ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

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 multimodal 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

Contrastive self-supervised methods [1], [2] align heterogeneous modalities by maximizing agreement between paired instances, but treat alignment as binary classification: paired samples are pulled together, unpaired pushed apart, without modeling intra-modal distance geometry. Representations thus align at the instance level yet disagree in how samples relate within each modality [3], which we identify as a *structural alignment gap*. This is a theoretical limitation: InfoNCE lower-bounds mutual information but is invariant to structural preservation [4]. The gap is critical in robotics, where trajectories form subclusters, contacts induce discontinuities [5], and proprioception follows physical constraints [6].

We introduce **UniOMA**, which augments contrastive learning with Gromov–Wasserstein (GW) barycenter regularization [7]. A dynamic GW barycenter captures structural consensus, and each modality is aligned to it via learned weights, reducing complexity from $O(M^2)$ to $O(M)$.

II. Method

Given M modalities $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(M)}$, we learn encoders $f^{(m)} : \mathcal{X}^{(m)} \to \mathbb{R}^d$ to produce embeddings $\mathbf{z}^{(m)} = f^{(m)}(\mathbf{x}^{(m)})$.

GW Distance. The GW distance [7] compares two distributions by matching their relational geometry without a cross-modal cost. For kernel matrices $\mathbf{K_x} \in \mathbb{R}^{I\times I}$, $\mathbf{K_y} \in \mathbb{R}^{J\times J}$, the empirical GW distance is:

$$\hat{d}_{gw}(\mathbf{K_x}, \mathbf{K_y}) = \max_{\mathbf{T}\in\Pi(\hat{\mathbf{p}}_\mathbf{x}, \hat{\mathbf{p}}_\mathbf{y})} \operatorname{tr}(\mathbf{K_x}^\top \mathbf{T}^\top \mathbf{K_y} \mathbf{T}), \quad (1)$$

where $\mathbf{T}$ is a doubly-stochastic transport plan.

Structural Consensus. Each modality's geometry is encoded by a kernel $\mathbf{K}_\mathbf{x}^{(m)} \in \mathbb{R}^{N_m \times N_m}$ (RBF for images, TCK [8] for time-series). The structural consensus is the GW barycenter:

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

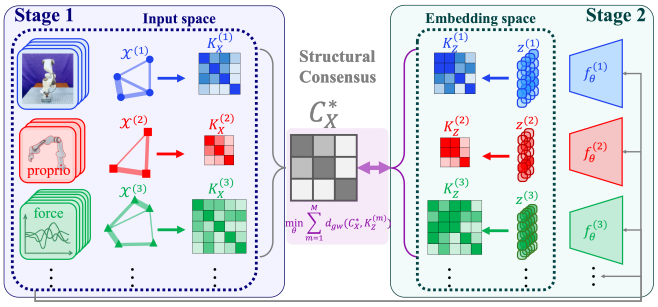

Fig. 1. **UniOMA overview.** Stage 1: compute input-space kernels $\mathbf{K}_\mathbf{x}^{(m)}$ and GW barycenter $\mathbf{C}_\mathbf{x}^*$. Stage 2: align embedding kernels $\mathbf{K}_\mathbf{z}^{(m)}$ to $\mathbf{C}_\mathbf{x}^*$ via Eq. 3.

where $\{\lambda_m\}$ are learnable weights with $\lambda_m \geq 0$, $\sum_m \lambda_m = 1$.

UniOMA Objective. We augment a contrastive loss with a structure-aware regularizer:

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

where $\mathbf{K}_\mathbf{z}^{(m)}$ is the embedding-space kernel and α balances the two terms.

Training alternates two stages per iteration (Fig. 1): (1) compute batch-wise kernels $\mathbf{K}_\mathbf{x}^{(m)}$ and estimate the consensus $\mathbf{C}_\mathbf{x}^*$; (2) encode the batch, form $\mathbf{K}_\mathbf{z}^{(m)}$, and update θ and $\{\lambda_m\}$ via SGD on $\mathcal{L}_{\text{UniOMA}}$.

III. Experiments

We evaluate on: (i) **VFD/VFP** [9] (Vision–Force–Depth/Proprioception): regression and classification; (ii) **MuJoCo Push** [10] (Vision–Force–Pose): position regression; (iii) **VAT** [11] (Vision–Audio–Tactile): cross-modal retrieval; (iv) **VIP** (Vision–IMU–Proprioception): real-world end-effector regression. All methods share identical backbones and training setup. Baselines: Pairwise [12], Symile [13], GRAM [14], OTLC [15], TRIANGLE [16], CoMM [17]; our GW regularizer is added plug-and-play ("+GW").

Results. Table I shows that adding the GW regularizer yields consistent gains across all base objectives and task types; every best result is a UniOMA variant. UniOMA also scales to 4–7 modalities with linear wall-clock growth vs. quadratic for pairwise baselines.

Interpretable Modality Weights. The learned $\{\lambda_m\}$ reveal dataset-specific modality salience (Fig. 2): depth dominates on VFD (spatial reasoning), proprioception on VFP (kinematics), vision on MuJoCo (contact/object state), and tactile

TABLE I

DOWNSTREAM RESULTS (MEAN±STD, 10 SEEDS). GRAY : UNIOMA. BEST PER-GROUP **BOLD**; OVERALL BEST BROWN.

	Regression ↓		Classification ↑(%)		VAT MAP ↑		
Method	V&F&D($\times 10^{-3}$)	MuJoCo	V&F&D	V&F&P	Vis→Aud	Vis→Tact	Tact→Aud
Pairwise [12]	1.27±.14	0.44±.07	89.59±.05	94.51±.02	0.25±.07	0.41±.11	0.10±.01
Pairwise+GW (ours)	**1.22**±.12	**0.38**±.09	**92.44**±.02	**94.68**±.03	0.36±.05	**0.60**±.03	**0.12**±.02
Symile [13]	2.81±.10	0.28±.04	90.02±.04	**93.94**±.06	0.10±.02	**0.21**±.05	0.08±.01
Symile+GW (ours)	**2.15**±.08	**0.23**±.02	**92.81**±.02	93.87±.03	**0.13**±.03	0.15±.03	**0.14**±.03
GRAM [14]	3.37±.09	0.52±.07	92.47±.04	93.65±.05	0.13±.02	0.34±.05	0.15±.01
GRAM+GW (ours)	**2.31**±.05	**0.30**±.06	**93.30**±.02	**93.91**±.04	**0.79**±.10	**0.58**±.04	**0.16**±.01
CoMM [17]	1.51±.05	0.26±.04	92.39±.01	94.13±.03	—	—	—
OTLC [15]	1.26±.11	0.40±.07	92.41±.02	94.66±.02	**0.37**±.05	0.58±.04	0.09±.01
TRIANGLE [16]	3.65±.09	0.41±.06	93.06±.04	93.82±.04	—	—	—

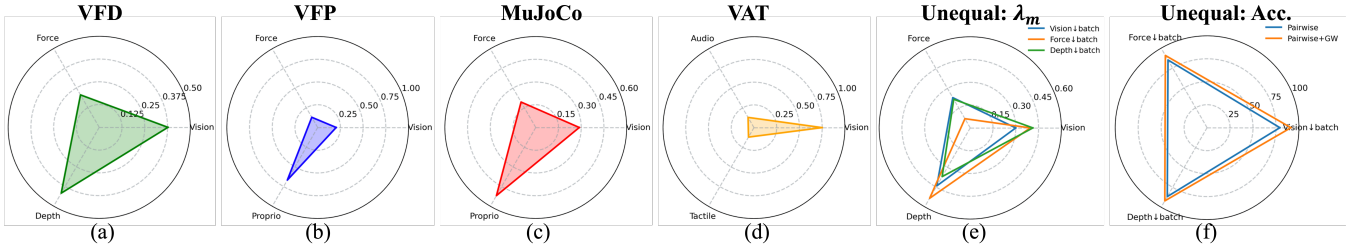

Fig. 2. **Learned modality weights** $\{\lambda_m\}$ across four benchmarks (left four). Right: ablation under unequal modality sampling—UniOMA redistributes $\{\lambda_m\}$ toward intact modalities.

on VAT (material retrieval). These weights, learned end-to-end, provide interpretable diagnostics of each modality's contribution.

IV. CONCLUSION

UniOMA closes the structural alignment gap by combining contrastive learning with a GW-barycenter regularizer, aligning 3+ modalities to a shared structural consensus with $O(M)$ complexity. Across five robotic benchmarks, it consistently improves regression, classification, and cross-modal retrieval.

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

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

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