j=1}^{n_t}$, with source labels $\ell_i^s \in \{1, \ldots, C\}$. We evaluate two full-resolution transfer pairs, the Lobby pair released with QGW, containing $1\,155\,072$ source points and $909\,312$ target points, and Area 5 Lobby 1 $\to$ Area 4 Lobby 2, containing approximately 1.16M source points and 2.05M target points. The same experimental configuration is used for both pairs without pair-specific re-tuning.

All methods receive the complete source and target point clouds, and every target point is retained for label transfer and evaluation. At this scale, correspondence is estimated through a reduced representation and then propagated back to the original points. For T-GROW, the randomized tree encoder is applied to the 3D coordinates, and the resulting leaf-level representation is used for correspondence estimation. QGW uses its random-anchor partition, while GW, EGW, and LR-GW are evaluated on the same anchor-based reduced representation with the same support size. FlowAlign and DepthAlign use their respective hierarchical representations and objectives, and SGW operates through one-dimensional projections. For cell-based methods, the reduced support size is approximately $m = 10^3$.

Label transfer. After correspondence estimation, all methods are evaluated using the same semantic label-transfer rule. Let $\Pi \in \mathbb{R}_+^{n_s \times n_t}$ denote the point-level correspondence induced by the method. The predicted label of target point y_j is

$$\hat{\ell}_j^t = \arg \max_{c \in \{1,\dots,C\}} \sum_{i=1}^{n_s} \Pi_{ij} \, \mathbf{1}[\ell_i^s = c] \,. \tag{62}$$

For methods that estimate correspondence on a reduced representation, the corresponding class masses are propagated back through the source and target assignments to the original points. For methods that return a hard correspondence, Π denotes the induced point-level matching. When several correspondences are produced, as in projection- or slice-based methods, the corresponding class scores are aggregated before the final $\arg\max$. For T-GROW, the slice-specific class scores are uniformly averaged according to Eq. 32. The same score-level averaging convention is used for the corresponding multi-projection baselines. In implementation, only the class-wise transported masses required for prediction are accumulated and the full $n_s \times n_t$ coupling therefore need not be explicitly materialized.

We report point-wise accuracy and mean intersection-over-union (mIoU) over all target points, with mIoU averaged over the semantic classes present in the target ground truth. Runtime is the wall-clock time from the two input point clouds to predicted labels for all target points, excluding data loading and metric evaluation. All methods are timed on the same machine. Stochastic methods are evaluated over three random seeds, and results are reported as mean $\pm$ standard deviation.

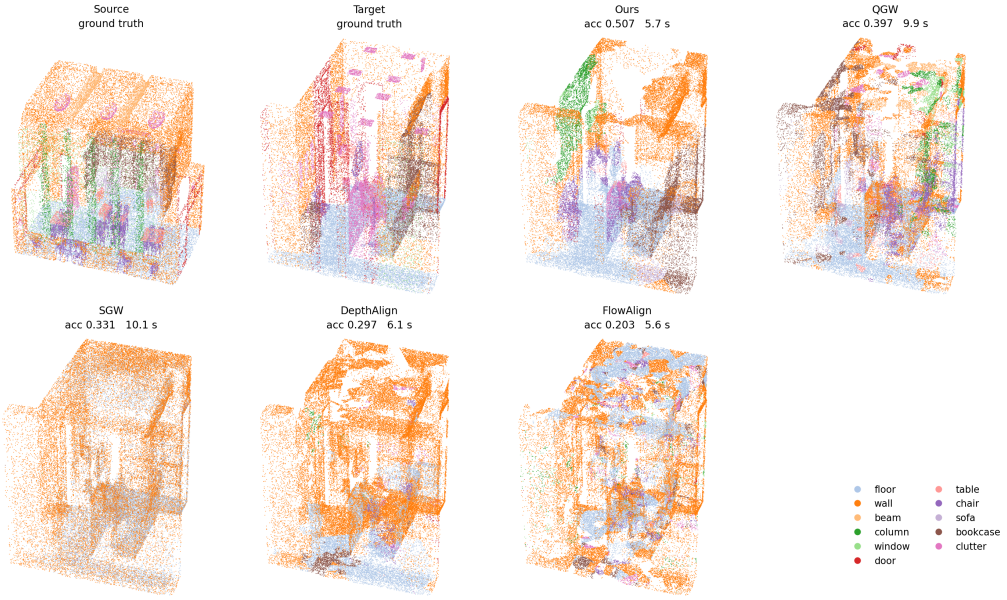


Figure 4: Large-scale semantic segmentation transfer on the released S3DIS Lobby pair ($1\,155\,072 \to 909\,312$ points). The first two panels show the source and target ground truth, followed by the semantic labels transferred by the competing methods. The visualization corresponds to one representative run.

Figure 4 provides a qualitative comparison on the released Lobby pair. The visualization complements the three-seed quantitative results reported in Table 1.

A.6 Ablation and Sensitivity Analysis

We conduct the ablation study on the blind vision–language matching benchmark, using the same experimental protocol as the corresponding main-text results, with $N = 100$ classes and five random seeds. The goal is to examine how the performance of T-GROW depends on its key structural choices and hierarchy parameters.

Proposition 2 shows that root-depth information alone does not determine the full tree geometry. The diagonal of the co-depth matrix records root depths, while its off-diagonal entries capture shared-ancestry information between pairs of leaves. We examine whether this additional relational information contributes to the empirical performance.

We compare the full method with a unary-only variant, in which the pairwise hierarchical representation used for matching is replaced by an additive representation based solely on the root depth of each node. This representation removes explicit pairwise interactions while keeping the hierarchy construction, randomization, and experimental protocol unchanged. In CIFAR-100 blind matching with $N = 100$, the full method achieves 0.656 ± 0.096 precision, while the unary variant reaches 0.314 ± 0.191. This substantial drop is consistent with Proposition 2. Unary hierarchical statistics alone do not retain the shared-ancestry information encoded by the full pairwise hierarchical representation.

A.6.1 Sensitivity to Hierarchy Resolution

The resolution of the hierarchy determines the scale at which the original support is represented. For a balanced hierarchy, the nominal number of leaves is $L = \kappa^H$. we span a broad range of resolutions from $L = 8$ to $L = 256$ to characterize the accuracy–runtime trade-off as the hierarchy becomes finer. To isolate the effect of the hierarchy resolution, we fix the number of randomized hierarchies at $S = 4$ throughout this study and keep all remaining experimental settings unchanged. Each configuration is evaluated over five random seeds.

Table 10: Sensitivity to the leaf resolution $L = \kappa^H$ on CIFAR-100 blind matching with $N = 100$, $S = 4$, and five random seeds (mean $\pm$ std).

κ	H	L	Accuracy	Runime (s)
2	3	8	0.584 ± 0.246	2.2
3	2	9	0.432 ± 0.335	1.3
4	2	16	0.358 ± 0.305	1.9
5	2	25	0.564 ± 0.282	2.0
3	3	27	0.458 ± 0.342	1.5
2	5	32	0.288 ± 0.257	2.4
6	2	36	0.404 ± 0.274	1.5
4	3	64	0.540 ± 0.241	1.8
3	4	81	0.656 ± 0.096	2.3
10	2	100	0.606 ± 0.127	2.9
5	3	125	0.462 ± 0.313	4.1
6	3	216	0.658 ± 0.063	5.5
4	4	256	0.572 ± 0.213	7.9

Table 10 shows that coarse hierarchies can lose substantial relational information, but increasing L beyond a moderate resolution does not consistently improve performance. In particular, $\kappa = 3$ and $H = 4$, corresponding to $L = 81$, achieves 0.656 accuracy in 2.1 seconds. Increasing the resolution to $L = 216$ produces a comparable mean accuracy of 0.658, but more than doubles the runtime to 5.5 seconds. The runtime results expose the complementary computational trade-off. The leaf-level second-order computation becomes increasingly expensive as the hierarchy resolution grows, consistent with the dependence on L in Appendix A.2. Consequently, excessively fine hierarchies can incur substantially greater computational cost without a corresponding gain in matching accuracy.

A.6.2 Sensitivity to the Number of Randomized Hierarchies

A single finite-depth hierarchy depends on a particular randomized partition of the input support. Equation 14 aggregates multiple randomized hierarchies to reduce the sensitivity to any individual partition. We isolate the effect of the number of randomized hierarchies. The hierarchy resolution is fixed at the selected operating point $\kappa = 3$ and $H = 4$, corresponding to $L = 81$, and only the number of sampled hierarchies S is varied.

Table 11: Sensitivity to the number of randomized hierarchies S on CIFAR-100 blind matching with the hierarchy resolution fixed at $\kappa = 3$, $H = 4$ ($L = 81$). All other settings are unchanged. Results are reported over five random seeds as mean $\pm$ standard deviation.

S	Accuracy	Runtime (s)
1	0.554 ± 0.153	1.8
2	0.552 ± 0.250	1.9
4	0.656 ± 0.096	2.3
8	0.648 ± 0.102	2.6
16	0.702 ± 0.026	4.0
32	0.658 ± 0.080	6.3
48	0.668 ± 0.052	14.4
64	0.658 ± 0.088	16.5
96	0.636 ± 0.107	28.0
128	0.710 ± 0.097	31.2
192	0.642 ± 0.101	54.1
256	0.668 ± 0.083	61.2

Table 11 shows that using only one or two randomized hierarchies leads to substantially greater variability. At $S = 1$ and $S = 2$, the accuracies are 0.554 and 0.552, respectively, indicating strong dependence on the particular randomized partition that is sampled. Once several hierarchies are aggregated, the performance becomes considerably more stable. From $S = 4$ onward, the mean accuracy remains within a relatively narrow range, from 0.636 to 0.710, despite increasing S by a factor of 64. Overall, aggregating several randomized hierarchies mainly improves robustness over the unstable single-hierarchy regime, rather than yielding monotonic gains as S increases. Runtime increases steadily with S, from 2.1 seconds at $S = 4$ to 61.2 seconds at $S = 256$, consistent with the complexity analysis in Appendix A.2. Thus a larger S beyond the plateau mainly increases computational cost without a consistent improvement in accuracy.

In practice, κ and H should therefore be chosen to achieve a moderate leaf resolution that captures sufficient hierarchical detail without unnecessarily increasing the leaf-level GW cost.