SheafAlign: A Sheaf-theoretic Framework for Decentralized Multimodal Alignment

Abdulmomen Ghalkha, Zhuojun Tian, Chaouki Ben Issaid, and Mehdi Bennis

***Abstract*—Conventional multimodal alignment methods assume mutual redundancy across all modalities, an assumption that fails in real-world distributed scenarios. We propose SheafAlign, a sheaf-theoretic framework for decentralized multimodal alignment that replaces single-space alignment with multiple comparison spaces. This approach models pairwise modality relations through sheaf structures and leverages decentralized contrastive learning-based objectives for training. SheafAlign overcomes the limitations of prior methods by not requiring mutual redundancy among all modalities, preserving both shared and unique information. Experiments on multimodal sensing datasets show superior zero-shot generalization, cross-modal alignment, and robustness to missing modalities, with 50% lower communication cost than state-of-the-art baselines.**

***Index Terms*—Distributed optimization, distributed sensing, sheaf theory, multimodal learning.**

I. Introduction

The rapid proliferation of multimodal sensor networks and distributed data collection systems mandates the development of learning frameworks capable of effectively integrating heterogeneous sources of information. Multimodal learning aims to integrate information from heterogeneous sources, such as images, audio, LiDAR, and wireless signals, for understanding a common event of interest. Central to this goal is multimodal alignment, where the learned embeddings must be semantically consistent across different modalities, facilitating downstream tasks.

Existing multimodal frameworks such as CLIP [1], AudioCLIP [2], and ImageBind [3] embed all modalities into a single, shared embedding space. In CLIP, alignment is established between images and text through contrastive learning. AudioCLIP extends this alignment to audio by adapting the CLIP framework to audio-text-image triplets. Similarly, ImageBind generalizes this further by binding six modalities (images, text, audio, depth, thermal, IMU) into one shared embedding space, relying solely on image-paired data to induce alignment.

While these unified embedding spaces enable strong performance in multimodal tasks, they have significant limitations, particularly in distributed settings where sensors have partially overlapping information. Collapsing all modalities into a single space can suppress modality-specific information and fail to capture nuanced relationships between certain modality pairs. Specifically, visual bias as binding all modalities via images risks distorting non-visual semantics, leading to suboptimal pairings since direct modality links (e.g., audio-text) are overlooked. Furthermore, in scenarios where certain modalities are absent, models trained within such a single space often struggle to infer or reconstruct missing information, particularly when the alignment relies on modalities with insufficient mutual redundancy or is biased towards dominant modalities such as vision. The key challenge is to move beyond these rigid global alignments and design a framework that can model cross-modal relationships, maintaining alignment and inference capabilities, when modalities exhibit limited redundancy or partial observability.

To address this challenge, we propose SheafAlign, a sheaf-theoretic multimodal alignment framework that introduces comparison spaces, latent spaces where embeddings from different modalities can be projected and directly aligned. Unlike single-space alignment approaches, SheafAlign models pairwise modality interactions using sheaves defined over a communication graph, where each edge corresponds to a shared comparison space between the two nodes it connects. Building on the formulation in [4], [5] and defining a sheaf over embeddings, this structure enables modality pairs to align locally without assuming mutual redundancy across all modalities, allowing richer and more pair-specific relationships. Furthermore, SheafAlign does not require the presence of a specific reference modality across all samples, enhancing robustness to missing or partially observed data during training and inference. The main contributions of this article are summarized as follows

- We introduce SheafAlign, a novel sheaf-theoretic framework for multimodal alignment that replaces a single embedding space with a network of comparison spaces.
- We develop a novel decentralized training procedure that enables clients to collaboratively learn local modality-specific embeddings and pairwise relations, enhancing scalability and communication efficiency.
- Experiments on real-world and synthetic multimodal datasets demonstrate superior cross-modal alignment, zero-shot and few-shot generalization, and robustness to missing modalities, while achieving up to 50% communication savings compared to other baselines.

The rest of this paper is organized as follows. Section II presents the system model and problem formulation. Section III introduces the proposed sheaf-theoretic framework for multimodal alignment. In Section IV, we evaluate the performance of SheafAlign through real-world and synthetic datasets.

A. Ghalkha, Z. Tian, C. B. Issaid, and M. Bennis are with the Center for Wireless Communications, University of Oulu, Oulu 90014, Finland. Email: {abdulmomen.ghalkha, zhuojun.tian, chaouki.benissaid, mehdi.bennis}@oulu.fi.

II. System Model and Problem Formulation

In the decentralized learning framework, the system is represented as a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V} = \{1, \ldots, N\}$ represents a client equipped with a single modality $m_i \in \mathcal{M} = \{1, \ldots, M\}$. The edges, denoted by $\mathcal{E}$, define the direct communication links between clients. Each client i collects its local observations $\boldsymbol{x}_{m_i}$ with its modality m_i.

All clients are assumed to observe the same underlying event or scene. We do not assume access to labels for these observations. Instead, clients learn collaboratively by exchanging information with their neighbors through the communication graph $\mathcal{G}$, allowing the system to discover relationships between modalities and capture cross-modal dependencies without centralized supervision. Each client applies a modality-specific feature extractor $f_i(\cdot)$ to obtain its embedding $\boldsymbol{h}_i = f_i(\boldsymbol{x}_{m_i})$, to learn aligned embeddings across modalities that satisfy $\boldsymbol{h}_i \approx \boldsymbol{h}_j, \forall i, j \in \mathcal{V}$, if $\boldsymbol{x}_{m_i}$ and $\boldsymbol{x}_{m_j}$ originate from the same sample.

A key challenge in decentralized multimodal alignment is achieving robust aligned representations across distributed sensors with heterogeneous modalities. Existing frameworks address this by embedding all modalities into a single shared embedding space, implicitly assuming equal information redundancy across modalities. However, in distributed settings, this assumption rarely holds as spatially separated sensors often capture views with different levels of overlap and correlation. As a result, visual bias, loss of modality-specific information, and weak alignment between non-visual pairs can occur, limiting fine-grained semantic consistency across modalities. To better illustrate these limitations, following the definition in [6], we define the information-theoretic decomposition of task-relevant information across modalities.

Definition 1 (Information-Theoretic Decomposition)**.** *Let $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M\}$ be a set of random variables representing M modalities observing a common phenomenon, and let Y denote a downstream task variable. The total task-relevant information can be decomposed as*

$$I(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M; Y) = \sum_i U_i + \sum_{i<j} R_{ij} + \cdots + R_{1\ldots M} + \sum_{i<j} S_{ij} + \cdots + S_{1\ldots M}, \quad (1)$$

where U_i denotes the information about Y that is unique to modality $\boldsymbol{X}_i$, while $R_{ij\ldots k}$ and $S_{ij\ldots k}$ represent, respectively, the redundant and synergistic information about Y that is shared among modalities $\{\boldsymbol{X}_i, \boldsymbol{X}_j, \ldots, \boldsymbol{X}_k\}$ but not by any subset thereof.

For each modality m_i, a modality-specific encoder $f_i : \boldsymbol{X}_{m_i} \to \boldsymbol{h}_i$ maps the raw observation $\boldsymbol{X}_{m_i}$ to a learned embedding $\boldsymbol{h}_i = f_i(\boldsymbol{X}_{m_i})$. The goal of representation learning is to learn $\{f_i\}_{i=1}^M$ such that, as in Definition 1, the mutual information $I(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M; Y)$ is maximized. However, when all embeddings are constrained to a single shared space, the retained information about Y is bounded by the smallest intersection across modalities, i.e., $I(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M; Y) \leq R_{1\ldots M}$, as shown in the bottom part of Figure 1. As a result, pairwise-shared and unique modality information cannot be fully captured, thereby hindering downstream task performance.

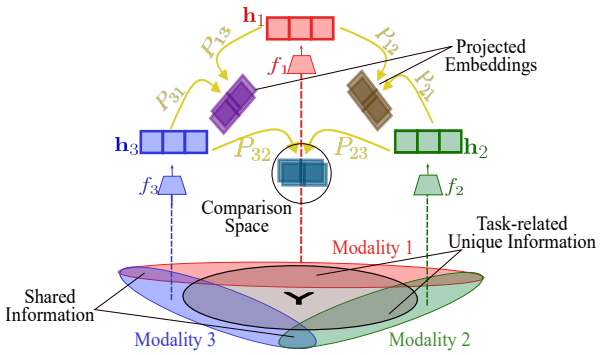


Fig. 1: **Bottom:** Illustration of the absence of mutual redundancies between modalities. **Top:** Schematic diagram of SheafAlign architecture denoting restriction maps and comparison space alignment.

III. Proposed Algorithm

To address the limitations of single-space alignment, we introduce multiple comparison spaces and adopt a pairwise contrastive learning strategy. With this, we allow each modality pair to align in its most informative subspace, rather than forcing all modalities into one shared embedding space. This approach captures shared information implicitly through overlapping pairwise comparisons while preserving non-redundant information. This formulation aligns naturally with sheaf theory, which provides a rigorous framework for representing local modality embeddings as locally consistent assignments and enforcing global consistency through the underlying sheaf structure. Inspired by that, we utilize the sheaf structure into the framework design, as illustrated in the top part of Figure 1.

A. Sheaf Structure

A sheaf structure, as described in [4], provides a principled framework for modeling interactions across decentralized agents by formalizing *local views* and their pairwise relationships. Building on this idea, we shift the focus from modeling relations between model parameters to *embeddings*. Specifically, we define a *cellular sheaf over the embedding space*, enabling pairwise alignment of embeddings between modalities while maintaining unique information of each modality, thus mitigating limitations of forcing all embeddings into a single shared space.

Intuitively, a sheaf structure assigns a vector space to each node and edge in the network, along with linear maps that relate them. In our context, these vector spaces represent *semantic embeddings*—intermediate feature representations extracted from local modality-specific observations. The sheaf structure captures how these local embeddings can be translated and compared between neighbors in the communication graph.

Formally, a *cellular sheaf* $\mathcal{F}$ of $\mathbb{R}$-vector spaces over a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of

- For each node $i \in \mathcal{V}$, a local embedding space $\mathcal{F}(i) = \mathbb{R}^{d_i}$.
- For each edge $e = (i, j) \in \mathcal{E}$, a shared comparison space $\mathcal{F}(e) = \mathbb{R}^{d_{ij}}$.
- For each edge e and incident vertex i, a linear *restriction map* $\mathcal{F}_{i \trianglelefteq e} : \mathcal{F}(i) \to \mathcal{F}(e)$, projecting local embeddings onto a shared comparison space for alignment.

Here, the stalk $\mathcal{F}(i)$ represents the *embedding space* of client i (e.g., $\boldsymbol{h}_i \in \mathbb{R}^{d_i}$), and the maps $\mathcal{F}_{i \trianglelefteq e}$ are used to project embeddings into a shared comparison space for *semantic comparison across neighbors*. Given an edge (i, j), we denote the matrix representation of $\mathcal{F}_{i \trianglelefteq e}$, with respect to a chosen basis such as the standard basis, by $\boldsymbol{P}_{ij}$, and let $\boldsymbol{P}_{ij}^T$ represent its transpose.

Building on this sheaf-theoretic formulation, local alignment between neighboring embeddings is realized by projecting each embedding into the corresponding shared comparison space, $\mathcal{F}(e)$ via $\mathcal{F}_{i \trianglelefteq e}(\boldsymbol{h}_i)$ and $\mathcal{F}_{j \trianglelefteq e}(\boldsymbol{h}_j)$. The discrepancy between neighboring embeddings is captured by the *sheaf Laplacian* $L_{\mathcal{F}}$, which measures the degree of consistency across all local assignments. The sheaf Laplacian is a block matrix, with each block L_{ji} defined as

$$L_{ji} = \begin{cases} \sum_{i \trianglelefteq e} \mathcal{F}^*_{i \trianglelefteq e} \circ \mathcal{F}_{i \trianglelefteq e}, & \text{if } i = j, \\ -\mathcal{F}^*_{j \trianglelefteq e} \circ \mathcal{F}_{i \trianglelefteq e}, & \text{if } e = (i, j) \in \mathcal{E}, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{2}$$

The quadratic form $\boldsymbol{h}_n^\top L_{\mathcal{F}} \boldsymbol{h}_n$ quantifies how inconsistent the neighboring embeddings are across the network

$$\boldsymbol{h}_n^\top L_{\mathcal{F}} \boldsymbol{h}_n = \sum_{e=(i,j)\in\mathcal{E}} \mathbb{1}_{(i,j),n} \|\mathcal{F}_{i \trianglelefteq e}(\boldsymbol{h}_{i,n}) - \mathcal{F}_{j \trianglelefteq e}(\boldsymbol{h}_{j,n})\|^2. \tag{3}$$

where $\boldsymbol{h}_n = \{\boldsymbol{h}_{i,n}\}_{i\in\mathcal{V}}$, and $\mathbb{1}_{(i,j),n}$ is 1 if both modalities i and j are observed for sample n, and 0 otherwise. This introduces an additional degree of freedom by allowing comparisons in dedicated edge-specific comparison spaces, while simultaneously preserving relationships and alignments between the original local embeddings.

B. Sheaf-enabled Multimodal Alignment

While the sheaf Laplacian enforces structural agreement, it alone cannot ensure that embeddings are semantically discriminative. To achieve this, we introduce a contrastive loss that operates directly within each shared comparison space $\mathcal{F}(e)$. Specifically, for an edge $e = (i, j)$ and a batch of B co-observed samples, we first obtain the local embeddings $\{\boldsymbol{h}_{i,n}\}_{n=1}^B$ and $\{\boldsymbol{h}_{j,n}\}_{n=1}^B$. These are then projected into the shared comparison space through $\boldsymbol{p}_{i,n}^{(e)} = \boldsymbol{P}_{i,j}\boldsymbol{h}_{i,n}$ and $\boldsymbol{p}_{j,n}^{(e)} = \boldsymbol{P}_{j,i}\boldsymbol{h}_{j,n}$.

We then apply an InfoNCE-style contrastive loss, treating $(\boldsymbol{p}_{i,n}^{(e)}, \boldsymbol{p}_{j,n}^{(e)})$ as a positive pair. The loss for edge e is then defined as

$$\mathcal{L}_{\text{contrast}}^{(e)} = -\frac{1}{B}\sum_{n=1}^{B} \mathbb{1}_{e,n} \log \frac{\exp(\text{sim}(\boldsymbol{p}_{i,n}^{(e)}, \boldsymbol{p}_{j,n}^{(e)})/\tau)}{\sum_{m=1}^{B} \exp(\text{sim}(\boldsymbol{p}_{i,n}^{(e)}, \boldsymbol{p}_{j,m}^{(e)})/\tau)}, \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity and τ is a temperature parameter. Following [7], [8], minimizing the InfoNCE-style contrastive loss $\mathcal{L}_{\text{contrast}}^{(e)}$ over projected embeddings $\boldsymbol{p}_i^{(e)}$ and $\boldsymbol{p}_j^{(e)}$ is equivalent to maximizing a lower bound on their mutual information, $I(\boldsymbol{p}_i^{(e)}; \boldsymbol{p}_j^{(e)})$. Such alignment in comparison spaces over the sheaf is shown in the top of Figure 1.

To improve cross-modal flexibility, we augment each edge $e = (i, j)$ with a dual map $\boldsymbol{Q}_{ij} : \mathcal{F}(e) \to \mathcal{F}(i)$, as a complement to the restriction map $\boldsymbol{P}_{ij}$, allowing us to avoid orthogonality assumptions on $\boldsymbol{P}_{ij}$. This dual map reconstructs local embeddings from shared comparison space projections, allowing inference of missing modalities $\tilde{\boldsymbol{h}}_{j\to i,n} = \boldsymbol{Q}_{ij}\big(\boldsymbol{P}_{ji}(\boldsymbol{h}_{j,n})\big)$. The dual maps are trained via the reconstruction loss at node i from neighbor j

$$\mathcal{L}_{\text{recon}}^{(i,j)} = \frac{1}{B}\sum_{n=1}^{B} \mathbb{1}_{(i,j),n} \big\|\boldsymbol{Q}_{ij}\boldsymbol{P}_{ji}(\boldsymbol{h}_{j,n}) - \boldsymbol{h}_{i,n}\big\|^2. \tag{5}$$

The final objective thus extends the sheaf Laplacian and contrastive alignment terms by incorporating bidirectional mappings between node and shared comparison spaces. In particular, the total loss is formally expressed as:

$$\mathcal{L}_{\text{total}} = \lambda \sum_{n=1}^{B} \boldsymbol{h}_n^\top L_{\mathcal{F}} \boldsymbol{h}_n + \beta \sum_{e\in\mathcal{E}} \mathcal{L}_{\text{contrast}}^{(e)} + \gamma \sum_{e\in\mathcal{E}} \mathcal{L}_{\text{recon}}^{(e)}, \tag{6}$$

where (λ, β, γ) are hyperparameters that balance consistency, discriminative alignment, and reconstruction of modality embeddings.

C. Training Procedure

The training process proceeds in a fully decentralized manner. Each node updates its parameters using only local data and the embeddings received from its immediate neighbors.

In each iteration, every node i computes its local embeddings $\boldsymbol{h}_{i,n} = f_i(\boldsymbol{x}_{m_i,n})$, projects them onto the incident comparison spaces and exchanges these projections with neighboring nodes to evaluate the local loss $\mathcal{L}_{i,n}$. The node then computes the gradients of its loss $\mathcal{L}_i$, and updates its parameters $\boldsymbol{\theta}_i$, $\boldsymbol{P}_{ij}$, and $\boldsymbol{Q}_{ij}$ using gradient descent.

Following this procedure, all the parameters, including restriction maps $\boldsymbol{P}_{ij}$ and dual maps $\boldsymbol{Q}_{ij}$, are jointly learned without requiring any central coordinator. The entire end-to-end procedure, including embedding exchange, loss computation, and decentralized parameter updates, is summarized in Algorithm 1.

This formulation preserves pairwise shared (R_{ij}), synergetic (S_{ij}), and unique (U_i) information, in contrast to conventional alignment methods, as discussed in section II. This is due to the fact that restriction maps $\boldsymbol{P}_{ij}$ act as filters extracting relevant features of each modality's embedding for alignment in the shared comparison space.

IV. Numerical Results

A. Training Settings

We evaluate the proposed SheafAlign framework on three multimodal datasets:

Algorithm 1 SheafAlign: Decentralized Sheaf-Theoretic Multimodal Alignment

Input: Communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, epochs T, learning rate η, batch size B, hyperparameters λ, β, γ, temperature τ.
Initialize: Feature extractors $\{\boldsymbol{\theta}_i\}_{i\in\mathcal{V}}$, restriction maps $\{\boldsymbol{P}_{ij}\}_{(i,j)\in\mathcal{E}}$, dual maps $\{\boldsymbol{Q}_{ij}\}_{(i,j)\in\mathcal{E}}$.
for epoch $= 1$ to T **do**
    Local step: Each node $i \in \mathcal{V}$ computes embeddings $\mathbf{h}_i = f_i(\mathbf{x}_i; \boldsymbol{\theta}_i)$ from local data.
    Edge alignment: For each edge $(i, j) \in \mathcal{E}$ (in parallel):
    Project embeddings to comparison space $\mathcal{F}(e)$: $\mathbf{p}_i^{(e)} = \boldsymbol{P}_{ij}\mathbf{h}_i$, $\mathbf{p}_j^{(e)} = \boldsymbol{P}_{ji}\mathbf{h}_j$.
    Nodes i and j exchange $\mathbf{p}_i^{(e)}$ and $\mathbf{p}_j^{(e)}$.
    Compute edge loss $\mathcal{L}^{(e)} = \lambda\mathcal{L}_{\text{Lap}}^{(e)} + \beta\mathcal{L}_{\text{contrast}}^{(e)} + \gamma\mathcal{L}_{\text{recon}}^{(e)}$ and the per node loss $\mathcal{L}_i$ using (3, 4), and (5).
    Parameter update: Each node i updates $\boldsymbol{\theta}_i$, $\boldsymbol{P}_{ij}$, and $\boldsymbol{Q}_{ij}$ with gradient descent.

- **DeepSense Multimodal Blockage Prediction** [9]. A mmWave communication dataset capturing transmitter–receiver links under vehicular blockages. It includes three sensing modalities—RGB images, 2D LiDAR scans, and RF power measurements—for classifying whether the link is blocked. For the feature encoders f_i, we use the same architectures described in [5].
- **Multi-view MNIST**. A synthetic dataset derived from MNIST digits with three transformed views (original, edge-filtered, pixel-inverted), used to study multimodal alignment under controlled redundancy levels. We use a feature encoder consisting of two convolutional layers followed by a fully connected layer.
- **Semantic Inpainting** [10]. It consists of images from three cameras and CSI data from nine sensors, each providing different views of the same environment with varying redundancy across modalities. A pre-trained ResNet50 architecture is used.

For the last two datasets, we use images with different views to mimic modalities with varying levels of redundancy. For each dataset $N = 3$ nodes are considered, and the communication graph is fully connected. To evaluate SheafAlign, we compare it with two baselines: (i) ImageBind [3], a state-of-the-art method aligning different modalities in a shared image embedding space; and (ii) a fully supervised learning baseline leveraging ground-truth labels during training. Model performance is assessed using zero-shot and few-shot accuracy (where the number of shots refers to the number of labeled samples), cross-modal retrieval recall (as defined in [2], [3]), and inference communication cost.

For all experiments, we use a batch size of $B = 128$ and train for 50 epochs. The temperature constant in the contrastive loss is set to $\tau = 0.1$, which is a standard choice. Adam optimizer is used with a learning rate of $\eta = 10^{-3}$. For SheafAlign, the training parameters are set as $\lambda = 1.0$, $\beta = 1.0$, and $\gamma = 0.1$.

B. Performance Comparison

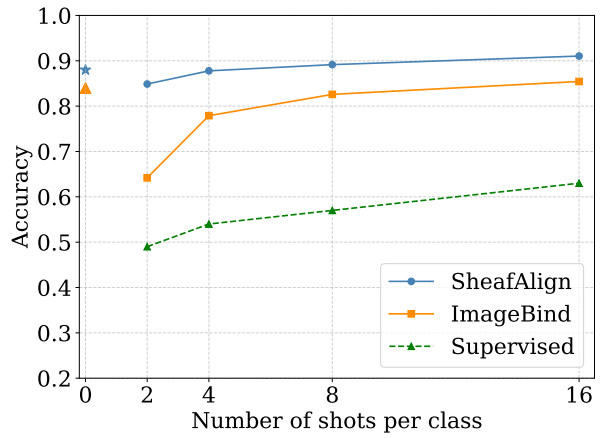
Figure 2 demonstrates the superior performance of SheafAlign compared to ImageBind and the supervised baselines in zero- and few-shot settings. For the DeepSense blockage prediction dataset (Figure 2a), accuracy saturates above four shots or samples per class, which can be attributed to the binary nature and simplicity of the task. In contrast, for the Multi-view MNIST dataset (Figure 2b), the accuracy continues to improve with the number of labeled examples, indicating that SheafAlign effectively leverages limited supervision for more complex tasks. In both cases, SheafAlign consistently outperforms the baseline methods, maintaining 5% accuracy improvement over the strongest baseline. The supervised models exhibit weaker generalization under limited data, reflecting their dependence on large labeled datasets. These results highlight that the embeddings produced by SheafAlign are both semantically meaningful and readily transferable to downstream tasks, even when only a small amount of labeled data is available for fine-tuning.

The experiment results presented in Figure 3 evaluate how well the learned embeddings are semantically aligned across modalities. Both datasets consist of multiple views and provide scenarios with varying degrees of redundancy. As shown in Figure 3a, SheafAlign achieves an average Recall@1 score that is 20% higher than ImageBind, and an 18% improvement in Recall@10. Similarly, Figure 3b demonstrates that SheafAlign attains more than 10% higher recall scores. These results indicate that approaches relying on a single embedding space may yield high recall for certain modality pairs at the expense of others, limiting their ability to consistently capture well-aligned semantic correspondences, particularly when mutual redundancy is insufficient or uneven across modalities. Overall, these findings highlight the superiority of SheafAlign over conventional multimodal alignment frameworks.

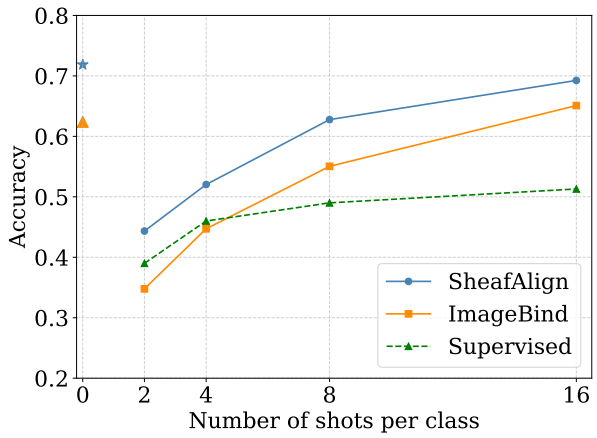
Finally, we evaluate the communication cost incurred when inferring missing embeddings under scenarios where sensors serving downstream tasks may drop with dropout probabilities $P_{\text{drop}} \in \{0.01, 0.1\}$. Semantically aligned embeddings can act as backup representations, allowing available sensors to share their embeddings to substitute the missing modality and maintain uninterrupted performance, particularly in real-time applications. As shown in Table I, SheafAlign achieves comparable accuracy while reducing communication overhead by approximately 50%. This efficiency arises naturally from the sheaf architecture, as the comparison space is designed to be half the size of the original embedding space. Consequently, the sheaf structure emphasizes extracting the most relevant information for each modality pair, while retaining other

TABLE I: Comparison of Average Accuracy and Communication Cost with sensors with dropout probability.

P_{drop}	Algorithm	Accuracy	Transmitted Bytes [KB]
0.1	SheafAlign	0.87	46.20
0.1	ImageBind	0.83	92.48
0.01	SheafAlign	0.90	4.86
0.01	ImageBind	0.84	9.72

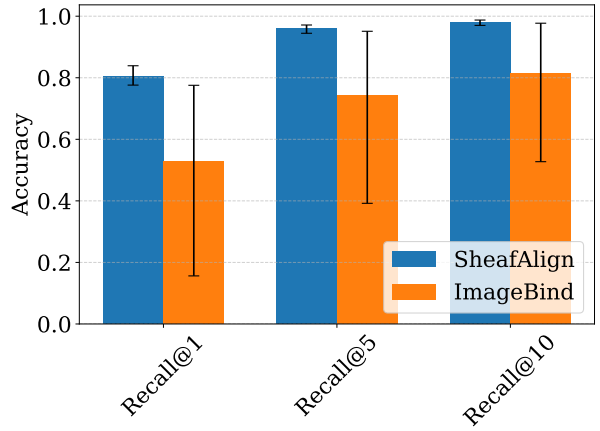


(a) DeepSense blockage prediction.

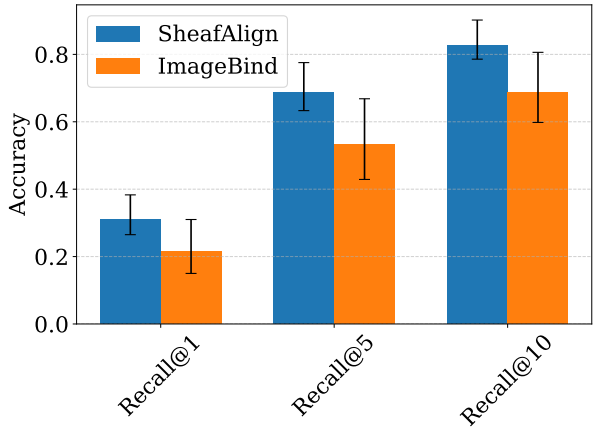


(b) Multi-view MNIST.

Fig. 2: Zero-shot and few-shot test accuracy for SheafAlign compared to ImageBind and supervised baselines across (a) DeepSense blockage prediction, and (b) Multi-View MNIST datasets, with zero-shot accuracies denoted as $\star$ and ▲ for SheafAlign and ImageBind, respectively.



(a) Multi-view MNIST dataset.



(b) Semantic Inpainting dataset.

Fig. 3: Cross-modal retrieval performance of SheafAlign compared to ImageBind across (a) Multi-view MNIST, and (b) semantic inpainting datasets. Bars show the mean Recall@K ($K = 1, 5, 10$), averaged across all modality pairs.

unique details that may be useful for alternative pairs. These properties make SheafAlign particularly promising for deployment in resource-constrained or latency-sensitive applications.

References

[1] 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.

[2] 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.

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

[4] C. B. Issaid, P. Vepakomma, and M. Bennis, "Tackling feature and sample heterogeneity in decentralized multi-task learning: A sheaf-theoretic approach," *Transactions on Machine Learning Research*.

[5] A. Ghalkha, Z. Tian, C. B. Issaid, and M. Bennis, "Sheaf-based decentralized multimodal learning for next-generation wireless communication systems," *arXiv preprint arXiv:2506.22374*, 2025.

[6] P. P. Liang, Z. Deng, M. Q. Ma, J. Y. Zou, L.-P. Morency, and R. Salakhutdinov, "Factorized contrastive learning: Going beyond multi-view redundancy," *Advances in Neural Information Processing Systems*, vol. 36, pp. 32 971–32 998, 2023.

[7] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] 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*. PMLR, 2019, pp. 5171–5180.

[9] S. Wu, C. Chakrabarti, and A. Alkhateeb, "Proactively predicting dynamic 6g link blockages using lidar and in-band signatures," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 392–412, 2023.

[10] T. Nishio, C. Chen, and M. Bennis, "A semantic inpainting framework for distributed cross-modal integrated sensing and communication," in *2025 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*. IEEE, 2025, pp. 1–6.