TRI-JEPA: INTERACTION-AWARE SELF-SUPERVISED MULTIMODAL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised multimodal pretraining aims to learn representations from unlabeled multimodal data that transfer across downstream tasks. Different tasks, however, may rely on different multimodal interactions: information *shared* across modalities, information *unique* to a single modality, and *synergistic* information that emerges only when modalities are considered jointly. Since the downstream task is unknown at pretraining time, encouraging all three may yield a more broadly useful prior. Existing methods either emphasize only a subset of these interactions or capture them implicitly through objectives over fused multimodal representations. We introduce **Tri-JEPA**, an interaction-aware joint-embedding predictive architecture with three predictive objectives defined by different source-target configurations. For a target modality, the *shared* objective predicts its representation from another modality, the *unique* objective predicts it from the same modality, and the *synergistic* objective predicts it from all other modalities jointly. These assignments are cycled across modalities for balanced coverage. The tailored objectives enable us to examine how different multimodal interactions are encouraged during pretraining and how their contributions vary across downstream tasks. Under controlled settings, Tri-JEPA consistently outperforms the multimodal self-supervised baselines considered across diverse downstream tasks. Overall, our results show that interaction-aware predictive objectives are a useful design principle for self-supervised multimodal pretraining.

1 INTRODUCTION

Self-supervised learning (SSL) has become a foundation for pretraining models that learn transferable representations across a wide range of tasks (Chen et al., 2020; Caron et al., 2021; He et al., 2022). By learning from large unlabeled datasets, SSL models aim to produce general-purpose representations that serve as priors for downstream tasks where labels are scarce or expensive to obtain (Bommasani, 2021; Balestriero et al., 2023). Real-world scenarios, however, often go beyond a single modality: benefiting from multiple sensors or sources that capture related but distinct aspects of the same underlying phenomenon (Zong et al., 2024). In such multimodal settings, SSL pretraining seeks to learn representations that integrate cues across modalities and support downstream tasks beyond what any single modality could provide (Radford et al., 2021; Girdhar et al., 2023).

To understand what multimodal pretraining captures, it is useful to consider the structure of information across modalities (Fig. 1, left). Multimodal information can be characterized by three types of interactions (Williams & Beer, 2010; Liang et al., 2023a;c): *shared* information available across modalities, *unique* information specific to a single modality, and *synergistic* information that emerges only when modalities are considered jointly. This characterization matters because different downstream tasks may rely on different interactions. Some may depend on information consistently observable across modalities, such as semantic content shared between image and text (Radford et al., 2021). Others may depend on modality-specific signals, such as fine-grained details visible only in one sensor (Liang et al., 2023b; Dhakal et al., 2025). Still others may rely on information that emerges through joint processing of multiple modalities, as in interpreting communicative intent from both spoken words and tone (Yu et al., 2024; Xin et al., 2025). Since the downstream task is unknown during pretraining, it is unclear which interaction will be most *task-relevant*. Hence, encouraging diverse interactions may yield a more broadly useful prior, as interactions weakly represented during pretraining may be less accessible for downstream transfer (Fig. 1, right).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

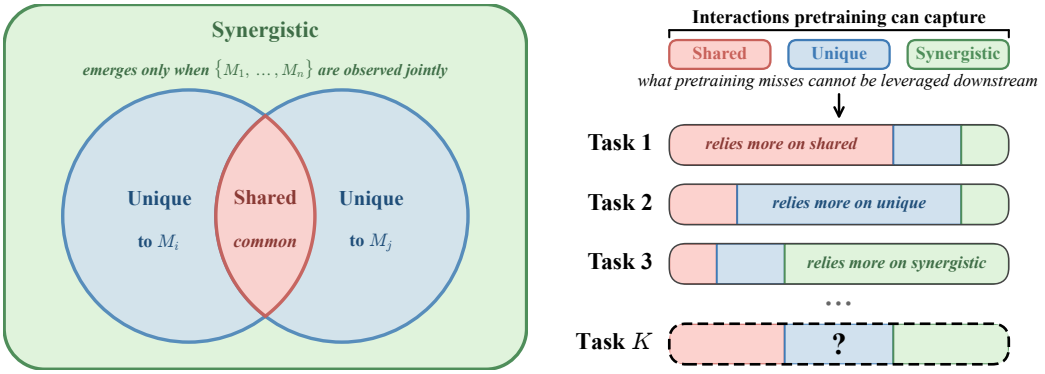


Figure 1: **(Left)** Three multimodal interactions during pretraining: *shared* information available across modalities, *unique* information specific to one modality, and *synergistic* information that emerges only when modalities are considered jointly. **(Right)** Different downstream tasks may depend on different interactions captured during pretraining. **(Takeaway)** Since the task is unknown during pretraining, aiming diverse multimodal interactions may yield a more broadly useful prior.

Prior multimodal SSL methods emphasize different interaction types. Contrastive approaches such as CLIP (Radford et al., 2021) and ImageBind (Girdhar et al., 2023) align paired modalities in a shared embedding space, emphasizing cross-modal information. These methods often rely on a multi-view redundancy assumption (Liang et al., 2023b), where shared information across modalities is sufficient for downstream transfer, which may be limiting when tasks depend on modality-specific or jointly emergent signals. FactorCL (Liang et al., 2023b) addresses this by factorizing representations into shared and unique components, while RANGE (Dhakal et al., 2025) augments contrastive embeddings with modality-specific features. More recently, CoMM (Dufumier et al., 2024) models shared, unique, and synergistic information jointly through contrastive learning over fused multimodal representations. InfMasking (Wen et al., 2025) extends this framework by masking modalities before fusion to strengthen learning from partial multimodal combinations.

Despite these advances, several questions remain open. Methods such as CoMM (Dufumier et al., 2024) and InfMasking (Wen et al., 2025) aim to capture shared, unique, and synergistic interactions within fused multimodal representations, but the three emerge jointly rather than through interaction-specific objectives. This makes it difficult to isolate how each interaction is encouraged during pretraining. InfMasking further strengthens synergy through an additional masking-based term, but still within the same fused framework. This raises a natural question: *can different multimodal interactions instead be encouraged through distinct pretext objectives?* Existing evidence is also largely post hoc and often based on synthetic benchmarks (Hermann & Lampinen, 2020), leaving their behavior on real-world data less clear. Finally, because the interactions remain entangled in the learned representation, it is difficult to determine how each contributes to downstream transfer and whether their importance changes across tasks. We organize our investigation around two pillars:

(1) Pretraining. *Can pretext objectives be designed to encourage shared, unique, and synergistic multimodal interactions during SSL pretraining?*

(2) Downstream transfer. *How do these objectives contribute to downstream performance, and how does their importance vary across tasks?*

We address these questions with **Tri-JEPA**, an interaction-aware multimodal joint-embedding predictive architecture built around three predictive objectives (LeCun et al., 2022; Assran et al., 2023). JEPA-style prediction operates in representation space rather than reconstructing raw inputs, avoiding commitment to input-space details that may be uninformative for downstream tasks (Balestriero & LeCun, 2024; Littwin et al., 2024; Van Assel et al., 2025), while also avoiding the strong augmentation assumptions of contrastive learning (Tian et al., 2020; Xiao et al., 2020). Tri-JEPA defines three source-target configurations. The *shared* objective predicts a target representation from another modality, encouraging cross-modal structure. The *unique* objective predicts it from the same modality, encouraging within-modality structure. The *synergistic* objective predicts it from all other

modalities jointly, encouraging the model to exploit information available from combined multimodal context. These assignments are cycled across modalities for balanced coverage.

We summarize our contributions:

- We propose **Tri-JEPA**, an interaction-aware multimodal SSL framework with cross-modal, within-modal, and joint-source predictive objectives that encourage shared, unique, and synergistic multimodal interactions (§3).
- We study how each multimodal interaction is emphasized during pretraining and how each contributes to downstream performance. We find that the objectives provide complementary, task-dependent contributions to downstream transfer (§5).
- Tri-JEPA consistently outperforms prior multimodal pretraining methods while using fewer pretraining epochs and less data, supporting interaction-aware predictive objectives as a useful design principle for multimodal pretraining (§5.2).

2 PRELIMINARIES

Goal of self-supervised learning (SSL). SSL aims to learn parameters θ_{pre} from an unlabeled dataset $\mathcal{D}_{pre}$ that serve as a useful prior for downstream tasks. Unlike supervised learning, where a single task label Y is provided and the model is optimized for that task, SSL pretraining derives its supervision from $\mathcal{D}_{pre}$ itself through a pretext task with loss $\mathcal{L}_{pre}$. The aim is not to minimize $\mathcal{L}_{pre}$ itself but to obtain parameters θ_{pre} that transfer well to downstream tasks $\{Y_k\}$. In multimodal SSL, $\mathcal{D}_{pre}$ contains co-occurring samples from n modalities $\{M_1, \ldots, M_n\}$ that capture related but distinct aspects of the same phenomenon, and the goal is to learn an encoder $f_{\theta_{pre}}$ that maps each modality to a latent representation $z_i = f_{\theta_{pre}}(M_i) \in \mathbb{R}^d$. After pretraining, the encoder serves as a prior for downstream tasks, each with its own labeled training data $\mathcal{D}_k$.

Pretraining objective and the prior. Since the downstream label Y is unavailable during pretraining, the pretext task derives its supervision from the modalities themselves. Unimodal SSL constructs this supervision from a single input through augmentation (Chen et al., 2020) or masking (He et al., 2022). Multimodal SSL has an additional supervision signal in the data: each modality is an informative observation of an underlying phenomenon, so predicting modality content encourages the encoder to capture what the modalities reveal (Girdhar et al., 2023; Chen et al., 2025). What is captured can be characterized by three interactions (Liang et al., 2023c): *shared* information available across modalities, *unique* information specific to one modality, and *synergistic* information that emerges when modalities are considered jointly. The choice of pretext task shapes which of these interactions are emphasized in the learned representation. If $\mathcal{L}_{pre}$ does not encourage a particular interaction, that interaction may be weakly represented, limiting downstream tasks that rely on it. Hence, **our goal** is to design $\mathcal{L}_{pre}$ that encourages all three through distinct predictive structures.

3 TRI-JEPA: INTERACTION-AWARE MULTIMODAL PRETRAINING

From goal to mechanism. Our goal is to design a pretext objective $\mathcal{L}_{pre}$ that encourages diverse multimodal interactions. We do so via prediction (LeCun et al., 2022): given a source subset $\mathcal{S}$ and target subset $\mathcal{T}$, the encoder must extract from $\mathcal{S}$ representations that are informative about $\mathcal{T}$. The choice of $\mathcal{S}$ and $\mathcal{T}$ therefore determines the structure of the prediction problem and which multimodal relationships are emphasized. Tri-JEPA uses three source-target configurations corresponding to cross-modal, within-modal, and joint-source prediction, motivating our shared, unique, and synergistic objectives, respectively (§2).

Why JEPA. Joint-embedding predictive architectures (JEPA) (Assran et al., 2023) provide a natural mechanism for these configurations: an encoder maps source and target observations into a shared latent space, and a predictor maps the source embedding toward the target embedding. Because source and target can be flexibly defined over modality subsets, the same framework accommodates all three prediction structures. JEPA predicts in representation space rather than reconstructing raw inputs, avoiding commitment to input-space details that may be uninformative for downstream tasks (Balestriero & LeCun, 2024; Littwin et al., 2024; Van Assel et al., 2025), and does not require augmentation assumptions common in contrastive learning (Tian et al., 2020; Xiao et al., 2020).

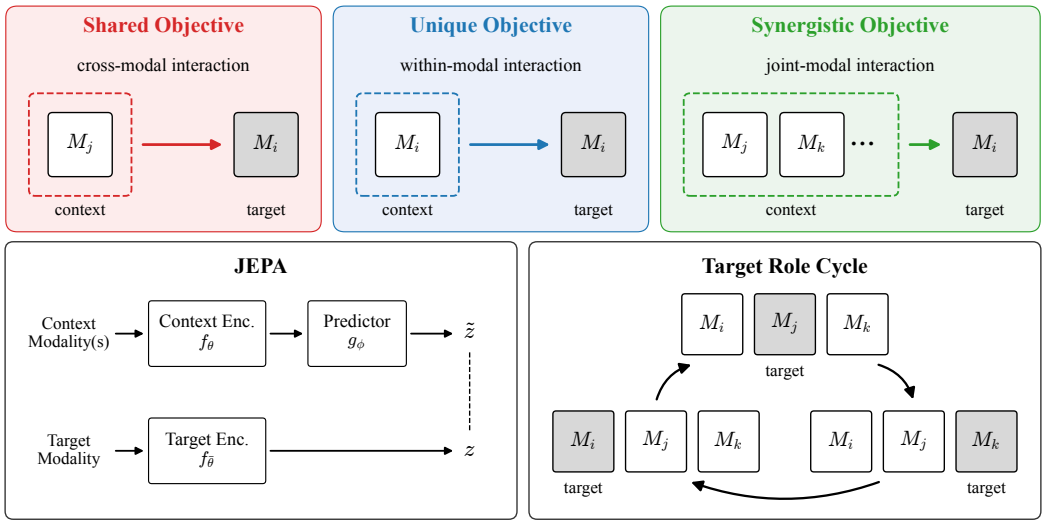


Figure 2: **Tri-JEPA.** Three predictive objectives use different source $\mathcal{S}$ and target $\mathcal{T}$ configurations to encourage distinct multimodal interactions. The *shared* objective predicts a target modality from another modality, the *unique* objective predicts held-out content from the same modality, and the *synergistic* objective predicts a target modality from all remaining modalities jointly.

JEPA framework. Given a source $\mathcal{S}$ and target $\mathcal{T}$, JEPA uses a context encoder f_θ, a target encoder $f_{\bar{\theta}}$, and a predictor g_ϕ that maps the source embedding toward the target embedding. The target encoder is updated as an exponential moving average of the context encoder, with gradients stopped through $f_{\bar{\theta}}$. The pretraining loss is defined in the embedding space:

$$\mathcal{L}_{\text{JEPA}}(\mathcal{S}, \mathcal{T}) = D(g_\phi(f_\theta(\mathcal{S})), f_{\bar{\theta}}(\mathcal{T}))\,, \tag{1}$$

where D measures agreement between predicted and target embeddings. We use token-level InfoNCE as D, which aligns corresponding prediction-target pairs and enforces batch uniformity, preventing representation collapse (Wang & Isola, 2020; Chen et al., 2025).

Tri-JEPA objectives. Tri-JEPA instantiates three JEPA losses that differ in how the source is constructed for a given target (Fig. 2). For each modality M_i, $i \in 1, \ldots, n$, we randomly partition its patches into a context subset M_i^{ctx} and a target subset M_i^{tgt}, and define:

$$\mathcal{L}_{\text{shared}}^{(i,j)} = \mathcal{L}_{\text{JEPA}}(M_j^{\text{ctx}},\, M_i^{\text{tgt}}), \quad j \neq i, \tag{2}$$

$$\mathcal{L}_{\text{unique}}^{(i)} = \mathcal{L}_{\text{JEPA}}(M_i^{\text{ctx}},\, M_i^{\text{tgt}}), \tag{3}$$

$$\mathcal{L}_{\text{syn}}^{(i)} = \mathcal{L}_{\text{JEPA}}\big(\{M_j^{\text{ctx}}\}_{j\neq i},\, M_i^{\text{tgt}}\big). \tag{4}$$

The *shared* objective predicts held-out patches of M_i from visible patches of another modality M_j, encouraging cross-modal predictability. The *unique* objective predicts them from visible patches of the same modality, encouraging preservation of within-modality structure. The *synergistic* objective predicts them from the visible patches of all remaining modalities jointly, encouraging the model to exploit their combined context. When the source contains multiple modalities, each is encoded by f_θ using its own patch embedding, and the resulting tokens are processed jointly by g_ϕ.

Cycling across modalities. Computing all three objectives for every target modality at each step would increase per-step cost with n. Instead, we cycle the target modality across training steps ($i = t \bmod n$ at step t) and evaluate the three objectives only for the current target:

$$\mathcal{L}_{\text{Tri-JEPA}} = \frac{1}{n-1} \sum_{j\neq i} \mathcal{L}_{\text{shared}}^{(i,j)} + \mathcal{L}_{\text{unique}}^{(i)} + \mathcal{L}_{\text{syn}}^{(i)}. \tag{5}$$

This ensures balanced coverage across modalities.

4 EXPERIMENTAL SETUP

4.1 PRETRAINING

Datasets. We use multi-sensor Earth observation (EO) as a controlled setting for studying multimodal interactions, since co-located sensors observe the same underlying scene while capturing complementary physical properties. We pretrain Tri-JEPA on a 130K-location subset of MOSAIC-CONUS (Potnis et al., 2026) across the contiguous United States, using three co-located imaging modalities: Sentinel-1 SAR (2 polarizations), Sentinel-2 multispectral imagery (10 bands), and WorldView optical imagery (4 bands: RGB + near-infrared). We deliberately use this geographically narrower setting to isolate the effect of objective design from pretraining scale and geographic coverage (Purohit et al., 2025; Kaur et al., 2026).

Implementation details. We pretrain Tri-JEPA with a ViT-Base (Dosovitskiy et al., 2020) for 100 epochs on 4 NVIDIA GH200 GPUs using AdamW with weight decay 0.05, a cosine learning-rate schedule, 10 warmup epochs, base learning rate 1.5×10^{-4}, batch size 128, and mixed precision. Each modality M_i uses a dedicated patch projection from its C_i input channels to the common token dimension d, while the transformer encoder is shared across modalities. The target encoder is an EMA of the context encoder with momentum 0.996. The predictor is a 4-layer transformer that predicts target features at masked positions from the encoded context. We mask 60% of patches in the cycling target modality and apply random masking at the same rate to context modalities. We use RandomResizedCrop (scale 0.3 to 1.0), horizontal flip, and light radiometric jitter, applied spatially consistently across modalities, with per-channel z-score normalization. For baselines (§4.3), we use the same encoder backbone and pretraining corpus while retaining each method's objective protocol.

4.2 DOWNSTREAM EVALUATION

Tasks & datasets. We evaluate Tri-JEPA across diverse downstream tasks. For *classification* and *semantic segmentation*, we use GEO-Bench (Lacoste et al., 2023), which consists of multispectral and SAR data. We report top-1 accuracy and mean Intersection over Union (mIoU), respectively. For *object detection*, we use DSSDD (Hu et al., 2021), a paired optical-SAR ship detection dataset, and report mAP@0.5. For *change detection*, we evaluate on LEVIR-CD (Chen & Shi, 2020), BRADD (Karaman et al., 2023), and GBv2-FoTW (Kerner et al., 2025), covering RGB building change, SAR disaster damage, and multispectral cropland change. We report change-class F1. All benchmarks use their official train, validation, and test splits.

Implementation details. We adopt a frozen-backbone evaluation protocol (Balestriero et al., 2023) to assess the quality of the pretrained representations. Classification uses a linear probe on mean-pooled patch features, and segmentation uses a UPerNet (Xiao et al., 2018) decoder on frozen patch features. For object detection, we attach a Faster R-CNN (Ren et al., 2015) head, and for change detection, we use UPerNet on the absolute difference between pre- and post-event patch features. All downstream heads are trained for 50 epochs with AdamW and a cosine learning-rate schedule. We sweep the learning rate using the validation set and report the test score from the best validation epoch.

4.3 BASELINES

We compare Tri-JEPA against general SSL methods, including ImageBind (Girdhar et al., 2023), MultiMAE (Bachmann et al., 2022), DINO (Caron et al., 2021), and I-JEPA (Assran et al., 2023). We also include the geospatial foundation model AnySat (Astruc et al., 2025) and interaction-based multimodal SSL methods, including FactorCL (Liang et al., 2023b), CoMM (Dufumier et al., 2024), and InfMasking (Wen et al., 2025). For controlled comparison, all baselines use the same ViT-Base backbone and pretraining data (Potnis et al., 2026) as Tri-JEPA.

5 Results & Discussion

5.1 Pretraining Analysis

We first examine how the three objectives behave during pretraining and the interaction-related structure they induce in the learned representations.

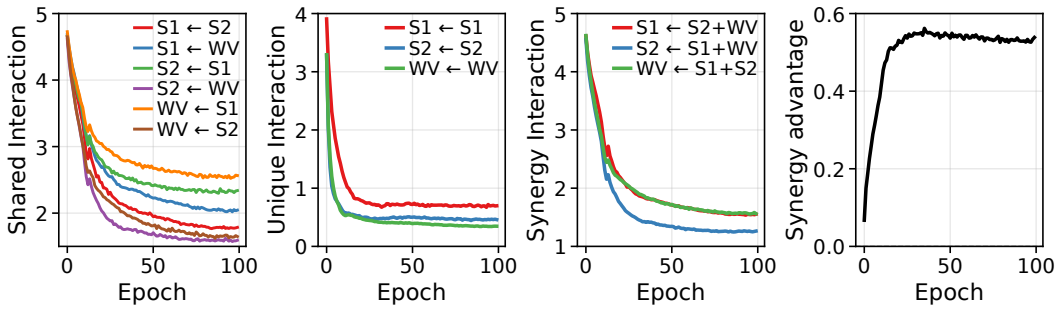


Figure 3: Training trajectories for the shared, unique, and synergistic objectives across source-target modality configurations. The rightmost panel shows the growing advantage of joint multimodal context over single-source cross-modal prediction during pretraining.

How do the objectives optimize during pretraining? Fig. 3 shows the trajectories of $\mathcal{L}_{\text{shared}}$, $\mathcal{L}_{\text{unique}}$, and $\mathcal{L}_{\text{syn}}$ for each target modality. All losses decrease steadily during pretraining, with an ordering $\mathcal{L}_{\text{unique}} < \mathcal{L}_{\text{syn}} < \mathcal{L}_{\text{shared}}$. This ordering is intuitive from the information available to each predictor: visible patches from the same modality provide the closest view of the target, combining multiple other modalities provides broader complementary context, and relying on a single different modality leaves the least direct information about the held-out target patches. Accordingly, within-modality prediction is easiest, joint-source prediction is intermediate, and single-source cross-modal prediction is the most difficult. $\mathcal{L}_{\text{unique}}$ stabilizes earlier, while $\mathcal{L}_{\text{shared}}$ and $\mathcal{L}_{\text{syn}}$ continue improving throughout training. The widening gap $\mathcal{L}_{\text{shared}} - \mathcal{L}_{\text{syn}}$ also shows that joint multimodal context becomes increasingly more predictive of the target than single-source cross-modal context.

Do the objectives induce the intended interaction structure? Loss trajectories alone do not establish whether the objectives shape representations as intended. We therefore examine two potential confounds: whether within-modality prediction also learns substantial shared structure, and whether joint-source prediction benefits simply from receiving more inputs.

Shared information should be recoverable across modalities (Liang et al., 2023a). We therefore estimate mutual information between paired modality representations on 1,000 held-out scenes from the pretraining corpus. For each scene and modality, we mean-pool the frozen patch tokens into a single representation and average the estimates across the three modality pairs. Shared-only yields substantially higher estimated MI than unique-only (0.741 vs. 0.050), indicating that cross-modal prediction induces markedly stronger cross-modal dependence than within-modality prediction.

We next test whether the synergistic objective captures more than the benefit of additional inputs. Following partial information decomposition (Williams & Beer, 2010; Liang et al., 2023a), for source modalities X_1, X_2 and target T,

$$\Delta = I(X_1, X_2; T) - I(X_1; T) - I(X_2; T) = S - R, \tag{6}$$

where S and R denote synergistic and redundant information. Thus, $\Delta > 0$ indicates that synergistic information exceeds redundant information. Using the same held-out scenes, we take each modality as T in turn, concatenate the other two source representations for the joint term, and use each source separately for the individual terms. Synergy-only yields $I(X_1, X_2; T) = 1.531$ compared with $I(X_1; T) + I(X_2; T) = 0.790$, giving an estimated $\Delta = +0.741$. In contrast, shared-only gives an estimated $\Delta = -0.653$. This provides evidence of a synergy-dominated interaction under the synergistic objective, rather than a gain explained solely by additional source inputs.

What complementary trends emerge under Tri-JEPA? To further test whether the three objectives provide complementary training signals, we evaluate cross-modal retrieval R@1 as a task-agnostic probe of how well the learned representations preserve scene identity across modalities. Given an embedding from one modality, we retrieve the matching scene from another modality among a set of candidates. We pretrain three *single-interaction variants*, each using only one of $\mathcal{L}_{\text{shared}}$, $\mathcal{L}_{\text{unique}}$, or $\mathcal{L}_{\text{syn}}$. Fig. 4 shows that the single-interaction variants yield different levels of cross-modal retrieval, while Tri-JEPA achieves the highest R@1. Notably, Tri-JEPA also exceeds the shared-only variant, suggesting that the objectives provide complementary signals when learned jointly.

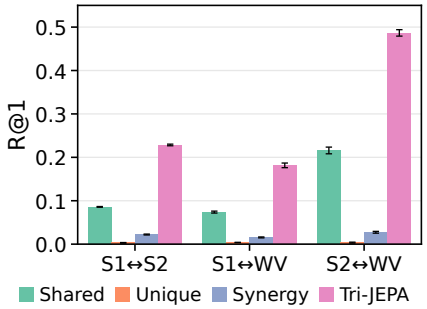


Figure 4: Cross-modal retrieval R@1.

5.2 Downstream Transfer Analysis

Having shown that the objectives induce interaction-related structure during pretraining (§5.1), we next examine how these representations transfer downstream. We first compare Tri-JEPA with multimodal SSL baselines under controlled settings, then analyze whether the three objectives contribute differently across tasks. Specifically, *we ask how the objectives contribute to downstream performance and whether their relative importance shifts across tasks.*

Main results. We compare Tri-JEPA against the baselines in §4.3. Tab. 1 shows that Tri-JEPA achieves consistently strong performance across the downstream suite. These results indicate that capturing multimodal interactions during pretraining yields features that transfer across many tasks.

Table 1: Downstream performance using frozen pretrained encoders. Mean ± standard deviation over 3 seeds, averaged across datasets to summarize task-level trends (§4.2).

	Image Classification	Semantic Segmentation	Object Detection	Change Detection
Method	Acc. ↑	mIoU ↑	mAP@0.5 ↑	F1 ↑
ImageBind	69.02 ±0.28	53.03 ±0.32	0.714 ±0.010	0.615 ±0.012
MultiMAE	70.68 ±0.25	54.60 ±0.30	0.722 ±0.013	0.729 ±0.011
DINO	68.30 ±0.31	57.83 ±0.33	0.780 ±0.012	0.716 ±0.013
I-JEPA	69.86 ±0.29	60.93 ±0.31	0.499 ±0.018	0.714 ±0.014
AnySat	62.59 ±0.33	56.34 ±0.35	0.527 ±0.006	0.661 ±0.015
FactorCL	61.41 ±0.36	45.12 ±0.40	0.658 ±0.015	0.598 ±0.016
CoMM	60.98 ±0.38	37.89 ±0.42	0.686 ±0.010	0.605 ±0.015
InfMasking	71.93 ±0.27	64.74 ±0.29	0.764 ±0.012	0.651 ±0.013
Tri-JEPA	**75.68** ±0.24	**66.20** ±0.27	**0.869** ±0.009	**0.739** ±0.011

How do individual objectives contribute to downstream transfer? We perform leave-one-out ablations by removing each objective from Tri-JEPA. Fig. 5 shows that their contributions are task-dependent. Removing the synergistic objective produces the largest drop in classification, while segmentation is more affected by removing either the unique or synergistic objective than the shared objective. For object detection, the differences are small relative to the observed variation, so we do not assign an importance ranking. Change detection degrades similarly when any objective is removed. Full Tri-JEPA achieves the highest mean performance across all four task families, supporting complementary contributions whose relative importance varies across tasks.

How label-efficient is Tri-JEPA? We evaluate few-shot transfer on GEO-Bench by varying the fraction of labeled downstream data. Fig. 6 shows that Tri-JEPA consistently matches or outperforms baselines that either emphasize only a subset of these interactions or model them jointly through bundled objectives, across label fractions for both classification and segmentation. The advantage is pronounced for classification in the low-label regime, while segmentation gains become clearer as more labels are available. Overall, Tri-JEPA retains strong performance with limited supervision.

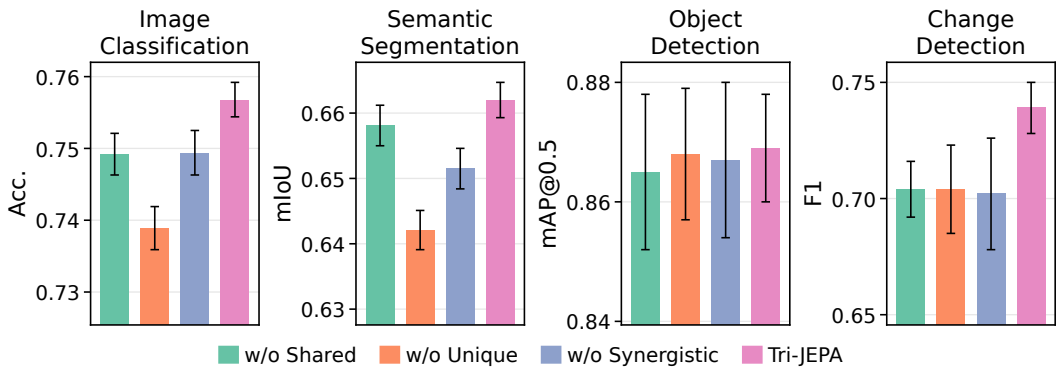


Figure 5: Leave-one-out objective ablations. Downstream performance after removing each Tri-JEPA objective, compared with the full model. The effect of each objective varies across task families, while full Tri-JEPA achieves the highest mean performance in all four.

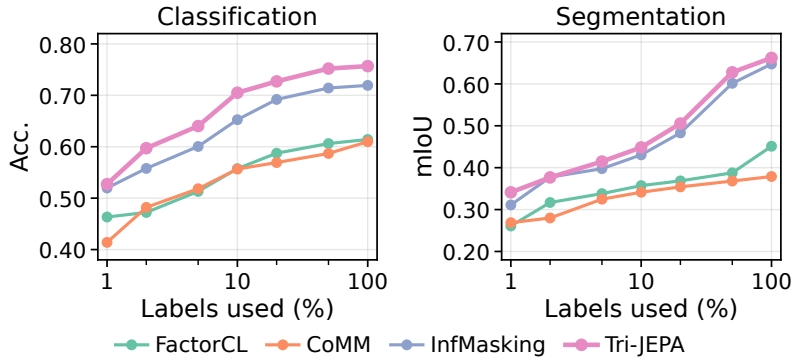


Figure 6: Few-shot experiments on GEO-Bench (Lacoste et al., 2023).

6 Related Work

Multimodal Self-Supervised Learning (SSL) Multimodal SSL learns transferable representations from co-occurring modalities without labels. Existing approaches broadly span joint-embedding, generative, and latent predictive objectives (LeCun et al., 2022; Assran et al., 2023). Joint-embedding methods align representations of compatible views (Chen et al., 2020; Caron et al., 2021). In multimodal settings, CLIP (Radford et al., 2021) and ImageBind (Girdhar et al., 2023) extend this idea to paired modalities, while DeCUR (Wang et al., 2024) and TaxaBind (Sastry et al., 2025) adapt it to geospatial data. Generative approaches instead reconstruct masked or corrupted inputs (He et al., 2022). MultiMAE (Bachmann et al., 2022) and 4M (Mizrahi et al., 2023) extend masked modeling across modalities, and SatMAE (Cong et al., 2022) and MMEarth (Nedungadi et al., 2024) bring this paradigm to Earth observation. Since reconstruction operates in input space, such methods can emphasize low-level details that are not always useful for downstream representation learning (Balestriero & LeCun, 2024; Van Assel et al., 2025). Joint-embedding predictive architectures (JEPA) address this by predicting target representations from source representations directly in latent space (LeCun et al., 2022; Assran et al., 2023). This predictive formulation has also been extended to multimodal settings, including VL-JEPA (Chen et al., 2025) for vision-language learning and AnySat (Astruc et al., 2025) for Earth observation.

Across these families, however, multimodal objectives typically align, reconstruct, or predict paired or fused representations without explicitly separating different types of multimodal interaction. Tri-JEPA instead uses three source-target configurations within a common predictive framework, encouraging cross-modal, within-modality, and joint-source structure associated with shared, unique, and synergistic interactions, respectively.

Modality Interactions in Multimodal Learning Multimodal information can be characterized in terms of shared, unique, and synergistic interactions (Williams & Beer, 2010; Liang et al., 2023a). In supervised learning, task labels provide a direct signal for identifying which interactions are predictive of the target. SMURF (Wörtwein et al., 2024) and MCR (Kontras et al., 2024) model redundancy and uniqueness through auxiliary correlation or mutual-information objectives, while MMoE (Yu et al., 2024), I2MoE (Xin et al., 2025), and MERGE (Han et al., 2025) model all three interactions using interaction-specific experts. In self-supervised learning, labels are unavailable and supervision must instead come from relationships among the modalities. Contrastive methods such as CLIP (Radford et al., 2021) emphasize shared cross-modal structure by aligning paired modalities under the multi-view assumption (Tsai et al., 2020; Tosh et al., 2021). FactorCL (Liang et al., 2023b) separates shared and unique information through mutual-information objectives, while CoMM (Dufumier et al., 2024) and InfMasking (Wen et al., 2025) further account for synergistic interactions through multimodal fusion and masking-based objectives.

These self-supervised approaches model multiple interactions within coupled or fused objectives, making the contribution of each interaction difficult to examine separately. Evidence for their capture has also relied largely on synthetic benchmarks (i.e., Trifeature (Hermann & Lampinen, 2020)). Tri-JEPA instead assigns each interaction its own predictive source-target configuration, enabling their effects to be studied separately during pretraining and downstream transfer.

7 Conclusion

We introduced Tri-JEPA, a self-supervised multimodal pretraining method that encourages shared, unique, and synergistic interaction-related structure through distinct predictive objectives. By varying the source-target configuration within a unified JEPA framework, Tri-JEPA enables these interactions to be examined separately during pretraining and downstream transfer. Across diverse downstream tasks, Tri-JEPA consistently outperforms the multimodal SSL baselines considered. Our analyses further show that the objectives induce distinct information relationships and provide complementary, task-dependent contributions downstream. Overall, these results support interaction-aware predictive objectives as a useful design principle for multimodal self-supervised pretraining.

Limitations & Future Work Tri-JEPA is motivated by partial information decomposition (PID) (Williams & Beer, 2010), but its objectives are not intended to recover exact PID components. Our mutual-information analyses instead provide initial evidence that they induce distinct cross-modal and synergy-dominated structure. Our experiments focus on image-based modalities, leaving extension to qualitatively different modalities such as text for future work.

Broader Impacts Our work studies general multimodal self-supervised pretraining. By improving the transferability of learned representations, Tri-JEPA may benefit applications where multimodal labels are limited or expensive to obtain. At the same time, stronger multimodal representations can lower barriers to dual-use applications, including surveillance. Practitioners should therefore consider the risks of the intended application when deploying models built on this work.

AI use statement

In this work, we did not use AI tools as an important, original, or non-standard component of the core methods.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19530–19540, 2025.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pp. 348–367. Springer, 2022.

Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24680–24689, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? *arXiv preprint arXiv:2409.07402*, 2024.

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.

Xing Han, Hsing-Huan Chung, Joydeep Ghosh, Paul Pu Liang, and Suchi Saria. Massively multimodal foundation models: A framework for capturing interactions with specialized mixture-of-experts. *arXiv preprint arXiv:2509.25678*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33: 9995–10006, 2020.

Yuxin Hu, Yini Li, and Zongxu Pan. A dual-polarimetric sar ship detection dataset and a memory-augmented autoencoder-based detection method. *Sensors*, 21(24):8478, 2021.

Kaan Karaman, V Sainte Fare Garnot, and Jan Dirk Wegner. Deforestation detection in the amazon with sentinel-1 sar image time series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:835–842, 2023.

Amandeep Kaur, Mirali Purohit, Gedeon Muhawenayo, Esther Rolf, and Hannah Kerner. Pretrain where? investigating how pretraining data diversity impacts geospatial foundation model performance. *arXiv preprint arXiv:2604.21104*, 2026.

Hannah Kerner, Snehal Chaudhari, Aninda Ghosh, Caleb Robinson, Adeel Ahmad, Eddie Choi, Nathan Jacobs, Chris Holmes, Matthias Mohr, Rahul Dodhia, et al. Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28151–28159, 2025.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Pu Liang, Matthew B Blaschko, and Maarten De Vos. Multimodal fusion balancing through game-theoretic regularization. In *Greeks in AI Symposium 2025*, 2024.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023a.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023b.

Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. *arXiv preprint arXiv:2306.04539*, 2023c.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems*, 37:91300–91336, 2024.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

Abhishek Potnis, Youssef Hussein, Waqwoya Abebe, Jang Hyeon Lee, Debvrat Varshney, Jacob Arndt, Philipe Ambrozio Dias, Dan Lu, Dalton Lunga, et al. Mosaic-conus: A multimodal, multi-temporally paired dataset for earth sciences. Technical report, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2026.

Mirali Purohit, Gedeon Muhawenayo, Esther Rolf, and Hannah Kerner. How does the spatial distribution of pre-training data affect geospatial foundation models? *arXiv preprint arXiv:2501.12535*, 2025.

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.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1765–1774. IEEE, 2025.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic learning theory*, pp. 1179–1206. PMLR, 2021.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*, 2025.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pp. 286–303. Springer, 2024.

Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. Infmasking: Unleashing synergistic information by contrastive multimodal interactions. *arXiv preprint arXiv:2509.25270*, 2025.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Torsten Wörtwein, Nicholas B Allen, Jeffrey F Cohn, and Louis-Philippe Morency. Smurf: Statistical modality uniqueness and redundancy factorization. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pp. 339–349, 2024.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *arXiv preprint arXiv:2505.19190*, 2025.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10006–10030, 2024.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2024.