MATCH: Multimodal Alignment Through Contrastive Learning of Heterogeneous Sensor Streams at Static Nodes

Julia Gersey*†, Nicola Ferrier*‡
*Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL
†Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI
‡Northwestern Argonne Institute of Science and Engineering, Northwestern University, Evanston, IL

***Abstract*—Multimodal environmental sensing networks routinely capture temporally co-occurring conditions across incompatible representations: an RGB frame, a thermal frame, an audio clip, and a suite of environmental readings share no common feature space, even when recorded by the same node in the same time window. We present MATCH, a contrastive learning framework that addresses this gap by mapping all four modalities into a shared 512-dimensional embedding space using frozen, domain-matched pretrained backbones and lightweight trainable projection heads. A multi-view InfoNCE loss aligns all six modality pairs without task-specific labels. Evaluated on a 195-observation held-out test set from a single deployed environmental node, retrieval improves over random across all six modality pairs, with RGB↔Thermal achieving the strongest alignment. Among its incorrect top-5 retrievals, 73.3% fall within ± 2 hours of the query's time of day on another day, compared with a 20.4% chance baseline, indicating that the learned space captures coarse diurnal structure. These preliminary results demonstrate the feasibility of within-node multimodal alignment as a foundation for scalable search and analysis across distributed static edge-sensing infrastructure, while motivating time-aware hard negatives and node-specific adaptation.**

***Index Terms*—multimodal sensing, contrastive learning, sensor fusion, edge computing, environmental monitoring**

I. Introduction

The Sage [1] environmental sensing network contains over 100 AI-enabled edge nodes collecting RGB, thermal, audio, and environmental time-series data. Scalable search and analysis across this distributed infrastructure require a common representation for its heterogeneous sensor streams. MATCH learns such a space, aligning temporally co-occurring observations without task-specific labels. Prior work has aligned image–text pairs (CLIP [2]), image-anchored sensory modalities (ImageBind [3]), and ecological observations (TaxaBind [4]); MATCH instead targets multimodal streams from a real-world, statically deployed sensor.

We first evaluate a single node to isolate alignment from site-identity confounds inherent to static sensors. It contains a Mobotix M16 RGB–thermal camera, an ML1-WS microphone, and a BME680 measuring temperature, humidity, pressure, and gas resistance. Nine weeks of hourly data yield 1,361 synchronized observations, each containing RGB, thermal, 30-second audio, and environmental measurements.

This work was supported in part by the National Science Foundation under Awards No. 2331263 and 2436842, and the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Award Number DE-SC0026073.

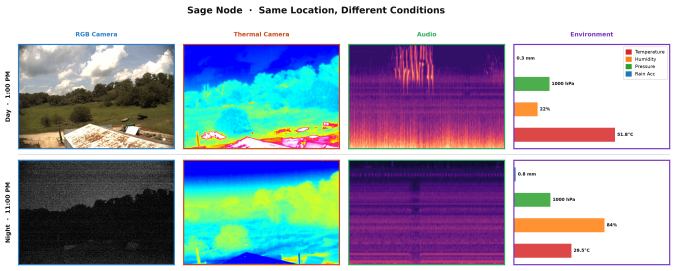


Fig. 1. A Sage node capturing four modalities simultaneously. A daytime observation (top) and a nighttime observation (bottom) from the same node illustrate how dramatically RGB, thermal, audio, and environmental signals vary across conditions, making cross-modal alignment non-trivial.

II. Approach & Methodology

We choose a simple, interpretable approach: pretrained foundation models serve as frozen feature extractors, with only lightweight per-modality projection heads trained on top. The overall architecture is shown in Figure 2.

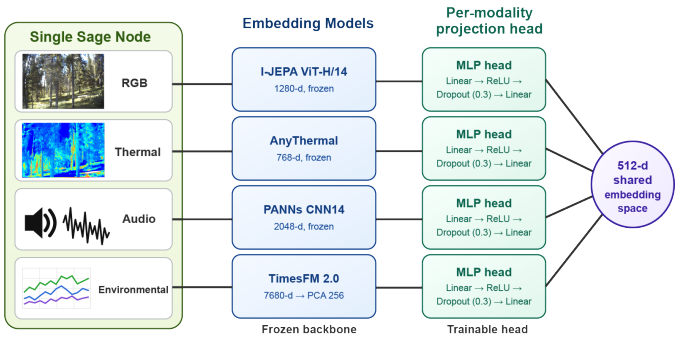


Fig. 2. MATCH architecture. Each modality from a single Sage node passes through a frozen, domain-matched backbone and a trainable per-modality projection head, mapping all four into a shared 512-d embedding space.

Backbone selection. Given roughly 1,000 training observations, we freeze the backbones to limit overfitting, and selection is grounded in per-modality domain fit.

RGB uses I-JEPA ViT-H/14 (1280-d) [5], a self-supervised masked latent prediction model; patch tokens are mean-pooled.

Thermal uses AnyThermal (768-d) [6], a DINOv2 ViT-B/14 distilled on thermal imagery.

Audio uses PANNs CNN14 (2048-d) [7], pretrained on AudioSet across 527 sound classes. Thirty-second clips are encoded as log-mel spectrograms.

Environmental time series use TimesFM 2.0-500M (7680-d) [8], Google's general-purpose time-series model. The four sensor channels are treated as a multivariate series, encoded jointly, then reduced to 256-d via PCA before training.

Modality-specific projection heads map each backbone output into a shared 512-d space. A multi-view InfoNCE objective jointly optimizes all six modality pairs using same-timestamp positives and in-batch negatives. Training uses Adam (3×10^{-4}), cosine annealing, $\tau = 0.07$, batch size 128, dropout 0.3, and early stopping with patience 40.

III. Results

We use a chronological 70/15/15 split to prevent adjacent observations from crossing train and test windows. Timestamp-paired training improves mean retrieval rank over random across all six modality pairs, with performance varying by pair (Table I).

Pair	z-score	Mean rank	Hub max rate
RGB ↔ Thermal	1.212	28.8	0.221
RGB ↔ Audio	1.017	39.7	0.308
RGB ↔ Env	1.082	36.1	0.262
Thermal ↔ Audio	0.846	49.4	0.349
Thermal ↔ Env	1.014	39.9	0.256
Audio ↔ Env	0.780	53.1	0.292

TABLE I
RETRIEVAL ON 195 HELD-OUT OBSERVATIONS. HIGHER Z-SCORE AND LOWER MEAN RANK AND HUB RATE INDICATE STRONGER ALIGNMENT; RANDOM MEAN RANK IS 97 ($\sigma \approx 56$).

Alignment remains modest ($z = 0.780$–1.212), with no pair reaching $z = 2$. RGB↔Thermal performs best across all three metrics, while Audio↔Environmental is weakest, likely reflecting the mismatch between AudioSet pretraining and outdoor acoustics.

A separate multi-node analysis confirms that site identity is strongly encoded across modalities. Cross-node transfer exhibits substantially greater hub domination than within-node training (0.631 vs. 0.163), motivating the within-node evaluation.

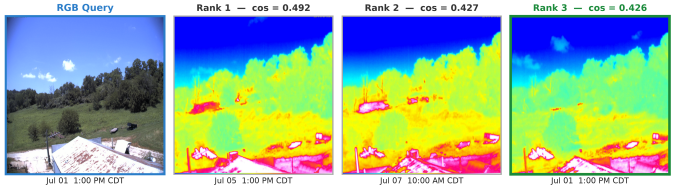


Fig. 3. RGB-to-thermal retrieval from the held-out test set. The correct thermal match ranks third, while nearby ranks show visually similar observations from other dates, illustrating the ambiguity induced by recurring static-site conditions.

Among incorrect top-5 retrievals, 73.3% fall within ±2 hours of the query's time of day on another day, compared with a 20.4% chance baseline, a 3.6× enrichment. Thus, MATCH captures coarse diurnal structure but struggles to distinguish observations within the same phase, motivating time-stratified hard negatives.

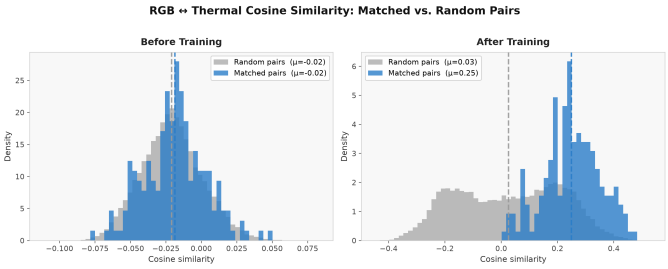


Fig. 4. RGB↔Thermal similarity before and after training. Matched and random pairs overlap before training ($\mu = -0.02$) but separate afterward ($\mu = 0.25$ vs. 0.03).

IV. Conclusion and Future Work

Shared multimodal embeddings provide a foundation for scalable search and analysis across distributed edge-science infrastructure. MATCH demonstrates preliminary feasibility via a single static edge node using frozen foundation models and lightweight contrastive training. This within-node framing provides a tractable starting point for separating cross-modal alignment from site identity. Future work will investigate node-specific adaptation, time-stratified hard negatives, soundscape-adapted audio encoders, anomaly detection, and cross-modal sensor imputation.

References

[1] "Home | Sage." [Online]. Available: https://sagecontinuum.org/

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, arXiv:2103.00020 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2103.00020

[3] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 15 180–15 190. [Online]. Available: https://ieeexplore.ieee.org/document/10203733/

[4] S. Sastry, S. Khanal, A. Dhakal, A. Ahmad, and N. Jacobs, "TaxaBind: A Unified Embedding Space for Ecological Applications," Nov. 2024, arXiv:2411.00683 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2411.00683

[5] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 15 619–15 629. [Online]. Available: https://ieeexplore.ieee.org/document/10205476/

[6] P. Maheshwari, J. Karhade, Y. Chawla, I. Adu, F. Heisen, A. Porco, A. Jong, Y. Liu, S. Pitla, S. Scherer, and W. Wang, "AnyThermal: Towards Learning Universal Representations for Thermal Perception," Feb. 2026, arXiv:2602.06203 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2602.06203

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," Dec. 2019, arXiv:1912.10211 [cs.SD] version: 1. [Online]. Available: http://arxiv.org/abs/1912.10211

[8] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," Apr. 2024, arXiv:2310.10688 [cs.CL]. [Online]. Available: http://arxiv.org/abs/2310.10688