UniOMA: Unified Optimal-Transport Multi-Modal Structural Alignment for Robot Perception

Xinrui Zu[1], Kevin Sebastian Luck[1], Shujian Yu[1]
[1]Vrije Universiteit Amsterdam, The Netherlands

***Abstract*—Contrastive objectives such as InfoNCE align multi-modal representations at the instance level but are unable to keep intra-modal geometries, which is called a structural alignment gap. We propose UniOMA, a multimodal structural alignment method using Gromov–Wasserstein (GW) barycenter regularizer to align each modality to a shared structural consensus, scaling linearly to 3+ modalities. Experiments on five robotic benchmarks (vision, force, depth, audio, tactile, proprioception) show consistent improvements in downstream tasks like regression, classification, and cross-modal retrieval.**

***Index Terms*—Robot perception, Multimodal alignment, Gromov-Wasserstein distance.**

I. Introduction

The integration of information from diverse sources or modalities has received increasing attention across a wide range of AI applications, including image/video/text generation (1; 2), healthcare (3), autonomous systems (4), and scientific discovery (5). Recent advances in contrastive self-supervised learning (CSSL) (6; 7; 8; 9), particularly those leveraging InfoNCE losses (10), have shown strong performance in aligning heterogeneous modalities into a shared representation space (11), enabling zero-shot cross-modal retrieval, transfer, generation, and completion (11; 12; 13; 14).

Despite their empirical success, existing CSSL methods fundamentally cast multimodal alignment as a binary classification problem (15): paired samples are encouraged to be close, while unpaired samples are pushed apart, without explicitly modeling the continuous intra-modal geometry. As a result, multimodal representations may appear statistically aligned at the population level, yet disagree in how samples relate to one another within each modality, a phenomenon we call the *structural alignment gap* (16). This gap is not an implementation artifact but a theoretical limitation of InfoNCE-style objectives: as lower bounds on mutual information, they are invariant to structure-distorting transformations that preserve structural correspondence (17; 18).

Fig. 1 illustrates this gap on a synthetic example: two modalities with near-identical InfoNCE similarity matrices can still have incompatible intra-modal geometries. The gap is particularly critical in robotics, where sensor streams are neither i.i.d. nor structureless: trajectories form subclusters (19), contact events induce discontinuities (20; 21), and proprioceptive signals follow physical constraints (22; 23). Preserving pointwise correspondences alone is therefore insufficient: effective multimodal representations must also align the relational structure of observations across modalities.

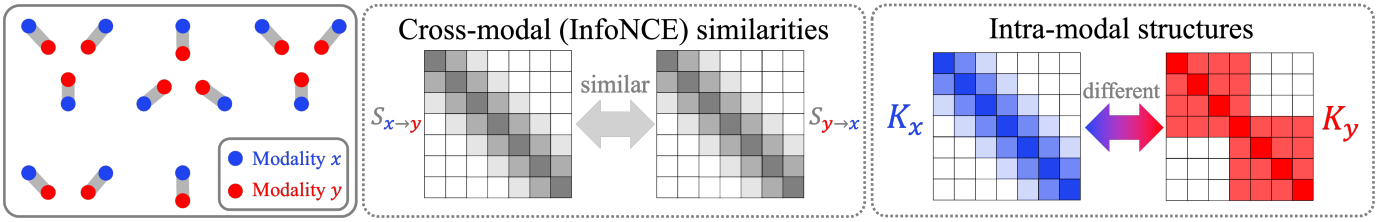
To close this gap, we introduce **UniOMA**—a **Uni**fied **O**ptimal-transport **M**ulti-modal structural **A**lignment framework that augments contrastive learning with a structure-aware regularizer based on Gromov–Wasserstein (GW) distances and barycenters (24; 25). Each modality is encoded as a metric space via intra-modal similarity matrices; a dynamic GW barycenter then forms the structural consensus, and each modality is softly aligned to it through weighted GW distances with end-to-end-learned modality weights. By replacing $O(M^2)$ pairwise couplings with $O(M)$ barycentric alignments, UniOMA scales naturally to three or more modalities.

Our main contributions are:

- C1 We formalize the structural alignment gap and show why InfoNCE-style objectives fail to preserve intra-modal geometry.
- C2 We propose UniOMA, a GW-barycenter regularizer that scales linearly to 3+ modalities.
- C3 We evaluate on five robotic benchmarks (including a new 6-DoF-manipulator dataset) across vision, audio, tactile, force, and proprioception, showing consistent gains in downstream tasks and structural consistency.

II. Method

Related work. A representative contrastive alignment, CLIP (11), trains modality-specific encoders with a symmetrized InfoNCE objective to identify the correct cross-modal pair among N candidates, and lower-bounds mutual information (17; 18). Pairwise extensions to three or more modalities (CMC (12; 13; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36)) sum all pairwise CLIP losses, neglecting higher-order dependencies. Symile (37) introduces triple-wise InfoNCE objectives; GRAM (38) replaces the dot-product similarity with the Gramian volume spanned by embeddings; CoMM (39) and TRIANGLE (40) further strengthen cross-modal matching. (41) introduces a standard optimal-transport (OT) regularizer over embedding distributions. All of these operate at the instance level or the embedding distribution; none explicitly align intra-modal relational geometry. In robotics, multimodal learning has focused primarily on fusion (42; 43; 44; 45; 46; 47); alignment under physically grounded sensors (force, tactile, proprioception) remains underexplored (19; 48; 49; 50), motivating alignment methods that preserve intra-modal geometry.



(a) 2-modality samples (b) InfoNCE view (c) Structure view

Fig. 1: **Structural alignment gap.** (a) Two modalities (x, y) with strong instance-level correspondence but different intra-modal structures. (b) Their cross-modal InfoNCE matrices $S_{x\to y}, S_{y\to x}$ are nearly identical. (c) Yet intra-modal kernels differ: K_y shows cluster structure absent from K_x. High InfoNCE alignment does not imply structural consistency.

A. Problem Statement

Let $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(M)}$ denote M modalities. The goal of multimodal alignment is to learn modality-specific encoders $f^{(m)} : \mathcal{X}^{(m)} \to \mathbb{R}^d$ that project inputs $\mathbf{x}^{(m)}$ into a shared latent space $\mathbf{z}^{(m)} = f^{(m)}(\mathbf{x}^{(m)})$, such that embeddings of the same underlying instance across modalities map to nearby latent vectors, i.e., $\mathbf{z}_i^{(1)} \approx \cdots \approx \mathbf{z}_i^{(M)}$. In robotics, $\mathcal{X}^{(m)}$ may include vision, audio, force-torque, proprioception, tactile sensing, and environment states; aligning them enables cross-modal reasoning, zero-shot transfer, and modality completion.

B. Gromov–Wasserstein Distance

The Gromov–Wasserstein (GW) distance (24; 25) extends optimal transport (OT) (51) to distributions lying in heterogeneous spaces, replacing a direct cross-modal cost with a relational cost that only requires two intra-modal similarity matrices.

Definition 1 (Gromov-Wasserstein Distance)**.** Let $\mathcal{X}_{d_\mathbf{x},\mu}$ and $\mathcal{Y}_{d_\mathbf{y},\nu}$ be two metric–measure spaces, with distance metrics $d_\mathbf{x}, d_\mathbf{y}$ and probability measures μ, ν. The GW distance is

$$d_{gw}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}^2\times\mathcal{Y}^2} c(d_\mathbf{x}(\mathbf{x},\mathbf{x}'), d_\mathbf{y}(\mathbf{y},\mathbf{y}'))\, d\pi(\mathbf{x},\mathbf{y}) d\pi(\mathbf{x}',\mathbf{y}')$$

where $c(\cdot,\cdot)$ measures the discrepancy between sample pairs.

Minimizing GW aligns distributions by matching relational geometry rather than raw coordinates, which is essential when modalities (e.g., vision and force) live in incomparable spaces. For discrete samples with uniform marginals, we use the empirical form (24; 25):

Theorem 1 (Empirical GW Distance)**.** *Let* $\mathbf{K_x} \in \mathbb{R}^{I\times I}$ *and* $\mathbf{K_y} \in \mathbb{R}^{J\times J}$ *be similarity matrices. Then* $\hat{d}_{gw}(\mathbf{K_x},\mathbf{K_y}) := \max_{\mathbf{T}\in\Pi(\hat{\mathbf{p}}_\mathbf{x},\hat{\mathbf{p}}_\mathbf{y})} \mathrm{tr}(\mathbf{K_x}^\top \mathbf{T}^\top \mathbf{K_y} \mathbf{T})$.

In practice, we estimate $\mathbf{T}^*$ via a Frank–Wolfe / network-simplex OT solver, and compute

$$\hat{d}_{gw}(\mathbf{K_x},\mathbf{K_y}) = \mathrm{tr}(\mathbf{K_x}^\top \mathbf{T}^{*\top} \mathbf{K_y} \mathbf{T}^*). \tag{1}$$

C. Structural Consensus and UniOMA Objective

We treat each modality $\mathcal{X}^{(m)}$ as a metric space and represent its geometry via a valid kernel matrix $\mathbf{K}_\mathbf{x}^{(m)} \in \mathbb{R}^{N_m\times N_m}$, where $(\mathbf{K}_\mathbf{x}^{(m)})_{ij} = \mathrm{sim}(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})$. For visual signals we embed inputs with a pretrained encoder and compute RBF similarities; for time-series (e.g., force-torque) we adopt a time-series clustering kernel (TCK; (52)).

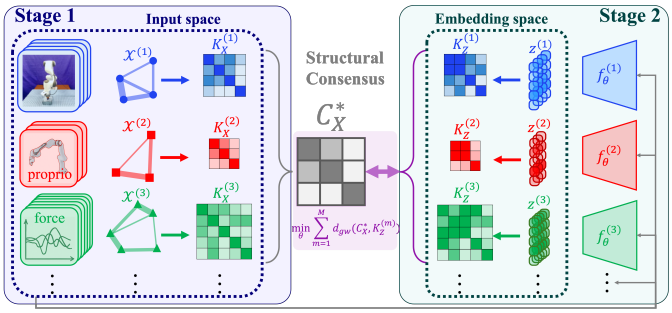


Fig. 2: **UniOMA in two stages.** *Stage 1* (left): for each modality $\mathcal{X}^{(m)}$ we form an input-space similarity matrix $\mathbf{K}_\mathbf{x}^{(m)}$ and estimate a GW barycenter $\mathbf{C}_\mathbf{x}^*$ as the *structural consensus*. *Stage 2* (right): encoders produce embeddings $\mathbf{z}^{(m)}$ inducing $\mathbf{K}_\mathbf{z}^{(m)}$, which are aligned to the consensus by minimizing $\sum_m \lambda_m\, d_{gw}(\mathbf{C}_\mathbf{x}^*, \mathbf{K}_\mathbf{z}^{(m)})$.

Definition 2 (Structural Consensus of Multimodal Data)**.** Given intra-modal kernel matrices $\{\mathbf{K}_\mathbf{x}^{(m)}\}_{m=1}^M$, the structural consensus is

$$\mathbf{C}_\mathbf{x}^* = \arg\min_{\mathbf{C_x}\in\mathcal{M}} \sum_{m=1}^{M} \lambda_m \cdot d_{gw}(\mathbf{C_x}, \mathbf{K}_\mathbf{x}^{(m)}), \tag{2}$$

where $\mathcal{M}$ is the space of symmetric positive definite (SPD) matrices, and $\{\lambda_m\}$ are learnable modality weights with $\lambda_m \geq 0, \sum_m \lambda_m = 1$.

$\mathbf{C}_\mathbf{x}^*$ is estimated via iterative barycenter updates (weighted-average couplings, (25)). UniOMA then augments a standard contrastive term with a structure-aware regularizer:

$$\mathcal{L}_{\text{UniOMA}}(\theta) = \mathcal{L}_\text{c}(\theta) + \alpha \sum_{m=1}^{M} \lambda_m \cdot d_{gw}(\mathbf{C}_\mathbf{x}^*, \mathbf{K}_\mathbf{z}^{(m)}), \tag{3}$$

where $\mathbf{K}_\mathbf{z}^{(m)}$ is the embedding-space similarity matrix of $\mathbf{z}^{(m)} = f_\theta^{(m)}(\mathbf{x}^{(m)})$. The scalar α balances contrastive discrimination and structural coherence, and learnable $\{\lambda_m\}$

quantify each modality's contribution. The full training procedure (stage 1: consensus estimation; stage 2: alignment update) is summarized in Alg. 1.

Algorithm 1 UniOMA Training($\{\mathcal{X}^{(m)}\}_{m=1}^{M}, \gamma, \alpha$)

Input: Multimodal dataset $\{\mathcal{X}^{(m)}\}_{m=1}^{M}$, learning rate γ, structural weight α

Initialize $\{f^{(m)}(\cdot)\}_{m=1}^{M}$, modality weights $\{\lambda_m\}_{m=1}^{M}$. **while** *not converged* **do**

```
// Stage 1: consensus estimation
```

Sample batch $\{\mathbf{x}_i^{(m)}\}_{i=1}^{N_m}$ for each modality **for** $m \leftarrow 1$ **to** M **do**

  Compute $\mathbf{K}_{\mathbf{x}}^{(m)}$ for the batch $\{\mathbf{x}_i^{(m)}\}_{i=1}^{N_m}$

Estimate the structural consensus $\mathbf{C}_{\mathbf{x}}^{*}$ via iterative GW barycenter updates

```
// Stage 2: alignment update
```

$\mathbf{z}_i^{(m)} \leftarrow f_\theta^{(m)}(\mathbf{x}_i^{(m)})$ for all i, m **for** $m \leftarrow 1$ **to** M **do**

  $\mathbf{T}^{(m)^*} \leftarrow \texttt{OTPlan}(\mathbf{C}_{\mathbf{x}}^{*}, \mathbf{K}_{\mathbf{z}}^{(m)})$ via Frank–Wolfe Compute $\hat{d}_{gw}(\mathbf{C}_{\mathbf{x}}^{*}, \mathbf{K}_{\mathbf{z}}^{(m)})$ via Eq. 1

Compute the contrastive learning loss $\mathcal{L}_{\text{c}}$ $\mathcal{L}_{\text{UniOMA}}(\theta) \leftarrow \mathcal{L}_{\text{c}}(\theta) + \alpha \sum_{m=1}^{M} \lambda_m \hat{d}_{gw}(\mathbf{C}_{\mathbf{x}}^{*}, \mathbf{K}_{\mathbf{z}}^{(m)})$ $\theta \leftarrow \theta - \gamma \nabla_\theta \mathcal{L}_{\text{UniOMA}}$ $\lambda_m \leftarrow \lambda_m - \gamma \nabla_{\lambda_m} \mathcal{L}_{\text{UniOMA}}$ for $m = 1, \ldots, M$

return $\{f_\theta^{(m)}\}_{m=1}^{M}, \{\lambda_m\}_{m=1}^{M}$

Why this design? (1) **scalable** to $M \geq 3$: aligning every modality to one consensus avoids $O(M^2)$ pairwise couplings. (2) **flexible** to heterogeneous and asynchronous modalities: GW compares intra-modal similarity matrices, not raw coordinates. (3) GW barycenters naturally handle unequal sample counts across modalities, which we validate in Sec. III-A.

Theoretical guarantee. A natural concern is whether a shared consensus collapses information-rich modalities to poorer ones. Near the barycenter, the GW gradient is positively colinear with that of the Dirichlet energy $\text{tr}(\mathbf{Z}^\top \mathbf{L}^* \mathbf{Z})$ induced by the consensus Laplacian $\mathbf{L}^*$; under $\|\mathbf{Z}\|_F^2 = c$, this is spectrally biased toward the low-frequency eigenmodes of $\mathbf{L}^*$ (53; 54; 55). Hence the regularizer constrains only the shared low-frequency structure, leaving higher-frequency modality-specific information intact for contrastive discrimination.

III. Experiments

Tasks and datasets. We evaluate UniOMA on: (i) **VFD** (Vision–Force–Depth) from Vision&Touch (56; 57): next-step end-effector orientation regression (4D) and binary trajectory-pair discrimination; (ii) **VFP** (Vision–Force–Proprioception): next-step contact prediction (binary); (iii) **MuJoCo Push** (22; 58): next-step 2-D object position regression from gray-scale image, current force, and end-effector pose; (iv) **VAT** (Vision–Audio–Tactile) from ObjectFolder 2.0 (48; 59): cross-modal retrieval with direction-specific MAP; and (v) **VIP** (Vision–IMU–Proprioception): we constructed a real-world dataset collected on a 6-DoF manipulator: current end-effector position regression (3D) from RGB, joint angles, and normalized IMU signals. To assess scalability, we also include a 4–7 modality trajectory-consistency classification on Vision&Touch (Sec. III-0d).

Implementation. Encoders, optimizer, temperature, and schedules are shared across methods (fusion heads differ in CoMM). We compute input-space kernels $\{\mathbf{K}_{\mathbf{x}}^{(m)}\}$ (RBF for images with tuned γ; TCK for time-series/force; RBF elsewhere) and estimate batch-wise $\mathbf{C}_{\mathbf{x}}^{*}$ with iterative barycenter updates. Hyperparameters, TCK settings, and convergence diagnostics are omitted for space.

Comparisons on downstream tasks. We compare UniOMA against Pairwise (CMC) (26), Symile (37), GRAM (38), OTLC (41), TRIANGLE (40), and CoMM (39), with matched optimizer, batch size, temperature, and epochs. For the contrastive baselines, we additionally report "+GW" plug-in variants. Results in Tables I and III show that UniOMA consistently improves over contrastive baselines across regression, classification, and retrieval tasks, with all task-best entries being UniOMA variants. The rare cases where a baseline slightly exceeds its GW variant trade a minor contrastive loss for structural coherence.

Efficiency and scalability. We evaluate scalability on trajectory-consistency classification with up to seven modalities. As M grows, pairwise contrastive and OT-based baselines degrade while UniOMA consistently achieves the highest accuracy (Table II). Because UniOMA aligns each modality independently to one consensus, its complexity is $O(M)$ rather than $O(M^2)$; wall-clock time per epoch grows approximately linearly and becomes strictly faster than the pairwise+OT baseline for $M \geq 6$, with identical peak memory. The GW barycenter converges stably with a small number of solver iterations ($T_{\max} = 5$).

A. Modality Weights and Ablation

UniOMA learns weights $\{\lambda_m\}$ quantifying each modality's contribution to the structural consensus (Fig. 3(a–d)). Vision dominates VAT retrieval, proprioception dominates VFP contact, and depth is critical for VFD orientation, suggesting $\{\lambda_m\}$ can be read as an *unsupervised sensor-importance score*. To test robustness to asynchronous perception, we downsample one modality per mini-batch by $\times\frac{1}{2}$ on VFD: UniOMA (Pairwise+GW) consistently outperforms Pairwise (Fig. 3(f)) and adaptively shifts weight toward unchanged modalities while retaining non-trivial mass on the downsampled one (Fig. 3(e)).

IV. Conclusion

We revisit multimodal alignment through *structural* consistency: pointwise correspondences may be tight while intra-modal geometries still disagree. UniOMA closes this gap by combining contrastive learning with a GW-barycenter regularizer that aligns 3+ modalities to a shared structural consensus at $O(M)$ cost. Across VFP, VFD, MuJoCo Push, VAT, VIP, and 4–7 modality settings, UniOMA improves downstream

TABLE I: Comparative results on downstream tasks (regression, classification, and cross-modal retrieval). Performance is measured by MSE ($\times 10^{-3}\downarrow$), Top-1 Acc. ($\%\uparrow$), and MAP ($\uparrow$) (mean $\pm$ std over 10 seeds). Arrows denote retrieval direction. Gray rows correspond to our method (UniOMA). Overall, our method consistently improves its corresponding baselines across most tasks, and all methods achieving the best performance for each task are UniOMA variants (highlighted in brown).

Method	Regression ↓		Classification ↑(%)		VAT MAP Score ↑		
	V&F&D ($\times 10^{-3}$)	MuJoCo	V&F&D	V&F&P	Vis→Aud	Vis→Tact	Tact→Aud
Pairwise (26)	1.27±0.14	0.44±0.07	89.59±0.05	94.51±0.02	0.25±0.07	0.41±0.11	0.10±0.01
Pairwise+GW(ours)	**1.22**±0.12	**0.38**±0.09	**92.44**±0.02	**94.68**±0.03	0.36±0.05	**0.60**±0.03	**0.12**±0.02
Symile (37)	2.81±0.10	0.28±0.04	90.02±0.04	**93.94**±0.06	0.10±0.02	**0.21**±0.05	0.08±0.01
Symile+GW(ours)	**2.15**±0.08	**0.23**±0.02	**92.81**±0.02	93.87±0.03	**0.13**±0.03	0.15±0.03	**0.14**±0.03
GRAM (38)	3.37±0.09	0.52±0.07	92.47±0.04	93.65±0.05	0.13±0.02	0.34±0.05	0.15±0.01
GRAM+GW(ours)	**2.31**±0.05	**0.30**±0.06	**93.30**±0.02	**93.91**±0.04	**0.79**±0.10	**0.58**±0.04	**0.16**±0.01
CoMM (39)	1.51±0.05	0.26±0.04	92.39±0.01	94.13±0.03	—	—	—
OTLC (41)	1.26±0.11	0.40±0.07	92.41±0.02	94.66±0.02	**0.37**±0.05	0.58±0.04	0.09±0.01
TRIANGLE (40)	3.65±0.09	0.41±0.06	93.06±0.04	93.82±0.04	—	—	—

TABLE II: **Scalability analysis with 4–7 modalities.** Trajectory-pair classification accuracy (mean $\pm$ std over 10 seeds) and wall-clock time per epoch. UniOMA (gray) achieves consistently higher accuracy and becomes faster than OT when $M \geq 6$.

Modality Combination	M	Pairwise		OTLC		Pairwise+GW(ours)	
		Acc.	Time	Acc.	Time	Acc.	Time
V+F+P+D	4	89.94±0.03	110.38±1.74s	92.07±0.03	135.57±2.92s	**92.39**±0.02	201.36±7.61s
V+F+P+D+A	5	90.72±0.03	129.44±1.92s	92.51±0.03	178.63±3.11s	**93.04**±0.02	225.89±5.44s
V+F+P+D+A+C	6	89.12±0.04	150.77±2.51s	91.03±0.03	268.41±6.83s	**92.11**±0.03	248.52±6.12s
V+F+P+D+A+C+O	7	87.95±0.05	171.42±3.12s	89.84±0.04	382.77±10.44s	**91.02**±0.03	273.36±7.40s

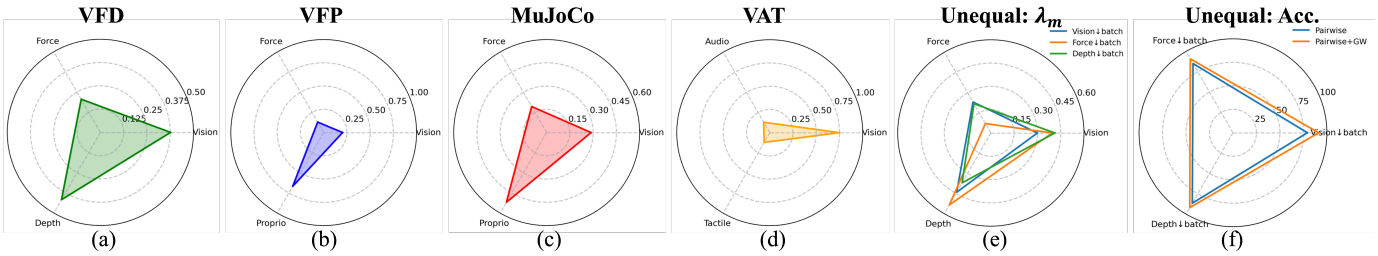


Fig. 3: (a–d) Final learned modality weights $\{\lambda_m\}$ for each task (VFD, VFP, MuJoCo Push, VAT), highlighting dataset-specific salience (e.g., depth in VFD, proprioception in VFP). (e) Ablation on VFD: one modality is downsampled by $\times\frac{1}{2}$ per batch; UniOMA adaptively redistributes $\{\lambda_m\}$ toward intact modalities. (f) Accuracy under the same ablation (Top-1, %); the outer polygon shows consistent gains of Pairwise+GW over Pairwise.

TABLE III: **End-effector position prediction on VIP.** Regression error in MSE ($\times 10^{-2}\downarrow$; mean $\pm$ std over 10 seeds). Gray row is ours.

Method	Modality	Regression ↓ ($\times 10^{-2}$)
		End-effector Pos.
Pairwise (26)	V+I+P	6.32±0.15
Symile (37)	V+I+P	4.59±0.12
GRAM (38)	V+I+P	6.10±0.18
TRIANGLE (40)	V+I+P	5.87±0.09
Pairwise+GW(ours)	V+I+P	**4.02**±**0.10**

tasks like regression, classification, and retrieval while learning interpretable modality weights.

Limitations and future work. UniOMA introduces additional computation due to the barycentric GW updates, but the runtime scales *linearly* with the number of modalities, rather than superlinearly, which makes UniOMA practical for 3+ modality and is a key advantage of the proposed method. A promising direction is large-scale real-robot alignment under heterogeneous sampling rates, missing segments, and asynchronous sensor streams. Another direction is extending the framework beyond symmetric similarity kernels to asymmetric similarities (e.g., temporal precedence or causal structure), enabling the consensus geometry to capture directional relations in interaction-rich robotics data.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[3] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical ai," *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, Sep. 2022.

[4] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[5] S. Steyaert, M. Pizurica, D. Nagaraj, P. Khandelwal, T. Hernandez-Boussard, A. J. Gentles, and O. Gevaert, "Multimodal data fusion for cancer biomarker discovery with deep learning," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 351–362, Apr. 2023.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[8] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent a new approach to self-supervised learning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[9] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[10] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[12] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

[13] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu, "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 842–72 866, 2023.

[14] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li *et al.*, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," *arXiv preprint arXiv:2310.01852*, 2023.

[15] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.

[16] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: understanding the modality gap in multi-modal contrastive representation learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[17] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, Jun. 2004.

[18] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International conference on machine learning*, 2019, pp. 5171–5180.

[19] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine, "Time-contrastive networks: Self-supervised learning from video," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3997350

[20] D. Stewart and J. Trinkle, "An implicit time-stepping scheme for rigid body dynamics with coulomb friction," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, ser. ROBOT-00, vol. 1. IEEE, pp. 162–169.

[21] M. Guo, Y. Jiang, A. E. Spielberg, J. Wu, and K. Liu, "Benchmarking rigid body contact models," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, N. Matni, M. Morari, and G. J. Pappas, Eds., vol. 211. PMLR, 15–16 Jun 2023, pp. 1480–1492. [Online]. Available: https://proceedings.mlr.press/v211/guo23b.html

[22] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg, "Multimodal sensor fusion with differentiable filters," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE. IEEE, Oct. 2020, pp. 10 444–10 451.

[23] G. Welch and G. Bishop, "An introduction to the kalman filter," USA, Tech. Rep., 1995.

[24] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-

wasserstein averaging of kernel and distance matrices," in *International conference on machine learning*. PMLR, 2016, pp. 2664–2672.

[25] F. Gong, Y. Nie, and H. Xu, "Gromov-wasserstein multimodal alignment and clustering," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, ser. CIKM '22. ACM, Oct. 2022, pp. 603–613.

[26] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*. Springer, 2020, pp. 776–794.

[27] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in neural information processing systems*, vol. 34, pp. 24 206–24 221, 2021.

[28] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," *Advances in neural information processing systems*, vol. 33, pp. 25–37, 2020.

[29] B. Chen, A. Rouditchenko, K. Duarte, H. Kuehne, S. Thomas, A. Boggust, R. Panda, B. Kingsbury, R. Feris, D. Harwath *et al.*, "Multimodal clustering networks for self-supervised learning from unlabeled videos," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8012–8021.

[30] J. Liu, S. Chen, X. He, L. Guo, X. Zhu, W. Wang, and J. Tang, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[31] J. Huang, Y. Li, J. Feng, X. Wu, X. Sun, and R. Ji, "Clover: Towards a unified video-language alignment and fusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 856–14 866.

[32] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2276–2289, 2022.

[33] S. Moon, A. Madotto, Z. Lin, A. Dirafzoon, A. Saraf, A. Bearman, and B. Damavandi, "Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text," *arXiv preprint arXiv:2210.14395*, 2022.

[34] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2022, pp. 20 020–20 029.

[35] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, and J. Luo, "Clip-vip: Adapting pre-trained image-text model to video-language representation alignment," *arXiv preprint arXiv:2209.06430*, 2022.

[36] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[37] A. Saporta, A. M. Puli, M. Goldstein, and R. Ranganath, "Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities," *Advances in Neural Information Processing Systems*, vol. 37, pp. 56 919–56 957, 2024.

[38] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminiello, "Gramian multimodal representation learning and alignment," *arXiv preprint arXiv:2412.11959*, 2024.

[39] B. Dufumier, J. Castillo-Navarro, D. Tuia, and J.-P. Thiran, "What to align in multimodal contrastive learning?" *arXiv preprint arXiv:2409.07402*, 2024.

[40] G. Cicchetti, E. Grassucci, and D. Comminiello, "A triangle enables multimodal alignment beyond cosine similarity," 2025.

[41] S. Zhu and D. Luo, *Enhancing Multi-modal Contrastive Learning via Optimal Transport-Based Consistent Modality Alignment*. Springer Nature Singapore, Nov. 2024, pp. 157–171.

[42] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.

[43] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," 2022.

[44] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: an embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[45] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," 2024.

[46] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong,

T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[47] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$_{0.5}$: a vision-language-action model with open-world generalization," 2025.

[48] J. Wojcik, J. Jiang, J. Wu, and S. Luo, "A case study on visual-audio-tactile cross-modal retrieval," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 472–12 478.

[49] A. Dutta, E. Burdet, and M. Kaboli, "Visuo-tactile based predictive cross modal perception for object exploration in robotics," 2024.

[50] M. Zambelli, Y. Aytar, F. Visin, Y. Zhou, and R. Hadsell, "Learning rich touch representations through cross-modal self-supervision," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 1415–1425. [Online]. Available: https://proceedings.mlr.press/v155/zambelli21a.html

[51] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2008, vol. 338.

[52] K. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen, "Time series cluster kernel for learning similarities between multivariate time series with missing data," *Pattern Recognition*, vol. 76, pp. 569–581, Apr. 2018.

[53] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[54] F. Chung, *Spectral Graph Theory*. American Mathematical Society, Dec. 1996.

[55] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Aug. 2007.

[56] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[57] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L.-P. Morency, "Multibench: Multiscale benchmarks for multimodal representation learning," 2021.

[58] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2012.

[59] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 598–10 608.