Learning Relational Invariance Requires Objective–Regime Compatibility

Anonymous ACL submission

Abstract

Relational and analogical reasoning requires invariance over *transformations* rather than over individual states. Yet such invariance does not emerge reliably from a single generic training objective. We argue that successful relational representation learning depends on *objective–regime compatibility*: some datasets induce *relation-discriminating* regimes, where multiple transformation families must be separated, while others induce *relation-sharing* regimes, where many apparent negatives remain semantically compatible and contrastive repulsion becomes harmful. To study this claim, we keep the encoder and relation extractor fixed and vary only the objective geometry in relation space. Across two relation-discriminating datasets (BATS, BLESS) and two relation-sharing subsets (SNLI entailment, SNLI contradiction), matched objectives consistently outperform mismatched ones. In discriminating regimes, contrastive training yields substantially higher verification and retrieval performance and stronger family separation; in sharing regimes, alignment-dominant training avoids the severe fragmentation induced by contrastive repulsion. These results suggest that progress in relational learning depends not only on model architecture, but also on matching the learning objective to the relational structure induced by the data.

1 Introduction

Relational and analogical reasoning concerns equivalence between *transformations*, not similarity between individual entities (Gentner, 1988; Gentner and Markman, 1997; Turney, 2005, 2006; Jurgens et al., 2012). Given two pairs (A, B) and (C, D), the central question is whether $A \rightarrow B$ expresses the same transformation as $C \rightarrow D$. This differs from standard representation learning, which is typically designed to preserve invariance over individual observations or states. Strong state representations therefore do not guarantee strong relational representations.

This distinction is visible across analogy research in NLP. Classical word analogy work often represents relations implicitly through vector offsets or retrieval heuristics (Mikolov et al., 2013; Turney, 2006), while later analyses showed that apparent analogy success can depend on embedding geometry, answer-space effects, or entity similarity rather than on transformation-level abstraction (Gladkova et al., 2016; Levy and Goldberg, 2014; Rogers et al., 2017; Vylomova et al., 2016). Recent work has extended analogical evaluation to contextualized and long-context settings (Ushio et al., 2021; Wijesiriwardene et al., 2023; Ye et al., 2024). What remains unclear is when relational invariance can be learned reliably.

Our claim is that the answer depends on *objective–regime compatibility*. Relational datasets do not impose the same learning demands. Some contain multiple transformation families and require separation across them. Others are dominated by a shared transformation pattern, so many apparent negatives remain semantically compatible. The same objective can therefore help in one setting and hurt in another. Contrastive learning is useful when negatives correspond to genuinely different relations, but harmful when many negatives are ambiguous or false (Saunshi et al., 2019; Chuang et al., 2020; Robinson et al., 2021). By contrast, objectives without explicit repulsion can preserve alignment in shared-relation settings while under-separating transformations that should remain distinct (Wang and Isola, 2020; Chen and He, 2021; Zbontar et al., 2021; Grill et al., 2020).

We formalize this contrast with two dataset-induced regimes. *Relation-discriminating* regimes contain multiple transformation families and require within-family invariance together with cross-family separation. *Relation-sharing* regimes are dominated by a shared transformation pattern, mak-

ing many in-batch negatives ambiguous or false. We argue that relational invariance emerges reliably only when the training objective matches the regime induced by the data.

To test this claim, we use a fixed relation representation pipeline that maps ordered pairs into explicit relation embeddings and applies learning directly in relation space. This design isolates objective effects by holding the encoder and relation extractor fixed, and reduces shortcuts based on entity similarity alone. Figure 1 summarizes the full experimental logic: the same frozen relation pipeline is applied across both regimes, only the objective changes, and mismatch produces a distinct geometric failure mode in each case.

Across four datasets, we find the same reversal. On relation-discriminating datasets, matched contrastive training improves verification, retrieval, and structural separation. On relation-sharing subsets, verification and retrieval saturate under both objectives, but the learned geometry diverges: contrastive training fragments relation space, whereas alignment-dominant training preserves a compact shared manifold. This reversal is the main empirical signature of objective–regime compatibility.

We therefore frame relational representation learning as a problem of objective–regime compatibility, identify two failure modes of mismatch — under-separation in relation-discriminating regimes and false-negative fragmentation in relation-sharing regimes — and show, with a fixed relation-space pipeline, that matched objectives outperform mismatched ones across BATS, BLESS, and SNLI-derived sharing subsets. We further show that simple dataset-level audit statistics recover the correct objective family for all four datasets.

2 Related Work

2.1 Relational Similarity and Analogical Representation Learning

Analogy has long been studied as a problem of *relational similarity* rather than attribute-level resemblance (Gentner, 1988; Gentner and Markman, 1997; Turney, 2005, 2006). In NLP, this view appears in word analogy benchmarks that test whether models recover consistent transformations across lexical pairs (Mikolov et al., 2013; Gladkova et al., 2016), in broader evaluations of graded relational similarity (Jurgens et al., 2012; Rogers et al., 2017), and in recent studies of analogy in contextualized and long-context language models (Ushio et al., 2021; Wijesiriwardene et al., 2023; Ye et al., 2024). However, much of this literature asks whether analogical structure is present, whereas we ask when it can be learned reliably.

2.2 Lexical Relation Learning from Pair Representations

A related literature studies lexical relations from explicit pair representations. Early approaches often relied on vector differences or other simple pairwise constructions, but later work showed that these signals are not uniformly reliable across relation types (Vylomova et al., 2016; Bouraoui et al., 2018). Subsequent models proposed more explicit relation-oriented spaces for distinguishing lexical relations such as synonymy, hypernymy, and meronymy (Wang et al., 2019). We build on this line of work in using explicit ordered-pair representations, but our focus is not better lexical relation classification. Instead, we ask how the learning objective reshapes relation space under different dataset regimes.

2.3 Contrastive and Non-Contrastive Representation Learning

Representation learning methods differ in the geometry they impose. Contrastive methods align positives and repel negatives (Oord et al., 2018; Chen et al., 2020; Saunshi et al., 2022; Tosh et al., 2021), while non-contrastive methods show that useful representations can also emerge without explicit negatives, through alignment, redundancy reduction, predictive bootstrapping, or distributional regularization (Wang and Isola, 2020; Grill et al., 2020; Chen and He, 2021; Zbontar et al., 2021; Assran et al., 2023). Most of this work studies invariance over instances or states. Our setting instead concerns invariance over transformations, so objective choice must be evaluated relative to the relational demands of the data.

2.4 False Negatives and Relational Regimes

False negatives are a central failure mode in contrastive learning: negatives may be semantically compatible rather than truly distinct (Saunshi et al., 2019; Chuang et al., 2020; Robinson et al., 2021). We connect this problem to relational learning by arguing that false negatives can be a structural property of a dataset's *relational regime*. In relation-sharing regimes, many distinct instances remain compatible at the transformation level, so explicit

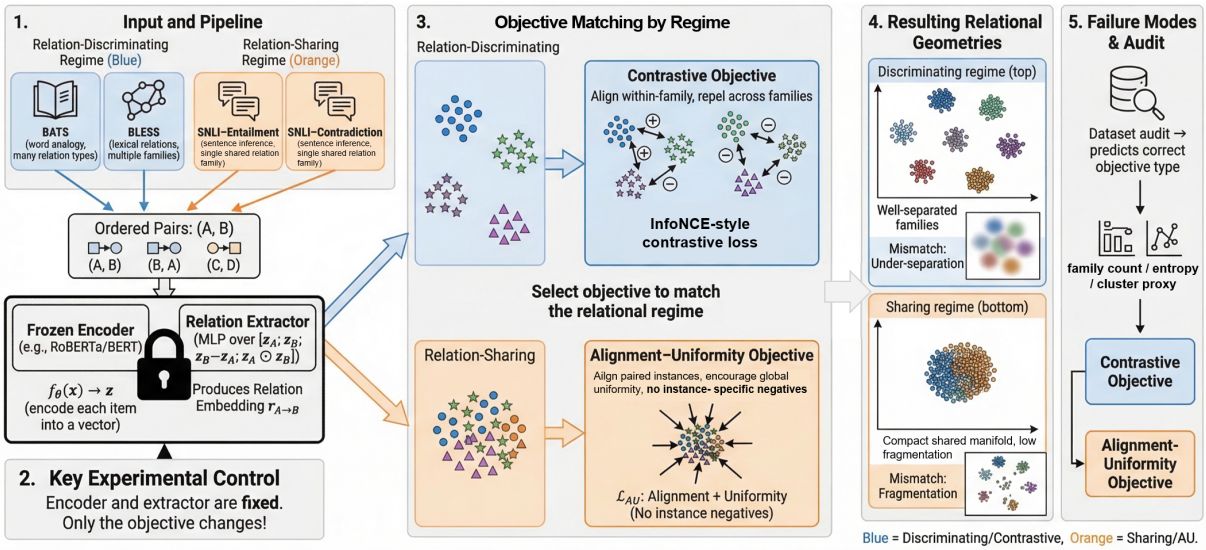

Figure 1: Overview of the objective–regime compatibility framework. The same fixed relation pipeline is used across relation-discriminating and relation-sharing regimes, while only the learning objective changes. Relation-discriminating regimes favor contrastive learning with explicit separation across families, whereas relation-sharing regimes favor alignment–uniformity learning without instance-specific negatives. These matched choices produce distinct relation geometries and can be predicted by simple dataset audit statistics.

repulsion is misaligned with the data. In relation-discriminating regimes, by contrast, separation across families is necessary, making contrastive pressure beneficial.

2.5 Sentence Representation Learning and NLI Supervision

Our use of SNLI-derived subsets also relates to sentence representation learning from natural language inference supervision. Sentence-BERT and SimCSE showed that paired sentence objectives can produce strong semantic sentence embeddings, and supervised SimCSE uses entailment pairs as positives and contradiction pairs as hard negatives (Reimers and Gurevych, 2019; Gao et al., 2021). We use SNLI differently. Rather than treating entailment and contradiction as supervision for generic sentence similarity, we treat each subset as inducing a relation-sharing regime and ask what geometric pressures are appropriate in that setting.

3 Problem Formulation

We study relational representation learning from ordered pairs without explicit relation labels in the loss.

States. Let $x \in \mathcal{X}$ denote an observation, such as a word or sentence. An encoder f_θ maps x to a latent state

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

Relations as transformations. Given an ordered pair (x_A, x_B) with states (z_A, z_B), a relation extractor ϕ_ψ produces a relation embedding

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

In our implementation, ϕ_ψ is an MLP over concatenated pair features,

$$r_{A\to B} = \mathrm{MLP}([z_A; z_B; z_B - z_A; z_A \odot z_B]). \tag{3}$$

Analogical equivalence and relational invariance. Two pairs (A, B) and (C, D) are analogically equivalent when $r_{A\to B}$ and $r_{C\to D}$ are close under a task-defined similarity metric. More generally, relational invariance maps instances of the same transformation family $\mathcal{T}$ to nearby points in relation space.

Relational regimes. We distinguish two dataset-induced regimes:

- **Relation-discriminating regime:** multiple transformation families coexist, so successful learning requires within-family invariance and cross-family separation.
- **Relation-sharing regime:** the data are dominated by a shared transformation pattern, so many apparent negatives remain semantically compatible and explicit repulsion can introduce widespread false negatives.

Objective–regime compatibility. In relation-discriminating regimes, objectives with explicit repulsion help separate distinct transformation families. In relation-sharing regimes, the same repulsion can conflict with the data by pushing apart compatible relation instances. Our central claim is that relational invariance learning requires *objective–regime compatibility*.

4 Method

4.1 Fixed Relation Representation Pipeline

We keep the encoder and relation extractor fixed across all experiments and vary only the learning objective. This lets us attribute differences in performance and geometry to objective choice rather than to architectural variation.

Each relation instance is an ordered pair $x_i = (a_i, b_i)$. We encode both elements with a frozen pretrained backbone,

$$z_i^{(a)} = f_\theta(a_i), \qquad z_i^{(b)} = f_\theta(b_i), \tag{4}$$

where the main backbone is frozen RoBERTa-base, and frozen BERT-base is used for robustness analysis. We then construct a relation embedding

$$\begin{aligned} r_i &= \phi_\psi\Big(z_i^{(a)}, z_i^{(b)}\Big) \\ &= \mathrm{MLP}\Big([z_i^{(a)}; z_i^{(b)}; z_i^{(b)} - z_i^{(a)}; z_i^{(a)} \odot z_i^{(b)}]\Big). \end{aligned} \tag{5}$$

4.2 Positive Construction

Training relies on constructed positive correspondence.

Relation-discriminating datasets. For BATS and BLESS, we use explicit family labels from dataset construction. Two relation instances form a positive pair iff they belong to the same family:

$$(x_i, x_j) \in \mathcal{P} \iff y_i = y_j,\ i \neq j. \tag{6}$$

All other families are available as candidate negatives under contrastive training.

Relation-sharing datasets. For SNLI entailment and SNLI contradiction subsets, each subset is treated as a single shared relation family. Positive correspondence is constructed within the subset rather than assumed over all same-subset instances. In the main setup, for each anchor instance we select a semantically similar partner from the same subset using nearest-neighbor pairing under frozen encoder features. Under the alignment-dominant objective, other batch elements contribute only to the global uniformity term and are not treated as instance-specific negatives.

4.3 Objective Families

Contrastive objective. For relation-discriminating regimes, we use a contrastive objective with explicit repulsion. Let

$$s_i^+ = \mathrm{sim}(r_i, r_i^+), \qquad s_{ij} = \mathrm{sim}(r_i, r_j), \tag{7}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity. We define

$$Z_i = \exp(s_i^+/\tau) + \sum_{j \in \mathcal{N}(i)} \exp(s_{ij}/\tau), \tag{8}$$

and optimize

$$\mathcal{L}_{\mathrm{ctr}}^{(i)} = -\log \frac{\exp(s_i^+/\tau)}{Z_i}, \tag{9}$$

where τ is a temperature parameter, r_i^+ is the positive relation embedding associated with anchor r_i, and $\mathcal{N}(i)$ is the negative set.

Alignment–uniformity objective. For relation-sharing regimes, we use an alignment-dominant objective without instance-specific negatives:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{B} \sum_{i=1}^{B} \|r_i - r_i^+\|_2^2, \tag{10}$$

$$\mathcal{L}_{\mathrm{uni}} = -\log \left(\frac{1}{B(B-1)} \sum_{i \neq j} \exp(-t\|r_i - r_j\|_2^2) \right), \tag{11}$$

and

$$\mathcal{L}_{\mathrm{AU}} = \lambda_{\mathrm{align}} \mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{uni}} \mathcal{L}_{\mathrm{uni}}. \tag{12}$$

4.4 Failure-Mode Metrics

To characterize mismatch, we track different geometric statistics for the two regimes.

Under-separation. For relation-discriminating datasets, we measure family separation margin:

$$\begin{aligned} \Delta_{\mathrm{sep}} = \mathbb{E}[d(r_i, r_j) \mid y_i \neq y_j] \\ - \mathbb{E}[d(r_i, r_j) \mid y_i = y_j]\,. \end{aligned} \tag{13}$$

Larger values indicate better between-family separation relative to within-family compactness.

Fragmentation. For relation-sharing datasets, we measure average within-subset spread as mean pairwise squared Euclidean distance:

$$\Delta_{\text{frag}} = \mathbb{E}\left[\|r_i - r_j\|_2^2 \mid x_i, x_j \in D_{\text{share}}\right]. \quad (14)$$

Higher values indicate stronger fragmentation of what should remain a coherent shared relation space.

5 Experiments

Our goal is to test a single claim: relational invariance learning requires objectives that match the relational regime induced by the data. We therefore keep the encoder and relation extractor fixed in the main experiments and vary only the objective geometry.

5.1 Datasets

We use four datasets or derived subsets. BATS is a word analogy benchmark with 40 explicit relation subcategories and 1,400 instances in our processed setting, and serves as a relation-discriminating dataset. BLESS is a lexical relation dataset with 6 explicit relation families and 37,173 instances in our processed setting, and serves as a second relation-discriminating dataset. For relation-sharing settings, we extract 50,000 entailment instances and 50,000 contradiction instances from SNLI, treating each subset as a single shared inference family. This differs from prior NLI-based sentence representation learning, where entailment pairs are often used as positives and contradiction pairs as hard negatives for sentence embedding supervision (Reimers and Gurevych, 2019; Gao et al., 2021). Here, we use SNLI instead for dataset-regime analysis in relation space.

5.2 Regime Audit

Before training, we compute simple audit statistics from the processed datasets. Table 1 shows that these statistics recover the intended regime assignments for all four datasets.

5.3 Experimental Setup and Evaluation

Our main backbone is frozen RoBERTa-base. For word-level datasets, words are encoded in a short template; for sentence-level datasets, we use mean pooling over token representations. We train a shared relation extractor on top of the frozen encoder. We compare two objective families: contrastive training with explicit negatives, and alignment–uniformity training without instance-specific negatives. For each dataset, matched training uses the objective predicted by the dataset regime, and mismatched training swaps the objective. All models are trained for up to 30 epochs with early stopping based on validation performance.

We report verification accuracy, Recall@1, and MRR on all datasets. For relation-discriminating datasets, we additionally report separation margin, silhouette, and NMI. For relation-sharing datasets, we report fragmentation. As ablations, we test backbone robustness by replacing frozen RoBERTa-base with frozen BERT-base, and relation-extractor robustness by replacing the MLP extractor with a difference-only extractor.

Table 1: Dataset audit statistics. Explicit family count and family entropy reflect dataset-level relation diversity. The audit correctly predicts the best objective family for all four datasets.

Dataset	# Fam.	Entropy	Cluster Proxy	Pred. Obj.
BATS	40	5.3219	-0.0635	Ctr
BLESS	6	2.2139	-0.0191	Ctr
SNLI-Ent	1	0.0000	0.0000	AU
SNLI-Contra	1	0.0000	0.0000	AU

6 Results

6.1 Objective–Regime Compatibility Is a Structural Requirement

Table 2 shows the central result. On both relation-discriminating datasets, matched contrastive training strongly outperforms mismatched alignment-dominant training. On both relation-sharing subsets, verification and retrieval saturate under both objectives, but the geometric behavior differs sharply: mismatched contrastive training produces severe fragmentation, whereas matched alignment-dominant training preserves a compact shared relation manifold.

The reversal is especially clear on BATS and BLESS. On BATS, contrastive training yields 0.556 verification accuracy and 2.027 separation margin, whereas mismatched AU training collapses to 0.106 accuracy and nearly zero separation. On BLESS, the gap is even larger: 0.933 versus 0.106 in verification accuracy. These results indicate that relation-discriminating settings require explicit separation across families.

For SNLI entailment and contradiction, verification and retrieval saturate under both objectives. In these settings, task performance alone is not in-

Table 2: Main results. In relation-discriminating datasets, matched contrastive training strongly outperforms mismatched AU training on task performance and structure. In relation-sharing datasets, task metrics saturate under both objectives, so fragmentation becomes the main diagnostic of geometric mismatch. Lower fragmentation is better.

Dataset	Setting	Verif. Acc	Recall@1	MRR	Sep. Margin	Silhouette	NMI / Frag.
BATS	Matched (Ctr)	0.5563	0.5563	0.6796	2.0275	0.0957	0.7219
BATS	Mismatched (AU)	0.1063	0.1063	0.1998	0.0004	-0.2786	0.2580
BLESS	Matched (Ctr)	0.9328	0.9323	0.9542	1.4598	0.6518	0.6365
BLESS	Mismatched (AU)	0.1056	0.2379	0.3460	-0.0000	-0.1731	0.0087
SNLI-Entailment	Mismatched (Ctr)	1.0000	1.0000	1.0000	–	–	20.5866
SNLI-Entailment	Matched (AU)	1.0000	1.0000	1.0000	–	–	0.0002
SNLI-Contradiction	Mismatched (Ctr)	1.0000	1.0000	1.0000	–	–	20.3312
SNLI-Contradiction	Matched (AU)	1.0000	1.0000	1.0000	–	–	0.0002

formative enough to diagnose relational structure. However, fragmentation differs dramatically: contrastive training yields fragmentation values above 20 on both subsets, whereas matched AU training keeps fragmentation near zero. This strongly supports the claim that explicit repulsion is geometrically harmful in relation-sharing regimes, even when downstream verification appears trivial.

6.2 Mismatch Produces Distinct Failure Modes

Mismatch produces qualitatively different failures in the two regimes.

In relation-discriminating datasets, AU training fails by *under-separation*. On BATS, separation margin drops from 2.0275 to 0.0004, silhouette drops from 0.0957 to -0.2786, and NMI drops from 0.7219 to 0.2580. On BLESS, the same pattern appears: matched contrastive training yields strong family recovery, while mismatched AU training collapses structure almost completely.

In relation-sharing datasets, contrastive training fails by *fragmentation*. Although verification remains saturated, the relation space becomes highly dispersed under explicit repulsion. On SNLI entailment, fragmentation increases from 0.0002 under AU to 20.5866 under contrastive training. The same holds for contradiction, where fragmentation increases from 0.0002 to 20.3312.

6.3 The Pattern Is Stable Across Backbones

Table 3 shows that the main pattern is not specific to RoBERTa. Replacing the backbone with frozen BERT-base preserves the same objective preference reversal.

Interestingly, BERT produces stronger absolute performance than RoBERTa in several settings, especially on BATS and BLESS. However, this changes neither the qualitative conclusion nor the failure signature: matched contrastive training remains superior on discriminating datasets, and matched AU training remains far less fragmented on sharing datasets. This supports our claim that the central phenomenon concerns objective–regime interaction rather than a particular pretrained encoder.

6.4 The Pattern Is Also Stable Across Relation Extractors

We next replace the MLP relation extractor with a simpler difference-only extractor. Table 4 shows that the same qualitative pattern remains, although the stronger MLP extractor improves absolute performance on the discriminating datasets.

The weaker extractor narrows some gaps, especially on BLESS, but does not remove the objective preference reversal. This suggests that extractor capacity affects the strength of relation learning, while objective–regime compatibility determines whether the geometry is qualitatively appropriate.

Robustness to correspondence and negative construction. To test whether the main findings depend on a specific positive-pair construction rule or a naive contrastive negative sampler, we conducted additional sensitivity analyses. Across multiple positive construction variants, including random same-family pairing, nearest-neighbor pairing, and filtered pairing, the qualitative objective preference reversal remained stable. We further tested alternative negative sampling strategies in the SNLI entailment sharing setting. Although entity-perturbed negatives reduced contrastive fragmentation relative to naive in-batch negatives, they remained far more fragmented than alignment-dominant training.

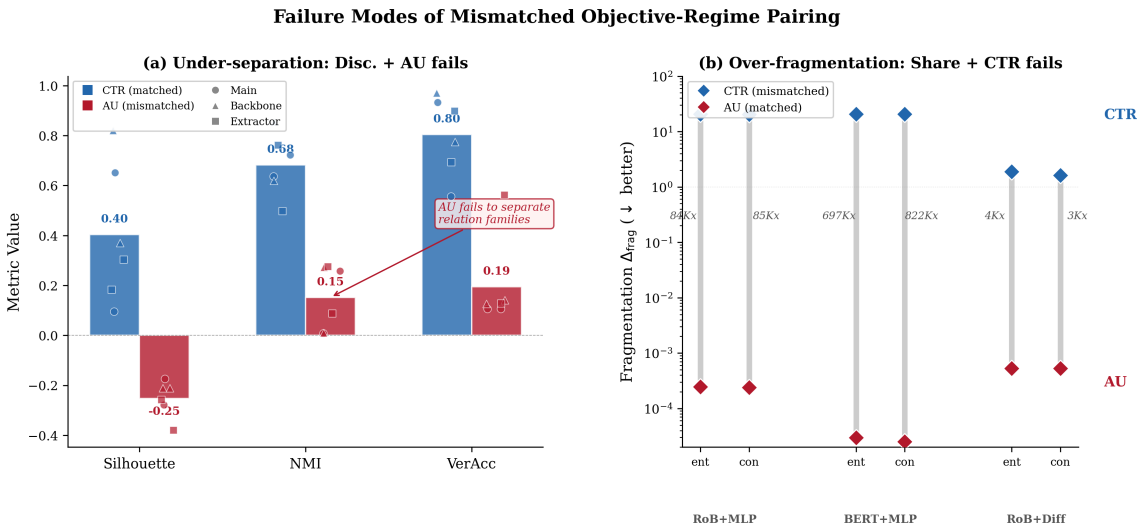

Figure 2: Failure modes under objective mismatch. Left: under-separation on relation-discriminating datasets, measured by separation margin. Right: false-negative fragmentation on relation-sharing datasets, measured by within-subset spread (lower is better).

Table 3: Backbone robustness with frozen BERT-base. The same objective preference reversal appears across all four datasets.

Dataset	Setting	Verif. Acc	Recall@1	MRR	Sep. Margin	Silhouette	NMI / Frag.
BATS	Matched (Ctr)	0.7750	0.7750	0.8397	2.3800	0.3695	0.8549
BATS	Mismatched (AU)	0.1406	0.1406	0.2326	0.0014	-0.2115	0.2719
BLESS	Matched (Ctr)	0.9704	0.9704	0.9790	4.4642	0.8204	0.6191
BLESS	Mismatched (AU)	0.1270	0.1797	0.3350	-0.0001	-0.2119	0.0102
SNLI-Entailment	Mismatched (Ctr)	1.0000	1.0000	1.0000	–	–	20.6238
SNLI-Entailment	Matched (AU)	1.0000	1.0000	1.0000	–	–	0.0000
SNLI-Contradiction	Mismatched (Ctr)	1.0000	1.0000	1.0000	–	–	20.7346
SNLI-Contradiction	Matched (AU)	1.0000	1.0000	1.0000	–	–	0.0000

These results indicate that the main trend is not an artifact of a single pairing heuristic or a naive negative construction rule; detailed sensitivity results are provided in Appendix E and Appendix F.

6.5 Regime Assignment Predicts the Better Objective

The dataset audit in Table 1 correctly predicts the preferred objective family for all four datasets. This matters because it moves the paper beyond a purely retrospective claim. We are not only observing that different datasets prefer different objectives after training. Rather, simple dataset-level statistics already identify whether relation learning is more likely to benefit from explicit separation or from alignment without repulsion.

7 Discussion

Our results support a simple but important claim: relational representation learning is not determined by architecture alone. It also depends on whether the learning objective matches the relational regime induced by the data. This reframes the central design question. The issue is not whether contrastive or non-contrastive learning is universally better, but which geometric pressure is appropriate for the transformation structure that a dataset actually contains.

This perspective helps explain two distinct failure modes. In relation-discriminating regimes, objectives without sufficient repulsion under-separate transformation families that should remain distinct. In relation-sharing regimes, explicit repulsion pushes apart relation instances that remain compatible at the transformation level, producing fragmentation through pervasive false negatives. These two mismatches are different in mechanism, but they follow the same principle: the wrong objective imposes the wrong geometry.

Our experiments also show why task accuracy

Table 4: Relation extractor ablation with a difference-only extractor. The same objective preference reversal remains visible, though absolute performance changes.

Dataset	Setting	Verif. Acc	Recall@1	MRR	Sep. Margin	Silhouette	NMI / Frag.
BATS	Matched (Ctr)	0.6938	0.6938	0.7669	0.2048	0.1827	0.7613
BATS	Mismatched (AU)	0.1281	0.1281	0.2121	0.0054	-0.2581	0.2754
BLESS	Matched (Ctr)	0.8975	0.8960	0.9239	0.1216	0.3035	0.4975
BLESS	Mismatched (AU)	0.5617	0.5905	0.6992	-0.0004	-0.3801	0.0881
SNLI-Entailment	Mismatched (Ctr)	1.0000	1.0000	1.0000	–	–	1.8822
SNLI-Entailment	Matched (AU)	1.0000	1.0000	1.0000	–	–	0.0005
SNLI-Contradiction	Mismatched (Ctr)	1.0000	1.0000	1.0000	–	–	1.6235
SNLI-Contradiction	Matched (AU)	1.0000	1.0000	1.0000	–	–	0.0005

alone can be misleading. On the SNLI sharing subsets, verification and retrieval saturate under both matched and mismatched objectives, yet the learned relation spaces differ radically in fragmentation. This means that apparent downstream success does not necessarily indicate that a model has learned an appropriate relational geometry. For relational learning, geometric diagnostics are therefore not merely auxiliary; in some regimes they are necessary to detect failure.

More broadly, the paper suggests a practical implication for future work. Progress in relational learning may depend less on searching for a single best objective and more on diagnosing the regime first. Our simple audit statistics are only a first step, but they show that dataset-level signals can already predict whether relation learning is more likely to benefit from explicit separation or from alignment without repulsion. In that sense, objective–regime compatibility is not only an explanatory framework for our results; it is also a practical criterion for choosing how relational learning should be trained.

8 Conclusion

We studied when relational invariance can be learned from ordered pairs without explicit relation labels in the loss. Our main claim is intentionally modest but consequential: a single generic objective does not suffice across relational datasets. Instead, successful relational learning depends on *objective–regime compatibility*. Relation-discriminating regimes require explicit separation among transformation families, whereas relation-sharing regimes are vulnerable to false negatives under contrastive training.

To test this claim, we used a fixed relation-space learning pipeline that keeps the encoder and relation extractor constant while varying only the objective geometry. Across BATS, BLESS, and two SNLI-derived sharing subsets, matched objectives consistently outperformed mismatched ones. In discriminating regimes, contrastive training improved verification, retrieval, and structural separation. In sharing regimes, alignment-dominant training prevented the severe fragmentation induced by explicit repulsion. Backbone and extractor ablations showed that the pattern remains stable across implementation choices.

These findings indicate that progress in relational representation learning depends not only on architectural design, but also on matching learning objectives to the relational structure induced by the data.

9 Limitations

Our work has several limitations.

First, our framework assumes access to constructed positive correspondence. Although BATS and BLESS provide explicit family structure, and SNLI provides reliable subset membership from which positive correspondence can be constructed, fully unsupervised discovery of analogous relation instances remains open.

Second, our sharing-regime evaluation reveals a measurement issue: on SNLI entailment and contradiction subsets, verification and retrieval saturate under both matched and mismatched objectives. This means that downstream task accuracy alone may be insufficient to assess relation quality in simple relation-sharing settings. Geometric diagnostics such as fragmentation become essential.

Third, our current regime audit is simple. It correctly predicts the best objective family in the present experiments, but more realistic settings may involve mixed or evolving regimes that require finer-grained diagnostics.

Fourth, although we evaluate several alternative negative sampling strategies and find that the main

contrastive disadvantage in relation-sharing settings remains, we do not exhaustively explore the full design space of contrastive objectives. More sophisticated adaptive or semantically informed negative construction procedures may partially mitigate fragmentation further.

Finally, our formulation focuses on single-step transformations represented by a single relation embedding. More complex analogies may involve multi-step or hierarchical structure that cannot be captured by a single relation vector.

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.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK. Association for Computational Linguistics.

Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2018. Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel R. 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dedre Gentner. 1988. Metaphor as structure mapping: The relational shift. *Child Development*, 59(1):47–59.

Dedre Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56.

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, San Diego, California. Association for Computational Linguistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

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*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Joshua D. Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

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 *Proceedings of the 39th 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 *Proceedings of the 36th International Conference on Machine Learning*, pages 5628–5637. PMLR.

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

Peter D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1136–1141.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

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, Online. Association for Computational Linguistics.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2019. SphereRE: Distinguishing lexical relations with hyperspherical relation embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1727–1737, Florence, Italy. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13060–13082, Miami, Florida, USA. Association for Computational Linguistics.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12310–12320. PMLR.

A Dataset Details

A.1 BATS

BATS (Gladkova et al., 2016) contains word analogy pairs organized into 4 super-families and 40 subcategories: inflectional morphology (10), derivational morphology (10), encyclopedic semantics (10), and lexicographic semantics (10). Each subcategory contains approximately 50 word pairs. We template each word as `"The word is {word}."` and encode it with the frozen backbone. After processing, BATS contains 1,400 instances across 40 families. We split 70%/15%/15% (train/dev/test) via stratified random sampling per family with seed 42.

A.2 BLESS

BLESS (Baroni and Lenci, 2011) contains concept–relatum pairs labeled with one of six relation types: HYPER, MERO, COORD, ATTRI, EVENT, and RANDOM. After stripping POS tags (e.g., `alligator-n` → `alligator`) and applying the same word template, the processed dataset contains 37,173 instances across 6 families with the same 70/15/15 stratified split.

Table 5: Data quality statistics. Entity leakage is the fraction of test entities seen in train. Lexical overlap is the fraction of test pairs whose *both* entities appear in training. FN proxy is the fraction of cross-family pairs sharing a super-family (BATS only).

Dataset	Total N	Entity Leak.	Lex. Overlap	FN Proxy
BATS	1,400	0.40	0.23	0.23
BLESS	37,173	0.90	0.94	–
SNLI-Ent	50,000	0.06	0.00	–
SNLI-Con	50,000	0.04	0.00	–

A.3 SNLI Entailment and Contradiction

We extract entailment and contradiction subsets from SNLI (Bowman et al., 2015), each capped at 50,000 training instances for tractability. We preserve the original SNLI train/dev/test splits. For positive construction, we use nearest-neighbor pairing: for each instance, we compute a pair embedding as the mean of encoded premise and hypothesis, then find the most similar pair embedding from a random pool of 100 candidates within the same subset. This provides a semantically grounded positive correspondence without introducing family-level structure.

A.4 Data Quality Statistics

Table 5 reports additional data quality statistics computed from the processed splits.

BLESS shows notably high entity leakage (0.90) and lexical overlap (0.94) between train and test, because its stratified split distributes pairs of the *same concept* across splits. This means that the same concept (e.g., *alligator*) appears in both train and test, but paired with different relata and under different relation families. While this allows entity-level memorization in principle, the objective preference reversal on BLESS is consistent with the pattern on BATS (which has much lower overlap), suggesting that the main phenomenon is not driven by entity leakage.

SNLI subsets show near-zero lexical overlap, confirming that the sentence-level datasets do not share entity-level information across splits.

B Implementation Details

B.1 Model Architecture

Encoder. The main encoder is frozen RoBERTa-base (125M parameters, hidden dimension 768). For the backbone ablation, we use frozen BERT-base-uncased (110M parameters, same hidden dimension). All parameters of the pretrained encoder remain frozen throughout training. We use mean pooling over token representations with a maximum sequence length of 64 tokens.

Relation extractor. The main relation extractor is an MLP over concatenated pair features:

$$r = \mathrm{MLP}([z_A;\ z_B;\ z_B - z_A;\ z_A \odot z_B]),$$

where the MLP consists of Linear($4 \times 768 \to 768$) $\to$ GELU $\to$ Dropout(0.1) $\to$ Linear($768 \to 256$). The output relation embedding has dimension $k = 256$.

For the extractor ablation, we replace the MLP with a difference-only projection:

$$r = W(z_B - z_A),$$

where $W \in \mathbb{R}^{256\times 768}$.

B.2 Training Protocol

All models are trained with AdamW (weight decay 10^{-4}) at learning rate 10^{-3} with cosine annealing and 3 epochs of linear warmup. Gradient norms are clipped at 1.0. We train for up to 30 epochs with evaluation every 5 epochs and select the best checkpoint by validation performance. Batch size is 128 for all runs. The random seed is fixed at 42 for all experiments.

B.3 Objective Hyperparameters

Contrastive. Temperature $\tau = 0.07$. For each anchor, we sample 1 positive (same family) and 15 negatives. In discriminating regimes, 30% of negatives are drawn from confusable (hard) families and 70% from random different families.

Alignment–uniformity. $\lambda_{\text{align}} = 1.0$, $\lambda_{\text{uni}} = 0.05$, uniformity temperature $t = 2.0$. Positives are constructed via nearest-neighbor pairing; no instance-specific negatives are used.

B.4 Batch Construction

For discriminating datasets, we use a balanced batch sampler that draws 8 families per batch with 16 instances per family (total batch size 128). This ensures sufficient within-batch diversity for contrastive negatives. For sharing datasets, we use simple random batching.

B.5 Computational Budget

Table 6 reports per-run training times for the main experiments. Total wall-clock time across all experimental runs and audit computations (8 main +

8 backbone + 8 extractor + 4 audit computations) was approximately 86,009 seconds (≈23.9 hours) on a single GPU.

Table 6: Training time (seconds) for main experiments.

Dataset	Ctr (s)	AU (s)
BATS	393	57
BLESS	12,188	3,050
SNLI-Ent	3,156	3,154
SNLI-Con	3,212	3,209

The large asymmetry on BATS and BLESS between Ctr and AU reflects the additional cost of hard negative mining and balanced batch sampling in the contrastive setting.

C Evaluation Metric Definitions

All metrics are computed on test-set relation embeddings.

Verification accuracy. For each test instance, we find its nearest neighbor by cosine distance (excluding self). Verification is correct if the neighbor shares the same family label:

$$\text{VerAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_{\text{NN}(i)} = y_i\right].$$

Recall@1. Recall@1 measures whether a same-family instance appears at rank 1 in the retrieval list. In many of our settings it tracks verification accuracy closely, but we report it separately because the retrieval and verification pipelines are computed independently and can differ slightly in practice.

MRR (Mean Reciprocal Rank).

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\text{rank}_i},$$

where rank_i is the rank of the first same-family neighbor in the cosine-distance ranking.

Separation margin (Δ_{sep}).

$$\begin{aligned}\Delta_{\text{sep}} = \mathbb{E}[\|r_i - r_j\|_2 \mid y_i \neq y_j] \\ - \mathbb{E}[\|r_i - r_j\|_2 \mid y_i = y_j].\end{aligned} \tag{15}$$

Higher values indicate better between-family separation relative to within-family compactness.

Silhouette score. Computed using `sklearn.metrics.silhouette_score` with cosine metric over relation embeddings, using ground-truth family labels.

NMI (Normalized Mutual Information). We run K-means clustering (with K equal to the number of true families) on relation embeddings and compute NMI between predicted clusters and true family labels.

Fragmentation (Δ_{frag}). For sharing datasets, we measure mean pairwise squared Euclidean distance within the shared subset:

$$\Delta_{\text{frag}} = \frac{1}{N(N-1)}\sum_{i \neq j} \|r_i - r_j\|_2^2.$$

Higher values indicate more fragmentation of the shared relation space.

D Consolidated Ablation Results

Table 7 consolidates the objective preference ratios across all three configurations.

Table 7: Objective preference ratios (matched / mismatched) across configurations. For discriminating datasets, the ratio is VerAcc(Ctr)/VerAcc(AU). For sharing datasets, the ratio is Frag(Ctr)/Frag(AU).

Dataset	Main	BERT	Diff
BATS (VerAcc)	5.2×	5.5×	5.4×
BLESS (VerAcc)	8.8×	7.6×	1.6×
SNLI-Ent (Frag)	84K×	697K×	3.6K×
SNLI-Con (Frag)	85K×	821K×	3.1K×

The reversal is consistent across all configurations. The difference-only extractor narrows the gap on BLESS (1.6× vs. 8.8×), likely because the simpler extractor has less capacity to exploit rich contrastive structure. On sharing datasets, even the weakest configuration (Diff) still produces fragmentation ratios exceeding 3,000×, confirming that the phenomenon is not extractor-specific.

E Sensitivity to Positive Correspondence Construction

To test whether the main objective–regime pattern depends on a specific positive construction heuristic, we repeated representative experiments under multiple positive pairing rules.

Protocol. For the relation-discriminating setting, we evaluated BATS under two positive construction rules: (i) *random same-family pairing*, where the positive relation was sampled uniformly from the same family, and (ii) *nearest-neighbor same-family pairing*, where the positive relation was selected as the closest same-family example under

frozen encoder features. For the relation-sharing setting, we evaluated the SNLI entailment subset under three positive construction rules: (i) nearest-neighbor pairing within the entailment subset, (ii) random same-label pairing, and (iii) filtered nearest-neighbor pairing, which excludes trivially similar candidates before selecting the nearest neighbor.

Results on BATS. Table 8 shows that the qualitative preference for the matched contrastive objective remains stable across positive construction rules. Although absolute performance changes, contrastive training consistently outperforms alignment-dominant training in verification, retrieval, and family separation.

Table 8: Positive construction sensitivity on BATS. Across multiple positive construction rules, contrastive training remains superior to AU on the relation-discriminating setting.

Positive Rule	Obj.	Verif. Acc	Recall@1	MRR
Random same-family	Ctr	0.459	0.459	0.615
Random same-family	AU	0.347	0.347	0.487
NN same-family	Ctr	0.475	0.475	0.613
NN same-family	AU	0.378	0.378	0.488

Table 9: Structural metrics for BATS under different positive construction rules. The matched contrastive objective consistently yields much larger separation margins and stronger clustering structure.

Positive Rule	Obj.	Sep. Margin	Silhouette	NMI
Random same-family	Ctr	2.627	0.028	0.715
Random same-family	AU	0.008	-0.084	0.561
NN same-family	Ctr	1.336	0.005	0.674
NN same-family	AU	0.007	-0.092	0.568

Results on SNLI entailment. Table 10 shows that the qualitative sharing-regime pattern is also stable across positive construction rules. Verification remains saturated under both objectives, but AU consistently produces a far more coherent relation space, with fragmentation remaining over two orders of magnitude lower than contrastive training across all three positive pairing strategies.

Taken together, these results show that the main objective–regime trend is not an artifact of a single positive construction heuristic.

Table 10: Positive construction sensitivity on SNLI entailment. Verification remains saturated, but AU consistently produces dramatically lower fragmentation than contrastive training across all positive construction variants.

Positive Rule	Obj.	Verif. Acc	Fragmentation
NN	Ctr	1.000	4.202
NN	AU	1.000	0.014
Random same-label	Ctr	1.000	3.482
Random same-label	AU	1.000	0.015
Filtered NN	Ctr	1.000	4.466
Filtered NN	AU	1.000	0.014

F Alternative Negative Sampling in Relation-Sharing Regimes

A natural alternative explanation for contrastive failure in relation-sharing settings is that the problem lies not in explicit repulsion itself, but only in the use of naive random in-batch negatives. To test this possibility, we compared several negative construction strategies on the SNLI entailment sharing subset while keeping the contrastive objective, backbone, and relation extractor fixed.

Protocol. We evaluated three contrastive negative strategies: (i) *random in-batch negatives*, (ii) *filtered in-batch negatives*, which exclude highly similar candidates before sampling negatives, and (iii) *entity-perturbed negatives*, which replace one side of the ordered pair to create a more obviously relation-breaking negative. We compare all three against the alignment-dominant objective, which uses no instance-specific negatives.

Results. Table 11 shows that more selective negatives reduce fragmentation only partially. In particular, entity-perturbed negatives reduce fragmentation from 4.202 to 2.147, but this remains far worse than the AU baseline, which achieves fragmentation 0.014. Thus, the sharing-regime disadvantage of contrastive training cannot be explained solely as an artifact of naive in-batch negative sampling.

Table 11: Alternative negative sampling strategies on the SNLI entailment sharing subset. Even more informative negatives remain far more fragmented than AU training.

Negative Strategy	Fragmentation	Relative to AU
Random in-batch	4.202	300× worse
Filtered in-batch	4.408	315× worse
Entity-perturbed	2.147	153× worse
AU (no negatives)	0.014	baseline

This result strengthens our main interpretation:

in relation-sharing settings, the mismatch arises not merely because one particular negative sampler is poorly chosen, but because explicit repulsion remains structurally misaligned with the data even under more informative negative construction rules.

G Relation Space Visualization

Figure 3 shows UMAP projections of test-set relation embeddings under matched and mismatched objectives.

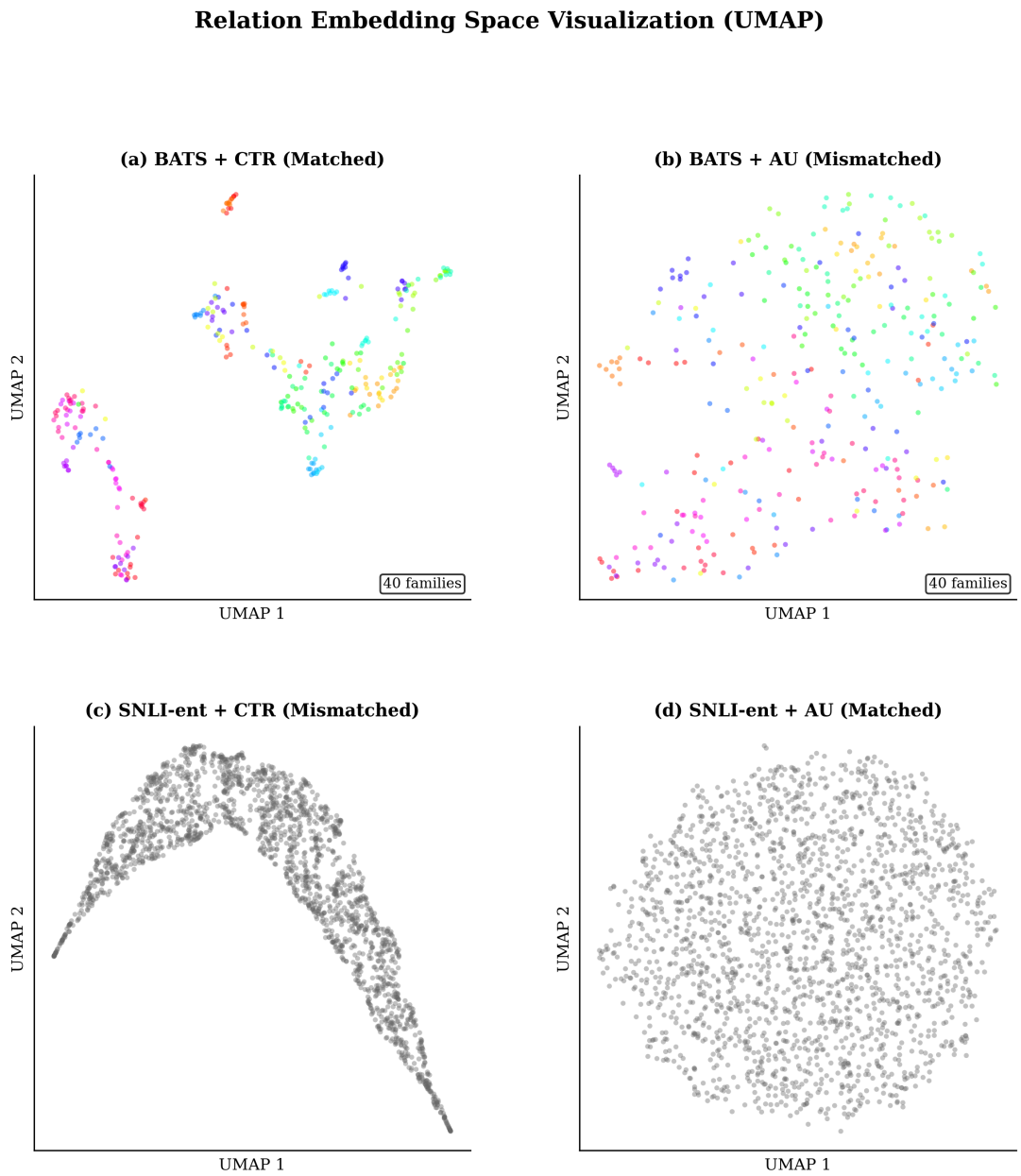

Figure 3: UMAP projections of relation embeddings. Top row: BATS (discriminating). Bottom row: SNLI-Entailment (sharing). Left column: matched objective. Right column: mismatched objective. In the discriminating regime, matched contrastive training produces well-separated clusters, while mismatched AU collapses structure. In the sharing regime, matched AU preserves a compact manifold, while mismatched Ctr fragments the space.

On BATS (discriminating), matched contrastive training produces well-separated family clusters, while mismatched AU training collapses all families into an undifferentiated mass. On SNLI-Entailment (sharing), matched AU training preserves a compact shared manifold, while mismatched contrastive training fragments the space into many spurious clusters driven by false-negative repulsion.

H Training Dynamics

Figure 4 shows training loss and validation metric curves across epochs.

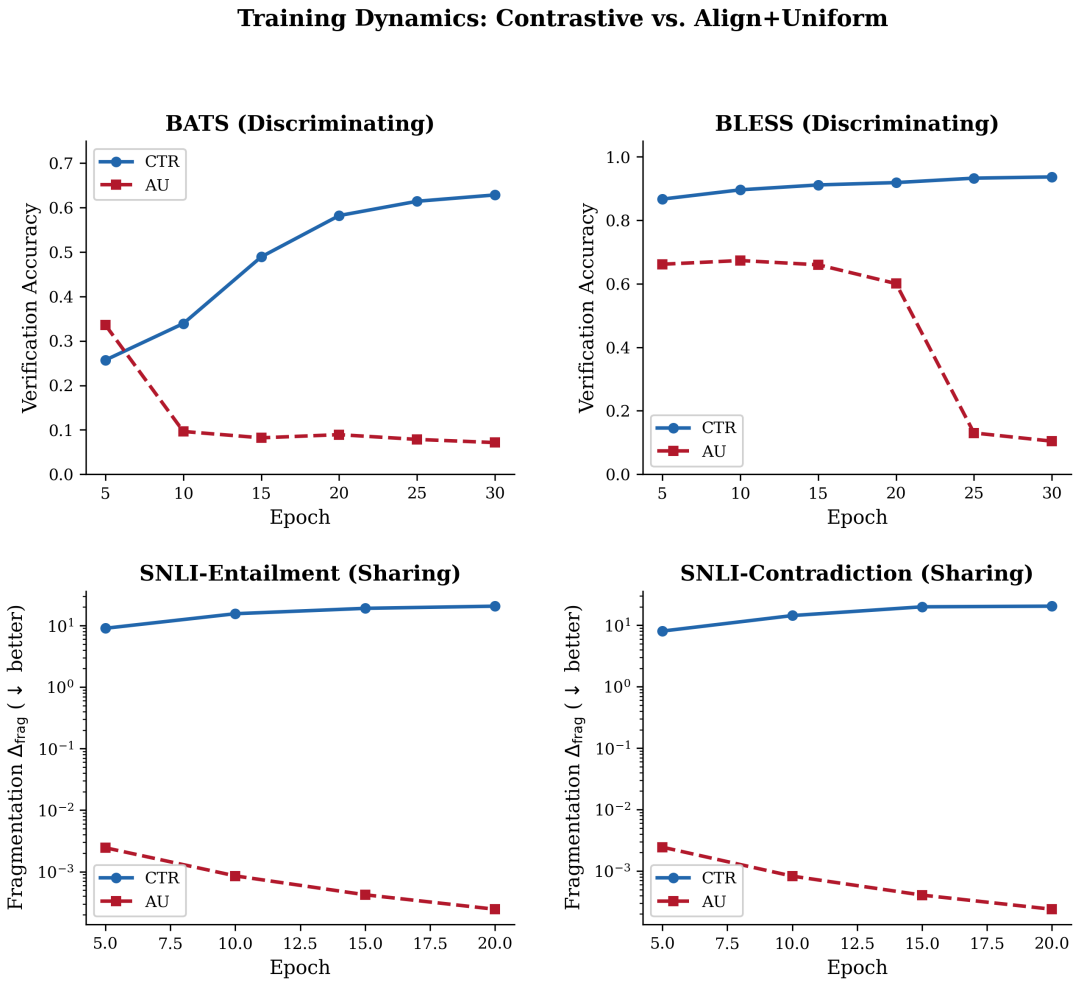

Figure 4: Training dynamics for all main experiments. Left two panels: discriminating datasets (BATS, BLESS) with training loss and verification accuracy. Right two panels: sharing datasets (SNLI-Ent, SNLI-Con) with training loss and fragmentation. Matched objectives converge to better final values; mismatched objectives either plateau early or diverge.

Several patterns emerge from the training curves:

- On discriminating datasets, mismatched AU training converges quickly in terms of loss but produces near-chance verification accuracy, indicating that the loss surface has local optima that do not correspond to useful relational structure.
- On sharing datasets, mismatched contrastive training achieves low loss values while producing high fragmentation, showing that optimizing the contrastive objective actively works against the desired geometry.
- Matched objectives exhibit more stable convergence, with validation metrics improving consistently through training.

I Sharing-Regime Evaluation Note

On SNLI entailment and contradiction, verification accuracy, Recall@1, and MRR all saturate at 1.0 under *both* matched and mismatched objectives. This occurs because the positive nearest-neighbor pairing used during training creates strong local alignment regardless of global geometry. As a result, task-level metrics cannot distinguish between a compact shared manifold (AU) and a severely fragmented space (Ctr) in these settings.

This is not a limitation of our claim. Rather, it demonstrates that standard downstream metrics may fail to reveal objective mismatch when the relational structure is simple. Geometric diagnostics such as fragmentation are necessary to assess the quality of the learned relation space in sharing regimes.