

Distributional Stabilization for Zero-Shot Relation Extraction

Yongxin Liu¹, Huaping Zhang¹, Qiuchi Li¹, Chunxiao Gao¹,
Haocheng Lv¹, Yan Ruohao², Baohua Zhang^{1*}, Lei Li¹

¹Beijing Institute of Technology, ²Xinjiang University

15936803692@163.com, kevinzhang@bit.edu.cn, liqiuchi@bit.edu.cn, Gao_chunxiao@bit.edu.cn
3120255822@bit.edu.cn, yanruohao@stu.xju.edu.cn, bhzhang_bit@163.com, lilei@bit.edu.cn

Abstract

Joint-encoding zero-shot relation extraction classifies all candidate relation labels in a single Transformer forward pass, achieving strong accuracy and high throughput. However, under a controlled full-vocabulary diagnostic, two *distributional* pathologies surface: unnormalized dot-product scoring causes *recall collapse*, with certain types receiving zero predictions, while the per-instance BCE loss leaves gold and negative scores entangled beyond any reliable threshold—*confidence ambiguity*. We propose **SCAR**, which keeps the baseline pathway intact and *stabilizes* both distributions via two auxiliary branches: a gradient-isolated coarse-to-fine cascade with Z-score normalization that restores cross-relation comparability, and a training-only contrastive alignment module that separates gold from negative scores. On Wiki-ZSL ($m=15$), SCAR raises Macro F1 from 75.9% to 83.0% and cuts zero-recall types from 35 to 1, suggesting that small-pool ZSRE protocols can obscure failure modes critical to large-vocabulary deployment.

1 Introduction

Relation extraction—identifying semantic relations between entity pairs in text—underpins knowledge-graph construction (Ji et al., 2022), question answering, and biomedical text mining. Real-world relation inventories are large and continually evolving, making it impractical to enumerate all types at training time. Zero-shot relation extraction (ZSRE) addresses this by treating candidate relations as natural-language labels supplied at inference.

Existing ZSRE methods include reading-comprehension (Levy et al., 2017) and entailment (Obamuyide and Vlachos, 2018; Sainz et al., 2021) approaches ($|\mathcal{R}|$ forward passes per sentence), prompt-based methods (Chia et al., 2022; Lv et al., 2023), and dual-encoder matching (Chen

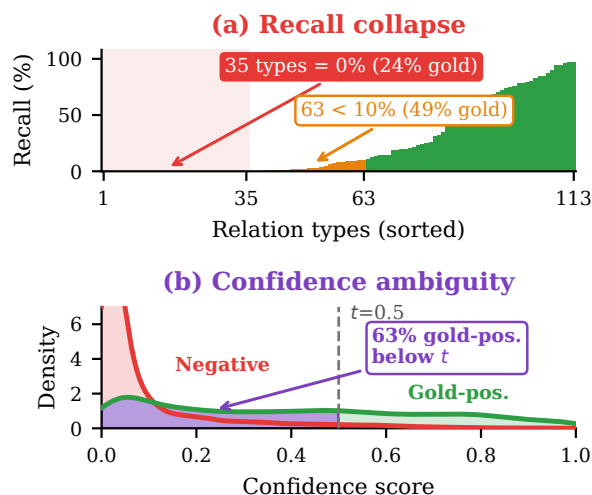


Figure 1: Baseline failures on Wiki-ZSL under the full-vocabulary diagnostic (113 types). **Top:** 35/113 types (24% of gold instances) receive zero recall. **Bottom:** gold and negative score distributions overlap heavily (gap = 0.30), making threshold selection brittle.

and Li, 2021). *Joint-encoding* methods (Boylan et al., 2025) concatenate all candidate labels with the input and encode them in a single Transformer pass, scoring every entity-pair-label combination at once—achieving strong Macro F1 at 254–300 sent/s in our reproduction, against the 1.4–27.8 reported for a pairwise method on other hardware (Table 2).

The standard ZSRE protocol evaluates on small held-out pools of 5–15 candidate types, under which joint-encoding methods perform well. Yet real deployments may present dozens to hundreds of candidates simultaneously. Under a *full-vocabulary diagnostic* that evaluates against all 113 Wiki-ZSL types at once (Figure 1)—a controlled stress test on the training ontology rather than a claim about open-world deployment—we uncover two structural pitfalls: (i) **Recall collapse**—35/113 types (covering 24% of gold instances) receive zero recall; mean per-relation recall is only

* Corresponding author.

24.1%. **(ii) Confidence ambiguity**—gold-positive and negative scores overlap heavily (gap = 0.30); even the best threshold yields only 18.2% Macro F1 vs. 75.9% on the $m=15$ pool. The root cause of (i) is that unnormalized dot-product scores are not scale-invariant: absolute magnitudes learned with ≤ 25 training candidates do not transfer to 113 types, systematically suppressing certain relations. The root cause of (ii) is that per-instance BCE does not constrain the global geometry of the representation space, leaving semantically close types (e.g. *instance of* vs. *part of*) entangled. These two pitfalls are *independent*: coarse scoring alone cuts zero-recall types from 35 to 22 yet leaves best-threshold Macro F1 at 18.2%; contrastive alignment separates scores but cannot recover types the fine scorer never surfaces (§4.4). Naive alternatives are also unsafe. Directly Z-score-normalizing $\mathbf{S}_{\text{fine}}$ turns BCE’s $|\mathcal{R}|$ independent binary decisions into a zero-sum competition, converts the absolute no-relation signal into a relative least-similar ranking, and breaks the pair–label inner-product alignment underlying zero-shot transfer. Replacing BCE with supervised contrastive learning is similarly problematic: it provides no per-(pair, label) decision score at inference, has no natural way to handle no-relation samples, and operates in a projection space that no longer maintains pair–label alignment. The challenge is therefore *how* to inject scale-invariance and global geometry *without* overwriting what the baseline does correctly.

Both pitfalls are therefore *distributional*: one stems from a scale-dependent score distribution, the other from an entangled gold/negative distribution. We propose **SCAR** (Stabilized Cascade and Alignment for Robust joint-encoding ZSRE), a *distributional stabilization* approach that keeps the baseline pathway intact and attaches two lightweight auxiliary branches that directly target the two root causes. To restore comparability across the score distribution, the *Coarse-to-Fine Cascade Augmentation* (CCA) branch attaches a gradient-isolated dual encoder whose Z-score normalization converts absolute scores into per-sample relative rankings within the candidate set, so a type drowned out by raw dot products surfaces once it ranks highly among its own candidates. To reshape the representation geometry, the *Interactive Supervised Contrastive Learning* (ISCL) branch adds a training-only contrastive objective that explicitly pulls same-relation representations together and pushes semantically close types apart, with

label-embedding-weighted hard negatives that target precisely the confusable pairs BCE leaves entangled. The qualifier *interactive* reflects the cross-attention from each entity pair to candidate label representations, distinguishing it from standard projection-based SupCon adaptations. Naïve integration is highly unstable (up to -23.3 pp; Table 3), so SCAR stabilizes both branches with gradient isolation, Z-score normalization, and curriculum warmup. ISCL is removed at inference ($\sim 0.1\%$ parameter, $\sim 5\%$ throughput overhead).

To our knowledge, this is the first systematic diagnosis showing that recall collapse and confidence ambiguity are independent pitfalls requiring distinct solutions. Across 30 runs on Wiki-ZSL and FewRel, SCAR improves Macro F1 by $+3.6$ to $+9.7$ pp and reduces the cross-run standard deviation by 49.8%, lifting the worst-case Baseline split from 69.24% to 77.88% ($+8.6$ pp worst-case gain). Our contributions:

- The **CCA branch**: gradient-isolated dual-encoder scoring with Z-score normalization that addresses recall collapse (§3.2).
- The **ISCL branch**: training-only contrastive alignment with label-aware anchoring and curriculum warmup that addresses confidence ambiguity (§3.3).
- **Experiments** on Wiki-ZSL and FewRel (30 runs): $+3.6$ – 9.7 pp Macro F1, zero-recall types $35 \rightarrow 1$, each stabilization mechanism proven essential via ablations, and a leakage-free protocol where the margin grows to $+11.0$ pp (§4).

Code, configurations, and split seeds: <https://github.com/qinglushengliangye/SCAR>.

2 Related Work

Zero-shot relation extraction. ZSRE methods span reading comprehension (Levy et al., 2017), textual entailment (Obamuyide and Vlachos, 2018), dual-encoder matching (Chen and Li, 2021), multiple-choice cross-encoding (Lan et al., 2023; Möller and Usbeck, 2024), prompt-based generation and discrimination (Chia et al., 2022; Lv et al., 2023), and more recent directions including schema alignment, multi-grained matching, customized representation learning, and prompt-contrastive learning (Li et al., 2024c,b; Zhang et al., 2025; Zhong et al., 2026). NLI-based reformulations (Obamuyide and Vlachos, 2018; Sainz et al., 2021) cast relation classification as entail-

ment over verbalized hypotheses; they generalize well to unseen labels but require hypothesis templates, are limited to sentence-level classification rather than span-level extraction, and their performance under large candidate sets remains underexplored. All these lines target *small-pool unseen-relation generalization*; direct cross-method comparison is nontrivial due to differences in backbone, split construction, and auxiliary encoders, so Table 1 reports only systems under comparable settings. GLiREL (Boylan et al., 2025) jointly encodes all candidate labels with the input in a single pass, a joint-encoding paradigm inspired by GLiNER (Zaratiana et al., 2024). Neither GLiNER nor GLiREL incorporates coarse-to-fine cascade scoring or contrastive alignment; we build on GLiREL and, to our knowledge, are the first to augment this joint-encoding paradigm with stabilized auxiliary branches, and also the first to expose recall collapse and confidence ambiguity as *paradigm-level* pitfalls rather than per-model artifacts.

LLM-based approaches. LLMs (Li et al., 2024a, 2025) offer flexible prompting, but the LLM data point in our controlled comparison shows a marked precision–recall imbalance. On Wiki-ZSL $m=15$, GPT-4o reaches 81% precision and only 32% recall—a $2.5\times$ asymmetry (Table 1). Generation-augmented retrieval (Li et al., 2025), which prompts an LLM to generate content that then drives retrieval, narrows the gap at the cost of additional generation and retrieval stages. Our approach stays inside the joint-encoding paradigm.

Dual encoders and contrastive learning. Dual encoders (Karpukhin et al., 2020; Reimers and Gurevych, 2019) and poly-encoders (Humeau et al., 2020) trade interaction depth for efficiency; CLIP (Radford et al., 2021) popularized learnable-temperature cosine scoring. Standard coarse-to-fine pipelines hard-prune candidates via Top- K ; in ZSRE, uncalibrated coarse scorers risk discarding low-confidence unseen relations before the joint encoder can recover them—CCA avoids this through soft additive fusion (§3.2). Supervised contrastive learning (Khosla et al., 2020) has been carried into NLP through supervised fine-tuning objectives (Gunel et al., 2021) and contrastive sentence-embedding objectives (Gao et al., 2021). In our experiments, applying it *directly* to the joint-encoding ZSRE pipeline degrades Macro F1 by up to -20.4 pp (worst run; Table 3), showing that the integration strategy—label-aware anchoring, cur-

riculum warmup, and inference-time removal—is as important as the loss itself (§3.3).

3 Method

Figure 2 previews SCAR, which stabilizes two distributions at once: the CCA branch (§3.2) restores cross-relation comparability of the score distribution via a Z-score-normalized parallel coarse pathway (Figure 2a), attacking recall collapse; the ISCL branch (§3.3) reshapes the representation geometry via training-only contrastive alignment removed at inference (Figure 2c), attacking confidence ambiguity. Both branches are lightweight, modular, and stabilized by gradient isolation, Z-score normalization, and curriculum warmup (Table 3).

3.1 Preliminaries and Baseline

Given text s , entity pair (e_h, e_t) , and candidate relation set $\mathcal{R}$, zero-shot RE predicts $r \in \mathcal{R} \cup \{\text{NoRel}\}$, where $\mathcal{R}$ may include types unseen during training. The joint-encoding baseline (Boylan et al., 2025) concatenates each r_l as a [REL] r_l prefix and jointly encodes the sequence with DeBERTa-v3-large (He et al., 2023) (1024-dim outputs projected to $d=768$), producing label representations $\mathbf{H}_{\text{type}} \in \mathbb{R}^{|\mathcal{R}| \times d}$ via an FFN over each [REL] token and entity-pair representations $\mathbf{H}_{\text{pair}} \in \mathbb{R}^{P \times d}$ (P : candidate pairs per sentence) via a BiLSTM over the head/tail entity-marker spans. Pair–label scores are unnormalized dot products:

$$\mathbf{S}_{\text{fine}} = \mathbf{H}_{\text{pair}} \cdot \mathbf{H}_{\text{type}}^{\top} \in \mathbb{R}^{P \times |\mathcal{R}|}, \quad (1)$$

trained with a per-instance weighted BCE ($w_+=2.0$, $w_-=1.0$); at inference $\sigma(\mathbf{S}_{\text{fine}}[p, l]) > t$ selects positives and the pair receives NoRel if none passes. Unnormalized scoring and the per-instance objective are precisely the structural sources of recall collapse and confidence ambiguity at full-vocabulary scale (§1).

Design philosophy: auxiliaries, not replacements. As argued in §1, directly normalizing $\mathbf{S}_{\text{fine}}$ or swapping BCE for SupCon would each destroy a baseline property (per-instance decisions, NoRel handling, or pair–label inner-product alignment). SCAR therefore keeps $\mathbf{S}_{\text{fine}}$ and BCE intact and attaches two stabilized auxiliaries: CCA injects a parallel normalized scoring stream bounded by a small α_{max} (0.05 on Wiki-ZSL, 0.5 on FewRel; Table 6), while ISCL acts only during training and is removed at inference. The guard-rails (gradient isolation, Z-score normalization, curriculum warmup)

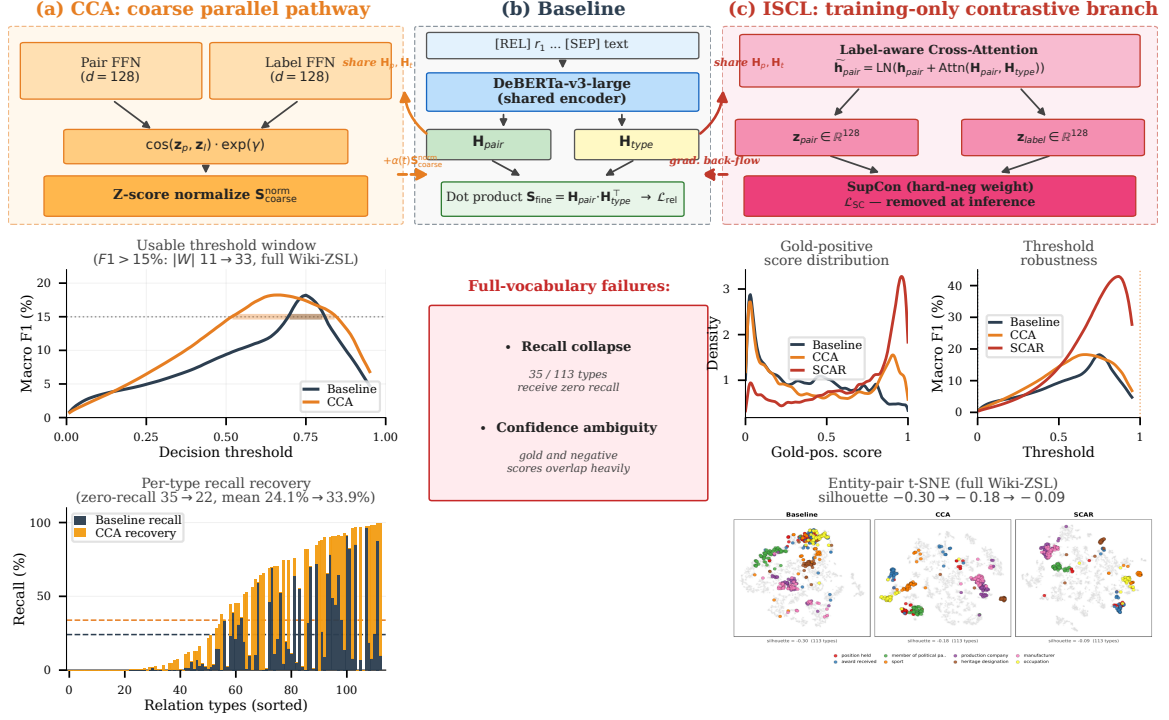


Figure 2: SCAR framework. All three modules share the DeBERTa-v3-large backbone and reuse $H_{\text{pair}}, H_{\text{type}}$. (a) **CCA**: gradient-isolated coarse scoring with Z-score normalization, merged into S_{fine} via curriculum-warmed $\alpha(t)$; inset plots show the widened usable threshold window (top) and per-relation recall recovery (bottom). (b) **Baseline**: fine-grained dot-product scoring; under full-vocabulary evaluation it suffers *recall collapse* and *confidence ambiguity*. (c) **ISCL**: training-only contrastive branch (label-aware cross-attention $\rightarrow$ SupCon with hard-negative weighting); removed at inference. Insets show gold-score separation (top-left), wider Macro F1 plateau (top-right), and tighter t-SNE clusters (bottom).

are empirically necessary: removing any single one costs up to -23.3 pp Macro F1 (Table 3).

3.2 CCA: Coarse-to-Fine Cascade Augmentation

CCA augments the fine-grained scorer with a parallel coarse scoring pathway. Two independent two-layer FFNs ($d=768 \rightarrow 2d_r \rightarrow d_r$, LayerNorm-GELU-Dropout_{0.1}; $\sim 0.1\%$ of backbone params) project h_{pair} and $h_{\text{type}}^{(l)}$ into a shared d_r -D space, and a learnable CLIP-style temperature (Radford et al., 2021) scales their cosine similarity:

$$\begin{aligned} \mathbf{v}_{\text{rel}} &= \ell_2\text{-norm}(f_{\text{rel}}(\mathbf{h}_{\text{pair}})), \\ \mathbf{v}_{\text{label}} &= \ell_2\text{-norm}(f_{\text{label}}(\mathbf{h}_{\text{type}})), \end{aligned} \quad (2)$$

$$S_{\text{coarse}}^{(b,p,l)} = (\mathbf{v}_{\text{rel}}^{(b,p)} \cdot \mathbf{v}_{\text{label}}^{(b,l)}) \cdot \exp(\gamma), \quad (3)$$

with γ initialized to $\ln(1/0.07)$ and clamped to $[1.0, 30.0]$. Coarse scores are calibrated per-sample by Z-score normalization along the *relation* axis (with $\mu_{bp} = \frac{1}{|\mathcal{R}|} \sum_l S_{\text{coarse}}^{(b,p,l)}$, $\sigma_{bp} = \text{std}_l S_{\text{coarse}}^{(b,p,l)}$):

$$[S_{\text{coarse}}^{\text{norm}}]_{bpl} = \text{clamp}\left(\frac{S_{\text{coarse}}^{(b,p,l)} - \mu_{bp}}{\sigma_{bp} + \epsilon}, -6, 6\right), \quad (4)$$

Intuitively, Z-score normalization converts absolute scores into *relative rankings* within each candidate set, so a type drowned out under raw dot products can surface if it ranks highly among its own candidates—directly addressing the scale mismatch identified in §1 (removing Z-score costs $-2.82/-8.73$ pp; Table 3). The result is additively fused into the fine pathway with a curriculum-warmed weight:

$$\begin{aligned} S_{\text{final}} &= S_{\text{fine}} + \alpha(t) S_{\text{coarse}}^{\text{norm}}, \\ \alpha(t) &= \min(t/T_w, 1) \alpha_{\text{max}}, \end{aligned} \quad (5)$$

where the dataset-specific values of d_r , α_{max} , and T_w are listed in Table 6. Training starts in pure fine mode ($\alpha(0)=0$); the coarse signal is phased in only after the dual encoders have stabilized.

An auxiliary InfoNCE-style loss monitors the coarse pathway:

$$\mathcal{L}_{\text{ret}} = -\frac{1}{|\mathcal{P}|} \sum_{(b,p)} \log \frac{\exp(S_{\text{cn}}^{(b,p,y_{bp})})}{\sum_l \exp(S_{\text{cn}}^{(b,p,l)})}, \quad (6)$$

where $S_{\text{coarse}}^{\text{norm}}$ is detached before entering $\mathcal{L}_{\text{ret}}$, so the auxiliary loss contributes *no* gradient to any

model parameter. The dual encoders are updated *only* through the main loss via eq. (5). This isolation is the most critical design choice in CCA: removing it costs a mean -13.2 pp on Wiki-ZSL (worst run -23.3 pp; Table 3). The CCA-stage objective is $\mathcal{L} = \mathcal{L}_{\text{rel}} + \lambda_{\text{ret}} \mathcal{L}_{\text{ret}}$, with λ_{ret} set per dataset (Table 6); the full three-term objective appears in §3.3.

3.3 ISCL: Interactive Supervised Contrastive Learning

The primary BCE loss optimizes per-instance classification but does not constrain the *global geometry* of the representation space: semantically similar types (e.g. *instance of* vs. *part of*) can remain entangled, producing the overlapping score distributions observed in §1. ISCL adds a training-only contrastive branch that pulls same-relation representations together and repels confusable negatives, explicitly reshaping the embedding space so that gold and negative scores separate. It differs from vanilla SupCon (Khosla et al., 2020) along three axes specific to joint-encoding ZSRE and, to our knowledge, absent from prior contrastive adaptations in NLP (Gunel et al., 2021; Gao et al., 2021):

- **Label-aware anchors:** anchors are produced by cross-attention from entity-pair to candidate-label representations (Equation (7)), not by independent projection;
- **Label-embedding hard negatives:** negative weights w_{ia} are derived from the cosine similarity of *label* embeddings (Equation (9)), exploiting the natural-language structure of relation descriptions rather than batch-level mining;
- **In-batch global-label-ID positives:** same-relation samples *within* each batch are aligned through a global relation-label ID mapping, so that the positive set is not fragmented by variation in the surface label string. No memory bank or cross-batch mechanism is involved.

We formalize each component below. Each entity-pair representation first attends to all candidate labels in the batch (single-head, residual, LN):

$$\begin{aligned} \mathbf{A} &= \text{softmax}\left(\mathbf{W}_Q \mathbf{h}_{\text{pair}} \cdot (\mathbf{W}_K \mathbf{H}_{\text{type}})^\top / \sqrt{d}\right), \\ \tilde{\mathbf{h}}_{\text{pair}} &= \text{LN}(\mathbf{h}_{\text{pair}} + \mathbf{W}_O \mathbf{A} \mathbf{W}_V \mathbf{H}_{\text{type}}), \end{aligned} \quad (7)$$

with $\mathbf{W}_{\{Q,K,V,O\}} \in \mathbb{R}^{d \times d}$. Two projection heads (architecture as in CCA, $d_c=128$) produce ℓ_2 -normalized $\mathbf{z}_{\text{pair}}$ from $\tilde{\mathbf{h}}_{\text{pair}}$ and $\mathbf{z}_{\text{label}}$ from $\mathbf{H}_{\text{type}}$.

This label-aware conditioning is critical: replacing it with a direct projection of $\mathbf{h}_{\text{pair}}$ degrades F1 by $-6.15/-9.46$ pp on Wiki-ZSL/FewRel (Table 3).

Relation strings are mapped to *global* IDs so that same-label samples across the batch are not falsely separated: $\text{pos_mask}_{i,j} = \mathbb{K}[\text{gid}(y_i)=\text{gid}(y_j)] \wedge (i \neq j)$. Without this alignment, F1 drops by $-11.73/-8.61$ pp on Wiki-ZSL/FewRel (Table 3).

For anchor i with positives $P(i)$ and negatives $N(i)$,

$$\mathcal{L}_{\text{SC}} = \frac{1}{|I_+|} \sum_{i \in I_+} \left[-\frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{a \neq i} w_{ia} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \right], \quad (8)$$

$\tau=0.1$, $I_+ = \{i : |P(i)| \geq 1\}$, with weights emphasizing label-embedding-similar negatives:

$$w_{ia} = \begin{cases} 1 + \beta \max(0, \mathbf{z}_{y_i}^{\text{lab}} \cdot \mathbf{z}_{y_a}^{\text{lab}}), & a \in N(i) \\ 1, & \text{else} \end{cases} \quad (9)$$

$\beta=1.0$. Deriving weights from *label-embedding* similarity exploits the natural-language structure of relation descriptions; empirical validation appears in the case study (appendix).

The contrastive weight follows a curriculum: $\lambda^{\text{sup}}(t) = \min(t/T_w^{\text{sup}}, 1) \lambda_{\text{max}}^{\text{sup}}$ with $\lambda_{\text{max}}^{\text{sup}}=0.05$ and $T_w^{\text{sup}}=1500$. Disabling the warmup (i.e., applying $\lambda_{\text{max}}^{\text{sup}}$ from step 0) degrades F1 by a mean $-11.5/-5.3$ pp on Wiki-ZSL/FewRel, with a worst-run drop of -20.4 pp on Wiki-ZSL (Table 3).

The full SCAR training objective is $\mathcal{L} = \mathcal{L}_{\text{rel}} + \lambda_{\text{ret}} \mathcal{L}_{\text{ret}} + \lambda^{\text{sup}}(t) \mathcal{L}_{\text{SC}}$. At inference, ISCL is removed entirely; only CCA’s additive fusion persists ($\sim 0.1\%$ parameter overhead, $\sim 5\%$ throughput overhead; Table 2).

4 Experiments

4.1 Setup

We evaluate on two standard ZSRE benchmarks: **Wiki-ZSL** (Chen and Li, 2021) (94,383 raw sentences, 113 relation types; 93,399 after preprocessing¹) and **FewRel** (Han et al., 2018), using the public 80-relation portion standard in ZSRE (56,000 instances; Chen and Li, 2021). Both provide candidate relation labels as natural-language strings.

Following Boylan et al. (2025), we denote by m the number of relation types held out from train-

¹Preprocessing retains only records that can be converted into a span-level format with at least one valid relation, resolvable entity spans, and a mapped Wikidata property label.

ing and used as unseen candidate labels at evaluation time (e.g. Wiki-ZSL $m=15$ trains on 98 types and holds out 15). We report results for $m \in \{5, 10, 15\}$, where larger m represents a harder task. For each dataset and m , every method is evaluated over **five independent runs**, each with a different random relation split, and we report the mean and standard deviation of Macro F1. Five independent experiments per method ensure that reported means are not contingent on any single favourable split; unpaired statistical tests are used accordingly (see Limitations).

All models use DeBERTa-v3-large (He et al., 2023) as the backbone ($d = 768$). We compare three configurations: (i) **Baseline**: GLiREL reproduction with its released default hyperparameters; (ii) **CCA**: Baseline + Coarse-to-Fine Cascade Augmentation; (iii) **SCAR**: CCA + ISCL (full framework). The Baseline reproduces GLiREL with $\text{lr}_{\text{enc}}=10^{-5}$, $\text{lr}_{\text{other}}=10^{-4}$, and warmup ratio 0.1. On Wiki-ZSL, CCA and SCAR use lower learning rates ($\text{lr}_{\text{enc}}=2 \times 10^{-6}$, $\text{lr}_{\text{other}}=2 \times 10^{-5}$, warmup ratio 0.03) to stabilize the auxiliary modules; on FewRel, all three configurations share the Baseline learning rates. A fully crossed control (three splits per cell) shows each configuration is best at its own rate: the Baseline drops from 75.9 to 70.7 under the CCA/SCAR rates, while CCA and SCAR drop to 63.1 and 64.5 under the Baseline rates. The gain is therefore not a learning-rate artifact. All runs use AdamW, cosine LR schedule, batch size 24 (Wiki-ZSL) or 32 (FewRel), and early stopping with patience 12 (Macro F1). All variants share the same backbone, optimizer, and training budget. Full hyperparameters are listed in the appendix. Our Baseline reproduces GLiREL from its released codebase and default hyperparameters, except the batch size, raised from the released 8 to 24/32 for our hardware. Boylan et al. (2025) likewise average over five random relation splits, but the partitions are not recoverable from the released configurations, so we re-run the Baseline under our own five-split protocol, letting Baseline, CCA and SCAR share identical data, splits, optimizer and budget.

Following the evaluation protocol of Chen and Li (2021); Möller and Usbeck (2024); Boylan et al. (2025), the held-out relation split serves as both dev and test set; the best Macro F1 checkpoint over a threshold grid $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.6\}$ is saved. Our conclusions rely on *relative* comparisons under this same protocol, supported by five

independent relation splits and permutation significance tests (see Limitations for discussion). The full-vocabulary diagnostic (§4.4) uses a broader sweep over $[0.01, 0.95]$. We compare against RelationPrompt (Chia et al., 2022), DSP-ZRSC (Lv et al., 2023), ZSRE (Tran et al., 2023), MC-BERT (Lan et al., 2023), TMC-BERT (Möller and Usbeck, 2024), and GPT-4o (Boylan et al., 2025).

4.2 Main Results

Table 1 reports Macro P/R/F1 averaged over 5 independent runs (each run a different random relation split) on both benchmarks.

CCA improves F1 in all 6 settings, by +3.4 to +8.8 pp. The gain is recall-led, with precision rising as well. On Wiki-ZSL $m=10$ recall climbs from 79.17 to 86.73 (+7.6 pp) and precision from 84.34 to 92.27 (+7.9 pp). The largest absolute gain is on Wiki-ZSL $m=5$ (+8.8 pp), where cross-run std also drops sharply ($7.13 \rightarrow 2.95$).

SCAR is the top performer in 3 of 6 settings and second-best in 2 more; together with CCA (which wins FewRel $m=15$), our methods take first place in 4 of 6 settings. On top of CCA’s recall, ISCL further raises precision (e.g. 93.71% vs. CCA’s 92.27% on Wiki-ZSL $m=10$); qualitative case studies (appendix) confirm that this gain comes from suppressing misclassification between semantically close relation types (e.g. *instance of* vs. *part of*). Mean cross-run std drops from 5.04 (Baseline) to 2.53 (SCAR), a **49.8%** reduction. Even the worst SCAR run across all 30 (77.88%, Wiki-ZSL $m=15$) still sits **8.6 pp** above the worst Baseline run (69.24%), indicating improved worst-case robustness across relation splits.

SCAR trails CCA in two of six settings: FewRel $m=15$ (−0.82 pp) and Wiki-ZSL $m=5$ (−0.46 pp), both within one std. FewRel contains one entity pair per sentence (Han et al., 2018) vs. ≈ 5 in Wiki-ZSL, yielding a $5 \times$ smaller effective contrastive batch (≈ 24 vs. 120 anchors). SCAR still beats Baseline by +6.07 pp there, and leads CCA on the remaining four settings. This provides a practical deployment guideline: the full SCAR framework is most beneficial on multi-pair corpora common in real-world scenarios, while CCA alone suffices on single-pair inputs where contrastive batch size is limited.

Our methods (CCA or SCAR) achieve the best reported F1 in 4 of 6 settings. At $m=15$, SCAR exceeds TMC-BERT by +9.3/+5.6 pp on Wiki-ZSL/FewRel and GPT-4o by +41.5/+15.9 pp.

Table 1: Macro P / R / F1 (%) on Wiki-ZSL and FewRel. Prior results are quoted from their original papers (GPT-4o and GLiREL* from [Boylan et al., 2025](#)); GLiREL* also pretrains on synthetic data, which we do not. Our variants run over 5 independent splits ($F1_{\pm std}$). Best per m in **bold**.

m	Model	Wiki-ZSL			FewRel		
		P	R	F1	P	R	F1
5	RelationPrompt (Chia et al., 2022)	70.66	83.75	76.63	90.15	88.50	89.30
	DSP-ZRSC (Lv et al., 2023)	94.10	77.10	84.80	93.40	92.50	92.90
	ZSRE (Tran et al., 2023)	94.50	96.48	95.46	96.36	96.68	96.51
	MC-BERT (Lan et al., 2023)	80.28	84.03	82.11	90.82	90.13	90.47
	TMC-BERT (Möller and Usbeck, 2024)	90.11	87.89	88.92	93.94	93.30	93.62
	GPT-4o (Boylan et al., 2025)	91.24	72.07	80.03	96.75	83.05	89.20
	GLiREL* (Boylan et al., 2025)	89.41	80.67	83.28	96.84	93.41	94.20
	Baseline (ours)	90.68	82.63	85.60 ± 7.1	93.82	84.78	86.89 ± 7.3
	CCA (ours)	96.93	92.65	94.41 ± 2.9	95.71	93.67	94.48 ± 3.6
	SCAR (ours)	96.04	92.31	93.95 ± 1.8	97.03	95.14	96.02 ± 2.1
10	RelationPrompt (Chia et al., 2022)	68.51	74.76	71.50	80.33	79.62	79.96
	DSP-ZRSC (Lv et al., 2023)	80.00	74.00	76.90	80.70	88.00	84.20
	ZSRE (Tran et al., 2023)	85.43	88.14	86.74	81.13	82.24	81.68
	MC-BERT (Lan et al., 2023)	72.81	73.96	73.38	86.57	85.27	85.92
	TMC-BERT (Möller and Usbeck, 2024)	81.21	81.27	81.23	84.42	84.99	85.68
	GPT-4o (Boylan et al., 2025)	77.62	66.14	68.35	84.07	58.00	66.20
	GLiREL* (Boylan et al., 2025)	89.87	81.56	83.67	91.09	87.42	87.60
	Baseline (ours)	84.34	79.17	80.14 ± 2.3	88.90	84.84	86.10 ± 2.2
	CCA (ours)	92.27	86.73	88.88 ± 3.7	91.89	88.54	89.52 ± 3.3
	SCAR (ours)	93.71	87.81	89.87 ± 2.1	91.94	88.89	89.73 ± 2.5
15	RelationPrompt (Chia et al., 2022)	63.69	67.93	65.74	74.33	72.51	73.40
	DSP-ZRSC (Lv et al., 2023)	77.50	64.40	70.40	82.90	78.10	80.40
	ZSRE (Tran et al., 2023)	64.68	66.01	65.30	66.44	69.29	67.82
	MC-BERT (Lan et al., 2023)	65.71	67.11	66.40	80.71	79.84	80.27
	TMC-BERT (Möller and Usbeck, 2024)	73.62	74.07	73.77	82.11	79.93	81.00
	GPT-4o (Boylan et al., 2025)	81.04	32.06	41.57	84.42	65.76	70.70
	GLiREL* (Boylan et al., 2025)	79.44	74.81	73.91	88.14	84.69	84.48
	Baseline (ours)	83.79	74.47	75.88 ± 5.1	84.66	80.32	80.50 ± 6.3
	CCA (ours)	86.86	81.13	82.40 ± 4.4	90.41	86.55	87.39 ± 3.6
	SCAR (ours)	85.84	81.67	83.04 ± 4.1	88.68	86.06	86.57 ± 2.5

The two exceptions are both on $m=5$, where ZSRE ([Tran et al., 2023](#)) leads by 0.5–1.1 pp. With only 5 candidate types, recall collapse and confidence ambiguity—the very failures SCAR targets—are mild, so the auxiliary branches offer less room for improvement. Table 2 reports inference throughput at $m=10$: SCAR reaches 285.0 and 254.2 sent/s on Wiki-ZSL and FewRel, $\sim 5\%$ below its own Baseline, whereas pairwise TMC-BERT is reported at 1.41 and 27.81 sent/s on other hardware. At $m=10$ and $m=15$ SCAR leads ZSRE by +3.1/+17.7 pp on Wiki-ZSL and +8.1/+18.7 pp on FewRel.²

Pooling all 30 runs, mean F1 improves by +7.00/+7.35 pp for CCA/SCAR; bootstrap 95% CIs are [+4.01, +9.95]/[+4.38, +10.14] pp, and a permutation test rejects equal means at $p \leq 10^{-4}$

²Per-run tables in the appendix show that even our *worst* SCAR run exceeds every prior system in Table 1 on Wiki-ZSL at $m=15$ and on FewRel at $m=10$.

Table 2: Macro F1 and speed (sent/s), both at $m=10$. F1 as in Table 1. Upper-block speeds are quoted from the $m=10$ GPU rows of [Boylan et al. \(2025, Table 3\)](#); ours average 3 timed runs on one NVIDIA GeForce RTX 4090 D (48 GB; batch 24). The blocks used different hardware, so speeds are not comparable across them.

Model	Wiki-ZSL		FewRel	
	F1	sent/s	F1	sent/s
RelationPrompt	71.50	2.06	79.96	1.72
TMC-BERT	81.23	1.41	85.68	27.81
GLiREL*	83.67	47.60	87.60	41.07
Baseline (ours)	80.14	299.8	86.10	265.5
CCA (ours)	88.88	289.9	89.52	259.6
SCAR (ours)	89.87	285.0	89.73	254.2

in aggregate (per setting, SCAR reaches $p < 0.05$ in five of six, FewRel $m=15$ excepted at $p=0.10$). Across all 6 dataset– m combinations, both CCA and SCAR consistently outperform the Baseline in Macro P, R, and F1.

Table 3: Ablation Δ Macro F1 (% , $m=15$): paired Δ = ablated – full per split over 3 runs; *worst* is the most negative, so the largest single-guard-rail cost is -23.3 pp. Negative Δ means the component helps.

Configuration	Wiki-ZSL		FewRel	
	mean	worst	mean	worst
<i>CCA components</i>				
w/o Z-score norm.	-2.82 ± 1.15	-3.65	-8.73 ± 5.65	-14.53
w/o gradient isolation	-13.23 ± 9.27	-23.26	-5.81 ± 5.08	-11.42
<i>ISCL components (atop CCA)</i>				
w/o cross-attention	-6.15 ± 3.68	-10.16	-9.46 ± 8.67	-15.11
w/o global alignment	-11.73 ± 8.17	-19.35	-8.61 ± 7.08	-15.37
w/o curriculum warmup	-11.51 ± 11.96	-20.35	-5.28 ± 6.32	-10.52

4.3 Ablation Study

Table 3 reports each ablation’s paired Δ over 3 runs at $m=15$. Removing gradient isolation (CCA $\rightarrow$ naive dual-encoder) and curriculum warmup (ISCL $\rightarrow$ naive contrastive integration) yield the deepest per-run drops: -23.3 pp and -20.4 pp on Wiki-ZSL, confirming that the instability diagnosed in §1 is real and that the stabilization mechanisms are essential. The large std (± 9 – 12 pp) on these rows further indicates that without stabilization the auxiliary branches introduce high variance across splits.

All 10 cells are negative in mean. *Gradient isolation* dominates CCA on Wiki-ZSL (-13.2 pp) while *Z-score normalization* dominates on FewRel (-8.7 pp); for ISCL, *cross-attention* is the heaviest contributor on FewRel (-9.5 pp). We ablate at $m=15$ to maximize discriminative power; gains at $m=5$ and $m=10$ (Table 1) provide complementary evidence.

4.4 Where Do the Gains Come From?

We trace the gains in Table 1 back to the two structural pitfalls diagnosed in the introduction. For each method, we run the best $m=15$ checkpoint on the full Wiki-ZSL corpus (93,399 samples, 113 relation types, threshold 0.5); this complements rather than replaces the zero-shot protocol in Table 1.

The Baseline assigns *zero* recall to **35** of 113 types, covering 24% of gold instances; another 28 types fall below 10% (49% of gold). Average per-relation recall is only **24.1%**. CCA cuts zero-recall types to 22 (avg 33.9%). SCAR reduces them to 1, grows the $\geq 80\%$ -recall set from 11 to **42**, and lifts average recall to **55.0%** ($+30.9$ pp); recall improves on 101 of 113 types. The 89 types below 50% Baseline recall gain $+20.0$ pp under CCA and $+40.4$ pp under SCAR on average, concentrating

the gains on the long-tail relations the Baseline collapses on (a small subset of high-recall relations regresses; see §4.6 and Figure 3).

One might suspect that the zero-recall types are simply the held-out unseen ones, but the arithmetic rules this out: the Baseline trains on 98 types with 15 held out, yet **35** collapse—so at least 20 *trained* types also fail, including well-supported relations like *occupation* (1,687 samples). The root cause is candidate-set scale mismatch: training uses ≤ 25 candidates per batch, but full-corpus evaluation pits each pair against all 113 types; unnormalized scores do not transfer across scales. CCA’s Z-score normalization (Equation (4)) supplies scale-invariance, explaining why removing it is the most damaging ablation on FewRel (-8.7 pp; Table 3). This result also confirms that the two pitfalls are genuinely independent: CCA alone cuts zero-recall types from 35 to 22 yet leaves the best-threshold Macro F1 unchanged at 18.2% (Table 14). Recovering recall does *not* resolve score overlap—the two problems require distinct mechanisms, and only ISCL’s global contrastive alignment lifts best-threshold F1 to 42.9%.

The Baseline’s gold and negative score distributions overlap heavily: the mean gap is only **0.296**, and the best threshold yields only **18.2%** Macro F1 (inside the narrow window $[0.70, 0.80]$). SCAR widens the gap to **0.560** ($+89\%$), lifts best-threshold Macro F1 to **42.9%**, and expands the $F1 > 15\%$ window $4.3\times$ in length ($0.10 \rightarrow 0.43$; $|W|$: $11 \rightarrow 44$ thresholds; Table 15), meaning that SCAR tolerates a much wider range of threshold settings at deployment time. The lift concentrates at borderline gold percentiles (p_{25} : $0.126 \rightarrow 0.424$; p_{50} : $0.372 \rightarrow 0.749$) while negative scores barely move: ISCL widens the margin rather than inflating gold scores uniformly, doubling the fraction of gold instances with score gap > 0.3 from 34.5% to 69.2%. The embedding space mirrors this separation: cosine silhouette rises from -0.30 (Baseline) to -0.09 (SCAR; Figure 2c), especially for type-confused pairs such as *instance of* vs. *part of* (cases in appendix).

An error decomposition at threshold 0.5 confirms the mechanism: SCAR doubles TP ($+106\%$) and halves FN (-52%); the rise in FP ($+38\%$) reflects SCAR’s higher gold-score distribution and shifts its optimal threshold to 0.87, at which Baseline $<$ CCA $<$ SCAR holds for both recall and F1 (appendix). Recall regresses on 12 of 113 types (-2.8 pp against a $+35.0$ pp gold-weighted net

Table 4: Reliability metrics on the full Wiki-ZSL corpus, reweighted to the true gold prevalence. Lower is better for all three. SCAR improves the proper scoring rules (Brier, NLL) while its ECE worsens, confirming that the gain is separation rather than probability calibration.

Method	ECE	Brier	NLL
Baseline	0.047	0.058	0.206
CCA	0.037	0.060	0.219
SCAR	0.059	0.046	0.169

gain); we characterize these cases in §4.6.

Separation, not probability calibration. Our effect is a change in how gold and negative scores are *distributed*, not an improvement in the reliability of the predicted probabilities. Table 4 reports ECE, Brier and NLL on the same full-vocabulary predictions, reweighted to the true gold prevalence ($\approx 7\%$). SCAR improves both proper scoring rules—Brier by 21%, NLL by 18%—while its ECE *worsens* from 0.047 to 0.059: under this class imbalance a model that rarely predicts a positive attains a low ECE while collapsing on recall, so the configuration that recovers recall is the one ECE penalizes. Post-hoc recalibrators (temperature, Platt, isotonic) cannot help either: as (near-)monotone maps they leave the ranking, and hence the zero-recall types, untouched. SCAR’s contribution is therefore distributional separation and threshold tolerance.

4.5 Leakage-Free Evaluation

Following Chen and Li (2021), the standard protocol lets the held-out relation split serve as both dev and test set. To confirm our gains do not depend on it, we additionally run a three-way, *relation-disjoint* split of Wiki-ZSL (83/15/15 relations for train/dev/test, $\approx 61\text{K}/10\text{K}/10\text{K}$ sentences on average); threshold and checkpoint are selected on dev only, and test is scored exactly once.

Table 5 reports the result over three independent splits. The advantage not only survives but grows: +10.4 pp for CCA and +11.0 pp for SCAR, against +6.5 and +7.2 pp on Table 1, and positive on all three splits (paired *t*-test, $p \approx 0.016$ for SCAR; CCA reaches only $p \approx 0.10$, not claimed significant at $n=3$). On the Baseline checkpoints the dev-selected threshold coincides exactly with the test-set oracle, so selection extracts no test information. Absolute values sit below Table 1 because this protocol trains on 83 instead of 98 relations—a capacity effect affecting all three configurations

Table 5: Leakage-free three-way relation-disjoint split (Wiki-ZSL, $m=15$; mean Macro F1 over three splits). Threshold and checkpoint are chosen on dev; test is scored once. Δ is the test-set margin over the Baseline, computed from unrounded means.

Method	dev F1	test F1	Δ
Baseline	73.5	53.5	–
CCA	83.1	64.0	+10.4
SCAR	82.0	64.5	+11.0

equally—so we treat it as a robustness check, not a replacement for the main comparison.

4.6 When and Why SCAR Fails

The regression follows directly from our own diagnosis. In 11 of the 12 cases the affected relation dominates a family sharing its entity-type signature—*drafted by* with *member of sports team*, *educated at* with *employer*, *date of death* with *date of birth*, *lyrics by* with *performer*—and under the Baseline it takes nearly all of the family’s recall while its siblings collapse to almost zero, exactly the recall collapse of §4.4. SCAR breaks that monopoly: the siblings return (*member of sports team* 0.0→94.8, *employer* 0.0→74.0, *performer* 19.3→91.6, *author* 20.9→87.1) and the dominant member gives up part of its share. The regression is therefore the direct counterpart of the recall SCAR recovers.

We identify the dominant member by each relation’s nearest neighbour in the learned representation space. For 11 of the 12 regressing relations that neighbour held the *lower* Baseline recall, by a median of 46.5 pp, against only 39% (median 0) among the 101 improving relations; in instances the 12 give up $\approx 3.8\text{K}$ while their neighbours gain $\approx 9.5\text{K}$. Relations at risk are thus identifiable in advance, and per-relation adaptive weighting is a natural remedy. Table 10 gives the full breakdown.

5 Conclusion

We identified two distributional pitfalls of joint-encoding ZSRE—*recall collapse* and *confidence ambiguity*—and proposed SCAR, which stabilizes them via a Z-score-normalized cascade and a training-only contrastive module. Across Wiki-ZSL and FewRel, SCAR consistently improves Macro F1 and cuts zero-recall types from 35 to 1; naïve coupling degrades F1 by 23.3 pp—stabilization matters as much as the signals. Full-vocabulary diagnostics merit broader adoption.

Limitations

We evaluate on two English sentence-level benchmarks (Wiki-ZSL, FewRel) under the $m \in \{5, 10, 15\}$ protocol (Chen and Li, 2021) with DeBERTa-v3-large; generalization beyond this setting (document-level, multilingual, other backbones) is unverified. Both branches attach to any joint encoder that exposes entity-pair and label representations, so SCAR is in principle backbone-agnostic, but we deliberately hold the backbone fixed to isolate the diagnosis from backbone effects, and gains on stronger encoders remain to be measured. The ISCL branch’s gain shrinks when each sentence carries only one entity pair (FewRel $m=15$: -0.82 pp vs. CCA). Following the standard ZSRE protocol (Chen and Li, 2021; Möller and Usbeck, 2024; Boylan et al., 2025), the relation-disjoint split serves as both dev and test set; we mitigate this with five independent relation splits per setting, worst-run reporting, permutation significance tests across all 30 runs ($p \leq 10^{-4}$ in aggregate; per-split F1 in appendix), and the leakage-free three-way split of §4.5. Under the full-vocabulary diagnostic, recall regresses on 12 of 113 types (8.8% of gold instances), costing -2.8 pp corpus recall against a $+35.0$ pp gold-weighted gain; these are the relations that monopolized a type-signature family under the Baseline (§4.6), and per-relation adaptive weighting or a type-conditional loss is left as future work.

Although SCAR adds only $\sim 0.1\%$ parameters at inference ($\sim 0.7\%$ during training, before ISCL is removed) and $\sim 5\%$ throughput overhead, the framework introduces two auxiliary branches with several hyper-parameters (e.g. cascade weight α , contrastive temperature τ , warmup schedule); tuning on new corpora may require additional effort. Our ablations do give transferable guidance: the two stabilizers (gradient isolation and curriculum warmup) are required, whereas the fusion weight $\alpha_{\max}$ and the contrastive loss weight did not show marked sensitivity in our preliminary tuning, though we did not sweep them systematically; and ISCL helps most on multi-pair corpora while CCA alone suffices on single-pair inputs. Finally, the full-vocabulary diagnostic is a synthetic stress test on the training-set ontology and does not directly measure generalization to truly novel relation types; its 113 relations also remain far smaller than the inventories of hundreds or thousands of evolving types found in real deployments. Its primary

value is revealing failure modes that small-pool protocols overlook.

Ethics Statement

Wiki-ZSL and FewRel are derived from Wikidata and Wikipedia, which carry known biases in demographic coverage and relation-type frequency (Mehrabi et al., 2021); per-relation audits are advised before deployment, especially for person-centric relation types. Our method does not introduce new data collection or annotation; all experiments use publicly available benchmarks. Training uses one NVIDIA GeForce RTX 4090 D (48 GB) and takes $\sim 1\text{--}3$ h/run across the ~ 120 runs behind Table 1 and Table 3, plus the 18 protocol and learning-rate control runs, a modest computational footprint.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFC3308101) and the Fundamental Research Funds for the Central Universities (2025CX11021). We thank the anonymous reviewers and the chairs of EMNLP 2026 for their constructive feedback, and the NLP community for the openly released datasets and code that made this study possible.

We used Claude and ChatGPT to polish the language of the manuscript and to help locate part of the related literature; all outputs were reviewed by the authors, who manually verified every cited reference against its original source and take full responsibility for the content of this paper.

References

- Jack Boylan, Chris Hokamp, and Demian Gholipour Ghalandari. 2025. **GLiREL - generalist model for zero-shot relation extraction**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8230–8245, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chih-Yao Chen and Cheng-Te Li. 2021. **ZS-BERT: Towards zero-shot relation extraction with attribute representation learning**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

- Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. [RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. [Supervised contrastive learning for pre-trained language model fine-tuning](#). In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.
- 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.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. [Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring](#). In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. [A survey on knowledge graphs: Representation, acquisition, and applications](#). *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. [Supervised contrastive learning](#). In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 18661–18673.
- Yuquan Lan, Dongxu Li, Yunqi Zhang, Hui Zhao, and Gang Zhao. 2023. [Modeling zero-shot relation classification as a multiple-choice problem](#). In *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Gold Coast, Australia.
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. [Zero-shot relation extraction via reading comprehension](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.
- Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024a. [Meta in-context learning makes large language models better zero and few-shot relation extractors](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6350–6358, Jeju, South Korea.
- Shilong Li, Ge Bai, Zhang Zhang, Ying Liu, Chenji Lu, Daichi Guo, Ruifang Liu, and Sun Yong. 2024b. [Fusion makes perfection: An efficient multi-grained matching approach for zero-shot relation extraction](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 79–85, Mexico City, Mexico. Association for Computational Linguistics.
- Zehan Li, Fu Zhang, and Jingwei Cheng. 2024c. [AlignRE: An encoding and semantic alignment approach for zero-shot relation extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2957–2966, Bangkok, Thailand. Association for Computational Linguistics.
- Zehan Li, Fu Zhang, Tianyue Peng, He Liu, and Jingwei Cheng. 2025. [Generation-augmented retrieval: Rethinking the role of large language models in zero-shot relation extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17928–17941, Suzhou, China. Association for Computational Linguistics.
- Bo Lv, Xin Liu, Shaojie Dai, Nayu Liu, Fan Yang, Ping Luo, and Yue Yu. 2023. [DSP: Discriminative soft prompts for zero-shot entity and relation extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5491–5505, Toronto, Canada. Association for Computational Linguistics.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. [A survey on bias and fairness in machine learning](#). *ACM Computing Surveys*, 54(6):1–35.
- Cedric Möller and Ricardo Usbeck. 2024. [Incorporating type information into zero-shot relation extraction](#). In *Joint Proceedings of the 3rd International Workshop on Knowledge Graph Generation from Text (TEXT2KG) and Data Quality Meets Machine Learning and Knowledge Graphs (DQMLKG)*, co-located

with the 21st Extended Semantic Web Conference (ESWC 2024), volume 3747 of *CEUR Workshop Proceedings*, Hersonissos, Greece.

Abiola Obamuyide and Andreas Vlachos. 2018. [Zero-shot relation classification as textual entailment](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

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.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. [Label verbalization and entailment for effective zero and few-shot relation extraction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Van-Hien Tran, Hiroki Ouchi, Hiroyuki Shindo, Yuji Matsumoto, and Taro Watanabe. 2023. [Enhancing semantic correlation between instances and relations for zero-shot relation extraction](#). *Journal of Natural Language Processing*, 30(2):304–329.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. [GLiNER: Generalist model for named entity recognition using bidirectional transformer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Fu Zhang, He Liu, Zehan Li, and Jingwei Cheng. 2025. [CE-DA: Custom embedding and dynamic aggregation for zero-shot relation extraction](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9814–9823, Abu Dhabi, UAE. Association for Computational Linguistics.

Xueyi Zhong, Liye Zhao, Licheng Peng, Guodong Yang, Kun Hu, and Wansen Wu. 2026. [Prompt-contrastive learning for zero-shot relation extraction](#). *Entropy*, 28(1):69.

A Hyperparameter Settings

Table 6 lists every hyperparameter used in the reported runs. Values are shared across Baseline, CCA and SCAR unless stated otherwise; the FewRel column records only the settings that differ from Wiki-ZSL.

B Per-Run Results

Tables 7 to 9 list the Macro F1 of every individual run behind Table 1: five independent random relation splits per method, dataset and value of m .

C Per-Run Ablation Δ

For each row of Table 3, three independent runs are drawn per (ablation, dataset) cell on Wiki-ZSL and FewRel at $m=15$. Unlike the main comparisons in Table 1 (where splits are not paired across Baseline/CCA/SCAR; see Limitations), each ablation run is matched to its corresponding full-model run on the *same* random split, so Δ = ablated – full model is reported as a paired difference. We report mean $\pm$ std of Δ across the 3 runs. The raw per-run best-F1 values for every (ablation, dataset, run) triple are written to

best_metrics.json files and are released with the code.

D Recall Collapse Evidence

The recall collapse phenomenon is evidenced by the per-run results in Tables 7 and 9: the Baseline shows the largest variance on Wiki-ZSL $m=5$ and FewRel $m=5$ (std 7.13 and 7.27; Table 9), with the worst Baseline run across all 30 reaching only 69.24% Macro F1 (Wiki-ZSL $m=15$ exp 4; Table 7), indicating that certain relation splits cause partial scoring failure.

The full-corpus confidence analysis (Table 14) further confirms this: Baseline gold-positive score is only $\mu = 0.392$, meaning many gold triplets receive scores barely above the decision threshold. CCA raises this to 0.462 and SCAR to 0.652, systematically recovering under-scored triplets across all relation types.

E Case Study

Table 11 shows representative examples where CCA recovers gold triplets that the Baseline misses entirely (confidence < 0.02), while Table 12 shows false positives that ISCL suppresses.

Table 6: Complete hyperparameter settings on Wiki-ZSL ($m=15$). The FewRel column lists values that differ from Wiki-ZSL and apply to *all three* configurations on FewRel (i.e., CCA and SCAR on FewRel also use the Baseline learning rates); “same” means the Wiki-ZSL value is reused.

Category	Hyperparameter	Baseline	CCA	SCAR	FewRel
Training	Learning rate (encoder)	1×10^{-5}	2×10^{-6}	2×10^{-6}	1×10^{-5}
	Learning rate (other)	1×10^{-4}	2×10^{-5}	2×10^{-5}	1×10^{-4}
	Warmup ratio	0.1	0.03	0.03	0.1
	Batch size	24	24	24	32
	Scheduler	Cosine with warmup			same
	Max steps	20,000	20,000	20,000	same
Model	Backbone	DeBERTa-v3-large ($d=768$)			same
	Dropout	0.4	0.4	0.4	same
	Scorer	dot	dot	dot	same
CCA	Retrieval dim d_r	–	128	128	256
	$\alpha_{\max}$	–	0.05	0.05	0.5
	Cascade warmup steps T_w	–	3,000	3,000	1,000
	Retrieval loss weight λ_{ret}	–	0.01	0.01	0.05
	Temperature init γ_0	–	$\ln(1/0.07)$	$\ln(1/0.07)$	same
ISCL	Contrastive dim d_c	–	–	128	same
	SupCon temperature τ	–	–	0.1	same
	Hard-negative weight β	–	–	1.0	same
	SupCon loss weight $\lambda_{\max}^{\text{sup}}$	–	–	0.05	same
	Curriculum warmup steps T_w^{sup}	–	–	1,500	same
	Cross-attention heads	–	–	1	same
Evaluation	Unseen relation types m	5, 10, or 15			
	Threshold search metric	Macro F1			
	Independent runs	5 random relation splits per setting			
	Early stopping patience	12 (Macro F1)			

Table 7: Per-run Macro F1 (%) for $m=15$ on Wiki-ZSL and FewRel. “exp k ” denotes the k -th independent random relation split.

	exp 1	exp 2	exp 3	exp 4	exp 5	Mean $\pm$ Std
<i>Wiki-ZSL</i>						
Baseline	79.84	81.53	76.43	69.24	72.37	75.88 $\pm$ 5.11
CCA	75.03	82.86	82.80	84.92	86.40	82.40 $\pm$ 4.39
SCAR	84.06	88.65	84.17	80.45	77.88	83.04 $\pm$ 4.10
<i>FewRel</i>						
Baseline	89.30	80.11	75.44	83.87	73.79	80.50 $\pm$ 6.32
CCA	90.28	89.58	90.08	83.60	83.43	87.39 $\pm$ 3.55
SCAR	83.59	88.96	85.27	89.47	85.57	86.57 $\pm$ 2.53

Table 8: Per-run Macro F1 (%) for $m=10$ on Wiki-ZSL and FewRel.

	exp 1	exp 2	exp 3	exp 4	exp 5	Mean $\pm$ Std
<i>Wiki-ZSL</i>						
Baseline	77.76	77.91	80.36	82.35	82.30	80.14 $\pm$ 2.25
CCA	93.46	90.05	88.56	89.04	83.28	88.88 $\pm$ 3.67
SCAR	91.03	90.27	90.40	86.19	91.44	89.87 $\pm$ 2.11
<i>FewRel</i>						
Baseline	86.08	89.37	83.53	84.90	86.63	86.10 $\pm$ 2.18
CCA	93.46	91.09	86.03	90.75	86.26	89.52 $\pm$ 3.25
SCAR	87.75	88.67	90.27	93.84	88.14	89.73 $\pm$ 2.49

The recall recovery cases (Table 11) involve relations with distinctive entity-type signatures (e.g., “production company” typically links films to studios) that the coarse dual-encoder branch captures effectively. The FP suppression cases (Table 12) are all “instance of” vs. semantically overlapping relations—precisely the type confusion category that ISCL’s hard-negative-weighted contrastive loss targets.

F Extended Analysis

This appendix provides extended diagnostics supporting §4.4. All results use the best checkpoint from each method on the complete Wiki-ZSL corpus (93,399 samples, 113 relation types, threshold 0.5 unless noted). The full-vocabulary diagnostics (score distribution, threshold robustness) and t-SNE projections of entity-pair representations are summarized in Figure 2c; here we report the supplementary error decomposition.

Table 13 reports TP, FP, and FN counts at a fixed threshold of 0.5 on the full Wiki-ZSL corpus. CCA reduces FN by 13.5% (from 89,937 to 77,821) and increases TP by 27.2%. SCAR reduces FN by 52.4% (to 42,814) and more than doubles TP (from 44,472 to 91,595), though at the cost of higher FP at this fixed threshold—consistent with its higher TP confidence pushing more predictions above 0.5 (the optimal SCAR threshold is in fact 0.87; see

Table 15). Qualitatively, the residual FPs cluster around type confusion between semantically close relations (e.g., *instance of* vs. *part of/use*; Table 12); ISCL’s label-embedding hard-negative weighting targets this category specifically.

G Threshold Sensitivity

Table 15 summarizes each method’s own optimal decision threshold on the full Wiki-ZSL corpus and the corresponding best Macro F1, verifying that the best-F1 ranking Baseline $\leq$ CCA $<$ SCAR holds when every method is allowed to pick its own threshold (the threshold robustness analysis referenced in §4.4). Figure 4 plots Macro F1 as a function of the threshold; CCA and SCAR exhibit broader high-F1 plateaus while the Baseline’s Macro F1 drops sharply away from its optimal threshold.

Table 9: Per-run Macro F1 (%) for $m=5$ on Wiki-ZSL and FewRel.

	exp 1	exp 2	exp 3	exp 4	exp 5	Mean $\pm$ Std
<i>Wiki-ZSL</i>						
Baseline	88.44	95.81	77.10	81.28	85.36	85.60 ± 7.13
CCA	97.46	96.84	90.10	93.42	94.24	94.41 ± 2.95
SCAR	96.84	94.45	93.56	92.59	92.33	93.95 ± 1.82
<i>FewRel</i>						
Baseline	83.36	81.24	97.77	90.88	81.19	86.89 ± 7.27
CCA	92.14	94.27	90.12	96.94	98.94	94.48 ± 3.55
SCAR	98.16	94.42	94.41	98.57	94.55	96.02 ± 2.14

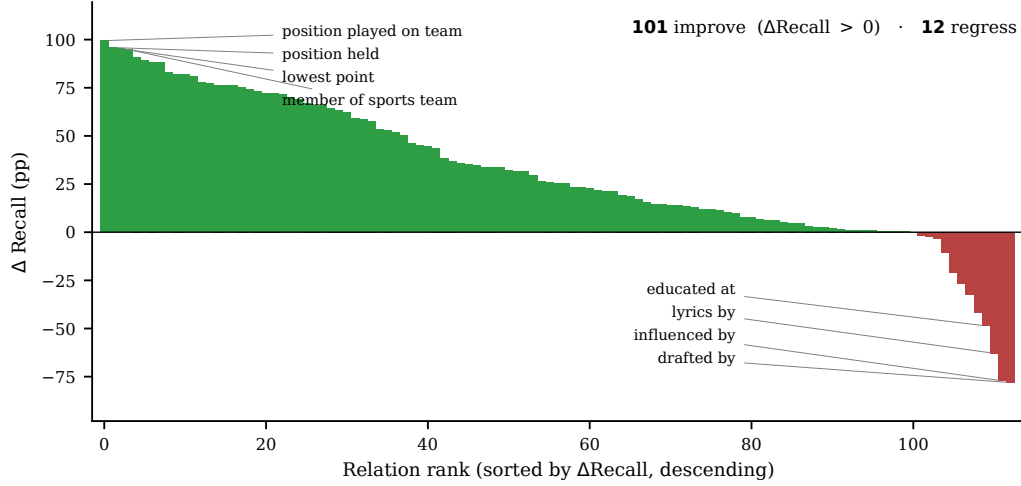


Figure 3: Per-relation Δ Recall on the full Wiki-ZSL corpus (SCAR – Baseline). Each bar corresponds to one of the 113 relation types, sorted by Δ Recall (descending). Green: improved; red: regressed. The four largest gainers (*position played on team / speciality*, *position held*, *lowest point*, *member of sports team*) move from 0% to $>94\%$ recall—these are precisely the types that the Baseline assigns zero recall, and that SCAR recovers. The four largest losses (*drafted by*, *influenced by*, *lyrics by*, *educated at*) are relations that monopolized a type-signature family under the Baseline; once cross-relation comparability is restored their previously suppressed siblings recover and they give back part of their share (§4.6). The net effect is overwhelmingly positive: 101 of 113 types improve, 12 regress.

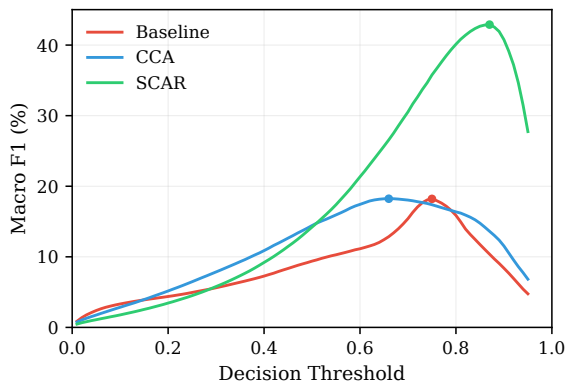


Figure 4: Macro F1 vs. decision threshold on Wiki-ZSL ($m=15$). CCA and SCAR maintain high Macro F1 over a wider threshold range than Baseline, indicating more robust deployment behavior.

Table 10: Full per-relation recall (%) on the complete Wiki-ZSL corpus (all 113 types, 134K gold instances, threshold 0.5), sorted by $\Delta\text{Recall} = \text{SCAR} - \text{Baseline}$ (descending). “Gold” is the number of gold instances of that relation. Regressions are shown in **red**. Recall improves on **101** of 113 types; the 12 regressing types carry 8.8% of gold instances and cost -2.8 pp against a $+35.0$ pp gold-weighted net gain (Section 4.6).

Relation	Gold	Base	SCAR	Δ	Relation	Gold	Base	SCAR	Δ
position played on team / spe...	1158	0.0	99.5	+99.5	genre	2252	64.4	89.6	+25.2
position held	1412	0.0	95.9	+95.9	based on	892	0.3	23.9	+23.5
lowest point	711	0.0	95.4	+95.4	child	1053	0.0	23.2	+23.2
member of sports team	1737	0.0	94.8	+94.8	cast member	2250	71.9	94.6	+22.7
located on astronomical location	1886	6.8	97.7	+90.9	official language	1180	68.6	90.4	+21.9
part of the series	1186	2.8	91.8	+89.0	occupant	898	56.1	77.2	+21.1
instrument	1221	9.8	98.3	+88.5	material used	504	74.4	95.4	+21.0
sports discipline competed in	459	0.0	88.2	+88.2	parent astronomical body	515	74.0	93.4	+19.4
league	1566	7.8	90.8	+83.0	producer	863	0.9	19.8	+18.9
platform	1476	10.9	93.0	+82.0	given name	592	0.0	17.2	+17.2
occupation	1687	0.0	81.7	+81.7	field of work	725	0.0	15.6	+15.6
subsidiary	1061	2.7	83.7	+81.0	winner	593	0.5	15.2	+14.7
family	858	0.5	78.1	+77.6	date of birth	1260	18.6	32.9	+14.4
award received	1256	21.1	98.2	+77.2	named after	1040	0.8	14.9	+14.1
original language of film or ...	1188	9.3	85.7	+76.3	subclass of	1479	0.0	14.1	+14.1
composer	1143	8.8	85.0	+76.2	connecting line	559	83.2	96.4	+13.2
headquarters location	1172	12.2	88.3	+76.1	from narrative universe	367	78.2	91.3	+13.1
languages spoken, written or ...	1184	19.0	94.3	+75.3	mouth of the watercourse	1195	46.0	57.9	+11.9
employer	1149	0.0	74.0	+74.0	heritage designation	441	87.5	99.3	+11.8
production company	624	9.9	83.2	+73.2	original broadcaster	1257	84.9	96.3	+11.4
licensed to broadcast to	1007	24.6	96.9	+72.3	location	1061	0.0	10.2	+10.2
performer	1717	19.3	91.6	+72.2	shares border with	2478	2.3	12.0	+9.8
director	1528	14.5	86.1	+71.6	organization directed from th...	417	0.0	7.9	+7.9
has quality	483	26.1	96.1	+70.0	part of	2063	3.8	11.7	+7.9
diplomatic relation	1543	2.7	71.5	+68.8	different from	609	0.0	6.6	+6.6
founded by	1133	2.1	69.2	+67.1	country	4558	79.5	85.9	+6.3
author	1408	20.9	87.1	+66.1	characters	1278	84.3	90.3	+6.0
record label	1120	13.6	79.5	+65.9	country of citizenship	3150	80.2	85.4	+5.2
developer	1333	30.8	95.3	+64.6	narrative location	453	0.0	4.9	+4.9
capital	1187	0.0	63.5	+63.5	said to be the same as	580	0.0	4.5	+4.5
instance of	3931	14.3	76.6	+62.3	industry	570	1.6	4.9	+3.3
contains administrative terri...	1034	7.8	67.2	+59.4	product or material produced	367	0.0	2.7	+2.7
participant of	1220	0.0	58.8	+58.8	main subject	669	0.0	2.4	+2.4
father	1106	0.5	58.1	+57.7	member of	1456	0.0	2.3	+2.3
located in the administrative...	2493	0.0	53.5	+53.5	residence	357	0.0	1.4	+1.4
sport	4775	44.4	97.5	+53.0	opposite of	1031	0.0	1.2	+1.2
item operated	1219	1.3	53.1	+51.8	publication date	1890	96.7	97.8	+1.2
manufacturer	1181	23.0	73.2	+50.1	applies to jurisdiction	1005	0.0	0.8	+0.8
spouse	1232	1.7	47.6	+45.9	use	377	0.0	0.8	+0.8
conflict	1292	34.4	79.5	+45.1	field of this occupation	738	0.0	0.7	+0.7
creator	1193	14.5	59.0	+44.5	facet of	453	0.0	0.7	+0.7
publisher	1258	4.5	48.2	+43.7	located in or next to body of...	1137	0.0	0.6	+0.6
religion	568	31.9	70.4	+38.5	is a list of	412	0.0	0.5	+0.5
country of origin	1422	0.0	37.0	+37.0	work location	762	0.1	0.3	+0.1
follows	1428	0.6	36.3	+35.8	owned by	1083	6.7	5.1	-1.6
notable work	1091	0.1	35.2	+35.1	constellation	632	96.5	94.0	-2.5
place of death	1446	19.6	54.1	+34.5	voice type	410	43.9	40.5	-3.4
military branch	1166	63.3	97.3	+34.0	continent	1873	94.0	83.2	-10.8
place of birth	2035	45.0	78.8	+33.8	movement	821	48.7	27.8	-20.9
student	702	0.0	33.5	+33.5	screenwriter	1144	46.4	19.6	-26.8
inception	1322	24.4	56.6	+32.1	located on terrain feature	1165	75.4	42.8	-32.6
family name	677	0.0	31.5	+31.5	date of death	1527	89.8	48.2	-41.6
mother	363	9.1	40.5	+31.4	educated at	1189	94.5	45.8	-48.7
successful candidate	1050	41.5	70.9	+29.3	lyrics by	396	63.6	0.8	-62.9
twinned administrative body	1355	0.0	26.4	+26.4	influenced by	1161	96.8	19.6	-77.2
replaces	377	0.0	26.0	+26.0	drafted by	392	78.1	0.0	-78.1
member of political party	1301	69.9	95.3	+25.4					

Table 11: CCA recall recovery: gold triplets with near-zero Baseline confidence that CCA correctly identifies. SCAR maintains the recovery.

Gold relation	Base	CCA	+ISCL
production company	0.01	0.97	0.93
member of political party	0.02	0.95	0.99
position held	0.02	0.95	0.90
part of	0.01	0.94	0.80

Table 12: ISCL precision refinement: high-confidence CCA false positives (type confusion) that ISCL suppresses. “Pred” is the incorrectly predicted relation.

Pred (wrong)	Gold (correct)	CCA	+ISCL
instance of	part of	0.93	0.01
instance of	use	0.93	0.06
instance of	use	0.93	0.10
instance of	use	0.93	0.02

Table 13: Error decomposition at threshold 0.5 on full Wiki-ZSL (134K gold triplets). $\Delta\%$ is relative to Baseline.

Method	TP		FP		FN	
	Count	Δ	Count	Δ	Count	Δ
Baseline	44,472	–	768K	–	89,937	–
CCA	56,588	+27%	593K	–23%	77,821	–13%
SCAR	91,595	+106%	1,063K	+38%	42,814	–52%

Table 14: Score statistics on full Wiki-ZSL (134K gold triplets, 113 types). Gold+ μ : mean score assigned to gold relation candidates; Neg μ : mean score assigned to non-gold candidates. Gap = Gold+ μ – Neg μ . Best F1: maximum Macro F1 over a global threshold sweep in $[0.01, 0.95]$ (an upper-bound diagnostic).

Method	Gold+ μ	Neg μ	Gap	Best F1
Baseline	0.392	0.096	0.296	18.2%
CCA	0.462	0.091	0.370	18.2%
SCAR	0.652	0.093	0.560	42.9%

Table 15: Each method’s optimal decision threshold t^* and the corresponding Macro F1 on full Wiki-ZSL ($m=15$ checkpoint). *Window* is the threshold range in $[0.01, 0.95]$ at which Macro F1 exceeds 15% (at 0.01 resolution; $|W|$ is the number of such thresholds). SCAR’s lead is preserved when every method picks its own threshold.

Method	t^*	Best F1	F1@0.5	Window	$ W $
Baseline	0.75	18.21	9.39	$[0.70, 0.80]$	11
CCA	0.66	18.23	14.43	$[0.52, 0.84]$	33
SCAR	0.87	42.90	14.21	$[0.52, 0.95]$	44