R-JEPA: Objective–Regime Compatible Relational Prediction for Self-Supervised Analogical Learning

Anonymous ACL submission

Abstract

Self-supervised relational invariance emerges only when the training objective is compatible with the relational structure of the data. Some analogy datasets contain multiple transformation families and require explicit discrimination, while others are dominated by a single transformation type in which negative samples are inherently ambiguous. Applying a single objective across these settings leads to systematic failure, even with identical architectures, as contrastive learning introduces false negatives in relation-sharing data and non-contrastive objectives collapse distinct relations when discrimination is required. We propose **Relational JEPA (R-JEPA)**, which represents transformations between paired observations as explicit relation embeddings and applies prediction directly in relation space with objectives selected according to the induced data regime. Across text-based analogy benchmarks, regime-matched training improves analogy verification, completion, and entity-disjoint transfer over state-based baselines, while mismatched objectives yield misleading geometric structure without relational invariance.

1 Introduction

Joint-Embedding Predictive Architectures (JEPAs) learn invariant representations by predicting targets in a shared embedding space, avoiding reconstruction (Assran et al., 2023). They are effective when invariance over *states* is sufficient.

Analogical reasoning, however, requires invariance over *transformations* (Gentner and Markman, 1997; Gentner, 1988). Given (A, B) and (C, D), the task is to assess whether $A \rightarrow B$ matches $C \rightarrow D$, rather than whether individual states are similar. Standard JEPA formulations predict latent states rather than transformations, leaving relational structure indirect and weakly constrained.

We argue that self-supervised relational learning is *conditional*: it succeeds only when the training objective is compatible with the relational regime induced by the data, and fails even with the same architecture otherwise. This aligns with theoretical analyses of contrastive learning, which show that representation separation depends critically on latent class structure and the informativeness of negatives (Saunshi et al., 2019). Crucially, datasets induce different *relational regimes*: some contain multiple transformation families and require *discrimination*, while others are dominated by a single transformation type and require *sharing* (Wijesiriwardene et al., 2023). Applying a single objective across regimes therefore leads to systematic failure.

This explains why contrastive objectives can produce misleading geometric structure. In relation-sharing settings, in-batch negatives introduce pervasive false negatives and fragment relation space (Chuang et al., 2020; Robinson et al., 2020). Conversely, in relation-discriminating settings, objectives that avoid explicit negatives can collapse distinct transformation families despite strong alignment (Wang and Isola, 2020; Chen and He, 2021; Zbontar et al., 2021).

To operationalize this principle, we propose **Relational JEPA (R-JEPA)**, which treats transformations as first-class prediction targets. R-JEPA constructs explicit relation embeddings, applies prediction directly in relation space, and selects objectives according to the dataset regime. Across analogy benchmarks, matched objectives improve verification, completion, and entity-disjoint transfer, while mismatches degrade performance despite identical architectures. This notion of generalization aligns with prior work on compositional and systematic generalization (Keysers et al., 2019).

Our contributions are:

- We frame self-supervised relational learning as conditional on *data regime* and *objective compatibility*.
- We identify objective–regime mismatch as

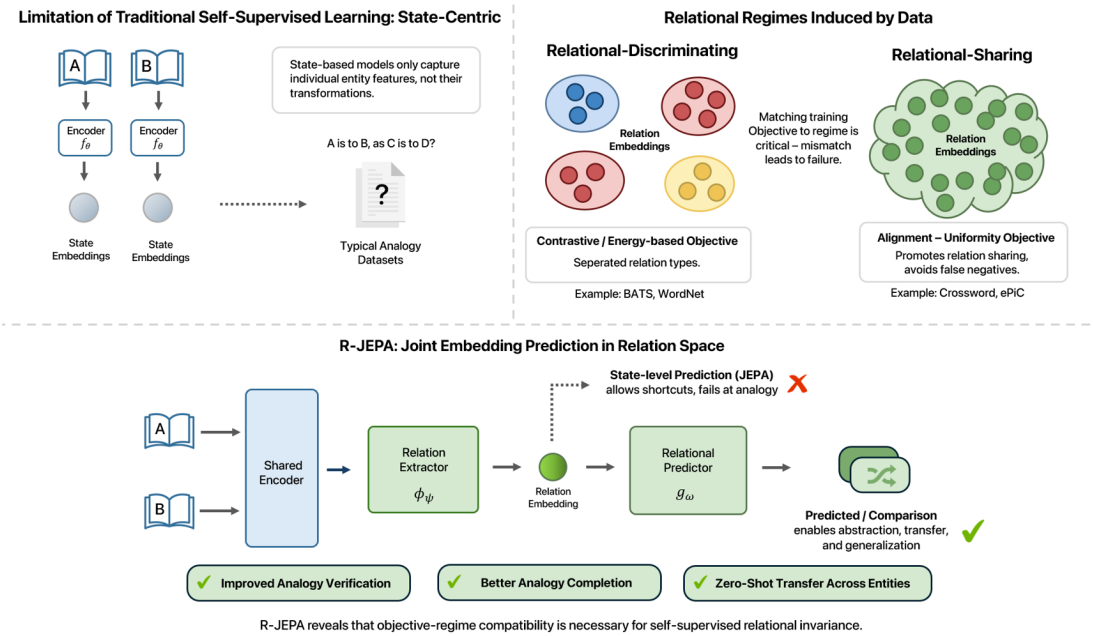

Figure 1: Objective–Regime Compatible Relational Learning with R-JEPA. Traditional self-supervised methods are state-centric, capturing individual entity features but leaving transformations implicit, which limits analogical reasoning. Analogy datasets induce different relational regimes: relation-discriminating regimes with multiple transformation families, where contrastive objectives are effective, and relation-sharing regimes dominated by a single transformation type, where contrastive learning introduces false negatives. R-JEPA addresses this by representing transformations explicitly as relation embeddings and applying prediction directly in relation space. By matching the training objective to the data regime, R-JEPA avoids state-level shortcuts and enables robust analogy verification, completion, and zero-shot transfer across entities.

a primary failure mode in analogy learning (Chuang et al., 2020; Robinson et al., 2020; Wang and Isola, 2020).

- We propose R-JEPA, which learns transformation-level invariances without relation labels by applying prediction directly in relation space under regime-matched objectives.

2 Related Work

2.1 Joint-Embedding Predictive Architectures

JEPAs learn representations by predicting targets in a shared embedding space, typically using architectural asymmetry, stop-gradient, and objectives that avoid collapse (Assran et al., 2023; Chen and He, 2021; Zbontar et al., 2021; Wang and Isola, 2020). However, existing formulations are state-centric: the predicted target is a latent state, and relations are only implicit (e.g., as differences between state embeddings). We extend JEPA by making transformations explicit and applying learning objectives directly in relation space. Related non-contrastive designs such as BYOL further show that stop-gradient asymmetry can stabilize representation learning even without explicit negatives or batch statistics (Richemond et al., 2020).

2.2 Predictive World Models and Latent Dynamics

World models learn latent dynamics to predict future states (Ha and Schmidhuber, 2018). While they model change, transitions are usually optimized as mechanisms for state prediction rather than as representational targets. Analogical reasoning instead requires abstraction over *types of change* (Gentner and Markman, 1997). Our approach represents transformations as embeddings and enforces invariance over transformation instances.

2.3 Relational and Analogical Reasoning

Analogy has long been characterized as relational similarity rather than attribute matching (Gentner and Markman, 1997; Gentner, 1988). In NLP, evaluation has traditionally focused on word-level and proportional analogies (Mikolov et al., 2013; Gladkova et al., 2016), and more recently on

whether pretrained language models encode analogical structure (Ushio et al., 2021; Wijesiriwardene et al., 2023).

Many embedding-based approaches define relations implicitly, for example via linear offsets or metric similarity in state embedding space. However, prior analyses show that such regularities can arise from embedding geometry without reflecting genuine relational abstraction (Levy and Goldberg, 2014; Ethayarajh et al., 2019). Similarly, contextual representations may encode rich linguistic information implicitly, but this structure is not explicitly organized for relational comparison (Peters et al., 2018).

These approaches assume that negatives are semantically meaningful and that similarity implies relational equivalence. We show that these assumptions break under relation-sharing regimes due to pervasive false negatives (Chuang et al., 2020; Robinson et al., 2020), motivating a regime-dependent view of relational learning and objective design.

3 Problem Formulation

We study learning relational representations from unlabeled paired observations, without explicit relation labels.

States. Let $x \in \mathcal{X}$ be an observation (e.g., a word or sentence). An encoder f_θ maps x to a latent state

$$z = f_\theta(x) \in \mathbb{R}^d. \tag{1}$$

State representations capture properties of individual observations, but analogical reasoning requires abstraction over *transformations* between states.

Relations as transformations. Given x_A, x_B with states z_A, z_B, a relation extractor ϕ_ψ produces a relation embedding

$$r_{A\to B} = \phi_\psi(z_A, z_B) \in \mathbb{R}^k. \tag{2}$$

$r_{A\to B}$ is intended to encode how A changes into B while reducing dependence on entity identity.

Analogical equivalence. For pairs (A, B) and (C, D), analogical equivalence is

$$r_{A\to B} \approx r_{C\to D}, \tag{3}$$

under a task-defined similarity metric in relation space, even if (z_A, z_C) and (z_B, z_D) are dissimilar. For analogy completion, given (A, B, C) and candidates $\mathcal{D}$,

$$D^* = \arg\min_{D\in\mathcal{D}} \operatorname{dist}(r_{A\to B}, r_{C\to D}), \tag{4}$$

where $\operatorname{dist}(\cdot,\cdot)$ is a relation-space distance.

Relational invariance. We define relational invariance as mapping instances of the same transformation family $\mathcal{T}$ (unknown during training) to nearby points:

$$\forall (A, B), (C, D) \in \mathcal{T} : \operatorname{dist}(r_{A\to B}, r_{C\to D}) \le \epsilon, \tag{5}$$

for small ϵ. This supports generalization across entities and contexts.

Relational regimes. Relational invariance is not always learnable from self-supervision; it depends on the dataset structure and the training objective. We distinguish (Wijesiriwardene et al., 2023):

- **Relation-discriminating regime:** multiple transformation families; learning requires invariance within families and separation across families; negatives are informative.
- **Relation-sharing regime:** data dominated by a single transformation family; in-batch negatives create pervasive false negatives and can destabilize learning (Chuang et al., 2020; Robinson et al., 2020).

These regimes are properties of the data distribution.

Objective–regime compatibility. In relation-discriminating regimes, contrastive/energy-based objectives can separate transformation families. In relation-sharing regimes, such objectives conflict with the data and objectives that avoid explicit negatives (e.g., separating alignment from global regularization) are preferred. Our central claim is that self-supervised relational invariance learning requires objective–regime compatibility.

Implication. This motivates models that represent transformations explicitly, apply learning directly in relation space, and select objectives according to the dataset regime. We instantiate this with R-JEPA.

4 R-JEPA Architecture

4.1 Overview

As illustrated in Figure 2, R-JEPA is a joint-embedding predictive architecture that treats *transformations between states* as the primary prediction

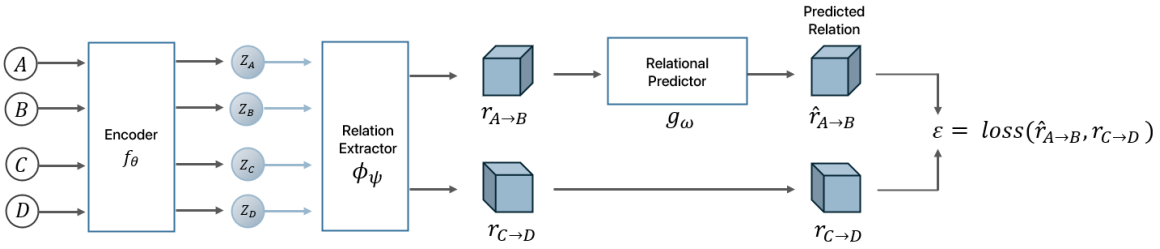

Figure 2: Overview of Relational JEPA (R-JEPA). A shared encoder f_θ maps observations to states (z_A, z_B, z_C, z_D). A relation extractor ϕ_ψ produces relation embeddings $r_{A\to B}$ and $r_{C\to D}$. A relational predictor g_ω operates in relation space for prediction and comparison.

target. Given a pair (A, B), observations are encoded into latent states and mapped to an explicit relation embedding $r_{A\to B} = \phi_\psi(z_A, z_B)$. Learning proceeds by predicting and comparing relation embeddings in a dedicated relation space.

Predictive learning is applied exclusively in relation space: the relational predictor g_ω does not access state embeddings directly, and losses are defined only over relation embeddings. This prevents state-level shortcuts (e.g., entity similarity) and forces learning signals to target transformation structure. The training objective is chosen according to the relational regime induced by the dataset (Section 4.4).

4.2 Architectural Components

The encoder f_θ maps each observation x to a latent state $z = f_\theta(x)$, capturing semantic properties of individual observations but not directly optimized for analogical comparison. Given a state pair (z_A, z_B), the relation extractor ϕ_ψ produces a relation embedding

$$r_{A\to B} = \phi_\psi(z_A, z_B), \tag{6}$$

intended to encode the transformation from A to B while reducing dependence on entity identity. The relational predictor g_ω then maps this embedding to a predicted relation

$$\hat{r} = g_\omega(r_{A\to B}), \tag{7}$$

introducing predictive structure over transformations. This step discourages degenerate solutions in which relations collapse to simple state differences and enforces consistency among relation instances.

4.3 Relation Extractor Design

We consider relation extractors of increasing expressivity, including linear difference-based mappings $r = W(z_B - z_A)$, concatenation-based MLPs over $(z_A, z_B, z_B - z_A)$, and more expressive bilinear or attention-based formulations. While these designs differ in capacity, they share the same role: producing a compact embedding that represents *how* one state changes into another. The explicit separation between state encoding and relation extraction is essential for isolating transformation-level structure.

4.4 Objective Families for Relational Learning

R-JEPA is objective-agnostic at the architectural level. Instead, the geometry of relation space is shaped by the training objective, which is selected based on the dataset's relational regime.

When datasets contain multiple distinct transformation families, relational learning requires explicit discrimination. We therefore use a contrastive energy-based objective (Oord et al., 2018; Saunshi et al., 2022; Tosh et al., 2021):

$$\mathcal{L}_{\text{EB}} = -\log \frac{\exp(-E(r, r^+))}{\sum_{r' \in \{r^+, r_1^-, \ldots, r_m^-\}} \exp(-E(r, r'))}, \tag{8}$$

where the energy function is defined as squared distance between normalized embeddings,

$$E(r, r') = \left\| \frac{r}{\|r\|} - \frac{r'}{\|r'\|} \right\|_2^2. \tag{9}$$

This objective pulls together instances of the same transformation while separating different transformation families.

In relation-sharing regimes dominated by a single transformation type, contrastive objectives introduce widespread false negatives (Chuang et al., 2020; Robinson et al., 2020). We therefore decouple positive alignment from global regularization using an alignment–uniformity (multi-objective) objective (Wang and Isola, 2020; Zbontar et al., 2021; Chen and He, 2021). Such multi-objective formulations have recently been shown to admit principled interpretations from an information-theoretic perspective, where alignment controls signal preservation and uniformity regulates representational entropy (Wang and Isola, 2020; Zhang et al., 2023).

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(r,r^+)} \left[\|r - r^+\|_2^2 \right], \tag{10}$$

$$\mathcal{L}_{\text{uni}} = -\log \mathbb{E}_{r_i,r_j} \left[\exp\left(-t\|r_i - r_j\|_2^2\right) \right]. \tag{11}$$

These terms are combined as

$$\mathcal{L}_{\text{MU}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{uni}}\mathcal{L}_{\text{uni}}. \tag{12}$$

This formulation encourages consistent alignment of analogous relations while preventing representational collapse.

Both objective families follow JEPA-style asymmetry with stop-gradient, where one branch provides a target relation embedding and the other predicts it, stabilizing training without reconstruction losses (Assran et al., 2023; Chen and He, 2021; Zbontar et al., 2021).

4.5 Training Data Construction

Training operates on pairs of pairs (A, B) and (C, D) intended to instantiate analogous transformations. In synthetic settings, transformation families are programmatically generated and support explicit relational discrimination. In weakly structured real data, analogical pairs are mined using coarse cues and often induce relation-sharing regimes, where avoiding false negatives is critical. To assess robustness, we inject controlled corruption into positive pairs and measure degradation in verification and transfer performance, isolating the effect of the relational objective from mining noise.

5 Experiments

Our experimental goal is to test the central claim of this paper: *relational invariance learning requires training objectives that match the relational regime of the data*. Accordingly, we organize experiments by whether datasets induce *relational discrimination* or *relational sharing*, fixing the architecture and using objective–regime alignment as a causal test of relational learning success or failure.

5.1 Datasets

We ground our experiments in the ANALOGICAL taxonomy, which organizes analogy datasets by relational complexity (Wijesiriwardene et al., 2023). From this space, we select datasets that instantiate relations as transformations between states, aligning with our focus on transformation-level relational invariance.

We exclude MSR, Google, SAT, and Quotes datasets, as they emphasize fixed proportional analogies, exam-style reasoning, or interpretive explanation rather than consistent state-to-state transformations (Gao et al., 2014; Mikolov et al., 2013; Rudrapal et al., 2017).

We distinguish two relational regimes induced by the data. **Relation-discriminating** datasets contain multiple transformation families and require separation across families, whereas **relation-sharing** datasets are dominated by a single transformation type, making in-batch negatives ambiguous (Chuang et al., 2020; Robinson et al., 2020).

Relation-discriminating datasets include BATS (Gladkova et al., 2016), WordNet (Miller, 1995), and syntactic transformations derived from SNLI (Bowman et al., 2015). Relation-sharing datasets include Crossword (Pwanson, 2016), entailment pairs from SNLI, and ePiC (Ghosh and Srivastava, 2022).

Detailed dataset statistics and regime annotations are provided in Appendix A.3.

5.2 Experimental Setup

All experiments use the same R-JEPA architecture (Figure 2), consisting of a shared encoder, relation extractor, and relational predictor. Unless otherwise noted, training objectives are chosen according to the dataset's relational regime (Section 4.4): contrastive energy-based objectives for relation-discriminating datasets, and alignment–uniformity objectives for relation-sharing datasets. We additionally report cross-regime ablations where objectives are intentionally mismatched.

5.3 Evaluation Tasks

We evaluate relational invariance using three tasks:

(1) Analogy Verification. Given (A, B) and (C, D), the model predicts whether $A \rightarrow B$ is analogous to $C \rightarrow D$. We report accuracy.

(2) Analogy Completion. Given (A, B, C) and candidates $\mathcal{D}$, the model selects $D \in \mathcal{D}$ such that $A \rightarrow B$ is analogous to $C \rightarrow D$. We report Recall@1 and Recall@K.

(3) Zero-shot Analogy Transfer. Models are trained on one set of entities and evaluated on disjoint entities, testing whether relation embeddings capture transformation structure rather than entity-specific cues.

5.4 Baselines

We compare R-JEPA with baselines that isolate state-level prediction, implicit relational reasoning, or explicit relational supervision. **State-JEPA** predicts z_B from z_A under the standard JEPA formulation, with analogical reasoning performed in state space. **Contrastive state embeddings** define relations implicitly via differences between instance-discriminative embeddings. In synthetic settings, we also report a **supervised relation classifier** as an approximate upper bound.

5.5 Ablation Studies

We conduct ablations over the relation extractor (difference, concat-MLP, bilinear/attention), the relational predictor, and negative sampling strategies. To directly test objective–regime compatibility, we swap contrastive and alignment–uniformity objectives across datasets. We further assess robustness by injecting controlled noise into weakly mined positive pairs. Performance is consistently strongest when both relation-space prediction is present and the training objective matches the dataset's relational regime.

6 Results and Analysis

Our results test a single claim: *self-supervised relational invariance emerges only when the training objective is compatible with the relational regime induced by the data.* We further show that achieving this requires applying prediction directly in relation space, and that the resulting representations generalize under entity shift. We report four main results.

6.1 Objective–Regime Compatibility Is a Structural Requirement

We directly test the core hypothesis that relational learning succeeds only when the training objective matches the dataset's relational regime. Table 1 compares R-JEPA trained with a matched objective and an intentionally mismatched one on two representative regimes: BATS (relation-discriminating) and Crossword (relation-sharing).

Performance depends sharply on objective–regime alignment. On BATS, the energy-based objective is required to separate transformation families; replacing it with alignment–uniformity collapses discrimination and degrades both verification and completion. On Crossword, alignment–uniformity is required; contrastive energy-based training substantially harms completion due to pervasive false negatives. Because architecture and relation extractor are held fixed, these results isolate objective choice as a structural condition rather than a tuning preference.

The same energy-based objective that succeeds on BATS fails on Crossword, while the alignment–uniformity objective exhibits the opposite behavior, indicating that performance is governed by objective–regime compatibility rather than by the objective family itself.

Geometric interpretation. Figure 3 visualizes relation embeddings learned under matched and mismatched objectives.

In BATS, matched objectives produce well-separated clusters corresponding to transformation families, whereas mismatched objectives collapse these distinctions. In Crossword, contrastive objectives fragment relation space into spurious clusters despite the absence of multiple relation types.

These visualizations are illustrative rather than diagnostic (Ethayarajh, 2019; Timkey and Van Schijndel, 2021). We therefore complement them with qualitative nearest-neighbor inspection under objective mismatch (Appendix A.4), which shows that tight clustering does not imply consistent transformation patterns.

6.2 Why Prediction Must Be Applied in Relation Space

We next ask what architectural constraint is necessary to realize the above condition. Even with an appropriate objective, prediction could be applied either to states (standard JEPA) or to relations (R-JEPA). Table 2 compares state-level JEPA variants

Table 1: Task-level impact of objective–regime alignment. R-JEPA uses an energy-based objective for BATS and an alignment–uniformity objective for Crossword. Mismatch refers to swapping these objectives.

Method	BATS (Relation-Discriminating)		Crossword (Relation-Sharing)	
	Verif. Acc	Recall@1	Verif. Acc	Recall@1
Pretrained only	0.79	0.31	0.74	0.28
Contrastive state baseline	0.82	0.35	0.76	0.30
R-JEPA (matched objective)	**0.94**	**0.58**	**0.91**	**0.55**
R-JEPA (mismatched objective)	0.88	0.46	0.79	0.34

Table 2: State-level vs relation-level prediction on BATS. All models share the same encoder, training data, and energy-based objective. The only difference lies in whether prediction is applied in state space or relation space. Align is an alignment loss (lower is better).

Method	**Verif. Acc**	**Verif. F1**	**Align**
State-JEPA	0.86	0.85	0.55
State-JEPA (parallel)	0.87	0.85	0.64
R-JEPA (ours)	**0.93**	**0.93**	**0.30**

Table 3: Entity-disjoint (zero-shot) evaluation. Δ denotes the difference between in-domain and zero-shot accuracy.

Model	In-dom	ZS	Δ
Contrastive state baseline	0.83	0.41	0.42
Difference-only relation	0.91	0.49	0.42
No relational predictor	0.93	0.51	0.42
R-JEPA (ours)	**0.94**	**0.58**	**0.36**

Table 4: Relation space structure quality on relation-discriminating datasets. Silhouette measures cluster separation, and NMI measures recovery of ground-truth relation types. Embedding diff is $(z_B - z_A)$.

Method	**BATS**	**WordNet**	**SNLI-Syn**
Pretrained only	0.01 / 0.43	0.00 / 0.06	0.20 / 0.63
Embedding diff	0.02 / 0.44	0.00 / 0.06	0.21 / 0.63
Diff-only	0.59 / 0.84	0.68 / 0.54	0.89 / 0.71
No predictor	0.63 / 0.84	0.71 / 0.56	0.93 / 0.73
R-JEPA (Ours)	**0.73 / 0.84**	**0.75 / 0.58**	**0.97 / 0.74**

against R-JEPA on BATS under the same encoder, training data, and energy-based objective.

State-level prediction permits shortcuts based on entity similarity, whereas relation-space prediction forces learning signals to align transformations directly.

State-level prediction allows the model to minimize loss by exploiting entity-level regularities, such as lexical or semantic similarity, without aligning transformations themselves. In contrast, relation-space prediction removes these shortcuts, forcing the predictive signal to operate directly on transformation structure.

6.3 Entity-Disjoint Transfer Indicates Transformation-Level Abstraction

If relation embeddings capture transformations rather than entity identity, they should transfer to unseen entities. Table 3 reports entity-disjoint evaluation.

R-JEPA retains significantly more accuracy under distribution shift and exhibits a smaller performance drop than state-based and weakened relation baselines. This suggests that the learned relation space encodes transformation-level invariances that generalize beyond memorized entities.

Notably, the reduced performance gap (Δ) for R-JEPA is consistent with remaining errors arising from transformation ambiguity rather than entity memorization. This behavior is consistent with relation embeddings encoding transformation-level abstractions.

6.4 Relation Space Structure as Supporting Evidence

Finally, as supporting evidence rather than a primary evaluation, we report diagnostic structure metrics on relation-discriminating datasets. Table 4 shows that R-JEPA induces coherent relation geometry (higher Silhouette and meaningful NMI recovery of transformation types), while pretrained-only and fixed-difference baselines exhibit negligible structure.

These results connect (i) objective–regime compatibility (Table 1), (ii) the necessity of relation-space prediction (Table 2), and (iii) generalization under entity shift (Table 3), with (iv) interpretable geometric organization as supporting evidence (Ta-

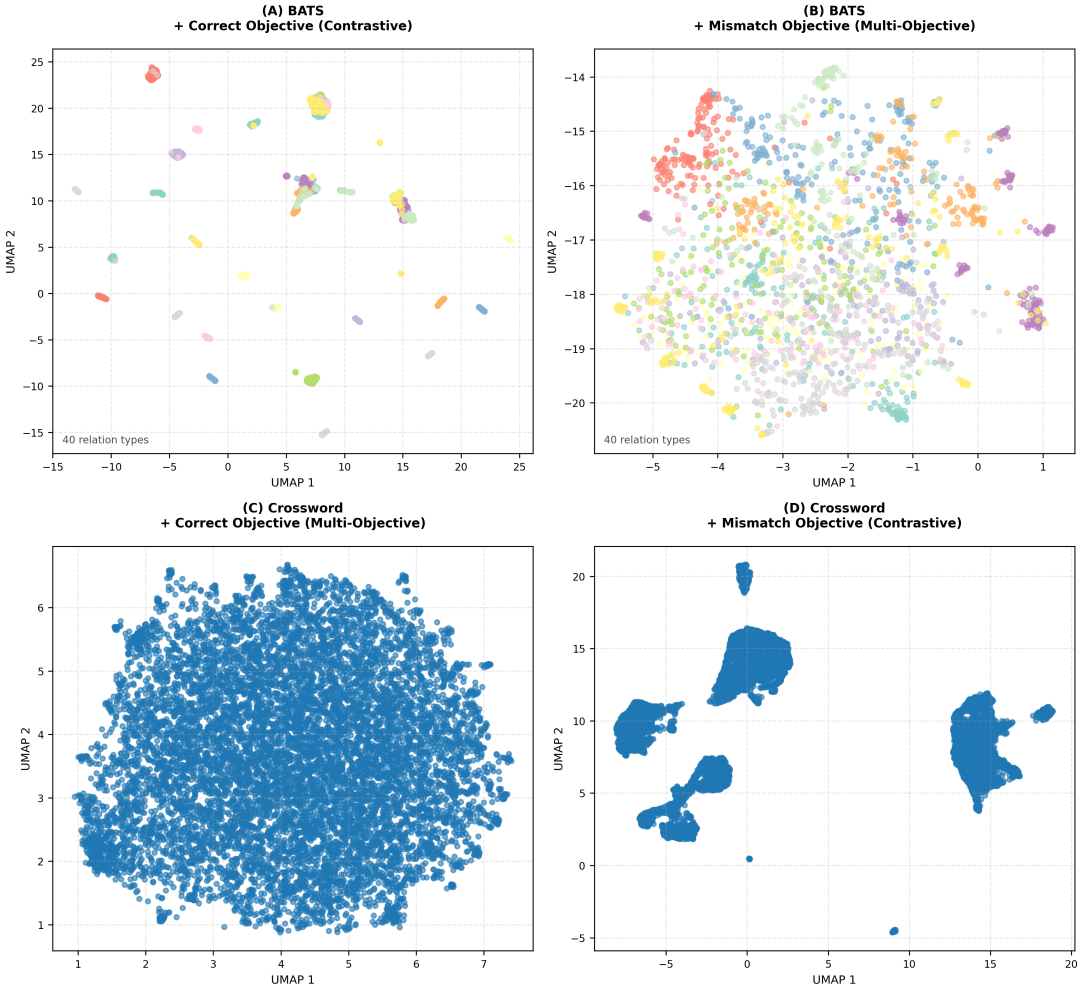

Figure 3: Objective–regime alignment in relation embedding space. Top: relation-discriminating setting (BATS). Bottom: relation-sharing setting (Crossword). Left: objective matched to the dataset regime. Right: mismatched objective. Colors indicate ground-truth transformation families (used only for visualization).

ble 4).

7 Conclusion

We investigated when self-supervised relational learning is possible and argued that its success is governed by structural conditions rather than by architecture alone. Specifically, we showed that relational invariance can only be learned when the training objective is compatible with the relational regime induced by the data. Failure to satisfy this condition leads to misleading structure in relation space, where apparent clustering does not correspond to meaningful transformations.

To instantiate relational learning under these conditions, we proposed Relational JEPA (R-JEPA), a joint embedding predictive architecture that treats transformations as first-class objects of prediction. By shifting the predictive target from states to relations and enforcing prediction exclusively in relation space, R-JEPA exposes relational learning signals when objective–regime compatibility holds. Across multiple analogy benchmarks, this enables improved verification, completion, and zero-shot transfer compared to state-based baselines.

More broadly, our findings suggest that progress in analogical and relational representation learning requires explicit attention to data structure and objective design. Rather than seeking universally applicable objectives, future work should characterize the relational regimes induced by data and design learning signals accordingly. We view R-JEPA not as a universal solution, but as a practical instantiation that succeeds precisely under the conditions where self-supervised relational learning is possible.

8 Limitations

Our work is subject to several limitations that point to promising directions for future research. First, R-JEPA assumes access to pairs of transformations that are approximately analogous. While synthetic data provides clean control over relational structure, mining analogical pairs from real-world corpora relies on weak heuristics and may introduce noise or bias. Although we explicitly evaluate robustness to noisy supervision, fully self-supervised discovery of high-quality relational pairs remains an open challenge.

Second, our formulation focuses on single-step transformations represented by a single relation embedding. More complex analogies may involve multi-step, hierarchical, or compositional transformations that cannot be captured by a single relation vector. Extending relational invariance learning to such settings likely requires compositional operators over relations or integration with longer-horizon predictive dynamics.

Third, our analysis highlights that relational learning is a conditional problem governed by data–objective compatibility. Datasets that do not instantiate sufficient relational variability may fundamentally limit what can be learned through self-supervision, regardless of architectural choices. While R-JEPA exposes relational learning signals when the necessary conditions are met, it does not eliminate these inherent data constraints.

Finally, we evaluate relational abstraction operationally through analogy verification, completion, and zero-shot transfer. These tasks probe systematic relational generalization but do not constitute a complete account of semantic understanding. Future work may explore complementary evaluations that capture broader aspects of relational reasoning.

References

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

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3253–3262.

Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. Wordrep: A benchmark for research on learning word representations. *arXiv preprint arXiv:1407.1640*.

Dedre Gentner. 1988. Metaphor as structure mapping: The relational shift. *Child development*, pages 47–59.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Sayan Ghosh and Shashank Srivastava. 2022. epic: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, and 1 others. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Saul Pwanson. 2016. Download crossword data. https://github.com/century-arcade/xd. Accessed: 2025-12-02.

Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, and 1 others. 2020. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Dwijen Rudrapal, Amitava Das, and Baby Bhattacharya. 2017. Quotology-reading between the lines of quotations. In *International Conference on Applications of Natural Language to Information Systems*, pages 292–296. Springer.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. 2022. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pages 5628–5637. PMLR.

William Timkey and Marten Van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv preprint arXiv:2109.04404*.

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

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624.

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

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. Analogical–a novel benchmark for long text analogy evaluation in large language models. *arXiv preprint arXiv:2305.05050*.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR.

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. 2023. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*.

A Additional Experimental Details and Ablations

This appendix provides supplementary experimental details and analyses that support the main claims of the paper. While the main text focuses on isolating the conditions under which relational invariance emerges, the appendix addresses additional ablations, alternative evaluation settings, and implementation details that may be of independent interest.

A.1 Training Protocol and Evaluation Setup

All models are trained using the same encoder architecture and relation extractor unless otherwise specified. For most datasets, models are trained for 10 epochs and evaluated on held-out test splits. For BATS ablation studies, several configurations are trained for 5 epochs and evaluated on validation splits due to computational constraints. We explicitly mark validation results in tables where applicable.

Entity-disjoint evaluation splits are constructed by ensuring that entities appearing in training do not overlap with those in evaluation. This setting tests whether relation embeddings encode transformation-level structure rather than memorizing entity-specific associations.

Unless otherwise stated, hyperparameters are shared across ablations to isolate architectural and objective-level effects.

A.2 Detailed Ablation Study on BATS

We provide a detailed ablation study on BATS to analyze the contribution of individual components in a relation-discriminating regime.

Table 5: BATS ablation results. Results marked with $*$ report validation accuracy after 5 epochs; others report test accuracy after 10 epochs.

Method	Acc. (%)	Ep.	Notes
R-JEPA (ours)	**93.7**	10	Full model
No predictor (contrastive)	91.3	10	Prediction removed
Entity-disjoint training	91.5*	5	Entities separated
No hard negatives	90.8*	5	Hard negatives removed
Reduced dim (64)	90.3*	5	Smaller relation space
Diff-only ($r = z_B - z_A$)	87.7*	5	Simple difference
Multi-objective (mismatch)	58.9	10	Objective mismatch

Table 6: Dataset characteristics and relational regimes.

Dataset	Regime	#Rel	Domain	Size	Obj.
BATS	Disc.	40	Word analogy	2.8K	Energy
WordNet	Disc.	8	Lexical relations	4K	Energy
SNLI-Syn	Disc.	3	Syntax transforms	13K	Energy
Entail.	Share	1	NLI	10K	Multi
Crossword	Share	1	Clue–Answer	10K	Multi
ePiC	Share	1	Proverb–Narrative	0.2K	Multi

These results indicate that while several components contribute incrementally, objective–regime alignment and relation-level prediction are the dominant factors for stable relational learning.

A.3 Dataset Statistics and Relational Regimes

Table 6 summarizes the datasets used in our experiments, their induced relational regimes, and the objective families that empirically perform best under each regime. Regime labels reflect relational variability and negative-sample semantics rather than intrinsic dataset annotations, and are provided to support the regime-based experimental design in the main text.

A.4 Qualitative Failure Modes Under Objective Mismatch

To complement the geometric analysis in the main text, we qualitatively inspect nearest neighbors from dense regions of the relation space learned under an objective–regime mismatch. Specifically, we analyze relation embeddings learned with a contrastive objective on the Crossword dataset, which induces a relation-sharing regime.

Table 7 presents representative nearest neighbors sampled from dense regions of the learned relation space. Although relation embeddings form tight geometric clusters under the contrastive objective, the retrieved neighbors do not exhibit a consistent transformation pattern. Pairs grouped together in relation space often correspond to semantically unrelated clue–answer mappings, indicating that apparent geometric structure does not imply relational invariance.

This qualitative evidence supports the quantitative results in the main text: under objective–regime mismatch, contrastive learning can induce visually coherent clusters that nevertheless fail to capture meaningful transformation-level regularities.

Table 7: Nearest neighbors under objective mismatch on Crossword. Although embeddings form tight clusters, nearest neighbors do not share a consistent transformation pattern.

Center Pair	Nearest Neighbors in Relation Space (Top-5)
Michigan city → FLINT	Seller of Toughskins kids' clothing → SEARS; Was corrosive → ATE; Leader of Islam → ALLAH; You Do the __ (Brad Paisley) → MATH; Ski-boot binding clip → FLANGE
Farewell act → SWANSONG	In the offing → OFFING; Canon and martial → LAWS; One way to go → RIDEAWAY; Site of a baseball merger? → SEAM; Way out? → AVEOFESCAPE
Sweet-talked → CHARMED	Tear it up → SLAY; Physicist James → JOULE; Tries to put out, as a small fire → STAMPSON; __ white → CHINESE; Have a meal at noon, perhaps → TAKELUNCH
Cooperated → PLAYEDBALL	Oktoberfest locale → BEERGARDEN; Cupid, in a popular song → LOVEBUG; Take out (the trash) → EMPTY; Popular YMCA class → YOGA; African fox → ASSE
Features of many wedding cakes → TIERS	Place for a shrine → APSE; Vast sums → MINTS; Proponents → FORS; Site of ancient Roman ruins → OSTIA; Beef, e.g. → REDMEAT

A.5 Limitations of Verification Metrics in Relation-Sharing Regimes

For datasets dominated by a single relation type (e.g., Crossword, ePiC), binary verification accuracy is an imperfect evaluation signal. In such settings, most pairs are trivially compatible, and nearest-neighbor retrieval or ranking-based metrics provide a more faithful assessment of relational structure. We leave a systematic exploration of alternative evaluation protocols for relation-sharing regimes to future work.