NYSTRÖM LANDMARKS FOR CROSS-GRAPH MERGING WITH OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Many multimodal prediction problems naturally give rise to *multiple graphs*—one per modality—yet learning is often performed on each modality graph separately, with cross-modal interaction occurring only indirectly. We study unsupervised structural graph merging for multimodal prediction: given modality-specific graphs associated with a supervised target, we construct a *sparse* set of cross-graph edges so that information can propagate directly across modalities during message passing. Our approach first learns *commensurate* node embeddings via self-supervision. It then constructs Nyström-style node-to-landmark fingerprints and an N-way low-rank landmark tensor that produces structured cross-modal context embeddings. Finally, for each modality pair, we instantiate cross-graph edges by solving an Earth Mover's Distance (optimal transport) problem between node distributions in this embedding space and extracting the support of the optimal coupling. The edge-construction procedure does not use downstream labels. We target the common setting of a small, fixed collection of complementary modalities and evaluate two- and three-modality configurations. Across heterogeneous multimodal graph classification settings, the resulting merged graphs can improve downstream performance, motivating sparse cross-graph structure as a first-class component of multimodal modeling.

1 INTRODUCTION

Modern prediction tasks often involve multiple complementary modalities—such as molecular, clinical, and histology data in precision medicine or radar and multispectral imagery in remote sensing—which can naturally be represented as separate graphs. However, most multimodal graph learning still processes these modality graphs in isolation, relying on representation fusion or alignment rather than adding explicit cross-graph structure, so cross-modal information only interacts indirectly instead of through direct message passing across graphs.

Problem. We address the following question: *Given a set of modality-specific graphs associated with the same supervised target, how can we construct cross-graph connectivity so that the resulting merged graph improves performance on the downstream task?*

From alignment to structural merging. Graph alignment methods estimate correspondences between node sets based on structural or embedding similarity and are often used as a precursor to transferring information across graphs. However, the output is commonly a matching or a soft correspondence rather than a graph that supports cross-modal message passing. In contrast, **graph merging** as we use the term constructs a sparse computation graph: the goal is not to recover a correspondence per se, but to create cross-graph links that enable message passing. We evaluate these label-free links by their downstream predictive utility, rather than by correspondence recovery alone.

Key challenge. Cross-modal edge construction is difficult because modality graphs can differ in node semantics, feature statistics, and topology, and ground-truth correspondences are rarely available. Naïve nearest-neighbor linking in a shared embedding space is brittle (e.g., hubness and many-to-one matches) and typically requires thresholding or a fixed k that may not transfer across modalities (Radovanovic et al., 2010).

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.

Approach overview. We propose a three-stage pipeline for unsupervised structural merging. (1) We learn commensurate node embeddings across modalities via self-supervision. (2) We build Nyström-like (Williams & Seeger, 2000) fingerprints using landmark nodes in each graph and combine them through an N-way low-rank landmark tensor to obtain cross-modal context embeddings. This construction incorporates context from all available modalities in a single representation stage, rather than requiring a separate representation design for each modality pair. (3) For each modality pair, we compute an optimal-transport coupling between the two node sets in this embedding space and extract cross edges from the support of the coupling. The transport constraints enforce global mass conservation, yielding balanced transport allocation and sparse cross-graph connectivity without ad hoc thresholding.

Beyond pairwise merging. Our method uses a single N-graph formulation for the small, fixed modality collections typical of multimodal applications. The N-way landmark tensor integrates information across all available graphs before pairwise OT couplings induce sparse cross-graph connectivity.

Contributions.

- We formulate **unsupervised structural graph merging**: a label-free procedure that constructs sparse cross-graph edges between modality graphs and is evaluated by downstream predictive utility.
- We introduce a practical **multi-graph** graph-construction pipeline for small, fixed modality collections that combines **Nyström landmark fingerprints** with a **low-rank N-way landmark tensor** to represent multiway cross-modal compatibility.
- We use **optimal-transport-based** edge induction to convert that compatibility into a sparse computation graph: cross edges are defined as the support of EMD couplings, providing balanced connectivity and an implicit sparsity control per modality pair.
- Starting from modality-specific input graphs, we evaluate whether this cross-graph topology construction procedure improves graph classification relative to alignment-derived edge construction across heterogeneous multimodal settings.

2 Related Work

Our work connects multimodal graph learning with network alignment and optimal transport (OT). Fusion methods combine modality representations or predictions; alignment methods such as REGAL (Heimann et al., 2018), CONE-Align (Chen et al., 2020), GradAlign+ (Park et al., 2022), FUGAL (Bommakanti et al., 2024), and CAPER (Zhu et al., 2022) estimate cross-graph correspondences using structural and/or attribute cues; and OT-based approaches provide globally balanced soft matchings under uncertainty (e.g., Zeng et al. (2023)). Rather than claiming a new embedding, landmark, or OT solver, we combine these components to construct a sparse cross-graph computation graph for message passing, then evaluate its predictive utility. Extended discussion and additional perspectives are provided in Section G.

3 Graph Merging

3.1 Algorithm

Problem setup. Let $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ be a set of N graphs, where

$$\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i).$$

Here, $\mathcal{V}_i$ denotes the node set of graph i, and $\mathcal{E}_i$ its (intra-graph) edge set. We write $\mathcal{V}_{ij}$ for node j in graph i, and $\mathcal{E}_{ijkl}$ for a *cross-graph* edge from node $\mathcal{V}_{ij}$ to node $\mathcal{V}_{kl}$ (with $i \neq k$). Each graph (or graph set) is associated with a target signal y for a downstream task. Our goal is to select cross-graph edges that improve downstream performance on the merged graph.

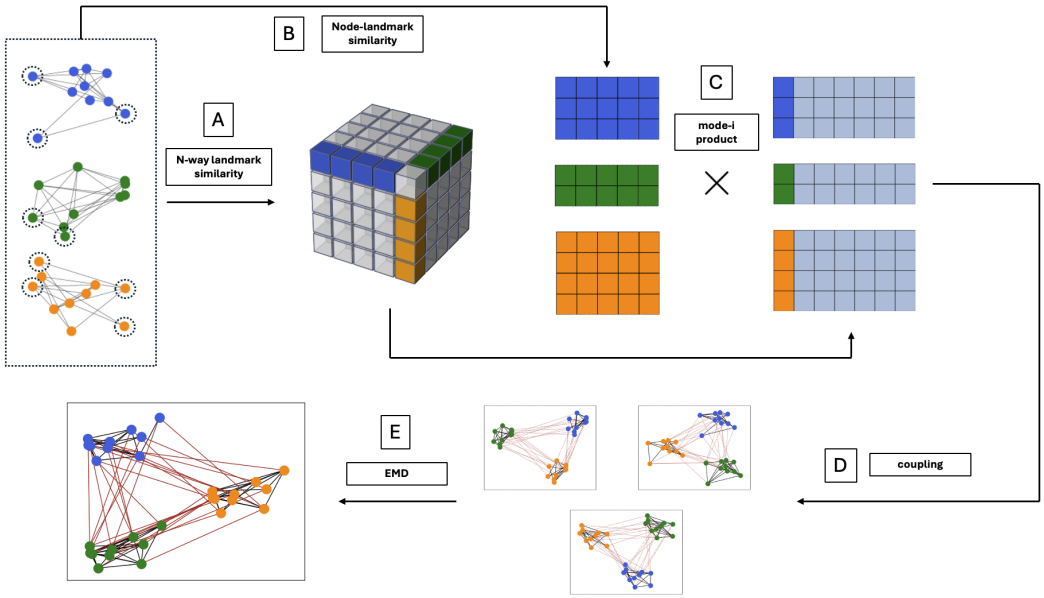


Figure 1: **Graph merging pipeline.** (A) Select landmark nodes in each input graph and compute an N-way landmark–landmark similarity tensor. (B) For every graph, compute node–landmark similarity matrices. (C) Obtain a shared (commensurable) embedding space by applying a mode-i tensor–matrix product between the similarity tensor and each node–landmark matrix. (D) Use the resulting embeddings to define an inter-graph coupling (soft alignment). (E) Optimize this coupling with Earth Mover's Distance (EMD) and use it to generate cross-graph edges, producing the merged graph.

Node features and shared representation. Assume all node features have the same dimensionality d. The merging stage receives precomputed node embeddings from an initial self-supervised learning (SSL) encoder. In our experiments, this encoder is GraphMAE (Hou et al., 2022); its output for each graph is the node-feature matrix consumed by landmark construction and OT. Each node is therefore represented by a feature vector $v \in \mathbb{R}^d$ in the encoder's shared representation space.

Landmarks. For each graph $\mathcal{G}_i$, select a landmark subset $\mathcal{L}_i \subset \mathcal{V}_i$ and collect landmark features in the matrix

$$L^{(i)} \in \mathbb{R}^{|\mathcal{L}_i| \times d}.$$

Landmark similarity tensor. Define the N-way *Kronecker-delta* tensor $T \in \mathbb{R}^{d \times d \times \cdots \times d}$ (order N) by

$$T_{r_1, r_2, \ldots, r_N} = \begin{cases} 1, & r_1 = r_2 = \cdots = r_N, \\ 0, & \text{otherwise.} \end{cases}$$

Using T, define the N-way landmark similarity tensor

$$S \in \mathbb{R}^{|\mathcal{L}_1| \times |\mathcal{L}_2| \times \cdots \times |\mathcal{L}_N|}$$

with entries

$$S_{j_1, j_2, \ldots, j_N} = \sum_{r_1=1}^{d} \sum_{r_2=1}^{d} \cdots \sum_{r_N=1}^{d} T_{r_1, r_2, \ldots, r_N} \prod_{k=1}^{N} L^{(k)}_{j_k, r_k}.$$

Equivalently (since T enforces $r_1 = \cdots = r_N = r$),

$$S_{j_1, j_2, \ldots, j_N} = \sum_{r=1}^{d} \prod_{k=1}^{N} L^{(k)}_{j_k, r}.$$

Node-to-landmark similarities. For each graph i, compute the node-to-landmark similarity matrix

$$L_S^{(i)} \in \mathbb{R}^{|\mathcal{V}_i| \times |\mathcal{L}_i|}$$

using cosine similarity between each node feature and each landmark feature.

Graph-specific embeddings via mode product. Let $\times_i$ denote the mode-i (Rabanser et al., 2017) tensor–matrix product. Define the embedding of nodes in graph i as

$$X^{(i)} = S \times_i L_S^{(i)},$$

so that

$$X^{(i)} \in \mathbb{R}^{|\mathcal{V}_i| \times |\mathcal{L}_{-i}|}, \qquad |\mathcal{L}_{-i}| \coloneqq \prod_{\substack{k=1 \\ k \neq i}}^{N} |\mathcal{L}_k|.$$

Optionally, the embedding dimensionality can be reduced to any

$$E \in \{1, 2, \ldots, |\mathcal{L}_{-i}|\}.$$

Cross-graph edge construction via EMD. For each pair of graphs $(\mathcal{G}_i, \mathcal{G}_j)$, connect nodes by minimizing the Earth Mover's Distance (EMD) between their empirical node distributions:

$$\mathrm{EMD}(P_{\mathcal{G}_i}, P_{\mathcal{G}_j}) = \inf_{\gamma \in \Pi(P_{\mathcal{G}_i}, P_{\mathcal{G}_j})} \mathbb{E}_{(\mathcal{V}_{im}, \mathcal{V}_{jn}) \sim \gamma}\big[d(\mathcal{V}_{im}, \mathcal{V}_{jn})\big],$$

where $P_{\mathcal{G}_i}$ is the uniform distribution over nodes in $\mathcal{V}_i$, i.e.

$$P_{\mathcal{G}_i}(\mathcal{V}_{im}) = \frac{1}{|\mathcal{V}_i|},$$

$\Pi(P_{\mathcal{G}_i}, P_{\mathcal{G}_j})$ is the set of couplings with marginals $P_{\mathcal{G}_i}$ and $P_{\mathcal{G}_j}$, and $d(\cdot, \cdot)$ is a chosen distance metric in the shared node-embedding space. All reported experiments solve this unregularized EMD problem as an exact linear program and add cross-graph edges for the support of the returned plan, $\{(u, v) : \gamma_{uv}^{(i,j)} > 0\}$. Each extracted edge is a label-free computational link: it encodes structural affinity under this learned representation and enables information exchange during downstream message passing. The main embedding construction procedure is summarized in Algorithm 1; the edge-extraction pseudocode for the OT/EMD stage is given in the appendix (Algorithm 2). We discuss the computational complexity in Section E.

3.2 Downstream target and label-free structural surrogate

Ideal downstream evaluation objective. Let $\mathcal{D} = \{(\{\mathcal{G}_i^{(s)}\}_{i=1}^N, y^{(s)})\}_{s=1}^{|\mathcal{D}|}$ be a dataset of samples, where each sample s provides N modality graphs and a shared label $y^{(s)}$ (graph-level classification in our experiments). For a candidate cross-edge set $\mathcal{E}_\times$, define the merged graph for sample s as

$$\mathcal{G}_{\mathrm{merge}}^{(s)}(\mathcal{E}_\times) = \Big(\bigsqcup_{i=1}^{N} \mathcal{V}_i^{(s)},\ \bigcup_{i=1}^{N} \mathcal{E}_i^{(s)} \ \cup\ \mathcal{E}_\times \Big),$$

and let g_ϕ be a GNN-based predictor operating on $\mathcal{G}_{\mathrm{merge}}^{(s)}(\mathcal{E}_\times)$. An ideal downstream objective would choose cross edges that improve generalization while keeping the merged graph sparse:

$$\min_{\mathcal{E}_\times \in \mathcal{A}} \ \min_{\phi} \ \frac{1}{|\mathcal{D}|} \sum_{s=1}^{|\mathcal{D}|} \ell\Big(g_\phi\Big(\mathcal{G}_{\mathrm{merge}}^{(s)}(\mathcal{E}_\times) \Big), y^{(s)} \Big), \tag{1}$$

where ℓ is the task loss and $\mathcal{A}$ is a constraint set encoding a sparsity/budget requirement on $\mathcal{E}_\times$ (discussed below).

Why direct optimization is intractable. Optimizing (1) over discrete cross-edge sets is combinatorial: potential cross edges scale as $O\big(\sum_{i \neq j} |\mathcal{V}_i||\mathcal{V}_j|\big)$, and the objective depends on $\mathcal{E}_\times$ only implicitly through non-linear message passing and training dynamics. Even continuous relaxations lead to a bilevel problem since the optimal parameters $\phi^\star(\mathcal{E}_\times)$ depend on the graph structure.

Algorithm 1 Multimodal graph merging: shared embeddings and landmark tensor construction

Input: modality graphs $\{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)\}_{i=1}^N$, node features $\{X_i\}$, SSL encoder f_θ, landmarks per graph $\{\mathcal{L}_i\}$, similarity $\mathrm{sim}(\cdot,\cdot)$
Output: per-graph node embeddings $\{X^{(i)}\}_{i=1}^N$
(1) Shared node representation (self-supervised).
for $i = 1$ **to** N **do**
  $H_i \leftarrow f_\theta(\mathcal{G}_i, X_i)$ $\{H_i \in \mathbb{R}^{|\mathcal{V}_i| \times d}\}$
  Select landmarks $\mathcal{L}_i \subset \mathcal{V}_i$; let $L^{(i)} \leftarrow H_i[\mathcal{L}_i]$ $\{L^{(i)} \in \mathbb{R}^{|\mathcal{L}_i| \times d}\}$
end for
(2) Build N-way landmark similarity tensor.
Construct $S \in \mathbb{R}^{|\mathcal{L}_1| \times \cdots \times |\mathcal{L}_N|}$ by
  $S_{j_1,\ldots,j_N} \leftarrow \sum_{r=1}^d \prod_{k=1}^N L^{(k)}_{j_k,r}$
(3) Node-to-landmark similarities and tensor-mode embeddings.
for $i = 1$ **to** N **do**
  Compute node-to-landmark similarity matrix $L_S^{(i)} \in \mathbb{R}^{|\mathcal{V}_i| \times |\mathcal{L}_i|}$:
    $(L_S^{(i)})_{u,j} \leftarrow \mathrm{sim}(H_i[u], L^{(i)}[j])$
  Compute graph-specific embedding via mode product:
    $X^{(i)} \leftarrow S \times_i L_S^{(i)}$
  (Optional) reduce dimensionality of $X^{(i)}$ to E (e.g., PCA)
end for

A label-free structural surrogate: representation → coupling → sparse support. Rather than optimize (1), our method decouples (i) learning a shared node representation and (ii) constructing cross edges by a globally consistent coupling in that representation space. Neither stage consumes the labels $y^{(s)}$ or the downstream loss ℓ; the objective above specifies the evaluation target, not the optimization performed by our edge-construction procedure. After SSL, each node $u \in \mathcal{V}_i$ has an embedding $h_u \in \mathbb{R}^d$, and we build cross-graph node embeddings $X^{(i)}$ via landmarks and tensor-mode products Algorithm 1. For each modality pair (i, j), we then compute an EMD coupling $\gamma^{(i,j)} \in \Pi(P_i, P_j)$ that minimizes expected transport cost:

$$\gamma^{(i,j)} \in \arg\min_{\gamma \in \Pi(P_i,P_j)} \sum_{u \in \mathcal{V}_i} \sum_{v \in \mathcal{V}_j} \gamma(u,v) \, d\Big(X^{(i)}[u], X^{(j)}[v]\Big). \tag{2}$$

with uniform node measures P_i, P_j. Cross edges are taken as the support of the coupling:

$$\mathcal{E}_\times^{(i,j)} \coloneqq \{(u,v) : u \in \mathcal{V}_i,\ v \in \mathcal{V}_j,\ \gamma^{(i,j)}(u,v) > 0\}, \qquad \mathcal{E}_\times \coloneqq \bigcup_{1 \le i < j \le N} \mathcal{E}_\times^{(i,j)}. \tag{3}$$

Implicit per-pair hard budget from EMD sparsity. The EMD problem (2) is a linear program over the transportation polytope, and therefore admits an optimal solution at an extreme point. Any extreme-point transport plan has at most $|\mathcal{V}_i| + |\mathcal{V}_j| - 1$ non-zero entries, implying

$$|\mathcal{E}_\times^{(i,j)}| = |\mathrm{supp}(\gamma^{(i,j)})| \le |\mathcal{V}_i| + |\mathcal{V}_j| - 1. \tag{4}$$

For the exact LP EMD solutions used in our experiments, this support-extraction rule induces a *hard* sparsity constraint per modality pair without a threshold hyperparameter, yielding an implicit budget set $\mathcal{A} = \{\mathcal{E}_\times : |\mathcal{E}_\times^{(i,j)}| \le |\mathcal{V}_i| + |\mathcal{V}_j| - 1,\ \forall i < j\}$.

Lemma 3.1 (EMD as an LP and sparsity of optimal couplings)**.** *Fix a modality pair (i, j) with $m = |\mathcal{V}_i|$ and $n = |\mathcal{V}_j|$. Define the cost matrix $C \in \mathbb{R}^{m \times n}$ by $C_{uv} = d\big(X^{(i)}[u], X^{(j)}[v]\big)$, and let $P_i \in \mathbb{R}^m$, $P_j \in \mathbb{R}^n$ be uniform measures. Let the transportation polytope be*

$$\Pi(P_i, P_j) \coloneqq \{\gamma \in \mathbb{R}_+^{m \times n} :\ \gamma \mathbf{1} = P_i,\ \gamma^\top \mathbf{1} = P_j\}.$$

Then the EMD problem

$$\gamma^\star \in \arg\min_{\gamma \in \Pi(P_i, P_j)} \langle \gamma, C \rangle, \qquad \langle \gamma, C \rangle \coloneqq \sum_{u=1}^{m} \sum_{v=1}^{n} \gamma_{uv} C_{uv}, \tag{5}$$

is a linear program over $\Pi(P_i, P_j)$ and admits an optimal solution at an extreme point. Moreover, any extreme-point optimal coupling satisfies

$$|\mathrm{supp}(\gamma^\star)| \leq m + n - 1. \tag{6}$$

Consequently, extracting cross edges as $\mathcal{E}_\times^{(i,j)} = \{(u, v) : \gamma^\star_{uv} > 0\}$ yields an implicit per-pair hard budget $|\mathcal{E}_\times^{(i,j)}| \leq |\mathcal{V}_i| + |\mathcal{V}_j| - 1$.

A proof is provided in Section A.1.

Why OT/EMD is preferable to greedy kNN. Greedy nearest-neighbor matching can suffer from hubness in high-dimensional embedding spaces, leading to many-to-one cross-modal connections (Section C). In contrast, the coupling constraints in OT/EMD enforce global mass conservation across node sets, producing balanced transport allocation and a sparse cross-edge scaffold that is well-suited for cross-modal message passing in the merged graph.

3.3 Landmarks and tensor embeddings as a low-rank cross-graph kernel

Our merging pipeline uses landmarks to (i) reduce cross-graph comparison cost and (ii) impose a structured, low-rank inductive bias on cross-modal similarity. This subsection explains why the landmark tensor and mode products can be viewed as constructing *kernel features* that are naturally suited for alignment and subsequent OT/EMD-based edge construction.

Landmarks as inducing points (Nyström intuition). Given shared-space node embeddings $h_u \in \mathbb{R}^d$ produced by the SSL encoder f_θ, we select a small landmark set $\mathcal{L}_i \subset \mathcal{V}_i$ per modality and store landmark embeddings $L^{(i)} \in \mathbb{R}^{|\mathcal{L}_i| \times d}$. For any node $u \in \mathcal{V}_i$, the node-to-landmark similarity vector

$$\phi_i(u) \;\coloneqq\; L_S^{(i)}[u, \cdot] \in \mathbb{R}^{|\mathcal{L}_i|}$$

acts as a *fingerprint* of u relative to a fixed reference set. This is analogous to Nyström kernel features: rather than comparing nodes across graphs directly (which is costly and sensitive to hubness in high dimension), we compare their *similarities to landmarks*, which is typically more stable and computationally cheaper.

Multiway landmark tensor as a low-rank compatibility kernel. The N-way landmark tensor

$$S_{j_1, \ldots, j_N} = \sum_{r=1}^{d} \prod_{k=1}^{N} L^{(k)}_{j_k, r}$$

scores the compatibility of a landmark tuple $(j_1, \ldots, j_N)$ across modalities. Crucially, S has a CP (CANDECOMP/PARAFAC) decomposition with rank at most d:

$$S = \sum_{r=1}^{d} a_r^{(1)} \otimes a_r^{(2)} \otimes \cdots \otimes a_r^{(N)}, \qquad a_r^{(k)} \in \mathbb{R}^{|\mathcal{L}_k|},$$

where $a_r^{(k)}$ is the r-th column of $L^{(k)}$. Thus, S is a *low-rank multiway kernel* over landmark indices: it restricts cross-modal compatibilities to be explained by a shared set of latent factors (the embedding dimensions $r = 1, \ldots, d$), regularizing against spurious affinity patterns.

Mode products yield cross-modal context signatures. For each modality i, the embedding

$$X^{(i)} = S \times_i L_S^{(i)}$$

maps each node $u \in \mathcal{V}_i$ to a feature vector whose coordinates correspond to *multi-modal landmark patterns* in the other modalities. In other words, $X^{(i)}[u, \cdot]$ describes how well node u aligns with consistent landmark configurations across the remaining modalities, rather than only measuring within-modality proximity. This is the representation that is subsequently compared across graphs when constructing OT/EMD couplings.

Lemma 3.2 (Landmark tensor induces low-rank cross-graph kernel features)**.** *Let* $\{L^{(k)} \in \mathbb{R}^{|\mathcal{L}_k|\times d}\}_{k=1}^{N}$ *be landmark embedding matrices and define the* N*-way landmark tensor* $S \in \mathbb{R}^{|\mathcal{L}_1|\times\cdots\times|\mathcal{L}_N|}$ *by*

$$S_{j_1,\ldots,j_N} \coloneqq \sum_{r=1}^{d}\prod_{k=1}^{N} L^{(k)}_{j_k,r}.$$

Then:

(i) CP low-rank structure. *S admits a CP decomposition with rank at most d:*

$$S = \sum_{r=1}^{d} a_r^{(1)} \otimes a_r^{(2)} \otimes \cdots \otimes a_r^{(N)}, \qquad a_r^{(k)} \coloneqq L^{(k)}_{:,r} \in \mathbb{R}^{|\mathcal{L}_k|}.$$

Hence S is a low-rank multiway kernel over landmark indices.

(ii) Nyström-like kernel features via mode products. *Given node-to-landmark similarity matrices $L_S^{(i)} \in \mathbb{R}^{|\mathcal{V}_i|\times|\mathcal{L}_i|}$, the mode product $X^{(i)} \coloneqq S \times_i L_S^{(i)}$ yields embeddings that can be written as a sum of d rank-1 terms across the remaining modes , so each node is represented by its similarity fingerprint to inducing landmarks filtered through S.*

A proof is provided in Section A.2.

Pairwise special case ($N = 2$): explicit low-rank kernel matrix. Let $N = 2$ with landmark matrices $L^{(1)} \in \mathbb{R}^{|\mathcal{L}_1|\times d}$ and $L^{(2)} \in \mathbb{R}^{|\mathcal{L}_2|\times d}$. The landmark tensor reduces to a matrix $S \in \mathbb{R}^{|\mathcal{L}_1|\times|\mathcal{L}_2|}$ with entries

$$\begin{aligned} S_{j_1,j_2} &= \sum_{r=1}^{d} L^{(1)}_{j_1,r}\, L^{(2)}_{j_2,r} \\ &= \big\langle L^{(1)}_{j_1,\cdot},\, L^{(2)}_{j_2,\cdot} \big\rangle. \end{aligned} \tag{7}$$

In matrix form,

$$S = L^{(1)}(L^{(2)})^{\top}, \qquad \operatorname{rank}(S) \le d. \tag{8}$$

The mode products simplify to standard matrix multiplication:

$$\begin{aligned} X^{(1)} = S \times_1 L_S^{(1)} &= L_S^{(1)}\, S \in \mathbb{R}^{|\mathcal{V}_1|\times|\mathcal{L}_2|}, \\ X^{(2)} = S \times_2 L_S^{(2)} &= L_S^{(2)}\, S^{\top} \in \mathbb{R}^{|\mathcal{V}_2|\times|\mathcal{L}_1|}. \end{aligned} \tag{9}$$

Thus, each node in graph 1 is represented as a weighted mixture (via its landmark fingerprint $L_S^{(1)}$) of a rank-$\le d$ kernel over landmarks in graph 2 (and symmetrically for graph 2). This provides a concrete interpretation of $X^{(i)}$ as Nyström-like cross-graph kernel features: nodes are embedded in a space indexed by *landmarks in the other graph*, enabling cross-graph comparison without constructing a dense $|\mathcal{V}_1| \times |\mathcal{V}_2|$ similarity matrix.

Why this representation helps downstream OT/EMD coupling. The subsequent OT/EMD stage relies on distances between node representations across graphs. Using $X^{(i)}$ instead of raw SSL embeddings H_i provides two benefits: (i) *stability*: similarity-to-landmark fingerprints reduce sensitivity to modality-specific distortions and high-dimensional hubness; and (ii) *cross-modal regularization*: the low-rank factorization of S biases representations toward cross-modal patterns that are consistent across shared latent dimensions. As a result, OT/EMD is applied in a space shaped by multi-modal affinity patterns, providing a structured basis for the induced cross edges used for graph merging.

3.4 Consistency of cosine-to-landmark fingerprints

Our construction uses node-to-landmark similarities as a modality-robust proxy for local structure. For each modality i, we map a node $u \in \mathcal{V}_i$ to its *cosine-to-landmark fingerprint* $\phi_i(u) \in \mathbb{R}^m$, where $m = |\mathcal{L}_i|$, by comparing u to a fixed set of landmark nodes in the shared embedding space.

Intuitively, if the SSL encoder yields commensurable representations across modalities (up to a modality-specific near-isometry), then cosine similarities to corresponding landmarks should be approximately preserved. The following proposition formalizes this stability and gives a sufficient margin condition under which nearest-neighbor relations in fingerprint space are preserved across modalities.

Proposition 3.3 (Stability of cosine-to-landmark fingerprints). *Let $h^{(i)}(u) \in \mathbb{R}^d$ be the embedding of node u in modality i, and let $\{\ell_a^{(i)}\}_{a=1}^m$ be the landmark embeddings in modality i. Define the cosine-to-landmark fingerprint $\phi_i(u) \in \mathbb{R}^m$ by*

$$\phi_i(u)_a \;=\; \cos\big(h^{(i)}(u), \ell_a^{(i)}\big), \qquad a = 1, \dots, m.$$

Assume there exist latent vectors $z(u), z_a \in \mathbb{R}^d$ and orthogonal matrices $\{Q_i\}$ such that

$$h^{(i)}(u) = Q_i z(u) + \epsilon_u^{(i)}, \qquad \ell_a^{(i)} = Q_i z_a + \eta_a^{(i)},$$

with $\|z(u)\| \geq \alpha$, $\|z_a\| \geq \alpha$, and $\|\epsilon_u^{(i)}\| \leq \delta$, $\|\eta_a^{(i)}\| \leq \delta$ for all u, a, i, where $\alpha > \delta > 0$. Then for all nodes u and modalities i, j,

$$\|\phi_i(u) - \phi_j(u)\|_\infty \;\leq\; \frac{8\delta}{\alpha - \delta}.$$

Moreover, let v be a nearest neighbor of u under modality i in fingerprint space, and define

$$\Delta \;\coloneqq\; \min_{w \neq v} \|\phi_i(u) - \phi_i(w)\|_2 \;-\; \|\phi_i(u) - \phi_i(v)\|_2.$$

If

$$\Delta \;>\; 4\sqrt{m}\,\frac{8\delta}{\alpha - \delta},$$

then v remains a nearest neighbor of u under modality j.

A detailed proof is provided in Section A.3, and numerical validation is discussed in Section D.

4 Experiments

We target the common setting in which a small, fixed collection of complementary modalities is available, and evaluate both two- and three-modality configurations.

We evaluate unsupervised structural graph merging on two heterogeneous multimodal graph classification benchmarks. In both, each modality is represented as its own graph with a shared graph-level label; labels are used only to train and evaluate the downstream predictor after the cross-graph edges have been constructed. We study (i) TCGA-BRCA for PAM50 breast-cancer subtype prediction using CNV, clinical, and WSI graphs (Amer et al., 2025), and (ii) SEN12MS for land-cover classification using SAR/VIS/SWIR graphs from co-registered Sentinel-1/2 patches (Schmitt et al., 2019; Schmitt & Wu, 2021). Starting from these separate input graphs, our empirical question is how to construct cross-graph topology for downstream message passing. We therefore compare methods that infer cross-graph relationships and convert them into edges: REGAL (Heimann et al., 2018), CONE-Align (Chen et al., 2020), GradAlign+ (Park et al., 2022), FUGAL (Bommakanti et al., 2024), CAPER (Zhu et al., 2022), and PARROT (Zeng et al., 2023). Representation-level fusion, gated fusion, and dense cross-attention define different multimodal architectures rather than edge-construction procedures, and are not part of this graph-topology comparison. For every run, GraphMAE (Hou et al., 2022) first produces the node embeddings supplied to the landmark/OT merge stage, and landmarks are sampled from those embeddings. SEN12MS preprocessing/graph construction is detailed in the appendix (Section B). We report all metrics as mean $\pm$ standard deviation, where the variability is taken over 10 folds for TCGA-BRCA and over 5 random seeds for SEN12MS. The main paper reports a compact view with single-modality baselines and the tri-modal setting (Tables 1 and 2); full pairwise and tri-modal results are reported in the appendix (Tables 4 and 5).

Results on SEN12MS. On SEN12MS, the compact comparison in Table 1 shows that our tri-modal model achieves the best Acc/F1/Kappa/AUC among the evaluated methods. Its reported score exceeds that of the strongest tri-modal alignment-derived edge baseline (CAPER) on every metric. The full appendix tables report the corresponding pairwise results across sensor combinations.

Table 1: Compact results on SEN12MS. Entries are reported as mean ± standard deviation over 5 random seeds. We report single-modality baselines and the tri-modal setting; full pairwise results remain in the appendix. Best multi-modal results are **bolded** and annotated with asterisk *.

MODEL	ACC	F1	KAPPA	AUC
SAR	0.6044 ± 0.006	0.6005 ± 0.009	0.5605 ± 0.007	0.9252 ± 0.001
VIS	0.6467 ± 0.008	0.6426 ± 0.008	0.6074 ± 0.009	0.9375 ± 0.002
SWIR	0.7113 ± 0.005	0.7079 ± 0.005	0.6792 ± 0.006	0.9590 ± 0.001
SAR+VIS+SWIR				
REGAL	0.7544 ± 0.009	0.7520 ± 0.007	0.7271 ± 0.01	0.9697 ± 0.0007
CONE-ALIGN	0.7637 ± 0.01	0.7622 ± 0.01	0.7374 ± 0.01	0.9706 ± 0.0007
GRADALIGN+	0.7548 ± 0.01	0.7536 ± 0.01	0.7276 ± 0.01	0.9706 ± 0.001
CAPER	0.7806 ± 0.007	0.7775 ± 0.008	0.7562 ± 0.008	0.9749 ± 0.001
FUGAL	0.7728 ± 0.004	0.7714 ± 0.005	0.7498 ± 0.005	0.9565 ± 0.0007
PARROT	0.7637 ± 0.0125	0.7636 ± 0.0112	0.7374 ± 0.0138	0.9715 ± 0.0011
OURS	$\mathbf{0.7855 \pm 0.007}^*$	$\mathbf{0.7830 \pm 0.008}^*$	$\mathbf{0.7616 \pm 0.008}^*$	$\mathbf{0.9752 \pm 0.0009}^*$

Table 2: Compact results on TCGA-BRCA. Entries are reported as mean ± standard deviation over 10 folds. We report single-modality baselines and the tri-modal setting; full pairwise results remain in the appendix. Best multi-modal results are **bolded** and annotated with asterisk *.

MODEL	ACC	F1	KAPPA	AUC
CNV	0.4328 ± 0.04	0.4001 ± 0.04	0.3610 ± 0.07	0.7836 ± 0.02
CLINICAL	0.5677 ± 0.04	0.5849 ± 0.03	0.4738 ± 0.07	0.8341 ± 0.03
WSI	0.4440 ± 0.03	0.4363 ± 0.04	0.3658 ± 0.07	0.7944 ± 0.04
CNV+CLINICAL+WSI				
REGAL	0.5303 ± 0.04	0.5089 ± 0.05	0.4945 ± 0.05	0.8585 ± 0.02
CONE-ALIGN	0.5581 ± 0.03	0.5350 ± 0.02	0.5364 ± 0.04	0.8621 ± 0.02
GRADALIGN+	0.5563 ± 0.04	0.5348 ± 0.03	0.5149 ± 0.05	0.8607 ± 0.02
CAPER	0.5418 ± 0.03	0.5298 ± 0.04	0.5072 ± 0.05	0.8657 ± 0.02
FUGAL	0.5299 ± 0.03	0.5099 ± 0.02	0.4985 ± 0.05	0.8484 ± 0.03
PARROT	0.5204 ± 0.0244	0.4995 ± 0.0323	0.4798 ± 0.0539	0.8352 ± 0.0302
OURS	$\mathbf{0.6529 \pm 0.05}^*$	$\mathbf{0.6593 \pm 0.05}^*$	$\mathbf{0.5863 \pm 0.06}^*$	$\mathbf{0.8890 \pm 0.02}^*$

Results on TCGA-BRCA. On BRCA, our method achieves the best results among the evaluated methods in the full CNV+Clinical+WSI setting, with the largest reported scores on all metrics. The appendix provides the full pairwise breakdown across all studied modality combinations.

Discussion. Overall, the results show that the evaluated graph-topology construction procedures can differ substantially across modality combinations. The appendix also shows variation among pairwise configurations, highlighting that the utility of induced cross-graph topology depends on the studied graphs and task. This motivates future work on *adaptive* merging—e.g., selectively pruning cross-edges, tuning coupling strength per modality pair, or using confidence-aware transport—to introduce cross-graph connectivity when the learned topology is sufficiently informative for the downstream task.

5 CONCLUSION

We present unsupervised structural graph merging, a label-free procedure that adds sparse cross-graph edges between modality-specific graphs for downstream message passing. Across the evaluated two- and three-modality configurations, multiway landmark context and OT/EMD support extraction achieve the strongest tri-modal results among the tested alignment-derived edge baselines. These results motivate sparse cross-graph topology for multimodal message passing and future work on adaptive coupling.

AI USE DISCLOSURE

We used generative AI to help develop theoretical models or conceptual frameworks and to assist in the writing of proofs. We did not use generative AI for the other tasks for which the ICLR 2027 policy requires disclosure. We also used generative AI tools for the following recommended-disclosure activities: creating or modifying scientific figures or images; creating or editing software code; drafting and editing portions of the manuscript; summarizing or analyzing existing literature; discovering research topics or identifying gaps; brainstorming; sourcing or searching for information; improving readability; and identifying relevant literature. The authors reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mohammed Amer, Mohamed A Suliman, Tu Bui, Nuria Garcia, and Serban Georgescu. Scalable and loosely-coupled multimodal deep learning for breast cancer subtyping. *arXiv preprint arXiv:2509.03408*, 2025.

Aditya Bommakanti, Harshith R Vonteri, Konstantinos Skitsas, Sayan Ranu, Davide Mottin, and Panagiotis Karras. Fugal: Feature-fortified unrestricted graph alignment. *Advances in Neural Information Processing Systems*, 37:19523–19546, 2024.

Xiyuan Chen, Mark Heimann, Fatemeh Vahedian, and Danai Koutra. Cone-align: Consistent network alignment with proximity-preserving node embedding. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1985–1988, 2020.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H196sainb. arXiv:1710.04087.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. doi: 10.48550/arXiv.1412.6568. URL https://arxiv.org/abs/1412.6568. Also appeared in ICLR Workshop (2015).

Roman Feldbauer and Arthur Flexer. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems*, 59:137–166, 2019. doi: 10.1007/s10115-018-1205-y. URL https://link.springer.com/article/10.1007/s10115-018-1205-y.

Mark Heimann, Haoming Shen, Tara Safavi, and Danai Koutra. Regal: Representation learning-based graph alignment. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 117–126, 2018.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Jin-Duk Park, Cong Tran, Won-Yong Shin, and Xin Cao. Gradalign+: Empowering gradual network alignment using attribute augmentation. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 4374–4378, 2022.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Stephan Rabanser, Oleksandr Shchur, and Stephan Günnemann. Introduction to tensor decompositions and their applications in machine learning. *arXiv preprint arXiv:1711.10781*, 2017.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pp. 865–872. ACM, 2009. doi: 10.1145/1553374.1553485. URL https://icml.cc/Conferences/2009/papers/360.pdf.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010.

Michael Schmitt and Yu-Lun Wu. Remote sensing image classification with the sen12ms dataset. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-2-2021, pp. 101–106, 2021. doi: 10.5194/isprs-annals-V-2-2021-101-2021.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2/W7, pp. 153–160, 2019. doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

Konstantinos Skitsas, Karol Orlowski, Judith Hermanns, Davide Mottin, and Panagiotis Karras. Comprehensive evaluation of algorithms for unrestricted graph alignment. In *EDBT*, pp. 260–272, 2023.

Qingqiang Sun, Xuemin Lin, Ying Zhang, Wenjie Zhang, and Chaoqi Chen. Towards higher-order topological consistency for unsupervised network alignment. In *2023 IEEE 39th international conference on data engineering (ICDE)*, pp. 177–190. IEEE, 2023.

Louis Vassaux and Laurent Massoulié. The feasibility of multi-graph alignment: a bayesian approach. *arXiv preprint arXiv:2502.17142*, 2025.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Liren Yu, Jiaming Xu, and Xiaojun Lin. Seedgnn: graph neural network for supervised seeded graph matching. In *International Conference on Machine Learning*, pp. 40390–40411. PMLR, 2023.

Zhichen Zeng, Si Zhang, Yinglong Xia, and Hanghang Tong. Parrot: Position-aware regularized optimal transport for network alignment. In *Proceedings of the ACM web conference 2023*, pp. 372–382, 2023.

Jing Zhu, Danai Koutra, and Mark Heimann. Caper: Coarsen, align, project, refine-a general multilevel framework for network alignment. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 4747–4751, 2022.

A PROOFS

A.1 LEMMA 3.1

Step 1: EMD is a linear program over the transportation polytope. Fix (i, j) with $m = |\mathcal{V}_i|$ and $n = |\mathcal{V}_j|$. Let $C \in \mathbb{R}^{m\times n}$ be defined by $C_{uv} = d(X^{(i)}[u], X^{(j)}[v])$. The EMD coupling variables are the entries of $\gamma \in \mathbb{R}^{m\times n}$, constrained by: (i) nonnegativity $\gamma_{uv} \geq 0$, (ii) row-sum constraints $\sum_{v=1}^{n} \gamma_{uv} = P_i(u)$ for all u, (iii) column-sum constraints $\sum_{u=1}^{m} \gamma_{uv} = P_j(v)$ for all v. Collecting these constraints yields the feasible set

$$\Pi(P_i, P_j) = \{\gamma \in \mathbb{R}^{m\times n} : \ \gamma \geq 0, \ \gamma\mathbf{1} = P_i, \ \gamma^\top \mathbf{1} = P_j\}.$$

This set is an intersection of finitely many affine half-spaces and hyperplanes, hence a convex polyhedron. Since $\sum_u P_i(u) = \sum_v P_j(v) = 1$ and $\gamma \geq 0$, it is bounded, therefore a polytope (the *transportation polytope*).

The objective is

$$\langle \gamma, C\rangle = \sum_{u=1}^{m}\sum_{v=1}^{n} \gamma_{uv} C_{uv},$$

which is linear in the decision variables $\{\gamma_{uv}\}$. Thus EMD is a linear program (LP) over the transportation polytope.

Step 2: An optimal solution exists at an extreme point. Because $\Pi(P_i, P_j)$ is a non-empty compact (closed and bounded) polytope and the objective $\langle \gamma, C\rangle$ is continuous and linear, the minimum is attained. A standard result in linear programming states that if an LP attains its optimum over a polytope, then at least one optimal solution is attained at a vertex (extreme point) of the feasible polytope. Hence there exists an optimal coupling $\gamma^\star$ that is an extreme point of $\Pi(P_i, P_j)$.

Step 3: Extreme points have support size at most $m + n - 1$. We next show that any extreme point $\gamma^\star$ of $\Pi(P_i, P_j)$ has at most $m + n - 1$ strictly positive entries.

The matrix $\gamma \in \mathbb{R}^{m\times n}$ has mn variables. The equality constraints consist of: m independent row-sum constraints and n column-sum constraints. However, one equation is redundant because summing all row constraints gives

$$\sum_{u=1}^{m}\sum_{v=1}^{n} \gamma_{uv} = \sum_{u=1}^{m} P_i(u) = 1,$$

and summing all column constraints gives the same total mass 1. Therefore, the number of linearly independent equality constraints is $m + n - 1$.

Extreme points of polytopes described by nonnegativity constraints and linear equalities correspond to *basic feasible solutions* (BFS) of the LP. Any BFS has at most as many basic (potentially nonzero) variables as the number of independent equality constraints; here that number is $m + n - 1$. Thus there exists an extreme-point optimal $\gamma^\star$ with

$$|\mathrm{supp}(\gamma^\star)| \leq m + n - 1.$$

(Equivalent graph intuition). Associate to γ a bipartite graph with left nodes $\mathcal{V}_i$ and right nodes $\mathcal{V}_j$, and an edge (u, v) whenever $\gamma_{uv} > 0$. For non-degenerate extreme points, this support graph is a forest (it contains no cycles); otherwise one could perturb mass around a cycle while preserving the marginals, which contradicts extremality. Any forest spanning $m + n$ vertices has at most $m + n - 1$ edges, giving the same bound.

Step 4: Implicit hard budget for support-extracted cross edges. Defining cross edges by support, $\mathcal{E}_\times^{(i,j)} = \{(u, v) : \gamma^\star(u, v) > 0\}$, immediately yields

$$|\mathcal{E}_\times^{(i,j)}| = |\mathrm{supp}(\gamma^\star)| \leq m + n - 1 = |\mathcal{V}_i| + |\mathcal{V}_j| - 1,$$

which is the claimed implicit per-pair hard budget.

A.2 LEMMA 3.2

(i) CP decomposition and rank bound. For each mode $k \in \{1, \dots, N\}$ and component $r \in \{1, \dots, d\}$, define the vector $a_r^{(k)} \in \mathbb{R}^{|\mathcal{L}_k|}$ as the r-th column of $L^{(k)}$, i.e., $a_r^{(k)} \coloneqq L^{(k)}_{\cdot,r}$. Consider the rank-1 tensor (outer product) $a_r^{(1)} \otimes \cdots \otimes a_r^{(N)}$, whose entries satisfy

$$\left(a_r^{(1)} \otimes \cdots \otimes a_r^{(N)}\right)_{j_1,\dots,j_N} = \prod_{k=1}^{N} a_r^{(k)}[j_k] = \prod_{k=1}^{N} L^{(k)}_{j_k,r}.$$

Summing over r yields

$$\left(\sum_{r=1}^{d} a_r^{(1)} \otimes \cdots \otimes a_r^{(N)}\right)_{j_1,\dots,j_N} = \sum_{r=1}^{d} \prod_{k=1}^{N} L^{(k)}_{j_k,r} = S_{j_1,\dots,j_N},$$

so

$$S = \sum_{r=1}^{d} a_r^{(1)} \otimes a_r^{(2)} \otimes \cdots \otimes a_r^{(N)}.$$

This is exactly a CP (CANDECOMP/PARAFAC) decomposition with at most d rank-1 terms, hence $\mathrm{rank}_{\mathrm{CP}}(S) \le d$.

(ii) Mode product yields low-rank kernel features. Fix a modality i and define $X^{(i)} \coloneqq S \times_i L_S^{(i)}$. By the definition of the mode-i product, for any node $u \in \mathcal{V}_i$ and any index tuple $j_{-i} = (j_1, \dots, j_{i-1}, j_{i+1}, \dots, j_N)$,

$$X^{(i)}_{u,j_{-i}} = \sum_{j_i=1}^{|\mathcal{L}_i|} (L_S^{(i)})_{u,j_i} \, S_{j_1,\dots,j_N}. \tag{10}$$

Substitute the CP form of S into (10):

$$X^{(i)}_{u,j_{-i}} = \sum_{j_i} (L_S^{(i)})_{u,j_i} \sum_{r=1}^{d} \prod_{k=1}^{N} L^{(k)}_{j_k,r} = \sum_{r=1}^{d} \left(\sum_{j_i} (L_S^{(i)})_{u,j_i} \, L^{(i)}_{j_i,r} \right) \prod_{k \neq i} L^{(k)}_{j_k,r}.$$

Define the scalar coefficient

$$b_r^{(i)}(u) \;\coloneqq\; \sum_{j_i=1}^{|\mathcal{L}_i|} (L_S^{(i)})_{u,j_i} \, L^{(i)}_{j_i,r} \;=\; \left\langle (L_S^{(i)})_{u,\cdot}, \, a_r^{(i)} \right\rangle.$$

Then

$$X^{(i)}_{u,j_{-i}} = \sum_{r=1}^{d} b_r^{(i)}(u) \prod_{k \neq i} a_r^{(k)}[j_k].$$

Equivalently, as a tensor over the remaining modes, each node embedding decomposes as

$$X^{(i)}[u, \cdot] = \sum_{r=1}^{d} b_r^{(i)}(u) \left(a_r^{(1)} \otimes \cdots \otimes a_r^{(i-1)} \otimes a_r^{(i+1)} \otimes \cdots \otimes a_r^{(N)} \right),$$

i.e., a sum of d rank-1 components across the other modalities. This shows that $X^{(i)}$ is a structured, low-rank set of *kernel features* computed from the node-to-landmark fingerprint $(L_S^{(i)})_{u,\cdot}$ and the multiway compatibility tensor S.

A.3 PROPOSITION 3.3

We first prove a perturbation bound for cosine similarity.

Lemma A.1 (Cosine perturbation bound)**.** *Let $x, y \in \mathbb{R}^d$ satisfy $\|x\| \ge \beta$ and $\|y\| \ge \beta$ for some $\beta > 0$. For perturbations $\Delta x, \Delta y$ with $\|\Delta x\| \le \rho$ and $\|\Delta y\| \le \rho$ and $\beta > \rho$, define $\tilde{x} = x + \Delta x$ and $\tilde{y} = y + \Delta y$. Then*

$$\left| \cos(\tilde{x}, \tilde{y}) - \cos(x, y) \right| \;\le\; \frac{4\rho}{\beta - \rho}.$$

Proof. By the reverse triangle inequality,

$$\|\tilde{x}\| \geq \|x\| - \|\Delta x\| \geq \beta - \rho, \qquad \|\tilde{y}\| \geq \|y\| - \|\Delta y\| \geq \beta - \rho.$$

For any nonzero vectors u and v,

$$\left\| \frac{u}{\|u\|} - \frac{v}{\|v\|} \right\| \leq \left\| \frac{u-v}{\|u\|} \right\| + \|v\| \left| \frac{1}{\|u\|} - \frac{1}{\|v\|} \right| \leq \frac{2\|u-v\|}{\|u\|},$$

where the last step uses $|\|v\| - \|u\|| \leq \|u-v\|$. Applying this with $(u,v) = (\tilde{x}, x)$ and $(u,v) = (\tilde{y}, y)$ gives

$$\left\| \frac{\tilde{x}}{\|\tilde{x}\|} - \frac{x}{\|x\|} \right\| \leq \frac{2\rho}{\beta - \rho}, \qquad \left\| \frac{\tilde{y}}{\|\tilde{y}\|} - \frac{y}{\|y\|} \right\| \leq \frac{2\rho}{\beta - \rho}.$$

Finally, writing $\hat{x} = x/\|x\|$ and similarly for the other vectors,

$$\begin{aligned} \big|\cos(\tilde{x}, \tilde{y}) - \cos(x, y)\big| &= \left| \hat{\tilde{x}}^\top \hat{\tilde{y}} - \hat{x}^\top \hat{y} \right| \\ &\leq \left| (\hat{\tilde{x}} - \hat{x})^\top \hat{\tilde{y}} \right| + \left| \hat{x}^\top (\hat{\tilde{y}} - \hat{y}) \right| \\ &\leq \left\| \hat{\tilde{x}} - \hat{x} \right\| + \left\| \hat{\tilde{y}} - \hat{y} \right\| \leq \frac{4\rho}{\beta - \rho}. \end{aligned}$$

This proves the claim. □

Proof of Proposition 3.3. Fix a node u and landmark index a. By assumption,

$$h^{(i)}(u) = Q_i z(u) + \epsilon_u^{(i)}, \qquad \ell_a^{(i)} = Q_i z_a + \eta_a^{(i)},$$

with $\|z(u)\| \geq \alpha$, $\|z_a\| \geq \alpha$ and $\|\epsilon_u^{(i)}\| \leq \delta$, $\|\eta_a^{(i)}\| \leq \delta$. Since Q_i is orthogonal, $\|Q_i z\| = \|z\|$ and

$$\cos(Q_i z(u), Q_i z_a) = \cos(z(u), z_a).$$

Apply Lemma A.1 with $x = Q_i z(u)$, $y = Q_i z_a$, $\Delta x = \epsilon_u^{(i)}$, $\Delta y = \eta_a^{(i)}$, $\beta = \alpha$ and $\rho = \delta$ to obtain

$$\left| \cos(h^{(i)}(u), \ell_a^{(i)}) - \cos(z(u), z_a) \right| \leq \frac{4\delta}{\alpha - \delta}.$$

The same bound holds for modality j. Therefore, by the triangle inequality,

$$|\phi_i(u)_a - \phi_j(u)_a| \leq |\phi_i(u)_a - \cos(z(u), z_a)| + |\phi_j(u)_a - \cos(z(u), z_a)| \leq \frac{8\delta}{\alpha - \delta}.$$

Taking a maximum over a yields

$$\|\phi_i(u) - \phi_j(u)\|_\infty \leq \frac{8\delta}{\alpha - \delta}.$$

Next, to show nearest-neighbor preservation, let v be the nearest neighbor of u under modality i in fingerprint space, and define

$$K \coloneqq \frac{8\delta}{\alpha - \delta}.$$

From the ℓ_∞ bound above, for any node x we have

$$\|\phi_j(x) - \phi_i(x)\|_2 \leq \sqrt{m}\, \|\phi_j(x) - \phi_i(x)\|_\infty \leq \sqrt{m}\, K.$$

Hence for any w,

$$\Big| \|\phi_j(u) - \phi_j(w)\|_2 - \|\phi_i(u) - \phi_i(w)\|_2 \Big| \leq \|\phi_j(u) - \phi_i(u)\|_2 + \|\phi_j(w) - \phi_i(w)\|_2 \leq 2\sqrt{m}\, K.$$

Let $\varepsilon \coloneqq 2\sqrt{m}\, K$. Then for any $w \neq v$,

$$\|\phi_j(u) - \phi_j(w)\|_2 - \|\phi_j(u) - \phi_j(v)\|_2 \geq \Big(\|\phi_i(u) - \phi_i(w)\|_2 - \|\phi_i(u) - \phi_i(v)\|_2 \Big) - 2\varepsilon.$$

Therefore, if the margin

$$\Delta \;\coloneqq\; \min_{w \neq v} \; \|\phi_i(u) - \phi_i(w)\|_2 \;-\; \|\phi_i(u) - \phi_i(v)\|_2$$

satisfies $\Delta > 2\varepsilon = 4\sqrt{m}\, K$, then

$$\|\phi_j(u) - \phi_j(v)\|_2 < \|\phi_j(u) - \phi_j(w)\|_2 \quad \text{for all } w \neq v,$$

so v remains a nearest neighbor of u under modality j. □

B SEN12MS

B.1 COMMON SETUP (ALL MODALITIES)

Let a SEN12MS patch be mapped to a regular grid of pixels. For each modality we construct a region graph $G = (V, E)$ with $|V| \approx 10^3$ nodes using SLIC superpixels, and edges defined by k-nearest neighbors (k-NN) in node-feature space. A node $v \in V$ corresponds to a superpixel mask $\Omega_v \subset \{1, \dots, H\} \times \{1, \dots, W\}$.

Training-set standardization. For any channel image X (per modality/band), we compute training statistics μ_X, σ_X over all training patches and pixels, and apply z-scoring:

$$\tilde{X}(p) = \frac{X(p) - \mu_X}{\sigma_X + \varepsilon}, \qquad \varepsilon > 0. \tag{11}$$

The same μ_X, σ_X are reused for validation/test.

Superpixel nodes. Given a multi-channel image stack I, SLIC produces masks $\{\Omega_v\}_{v=1}^{|V|}$ with $n_{\text{segments}} \approx 1000$ and compactness ≈ 10.

Node features and aggregation. For each node v, we compute a feature vector $\mathbf{x}_v \in \mathbb{R}^d$ by aggregating per-pixel maps inside Ω_v. For any per-pixel feature map $f(p)$, we use mean (and std where applicable):

$$\mu_f(v) = \frac{1}{|\Omega_v|} \sum_{p \in \Omega_v} f(p), \qquad \sigma_f(v) = \sqrt{\frac{1}{|\Omega_v|} \sum_{p \in \Omega_v} \big(f(p) - \mu_f(v)\big)^2}. \tag{12}$$

k-NN edges. Edges are constructed by connecting each node to its k nearest neighbors under a distance in feature space:

$$E = \{(u, v) \mid v \in \mathrm{NN}_k(u)\}. \tag{13}$$

B.2 SAR GRAPH G_{SAR} (SENTINEL-1 VV/VH)

Source. Sentinel-1 dual-pol SAR (VV, VH) co-registered to the optical patch.

Preprocess.

- Light denoising to reduce speckle while preserving edges.
- Standardize each channel over the training set: $\widetilde{\mathrm{VV}}, \widetilde{\mathrm{VH}}$.

Nodes. SLIC over $\widetilde{\mathrm{VV}} \oplus \widetilde{\mathrm{VH}}$.

Node features (18D). All features are computed within Ω_v.

- **First-order (4D):** μ, σ for VV and VH: $\mu_{\mathrm{VV}}, \sigma_{\mathrm{VV}}, \mu_{\mathrm{VH}}, \sigma_{\mathrm{VH}}$.
- **Polarization contrast (1D):** $\Delta = \mathrm{VV} - \mathrm{VH}$.
- **Texture (8D):** GLCM per channel (contrast, homogeneity, energy, entropy) on quantized VV and VH (distance d=1, 4 directions).
- **Gradient/edges (3D):** Sobel magnitude mean, Sobel magnitude std, edge density.
- **Shape/geometry (2D):** compactness and elongation (major/minor axis ratio) from Ω_v.

B.3 VIS GRAPH G_{VIS} (SENTINEL-2 B2/B3/B4)

Source. Sentinel-2 visible bands: B2 (Blue), B3 (Green), B4 (Red).

Preprocess.

- Per-band scaling: z-score per band.
- Mild Gaussian smoothing ($\sigma = 1$) prior to SLIC.

Nodes. SLIC over $B \oplus G \oplus R$.

Node features (25D). Let B, G, R denote the normalized bands (B2,B3,B4). Compute per-pixel maps then aggregate within Ω_v.

- **(A) Band stats (6D):** $\mu_B, \mu_G, \mu_R, \sigma_B, \sigma_G, \sigma_R$.
- **(B) Chromaticity & luminance (4D):**

$$Y = 0.2126R + 0.7152G + 0.0722B, \tag{14}$$

$$r = \frac{R}{R+G+B+\varepsilon}, \qquad g = \frac{G}{R+G+B+\varepsilon}, \tag{15}$$

$$S = 1 - \frac{\min(R,G,B)}{Y+\varepsilon}. \tag{16}$$

  Aggregate $\mu_Y, \mu_r, \mu_g, \mu_S$.
- **(C) VIS-only indices (4D):**

$$\mathrm{VARI} = \frac{G-R}{G+R-B+\varepsilon}, \tag{17}$$

$$\mathrm{ExG} = 2G - R - B, \tag{18}$$

$$\mathrm{ExR} = 1.4R - G, \tag{19}$$

$$\mathrm{CIVE} = 0.441R - 0.811G + 0.385B + 18.78745. \tag{20}$$

  Aggregate mean per region.
- **(D) Texture on luminance (4D):** GLCM on quantized Y (distance d=1, 4 directions): contrast, homogeneity, energy, entropy.
- **(E) Gradient/edges (3D):** Sobel magnitude mean/std and edge density on normalized Y.
- **(F) Shape/geometry (2D):** compactness and elongation from Ω_v.
- **(G) Robustness (2D):** whiteness and/or shadow score, $W = \|[R,G,B] - Y\mathbf{1}\|$ (region mean), and fraction of pixels with Y below a low quantile.

B.4 SWIR GRAPH G_{SWIR} (SENTINEL-2 B11/B12)

Source. Sentinel-2 SWIR bands B11 (SWIR1) and B12 (SWIR2).

Preprocess.

- Z-score per band across the training set.

Nodes. SLIC over $\mathrm{B11} \oplus \mathrm{B12}$.

Node features (18D). Compute within each Ω_v.

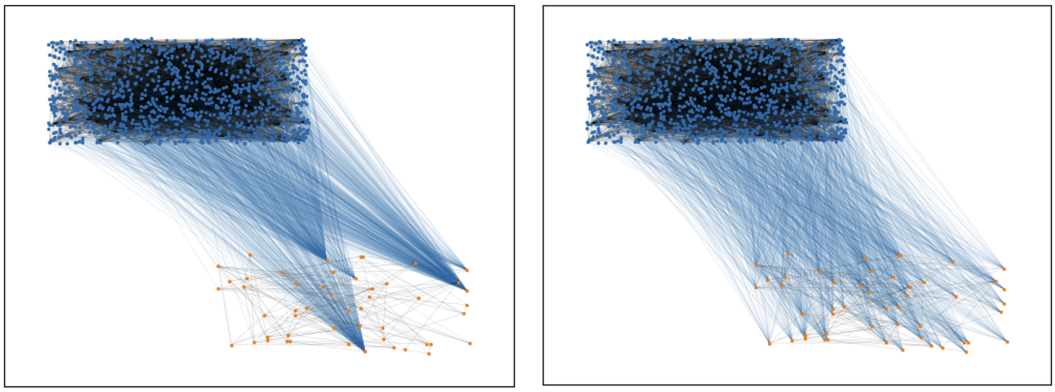
- **(A) First-order stats (4D):** $\mu_{S_1}, \mu_{S_2}, \sigma_{S_1}, \sigma_{S_2}$.
- **(B) SWIR composites/indices (5D):**

$$B = \tfrac{1}{2}(S_1 + S_2) \quad \Rightarrow\ (\mu_B, \sigma_B), \tag{21}$$

$$D = S_1 - S_2 \quad \Rightarrow\ \mu_D, \tag{22}$$

$$\mathrm{ND}_{\mathrm{SWIR}} = \frac{S_1 - S_2}{S_1 + S_2 + \varepsilon} \quad \Rightarrow\ \mu_{\mathrm{ND}}, \tag{23}$$

$$R = \frac{S_1}{S_2 + \varepsilon} \quad \Rightarrow\ \mu_R. \tag{24}$$



(a) **SOTA baseline exhibits hubness.** Many-to-one fan-in onto a few target hubs.

(b) **Our method reduces hubness.** Edges spread more evenly across targets.

Figure 2: **Hubness in cross-graph linking: baseline vs. ours.** The SOTA baseline shows hub targets with disproportionately high k-occurrence, while our method reduces this fan-in and improves target coverage.

- **(C) Texture (4D):** GLCM (distance d=1, 4 directions) on quantized brightness B (contrast, homogeneity, energy, entropy).
- **(D) Gradient/edges (3D):** Sobel magnitude mean/std and edge density on normalized brightness B.
- **(E) Shape/geometry (2D):** compactness and elongation from Ω_v.

Geometry definitions. For a region mask Ω_v with area $A(v)$ (pixels) and perimeter $P(v)$ (pixels),

$$\text{Compactness}(v) = \frac{4\pi A(v)}{P(v)^2 + \varepsilon}, \qquad \text{Elongation}(v) = \frac{a_{\max}(v)}{a_{\min}(v) + \varepsilon}, \tag{25}$$

where $a_{\max}, a_{\min}$ are the major/minor axis lengths from second-moment analysis of the mask.

Target task (SEN12MS land-cover classification). For SEN12MS, the target task is **land-cover classification** at the patch level. We formulate this as a C-way classification problem with $C = 10$ land-cover classes. To obtain a balanced benchmark, we subsample **1000 patches per class** (10,000 total samples) from the dataset and use the resulting labels as graph-level targets for each modality-specific graph constructed from the corresponding patch. The selected classes are:

1. Evergreen Needleleaf Forests
2. Evergreen Broadleaf Forests
3. Deciduous Broadleaf Forests
4. Mixed Forests
5. Open Shrublands
6. Woody Savannas
7. Savannas
8. Grasslands
9. Croplands
10. Urban and Built-up Lands

C Appendix: Hubness in cross-graph nearest-neighbor linking

When we connect nodes across graphs by nearest-neighbor (NN) retrieval in a shared embedding space (e.g., for two modalities), we often observe *hubness*: a small number of target nodes become nearest neighbors of an unusually large number of source nodes. Hubness is a well-studied

facet of the *curse of dimensionality* in similarity search and kNN graphs, originally formalized via the *k-occurrence* distribution—the number of times a point appears in other points' kNN lists. (Radovanovic et al., 2010; Radovanović et al., 2009) show that in high-dimensional spaces, this distribution becomes highly skewed, producing *hubs* (very frequent neighbors) and *antihubs* (rarely/never retrieved neighbors).

Formal definition (k-occurrence). Let X (source) and Y (target) be node embeddings, and let $N_k(x) \subseteq Y$ denote the k nearest targets of $x \in X$ under a distance/similarity metric. The *k-occurrence* of a target $y \in Y$ is

$$N_k(y) \;=\; |\{x \in X \;:\; y \in N_k(x)\}|\,. \tag{26}$$

A *hub* is a y with unusually large $N_k(y)$; an *antihub* has unusually small $N_k(y)$ (often 0). Hubness can be summarized by statistics of $\{N_k(y)\}_{y\in Y}$, e.g., its skewness or the fraction of mass captured by the top-m targets (Radovanovic et al., 2010; Feldbauer & Flexer, 2019).

Why hubness emerges (intuition). Hubness is closely tied to *distance concentration*: as dimensionality grows, pairwise distances become less informative, and small geometric biases get amplified. Points closer to the global (or local) centroid tend to be slightly closer to many other points and thus become *popular neighbors* (hubs) (Radovanovic et al., 2010; Feldbauer & Flexer, 2019). In cross-domain settings, an additional driver is *mapping-induced anisotropy*: learned projections (or joint encoders) can compress variance along certain directions, causing many source points to crowd toward regions where a few targets lie, worsening many-to-one NN matches. This effect is documented in cross-lingual and zero-shot retrieval, where NN neighborhoods become "polluted by hubs," pushing correct matches down the neighbor list (Dinu et al., 2014; Conneau et al., 2018).

Hubness in graph merging: SOTA vs. our method. In bipartite linking between graphs, hubness manifests as *star-like fans* of edges into a few targets: many source nodes select the same target(s) as their nearest neighbor(s). This failure mode is visible in Figure 2a (a representative SOTA baseline): a dense set of source nodes (blue, top-left) collapses onto a small subset of target nodes (orange, bottom-right), producing extreme fan-in and a heavy-tailed k-occurrence distribution (a few targets have very large $N_k(y)$).

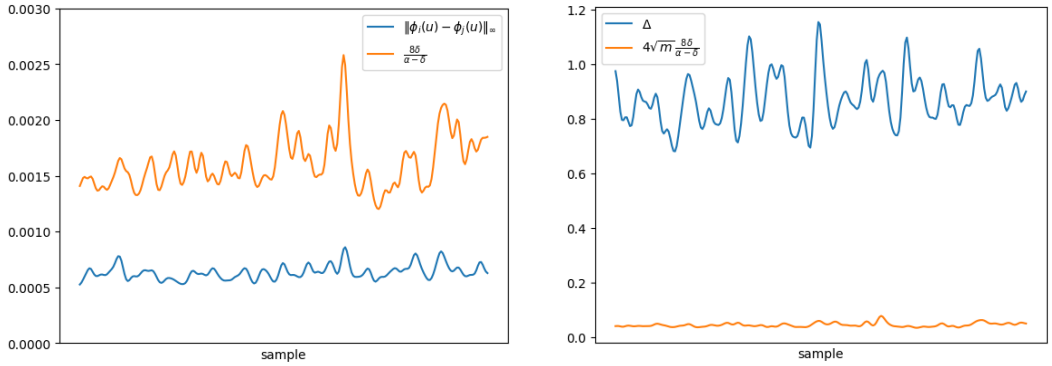
In contrast, Figure 2b (our method) shows *no or substantially reduced hubness*: cross-graph edges are distributed more evenly across target nodes, with markedly fewer star-like structures. Interpreted through k-occurrence, our method reduces the mass captured by the top-m targets and increases coverage of non-hub targets (fewer antihubs), yielding a more balanced bipartite connectivity pattern.

Why this matters. Hubness is harmful for merging because it induces:

- **Many-to-one correspondences** (collapse of alignment diversity),
- **Brittle linking** (small score changes flip large numbers of links),
- **Biased cross-modal context** (message passing dominated by hub targets),
- **Reduced coverage** (many targets become antihubs and never get linked).

How our method mitigates hubness. Naïve NN linking (used by many baselines) treats each source node independently and thus has no mechanism to prevent many-to-one collapse. Our method mitigates hubness by *coupling* cross-graph decisions and encouraging *balanced* correspondences, which directly counteracts hub formation. Concretely, our procedure:

- **Uses context-aware matching rather than raw NN.** By injecting cross-graph context into similarity, we increase discriminability and reduce spurious attraction to globally "central" targets that become hubs.
- **Introduces competition / normalization across candidates.** By discouraging one target from being selected by many sources (implicitly or explicitly), we reduce the heavy tail of $N_k(y)$ and prevent star-like fan-in.



(a) **Fingerprint stability.** Empirical deviation $\|\phi_i(u) - \phi_j(u)\|_\infty$ (blue) versus the theoretical bound $\frac{8\delta}{\alpha-\delta}$ (orange) across sampled nodes.

(b) **NN preservation via margin.** Nearest-neighbor margin Δ (blue) compared to the sufficient threshold $4\sqrt{m}\,\frac{8\delta}{\alpha-\delta}$ (orange).

Figure 3: **Numerical simulation of Proposition 3.3.** Synthetic two-modality embeddings are generated by an orthogonal transform plus small perturbations, and cosine-to-landmark fingerprints are computed in each modality. The predicted stability bound controls the cross-modality fingerprint discrepancy (left), and the sufficient margin condition for nearest-neighbor preservation is satisfied in this regime (right).

D Numerical Simulation of Proposition 3.3

To empirically validate the stability guarantee in Proposition 3.3, we construct a controlled two-modality toy setting that matches the proposition's near-isometry + bounded-noise model. We first sample latent vectors $\{z(u)\}_{u=1}^n \subset \mathbb{R}^d$ (and reuse them as landmark latents $\{z_a\}_{a=1}^m$ for simplicity), then generate two "modalities" by applying an orthogonal transform and adding small perturbations:

$$h^{(i)}(u) = Q_i z(u) + \epsilon_u^{(i)}, \qquad \ell_a^{(i)} = Q_i z_a + \eta_a^{(i)},$$

with Q_i sampled uniformly from the orthogonal group and ϵ, η drawn as small Gaussian noise. We then compute cosine-to-landmark fingerprints $\phi_i(u)_a = \cos\big(h^{(i)}(u), \ell_a^{(i)}\big)$ and evaluate the discrepancy between modalities for each node u.

Verifying the uniform stability bound. For each u, we measure the empirical deviation $\|\phi_i(u) - \phi_j(u)\|_\infty$ and compare it to the theoretical upper bound $\frac{8\delta}{\alpha-\delta}$ from Proposition 3.3. In the simulation we compute δ from the realized perturbation magnitudes and take $\alpha = \|z(u)\|$ per sample (an instance-wise proxy for the global lower-bound assumption $\|z(u)\| \geq \alpha$). As shown in Figure 3a, the observed deviation (blue) stays below the predicted bound (orange) across all samples, illustrating the Lipschitz-type stability of cosine-to-landmark fingerprints under orthogonal transforms and bounded noise.

Checking the sufficient margin condition for NN preservation. Next, we evaluate the margin

$$\Delta = \min_{w \neq v} \|\phi_i(u) - \phi_i(w)\|_2 - \|\phi_i(u) - \phi_i(v)\|_2,$$

where v is a nearest neighbor of u under modality i in fingerprint space. We compare Δ to the sufficient threshold $4\sqrt{m}\,\frac{8\delta}{\alpha-\delta}$ from Proposition 3.3. In Figure 3b, the empirical margins are substantially larger than this threshold in our noise regime, indicating that the proposition's sufficient condition is satisfied and thus nearest-neighbor relations are predicted to be preserved across modalities. While the condition is sufficient (and therefore conservative), the observed gap illustrates that the landmark-fingerprint representation can robustly preserve local neighborhood structure even under small cross-modality perturbations.

E COMPUTATIONAL COMPLEXITY

We analyze the time and memory complexity of Algorithms 1 and 2. Let N be the number of modality graphs, $n_i := |\mathcal{V}_i|$ the number of nodes in graph i, $\ell_i := |\mathcal{L}_i|$ the number of landmarks, d the SSL embedding dimension, and E the (optional) post-compression dimension of $X^{(i)}$ used for OT/EMD distances. We *exclude* the SSL pretraining/encoding cost (GraphMAE) and start after node embeddings H_i are available.

E.1 ALGORITHM 1: LANDMARKS AND TENSOR EMBEDDINGS

(2) Landmark tensor construction. Algorithm 1 constructs the N-way landmark tensor $S \in \mathbb{R}^{\ell_1 \times \cdots \times \ell_N}$ with entries $S_{j_1,\ldots,j_N} = \sum_{r=1}^{d} \prod_{k=1}^{N} L^{(k)}_{j_k,r}$. A direct construction costs

$$T_S \;=\; O\left(d \prod_{k=1}^{N} \ell_k \right), \tag{27}$$

and requires

$$M_S \;=\; O\left(\prod_{k=1}^{N} \ell_k \right) \tag{28}$$

memory to store S explicitly.

(3a) Node-to-landmark similarity matrices. For each graph i, we compute cosine similarities between n_i nodes and ℓ_i landmarks in $\mathbb{R}^d$. This costs

$$T_{\text{sim}} \;=\; \sum_{i=1}^{N} O(n_i\, \ell_i\, d), \qquad M_{\text{sim}} \;=\; \sum_{i=1}^{N} O(n_i\, \ell_i), \tag{29}$$

where the memory term is for storing the matrices $L_S^{(i)} \in \mathbb{R}^{n_i \times \ell_i}$.

(3b) Mode-i tensor–matrix product. If S is materialized, the mode product $X^{(i)} = S \times_i L_S^{(i)}$ produces $X^{(i)} \in \mathbb{R}^{n_i \times \prod_{k \neq i} \ell_k}$ and costs

$$T_{\times} \;=\; \sum_{i=1}^{N} O\left(n_i\, \ell_i \prod_{k \neq i} \ell_k \right), \qquad M_X \;=\; \sum_{i=1}^{N} O\left(n_i \prod_{k \neq i} \ell_k \right), \tag{30}$$

Remarks on scaling. The landmark-tensor stage provides one representation construction for a fixed collection of N modalities. We target the common multimodal regime with a small number of complementary input graphs; the reported evaluations cover two- and three-modality configurations. The implementation used in this work explicitly materializes S (and the corresponding mode products), so its tensor-stage cost includes factors of $\prod_k \ell_k$. Small landmark budgets and downstream dimensionality reduction (to E) provide practical controls for these evaluated settings.

E.2 ALGORITHM 2: OT/EMD EDGE INDUCTION

Algorithm 2 iterates over all modality pairs (i, j), with $m = n_i$ and $n = n_j$.

Cost matrix computation. Computing pairwise distances between embeddings (dimension E after optional compression) costs

$$T_C^{(i,j)} \;=\; O(n_i\, n_j\, E), \qquad M_C^{(i,j)} \;=\; O(n_i\, n_j), \tag{31}$$

to store the dense cost matrix.

EMD solve. The dominant cost is the exact EMD solver. For this transportation LP, the runtime depends on the solver (e.g., network simplex / min-cost flow variants) and is typically superlinear in the number of nonzeros of the cost matrix (Peyré et al., 2019); in the worst case it is polynomial but can be expensive for large $n_i n_j$. All experiments reported in this paper use exact LP EMD to induce cross-graph edges.

Support extraction and sparsity. Extracting edges from the returned exact-LP coupling support is linear in the number of nonzeros. An extreme-point optimal plan satisfies $|\mathrm{supp}(\gamma^{(i,j)})| \le n_i + n_j - 1$ (Lemma 3.1), so the extracted cross-edge set per pair is *sparse*:

$$|\mathcal{E}_{\times}^{(i,j)}| \le n_i + n_j - 1. \tag{32}$$

E.3 Overall complexity

Let $\mathcal{P} = \{(i,j) : 1 \le i < j \le N\}$ denote all modality pairs.

Time. Excluding SSL, the end-to-end time is

$$T_{\text{total}} = \underbrace{\sum_{i=1}^{N} O(n_i\,\ell_i\,d)}_{\text{node–landmark sims}} + \underbrace{O\left(d\prod_{k}\ell_k\right)}_{S\text{ construction}} + \underbrace{\sum_{i=1}^{N} O\left(n_i\,\ell_i\prod_{k\neq i}\ell_k\right)}_{\text{mode products}} + \underbrace{\sum_{(i,j)\in\mathcal{P}}\left(O(n_i n_j E) + T_{\text{EMD}}^{(i,j)}\right)}_{\text{pairwise OT/EMD}}. \tag{33}$$

Here $T_{\text{EMD}}^{(i,j)}$ is the runtime of the exact LP EMD solver. If dimensionality reduction to E is applied, it affects only the OT cost-matrix term $O(n_i n_j E)$.

Memory. The peak memory is dominated by (i) storing node–landmark matrices, (ii) storing S, (iii) the mode-product outputs, and (iv) the OT cost/coupling matrices:

$$M_{\text{total}} = \underbrace{\sum_{i=1}^{N} O(n_i\,\ell_i)}_{\text{store } L_S^{(i)}} + \underbrace{O\left(\prod_{k}\ell_k\right)}_{\text{store } S} + \underbrace{\sum_{i=1}^{N} O\left(n_i\prod_{k\neq i}\ell_k\right)}_{\text{store } X^{(i)}} + \underbrace{\max_{(i,j)\in\mathcal{P}} O(n_i\,n_j)}_{\text{OT cost/coupling}}. \tag{34}$$

In practice, memory can be reduced by (a) streaming OT pairwise (one (i,j) at a time), and (b) avoiding simultaneous storage of all intermediate OT matrices.

F Ablation studies

We report ablated configurations of two pipeline components: (i) the *multi-graph (N-way) landmark tensor context* construction (Algorithm 1), and (ii) *OT/EMD-based edge induction* (Algorithm 2). Results are shown for the tri-modal settings **SEN12MS (SAR+VIS+SWIR)** and **TCGA-BRCA (CNV+Clinical+WSI)**.

Ablated variants. Let **Full (Ours)** denote the complete method (Algorithms 1 and 2). We consider:

- **No multi-graph context ($-$ N-way):** remove Algorithm 1 and compute cross-graph distances directly in the raw SSL embedding space (GraphMAE), i.e., use H_i instead of $X^{(i)}$.
- **No OT coupling ($-$ OT / kNN edges):** keep Algorithm 1 to form context embeddings, but replace OT/EMD edge induction with greedy kNN linking in the resulting embedding space.
- **No context + no OT ($-$ N-way & OT):** remove Algorithm 1 and replace OT/EMD with kNN linking directly on SSL embeddings.

Results. Table 3 reports the full and ablated configurations on each dataset.

Multi-graph context comparison (Algorithm 1). The Full configuration has higher reported scores than *No N-way* on both datasets. On SEN12MS, Acc is 0.7855 for Full and 0.7736 for *No N-way*. On BRCA, Acc is 0.6529 versus 0.6255, and F1 is 0.6593 versus 0.6080. These comparisons characterize the two configurations on the evaluated datasets.

OT/EMD edge-induction comparison (Algorithm 2). The Full configuration also has higher reported scores than *No OT (kNN)* on both datasets. On BRCA, Acc is 0.6529 for Full and 0.5358 for

Table 3: **Ablation study on tri-modal graph merging.** We evaluate (i) SEN12MS *SAR+VIS+SWIR* and (ii) TCGA-BRCA *CNV+Clinical+WSI*. Entries are reported as mean ± standard deviation, using 5 random seeds for SEN12MS and 10 folds for TCGA-BRCA. **Full** is the complete pipeline (multi-graph N-way landmark context + OT/EMD edge induction). **No N-way** removes Algorithm 1 and uses raw SSL embeddings for cross-graph distances. **No OT** replaces Algorithm 2 with greedy kNN cross-graph linking. **No N-way + No OT** removes both components, i.e. raw SSL embeddings plus kNN linking. Best results are **bolded** and marked with *.

VARIANT	ACC	F1	KAPPA	AUC
SEN12MS				
FULL (OURS)	$\mathbf{0.7855 \pm 0.007}^*$	$\mathbf{0.7830 \pm 0.008}^*$	$\mathbf{0.7616 \pm 0.008}^*$	$\mathbf{0.9752 \pm 0.0009}^*$
NO N-WAY	0.7736 ± 0.005	0.7713 ± 0.005	0.7506 ± 0.006	0.9559 ± 0.001
NO OT	0.7646 ± 0.003	0.7622 ± 0.005	0.7384 ± 0.003	0.9709 ± 0.0004
NO N-WAY + NO OT	0.7594 ± 0.008	0.7576 ± 0.007	0.7326 ± 0.009	0.9701 ± 0.0004
TCGA-BRCA				
FULL (OURS)	$\mathbf{0.6529 \pm 0.05}^*$	$\mathbf{0.6593 \pm 0.05}^*$	$\mathbf{0.5863 \pm 0.06}^*$	$\mathbf{0.8890 \pm 0.02}^*$
NO N-WAY	0.6255 ± 0.05	0.6080 ± 0.07	0.5671 ± 0.07	0.8885 ± 0.02
NO OT	0.5358 ± 0.02	0.5169 ± 0.04	0.4947 ± 0.04	0.8461 ± 0.02
NO N-WAY + NO OT	0.5454 ± 0.04	0.5298 ± 0.05	0.5043 ± 0.06	0.8538 ± 0.02

No OT, while AUC is 0.8890 versus 0.8461; on SEN12MS, Acc is 0.7855 versus 0.7646. The two edge-construction procedures differ in their connectivity patterns: kNN is susceptible to hubness and many-to-one matches (Section C), whereas OT/EMD enforces global mass conservation and yields balanced sparse supports.

Combined ablation. On SEN12MS, *No N-way + No OT* has the lowest reported Acc among the ablated configurations (0.7594, compared with 0.7855 for Full). On BRCA, its Acc is 0.5454, compared with 0.6529 for Full; *No OT* reports Acc 0.5358 and AUC 0.8461. Together, these values describe how the evaluated configurations differ when the context construction and edge-induction stages are varied.

Summary: Across the reported ablation configurations, Full has the highest scores on both datasets; the score gaps differ between SEN12MS and BRCA.

G EXTENDED RELATED WORK

Our work sits at the intersection of multimodal graph learning, network alignment, and optimal-transport-based matching. The key distinction is that we perform *unsupervised structural integration*: we do not primarily seek a correspondence as an end goal, but instead construct a sparse set of cross-graph edges for downstream message passing and evaluate it by predictive utility.

Multimodal learning with graphs: fusing representations vs. merging structure. A common approach to multimodal prediction is to train modality-specific encoders and combine modalities through early fusion (feature concatenation), late fusion (logit aggregation), shared latent spaces, or cross-modal attention. When modalities are represented as graphs, many such pipelines retain separate intra-graph message passing and exchange information through fusion or attention modules. In contrast, we make cross-modal interaction *structural*: we explicitly add cross-graph edges so that message passing can propagate evidence between modality graphs inside a single merged computation graph. The distinction is not that fusion cannot support intermediate interaction, but that our method represents selected interactions as a sparse graph topology rather than as a learned fusion or attention operator.

Network alignment and graph matching. Network/graph alignment aims to recover correspondences between node sets, often as a (partial) bijection that preserves structure and/or attributes. Early scalable approaches such as REGAL (Heimann et al., 2018) learn structural representations (optionally using attributes) and align nodes by similarity in the learned space. CONE-Align (Chen

et al., 2020) strengthens neighborhood consistency by learning proximity-preserving embeddings and aligning embedding spaces before matching. More recent methods improve robustness under weak attributes or noisy settings: GradAlign+ (Park et al., 2022) introduces gradual alignment and attribute augmentation, FUGAL (Bommakanti et al., 2024) emphasizes robustness through feature augmentation, and CAPER (Zhu et al., 2022) uses a multilevel coarsen-align-refine strategy for scalability and improved match quality. These methods are natural baselines for our setting because they produce a correspondence signal that can be converted into cross-graph links.

How we differ. Alignment methods optimize a matching objective and are commonly evaluated by correspondence quality. Our method instead constructs sparse cross-graph connectivity for message passing and evaluates it by downstream predictive utility. The edge-induction procedure itself remains label-free; it is intended as a structural alternative to turning alignment outputs into a computation graph, with balanced connectivity, no hubness-driven collapse, and implicit sparsity control.

Optimal transport for correspondence. Optimal transport (OT) provides a principled relaxation of matching by replacing hard correspondences with a soft coupling constrained by global mass conservation. OT-style formulations have been used for alignment, especially when correspondences are partial or uncertain; PARROT (Zeng et al., 2023), for example, proposes a regularized OT approach that incorporates positional information. OT is appealing in our setting because its coupling constraints mitigate many-to-one hubness and yield globally consistent correspondences.

How we differ. We use OT not merely to output a correspondence, but as an *edge generator* for graph merging: cross-graph edges are defined as the support of the EMD coupling. This choice has two consequences that are central to our framing: (i) it produces a sparse cross-edge scaffold suitable for message passing, and (ii) because EMD is a linear program, extreme-point solutions admit an *implicit per-pair hard budget* on the number of nonzeros, avoiding ad hoc thresholding.

Seeded and learning-based matching. When anchor pairs are available, learning-based methods can exploit seeds to propagate correspondence information through graph neighborhoods. SeedGNN (Yu et al., 2023) is an example of a GNN-based seeded matching approach that learns over node-pair representations conditioned on known correspondences. These methods highlight the value of message passing for correspondence reasoning.

How we differ. We consider the common multimodal regime where explicit node-level seeds are unavailable and node semantics can differ across modalities. Rather than learning a pairwise classifier over candidate matches, we construct a merged graph that enables cross-graph message passing for the *downstream* task.

Unrestricted alignment, evaluation protocols, and multi-graph perspectives. Practical alignment often deviates from idealized assumptions (different node sets, noise, weak attributes), motivating unrestricted settings and careful evaluation. Skitsas et al. (2023) provide a broad empirical evaluation of unrestricted alignment algorithms, while FUGAL (Bommakanti et al., 2024) targets robustness through feature augmentation. Beyond pairwise alignment, higher-order and multi-graph viewpoints are increasingly studied, including higher-order topological consistency (Sun et al., 2023) and feasibility thresholds for multi-graph alignment (Vassaux & Massoulié, 2025). These perspectives are relevant to multimodal scenarios where more than two graphs are present.

How we differ. While multi-graph alignment aims to jointly recover correspondences across several graphs, we construct a merged structure for downstream message passing and evaluate it by predictive utility. For small, fixed modality collections, our Nyström-style landmarks and low-rank multiway landmark tensor capture multi-modal compatibility patterns before inducing pairwise OT couplings, bridging multi-graph structure with practical edge construction.

Summary. In short, prior alignment and OT-based matching methods provide powerful tools for estimating cross-graph correspondences, but they are typically evaluated by match quality rather than downstream predictive utility. We recast the problem as **unsupervised structural graph merging**: for a small, fixed collection of $N \geq 2$ modalities, learn a representation that integrates multiway context through landmark fingerprints and a low-rank *multiway* tensor, then induce sparse, balanced cross-graph edges via pairwise OT so that cross-modal interaction becomes *structural* within a single message-passing model.

Algorithm 2 Cross-graph edge construction via Earth Mover's Distance (EMD): support extraction

Input: per-graph node embeddings $\{X^{(i)}\}_{i=1}^{N}$, node sets $\{\mathcal{V}_i\}$, distance $d(\cdot,\cdot)$, EMD solver
Output: cross-graph edge set $\mathcal{E}_\times$
Initialize $\mathcal{E}_\times \leftarrow \emptyset$
for all pairs (i,j) with $1 \le i < j \le N$ **do**
  Define uniform node measures $P_i(u) = 1/|\mathcal{V}_i|$ and $P_j(v) = 1/|\mathcal{V}_j|$
  Solve EMD coupling:

$$\gamma^{(i,j)} \leftarrow \arg\min_{\gamma \in \Pi(P_i,P_j)} \sum_{u\in\mathcal{V}_i}\sum_{v\in\mathcal{V}_j} \gamma(u,v)\, d\Big(X^{(i)}[u], X^{(j)}[v]\Big)$$

  Add cross edges as the support of $\gamma^{(i,j)}$:
    $\mathcal{E}_\times \leftarrow \mathcal{E}_\times \cup \{(u,v) : \gamma^{(i,j)}(u,v) > 0\}$
end for

H Algorithms

This appendix algorithm makes explicit how pairwise transport plans are computed from the learned graph-specific embeddings and then converted into sparse cross-graph edges through support extraction.

The procedure is applied independently to each modality pair after the shared embedding stage in Algorithm 1. Together, the two algorithms define the full merging pipeline used throughout the experiments and ablations.

I Full Benchmark Results (Including Unimodal Baselines)

Tables 4 and 5 report the complete benchmark suite for SEN12MS and TCGA-BRCA, including unimodal baselines alongside all multi-modal configurations and alignment/merging methods. These tables are provided for completeness and to contextualize the gains from cross-graph merging relative to (i) single-modality performance and (ii) representative alignment-based baselines.

When does merging help most? The strongest reported results occur in the tri-modal settings. On SEN12MS, the tri-modal configuration yields the best overall performance among the evaluated methods. On BRCA, the proposed method likewise achieves its strongest reported result in the full CNV+Clinical+WSI configuration. The full tables below show that pairwise outcomes vary across modality combinations, reinforcing that the value of induced cross-graph topology depends on the specific graphs and task. This variation motivates adaptive extensions (e.g., confidence-aware transport, selective edge pruning, or per-pair coupling modulation) that introduce cross-graph connectivity when the learned topology is sufficiently informative for the downstream task.

J Limitations

Our results suggest that unsupervised structural graph merging is useful when modality-specific graphs contain complementary information and share enough latent structure for meaningful cross-graph coupling. This condition is not guaranteed. When modalities encode weakly related signals, have incompatible node semantics, or contain severe registration and preprocessing errors, the induced cross-graph edges can propagate noise rather than useful evidence. The pairwise results in Tables 4 and 5 illustrate this boundary condition: gains are strongest in some modality combinations, while other pairs benefit less or favor an alignment baseline on selected metrics.

Scalability. The method is designed for the small, fixed modality collections considered in this work. Within this regime, its N-way landmark construction incorporates information from all available modalities before pairwise OT/EMD edge induction. The evaluated implementation explicitly materializes the multiway landmark tensor, so its tensor-stage cost depends on the product of the

Table 4: Performance on SEN12MS. Entries are reported as mean ± standard deviation over 5 random seeds. Best results are **bolded** and annotated with asterisk *.

MODEL	ACC	F1	KAPPA	AUC
SAR	0.6044 ± 0.006	0.6005 ± 0.009	0.5605 ± 0.007	0.9252 ± 0.001
VIS	0.6467 ± 0.008	0.6426 ± 0.008	0.6074 ± 0.009	0.9375 ± 0.002
SWIR	0.7113 ± 0.005	0.7079 ± 0.005	0.6792 ± 0.006	0.9590 ± 0.001
SAR+VIS				
REGAL	0.6986 ± 0.009	0.6974 ± 0.007	0.6652 ± 0.01	0.9529 ± 0.002
CONE-ALIGN	0.7004 ± 0.006	0.6994 ± 0.006	0.6672 ± 0.007	0.9550 ± 0.001
GRADALIGN+	0.6989 ± 0.008	0.6951 ± 0.007	0.6655 ± 0.009	0.9545 ± 0.001
CAPER	0.7157 ± 0.009	0.7173 ± 0.008	0.6842 ± 0.01	0.9596 ± 0.0007
FUGAL	0.7178 ± 0.008	0.7156 ± 0.01	0.6887 ± 0.009	0.9433 ± 0.002
PARROT	0.7068 ± 0.012	0.7046 ± 0.0127	0.6742 ± 0.0133	0.9558 ± 0.0014
OURS	$\mathbf{0.7391 \pm 0.006}^*$	$\mathbf{0.7381 \pm 0.005}^*$	$\mathbf{0.7102 \pm 0.007}^*$	$\mathbf{0.9623 \pm 0.001}^*$
SAR+SWIR				
REGAL	0.7472 ± 0.009	0.7454 ± 0.008	0.7192 ± 0.01	0.9675 ± 0.001
CONE-ALIGN	0.7562 ± 0.007	0.7518 ± 0.007	0.7291 ± 0.007	0.9689 ± 0.001
GRADALIGN+	0.7541 ± 0.006	0.7516 ± 0.006	0.7267 ± 0.007	0.9684 ± 0.0008
CAPER	0.7636 ± 0.004	0.7618 ± 0.006	0.7373 ± 0.005	0.9709 ± 0.0006
FUGAL	0.7494 ± 0.005	0.7489 ± 0.005	0.7238 ± 0.005	0.9532 ± 0.001
PARROT	0.7623 ± 0.0062	0.7608 ± 0.0066	0.7359 ± 0.0068	0.9694 ± 0.0015
OURS	$\mathbf{0.7674 \pm 0.009}^*$	$\mathbf{0.7645 \pm 0.009}^*$	$\mathbf{0.7416 \pm 0.01}^*$	$\mathbf{0.9713 \pm 0.001}^*$
VIS+SWIR				
REGAL	0.7380 ± 0.007	0.7350 ± 0.006	0.7088 ± 0.008	0.9636 ± 0.001
CONE-ALIGN	0.7317 ± 0.01	0.7284 ± 0.01	0.7020 ± 0.01	0.9647 ± 0.001
GRADALIGN+	0.7298 ± 0.006	0.7287 ± 0.006	0.6998 ± 0.006	0.9628 ± 0.001
CAPER	0.7515 ± 0.01	0.7497 ± 0.009	0.7240 ± 0.01	$\mathbf{0.9692 \pm 0.001}^*$
FUGAL	0.7519 ± 0.004	0.7508 ± 0.003	0.7266 ± 0.004	0.9530 ± 0.001
PARROT	0.7338 ± 0.0079	0.7294 ± 0.0087	0.7042 ± 0.0088	0.9645 ± 0.0009
OURS	$\mathbf{0.7568 \pm 0.005}^*$	$\mathbf{0.7527 \pm 0.005}^*$	$\mathbf{0.7298 \pm 0.005}^*$	0.9687 ± 0.0005
SAR+VIS+SWIR				
REGAL	0.7544 ± 0.009	0.7520 ± 0.007	0.7271 ± 0.01	0.9697 ± 0.0007
CONE-ALIGN	0.7637 ± 0.01	0.7622 ± 0.01	0.7374 ± 0.01	0.9706 ± 0.0007
GRADALIGN+	0.7548 ± 0.01	0.7536 ± 0.01	0.7276 ± 0.01	0.9706 ± 0.001
CAPER	0.7806 ± 0.007	0.7775 ± 0.008	0.7562 ± 0.008	0.9749 ± 0.001
FUGAL	0.7728 ± 0.004	0.7714 ± 0.005	0.7498 ± 0.005	0.9565 ± 0.0007
PARROT	0.7637 ± 0.0125	0.7636 ± 0.0112	0.7374 ± 0.0138	0.9715 ± 0.0011
OURS	$\mathbf{0.7855 \pm 0.007}^*$	$\mathbf{0.7830 \pm 0.008}^*$	$\mathbf{0.7616 \pm 0.008}^*$	$\mathbf{0.9752 \pm 0.0009}^*$

landmark budgets, and the OT/EMD stage requires pairwise cost and coupling matrices. Extending structural merging to larger modality collections is a promising direction for future work. Approximate OT solvers, sparsified candidate sets, streaming implementations, adaptive edge budgets, and direct low-rank-factor contractions provide practical avenues to investigate.

Sensitivity to graph construction and landmarks. The approach assumes that the per-modality graphs are informative for the downstream task before merging. Poor graph construction, noisy node features, or unrepresentative landmarks can degrade the learned context embeddings and consequently the transport plan. In this work we use random landmark sampling for simplicity; more robust landmark selection, uncertainty-aware edge pruning, or validation-driven coupling strength could improve reliability, especially when modality quality varies.

Evaluation scope. We evaluate on two multimodal graph classification benchmarks: SEN12MS and TCGA-BRCA. These datasets cover different domains, but they do not exhaust the range of settings where graph merging might be applied. Our study is centered on constructing cross-graph topology from separate modality graphs; accordingly, it compares alignment-derived edge construction rather

Table 5: Performance on TCGA-BRCA. Entries are reported as mean ± standard deviation over 10 folds. Best results are **bolded** and annotated with asterisk *.

MODEL	ACC	F1	KAPPA	AUC
CNV	0.4328 ± 0.04	0.4001 ± 0.04	0.3610 ± 0.07	0.7836 ± 0.02
CLINICAL	0.5677 ± 0.04	0.5849 ± 0.03	0.4738 ± 0.07	0.8341 ± 0.03
WSI	0.4440 ± 0.03	0.4363 ± 0.04	0.3658 ± 0.07	0.7944 ± 0.04
CNV+CLINICAL				
REGAL	0.5501 ± 0.04	0.5406 ± 0.05	0.4987 ± 0.05	0.8583 ± 0.02
CONE-ALIGN	0.5433 ± 0.04	0.5288 ± 0.07	0.4954 ± 0.07	0.8553 ± 0.02
GRADALIGN+	0.5324 ± 0.03	0.5238 ± 0.07	$\mathbf{0.5014 \pm 0.05}^*$	0.8544 ± 0.02
CAPER	0.5231 ± 0.03	0.5096 ± 0.04	0.4868 ± 0.05	0.8542 ± 0.01
FUGAL	0.5134 ± 0.0398	0.4971 ± 0.0347	0.4828 ± 0.0543	0.8549 ± 0.0214
PARROT	$\mathbf{0.5773 \pm 0.0464}^*$	$\mathbf{0.5632 \pm 0.0405}^*$	0.4940 ± 0.0355	$\mathbf{0.8624 \pm 0.0270}^*$
OURS	0.5336 ± 0.04	0.5216 ± 0.04	0.4908 ± 0.04	0.8580 ± 0.01
CNV+WSI				
REGAL	$\mathbf{0.4804 \pm 0.04}^*$	$\mathbf{0.4650 \pm 0.04}^*$	$\mathbf{0.4310 \pm 0.08}^*$	0.7885 ± 0.03
CONE-ALIGN	0.4700 ± 0.02	0.4470 ± 0.02	0.4112 ± 0.05	0.7900 ± 0.03
GRADALIGN+	0.4617 ± 0.04	0.4426 ± 0.03	0.3986 ± 0.06	0.7737 ± 0.04
CAPER	0.4722 ± 0.03	0.4500 ± 0.04	0.4185 ± 0.06	0.7912 ± 0.04
FUGAL	0.4681 ± 0.0366	0.4515 ± 0.0468	0.4135 ± 0.0615	0.7746 ± 0.0351
PARROT	0.4604 ± 0.0314	0.4289 ± 0.0270	0.3879 ± 0.0643	0.7783 ± 0.0332
OURS	0.4560 ± 0.03	0.4318 ± 0.03	0.3807 ± 0.06	$\mathbf{0.7971 \pm 0.03}^*$
CLINICAL+WSI				
REGAL	0.4965 ± 0.03	0.4772 ± 0.03	0.4379 ± 0.04	0.8331 ± 0.03
CONE-ALIGN	0.5295 ± 0.05	0.5159 ± 0.06	0.4708 ± 0.06	0.8378 ± 0.02
GRADALIGN+	0.5253 ± 0.03	0.5113 ± 0.04	0.4730 ± 0.05	0.8309 ± 0.02
CAPER	0.5311 ± 0.06	0.5245 ± 0.07	0.4609 ± 0.09	0.8332 ± 0.03
FUGAL	0.5115 ± 0.0513	0.4959 ± 0.0552	0.4465 ± 0.0755	0.8236 ± 0.0371
PARROT	0.5274 ± 0.0503	0.5127 ± 0.0678	0.4543 ± 0.0546	0.8174 ± 0.0278
OURS	$\mathbf{0.5736 \pm 0.04}^*$	$\mathbf{0.5597 \pm 0.07}^*$	$\mathbf{0.4938 \pm 0.04}^*$	$\mathbf{0.8620 \pm 0.02}^*$
CNV+CLINICAL+WSI				
REGAL	0.5303 ± 0.04	0.5089 ± 0.05	0.4945 ± 0.05	0.8585 ± 0.02
CONE-ALIGN	0.5581 ± 0.03	0.5350 ± 0.02	0.5364 ± 0.04	0.8621 ± 0.02
GRADALIGN+	0.5563 ± 0.04	0.5348 ± 0.03	0.5149 ± 0.05	0.8607 ± 0.02
CAPER	0.5418 ± 0.03	0.5298 ± 0.04	0.5072 ± 0.05	0.8657 ± 0.02
FUGAL	0.5299 ± 0.0254	0.5099 ± 0.0226	0.4985 ± 0.0536	0.8484 ± 0.0286
PARROT	0.5204 ± 0.0244	0.4995 ± 0.0323	0.4798 ± 0.0539	0.8352 ± 0.0302
OURS	$\mathbf{0.6529 \pm 0.05}^*$	$\mathbf{0.6593 \pm 0.05}^*$	$\mathbf{0.5863 \pm 0.06}^*$	$\mathbf{0.8890 \pm 0.02}^*$

than representation-level fusion, gated fusion, or dense cross-attention architectures. We also do not evaluate node-level prediction tasks, dynamic graphs, very large graphs, missing-modality regimes, or cases with explicit ground-truth node correspondences. The reported conclusions should therefore be interpreted as evidence for task-level graph classification under the studied graph construction pipelines, not as a universal guarantee for all multimodal graph problems.

Interpretability and uncertainty. OT produces balanced couplings and sparse edge supports. The resulting cross-graph edges summarize geometric, structural, and feature-based affinity in the learned representation, and enable downstream message passing. Their interpretation is therefore tied to the graph-construction and embedding choices; future work can study how these learned affinities relate to domain-specific cross-modal relationships and can attach confidence estimates to support downstream use.

Data and societal considerations. The method inherits the limitations of the datasets and preprocessing pipelines used to construct the input graphs. In biomedical settings such as TCGA-BRCA, deployment would require careful validation across institutions, acquisition protocols, and demo-

graphic subgroups before any clinical use. In remote sensing settings, performance may shift under seasonal, geographic, or sensor-distribution changes. We do not make claims about fairness, privacy preservation, or clinical decision support; these require dedicated evaluation beyond the scope of this work.

K Compute Resources

All reported experiments were run on GPU workers equipped with NVIDIA A10 GPUs. Each A10 GPU provides 24GB of GPU memory. For SEN12MS, one complete experimental run corresponds to evaluating a method and modality configuration across the 5 random seeds reported in the tables. For TCGA-BRCA, one complete experimental run corresponds to evaluating a method and modality configuration across the 10 cross-validation folds reported in the tables. In both benchmarks, a complete repeated evaluation of this form took approximately one day of wall-clock time.

The total compute required to reproduce the full benchmark therefore scales with the number of methods and modality configurations in Tables 4 and 5. Pairwise and tri-modal configurations were run separately, and preliminary or diagnostic experiments outside the reported benchmark suite required additional compute beyond these reproduction estimates.