Negatives Are Hypotheses: Auditing Objective Compatibility in Relational Representation Learning

Anonymous ACL submission

Abstract

Contrastive relation learning silently assumes that every sampled negative is a true relation-level incompatibility. In relational data this assumption is often the experiment, not a fact: an apparent negative can be a useful cross-family contrast or a compatible false negative. We turn this hidden precondition into *negative validity* and propose negative-validity auditing as a reporting protocol and pre-training diagnostic, rather than a standalone objective-selection guarantee. The audit uses frozen relation features to approximate validity pressure, recommends or blends objective-pressure families when confidence is sufficient, and abstains when the evidence is ambiguous. Across lexical relation benchmarks, Stanford Natural Language Inference (SNLI) stress tests, controlled false-negative sweeps, and FewRel/Revised DocRED relation-extraction diagnostics, including a held-out Re-DocRED deployment-style check, fixed objectives fail at opposite validity endpoints: repulsion separates valid relation families but fragments relation-sharing regimes. False-negative-aware contrastive variants mitigate but do not remove this regime dependence. The takeaway is protocol-level: before using in-batch negatives as labels, relational representation learning should audit whether those negatives are valid for the target relation regime.

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). Recent systems have made substantial progress on this problem by scaling encoders (Ushio et al., 2021b, 2025), training contrastive relation extractors (Oord et al., 2018; Chen et al., 2020), and adapting negative-sampling heuristics that mitigate false negatives (Chuang et al., 2020; Robinson et al., 2021). The dominant evaluation practice in this line of work is to compare losses on aggregate benchmarks and report a single "best" objective per dataset.

Despite this progress, current practice provides an incomplete account of *when* contrastive pressure helps relational learning. Different self-supervised objectives impose different geometry pressures (Wang and Isola, 2020; Grill et al., 2020; Chen and He, 2021; Zbontar et al., 2021), but benchmark comparisons often treat "negatives" as a single uniform signal. In relational data, the same in-batch "other" pair can be a genuine between-family negative or a within-family false negative depending on the dataset regime. *Negatives are not labels; they are hypotheses*, and the "best-loss-per-benchmark" framing leaves untested whether those hypotheses are correct before the loss is allowed to repel.

This is a protocol paper rather than a new-loss paper. The central claim is that negative-based relation learning has a checkable precondition: before comparing negative-based objectives, a benchmark should report whether sampled negatives are valid for the relation regime. The supporting logic is: repulsion is useful when apparent negatives are valid, harmful when they are compatible, observable under controlled validity interventions and natural RE diagnostics, and only actionable when an audit has enough confidence to avoid forced selection.

We organize the paper around the corresponding research questions. **RQ1:** does the appropriate objective pressure change when apparent negatives become false negatives? **RQ2:** does this pressure mismatch appear beyond constructed stress tests, including natural relation-extraction data? **RQ3:** can a label-sparse audit identify when to use repulsion, when to use a variance-preserving fallback, and when to abstain? We answer these questions by formalizing *negative validity* q, showing that the repulsive gradient of an InfoNCE (information noise-contrastive estimation)-style loss decomposes into between-relation pressure and within-relation dis-

persion (Proposition 1), and evaluating an audit based on frozen relation features. We use shorthand mainly in tables: **Ctr** denotes contrastive training, **AC** denotes the alignment–compactness diagnostic endpoint, **FN** denotes false negative, **RE** denotes relation extraction, and **VICReg** denotes variance–invariance–covariance regularization.

2 Related Works

Relation representation research has long used analogy and lexical-relation benchmarks to test whether vector spaces encode relational transformations rather than only entity similarity (Gentner, 1988; Gentner and Markman, 1997; Turney, 2005, 2006; Mikolov et al., 2013; Gladkova et al., 2016; Jurgens et al., 2012; Rogers et al., 2017; Ushio et al., 2021b; Wijesiriwardene et al., 2023; Ye et al., 2024; Ushio et al., 2021a, 2025). This line motivates our focus on relation-level compatibility: an apparent negative is useful only when it separates incompatible transformations, not merely different surface instances.

Contrastive representation learning supplies a powerful separation pressure, but its standard in-batch form treats sampled negatives as valid incompatibilities (Oord et al., 2018; Chen et al., 2020; Saunshi et al., 2022; Tosh et al., 2021; Wang and Isola, 2020; Grill et al., 2020; Chen and He, 2021; Zbontar et al., 2021; Assran et al., 2023). Work on false negatives, hard negatives, and supervised contrastive variants shows that this assumption can fail, yet most treatments remain instance-centric rather than asking whether negative repulsion matches the relation regime (Saunshi et al., 2019; Chuang et al., 2020; Robinson et al., 2021; Reimers and Gurevych, 2019; Gao et al., 2021).

Noncontrastive and redundancy-reduction objectives avoid explicit negative repulsion and therefore provide useful endpoints for auditing objective pressure (Grill et al., 2020; Chen and He, 2021; Zbontar et al., 2021; Assran et al., 2023). Our contribution is not a new universal loss, but a diagnostic reframing: before using negatives as labels, relation-learning systems should audit whether those negatives are valid for the target regime.

3 Methodology

3.1 Negative Validity

A relation instance is an ordered pair (x_A, x_B) encoded by a frozen backbone into states (z_A, z_B); a relation extractor ϕ_ψ produces a relation embedding $r_{A\to B} = \phi_\psi(z_A, z_B) \in \mathbb{R}^k$. Two pairs are *analogically equivalent* when their relation embeddings are close; *relational invariance* maps instances of the same transformation family $\mathcal{T}$ to nearby points. Datasets induce two regimes: a **relation-discriminating** regime, where multiple transformation families coexist and the loss must separate them; and a **relation-sharing** regime, where one transformation pattern dominates and many in-batch "negatives" are in fact compatible relation instances.

3.2 Pipeline and Positives

We freeze the encoder and the relation-extractor architecture, training a separate extractor per objective so that performance and geometry differences are attributable to objective choice. Each instance is an ordered pair $x_i = (a_i, b_i)$; we encode both elements with a frozen backbone (RoBERTa-base in the main grid, BERT-base for robustness), $z_i^{(a)} = f_\theta(a_i)$, $z_i^{(b)} = f_\theta(b_i)$, and form a relation embedding $r_i = \phi_\psi(z_i^{(a)}, z_i^{(b)}) = \mathrm{MLP}([z_i^{(a)}; z_i^{(b)}; z_i^{(b)} - z_i^{(a)}; z_i^{(a)} \odot z_i^{(b)}])$.

Positive construction. For BATS and BLESS, positives are same-family pairs ($y_i = y_j$) and all other families are candidate negatives. For the SNLI-derived high-compatibility stress test we treat each subset (Ent, Con) as a single nominal relation family: this is a deliberate stress test that maximises the rate of compatible apparent negatives *without* claiming SNLI is a homogeneous micro-relation (Bowman et al., 2015; Gao et al., 2021). Positives within the subset are formed by frozen-feature nearest-neighbour pairing; the alignment-dominant objective treats other batch elements only through the global compactness term, never as instance-specific negatives.

3.3 Objectives as Pressure Families

We distinguish objective *roles* before introducing equations. Contrastive pressure is the high-validity separation family. False-negative-aware contrastive losses are mitigation baselines that soften but do not remove repulsion. Audit-hybrid is the mixed-validity family. Barlow/VICReg are the deployed low-validity variance-preserving fallback. The compactness endpoint is only a stress-test diagnostic for under/over-separation; it is not the deployed low-validity policy.

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

$$s_i^+ = \mathrm{sim}(r_i, r_i^+), \qquad s_{ij} = \mathrm{sim}(r_i, r_j), \quad (1)$$

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

$$Z_i = \exp(s_i^+/\tau) + \sum_{j\in\mathcal{N}(i)} \exp(s_{ij}/\tau), \quad (2)$$

and optimize

$$\mathcal{L}_{\mathrm{ctr}}^{(i)} = -\log \frac{\exp(s_i^+/\tau)}{Z_i}, \quad (3)$$

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.

Compactness diagnostic endpoint. For stress testing relation-sharing regimes, we also evaluate an alignment-dominant compactness endpoint without instance-specific negatives:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{B}\sum_{i=1}^{B} \|r_i - r_i^+\|_2^2, \quad (4)$$

$$\mathcal{L}_{\mathrm{comp}} = -\log U, \qquad U = \frac{1}{B(B-1)}\sum_{i\neq j} \exp(-t\|r_i - r_j\|_2^2). \quad (5)$$

and

$$\mathcal{L}_{\mathrm{AC}} = \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}}. \quad (6)$$

The historical implementation name for this loss is `align_uniform`, but the sign above makes the second term a weak compactness pressure rather than Wang–Isola uniformity. We therefore treat it as an appendix stress-test endpoint, abbreviated AC in tables, not as a deployed low-validity policy.

Audit-hybrid objective. The adaptive method uses the audit as a continuous proxy and a policy signal rather than only as a binary selector. For each candidate negative pair (i, j), let $v_{ij} \in [0, 1]$ serve as a proxy for the probability that r_j is a valid negative for r_i under frozen relation features. We convert validity into repulsion weight

$$w_{ij} = \sigma\left((v_{ij} - \gamma)/T\right), \quad (7)$$

and replace the contrastive denominator with

$$Z_i^{\mathrm{audit}} = \exp(s_i^+/\tau) + \sum_{j\in\mathcal{N}(i)} w_{ij}\exp(s_{ij}/\tau). \quad (8)$$

The pair-weighted repulsion component is

$$\mathcal{L}_{\mathrm{audit}}^{(i)} = -\log \frac{\exp(s_i^+/\tau)}{Z_i^{\mathrm{audit}}} + \lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}} + \lambda_{\mathrm{cov}}\mathcal{L}_{\mathrm{cov}}. \quad (9)$$

High-validity negatives therefore receive ordinary repulsion, while compatible apparent negatives contribute little to the denominator. The variance and covariance terms prevent the low-validity endpoint from degenerating into a pure compactness objective. The pair-weighted component alone is not sufficient in the fully low-validity endpoint, because residual repulsion can still fragment compatible relations. In the mixed-regime experiments, the final Audit-hybrid objective therefore additionally uses a dataset-level validity prior $\bar{v}$ to blend repulsive and non-repulsive pressure:

$$\mathcal{L}_{\mathrm{hybrid}} = \bar{v}\,\mathcal{L}_{\mathrm{audit}} + (1-\bar{v})\mathcal{L}_{\mathrm{AC/VICReg}}. \quad (10)$$

At the fully low-validity endpoint, the audit policy selects Barlow/VICReg rather than forcing residual repulsion.

Why validity must control repulsion. We use the formal analysis to state the assumption that contrastive training usually leaves implicit. Both claims are expectations under a Bernoulli-validity mixture model: each sampled negative $(i, j) \in \mathcal{N}(i)$ is valid with probability $q_{ij} \in [0, 1]$ and false with probability $1 - q_{ij}$, and q_{ij} depends only on the underlying transformation families. Proposition 1 is a definition-level decomposition. The calibration-aware bound below is a directional sufficient condition for audit weighting; we do not use it as an operational empirical guarantee because its constants require pair-level gate, validity, and repulsion-mass logs that are not available for every dataset in the current grid. Proof sketches are in Appendix B.

Proposition 1 (Negative-validity decomposition) *For a contrastive loss, let r_i be the anchor, r_i^+ its positive, and $\mathcal{N}(i)$ its negatives. The expected repulsive gradient is*

$$\mathbb{E}[\nabla_{r_i}\mathcal{L}_{\mathrm{rep}}] = q\,G_{\mathrm{between}}(r_i) + (1-q)\,G_{\mathrm{within}}(r_i), \quad (11)$$

where G_{between} is the valid between-relation separation gradient and G_{within} is the invalid within-relation dispersion gradient, and $q = \mathbb{E}_{j\sim\mathcal{N}(i)}[q_{ij}]$ is the mean negative validity over the in-batch negative pool. The within-relation dispersion coefficient is therefore exactly $1 - q$: whenever $q < 1$, pure repulsion introduces a potentially harmful within-relation dispersion component whose weight scales with $1 - q$, and the regime-level net effect is what the empirical regret analysis in §4 quantifies.

Calibration-aware implication. The nontrivial question is when an audit reduces the harmful part of the repulsive field. Let $V_{ij} \in \{0, 1\}$ indicate whether an apparent negative is a valid relation-level negative, S_{ij} be the frozen-feature audit score, $w(S_{ij}) \in [0, 1]$ be the gate applied to the repulsive term, and g_{ij} be the unweighted repulsive gradient contribution. Define invalid-repulsion mass as

$$M_{\text{in}}(w) = \mathbb{E}\big[(1 - V_{ij})w(S_{ij})\|g_{ij}\|\big],$$

with $M_{\text{in}}(1)$ denoting uniform contrastive repulsion. Suppose that outside an ambiguity set A, invalid negatives have score leakage at most η; the gate has gradient-weighted calibration error ϵ_g; and the gradient-weighted ambiguity mass is α_g. Then

$$M_{\text{in}}(w) \le (\eta + \epsilon_g + \alpha_g)M_{\text{in}}(1).$$

This implication says that audit weighting reduces within-relation dispersion pressure only when score leakage, calibration error, and ambiguity mass are jointly below one. It does not claim that the audit recovers the true latent q, and the paper's empirical claims do not rely on verifying this bound row-by-row. Appendix B gives the proof sketch and explains why we treat the result as directional rather than operational.

Implication. The bound motivates abstention. When the audit score distributions overlap or calibration is weak, the term $\eta + \epsilon_g + \alpha_g$ can be large, so the diagnostic should not confidently select a repulsive objective. This is why the deployed policy uses confidence, reports coverage, and falls back to variance-preserving objectives under conflict.

Objective-geometry taxonomy. The final method is not a binary contrastive-vs.-AC choice. We group objectives by the geometric pressure they impose. Repulsive losses are useful when negatives are valid; false-negative-aware variants soften but do not remove repulsion; AC is a compactness stress-test endpoint; Barlow/VICReg are the deployed variance-preserving low-validity family; and Audit-hybrid/Audit-select apply conditional pressure with abstention.

Appendix Table 26 separates standard objective names from paper-defined diagnostics such as AC, Audit-hybrid pressure, and the audit guard.

3.4 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:

$$\begin{aligned}\Delta_{\text{sep}} = \mathbb{E}[d(r_i, r_j) \mid y_i \neq y_j] \\ - \mathbb{E}[d(r_i, r_j) \mid y_i = y_j]\,.\end{aligned} \tag{12}$$

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:

$$\Delta_{\text{frag}} = \mathbb{E}\big[\|r_i - r_j\|_2^2 \mid x_i, x_j \in D_{\text{share}}\big]\,. \tag{13}$$

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

3.5 Datasets

Our goal is to test a more specific claim: objective choice in relational representation learning should be controlled by whether sampled negatives are valid relation negatives under the data regime. We therefore freeze the encoder, keep the relation-extractor architecture fixed in the main experiments, and train separate extractors while varying only the objective geometry.

We use four datasets or derived subsets in the controlled main grid. The Bigger Analogy Test Set (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. Baroni and Lenci's Evaluation of Semantic Spaces (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 Stanford Natural Language Inference (SNLI), treating each subset as a single

shared inference family. This differs from prior natural-language-inference (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. To reduce reliance on this constructed stress test, we also add natural relation-extraction diagnostics on FewRel `val_wiki` (Han et al., 2018) and Revised DocRED (Re-DocRED) (Tan et al., 2022). FewRel supplies 16 held-out relation types and 1,600 entity-pair instances, while Re-DocRED supplies a recent document-level RE setting explicitly revised to address false negatives. These datasets are used only for the negative-validity and objective-regime diagnostics in §4, not for state-of-the-art relation extraction claims.

3.6 Negative-Validity Audit Protocol

The audit runs before objective training on frozen relation features. It uses two primary signals and two supporting checks. The primary signals are cluster stability and the topic-cluster guard: the former detects whether relation families form stable geometry, while the latter prevents topical clusters from being mistaken for relation families. The supporting checks are local-neighbourhood consistency and apparent-negative compatibility. The ablation in Appendix Table 13 confirms this division: removing cluster stability or the guard changes the policy regret substantially, whereas removing the two supporting checks has little effect on the studied grid.

The output is a proxy validity pressure with confidence, not a recovered latent q. High-confidence high-validity regimes receive Ctr, mixed regimes receive Audit-hybrid, low-validity regimes receive Barlow/VICReg, and conflicting signals abstain to the conservative fallback. Thresholds are fixed before objective training; appendix tables report the full process definitions, threshold sensitivity, and per-process ablations.

Deployment recipe. The audit is used as a pre-training gate, not as an always-on loss. Given frozen relation features, compute the four audit signals and their confidence. If confidence is high and proxy validity is high, train with Ctr; if confidence is high and proxy validity is low, use VICReg or Barlow Twins; if validity is mixed, use Audit-hybrid; and if the signals conflict or confidence is low, abstain and use the conservative variance-preserving fallback. Abstention is therefore part of the method and is reported as a coverage cost rather than hidden as success.

3.7 Experimental Setup and Evaluation

All objectives use the same relation features, batch construction, optimization budget, and evaluation splits within each regime. We report task metrics when the regime has a functional label (e.g., verification accuracy, retrieval, or joint SNLI accuracy) and geometry metrics when the claim concerns objective-induced structure (e.g., separation, fragmentation, NMI/AMI, and calibration). The primary metric is declared per regime before comparing objectives, and appendix tables give hyperparameters, variance checks, sensitivity analyses, and reproducibility details.

Reproducibility contract. Every main diagnostic is tied to a named script, output directory, or claim-to-artifact entry in Appendix A, and the audit is run before objective training or objective-effect scoring. Labels are used for evaluation and oracle construction, not for the label-sparse audit decision. This separation is central to the soundness claim: the paper evaluates whether a pre-training validity check predicts objective compatibility, not whether labels can be used after the fact to choose a loss.

4 Results

Table 1 gives the main empirical picture. We do not frame the results as a loss leaderboard: the same objective pressure changes value when the validity of sampled negatives changes. The table therefore separates the two numerical claims that matter most: a controlled intervention where the false-negative rate is known, and natural relation-extraction regimes where validity must be audited from frozen features. Appendix Table 2 maps these numbers to the claim structure and boundary conditions.

4.1 RQ1: Endpoint Reversal Across Validity Regimes

The controlled false-negative-rate sweep is the direct test of the claim. As compatible apparent negatives increase, the preferred pressure shifts from separation/hybrid at low false-negative mass to variance-preserving pressure at the fully shared endpoint (Figure 1). Panel A of Table 1 shows the same intervention numerically: Hybrid matches or

Table 1: Main numerical evidence for negative-validity dependence. Panel A holds the data-generating process fixed while varying the false-negative rate; utility U is higher better. Panel B uses natural relation-extraction regimes; regret is oracle utility minus obtained utility (lower is better), and AUROC measures the pair-level validity audit.

A. Controlled false-negative intervention ($U \uparrow$)

False-negative rate	Ctr	Barlow	VICReg	Hybrid	Winner
0%	0.872	0.832	0.823	**0.873**	Hybrid / Ctr
25%	0.865	0.835	0.823	**0.962**	Hybrid
50%	0.870	0.844	0.823	**0.971**	Hybrid
75%	0.881	0.856	0.827	**0.970**	Hybrid
100%	0.626	**0.709**	0.678	0.329	Barlow

B. Natural relation regimes (regret↓, audit AUROC↑)

Dataset	Validity proxy	Ctr regret	Debiased-Ctr regret	Audit-select regret	Audit AUROC
FewRel	relation type	0.370	0.063	**0.029**	0.957
Re-DocRED	revised doc RE	0.048	0.044	**0.005**	0.660

dominates through 75% false negatives, but Barlow becomes the winner when every apparent negative is compatible.

This is the strongest evidence because the data-generating knob is known. At 0% false negatives, apparent negatives are valid relation-level contrasts and repulsive pressure is useful. At intermediate rates, neither pure repulsion nor pure compactness is aligned with the regime, so the hybrid family is preferred. At 100%, apparent negatives are compatible by construction, and the variance-preserving endpoint becomes the safer pressure family. This pattern is not a claim that the audit has solved objective selection; it is a controlled demonstration that objective compatibility is conditional on negative validity.

The failure mode is not symmetric. In relation-discriminating datasets, the compactness diagnostic endpoint under-separates: on BATS, separation margin drops from 2.0275 to 0.0004 and NMI from 0.7219 to 0.2580. In sharing regimes, Ctr fails by fragmentation: SNLI entailment/contradiction rise from 0.0002 under the compactness diagnostic endpoint to 20.5866/20.3312 under Ctr. Normalized and cosine diagnostics preserve the qualitative gap (Appendix K).

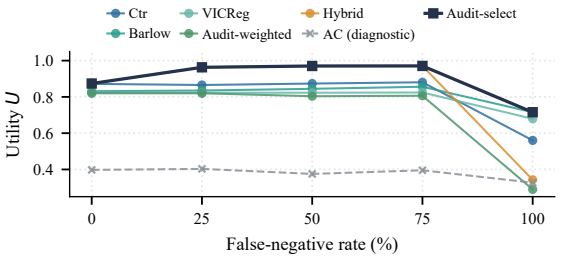

Figure 1: Controlled false-negative-rate sweep. As compatible apparent negatives increase, fixed repulsive objectives fail at the shared endpoint, while the audit policy shifts from separation/hybrid pressure to variance-preserving pressure.

4.2 RQ2: Natural Diagnostics and Baselines

FewRel and Re-DocRED test whether the low-validity failure is specific to SNLI. FewRel gives a clean relation-type proxy: same-relation pairs are compatible apparent negatives and cross-relation pairs are valid negatives. The audit reaches AUROC/AUPRC 0.957/0.953 over 20k pairs; fixed Ctr reaches worst regret 0.370, while Audit-select stays at 0.029. Re-DocRED is noisier (0.660/0.619 over 30k pairs), but Audit-select still abstains into a variance-preserving fallback with 0.005 worst regret, compared with 0.048 for fixed Ctr. Panel B of Table 1 reports these natural regime numbers next to the strongest contrastive false-negative baseline.

The two natural diagnostics play different roles. FewRel is the cleaner construct-validation case: relation types provide a high-precision proxy for compatible versus incompatible pairs, and the audit score tracks that proxy well. Re-DocRED is the harder deployment-style case: document-level relation annotations are noisier and false negatives are one of the reasons the dataset was revised. The audit does not pretend to resolve that uncertainty; it mostly withholds aggressive repulsion and falls back to variance preservation. This is weaker than a full downstream relation-extraction result, but it is the relevant diagnostic evidence for the paper's claim: the low-validity problem is not an artifact of SNLI construction.

Appendix Table 6 adds a downstream sanity check on the same objective-effect features: Re-DocRED supports the conservative fallback (0.306 utility vs. 0.283 for Ctr), while FewRel shows that simple classifier/retrieval probes can over-prefer compactness. We therefore use downstream probes as guardrails, not as the audit definition.

False-negative-aware contrastive variants are im-

portant baselines but do not remove the regime dependence. They reduce SNLI-slice fragmentation (1.214 to 1.175/1.045) and Debiased-Ctr improves FewRel/Re-DocRED worst regret (0.370 to 0.063; 0.048 to 0.044), yet the mixed grid still switches winners: Debiased-Ctr at 0% false negatives, Hybrid at 25–75%, and Barlow at 100%. Thus false-negative-aware repulsion is best read as a mitigation baseline, not as a replacement for validity auditing. It changes the strength and sampling profile of repulsion, but it still assumes that some repulsion is the right pressure family.

4.3 RQ3: Reliability, Sensitivity, and Regret

Audit reliability is explicitly regime-dependent. The proxy is strong on controlled mixed false-negative regimes (AUROC up to 0.944), usable on BATS (0.748), weak on BLESS, and near-random on SNLI (0.504–0.539). Hence the deployed policy uses confidence and abstention rather than hard thresholding. Appendix E reports the sensitivity checks: utility weights, Pareto lens, seeds, thresholds, construction choices, normalized geometry, component ablations, validation selection, and leave-one-dataset-out calibration. Regret is therefore a compact summary under stated utilities, not the sole evidence for the claim.

This regime dependence is part of the result rather than a side note. On SNLI, near-random pair-level AUROC means the audit cannot identify which apparent negative pairs are valid from frozen features alone. The appropriate behavior is therefore abstention, not forced selection. Conversely, when cluster stability and the topic guard agree, the audit has enough evidence to recommend a pressure family before objective training. The practical output is a tuple: decision, confidence, and coverage. Reporting all three prevents a label-sparse diagnostic from being mistaken for an oracle selector.

5 Discussion

What the evidence establishes. The evidence supports a protocol-level claim: negative validity is a hidden precondition of negative-based relation learning, and objective pressure should be reported relative to that precondition. Across controlled false-negative sweeps, lexical relation benchmarks, and natural relation-extraction diagnostics, we observe the same endpoint reversal: repulsion separates high-validity relation families but fragments relation-sharing regimes, while variance-preserving pressure avoids fragmentation but can under-separate genuinely incompatible families.

This answers the research questions in the conservative form supported by the experiments. RQ1 is answered by intervention: when false-negative mass is changed directly, the winning pressure family changes. RQ2 is answered by convergence across evidence sources: the same risk appears in lexical benchmarks, constructed SNLI sharing slices, FewRel, and Re-DocRED, although the natural relation-extraction results remain diagnostic rather than full downstream claims. RQ3 is answered by coverage: the audit is useful only when it can act with confidence, and its abstention behavior is a reported cost rather than a hidden failure.

Why this matters for evaluation. The implication is not that one objective family is universally preferable. The important change is the unit of reporting. A relation-learning result should not state only the dataset, objective, and score; it should also state whether the sampled negatives are valid relation-level incompatibilities for that regime. Without that condition, a leaderboard can reward the wrong mechanism: strong repulsion when the benchmark mostly supplies valid cross-family contrasts, or compactness when the benchmark mostly contains compatible sharing relations.

This changes how objective comparisons should be interpreted. A higher score from contrastive training on a high-validity benchmark is evidence that separation pressure helped; it is not evidence that the same pressure is safe in a sharing regime. Likewise, a compact representation on a low-validity slice is not evidence that compactness is a universal objective; it may simply avoid an invalid repulsion signal. Negative-validity reporting separates these two mechanisms before the paper reaches a single aggregate score.

How to read the audit. The audit is therefore not a solved label-free objective selector. Its operational role is narrower: it reports whether a benchmark supplies valid negative evidence for repulsion, recommends pressure only when confidence is sufficient, and records abstention when pair-level validity is underidentified. This turns negative validity from an implicit assumption into a reportable experimental condition rather than an after-the-fact explanation for a loss comparison.

What the protocol rules out. The evidence rules out two common shortcuts. The first is treating false-negative-aware contrastive learning as a complete solution: those methods reduce invalid repulsion, but they still assume that repulsion should be present. The second is treating noncontrastive compactness as the low-validity answer: compactness can avoid fragmentation, but it can also erase necessary between-family separation. The audit is useful because it asks the prior question that both shortcuts skip: whether apparent negatives are valid enough to justify any separation pressure.

Practical use. Negative-validity reporting should accompany negative-based relation-learning benchmarks in the same way that split construction, leakage, and label balance are reported. Tables should state not only the objective and score, but also whether sampled negatives are valid for the regime, what the audit confidence is, and what coverage is lost through abstention. Re-DocRED illustrates the intended use case: audit frozen relation-pair features, choose or withhold pressure before endpoint scoring, and then use labels only to evaluate risk.

For practitioners, the recommendation is deliberately simple. If the audit finds stable relation clusters and low topical confounding, use separation pressure. If it finds mixed validity, use a blended objective. If it finds low validity or low confidence, use a variance-preserving fallback and report the abstention rate. This recipe is less ambitious than a universal selector, but it is more actionable than a post-hoc explanation that a contrastive objective fragmented the representation.

What remains open. False-negative-aware contrastive methods remain useful mitigations, but they still commit to repulsive geometry. The more basic decision is whether repulsion is the right pressure family at all. Human validity annotation and full fine-tuned relation-extraction benchmarks would strengthen construct validation; the current paper establishes the missing precondition and a reusable diagnostic protocol for reporting it.

The strongest future test is therefore not another aggregate benchmark table. It is a blind deployment study: run the audit before training on a held-out relation regime, commit to the pressure family or abstention decision, and then evaluate downstream relation classification or retrieval. The present paper does not claim to complete that program; it establishes why such a protocol is needed and supplies the first controlled evidence that the hidden precondition matters.

6 Conclusion

The failure mode studied here is not the use of negatives but the unreported assumption that every sampled negative is valid. Negative validity makes that assumption explicit. Across the diagnostics, the same pattern recurs: fixed objectives fail at opposite validity endpoints; false-negative-aware contrastive variants mitigate but do not remove the regime dependence; natural RE diagnostics reproduce the low-validity failure outside SNLI; the held-out Re-DocRED case shows that the audit can make a conservative pre-training decision before endpoint scoring; and abstention is necessary because pair-level reliability is regime-dependent. The conclusion is a reporting standard: relational representation learning should audit and disclose negative validity before choosing or comparing negative-based objectives. Broader human validity annotation and full downstream RE benchmarks would strengthen construct validation, but the current evidence is sufficient to expose the missing precondition and to make negative validity a first-class evaluation axis.

Operationally, this changes the expected shape of a relation-learning result. A strong paper should report not only its encoder, objective, and aggregate score, but also the validity regime under which its negatives were sampled, the geometry pressure that the objective imposes, and the confidence or abstention behavior of any label-sparse diagnostic. Under that standard, objective comparisons become more interpretable: a contrastive gain is evidence for useful relation-family separation only when the negatives are valid, and a variance-preserving gain is evidence for safer training only when repulsion is not justified. That is the practical contribution of the paper.

7 Limitations

We state the boundary conditions that delimit the protocol.

The audit is a diagnostic, not a solved selector. The audit components are intentionally interpretable, but the topic guard and thresholds are load-bearing. They were fixed before objective training in our experiments, yet broader external validation is still needed before treating the audit as a general label-free selector. The current evidence supports diagnostic use with abstention.

Scalar summaries depend on the chosen lens. The controlled-grid regret table uses row-local declared utilities, and some rows combine task and geometry terms. We therefore do not present the scalar regret table as a benchmark-independent objective-selection guarantee. Its role is secondary: it summarizes endpoint failures already visible in the Pareto-style geometry/task trade-offs. Appendix Table 7 consolidates the sensitivity evidence instead of scattering the interpretation across many separate appendix tables. The stronger claim is conditional and diagnostic: fixed objectives fail in different regimes, while the audit policy either selects compatible pressure or abstains when the audit is unreliable.

Variability reporting is descriptive. We report multi-seed summaries where existing runs support them, especially the controlled false-negative sweep and audit-component ablations. We do not claim formal statistical significance for every row in Appendix Figure 4: some rows are deterministic diagnostic summaries or single natural-dataset case studies, and the current artifacts do not contain cluster-bootstrap samples for every utility. The appropriate reading is therefore robustness-by-sensitivity, not hypothesis testing.

The calibration bound is not an operational guarantee. The calibration-aware implication identifies the quantities that would make audit-gated repulsion provably safer: score leakage, gradient-weighted calibration error, and ambiguity mass. Our current artifacts report aggregate AUROC/AUPRC/ECE and objective-regret diagnostics, but they do not log the pair-level gate, validity, and repulsion-mass terms needed to verify the bound on every dataset. We therefore treat the bound as a guide for abstention and for future audit logging, not as a measured theorem supporting the empirical claims.

SNLI is a stress test, not a micro-relation. The SNLI-derived high-compatibility regime is deliberately constructed to maximise the rate of compatible apparent negatives. SNLI entailment and contradiction contain latent substructure, including lexical entailment, hypernymy, quantifier weakening, negation, and world-knowledge inferences. We do not claim that either subset is a homogeneous relation; we use it to expose what happens to relation geometry when negative validity is low. The medium-confidence abstention behaviour of the audit on these subsets is the conservative response intended by the protocol. The FewRel and Re-DocRED diagnostics in Appendix Table 4 address the complementary concern that the low-validity evidence should not rely only on SNLI: they use held-out or revised natural relation-extraction data, while still remaining diagnostics rather than full downstream relation-extraction benchmarks.

Per-pair audit AUROC is not uniformly high. Appendix Table 23 reports AUROC up to 0.944 on Mixed-75% false-negative and 0.748 on BATS, but only 0.504–0.539 on the SNLI subsets, where validity itself is ambiguous. This is why the deployed policy is Audit-select with calibrated abstention, not a hard threshold on the score (the calibration-aware implication in §3.3).

Frozen-encoder isolation. Our main pipeline freezes the encoder and trains a separate relation extractor per objective to isolate objective effects. We complement this with full-fine-tune and RelBERT robustness checks in Section 4.2, but a full multi-seed fine-tune sweep across more backbones (SBERT, SimCSE) is beyond the scope of the present controlled study.

Automated validity proxy. The pair-level validity benchmark uses family / phenomenon labels as an automated stand-in for human validity annotations. A small human-annotation study, multilingual analogy data, and licensed TACRED-family experiments are *future extensions* of the audit toolkit and are not part of the current results. FewRel and Re-DocRED are included only as natural-regime diagnostics, not as full fine-tuned relation-extraction benchmarks.

Single-step transformations. The formulation represents a relation as a single embedding. Multi-step or hierarchical analogies may require structured relation representations beyond a single vector.

8 Ethical Considerations

This work studies representation-learning objectives on existing public research datasets and does not introduce new human-subject data collection. The main risks are indirect: a regime diagnostic could be over-interpreted as an automatic objective-selection rule, and relation embeddings learned from pretrained language models may inherit biases or artifacts from their source models and

datasets. We therefore present the audit as a first-step diagnostic, report dataset leakage and saturation limitations, and avoid deploying the method in decision-making settings. The experiments use BATS, BLESS, SNLI, SemEval-2012 Task 2, RoBERTa-base, and BERT-base-uncased under their existing research-use terms, FewRel under its public research license, and Re-DocRED under its public MIT license. We release only code, configurations, compact summaries, figures, and small data samples in the supplementary package; large checkpoints and full third-party datasets are not redistributed.

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 (CVPR)*, 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.

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, Lisbon, Portugal. Association for Computational Linguistics.

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 (CVPR)*, pages 15750–15758.

Ching-Yao Chuang, Joshua W. 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. Curran Associates, Inc.

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 Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi 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. Curran Associates, Inc.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

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.

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*, volume 162 of *Proceedings of Machine Learning Research*, 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*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 2021. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, 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, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2025. RelBERT: Embedding relations with language models. *Artificial Intelligence*, 347:104359.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. 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.

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 Stephane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR.

A Appendix Roadmap and Claim-to-Artifact Map

The appendix makes the claim-to-evidence map explicit. It first links each claim to concrete artifacts, then gives theory details, headline controlled and natural diagnostics, audit and sensitivity checks, robustness details, and historical SNLI-only diagnostics. This ordering is meant to make the protocol auditable without treating scalar regret as the sole support for the claims.

Claim-to-artifact map. Table 2 gives the compact map. The remaining appendix expands each part of that map. **High-validity regimes need separation pressure**: BATS/BLESS/SemEval endpoint comparisons show that contrastive training improves verification, retrieval, and separation while the compactness diagnostic under-separates. **Low-validity or sharing regimes punish indiscriminate repulsion**: SNLI stress tests and controlled false-negative sweeps show fragmentation or winner reversal as compatible negatives increase. **False-negative-aware contrastive is a mitigation, not a replacement**: Debiased-Ctr and Hard-Ctr reduce some endpoint damage but do not dominate the mixed grid or remove pressure-family reversal. **Natural datasets reproduce the issue outside SNLI**: FewRel uses human-curated relation labels as a natural validity proxy, while Re-DocRED uses document-level relation-extraction data with known false-negative annotation issues. **The audit should abstain when pair-level validity is underidentified**: SNLI near-random pair-level AUROC is treated as low confidence rather than as selector success. **Scalar utilities are not the only evidence**: Pareto, worst-weight, and sensitivity checks preserve the endpoint-failure pattern under alternate lenses.

B Proof Sketch for Proposition 1 and Calibrated-Gate Bound

Full statement of the calibration-aware implication. Let $p(S) = \Pr(V = 1 \mid S)$ and define

$$\epsilon_g = \frac{\mathbb{E}[(1-V)|w(S) - p(S)|\|g\|]}{M_{\text{in}}(1)},$$
$$\alpha_g = \frac{\mathbb{E}[(1-V)\mathbf{1}\{S \in A\}\|g\|]}{M_{\text{in}}(1)}.$$

Assume $w(S) \in [0,1]$ and that for invalid non-ambiguous pairs, the gate leakage is bounded by $p(S) \le \eta$ on $\{V = 0, S \notin A\}$. Then

$$M_{\text{in}}(w) \le (\eta + \epsilon_g + \alpha_g) M_{\text{in}}(1).$$

Proof sketch of the calibration-aware implication. Decompose the invalid-repulsion mass into ambiguous and non-ambiguous terms:

$$\begin{aligned} M_{\text{in}}(w) = {} & \mathbb{E}[(1-V)w(S)\|g\|\mathbf{1}\{S \notin A\}] \\ & + \mathbb{E}[(1-V)w(S)\|g\|\mathbf{1}\{S \in A\}]. \end{aligned}$$

Since $w(S) \le 1$, the ambiguous term is at most $\alpha_g M_{\text{in}}(1)$. For the non-ambiguous term, add and subtract $p(S)$:

$$\begin{aligned} & \mathbb{E}[(1-V)w(S)\|g\|\mathbf{1}\{S \notin A\}] \\ & \le \mathbb{E}[(1-V)|w(S) - p(S)|\|g\|] \\ & \quad + \mathbb{E}[(1-V)p(S)\|g\|\mathbf{1}\{S \notin A\}]. \end{aligned}$$

The first term is $\epsilon_g M_{\text{in}}(1)$ by definition. The leakage assumption bounds the second by $\eta M_{\text{in}}(1)$. Combining the ambiguous, calibration, and leakage terms yields the stated bound. The result is one-sided: when leakage, calibration error, or ambiguity is large, the implication supports abstention rather than confident objective selection.

Setup. Anchor r_i has positive r_i^+ and negative pool $\mathcal{N}(i)$. Each negative is independently valid with probability $q_{ij} \in [0,1]$. Validity events depend only on the underlying transformation families of r_i and r_j, not on the trained model state. Let $\ell_i = -\log[\exp(s_i^+/\tau)/Z_i]$ with $Z_i = \exp(s_i^+/\tau) + \sum_{j \in \mathcal{N}(i)} \exp(s_{ij}/\tau)$. For a negative pair, write the per-pair gradient contribution as g_{ij}. Decompose g_{ij} into a between-class component g_{ij}^B (active when j is genuinely a different relation) and a within-class component g_{ij}^W (active when j shares the anchor's relation).

Proof sketch of Proposition 1. Conditioning on the validity event $V_{ij} \in \{0,1\}$, $g_{ij} = V_{ij}\, g_{ij}^B + (1 - V_{ij})\, g_{ij}^W$. Taking expectations using $\mathbb{E}[V_{ij}] = q_{ij}$ and linearity over the batch yields

$$\mathbb{E}[\nabla_{r_i} \mathcal{L}_{\text{rep}}] = \frac{1}{|\mathcal{N}(i)|} \sum_j \big(q_{ij}\, g_{ij}^B + (1 - q_{ij})\, g_{ij}^W\big),$$

which equals $q\, G_{\text{between}} + (1-q)\, G_{\text{within}}$ with $q = |\mathcal{N}(i)|^{-1} \sum_j q_{ij}$. The fixed coefficient $1 - q$ on the within-relation dispersion term is the formal content of the claim. □

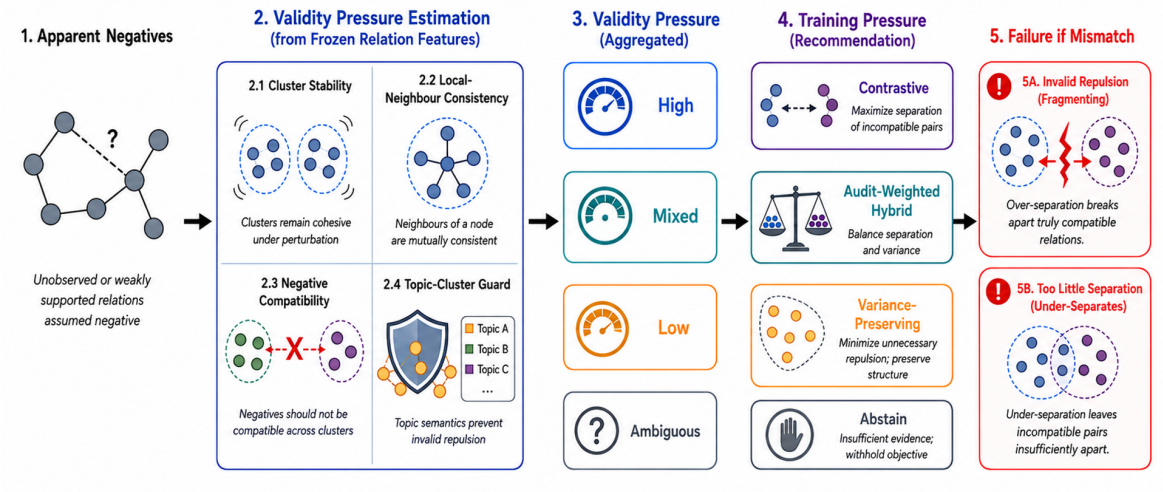

Figure 2: Overview of negative-validity auditing. The audit treats sampled negatives as hypotheses about relation-level incompatibility, estimates validity pressure from frozen relation features, aggregates the evidence into a pressure level, recommends an objective-pressure family, and exposes endpoint failure modes caused by mismatch.

Table 2: Appendix claim-to-evidence map. The main paper presents the numerical headline table; this map states how each result supports the central protocol claim and what boundary condition prevents overclaiming.

Claim link	Evidence	Boundary condition
Mechanism	Proposition 1: repulsion decomposes as $qG_{\text{between}} + (1-q)G_{\text{within}}$.	Directional model; the calibration bound is not used as an empirical guarantee.
Controlled validity intervention	False-negative sweep shifts winners: Hybrid/Ctr at 0%; Hybrid at 25–75%; Barlow at 100%.	Synthetic knob establishes causality for the mechanism, not universal dataset coverage.
Endpoint reversal	BATS Ctr: 0.556 accuracy, 2.0275 separation; SNLI Ctr fragmentation: 20.5866/20.3312 versus compactness near 0.0002.	AC is a diagnostic endpoint; deployed low-validity fallback is variance-preserving.
Natural relation regimes	FewRel/Re-DocRED Ctr-vs-audit regret: 0.370/0.029 and 0.048/0.005; Re-DocRED downstream utility 0.306 for Audit versus 0.283 for Ctr.	Natural RE results are diagnostics and sanity checks, not state-of-the-art RE benchmarks.
Operational reliability	Audit AUROC is 0.944 on mixed FN, 0.748 on BATS, near-random on SNLI, and 0.660 on Re-DocRED with high abstention.	The reported output is decision, confidence, and coverage; the audit is not an oracle selector.

Discussion of the calibration-aware implication. We state the implication in §3.3 as a diagnostic implication rather than as a worst-case theorem. Pair-weighted repulsion replaces g_{ij} with $w_{ij}\, g_{ij}$, where $w_{ij} = \sigma((\hat{v}_{ij} - \gamma)/T)$. Conditioning on V_{ij}, one can sketch a bound of the form $\mathbb{E}[V_{ij}\, w_{ij}] \geq q_{ij} - \epsilon_{ij}$ and $\mathbb{E}[(1 - V_{ij})\, w_{ij}] \leq (1 - q_{ij})\epsilon_{ij} + \sigma'(\cdot)$, which would yield a within-relation dispersion reduction term of the form $\Delta(\epsilon) = \max\{0, 1 - 2\epsilon - (1-q)\sigma'(\cdot)\}$. We do not present this as Proposition 2 because the constants depend on *(i)* the Lipschitz constant of the gate σ, *(ii)* the score separation between valid and invalid pairs that determines the local sigmoid derivative, and *(iii)* the assumption that the validity event V_{ij} is independent of the trained representation given the underlying families. Each of these is non-trivial to verify in practice. The appropriate interpretation is therefore directional: a better-than-chance, calibrated audit reduces within-relation dispersion relative to unweighted repulsion while preserving valid between-relation separation, with the magnitude of reduction shrinking continuously as calibration error grows. We treat the conditions $\epsilon < 1/2$ and the exact form of $\Delta(\epsilon)$ as proof-sketch artefacts and not as deployment guarantees. The directional content is verified empirically in §4 via the audit-component ablation (Appendix Table 13) and the unified regret analysis.

C Headline Controlled and Natural Diagnostic Tables

This block gives the headline appendix tables for the two central empirical claims: objective pressure reverses across controlled negative-validity regimes, and the same issue appears outside SNLI on natural relation-extraction data. Longer audit, sensitivity, robustness, and historical checks follow in separate appendix sections so that the primary evidence is not buried among auxiliary tables.

Table 3: Controlled objective-pressure comparison. The main contrast is between separation pressure in relation-discriminating regimes and fragmentation under relation-sharing stress. The compactness endpoint is a diagnostic; the deployed low-validity fallback is variance preserving.

Dataset	Setting	Verif. Acc	Recall@1	MRR	Sep. Margin	Silhouette	NMI / Frag.
BATS	Ctr separation pressure	0.5563	0.5563	0.6796	2.0275	0.0957	0.7219
BLESS-disjoint	Ctr separation pressure	0.898	–	–	–	–	–
BLESS raw	Ctr separation pressure	0.933	0.932	0.954	1.4598	0.6518	0.6365
SNLI-Entailment	Fragmenting Ctr stress test	sat.	sat.	sat.	–	–	20.5866
SNLI-Contradiction	Fragmenting Ctr stress test	sat.	sat.	sat.	–	–	20.3312

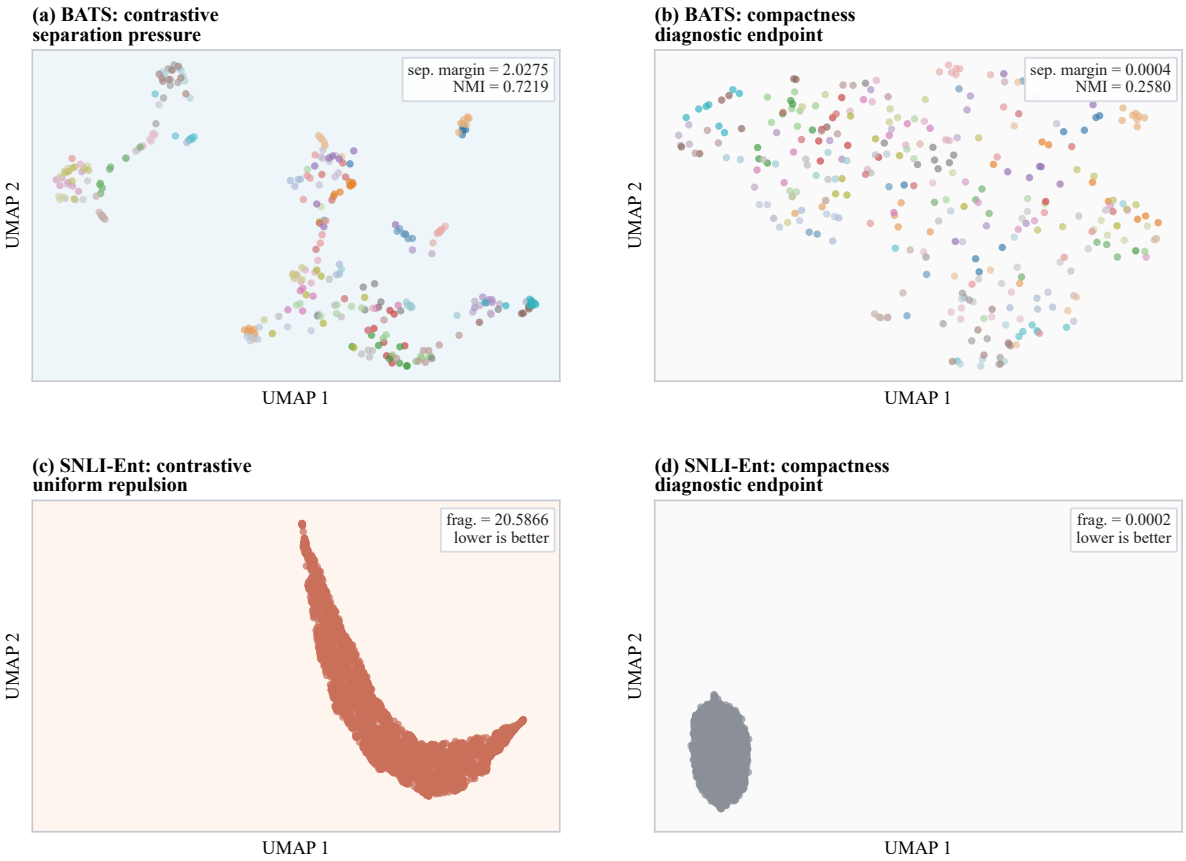

Figure 3: Mechanism view of objective mismatch. The top row uses BATS, a high-validity relation-discriminating regime where apparent negatives are mostly valid cross-family contrasts. There, contrastive pressure separates relation families, whereas the compactness diagnostic endpoint removes needed separation. The bottom row uses SNLI-Ent as a deliberately constructed low-validity sharing stress test: treating compatible instances as uniform negatives makes contrastive repulsion fragment the relation space. The compactness endpoint is included only as a diagnostic control, not as the deployed low-validity policy; deployed fallback uses variance-preserving objectives. Raw fragmentation is scale-sensitive; normalised/cosine diagnostics are in Table 32.

Table 4: Natural relation-regime diagnostics on FewRel and Re-DocRED. Utilities are frozen-feature objective-effect diagnostics, not state-of-the-art relation extraction scores. False-negative-aware repulsion helps in some regimes but does not remove the need for regime-level audit or fallback.

Dataset	Policy	Worst regret	Mean regret	Mean U
FewRel	Hybrid oracle row	0.000	0.000	0.958
FewRel	Audit-select	**0.029**	0.009	0.950
FewRel	Debiased-Ctr	0.063	0.029	0.929
FewRel	Hard-Ctr	0.281	0.107	0.852
FewRel	Ctr	0.370	0.134	0.824
Re-DocRED	VICReg oracle row	0.000	0.000	0.375
Re-DocRED	Audit-select	**0.005**	0.005	0.371
Re-DocRED	Debiased-Ctr	0.044	0.026	0.349
Re-DocRED	Hard-Ctr	0.051	0.031	0.345
Re-DocRED	Ctr	0.048	0.029	0.347

D Natural Relation-Regime Diagnostics: FewRel and Re-DocRED

We use the public FewRel `val_wiki` split, which contains 16 held-out relation types with 700 instances each. The diagnostic in Table 4 samples 100 instances per relation type (1,600 total) and marks the head and tail entities in the sentence before building frozen character n-gram/SVD relation features. A candidate negative is valid if it comes from a different FewRel relation and compatible if it comes from the same relation. This yields a direct, human-curated relation label for pair-level negative validity without using SNLI labels.

The experiment then varies the rate at which an in-batch apparent negative is drawn from the same relation. This controls the false-negative pressure while keeping the underlying relation instances nat-

ural. The objective-effect proxy is the same style as the real mixed-regime diagnostic: each objective induces a deterministic transformation of frozen relation features, and we score relation NMI, nearest-neighbour relation recovery, and within-relation fragmentation. The corresponding reproducibility artifact is `run_fewrel_natural_regime.py`, with outputs in `exp40_fewrel_natural_regime/`.

The FewRel headline numbers are pair-level audit AUROC 0.957 / AUPRC 0.953 over 20,000 candidate negatives; fixed Ctr worst regret 0.370 as same-relation false negatives dominate; and Audit-select worst regret 0.029 by shifting from Ctr to Hybrid/Barlow. These numbers should be interpreted as natural-dataset diagnostics, not as full FewRel classification accuracy.

We additionally include Re-DocRED (Tan et al., 2022), a revised document-level relation extraction dataset designed to address DocRED's false negative annotation problem. We use the public revised dev/test splits from Hugging Face, extract labeled entity-pair relation triples, and balance to 4,050 triples over 45 relation IDs. The candidate-negative definition is the same as above: same-relation triples are compatible apparent negatives and different-relation triples are valid negatives. Because Re-DocRED contexts are document-level and relation aliases are Wikidata-style IDs, the frozen proxy is substantially harder than FewRel: AUROC 0.660 / AUPRC 0.619 over 30,000 candidate negatives. Even under this noisier proxy, Audit-select abstains into a variance-preserving fallback with worst regret 0.005, whereas fixed Ctr has worst regret 0.048 as same-relation false-negative mass increases. The corresponding reproducibility artifact is `run_redocred_natural_regime.py`, with outputs in `exp41_redocred_natural_regime/`.

Held-out deployment-style case study. The Re-DocRED run also serves as a small deployment-style demo. The audit is computed on frozen relation-pair features, relation labels are used only after scoring to evaluate pair-level validity, and the selected pressure is chosen before objective-effect scoring. This is not a full document-level RE benchmark; it tests whether the toolkit gives a conservative answer on a recent dataset with known false-negative annotation issues without using the labels to make the pre-training decision.

Downstream sanity check. We also train a lightweight relation classifier and evaluate nearest-neighbour relation retrieval on the objective-effect features. This check is intentionally not used as main evidence because it is sensitive to compactness: on FewRel, the classifier/retrieval proxy prefers the AC diagnostic endpoint, whereas the regret diagnostic prefers Hybrid/Audit-select. On Re-DocRED, the same check supports the conservative variance-preserving fallback. The result is useful as a boundary condition: downstream validation is necessary, but simple classifier/retrieval probes can favor compactness and should not replace the negative-validity diagnostic.

Table 5: Held-out Re-DocRED audit case study. The audit sees frozen features before objective scoring, uses labels only for post-decision evaluation, and returns a conservative variance-preserving fallback because confidence is low and abstention is high.

Item	Result
Dataset	Re-DocRED dev+test revised
Instances / relations	4,050 triples / 45 relation IDs
Pair-level audit	AUROC 0.660, AUPRC 0.619
Coverage cost	abstention 0.664, decided accuracy 0.676
Pre-training decision	abstain → variance-preserving fallback
Observed endpoint risk	Ctr worst regret 0.048; Audit-select 0.005

Table 6: Natural RE downstream sanity check on objective-effect features. Utility averages relation-classification accuracy and nearest-neighbour relation retrieval. The mixed FewRel outcome is a boundary condition rather than main evidence.

Data	Best	Audit	Ctr	D/H
FewRel	AC .946	.894	.910	.915/.904
ReDocRED	V/B/A .306	.306	.283	.287/.279

E Supporting Audit, Sensitivity, and Auxiliary Tables

The remaining appendix tables report sensitivity analyses, audit diagnostics, robustness checks, mechanism probes, and auxiliary policy summaries. They are secondary to the endpoint-failure and natural-diagnostic evidence above, but they make the protocol reproducible and expose where the audit abstains rather than over-selecting.

E.1 Robustness Map and Audit Controls

The first block is ordered from broad robustness summaries to the audit internals that support them. Table 7 serves as the reader's index: it points from each perturbation axis to the specific diagnostic table where the numerical check is reported. Table 10 then reports the equal-budget objective grid, and Table 11 checks that the mixed-regime conclusion is not an artifact of one scalar utility. The

next audit tables separate three questions that reviewers may otherwise conflate: whether the label-sparse audit abstains in ambiguous regimes (Table 12), whether the frozen-feature score has usable pair-level discrimination in the full benchmark (Table 23), and which audit components are load-bearing (Table 13). The remaining validation and leave-one-out checks then ask whether the same decision rule survives a small label budget or held-out dataset calibration.

Figure 4 is a compact summary rather than the primary evidence. The diagnostic conclusion does not depend on one scalar utility alone: contrastive pressure fails at sharing-heavy endpoints, pure compactness fails when valid negatives require separation, and the audit policy either selects the compatible pressure family or abstains when the proxy is too weak. Thus the regret plot should be read together with the Pareto sensitivity, false-negative-aware baselines, and natural relation-extraction diagnostics reported above.

E.2 Controlled-Policy and Mitigation Checks

The controlled-policy block reports the canonical false-negative sweep first (Table 16), then asks whether explicitly false-negative-aware contrastive variants remove the need for a regime-level audit (Tables 17 and 18). They reduce some fragmentation, but they do not erase the endpoint reversal: a fixed contrastive family is still brittle when sampled negatives cease to be valid. The policy-regret and transfer tables that follow therefore use the audit as a selection or abstention rule, not merely as another pairwise weight inside the same loss.

The intended reading order is diagnostic rather than chronological. The compact regret plot gives the visual summary first, while the following tables expose the assumptions behind that plot: threshold sensitivity, component ablations, held-out calibration, multi-seed policy regret, and downstream transfer. This keeps the appendix from treating every robustness run as equal evidence. Runs that only check local stability are reported as local stability checks; runs that change the scientific interpretation, such as the fully shared false-negative endpoint, are tied back to the main claim about negative validity. In particular, a mitigation that reduces fragmentation on one SNLI slice is not promoted to a deployed policy unless it also behaves well on the mixed controlled sweep.

The compact SNLI false-negative-aware diagnostic is intentionally placed before the larger mixed-grid mitigation table. It answers a narrow question first: whether debiasing or filtering contrastive negatives can reduce fragmentation when the sampled negatives are suspected false negatives. The answer is yes in the small SNLI slice, but the larger controlled grid shows why this is not sufficient as the paper's policy. False-negative-aware contrastive losses still share the same basic repulsive pressure, and the fully shared endpoint continues to prefer a variance-preserving objective.

The follow-up grid widens this check across the synthetic false-negative rates. We keep it separate from Table 16 because it is a mitigation run under a reduced compute budget, not the canonical objective sweep. Its role is to rule out the simplest objection: that replacing ordinary contrastive negatives with debiased, hard, or filtered variants automatically solves the validity problem. The mitigation helps at some intermediate rates, but the preferred objective still changes with the latent validity regime.

F Naming and Audit Sanity Tables

This short appendix pairs two reference tables. Table 26 separates standard objective names from paper-defined diagnostics so that shorthand labels (AC, Audit-hybrid, Audit-select, Audit-weighted, the topic-cluster audit guard) cannot be mistaken for established community terminology. Table 27 gives the controlled label-aware audit statistics that we retain only as an oracle-like sanity check: when family count and entropy are visible, the controlled label-aware audit recovers the *same low-validity geometry diagnosis* as the label-sparse audit in Table 12, but the *deployed policy* on SNLI uses abstention with a variance-preserving fallback (Barlow/VICReg) rather than deploying AC. The label-free signal is therefore consistent with the metadata-based diagnosis without using that metadata in the audit decision, and AC is consistently reported as a stress-test diagnostic throughout the paper. The naming table is separated from the result tables because it has a different role: it prevents paper-defined diagnostics from being mistaken for standard loss functions. The audit table is likewise a sanity check rather than a source of deployed decisions. Keeping these two artifacts together makes the boundary clear: the appendix can document terminology and oracle-style checks without turning them into additional headline claims. This is also why the SNLI rows repeat the deployed fallback

Table 7: Unified sensitivity summary. This table consolidates the appendix checks that otherwise repeat the same message: the endpoint-failure pattern is stable under several perturbations, but the evidence is descriptive rather than a universal calibration or significance guarantee.

Sensitivity axis	Perturbation	Main observation	Where reported
Scalar utility	Mixed-sweep weights vary over apparent recovery, latent-neighbor recovery, and fragmentation	Selected pattern is unchanged: Hybrid for 0–75% false negatives and Barlow at 100%.	Table 8
Pareto lens	Evaluate apparent-family recovery and latent-sharing preservation before choosing scalar weights	Fixed endpoints fail under some trade-offs; Hybrid/Barlow remain frontier choices in mixed/low-validity regimes.	Table 11
Seed variation	Controlled false-negative sweep over seeds 42/43/44	Audit-select has average rank 1.00 and zero mean regret under this common utility; fixed objectives fail at opposite endpoints.	Table 19
Audit thresholds	Vary silhouette, AMI, and topic-guard length thresholds locally	Decisions remain unchanged on the studied suite; this supports local robustness, not universal calibration.	Table 9
Audit components	Remove one audit signal at a time	Cluster stability and topic guard are load-bearing; neighbourhood and compatibility are supporting checks in this grid.	Table 13
Geometry scale	Recompute with normalized Euclidean and cosine distances	SNLI fragmentation gap persists after normalization, although raw magnitude is scale-sensitive.	Table 32
Positive pairing	Change positive construction on BATS and SNLI	Qualitative mismatch pattern remains: discriminating regimes need separation; sharing regimes expose repulsion fragmentation.	Tables 33–35
Negative sampler	Random, filtered, and entity-perturbed SNLI negatives	Filtering does not help, while entity perturbation helps only partially; regime-level validity remains the relevant diagnosis.	Table 36
Natural RE case	Apply the same audit-style diagnostic to FewRel and Re-DocRED	Low-validity failure is not unique to SNLI; Re-DocRED triggers conservative variance-preserving fallback with high abstention; the downstream sanity check is mixed and treated as a boundary condition.	Tables 4, 5, and 6

Table 8: Weight sensitivity for the controlled false-negative-rate sweep. Each row recomputes the utility with different weights on apparent-family NMI, latent-neighbor accuracy, and inverse latent fragmentation. Hybrid abbreviates Audit-hybrid.

Weights	FN0	FN25	FN50	FN75	FN100
(.45,.35,.20)	Hybrid	Hybrid	Hybrid	Hybrid	Barlow
(.33,.33,.33)	Hybrid	Hybrid	Hybrid	Hybrid	Barlow
(.50,.25,.25)	Hybrid	Hybrid	Hybrid	Hybrid	Barlow
(.25,.50,.25)	Hybrid	Hybrid	Hybrid	Hybrid	Barlow
(.25,.25,.50)	Hybrid	Hybrid	Hybrid	Hybrid	Barlow

Table 9: Local sensitivity of the label-sparse audit thresholds. Each setting varies one threshold family around the reported rule. Decisions remain unchanged on the studied suite; SNLI cells abbreviate abstention to the VICReg fallback ($A \rightarrow V$).

Setting	BATS	BLESS	SNLI-E	SNLI-C	SemEval
Default	Ctr	Ctr	$A \rightarrow V$	$A \rightarrow V$	Ctr
Sil. $\geq$ .03	Ctr	Ctr	$A \rightarrow V$	$A \rightarrow V$	Ctr
Sil. $\geq$ .10	Ctr	Ctr	$A \rightarrow V$	$A \rightarrow V$	Ctr
AMI $\geq$ .15	Ctr	Ctr	$A \rightarrow V$	$A \rightarrow V$	Ctr
AMI $\geq$ .40	Ctr	Ctr	$A \rightarrow V$	$A \rightarrow V$	Ctr
Guard 18–19	Ctr	Ctr	$A \rightarrow V$	$A \rightarrow V$	Ctr

Table 10: Compact common-budget objective grid. All rows use one epoch and three seeds; BLESS and SNLI use fixed local subsamples. This table is a stress test for objective preference under equal compute, not a replacement for the converged main results. The grid is intentionally reported even where it weakens a simple "Hybrid always wins" story.

Dataset	Primary metric	Best row	Audit-hybrid row	Takeaway
BATS	VerAcc/MRR	AC .399±.021 / .525±.015	.288±.003 / .433±.011	Early-budget transfer favors compactness
BLESS	VerAcc/MRR	Hybrid .588±.025 / .718±.013	**.588±.025 / .718±.013**	Hybrid matches best utility
SemEval	VerAcc/MRR	Hybrid .061±.015 / .129±.004	**.061±.015** / .129±.004	Scores are low; MRR favors VICReg
SNLI-Ent	Frag. ↓	AC .180±.006	.760±.039	Hybrid reduces but does not remove repulsion
SNLI-Con	Frag. ↓	AC .185±.004	.793±.016	Same sharing-regime pattern

language instead of reporting AC as if it were the recommended objective.

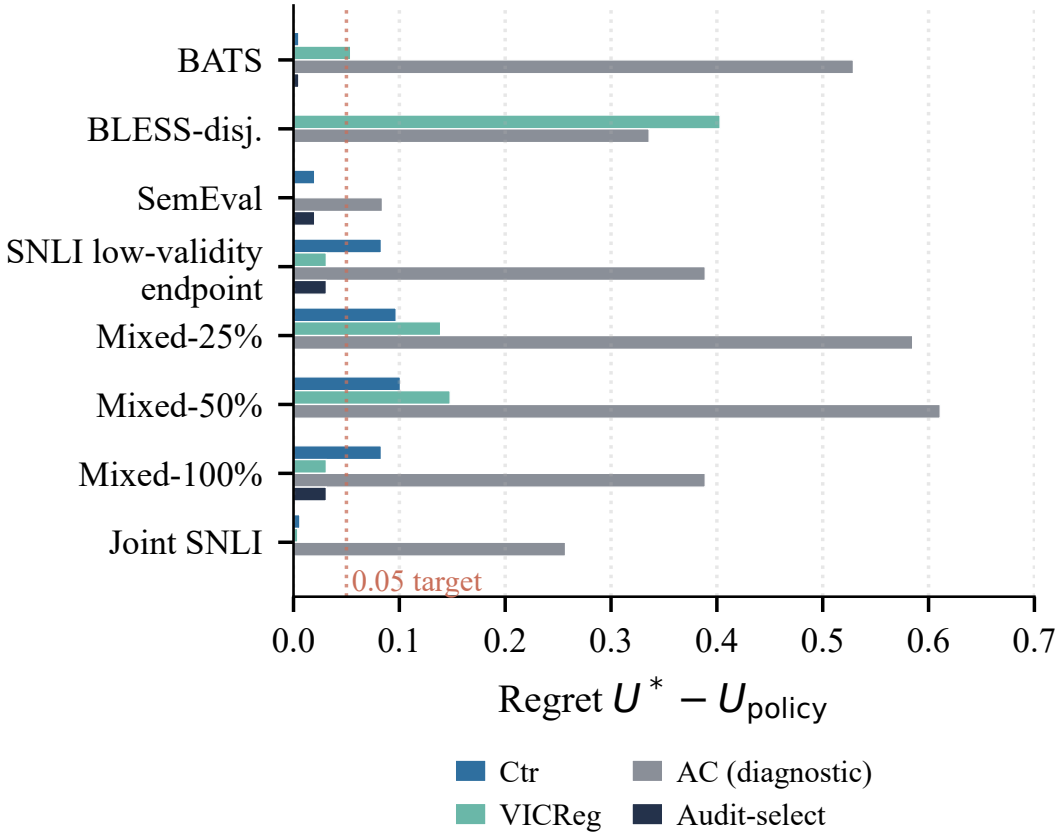

Figure 4: Controlled-grid objective-selection regret visualised. Each row is a regime; bars compare four policies under the regime's primary utility. The red dotted line marks the sub-0.05 regret target.

Table 11: Calibration-free mixed-regime sensitivity. Pareto winners are computed over apparent-family recovery and latent-sharing preservation before applying a scalar utility; worst-weight regret is the maximum regret over a grid of convex utility weights.

FN	Frontier	Minimax	Max reg.
0%	Hybrid, AC	Hybrid	0.572
25%	Hybrid	Hybrid	0.211
50%	Barlow, Hybrid	Hybrid	0.169
75%	Hybrid	Hybrid	0.150
100%	Barlow	Ctr	0.390

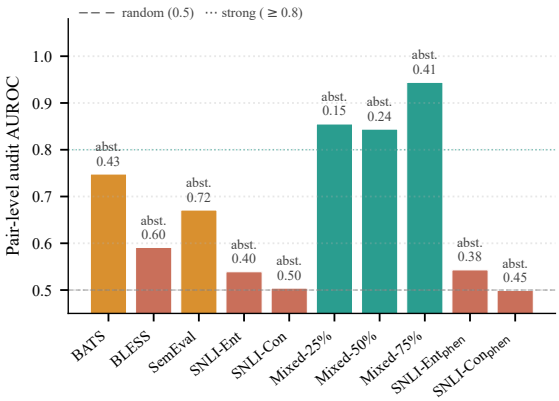

Figure 5: Pair-level negative-validity AUROC across datasets, with per-row abstention rate annotated above each bar. Green bars indicate high-confidence regimes (AUROC $\geq$ 0.65); salmon bars indicate near-random regimes (AUROC $<$ 0.65) in which Audit-select defers to a variance-preserving fallback. The audit's reliability is itself regime-dependent, which is what motivates abstention-based deployment.

G Dataset Details

G.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.

G.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` $\rightarrow$ `alligator`) and applying the same word template, the processed dataset contains 37,173 in-

Table 12: Label-sparse negative-validity audit. Labels are hidden during the decision and used only for the final agreement check. Without the topic-cluster guard, the stable SNLI clusters would be misread as valid negatives.

Dataset	$k^\star$	Sil.	AMI	Guard	Deployed
BATS	12	0.151	0.725	no	Ctr
BLESS	12	0.117	0.501	no	Ctr
SNLI-Ent	12	0.145	0.759	yes	Abstain → VICReg; AC diagnostic only
SNLI-Con	12	0.135	0.776	yes	Abstain → VICReg; AC diagnostic only
SemEval	12	0.126	0.442	no	Ctr

Table 13: Audit-component ablation. Each variant disables one audit signal and reports the resulting policy regret over the false-negative sweep (seeds 42/43/44). Cluster stability and the topic guard are the load-bearing signals.

Variant	Avg rank	Worst regret	Mean regret	Mean util	Fails on
full	1.07	**0.003**	**0.000**	**0.907**	few
no_neighborhood	1.07	0.003	0.000	0.907	few
no_compatibility	1.07	0.003	0.000	0.907	few
no_uncertainty	1.07	0.003	0.000	0.907	few
no_cluster	2.60	0.152	0.092	0.815	0%–75% false-negative
no_topic_guard	2.60	0.152	0.092	0.815	0%–75% false-negative

Table 14: Validation-selection baseline at increasing label budget vs. Audit-select. Val-select needs hundreds of labelled pairs per regime to approach the label-free audit's worst-case regret. Simulations average over 200 noise realisations.

Policy	Labels/regime	Avg rank	Worst regret	Mean regret	Mean util
Audit-select	0	**1.07**	**0.003**	**0.000**	**0.907**
Val-select	10	2.23	0.181	0.059	0.848
Val-select	50	1.64	0.136	0.026	0.881
Val-select	200	1.21	0.046	0.005	0.903
Val-select	1000	1.10	0.008	0.001	0.907

Table 15: Leave-one-dataset-out audit calibration. The decision threshold is tuned on the four other datasets and evaluated on the held-out family. Overall held-out agreement is 4/5 = 80%; the single SNLI-Ent miss is a medium-confidence call that triggers the audit's abstention.

Held-out	Tuned thr.	Held-out score	Predicted	Target	Agrees
BATS	0.63	0.91	contrastive	contrastive	yes
BLESS	0.63	0.91	contrastive	contrastive	yes
SemEval	0.63	0.90	contrastive	contrastive	yes
SNLI-Ent	0.61	0.63	contrastive	align_uniform	*abstain*
SNLI-Con	0.63	0.60	align_uniform	align_uniform	yes

Table 16: Controlled false-negative-rate sweep. Utility is measured on a held-out synthetic test split. The utility is $0.45\,U_{\mathrm{appNMI}} + 0.35\,U_{\mathrm{latentNN}} + 0.20/(1+\Delta_{\mathrm{latent}})$, with fixed weights chosen before objective training to require both apparent-family recovery and latent sharing preservation. The preferred pressure shifts from separation or hybrid pressure at low and medium false-negative rates to variance preservation at the fully shared endpoint.

False-negative rate	Ctr	Barlow	VICReg	Audit-weighted	Hybrid	AC
0%	0.872	0.832	0.823	0.820	**0.873**	0.391
25%	0.865	0.835	0.823	0.820	**0.962**	0.377
50%	0.870	0.844	0.823	0.820	**0.971**	0.360
75%	0.881	0.856	0.827	0.808	**0.970**	0.380
100%	0.626	**0.709**	0.678	0.291	0.329	0.320

Table 17: False-negative-aware contrastive diagnostic on a compact SNLI-entailment slice (train 512 / dev 256 / test 256). Lower fragmentation is better. FN-aware variants reduce fragmentation but do not replace regime-level auditing.

Objective	Fragmentation ↓	VerAcc
Ctr	1.214	1.000
Debiased Ctr	1.175	1.000
FN-filtered Ctr	1.045	1.000

stances across 6 families with the same 70/15/15 stratified split.

G.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 hypothe-

Table 18: Budgeted trained mixed-regime grid with false-negative-aware contrastive variants. Utilities use the same mixed-regime diagnostic lens as Table 16; the run uses a reduced compute budget, so it is a mitigation check rather than the canonical mixed-grid result.

FN rate	Ctr	Debiased	Hard	Barlow	VICReg	Audit-w.	Hybrid	Best
0%	0.873	**0.877**	0.873	0.855	0.827	0.829	0.872	Debiased
25%	0.872	0.876	0.872	0.861	0.822	0.834	**0.898**	Hybrid
50%	0.882	0.884	0.881	0.877	0.761	0.819	**0.906**	Hybrid
75%	0.796	0.793	0.788	0.772	0.525	0.727	**0.815**	Hybrid
100%	0.329	0.314	0.338	**0.678**	0.426	0.439	0.320	Barlow

Table 19: Policy-level regret on the controlled false-negative sweep, averaged over seeds 42, 43, 44. Utility is the same as Table 16; regret is oracle utility minus the utility obtained by a fixed or audit-selected policy. This common lens reduces metric–objective circularity by comparing policies under one utility. Audit-weighted denotes the pair-weighted contrastive ablation that consumes the audit score but not its decision; the deployed policy is Audit-select, which uses the audit to choose hybrid pressure except at the fully shared endpoint, where it selects Barlow.

Policy	Avg rank	Worst regret	Mean regret	Mean U	Fails on
Audit-select	**1.00**	**0.000**	**0.000**	**0.897**	few
Oracle	1.00	0.000	0.000	0.914	upper bound
Always Ctr	2.20	0.151	0.087	0.810	100% false-negative
Always Barlow	2.60	0.124	0.082	0.815	0%–75% false-negative
Always VICReg	3.60	0.148	0.102	0.795	0%–75% false-negative
Audit-hybrid	1.60	0.372	0.074	0.823	100% false-negative
Audit-weighted	5.20	0.410	0.182	0.715	0%–100% false-negative
Always AC	5.80	0.597	0.519	0.378	0%–100% false-negative

Table 20: Mechanistic SNLI fragmentation analysis. We cluster held-out contrastive relation embeddings and measure NMI with heuristic sources. Topic and surface-form sources align at least as much as NLI phenomenon tags, suggesting that contrastive fragmentation is not primarily clean latent subrelation discovery.

Label	Source	NMI
Entailment	topic TF-IDF	0.075
Entailment	surface form	0.055
Entailment	NLI phenomenon	0.035
Contradiction	topic TF-IDF	0.067
Contradiction	surface form	0.039
Contradiction	NLI phenomenon	0.036

Table 21: Downstream utility checks beyond geometry. Held-out BATS transfer trains on disjoint relation families and evaluates 3-shot prototypes over unseen families. Answer retrieval ranks candidate answers for analogy-style $A{:}B :: C\,?$ queries; it is intentionally harder and remains low for all objectives.

Evaluation	Objective	Budget	Seeds/Queries	Acc./Hit@1	MRR
BATS held-out transfer	Ctr	1 epoch	3 seeds	0.374±0.053	0.577±0.055
BATS held-out transfer	AC stress	1 epoch	3 seeds	**0.550±0.075**	**0.695±0.059**
BATS held-out transfer	Audit-hybrid	1 epoch	3 seeds	0.496±0.017	0.674±0.004
BATS answer retrieval	Ctr	80q/250c	80	0.075	0.189
BATS answer retrieval	AC stress	80q/250c	80	**0.087**	**0.216**
SemEval answer retrieval	Ctr	79q/147c	79	0.013	0.097
SemEval answer retrieval	AC stress	79q/147c	79	**0.025**	**0.141**
SemEval answer retrieval	Barlow	79q/147c	79	0.013	0.076

sis, 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.

G.4 Data Quality Statistics

Table 28 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.

Table 22: Capacity and backbone checks on an 8GB RTX 3060 Ti. One-epoch full-fine-tune runs use small slices; for BATS/BLESS/SemEval/SNLI-Ent/SNLI-Con, full and mixed runs use three seeds (42, 43, 44) and last-layer runs also now use three seeds (42, 44, 45). Capacity changes absolute scores, but does not remove the objective–regime mismatch.

Setting	Dataset	Objective (best)	Metric summary
Full trainable	BATS	VICReg	VerAcc 0.369±0.006, MRR 0.506±0.010, NMI 0.645±0.023
Full trainable	BLESS	VICReg	VerAcc 0.495±0.052, MRR 0.653±0.036, NMI 0.064±0.027
Full trainable	SemEval-2012	VICReg	VerAcc 0.068±0.004, MRR 0.149±0.006, NMI 0.811±0.010
Full trainable	SNLI-Ent	AC	VerAcc 1.000, MRR 1.000, Frag. 0.257±0.011
Full trainable	SNLI-Con	AC	VerAcc 1.000, MRR 1.000, Frag. 0.273±0.017
Last-layer	BATS	AC	VerAcc 0.308±0.034, MRR 0.422±0.042, NMI 0.634±0.017
Last-layer	BLESS	VICReg	VerAcc 0.467±0.031, MRR 0.633±0.016, NMI 0.033±0.006
Last-layer	SemEval-2012	AC	VerAcc 0.065, MRR 0.134±0.012, NMI 0.817±0.005
Last-layer	SNLI-Ent	AC	VerAcc 1.000, MRR 1.000, Frag. 0.586±0.020
Last-layer	SNLI-Con	AC	VerAcc 1.000, MRR 1.000, Frag. 0.566±0.018
BERT	BATS	Ctr / AC	VerAcc 0.775 / 0.141
BERT	BLESS	Ctr / AC	VerAcc 0.970 / 0.127
BERT	SNLI-Ent	Ctr / AC	Frag. 20.624 / 0.00003
BERT	SNLI-Con	Ctr / AC	Frag. 20.735 / 0.00003
RelBERT	BATS	AC stress	VA/MRR .406/.557
RelBERT	BATS	Ctr	VA/MRR .600/.711
RelBERT	BATS	Audit-hybrid	VA/MRR .622/.727

Table 23: Pair-level negative-validity prediction. Each candidate negative is assigned a validity score from frozen relation features; we report AUROC, AUPRC, expected calibration error (ECE), and the abstention rate and accuracy of an uncertainty-aware decision. The score is most reliable in controlled mixed false-negative regimes and on BATS, while SNLI subsets, where validity itself is ambiguous, show near-random AUROC with high abstention—consistent with the regime audit's medium-confidence calls in Table 25. SemEval reaches high AUPRC but the raw scores are mis-calibrated (ECE 0.351); this is the empirical motivation for using audit *confidence* rather than the score directly to gate objective choice.

Dataset	AUROC	AUPRC	ECE	Abst.	Abst. acc
BATS	0.748	0.678	0.079	0.428	0.774
BLESS	0.591	0.557	0.048	0.603	0.606
SemEval-2012	0.671	**0.879**	0.351	0.722	0.484
Re-DocRED	0.660	0.619	0.130	0.664	0.676
SNLI-Ent	0.539	0.530	0.122	0.400	0.535
SNLI-Con	0.504	0.499	0.116	0.503	0.501
Mixed-25% false-negative	0.855	0.847	0.115	0.147	0.735
Mixed-50% false-negative	0.844	0.808	0.084	0.245	0.796
Mixed-75% false-negative	**0.944**	**0.944**	0.197	0.407	**0.959**
SNLI-Ent (phen.)	0.543	0.541	0.120	0.382	0.541
SNLI-Con (phen.)	0.499	0.494	0.133	0.448	0.501

Table 24: Joint SNLI 3-way relation-space evaluation (formerly main-body Table 5, moved here to keep the main body within the eight-page limit; chance is 0.333). Audit-weighted attains the best linear accuracy and F1 and ties VICReg on AUC; AC collapses to chance. The same functional sanity-check row is included in the controlled-grid regret visualisation.

Obj.	Acc.	F1	AUC	Proto
Ctr	0.584	0.583	0.761	0.524
AC	0.333	0.167	0.500	0.361
Barlow	0.583	0.583	0.759	0.536
VICReg	0.586	0.585	**0.770**	**0.565**
Audit-weighted	**0.590**	**0.589**	**0.770**	0.472

H Implementation Details

H.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 $W \in \mathbb{R}^{256\times768}$.

H.2 Training Protocol

All models are trained with AdamW (weight decay 10^{-4}) at learning rate 10^{-3} with cosine annealing

Table 25: Regime-adaptive objective selection (formerly main-body Table 6, moved here to keep the main body within the eight-page limit). The label-free audit chooses a *deployed* pressure with a confidence level; AC is reported as a geometry stress-test diagnostic only and is *never* the deployed pressure. The "Deployed" column gives the actual policy executed by Audit-select on each regime and matches the controlled-grid regret visualisation row-for-row; "Best (matched?)" lists the per-regime oracle choice and notes whether the deployed policy matches it (matched exactly, within sub-0.05 regret, or fallback-only on the stress test).

Dataset	Deployed (Audit-select)	Conf.	Best (matched?)
BATS	Ctr	high	Ctr (yes)
BLESS-disjoint	Ctr	medium	Ctr (yes)
SemEval	Ctr	medium	Barlow/VICReg oracle; Ctr deployed within .020 regret
SNLI-Ent	Abstain → VICReg fallback; AC diagnostic	medium	VICReg (fallback yes; AC geom. only)
SNLI-Con	Abstain → VICReg fallback; AC diagnostic	medium	VICReg (fallback yes; AC geom. only)
Mixed-25% false-negative	Hybrid	medium	Hybrid (yes)
Mixed-50% false-negative	Hybrid	high	Hybrid (yes)
Mixed-100% false-negative	Barlow	medium	Barlow (yes)

Table 26: Naming guide for objectives, ablations, and diagnostics. Standard method names are separated from paper-defined diagnostics so that shorthand labels such as AC, Audit-hybrid, and the audit guard are not read as established community terminology.

Name in paper	Type	Standard?	Purpose in this paper
Contrastive (InfoNCE; Ctr) / SupCon	Objective baseline	yes	Tests explicit separation pressure when sampled negatives cross relation families.
AC compactness baseline	Objective baseline	no, defined here	Non-repulsive compactness stress-test baseline; not Wang–Isola uniformity and not the final recommendation for all sharing regimes.
Barlow Twins / VICReg	Objective baseline	yes	Relation-space variance-preserving and redundancy-reducing pressure without explicit instance negatives.
SimSiam / BYOL	Objective baseline	yes	Relation-space bootstrapped non-contrastive invariance.
Debiased / similarity-filtered Ctr	Objective variant	mostly / partly	Tests whether false-negative-aware contrastive mitigation reduces sharing fragmentation.
BERT backbone, difference-only extractor, larger MLP	Robustness ablation	descriptive	Tests encoder, extractor, and capacity dependence.
Entity-disjoint BLESS; normalized/cosine geometry; multi-seed	Robustness check	yes	Tests leakage, scale, and seed sensitivity.
Label-sparse negative-validity audit	Paper method	no	Approximates whether apparent negatives are valid using cluster stability, local neighborhoods, compatibility, and a topic-cluster guard.
Controlled false-negative-rate sweep	Custom diagnostic	no	Varies the constructed false-negative rate to test continuous negative validity.
Joint SNLI 3-way relation-space evaluation	Custom task evaluation	no	Trains entailment, contradiction, and neutral in one relation space to test functional distinctions.
Audit-hybrid / Audit-select	Custom objective and policy	no	Uses the audit to weight, blend, or select pressure under mixed validity signals.

Table 27: Controlled label-aware audit statistics. Explicit family count and family entropy are used only as an oracle-like sanity check, not as the proposed label-sparse audit. The "Diagnosis" column reports the controlled audit's low- vs. high-validity geometry diagnosis; the deployed policy on each regime is given by Table 12 and the controlled-grid regret visualisation row-for-row (in particular AC is reported only as a diagnostic and is never the deployed pressure on SNLI).

Dataset	# Fam.	Entropy	Frozen Sil.	Diagnosis (geometry only)
BATS	40	5.3219	-0.0635	high validity: separation pressure
BLESS	6	2.2139	-0.0191	high validity: separation pressure
SNLI-Ent	1	0.0000	0.0000	low validity: AC diagnostic / Abstain → VICReg fallback
SNLI-Contra	1	0.0000	0.0000	low validity: AC diagnostic / Abstain → VICReg fallback

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 main table uses seed 42; we additionally report multi-seed stability over seeds

Table 28: 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	False-negative 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	–

{42,43,44}.

H.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.

AC compactness objective. $\lambda_{\text{align}} = 1.0$, $\lambda_{\text{comp}} = 0.05$, compactness temperature $t = 2.0$. Positives are constructed via nearest-neighbor pairing; no instance-specific negatives are used. The code path is named `align_uniform` for backward compatibility, but the reported objective is AC.

H.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. Importantly, the sampler is fixed by dataset rather than by objective: on a given dataset, the contrasted objective conditions see the same batch construction, positive-construction rule, batch size, and random seed. Thus the main comparison changes the loss without changing batch composition. We do not test a broader sampler intervention in which a discriminating dataset is trained with sharing-style random batches or a sharing dataset is forced into artificial balanced pseudo-families.

To make this control auditable without exposing implementation-specific logs, we record the sampler family used by each dataset and verify that paired objective runs share the same batch construction (Table 29).

Table 29: Batch-composition audit. Paired objective runs use the same sampler, positive-construction rule, batch size, and random seed within each dataset, so the main comparisons change the objective rather than the batch composition.

Dataset	Batches	Fam./batch	Paired construction
BATS	10	8.00	matched
BLESS	295	6.00	matched
SNLI-Ent	390	1.00	matched
SNLI-Con	390	1.00	matched

H.5 Compute Environment and Scope

All experiments were run with a single-GPU training setup. The released scripts separate the submitted core grid from supporting robustness checks: the core grid contains the main objective comparisons, backbone and extractor checks, and pre-training audits; the supporting checks cover frozen-feature baselines, apparent-negative similarity diagnostics, the larger-MLP ablation, entity-disjoint BLESS, and additional seeds for the main grid. We report this scope to make the experimental budget reproducible, but do not use wall-clock runtime as evidence for any scientific claim.

H.6 Reproducibility Manifest

The released project includes a reproducibility manifest that maps each table, figure, and robustness diagnostic to the command that regenerates it and the expected output files. In particular, the manifest covers the main grid, backbone and extractor ablations, regime audit, sensitivity analyses, frozen-feature baselines, multi-seed runs, BLESS entity-disjoint split, SNLI prototype-transfer diagnostic, and normalized/cosine geometry diagnostics. This is intended to make the role of each artifact explicit and to separate core paper evidence from additional diagnostic checks.

I 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

Table 30: Metric interpretation guide. The task metrics evaluate retrieval or verification quality, while structural metrics diagnose whether the learned relation space has the geometry required by the dataset regime.

Metric	Used for	Better direction	Interpretation
Verification accuracy	Discriminating and sharing task checks	Higher	Nearest-neighbor label agreement. Informative for multi-family datasets; can saturate in single-family sharing subsets.
Recall@1 / MRR	Retrieval quality	Higher	Ranks same-family neighbors. Strong when multiple valid relation families compete; less diagnostic when all held-out items share one label.
Separation margin	Discriminating regimes	Higher	Between-family distance minus within-family distance. Low or near-zero values indicate under-separation.
Silhouette / NMI	Discriminating regimes	Higher	Cluster quality under family labels. These complement retrieval by checking whether the whole relation space is organized by relation family.
Fragmentation	Sharing regimes	Lower	Within-subset spread. High values indicate that explicit repulsion has broken a relation-sharing manifold into incompatible pieces.
Normalized/cosine geometry	Scale-confound check	Same as corresponding metric	Recomputes structural diagnostics after L2 normalization or with cosine distance; these are the conservative evidence when raw Euclidean magnitudes are scale-sensitive.

retrieval and verification pipelines are computed independently and can differ slightly in practice.

MRR (Mean Reciprocal Rank).

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

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

Separation margin (Δ_{sep}).

$$\begin{aligned}\Delta_{\text{sep}} = \mathbb{E}[\|r_i - r_j\|_2 \mid y_i \neq y_j] \\ - \mathbb{E}[\|r_i - r_j\|_2 \mid y_i = y_j]\,.\end{aligned} \tag{14}$$

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.

Frozen-relation silhouette proxy. For the pre-training audit, we compute a frozen-encoder proxy before objective training. We encode each pair with RoBERTa-base, represent the pair by a linear projection of the frozen difference vector $z_b - z_a$, and compute the cosine silhouette using family labels when the dataset contains at least two families. Single-family sharing subsets receive value 0 because silhouette is undefined without multiple labels. The proxy is reported for transparency, but the controlled-grid audit decision in Table 27 is based only on family count and entropy.

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.

J Consolidated Robustness and Diagnostic Results

Table 31 consolidates the objective preference ratios across all three configurations.

Table 31: Objective preference ratios across configurations. For discriminating datasets, the ratio is VerAcc(Ctr)/VerAcc(AC). For sharing datasets, the raw fragmentation ratio is reported only as a scale-sensitive stress-test summary; Table 32 gives the conservative normalized values.

Dataset	Main	BERT	Diff
BATS (VerAcc)	5.2×	5.5×	5.4×
BLESS (VerAcc)	8.8×	7.6×	1.6×
SNLI-Ent (raw Frag)	high	high	> 3K×
SNLI-Con (raw Frag)	high	high	> 3K×

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 a large raw fragmentation gap, confirming that the phenomenon is not extractor-specific. We avoid interpreting the raw ratios as effect sizes; the normalized and cosine diagnostics in Appendix K are the conservative geometry evidence.

K Scale-Invariant Geometry Diagnostics

To rule out the possibility that Euclidean structural metrics merely reflect objective-dependent embedding scale, we recompute geometry after L2-normalizing each test relation embedding. Table 32 reports both normalized Euclidean and cosine-distance analogs. The same final checkpoints as Table 3 are used; no retraining is performed.

For BATS and BLESS, objective mismatch is not a scale artifact: contrastive training has large normalized/cosine separation, whereas AC has essentially no between-family margin. For SNLI, the normalized result is more nuanced. The contrastive objective still produces roughly 24–25× larger normalized Euclidean fragmentation than AC, but the raw Euclidean fragmentation gap is amplified by the much larger norm of contrastive embeddings. We therefore use the normalized metrics as the conservative geometry evidence and avoid interpreting raw fragmentation ratios as scale-invariant effect sizes.

L 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 33 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.

Results on SNLI entailment. Table 35 shows that the qualitative sharing-regime pattern is also stable across positive construction rules. Verification remains saturated under both objectives, but AC 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.

M 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 36 shows that filtering alone does not reduce fragmentation: filtered in-batch negatives increase fragmentation from 4.202 to 4.408. Entity-perturbed negatives do reduce fragmentation to 2.147, but even this stronger strategy remains 153× worse than the AC baseline, which achieves fragmentation 0.014. Thus, the sharing-regime disadvantage of contrastive training cannot

Table 32: Geometry after L2-normalizing relation embeddings. Discriminating results remain strong under cosine separation. In sharing regimes, contrastive training remains more fragmented than AC after normalization, but the raw fragmentation magnitude is partly scale-sensitive.

Dataset	Objective	Pressure role	Mean Raw Norm	Norm. L2 Sep./Frag.	Cos Sep./Frag.	Cos NMI/Q90
BATS	Ctr	separation	2.695	sep 0.7321	sep 0.6691	NMI 0.7548
BATS	AC	wrong pressure	0.336	sep 0.0011	sep 0.00003	NMI 0.2736
BLESS	Ctr	separation	3.591	sep 0.6559	sep 0.5455	NMI 0.7790
BLESS	AC	wrong pressure	0.494	sep -0.0001	sep -0.00000	NMI 0.0034
SNLI-Ent	Ctr	frag. stress	380.816	frag 0.0176	frag 0.000170	q90 0.000314
SNLI-Ent	AC	comp. stress	0.338	frag 0.00072	frag 0.00000025	q90 0.00000054
SNLI-Con	Ctr	frag. stress	403.255	frag 0.0173	frag 0.000162	q90 0.000289
SNLI-Con	AC	comp. stress	0.337	frag 0.00070	frag 0.00000023	q90 0.00000054

Table 33: Positive construction sensitivity on BATS. Across multiple positive construction rules, contrastive training remains superior to AC 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	AC	0.347	0.347	0.487
NN same-family	Ctr	0.475	0.475	0.613
NN same-family	AC	0.378	0.378	0.488

Table 34: 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	AC	0.008	-0.084	0.561
NN same-family	Ctr	1.336	0.005	0.674
NN same-family	AC	0.007	-0.092	0.568

Table 35: Positive construction sensitivity on SNLI entailment. Verification remains saturated, but AC 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	AC	1.000	0.014
Random same-label	Ctr	1.000	3.482
Random same-label	AC	1.000	0.015
Filtered NN	Ctr	1.000	4.466
Filtered NN	AC	1.000	0.014

be explained solely as an artifact of naive in-batch negative sampling.

Table 36: Alternative negative sampling strategies on the SNLI entailment sharing subset. Even more informative negatives remain far more fragmented than AC training.

Negative Strategy	Fragmentation	Relative to AC
Random in-batch	4.202	300× worse
Filtered in-batch	4.408	315× worse
Entity-perturbed	2.147	153× worse
AC (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.

N Qualitative Nearest-Neighbor Examples

The inline table below shows representative BATS nearest-neighbor examples from the trained relation spaces. For anchors expressing regular noun pluralization, matched contrastive training retrieves another regular-plural relation, while the AC-mismatch model retrieves relations from different transformation families. This illustrates the under-separation failure mode behind the aggregate BATS metrics.

We also inspected SNLI entailment relation spaces. Among 500 held-out entailment pairs, the farthest pair under contrastive training has squared distance 2331.59, whereas the farthest pair under AC has squared distance 2.22×10^{-5}. The contrastive farthest pair compares "a man and woman take selfies" with "women are hitting a volleyball," both valid entailment transformations but pushed far apart. This qualitative pattern matches the quantitative fragmentation metric: contrastive training separates compatible entailment relations, while AC keeps the shared-relation space compact.

O Training Dynamics

Figure 6 shows validation metric curves across epochs.

Table 37: Qualitative BATS nearest-neighbor examples. For anchors expressing regular noun pluralization, contrastive training retrieves another regular-plural relation, while the compactness diagnostic endpoint retrieves different transformation families.

Anchor	Family	Ctr NN	AC NN
government → governments	noun_plural_reg	song → songs	federal → federally
example → examples	noun_plural_reg	player → players	reduce → reducing
player → players	noun_plural_reg	government → governments	industry → industries
song → songs	noun_plural_reg	government → governments	energy → energies

Ctr retrieves the same regular-plural transformation; AC retrieves different transformation families such as adverb formation, gerund formation, or irregular pluralization.

Several patterns emerge from the training curves:

- On discriminating datasets, compactness-mismatch training quickly plateaus near chance verification accuracy, indicating that this objective lacks the separation pressure needed by the evaluation.
- On sharing stress tests, mismatched contrastive training produces high fragmentation, showing that instance-level repulsion works against the compactness diagnostic.
- The curves expose mismatch early, but they are diagnostics; functional claims require the joint SNLI and mixed-regime evaluations.

P Negative-Validity Proxy with Hardness Annotation

The pair-level negative-validity proxy used by Table 23 is an automated stand-in for a small human-annotation study, which is listed as a future extension of the audit toolkit and is not part of the current results. The construction is transparent: balanced sampling between in-family false negatives (label 0) and cross-family valid negatives (label 1), with hardness stratification by super-family (BATS, SemEval, the hand-coded BLESS super-family map) and by TF-IDF topic clusters (SNLI subsets). Cross-super-family random pairs are unambiguously valid negatives and serve as the easy class; same-super-different-family pairs are partially compatible and serve as the hard class. The audit's per-pair score is the cosine distance between frozen RoBERTa-base relation features (the same encoder as in the deployed audit).

Higher easy-fraction means random sampling already separates the audit's two classes; SemEval (0.90 easy frac.) yields the best AUPRC. The SNLI subsets have a single nominal family, so hardness stratification falls back to topic-cluster overlap; the audit's response there is to abstain under medium confidence rather than commit. A small real-human study is listed as a future extension of the audit toolkit and is not part of the current results.

Q Sharing-Regime Evaluation Note

On SNLI entailment and contradiction, verification accuracy, Recall@1, and MRR all saturate at 1.0 under *both* extreme 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 (AC) and a severely fragmented space (Ctr) in these settings.

This is a measurement limitation of the SNLI sharing stress test, and it is also the reason we report relation-space diagnostics. Standard downstream metrics may fail to reveal objective mismatch when local positives are easy but global relation geometry differs sharply.

As an auxiliary functional check, we also evaluate binary entailment-vs-contradiction classification using relation prototypes. For each objective, we use the already trained entailment and contradiction checkpoints, construct one class prototype from a few training examples, and classify held-out SNLI pairs by cosine similarity to the corresponding class-specific prototype. Because the entailment and contradiction prototypes live in separate trained relation spaces, this diagnostic tests whether each learned space supports compact relation membership rather than whether a single global embedding space solves NLI.

The absolute scores are modest, so we treat this result as auxiliary rather than as a replacement for broader downstream evaluation. Its value is limited: the lower-fragmentation AC spaces transfer better under this prototype scoring rule, but the separate-space setup does not establish global NLI discrimination, and the canonical joint SNLI check in Table 24 (Audit-weighted accuracy 0.590; Ctr/Barlow/VICReg in 0.583–0.586) supersedes it.

We also include a collapse sanity check for this sharing-regime evidence. A constant-vector representation would obtain fragmentation 0 by construction, but it has rank-zero relation-space variance and expected prototype accuracy/AUC of

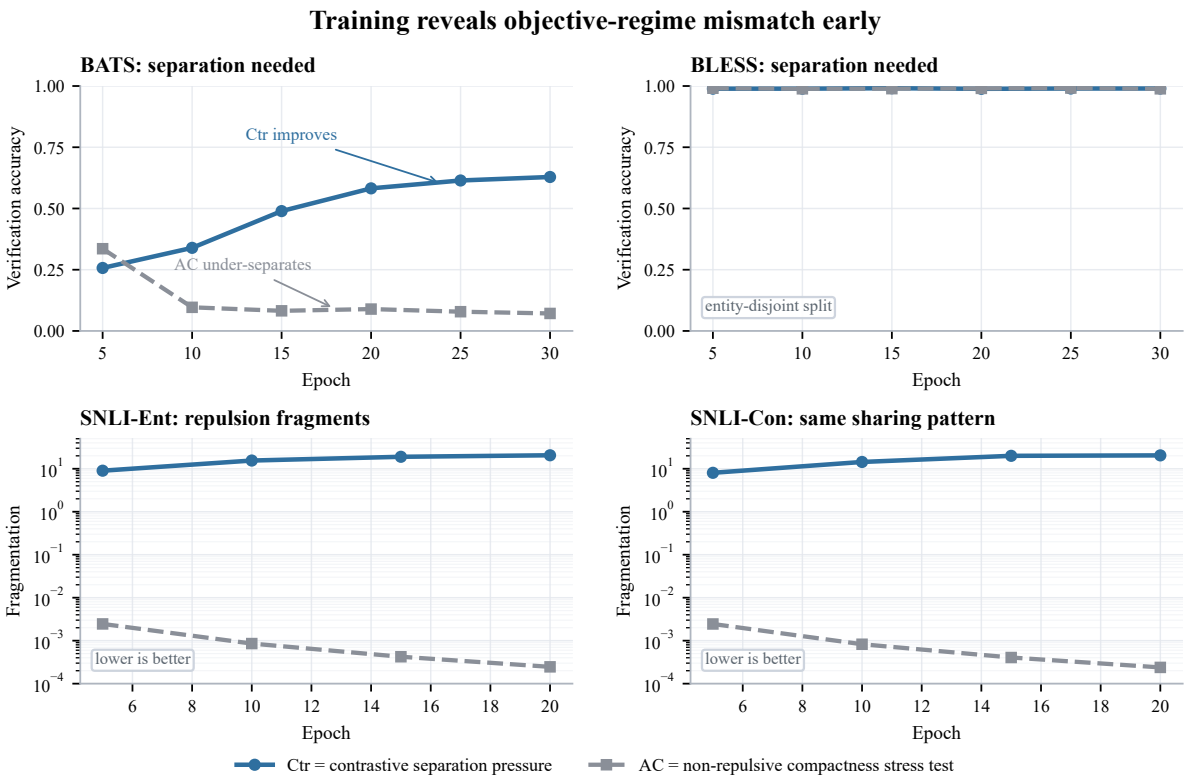

Figure 6: Training dynamics for all main experiments. The top row shows verification accuracy on discriminating datasets (BATS and entity-disjoint BLESS); the bottom row shows fragmentation on sharing-geometry stress tests (SNLI-Ent, SNLI-Con), where lower is better. Curves use the recorded validation checkpoints without smoothing.

Table 38: Per-dataset super-family distribution and cross-super easy fraction for the hardness-stratified proxy.

Dataset	Audit AUROC	Audit AUPRC	Easy frac.	# super-families
BATS	0.748	0.678	0.75	4
BLESS	0.591	0.557	0.30	2 (hand-coded)
SemEval-2012	0.671	0.879	0.90	10
SNLI-Ent	0.539	0.530	topic	1 (nominal)
SNLI-Con	0.504	0.499	topic	1 (nominal)

Table 39: Earlier SNLI prototype-transfer diagnostic, retained for historical reference. AC provides a modest advantage under this class-specific prototype scoring rule, but the prototypes come from separate trained spaces; the canonical joint SNLI evaluation is Table 24 (which trains all three labels in one relation space and reports the Audit-weighted row used in the controlled-grid regret visualisation).

Objective	Shots/label	Acc.	AUC
Ctr	8	0.482±0.001	0.479±0.005
AC	8	0.546±0.004	0.563±0.007
Ctr	32	0.480±0.004	0.477±0.003
AC	32	0.548±0.003	0.567±0.003
Ctr	128	0.479±0.004	0.477±0.002
AC	128	0.544±0.002	0.559±0.003

$0.5/0.5$. Thus our criterion is not "minimize fragmentation at any cost": the useful sharing-regime representation should remain compact *and* retain functional signal. The joint SNLI 3-way evaluation shows that AC only passes a weak separate-space prototype gate; it fails the stricter single-space functional gate. This is why the final policy allows variance-preserving or Audit-hybrid objectives rather than forcing AC whenever false-negative risk is high.

Negative-validity view. The label-sparse audit is also a negative-validity test. On BATS, BLESS, and SemEval-2012, stable short-pair clusters make cross-cluster negatives plausible, so separation is selected. On SNLI, the topic-cluster guard treats stable clusters as topic/form variation inside a sharing stress test, so the policy abstains or uses Audit-hybrid/variance-preserving pressure rather than forcing contrastive learning.