Collaborative Unpaired Multimodal Learning for Image Classification

Anonymous authors
Paper under double-blind review

Abstract

Multimodal learning typically requires expensive paired data for training and assumes all modalities are available at inference. Many real-world scenarios, however, involve unpaired and heterogeneous data distributed across institutions, making collaboration challenging. We introduce *Unpaired Multimodal Learning (UML)* as the problem of leveraging semantically related but unaligned data across modalities, without requiring explicit pairing or multimodal inference. This setting naturally arises in collaborative scenarios such as satellite imagery, where institutions collect data from diverse sensors (optical, multispectral, SAR), but paired acquisitions are rare and data sharing is restricted. We propose a collaborative framework that combines modality-specific projections with a shared backbone, enabling cross-modal knowledge transfer without paired samples. A key element is post-hoc batch normalization calibration, which adapts the shared model to each modality. Our framework also extends naturally to federated training across institutions. Experiments on multiple satellite benchmarks and additional visual datasets show consistent improvements over unimodal baselines, with particularly strong gains for weaker modalities and in low-data regimes.

1 Introduction

Computer vision and sensing tasks often benefit from integrating data from multiple sources that provide complementary information. In remote sensing, for example, optical and radar imagery capture different physical properties of the Earth's surface. While optical sensors offer high spatial resolution, radar can penetrate cloud cover and better characterize surface structure. Similar multimodal configurations also arise in medical imaging (*e.g.*, CT–MRI) and robotics (*e.g.*, RGB–depth). Leveraging such modality diversity during training leads to *multimodal learning* and has the potential to provide more robust and accurate models.

Despite this promise, most existing approaches make two strong assumptions: paired samples are available across modalities during training, and all modalities are accessible at inference. In practice, neither assumption holds. Data are often collected independently by different institutions, coverage is disjoint, sensors are heterogeneous, and privacy constraints limit sharing. As a result, multimodal datasets are frequently *unpaired*, fragmented, and distributed.

Existing strategies often fall into two extremes. On one side, an unimodal training approach which learns models from a single sensor or modality, without requiring any alignment or collaboration (Helber et al., 2019; Sumbul et al., 2021). On the other side, fully paired fusion that improves accuracy, yet requires costly alignment, depends on scarce co-acquisitions, and typically assumes multimodal inputs at inference (Schmitt et al., 2019; Baltrušaitis et al., 2018). A new line of work explores intermediate solutions that reduce the reliance on pairing. Some methods combine a small set of paired samples with a large unpaired corpus to align shared representations (Yacobi et al., 2025). Others address missing modalities during training or inference by designing imputation strategies or learning models that are robust to incomplete inputs (Wu et al., 2024). There are also approaches for unpaired alignment through pseudo-pairs, cycle consistency, or distillation, particularly in medical imaging and multimodal representation learning (Dou et al., 2020;

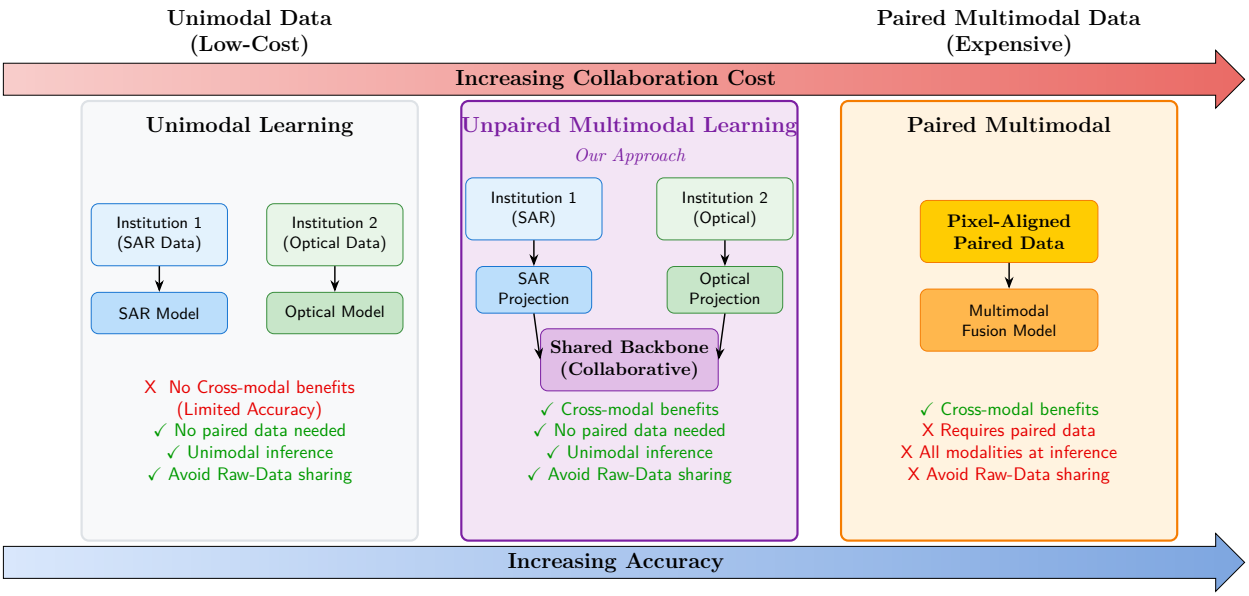


Figure 1: Motivation for unpaired multimodal collaborative learning. At one end, unimodal training is inexpensive and avoid data sharing but fails to exploit complementary information. At the other end, paired multimodal fusion improves accuracy but requires costly aligned data and multimodal inference. Our framework targets the middle ground, enabling cross-modal benefits without paired data or multimodal inputs at test time.

Timilsina et al., 2024). These efforts demonstrate the importance of the unpaired setting, but they typically assume partial pairing, multimodal inference, or homogeneous encoders. In contrast, our goal is to develop a *practical framework for fully unpaired multimodal collaboration* that requires no paired samples, supports unimodal inference, and is simple enough to deploy in realistic distributed settings.

This motivates our central question: *Can we transfer knowledge across modalities without paired data, without multimodal inference, and while respecting heterogeneous architectures and privacy constraints?* Figure 1 illustrates this trade-off.

Motivating example. Satellite imagery provides a concrete example. Sentinel-1, which carries a *Synthetic Aperture Radar (SAR)* instrument, captures surface structure and is robust to clouds and illumination. Sentinel-2, in contrast, provides multispectral optical imagery with high spatial detail[1]. Each modality is useful on its own, but when paired, they provide shared and unique information. In practice, paired acquisitions across these missions are rare, co-registration is error-prone, and institutions often cannot share raw data. Similar challenges arise in medical imaging and robotics.

Our approach. We propose a framework that enables cross-modal knowledge transfer without requiring paired samples or multimodal inference. Each modality is equipped with a projection into a shared representation, where a backbone learns modality-agnostic semantics. After training, *post-hoc batch normalization calibration* adapts the backbone to each modality, yielding strong unimodal performance.

Contributions. This paper makes three contributions:

- **Problem formulation.** We define *Unpaired Multimodal Learning (UML)* as the task of leveraging semantically related but unaligned data across modalities, without requiring paired samples or multimodal inference. Unlike prior settings that assume partial pairing, missing-modality models, or homogeneous encoders, UML captures realistic constraints faced in satellite imagery and other distributed domains.

[1]**Sentinel missions.** Sentinel-1 (radar) and Sentinel-2 (optical) are part of the European Copernicus program, which provides freely available global Earth observation data at large scale (Torres et al., 2012; Drusch et al., 2012).

- **Method.** We propose a lightweight collaborative framework for UML that combines modality-specific projections, a shared backbone, and post-hoc batch normalization (BN) calibration (Ioffe & Szegedy, 2015). The design requires no paired data, supports unimodal inference, and we can perform federated training across institutions while not sharing raw data.

- **Empirical findings.** Across three satellite benchmarks and additional visual datasets, our approach consistently improves over unimodal baselines. Gains are largest for weaker modalities and in low-data regimes, and we show that BN calibration is critical to stable performance. These results establish clear principles for when and why collaboration is most beneficial.

Together, these contributions establish a practical and general framework for multimodal collaboration under the realistic constraints of unpaired, heterogeneous, and distributed data.

2 Problem Formulation

We consider a collaborative learning scenario with K institutions, where each institution $k \in \{1, 2, \ldots, K\}$ holds data from a distinct modality. Let $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ denote the dataset at institution k, where $x_i^k \in \mathcal{X}_k$ represents an input sample from modality k, $y_i^k \in \mathcal{Y}$ is the corresponding label, and N_k is the number of samples. Each input space $\mathcal{X}_k$ is heterogeneous, with $\mathcal{X}_k \subset \mathbb{R}^{d_k}$, where the dimensionality d_k may vary across modalities.

Key constraint (Unpaired Samples). The datasets are *unpaired* across modalities. This means we do not possess aligned instances (x^j, x^k) that observe the exact same physical entity or location. This eliminates the possibility of pixel-wise or sample-wise alignment.

Objective. Each institution aims to learn an improved classifier $h_k(x^k; \omega_k) : \mathcal{X}_k \to \mathcal{Y}$ for its own modality, where h_k is a neural network parameterized by ω_k. The goal is to achieve this improvement by collaborating with other institutions, without sharing raw data or requiring paired samples.

Key assumptions. We make two key assumptions that enable effective collaboration in this challenging setting:

Assumption 1 (Semantic Coherence). Although data are unpaired, modalities capture semantically related phenomena and share common high-level semantic structures. Formally, samples from different modalities can be mapped to a shared semantic space $\mathcal{S}$ via functions $\psi_k : \mathcal{X}_k \to \mathcal{S}$. Crucially, these semantic mappings are highly non-injective (many-to-one). Because multiple diverse inputs map to the same semantic concept, the sample-level inverse mapping ψ_k^{-1} is undefined. Therefore, one cannot construct a valid sample-to-sample mapping via composition $(\psi_k^{-1} \circ \psi_j)$, preserving the strictly unpaired nature of the datasets. In our satellite imagery context, this assumption holds because both SAR and optical sensors observe the same Earth surface phenomena, unlike more disparate modality combinations (*e.g.*, text-image).

Assumption 2 (Shared Label Space). All institutions operate on the same classification task with identical label space $\mathcal{Y}$. This enables knowledge transfer through supervised learning signals without requiring explicit sample correspondences. While we focus on a shared label space to isolate the effects of multimodal collaboration, this assumption can be relaxed in practice by using task-specific classifier heads for heterogeneous taxonomies.

These assumptions are realistic in many collaborative scenarios: (i) Earth observation, where different institutions collect complementary sensor data (SAR, optical, hyperspectral) for the same land cover classification objectives; (ii) medical imaging, where hospitals may specialize in different modalities (CT, X-ray, MRI) for the same diagnostic task (*e.g.*, lung disease classification), but acquiring paired scans from the same patient across all modalities is rare due to cost and patient burden.

Algorithm 1 Centralized Unpaired Multimodal Learning

Require: epochs E, batch size b, number of modalities K, learning rate η
Require: per-modality datasets $\{\mathcal{D}_k\}_{k=1}^{K}$ with $|\mathcal{D}_k| = N$ (balanced); hence $M \triangleq N/b$ mini-batches per epoch
Initialize: shared backbone $g(\cdot;\theta)$; modality-specific projections $\{f_k(\cdot;\phi_k)\}_{k=1}^{K}$
for $e = 1$ to E **do** ▷ epoch
  Shuffle each $\mathcal{D}_k$ and form M mini-batches of size b
  for $m = 1$ to M **do** ▷ mini-batch within epoch
    for $k = 1$ to K **do**
      Sample mini-batch $\{(x_i^k, y_i^k)\}_{i=1}^{b}$ from $\mathcal{D}_k$
      $\mathcal{L}_k = \frac{1}{b}\sum_{i=1}^{b} \ell\Big(g\big(f_k(x_i^k;\phi_k);\theta\big),\, y_i^k\Big)$ ▷ calculate mini-batch loss of each modality
    Total loss (per step): $\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k$
    Update: $(\theta, \phi_1, \ldots, \phi_K) \leftarrow (\theta, \phi_1, \ldots, \phi_K) - \eta\, \nabla_{(\theta,\phi_1,\ldots,\phi_K)}\mathcal{L}$
Post-training: Perform Algorithm 2 for BN calibration

3 Proposed Approach

3.1 Model Architecture

Since our data consist of heterogeneous modalities distributed across many institutions, lets say K institutes, we decompose the modality-specific classifier $h_k(x^k;\omega_k)$ into two components: (i) **Modality-specific projection** $f_k(x^k;\phi_k) : \mathcal{X}_k \to \mathcal{Z}$, which maps raw input x^k to a shared latent space $\mathcal{Z}$; and (ii) **Shared backbone** $g_\theta : \mathcal{Z} \to \mathcal{Y}$, which performs classification in the common representation space. The complete model for modality k is:

$$h_k(x;\omega_k) = g_\theta\big(f_k(x;\phi_k)\big) \qquad \text{s.t.} \qquad x \in \mathcal{X}_k, \tag{1}$$

where $\omega_k = \{\phi_k, \theta\}$ denotes the set of parameters involved for modality k.

Design rationale. The projection f_k handles modality-specific characteristics (*e.g.*, different channel dimensions, sensor properties), while the shared backbone g learns modality-agnostic semantic features that generalize across modalities.

3.2 Training Objective

In the centralized setting, we minimize the empirical risk across all modalities:

$$\mathcal{L}(\phi_1, \ldots, \phi_k, \theta) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}\ell\Big(h_k\big(x_i^k\big), y_i^k\Big), \tag{2}$$

and $\ell(\cdot)$ is the cross-entropy loss. Joint optimization encourages g_θ to learn shared semantics across modalities, akin to multi-task learning where each modality is treated as a related task. Specifically, the shared backbone learns semantically aligned features by observing diverse modality inputs, enabling *implicit cross-modal knowledge transfer*. Weaker modalities benefit from richer representations learned from stronger ones, while stronger modalities gain robustness.

Training Procedure. Algorithm 1 presents our training procedure. A critical challenge in multimodal learning is that the batch normalization (BN) statistics computed during training may be biased toward dominant modalities or inappropriate for individual modalities during inference. To address this, we propose a simple yet effective post-training calibration procedure (Algorithm 2). Standard BN uses exponential moving averages to track running statistics during training. However, in the multimodal setting, these statistics represent a mixture across modalities and may not be optimal for any single modality during inference.

Algorithm 2 Post-hoc Batch Normalization (BN) Calibration

Require: Trained backbone $g(\cdot\,;\theta)$, projections $\{f_k(\cdot\,;\phi_k)\}_{k=1}^{K}$
Require: Calibration epochs E_{cal}, per-modality datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, batch size B
for $k = 1$ to K **do**
  $g_k(\cdot\,;\theta_k) \leftarrow \text{copy}(g(\cdot\,;\theta))$ ▷ independent copy
  Freeze all parameters in g_k and f_k ▷ weights fixed
  Reset BN running statistics in g_k and f_k to zero
  Set BN layers to accumulate statistics without momentum (with CMA)
  for $e = 1$ to E_{cal} **do**
    for mini-batch $\{x_i^k\}_{i=1}^{B} \sim \mathcal{D}_k$ **do**
      $Z \leftarrow f_k(\{x_i^k\}_{i=1}^{B}; \phi_k)$ ▷ batch processing
      $g_k(Z; \theta_k)$ ▷ forward only, updates BN stats
  Output: Calibrated model $h_k = g_k \circ f_k$

After training, we create modality-specific copies of the shared backbone and recalibrate their BN layers using only data from the corresponding modality. Specifically, we freeze all learned parameters to preserve trained weights, reset BN running statistics, and disable exponential moving averages Finally, we recompute statistics using cumulative moving averages (CMA) over multiple calibration epochs ($E_{cal} \ll E$). We use a larger batch size $B \gg b$ during this post-hoc step to minimize sampling noise and improve the accuracy of mean and variance estimates—an approach made possible here because weights are frozen, reducing GPU memory overhead. CMA ensures each batch contributes equally to the final estimate, resulting in more stable statistics than standard exponential moving averages. This procedure requires no additional parameter training and maintains privacy, as statistics are computed independently for each modality.

Benefits. The calibrated models $h_k(\cdot;\omega_k)$ maintain the representational power of the shared backbone while providing modality-appropriate normalization statistics, leading to improved performance without additional learnable parameters or privacy concerns.

Connection to multi-task learning. Our training objective resembles multi-task learning with a shared backbone, where each modality constitutes a distinct "task." This similarity is intentional—it allows our framework to inherit the well-established regularization benefits of multi-task learning, where the shared parameters prevent overfitting to any single modality while learning generalizable representations across tasks.

Federated Extension. Our approach naturally extends to federated learning settings (Algorithm 3). The key insight is selective parameter sharing: only the shared backbone parameters θ are aggregated across clients using FedAvg, while modality-specific projections(ϕ_k remain local to preserve data heterogeneity and privacy. This design ensures that raw data never leaves client devices, with only model parameters being exchanged. Following federated training, each client performs the same BN calibration procedure (Algorithm 2) using local data to obtain personalized models. Our framework is agnostic to the choice of federated aggregation algorithm (FedProx, SCAFFOLD) and optimizer (Adam, SGD), making it broadly applicable to various federated scenarios.

3.3 Theoretical Intuition

Why this works. The shared backbone g observes diverse feature representations from all modalities during training, acting as a regularizer that prevents overfitting to any single modality's characteristics. Simultaneously, the common classification objective provides a supervisory signal that aligns class-level decision structures across modalities without requiring explicit sample correspondences.

Cross-modal knowledge transfer. Weaker modalities benefit from the richer representations learned by stronger modalities through the shared backbone, while stronger modalities gain robustness through exposure to diverse feature patterns.

BN calibration necessity. Collaborative training causes Batch Normalization (BN) statistics to drift toward dominant modalities. Post-hoc recalibration using modality-specific data re-centers these distributions, ensuring the global model is correctly mapped to each client's specific feature manifold.

4 Experimentation

4.1 Datasets and Experimental Setup

We evaluate on three multimodal Earth observation benchmarks: *BigEarthNet-MM* (Sumbul et al., 2021), *EuroSAT-S1-RGB* (Helber et al., 2019; Wang et al., 2023), and *SEN12MS* (Schmitt et al., 2019). All datasets are originally imbalanced and multi-label; we construct class-balanced subsets and recast them as single-label classification to isolate unpaired multimodal learning effects. Each dataset provides Sentinel-1 (S1) SAR and Sentinel-2 (S2) multispectral imagery with varying spectral richness: BigEarthNet-MM (2 SAR + 12 S2 bands), SEN12MS (2 + 13), and EuroSAT-S1/RGB (2 + 3 RGB). The complete data set statistics and band descriptions are in Appendix D and E.

Training protocol. We implement our framework with Algorithm 1 primarily in a *centralized* setting. All methods use identical ResNet-18 (He et al., 2015) computational budgets: unimodal baselines train one ResNet-18 per modality, while our method decomposes ResNet-18 into modality-specific projections f_k plus shared backbone g with matching total parameters. Post-training, we recalibrate BatchNorm statistics per modality using local data. We control for all other factors: optimizer, schedule, augmentations, epochs, and early stopping (details in Appendix 7). We report top-1 accuracy (correct predictions/total samples), which is appropriate given the balanced nature of all datasets.

Experimental configurations. (i) *Fine-grained*: Each spectral band/channel becomes a separate client ($K = 5$ to 15), demonstrating scalability; (ii) *Bi-modal*: The standard setting of two clients ($K = 2$), reflecting practical deployment scenarios.

4.2 Main Results: Cross-Modal Collaboration Benefits

Fine-grained collaboration (Experiment 1). Figure 2a and Figure 5 compare unimodal baselines (blue) to our method (orange) for every band/channel. Our approach improves mean test accuracy on all three datasets: **+13.84** percentage points (pp) on BigEarthNet-MM, **+4.59** pp on SEN12MS, and **+6.24** pp on EuroSAT-S1/RGB.

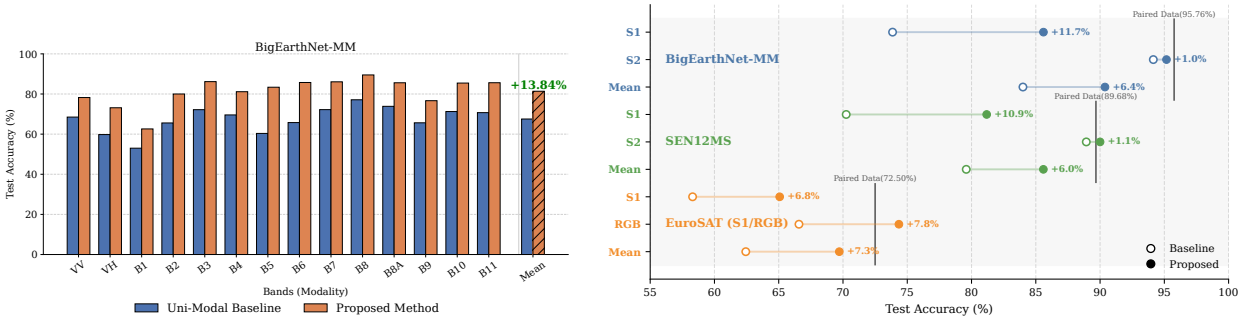
Bi-modal collaboration (Experiment 2). Figure 2b summarizes the bi-modality case. S2 baselines are already high on BigEarthNet-MM and SEN12MS, so adding S1 yields marginal uplifts for S2. In contrast, S1 benefits greatly from collaboration: **+11.7** pp (BigEarthNet-MM), **+10.9** pp (SEN12MS), and **+6.8** pp (EuroSAT-S1-RGB). On EuroSAT, RGB also improves by **+7.8** pp. Vertical gray markers indicate paired-data references where available; in SEN12MS and EuroSAT, the stronger modality occasionally exceeds that paired, consistent with reduced overfitting when collaborating without pixel-level pairing. *Takeaway:* collaboration predominantly lifts the weaker modality, while already-informative S2 bands see small but stable changes.

4.3 Analysis: Sources of Improvement

Regularization vs. semantic transfer. We disentangle two potential mechanisms through controlled experiments (Tables 1, 2). Collaborating semantically related modalities (S1 ↔ S2, MNIST ↔ SVHN) yields larger gains than unrelated pairs (satellite ↔ natural images, digits ↔ fashion). However, even semantically distant collaborators provide positive regularization effects (*e.g.*, SAR+Imagenette[2]: +9.1 pp vs. SAR+optical: +11.7 pp), confirming that both mechanisms contribute.

Data efficiency and BN calibration. Figure 3 demonstrates that collaboration benefits are most pronounced in low-data regimes, with diminishing returns as per-modality data increases. Removing BN cal-

[2]ImgNette is a subset of Imagenet Dataset containing natural images: `https://github.com/fastai/imagenette`



(a) Experiment 1: BigEarthNet-MM (fine-grained)

(b) Experiment 2: Bi-modal (S1 vs. S2/RGB)

Figure 2: Results of unpaired multimodal learning. (Left) Experiment 1: BigEarthNet-MM fine-grained modalities, showing consistent gains over unimodal baselines (blue vs. orange). (Right) Experiment 2: Bi-modal setting, where Sentinel-1 (SAR) benefits strongly from collaboration. Results for SEN12MS and EuroSAT-S1/RGB in Experiment 1 are provided in the Appendix B.

Method	BigEarthNet-MM			SEN12MS			EuroSAT S1–RGB		
	S1	S2	Mean	S1	S2	Mean	S1	RGB	Mean
Unimodal Baseline	73.86	94.14	84.00	70.25	88.92	79.59	58.30	66.57	62.44
ISCATimilsina et al. (2024)	80.76	93.45	87.10	70.67	89.29	79.98	62.75	69.35	66.05
Proposed	**85.59**	**95.16**	**90.38**	**81.17**	**90.00**	**85.59**	**65.07**	**74.35**	**69.71**
Paired Data (Multimodal inference)	-		95.76	-		89.68	-		72.50

Table 3: Comparison with unpaired multimodal learning methods. Classification accuracy (%) across remote sensing datasets. Our method outperforms ISCA, the only existing unpaired multimodal approach, and approaches the performance of paired multimodal methods (shown in gray for reference)

ibration severely degrades performance across all data scales, with larger degradation at high data volumes—consistent with BN statistics drifting toward dominant modalities during collaborative training.

4.4 Baseline Comparisons

Unpaired multimodal methods. We compare our approach against Identifiable Shared Component Analysis (ISCA), the only existing method to our knowledge designed for completely unpaired multimodal learning. As shown in Table 3, our method consistently outperforms ISCA across all datasets and modalities.

Domain adaptation baselines. Standard DA methods assume shared encoders and often unsupervised targets. We adapt DANN, CDAN, MCC, and MDD to our supervised, heterogeneous-modality setting by using separate encoders per modality with shared classifiers and providing the labels for both modalities(domains). Table 4 shows that our method outperforms all DA variants. Critically, several DA methods underperform unimodal baselines, indicating that standard domain alignment objectives are ill-suited for heterogeneous modalities with different architectural requirements.

Federated Learning Extension As discussed earlier, our proposed method naturally extends to privacy-aware (no raw data exchange) federated settings, as demonstrated in Algorithm 3. Figure 4 shows performance across different local epochs ($L \in \{2, 5, 10, 25, 50\}$). Our method maintains effectiveness across communication-efficiency trade-offs, with optimal performance typically achieved around $L = 5$–10 depending on dataset characteristics. Higher local epoch values reduce communication frequency between clients and the server. We denote the number of communication rounds by R, where each round involves clients sharing their local backbone weights θ_k for aggregation. For fair comparison with the centralized setting (200 training epochs), we set $R \in \{100, 40, 20, 8, 4\}$ such that $R \times L = 200$ remains constant across all federated experiments (see Appendix 7 for details).

Evaluated Modality	Supplementary Modality		
	MNIST	SVHN	FMNIST
MNIST	-	+1.00	-1.00
SVHN	+3.67	-	+2.00
FMNIST	-2.00	+1.67	-

Table 1: Cross-modal collaboration benefits on digit datasets.Delta accuracy (%) when adding supplementary modalities. Semantically similar datasets (MNIST+SVHN) show mutual benefits

Evaluated Modality	Supplementary Modality		
	BGE-S1	BGE-S2	ImgNette
BGE-S1	-	+11.7	+9.1
BGE-S2	+1.0	-	+0.8
ImgNette	+8.4	+8.7	-

Table 2: Cross-modal collaboration on remote sensing data. Delta accuracy (%) with supplementary modalities. BGE-S1/S2 gain most from each other due to semantic similarity.

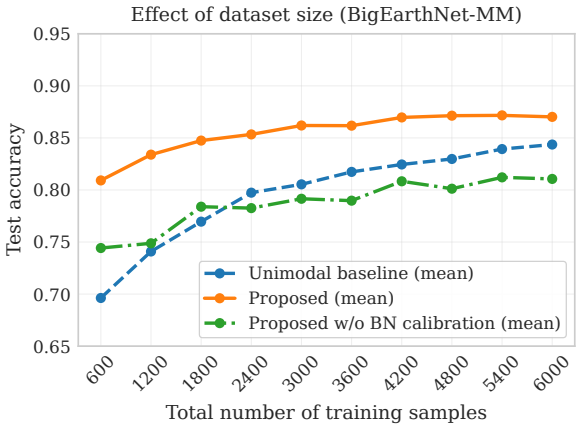
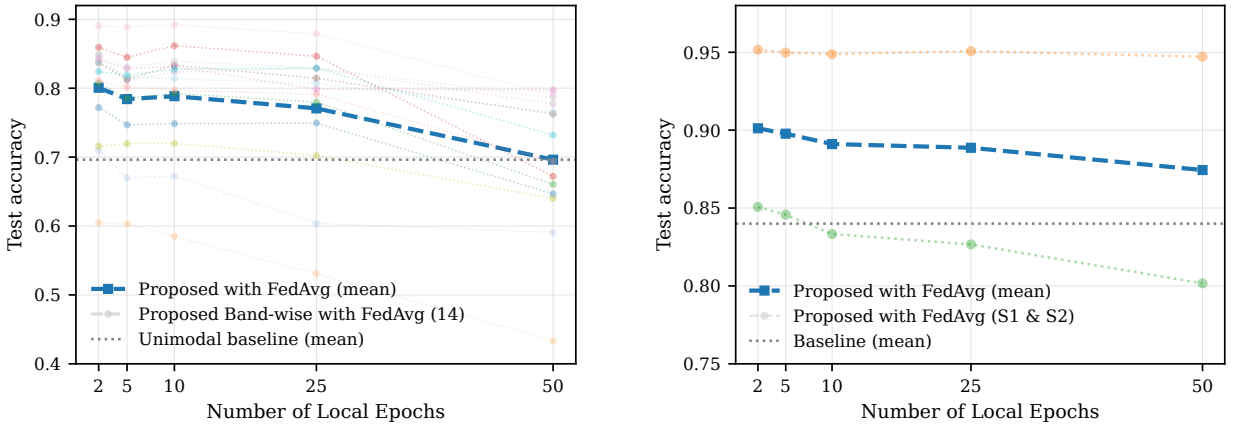


Figure 3: Effect of dataset size on BigEarthNet-MM. Collaboration yields larger gains in the low-data regime, while BN calibration remains critical at all scales.

Method	BigEarthNet-MM			SEN12MS			EuroSAT S1–RGB		
	S1	S2	Mean	S1	S2	Mean	S1	RGB	Mean
Unimodal Baseline	73.86	94.14	84.00	70.25	88.92	79.59	58.30	66.57	62.44
DANN	71.31	94.09	82.70	69.71	88.79	79.25	58.67	68.25	63.46
CDAN	77.81	93.76	85.79	71.25	86.18	78.71	58.07	68.92	63.50
MCC	76.95	93.47	85.21	70.00	88.64	79.32	59.67	66.00	62.84
MDD	70.36	94.16	82.26	67.25	87.29	77.27	62.17	66.42	64.30
Proposed	**85.59**	**95.16**	**90.38**	**81.17**	**90.00**	**85.59**	**65.07**	**74.35**	**69.71**

Table 4: **Comparison with domain adaptation baselines.** Classification accuracy (%) across three remote sensing datasets. Our method consistently outperforms DA approaches and unimodal baselines. Results demonstrate that our framework achieves superior cross-modal knowledge transfer compared to traditional domain adaptation techniques that require source-target domain alignment.

Limitations. Our approach has several important limitations that suggest directions for future work. First, the benefits of collaborative learning diminish as per-modality data becomes abundant (Figure 3). Second, our experimental evaluation focuses on CNN architectures, which are well-suited to the limited-data scenarios that motivate our approach. Modern transformer architectures (Dosovitskiy et al., 2021) with LayerNorm (Ba et al., 2016) may not require BN calibration, though they typically require substantially larger datasets for effective training from scratch. Third, our evaluation primarily considers balanced per-modality data distributions. In severely imbalanced scenarios—where one modality contains orders of magnitude more data than others—the collaboration dynamics between modalities may shift significantly, potentially leading to domination effects that our current framework does not explicitly address. Fourth, our method is designed for the purely unpaired setting and lacks a principled mechanism to leverage partially paired data when available. If some pixel-wise aligned samples exist across modalities, our current framework cannot systematically incorporate this valuable supervisory signal, representing a missed opportunity for improved performance in hybrid scenarios. Finally, we assume a shared label space; extending this to heterogeneous taxonomies or label skew would require replacing the common classifier with institution-specific heads.



(a) Extension of Experiment 1 (fine-grained)

(b) Extension of Experiment 2 (bi-modal)

Figure 4: Results on BigEarthNet-MM dataset showing the effect of local epochs L on test accuracy under FedAvg with fixed total training budget ($R \times L = 200$). Our method maintains superiority over unimodal baselines across different communication frequencies, with optimal performance at $L = 5$ to 10. Higher local epochs reduce communication overhead but may degrade performance due to client drift, with convergence to baseline performance at $L = 50$ in both scenarios.

5 Related Work

Multimodal representation learning. Multimodal learning is the process of jointly leveraging information from multiple data modalities (*e.g.*, images, text, audio, or sensor signals) to learn richer and more robust representations than from any single modality alone (Baltrušaitis et al., 2018; Liang et al., 2023). A central challenge is dealing with heterogeneity across modalities and the need for alignment. Traditional fusion methods can be categorized into early, middle, and late fusion. Early fusion concatenates modalities at the input level, middle fusion shares intermediate representations, and late fusion aggregates modality-specific predictions. These approaches often assume paired data and require all modalities during inference. Recent dual-encoder frameworks such as CLIP (Radford et al., 2021), OneLLM (Han et al., 2024), VLMo (Bao et al., 2022), SIMVLM (Wang et al., 2021), ImageBind (Girdhar et al., 2023), VL-GPT (Zhu et al., 2023), and CROMA (Fuller et al., 2023) address modality heterogeneity by assigning each modality its own encoder and aligning representations via contrastive objectives. While effective across heterogeneous modalities, these methods still depend on paired data for training in contrast to our method where we do not need any paired data.

Domain Adaptation Our problem of unpaired multimodal learning shares connections with domain adaptation (DA), where the goal is to align feature spaces across source(s) and target(t) domains so that $f(x^{(s)})$ and $f(x^{(t)})$ yield consistent representations when semantically similar (Wilson & Cook, 2020). Popular DA methods include adversarial alignment, *e.g.*, DANN (Ganin et al., 2016) and CDANN (Long et al., 2018), and discrepancy-based approaches such as MDD (Li et al., 2020a) and MCC (Jin et al., 2020). These methods encourage domain-invariant features through a single shared encoder and have proven effective for homogeneous domains. However, they are not directly applicable to multimodal settings, where each modality requires a distinct encoder due to heterogeneous input structures. In our work, we adapt representative DA baselines by equipping each modality with its own encoder. Empirically, these adapted methods struggle to close the modality gap, highlighting their design limitations for unpaired multimodal learning, whereas our approach achieves stronger cross-modal knowledge transfer (see Table 4).

Multimodal Learning with Missing or Unpaired Data Several works extend multimodal learning beyond the fully paired assumption. Nakada et al. (2023) introduce a contrastive framework that integrates unpaired samples into training. Kim & Kim (2024) propose predicting embeddings of missing modalities in the joint representation space to handle incomplete inputs during inference. Ma et al. (2021) address

scenarios with severely missing modalities using a meta-learning approach, while large-scale vision-language models such as Singh et al. (2022) leverage self-supervised learning to train on a mix of paired and unpaired data. Timilsina et al. (2024) introduce Identifiable Shared Component Analysis, which disentangles shared and private components from unpaired multimodal distributions to enable a joint classifier. We benchmark against this approach and find that our framework achieves stronger performance while avoiding reliance on paired data (see Table 3)

Federated learning with heterogeneous data Classical FL algorithms such as FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020b), and SCAFFOLD (Karimireddy et al., 2020) assume homogeneous architectures across clients, allowing parameter averaging. Personalization-oriented methods like FedPer Arivazhagan et al. (2019) and FedRep (Collins et al., 2021) relax this by sharing early layers or representations while keeping task-specific heads local, but they still assume a homogenous data input structure. More recent approaches, including HeteroFL (Diao et al., 2020) and LG-FedAvg (Liang et al., 2020), support heterogeneous client models or backbone splits, yet they are not designed for fundamentally different modalities. A key challenge in such settings is handling batch normalization (BN) statistics, which become inconsistent across modalities. In our proposed method, running statistics of BN layers are not tracked during training similar to *static Batch Normalization* (sBN) (Diao et al., 2020). We perform post-hoc BN calibration. This design has two main advantages: (i) it preserves privacy by avoiding the exchange of first- and second-order statistics across clients, thus addressing one of the key limitations highlighted in Diao et al. (2020); and (ii) it ensures that BN statistics are well aligned with the distributional characteristics of each modality. Notably, this differs from Li et al. (2021) where both the affine parameters and BN statistics are client-specific. In contrast, our approach keeps affine parameters shared while only calibrating BN statistics post-training.

6 Conclusion and Future Work

We addressed the problem of *Unpaired Multimodal Learning (UML)*, where data from different sensors or modalities are semantically related but not aligned across samples. While prior work has explored partial solutions, such as semi-paired training or missing-modality models, a practical framework for fully unpaired collaboration remained elusive. We proposed a simple yet effective approach that combines modality-specific projections, a shared backbone, and post-hoc BN calibration. This design enables cross-modal knowledge transfer without requiring paired data or multimodal inference, and scales naturally to distributed training across institutions without sharing raw data. Our experiments on three satellite benchmarks, complemented by digits and natural image datasets, demonstrate consistent gains over unimodal baselines, with the strongest improvements for weaker modalities and in low-data regimes. These results show that even lightweight architectural changes can unlock significant cross-modal benefits in realistic unpaired settings. Looking ahead, extensions to transformer backbones, alternative normalization schemes, and richer modality combinations promise to broaden the scope of this framework. This work is a step towards multimodal collaboration under the real-world constraints of unpaired, heterogeneous, and private data.

Reproducibility statement

We implement all experiments in PyTorch following official reproducibility guidelines[3] with fixed random seeds and deterministic operations. Complete hyperparameters are provided in Appendix 7, and code will be made publicly available upon acceptance.

Broader Impact Statement

This work introduces a privacy-preserving framework for collaborative multimodal learning that eliminates the need for raw data sharing or sample-level pairing. A primary positive impact of this approach is democratizing machine learning capabilities. It enables institutions with isolated or unimodal datasets to collectively train robust models while strictly maintaining data sovereignty, which is particularly beneficial for fields like Earth observation, healthcare, and scientific research where data silos are prevalent.

[3] `https://docs.pytorch.org/docs/stable/notes/randomness.html`

References

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=bydKs84JEyw`.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: ESA's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. doi: 10.1016/j.rse.2011.05.026.

Anthony Fuller, Koreen Millard, and James R Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European conference on computer vision*, pp. 464–480. Springer, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pp. 171–187. Springer, 2024.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):3918–3930, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020b.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023. URL `https://arxiv.org/abs/2209.03430`.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2302–2310, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4348–4380. PMLR, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL `http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 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, 2019. URL https://arxiv.org/abs/1906.07789.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15638–15650, 2022.

Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begum Demir, and Volker Markl. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, September 2021. ISSN 2373-7468. doi: 10.1109/mgrs.2021.3089174. URL http://dx.doi.org/10.1109/MGRS.2021.3089174.

Subash Timilsina, Sagar Shrestha, and Xiao Fu. Identifiable shared component analysis of unpaired multimodal mixtures, 2024. URL https://arxiv.org/abs/2409.19422.

Ramon Torres, Paul Snoeij, Detlef Geudtner, David Bibby, Michael Davidson, Erik Attema, Pascal Potin, Bert Rommen, Nicolas Floury, Michelle Brown, Isabel Traver, Pascal Deghaye, Bernhard Duesmann, Benito Rosich, Nuno Miranda, Carlo Bruno, Marcello L'Abbate, Riccardo Croci, Adriano Pietropaolo, Matthias Huchler, and Fabrice Rostan. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120: 9–24, 2012. doi: 10.1016/j.rse.2011.05.028.

Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint arXiv:2310.18653*, 2023.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), July 2020. ISSN 2157-6904. doi: 10.1145/3400066. URL https://doi.org/10.1145/3400066.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

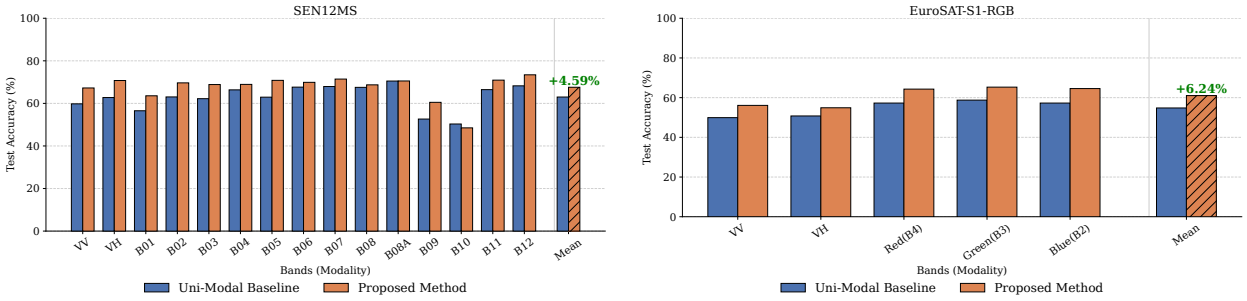
Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. *arXiv preprint arXiv:2505.21524*, 2025.

Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. VL-GPT: A generative pre-trained transformer for vision and language understanding and generation, 2023. URL https://arxiv.org/abs/2312.09251.

A Appendix

B Additional Experiment 1 Results

In the main text (Section 4.2), we presented Experiment 1 results on BigEarthNet-MM and discussed the general trends across datasets. For completeness, Figure 5 provides the corresponding results for SEN12MS and EuroSAT-S1/RGB. These follow the same pattern: our method consistently improves over mean unimodal baselines.



(a) SEN12MS

(b) EuroSAT S1–RGB

Figure 5: Experiment 1 results for SEN12MS and EuroSAT-S1/RGB. Unimodal baselines (blue) vs. proposed method (orange). Trends are consistent with those reported in the main text.

C Federated extension of unpaired multimodal learning

We demonstrate the Federated learning extension of our proposed method in Algorithm 3. We also performed experiments on the BigEarthNet-MM dataset. The results are shown in Figure 4.

Algorithm 3 Federated Unpaired Multimodal Learning

Require: Communication rounds R, local epochs L, learning rate η batch size b
Require: per-modality datasets $\{\mathcal{D}_k\}_{k=1}^{K}$ with $|\mathcal{D}_k| = N$ (balanced); hence $M \triangleq N/b$ mini-batches per epoch
Initialize: Shared backbone parameters $g(\cdot\,;\theta^0)$, modality-specific projections $\{f_k(\cdot;\phi_k)\}_{k=1}^{K}$
for round $r = 1$ to R **do**
  Server broadcasts $g(\cdot;\theta^{r-1})$ to all clients
  for each client k **in parallel do**
    Receive $g(\cdot;\theta^{r-1})$ from server
    Initialize local backbone: $g(\cdot;\theta^{k,r}) \leftarrow g(\cdot;\theta^{r-1})$
    for local epoch $l = 1$ to L **do**
      Shuffle each $\mathcal{D}_k$ and form M mini-batches of size b
      for $m = 1$ to M **do** ▷ mini-batch within epoch
        Compute loss: $\mathcal{L}_k = \frac{1}{b}\sum_{i=1}^{b} \ell\Big(g\big(f_k(x_i^k;\phi_k);\theta^{k,r}\big),\, y_i^k\Big)$
        Update: $\phi^k, \theta^{k,r} \leftarrow (\phi^k, \theta^{k,r}) - \eta\nabla_{(\theta^{k,r})}\mathcal{L}_k$
    Send $\theta^{k,r}$ to server ▷ Only backbone parameters
    Server aggregates: $\theta^r \leftarrow \frac{1}{K}\sum_{k=1}^{K}\theta^{k,r}$ ▷ FedAvg
post-training: Perform **Algorithm 2** for BN calibration

D Information about Bands

For completeness, we list the spectral bands of Sentinel-1 and Sentinel-2 along with their spatial resolution and typical applications. The information was compiled from Sentinel Hub documentation[4] [5] and TorchGeo dataset documentation[6].

[4] https://custom-scripts.sentinel-hub.com/sentinel-2/bands/
[5] https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel/sentinel-1/
[6] https://torchgeo.readthedocs.io/en/stable/api/datasets.html#bigearthnet

Satellite	Bands	Pixel size (m)	Typical Application
Sentinel-2	B01 (Aerosol)	60	Aerosol detection
	B02 (Blue)	10	Visible range (RGB)
	B03 (Green)	10	Visible range (RGB)
	B04 (Red)	10	Visible range (RGB)
	B05 (Red Edge 1)	20	Vegetation
	B06 (Red Edge 2)	20	Vegetation
	B07 (Red Edge 3)	20	Vegetation
	B08 (NIR)	10	Shorelines, biomass
	B8A (Narrow NIR)	20	Vegetation
	B09 (Water Vapour)	60	Water vapour detection
	B10 (Cirrus)	60	Cloud detection
	B11 (SWIR 1)	20	Snow, moisture
	B12 (SWIR 2)	20	Snow, moisture
Sentinel-1	VV	10	Texture, backscatter
	VH	10	Moisture, vegetation structure

Table 5: Spectral bands of Sentinel-1 and Sentinel-2 with pixel size and typical applications.

E Information about Datasets Used

In this section, we provide details of the datasets employed in our experiments. We describe the modalities, selected classes, and experimental splits. Table 6 summarizes the dataset statistics, and Figure 6 shows representative examples from the visible spectrum.

BigEarthNet-MM. The BigEarthNet-MM dataset consists of co-registered Sentinel-1 (SAR) and Sentinel-2 (multispectral optical) image patches.

- **Features:** Sentinel-1 provides two polarization bands (VV, VH), while Sentinel-2 provides 12 spectral bands (B01–B12, excluding B10) with spatial resolutions of 10–60 m/pixel. All bands were upsampled to 10 m and cropped into 120×120 patches.
- **Format:** Each image patch is provided as multiple single-channel GeoTIFFs. Labels are originally multi-class, but we retain only single-label samples.
- **Classes:** We selected six representative classes: *Arable land*, *Broad-leaved forest*, *Coniferous forest*, *Marine waters*, *Pastures*, and *Urban fabric*.

Figure 6(a) shows a visible-spectrum sample, and Table 6 lists the distribution.

SEN12MS. SEN12MS is a large-scale dataset of Sentinel-1 and Sentinel-2 patches annotated with MODIS land cover labels. We utilize only single-label International Geosphere-Biosphere Program (IGBP) labels provided by the authors.

- **Features:** Sentinel-1 provides VV and VH backscatter (dB scale), while Sentinel-2 provides 13 spectral bands (B01–B12). Patches are 256×256.
- **Preprocessing:** The original dataset is seasonally partitioned; we retain only the summer subset.
- **Classes:** We select seven land cover types: *Evergreen broadleaf forest*, *Open shrublands*, *Savannas*, *Grasslands*, *Croplands*, *Urban and built-up*, and *Water bodies*.

A representative sample is shown in Figure 6(b).

Dataset	Classes	Train	Validation	Test	Resolution
BigEarthNet-MM	6	1800 (100)	400	700	120×120
SEN12MS	7	300 (100)	400	400	256×256
EuroSAT S1–RGB	10	1000 (100)	400	400	64×64

Table 6: Dataset statistics used in our experiments. The numbers in parentheses denote the reduced training samples used in low-data settings (Experiment 1 and Experiment 2).

EuroSAT S1–RGB. The EuroSAT dataset contains Sentinel-2 imagery with 10 target classes. For our experiments, we also include Sentinel-1 SAR patches aligned to the same grid, forming an S1–RGB subset. Original EuroSAT has only RGB images, we obtain the SAR from Wang et al. (2023).

- **Features:** RGB bands (B02, B03, B04) from Sentinel-2 and SAR (VV, VH) from Sentinel-1. Each patch is 64×64.
- **Classes:** All 10 original classes are retained, including *Annual Crop, Forest, Herbaceous Vegetation, Residential, Sea Lake, Highway,Permanent Crop, Industrial, River, Pasture.*

Figure 6(c) illustrates an RGB example.

ImgNette. ImgNette is a curated subset of ImageNet designed to reduce label noise and simplify evaluation.

- **Features:** RGB natural images, resized to 224×224.
- **Classes:** Ten classes corresponding to high-level object categories. For our work, shown in 2, we only used 6 classes.

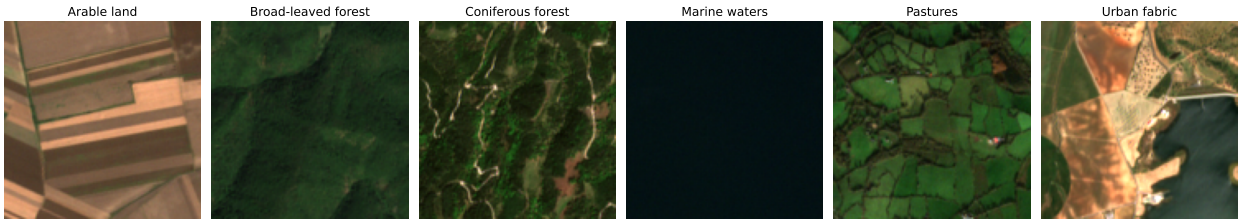
MNIST (LeCun, 1998), SVHN (Netzer et al., 2011), and FMNIST (Xiao et al., 2017). We additionally include canonical vision benchmarks for results shown in 1:

- **MNIST:** Grayscale handwritten digits (10 classes, 28×28).
- **SVHN:** RGB house numbers trimmed from Google Street View (10 classes, 32×32). We crop these image to 28×28 while collaborating with MNIST and Fashion-MNIST.
- **Fashion-MNIST:** Grayscale fashion items (10 classes, 28×28).

F Size of the projection layers vs accuracy

We conducted an ablation to study how the capacity of the modality-specific projection versus the shared backbone affects performance. Specifically, we varied the number of ResNet basic blocks allocated to the projection layer, while keeping the total capacity (projection + backbone) equal to a ResNet-18. As shown in Figure 7, when the projection layer is too deep (and the backbone correspondingly shallow), performance drops significantly (down to $\sim$69%), falling below the unimodal baseline. In contrast, allocating fewer blocks to the projection and more to the backbone yields higher test accuracy (blue curve). This highlights the importance of preserving sufficient depth in the shared backbone.

The orange curve shows results without batch normalization (BN) calibration. The gap clearly demonstrates that BN calibration is critical, especially when the backbone is deeper and contains more BN layers that can drift under multimodal training. Overall, this experiment emphasizes the need for careful capacity allocation between modality-specific and shared components, as well as the necessity of BN calibration.



(a) BigEarthNet-MM

(b) SEN12MS

(c) EuroSAT

Figure 6: Representative visible-spectrum examples for selected classes from each dataset.

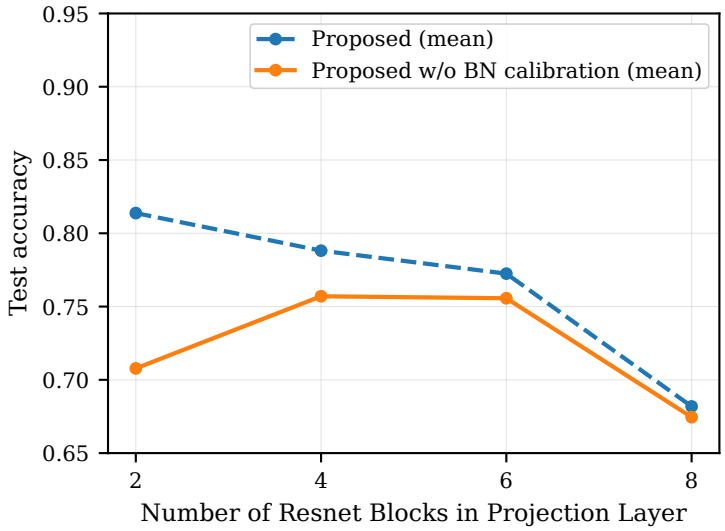


Figure 7: Effect of projection layer depth (in ResNet blocks) on BigEarthNet-MM fine-grained clients (Experiment 1). The blue curve shows results with BN calibration, while the orange curve shows results without.

G Use of Large Language Models

We used large language models (ChatGPT-5 and Claude Sonnet 4) as writing assistance tools to improve grammatical correctness and sentence clarity. The LLMs were employed solely for language polishing and did not contribute to research ideation, experimental design, technical methodology, or scientific content generation.

Listing 1: Architecture of the ResNet BasicBlock used in our experiments.

```
class BasicBlock(nn.Module):
    expansion = 1

    def __init__(self, in_planes, planes, stride=1):
        super(BasicBlock, self).__init__()
        self.conv1 = nn.Conv2d(
            in_planes, planes, kernel_size=3, stride=stride,
            padding=1, bias=False)
        self.bn1 = nn.BatchNorm2d(planes)
        self.conv2 = nn.Conv2d(
            planes, planes, kernel_size=3, stride=1,
            padding=1, bias=False)
        self.bn2 = nn.BatchNorm2d(planes)
        self.shortcut = nn.Sequential()
        if stride != 1 or in_planes != self.expansion * planes:
            self.shortcut = nn.Sequential(
                nn.Conv2d(in_planes, self.expansion * planes,
                          kernel_size=1, stride=stride, bias=False),
                nn.BatchNorm2d(self.expansion * planes)
            )

    def forward(self, x):
        out = F.relu(self.bn1(self.conv1(x)))
        out = self.bn2(self.conv2(out))
        out += self.shortcut(x)
        out = F.relu(out)
        return out
```

H Hyperparameters

The hyperparameters selected for different experiments in our Unpaired multimodal learning are shown in the Table 7

Table 7: Hyperparameter selection for different experimental settings on BigEarthNet-MM.

Hyperparameter	**Unimodal**	**Unpaired Multimodal (Exp. 1 & 2)**	**Federated UML**
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.01	0.01	0.01
Initial Learning Rate	0.001	0.001	0.001
Batch Size	32	32	32
Scheduler	Step	Step	Step
Total Epochs	200	200	200
Scheduler (step, decay)	(150, 0.1)	(150, 0.1)	(150, 0.1)
Augmentations	None (Norm. only)	None (Norm. only)	None (Norm. only)
Early Stopping	Best Val ACC after 190 ep.	Best Val ACC after 190 ep.	Best Val ACC after $R \times L > 190$
ResNet blocks (projection)	8	2	2
ResNet blocks (backbone)	–	6	6
Communication Rounds R	–	–	(100, 40, 20, 8, 4)
Local Epochs L	–	–	(2, 5, 10, 25, 50)
BN Calibration Epochs	–	10	10
BN Calibration Batch Size	–	600	600