DEBAR: Mitigating Contextual Bias in Cross-Document Relation Extraction via Dual-Stream Decoupling

Zhixuan Yang, Fu Zhang*, Huangming Xu, Jingwei Cheng
School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
yangzx489@gmail.com, zhangfu@neu.edu.cn

Abstract

Cross-document Relation Extraction (CodRE) requires reasoning over scattered evidence to identify relations between target entities across multiple documents. Existing methods indiscriminately fuse target entities and the intermediate bridge entities that link them into a unified representation. This leads to intermediate evidence that often aligns with only one side of the entity pair, resulting in one-sided relation transfer contextual bias and incomplete reasoning chains. Moreover, these methods typically employ a global threshold to determine relation existence for all entity pairs, limiting the model's reasoning performance. To address these issues, we propose DEBAR (Dual-stream Entity Bias Reduction), a framework designed to explicitly decouple and preserve bidirectional bridge evidence, combined with a novel dynamic loss optimization objective. Specifically, DEBAR employs a bridge-aware input construction strategy and a dual-stream graph reasoning network to separately encode head and tail contexts, preventing semantic interference while capturing global dependencies through iterative message passing. Furthermore, we introduce a curriculum-aware ranking optimization objective that progressively tightens classification constraints to stabilize training and enforce discriminative decision boundaries. Experiments on the CodRE benchmarks show that DEBAR achieves state-of-the-art performance while effectively mitigating cross-document contextual bias. Moreover, extensive experiments on our proposed loss across backbones confirm its generalization, suggesting it as a reliable replacement for existing CodRE losses.[1]

1 Introduction

Relation Extraction (RE) is a fundamental task in information extraction (Han et al., 2020), aiming to identify the relation between a pair of target head and tail entities from text. Existing methods focus on sentence-level (Wang et al., 2016; Zhu et al., 2019; Zhou and Chen, 2022) or single-document settings (Yao et al., 2019; Zaporojets et al., 2021; Zhou et al., 2021). Specifically, in sentence-level RE, the entity pair appears within a sentence; in single-document RE, the entity pair may be distributed across sentences within a document. However, in many scenarios, target entities are often scattered across different documents. This requires the integration of information from multiple sources, giving rise to the task of **cross-document RE** (CodRE) (Yao et al., 2021).

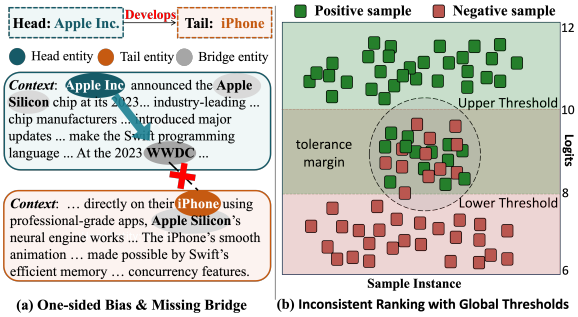

Figure 1: Illustration of relation transfer contextual bias and the limitations of global threshold optimization in existing CodRE methods.

The existing work (Yao et al., 2021) formulates CodRE as a path-based inference task, where the model is required to connect *target entities* via intermediate *bridge entities* scattered across documents. However, such path-based heuristics often suffer from noise introduced by spurious connections. To mitigate this, subsequent research typically falls into two directions: entity-centric frameworks (Wang et al., 2022; Jain et al., 2024; Yue et al., 2024) that aggregate global context across documents to suppress noise, and advanced multi-hop reasoning models (Lu et al., 2023b) that aim to distill more reliable evidence chains.

*Corresponding author.
[1]Code: https://github.com/newyuyou/DEBAR.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23804–23815
July 2-7, 2026

Despite these advances, most existing methods indiscriminately fuse the target entities and bridge entities into a unified representation. In such a space, the intermediate evidence often *aligns with only one side of the entity pair* (e.g., highly correlated with the head but irrelevant to the tail), inevitably leading to biased reasoning. As illustrated in Fig. 1, the bridge entity (WWDC) is contextually bound to the head (Apple Inc.) yet isolated from the tail (iPhone). When the model encodes this imbalanced structure, it may correctly capture the Head-Bridge relation but fail to extend this understanding to the Bridge-Tail connection. This disruption in entity connections prevents valid deduction of the target relation, introducing a relation transfer bias where semantic information fails to propagate to the tail entity.

Moreover, as a multi-label classification task, these existing CodRE methods typically rely on a global threshold to determine the existence of relations. Specifically, *a fixed pair of upper and lower thresholds is applied to all entity pairs* as shown in Fig. 1: the model is optimized to ensure that the prediction scores of positive samples exceed the lower bound, while those of negative samples remain below the upper bound. Although this interval-based strategy (i.e., setting a *tolerance margin* between the upper and lower thresholds) facilitates model optimization, it *fails to guarantee that the scores of positive samples consistently exceed those of negative ones*, which inevitably impacts the model's reasoning performance.

To address the issues, we propose a novel **D**ual-stream **E**ntity **Bia**s **R**eduction framework (**DEBAR**). DEBAR is designed to explicitly decouple and preserve bidirectional bridge evidence, combined with a novel dynamic loss optimization objective. Our framework is composed of three main components: 1) **Bridge-Aware Input Construction**. This module aims to address the imbalance where bridge entities only connect to one side of the entity pair. Specifically, we quantify the importance of each bridge entity based on its joint co-occurrence with both the head and tail entities. This importance score is then used to assess the relevance of sentences to the target entities. Ultimately, the highest-scoring sentences are selected to construct *two distinct contextual input streams* for the head and tail entities. 2) **Dual-Stream Graph Reasoning Network**. We design a dual-stream encoder to process the head and tail contexts independently, avoiding early interaction while capturing local head-bridge and tail-bridge associations. Based on these encoded representations, a global entity graph is constructed. Finally, a Graph Recurrent Network (GRN) conducts iterative message passing over the graph to synthesize the separated contexts and capture the global relationship between the target entity pair. 3) **Curriculum-Aware Ranking Optimization**. To address the limitation of fixed global thresholds, we propose a curriculum-inspired dynamic training strategy. Instead of using fixed intervals, our method gradually tightens the tolerance margin during training. Additionally, we introduce a ranking penalty to ensure that positive samples consistently outscore negative ones, resolving order ambiguity and improving inference precision. The main contributions of this work are as follows:

- We propose a bridge-aware input construction strategy that filters irrelevant context via joint co-occurrence assessment, ensuring balanced evidence retrieval for head and tail entity pairs.
- We further propose a dual-stream graph reasoning framework that captures bidirectional dependencies between the head and tail entities in a more balanced manner, effectively mitigating relation transfer bias.
- We design a novel curriculum-aware ranking optimization objective, which dynamically tightens classification constraints to stabilize training while enforcing a rigorous relative order between positive and negative samples.
- Experiments on the CodRE benchmark under closed and open settings demonstrate that DEBAR achieves state-of-the-art performance. In addition, integrating our ranking optimization objective into other backbones further enhances their performance, validating its effectiveness and generality.

2 Related Work

2.1 Document-level Relation Extraction

Document-level RE (DocRE) extends sentence-level RE by identifying relations between entities scattered across multiple sentences *within a single document* (Yao et al., 2019). Prior works primarily explore three paradigms: Transformer-based models (Zhou et al., 2021; Ma et al., 2023) that im-

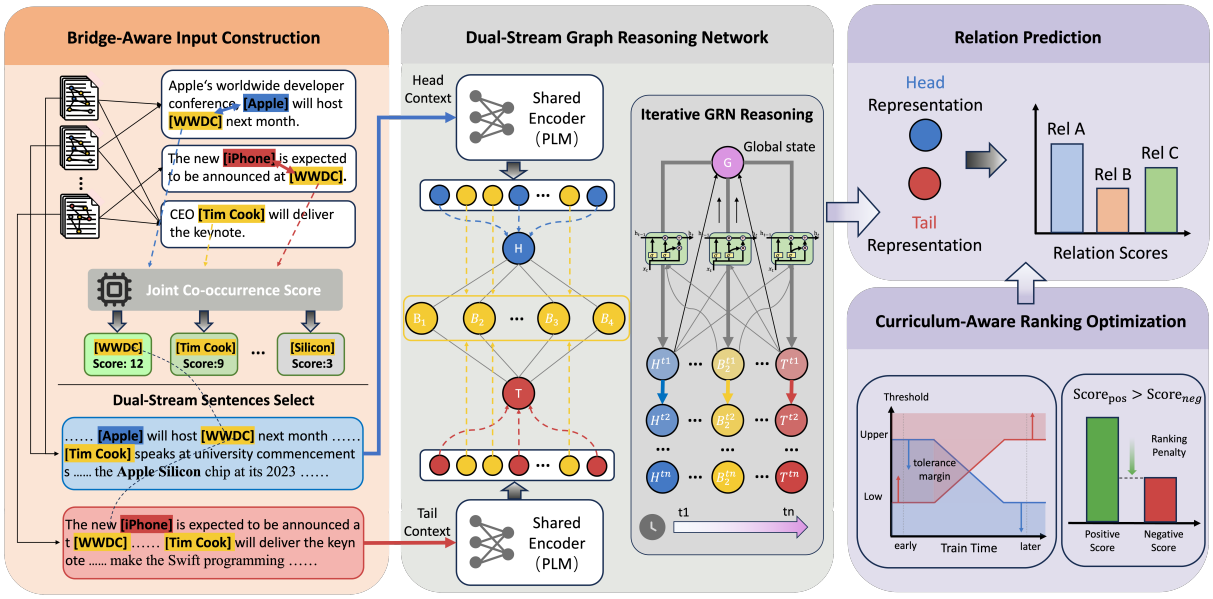

Figure 2: Overview of the DEBAR architecture. It comprises: (1) **Bridge-Aware Input Construction**, which utilizes joint co-occurrence to generate two balanced input streams, effectively mitigating relation transfer bias; (2) a **Dual-Stream Graph Reasoning Network** for independent context encoding and global synthesis via GRN; and (3) a Relation Prediction module using **Curriculum-Aware Ranking Optimizatio**n, which enforces strict sample ordering by dynamically tightening constraints.

plicitly capture long-range dependencies via self-attention; Graph-based methods (Zeng et al., 2020; Li et al., 2020; Lu et al., 2023a) that explicitly model structural interactions through heterogeneous entity graphs; and emerging LLM-based generative methods (Xue et al., 2024; Zhang et al., 2025) that exploit in-context learning capabilities. Despite their success in intra-document reasoning, these methods are intrinsically constrained to a single document context and struggle to reason over entities distributed across different documents.

2.2 Cross-Document Relation Extraction

Cross-document RE (CodRE) (Yao et al., 2021) extends DocRE to scenarios where *the target entity pair is mentioned in different documents*. The existing CodRE methods primarily focus on organizing scattered evidence into coherent inputs. MR.COD (Lu et al., 2023b) constructs passage graphs via multi-hop retrieval to link latent evidence chains. REIC (Na et al., 2024) employs reinforcement learning to filter noise and refine evidence chains from documents. To further enhance entity interactions, ECRIM (Wang et al., 2022) synthesizes information across distinct paths via entity-centric representations, while NEPD (Yue et al., 2024) explicitly models global dependencies through a unified entity graph. Additionally, KD-CodRED (Jain et al., 2024) augments entity representations by incorporating external background knowledge as additional evidence.

Nevertheless, existing methods suffer from structural imbalance in input representations due to indiscriminate entity fusion and the threshold-induced ranking ambiguity of positive/negative samples, as detailed in Section 1. To address this, our method constructs independent context streams via bridge-aware input construction and captures dependencies through a dual-stream graph reasoning network. Finally, a curriculum-inspired dynamic training strategy is proposed to enforce highly discriminative decision boundaries.

3 Methodology

In this section, we present the DEBAR framework, as illustrated in Fig. 2. The framework is designed to effectively mitigate relation transfer bias by explicitly decoupling and preserving bidirectional bridge evidence, while also addressing the threshold-induced ranking inconsistency by leveraging a curriculum-aware ranking penalty.

3.1 Problem Definition

Given a document corpus $\mathcal{D}$, CodRE aims to predict the relation $r \in \mathcal{R}$ for each target entity pair (e_h, e_t), where $\mathcal{R}$ is a predefined set of relation types. CodRE requires reasoning over *bridge enti-*

ties e_b, which serve as intermediate evidence connecting e_h and e_t across different documents.

Typically, the target entity pair (e_h, e_t) exhibits a disjoint distribution: e_h appears in document subset $\mathcal{D}_h \subset \mathcal{D}$, while e_t appears in $\mathcal{D}_t \subset \mathcal{D}$, with $\mathcal{D}_h \cap \mathcal{D}_t = \emptyset$. To model their interaction, we construct a bag of N text paths $\mathcal{B} = \{p_i\}_{i=1}^{N}$. Each path is defined as a pair of documents $p_i = (d_h^i, d_t^i)$, where $d_h^i \in \mathcal{D}_h$, $d_t^i \in \mathcal{D}_t$, and both documents share at least one bridge entity. The goal is to infer the relation based on the bag of all text paths.

3.2 Bridge-aware Input Construction (BIC)

This module aims to construct a compact and refined input context, which is essential for efficient cross-document reasoning. A key challenge lies in identifying which bridge entities effectively facilitate reasoning. Prior work like ECRIM (Wang et al., 2022) assesses entity importance based on global statistics across the retrieved documents. However, these methods overlook the *structural imbalance* of the bridge entities: a bridge entity might be globally salient but connected exclusively to the head entity, thereby failing to establish a valid link to the tail (and vice versa). To address this, we propose a balanced relevance scoring mechanism. Our core assumption is that valid bridge entities need to exhibit strong dual-connectivity, effectively linking both the head and tail entities. Consequently, we explicitly quantify and balance the co-occurrence strength from both directions, ensuring that the selected contexts preserve a complete, bidirectional evidence path rather than a skewed, one-sided context.

Formally, for a bridge entity e_b and a target entity e_k (where $k \in \{h, t\}$), we compute the connection score $\mathcal{S}_k(e_b)$ as a weighted sum of co-occurrence frequencies at three levels:

$$\mathcal{S}_k(e_b) = \alpha \cdot \phi_s(e_b, e_k) + \beta \cdot \phi_p(e_b, e_k) + \gamma \cdot \phi_d(e_b, e_k) \tag{1}$$

where ϕ_s, ϕ_p, ϕ_d represent co-occurrence counts exclusively at the sentence, paragraph, and document levels. Specifically, let $\mathcal{U}_s$ be the set of all sentences. To avoid redundancy, we apply a hierarchical exclusion strategy to define the effective paragraph set $\mathcal{U}_p$ and document set $\mathcal{U}_d$. Here, $\mathcal{U}_p$ consists of paragraphs containing the pair but excluding those with sentence-level matches, and $\mathcal{U}_d$ follows the same logic for documents. The counts are calculated as:

$$\phi_s(e_b, e_k) = |\{u \in \mathcal{U}_s \mid e_b, e_k \in u\}| \tag{2}$$

$$\phi_p(e_b, e_k) = |\{u \in \mathcal{U}_p \mid e_b, e_k \in u\}| \tag{3}$$

$$\phi_d(e_b, e_k) = |\{u \in \mathcal{U}_d \mid e_b, e_k \in u\}| \tag{4}$$

where α, β, γ in Eq. (1) are hyper-parameters. We set $\alpha > \beta > \gamma$ to prioritize semantic locality, reflecting the intuition that sentence-level syntactic dependencies are stronger than paragraph-level discourse relations and broader thematic context.

Finally, to penalize one-sided bias and enforce structural balance, we synthesize the scores from both sides using the geometric mean:

$$S_{bri}(e_b) = \sqrt{\mathcal{S}_h(e_b) \cdot \mathcal{S}_t(e_b)} \tag{5}$$

This formulation ensures that $S_{bri}(e_b)$ is non-zero if and only if e_b maintains connectivity to both the head and tail entities, effectively filtering out spurious single-side links.

After quantifying the importance of each bridge entity via the balanced metric $S_{bri}(e_b)$, we propagate these scores to the sentence level to identify the most informative context. We define the saliency score $S_{sen}(s)$ for a sentence s by aggregating the contributions of the entities it contains.

Formally, let $\mathcal{E}_b(s)$ denote the set of bridge entities present in the sentence s. The sentence saliency score is computed as:

$$S_{sen}(s) = \begin{cases} +\infty & \text{if } \{e_h, e_t\} \subseteq s \\ \sum_{e_b \in \mathcal{E}_b(s)} S_{bri}(e_b) & \text{else if } \mathcal{E}_b(s) \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where the first condition ensures that sentences containing both target entities are assigned the highest priority. For other sentences, the score reflects the cumulative quality of the bridge entities they contain. This mechanism favors contexts rich in highly relevant, dual-connected bridge entities.

Finally, we rank all sentences in the documents by $S_{sen}(s)$ in descending order and retain the top-ranked sentences to construct the input (i.e., *two input streams* with respect to the given entity pair) for the subsequent module, up to the maximum sequence length allowed by the pre-trained language model.

3.3 Dual-Stream Graph Reasoning Network (DSGN)

To capture complex cross-document dependencies of two input streams while mitigating signal interference, we propose a unified reasoning module.

This module employs a decoupled encoding strategy to generate entity-centric representations for constructing a global entity graph, which then initializes a Graph Recurrent Network (GRN) for iterative, multi-hop reasoning.

Dual-Stream Encoding. Instead of fusing the disjoint contexts of head and tail entities into a unified sequence, we adopt a dual-stream architecture to independently encode the evidence on each side. Specifically, we construct two input sequences, X_h and X_t, using the top-ranked sentences from Section 3.2. We employ a pre-trained language model (PLM) to encode these sequences:

$$\mathbf{H}_h = \text{PLM}(X_h), \quad \mathbf{H}_t = \text{PLM}(X_t) \tag{7}$$

From the output embeddings, we extract the initial representations via max-pooling over the corresponding token spans. Specifically, we obtain the head entity representation $\mathbf{e}_h^{(0)}$ from X_h and the tail entity representation $\mathbf{e}_t^{(0)}$ from X_t. Additionally, for each bridge entity, we independently extract its context-specific representations from both the head and tail input streams.

Iterative Reasoning via GRN. We construct a global entity graph $\mathcal{G}$ initialized with the extracted dual-stream embeddings, where edges encode semantic co-occurrences. Let $\mathcal{N}(v)$ denote the neighbor set of node v. We explicitly set the initial state $\mathbf{h}_v^{(0)}$ of each node to its corresponding entity embedding (i.e., $\mathbf{e}_h^{(0)}$, $\mathbf{e}_t^{(0)}$, or $\mathbf{e}_b^{(0)}$) and simultaneously initialize a global state $\mathbf{g}^{(0)}$ via mean-pooling over all node states.

We employ GRN that iteratively refines these states over L steps utilizing a Gated Recurrent Unit (GRU). At step l, we first perform gated aggregation to filter noisy neighbors, computing a message vector $\mathbf{m}_v^{(l)}$ as a weighted sum $\sum_{u \in \mathcal{N}(v)} \alpha_{vu} \mathbf{h}_u^{(l-1)}$, where α_{vu} is learnable coefficient. To capture global dependencies, we update both the local and global states via coupled GRUs:

$$\mathbf{h}_v^{(l)} = \text{GRU}_{node}\left(\mathbf{m}_v^{(l)} \oplus \mathbf{g}^{(l-1)}, \mathbf{h}_v^{(l-1)}\right) \tag{8}$$

$$\mathbf{g}^{(l)} = \text{GRU}_{global}\left(\text{Pool}(\{\mathbf{h}_v^{(l)}\}), \mathbf{g}^{(l-1)}\right) \tag{9}$$

where $\oplus$ denotes concatenation. This mechanism facilitates the continuous exchange of local information and global context over L iterations.

3.4 Prediction and Curriculum-aware Ranking Optimization (CAO)

Based on the refined entity representations from GRN, we perform relation classification using an aggregation strategy, which is optimized via a novel curriculum-aware learning objective.

Relation Prediction. Following ECRIM (Wang et al., 2022), we first construct path-specific relation representations. Let $\mathbf{h}_h$ and $\mathbf{h}_t$ denote the refined representations of the head and tail entities obtained from GRN in Eq. (8). We derive a relation embedding $\mathbf{r}_k$ for the $\mathbf{k}$-th text path, and subsequently apply a cross-path self-attention mechanism to capture dependencies among distinct paths. The path-level probability distribution is predicted via an Multi-Layer Perceptron (MLP), and the final bag-level score $\hat{y}(r)$ is obtained via max-pooling. Finally, a relation r is predicted as positive if score exceeds the decision threshold:

$$\hat{y}(r) = \max_k \text{MLP}(\mathbf{r}_k)^{(r)} \tag{10}$$

Curriculum-aware Ranking Optimization (CAO). The existing loss in all CodRE models employs a fixed pair of upper and lower thresholds for all entity pairs, and the *tolerance margin between the upper and lower thresholds fails to guarantee that the scores of positive samples consistently exceed those of negative ones*, as illustrated in Fig. 1. Specifically, a rigid margin can hinder convergence in the initial phase, while a relaxed margin lacks the strictness required to separate hard negatives during the final phase.

To address this, we propose a dynamic training strategy inspired by curriculum learning, which guides the model from a lenient optimization constraint setting to a strict one. Instead of fixed thresholds, we employ dynamic scalar margins $m_n^{(t)}$ and $m_p^{(t)}$ to serve as the upper and lower bounds, which progressively tighten to enforce stricter discrimination. Initially, lenient constraints facilitate convergence on coarse patterns; as training proceeds, the constraints tighten to enforce strict separation. The loss is defined as:

$$\mathcal{L}_{CL} = \log\left(e^{m_n^{(t)}} + \sum_{r \in \Omega_{neg}} e^{\hat{y}(r)}\right) + \log\left(e^{-m_p^{(t)}} + \sum_{r \in \Omega_{pos}} e^{-\hat{y}(r)}\right) \tag{11}$$

		Train	Dev	Test
Bags	Pos	2733	1010	1012
	NA	16668	4558	4523
Text paths	Pos	8263	2558	38019
	NA	120925	38182	2505

Table 1: Statistics of the CodRED dataset.

where Ω_{pos} and Ω_{neg} denote the sets of positive relations r_p and negative relations r_n, respectively. The boundaries shift linearly from a relaxed state ($m_n^{\text{start}} > m_p^{\text{start}}$) to a rigorous margin constraint ($m_p^{\text{end}} > m_n^{\text{end}}$) as the training epoch t increases.

Furthermore, while $\mathcal{L}_{\text{CL}}$ separates classes via global boundaries, it does not strictly guarantee the relative order of positive-negative pairs.

To explicitly enforce that positive relations outscore negative ones, we incorporate a weighted pairwise ranking penalty. First, we define the weighted penalty $\ell(r_p, r_n)$ for a specific positive-negative pair as:

$$\ell(r_p, r_n) = \sigma\big(\hat{y}(r_n) - \hat{y}(r_p)\big) \cdot \max\big(0, \delta + \hat{y}(r_n) - \hat{y}(r_p)\big) \quad (12)$$

where δ is the margin and $\sigma(\cdot)$ is the sigmoid function that acts as a difficulty-aware weighting mechanism. The final ranking objective is then computed by averaging over all pairs:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\Omega_{\text{pos}}||\Omega_{\text{neg}}|} \sum_{r_p \in \Omega_{\text{pos}}} \sum_{r_n \in \Omega_{\text{neg}}} \ell(r_p, r_n) \quad (13)$$

The final objective is $\mathcal{L}_{\text{CAO}} = \mathcal{L}_{\text{CL}} + \lambda \mathcal{L}_{\text{rank}}$, where λ balances the two terms.

4 Experimental setup

4.1 Dataset and Metrics

Consistent with prior studies, we also conduct experiments on the standard **CodRED** benchmark (Yao et al., 2021), a large-scale dataset derived from Wikipedia and Wikidata specifically designed for cross-document relational reasoning. The statistics of CodRED are shown in Table 1. It features 276 distinct relation types and long documents (averaging 4,939 tokens), requiring multi-hop inference across disjoint texts. We conduct experiments under both the **closed** setting, where the valid text paths are explicitly provided, and the **open** setting, where the paths are not given and must be retrieved from the corpus. Performance is assessed using four standard metrics: Area Under the Curve (AUC), F1 score, and Precision at K (P@500, P@1000).

4.2 Implementation Details

We implement DEBAR using BERT-base and RoBERTa-large as backbones. The model is optimized via AdamW (Loshchilov and Hutter, 2019). We set the learning rate to 3e-5 and train for 10 epochs. The hidden dimension is set to 768 for BERT-base and 1024 for RoBERTa-large. The full set of hyperparameters is reported in **Appendix A**.

4.3 Main Results

Results on CodRED Our model demonstrates state-of-the-art performance across settings and backbones. First, using the BERT-base backbone (Table 2), DEBAR achieves consistent gains over the previous baselines. Under the closed setting, it outperforms the best-performing baselines across all metrics, yielding improvements of 0.95 in F1, 0.42 in AUC, and 0.50 in both P@500 and P@1000 on the dev set. Crucially, this superiority extends to the open setting, where DEBAR surpasses the leading baseline by 0.24 in F1, demonstrating its robustness against retrieval noise.

Furthermore, employing RoBERTa-large to test generalization (Table 3), DEBAR surpasses LGCR by 1.55 in AUC and 2.06 in F1 on the dev set. This superiority is sustained on the test set, where DEBAR outperforms the baseline by 3.73 in AUC and 1.63 in F1.

Results on CodRED with Single-Document Subset Removed To evaluate DEBAR in a strictly cross-document relation extraction setting, following prior work, we remove CodRED's single-document subset and train/test using only cross-document text-path instances (Table 4). DEBAR remains strong without single-document data, achieving 39.32 F1 and 33.75 AUC. It outperforms NEPD by 1.10 F1 and 0.60 AUC, and exceeds ECRIM by 2.91 F1. These results indicate that DEBAR can capture high-quality reasoning evidence directly from disjoint documents.

4.4 Ablation Study

We conduct ablation experiments to verify the effectiveness of each component of our model, with the ablation results presented in Table 5.

Model	Closed Setting						Open Setting			
	Dev				Test		Dev		Test	
	AUC	F1	P@500	P@1000	AUC	F1	AUC	F1	AUC	F1
Pipeline (Yao et al., 2021)	17.45	30.54	30.60	26.70	18.94	32.29	14.07	26.45	16.26	28.70
End-to-end (Yao et al., 2021)	47.94	51.26	62.80	51.00	47.46	51.02	40.86	47.23	39.05	45.06
ECRIM (Wang et al., 2022)	61.23	60.85	78.04	60.13	60.35	62.28	48.73	50.94	49.27	51.36
MR.COD (Lu et al., 2023b)	59.22	61.20	–	–	61.68	62.53	51.00	53.06	53.30	**57.88**
LGCR (Wu et al., 2023)	63.17	61.67	76.65	61.84	61.08	60.75	51.48	52.96	50.15	53.45
NEPD (Yue et al., 2024)	65.01	63.63	77.84	64.03	66.23	64.41	**54.92**	54.49	55.87	56.68
DEBAR	**65.43**	**64.58**	**78.54**	**64.53**	**66.47**	**64.72**	53.36	**54.73**	**57.15**	57.21

Table 2: Performance comparison with baselines on CodRED using **BERT-base**. The left block reports results under the **Closed Setting**, while the right block reports the **Open Setting**. Results of the baselines are cited from original papers.

Model	Dev				Test	
	AUC	F1	P@500	P@1000	AUC	F1
End-to-end	56.47	57.49	72.60	57.10	56.32	56.05
ECRIM	59.36	61.81	**79.60**	61.90	62.12	63.06
LGCR	64.76	63.18	77.25	63.74	63.03	63.79
DEBAR	**66.31**	**65.24**	78.73	**64.82**	**66.76**	**65.42**

Table 3: Comparison using **RoBERTa-large** under the Closed Setting.

Model	Dev		Test	
	F1	AUC	F1	AUC
End-to-end	26.56	15.67	–	–
ECRIM	39.19	29.85	36.41	27.40
LGCR	40.73	32.81	36.67	28.01
NEPD	42.26	34.13	38.22	33.15
DEBAR	**43.35**	**35.08**	**39.32**	**33.75**

Table 4: Experimental results when removing CodRED's single-document subset and train/test using only cross-document text-path instances.

Model	AUC	F1	P@500	P@1000
DEBAR (Full)	**65.43**	**64.58**	**78.54**	**64.53**
w/o BIC	59.11	58.84	73.57	58.36
w/o Dual-Encoder	63.54	62.42	76.71	62.36
w/o GRN	64.59	63.85	77.53	64.27
w/o Dual-Encoder & GRN	61.45	60.36	75.83	59.91
w/o CAO	64.57	64.14	77.81	64.16

Table 5: Ablation studies on the CodRED dev set.

w/o BIC. To evaluate the contribution of the BIC module, we conduct an ablation study by removing this module. As shown in Table 5, this removal precipitates a sharp performance drop on the CodRED development set, decreasing F1 by 5.74 and AUC by 6.32. These results indicate that BIC effectively addresses the imbalance where bridge entities connect to only one side of the entity pair.

w/o Dual-Encoder. To evaluate the impact of early decoupling, we remove the Dual-Encoder architecture. As shown in Table 5, this removal decreases F1 by 2.16 and AUC by 1.89. This indicates that maintaining feature independence is crucial for mitigating relation transfer bias.

w/o GRN. To assess the contribution of graph reasoning, we remove the GRN module. As shown in Table 5, this removal decreases F1 by 0.73 and AUC by 0.84. This indicates that the GRN is effective in handling cross-document evidence fusion and capturing global dependencies.

w/o Dual-Encoder & GRN. We substituted it with the single-stream input structure used in ECRIM. This modification reduces the F1 score by 4.19 and AUC by 3.75. These results indicate that the dual-stream design better captures bidirectional headtail dependencies, mitigating relation transfer contextual bias.

w/o CAO. To evaluate the impact of the CAO module, we replace CAO with the original loss. As shown in Table 5, this change leads to consistent performance drops across all metrics, with F1 decreasing by 0.44 and AUC by 0.86. These results indicate that our loss enforces a stricter relative ordering between positive and negative samples, yielding a more effective optimization objective.

Model	Time (hours/epoch)	F1 Score
End-to-end	6.5	51.26
ECRIM	9.7	60.85
DEBAR	10.2	64.58

Table 6: Comparison of efficiency and performance of baseline models.

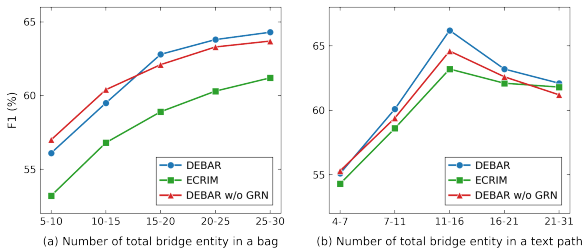

Figure 3: Experimental results on F1 based on the number of bridge entities under two settings, (a) is the max number of bridge entities in a bag, (b) is the average number of bridge entities in a text path.

4.5 Cost Analysis

As shown in Table 6, we compare the training efficiency and performance of different models. While the End-to-end baseline requires the least training time (6.5 hours/epoch), it yields significantly lower performance (51.26 F1). In contrast, compared to ECRIM, DEBAR incurs only a marginal increase in training time (approximately 5.2%) but improves the F1 score from 60.85 to 64.58. These results demonstrate that DEBAR effectively balances computational cost and reasoning capability, achieving superior performance with modest time overhead.

4.6 Impact of Bridge Entity Number

To investigate the impact of the number of bridge entities on model performance, we conduct experiments by controlling the maximum number of bridge entities per bag. As shown in Fig. 3(a), the F1 score of our model improves steadily as the number of bridge entities increases, consistently outperforming the ECRIM baseline. Notably, the full DEBAR model exhibits a steeper performance improvement compared to the GRN-free variant. This demonstrates that the GRN module effectively leverages the additional bridge entities to enhance inter-entity reasoning.

Furthermore, Fig. 3(b) presents the performance across varying quantities of bridge entities. While both models show an initial improvement followed by a decline, DEBAR maintains a significantly larger margin over ECRIM as the bridge entity count grows. This confirms DEBAR's effectiveness in modeling complex multi-hop interactions even under noisy or dense conditions.

4.7 Impact of the CAO Loss

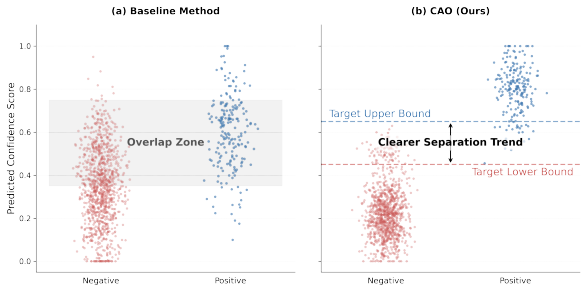

Figure 4: Comparison of normalized confidence score distributions between the baseline loss and CAO.

Model	Optimization Strategy	F1
End-to-end	Origin	51.26
	+ CAO	**52.37**
ECRIM	Origin	60.85
	+ CAO	**61.73**

Table 7: Generalization analysis of the CAO loss.

We validate the effectiveness and generalization of the CAO loss through visualization and comparative experiments. As shown in Fig. 4, using the original loss as a baseline in our model suffers from a "Overlap Zone" where positive and negative samples are indistinguishable. In contrast, CAO effectively disentangles this overlap by suppressing negative scores and elevating positive ones, establishing a clear decision boundary. Furthermore, Table 7 confirms the module's generalization: applying CAO consistently boosts F1 scores for both the End-to-end (51.26→52.37) and ECRIM (60.85→61.73) baselines, demonstrating its robustness as a generalized strategy.

4.8 Comparison with Large Language Models

We conduct comparative experiments following NEPD (Yue et al., 2024), which provides the latest, widely recognized setup for evaluating Large Language Models (LLMs) in CodRE. We evaluate the models under the Closed Setting on the development set, adopting Micro-F1. This accounts for

Model	Micro-F1 (Dev)
End-to-end*	78.24
ECRIM*	82.59
LGCR*	82.07
NEPD*	84.35
GPT-3.5-turbo*	28.05
InstructUIE*	70.34
InstructUIE-FT*	80.72
DEBAR	**84.63**

Table 8: Comparison with LLMs on the CodRED dev set (Closed Setting). * indicates results from NEPD (Yue et al., 2024).

generative LLMs not natively producing the calibrated probability distributions across all relations required for threshold-based Max-F1. The NEPD protocol ensures rigorous head-to-head precision comparison.

As shown in Table 8, following the NEPD protocol, DEBAR outperforms the fine-tuned InstructUIE-FT by 3.91 Micro-F1. This margin demonstrates that our specialized structural constraints mitigate relation transfer bias more effectively than general-purpose generative reasoning, achieving superior multi-hop accuracy with substantially lower computational overhead.

4.9 Case Study

To explicitly demonstrate how DEBAR mitigates Relation Transfer Bias, we present a qualitative analysis in Table 9 for the triplet (*Sexxx Dreams*, *Synth-pop*, *genre*). The inference relies on establishing a semantic path from the head (song) to the tail (genre) via the bridge entity (*Lady Gaga*). As shown, both models successfully retrieve the foundational Head-Bridge connection (S1). However, a significant divergence observed in Tail-Side Evidence Selection underscores a critical limitation of the baseline. Specifically, the ECRIM model exhibits a pronounced relation transfer bias, leading to issues with distraction and disconnection during evidence integration. Driven by surface-level relevance to the tail entity, it erroneously selects a noisy sentence (S2) discussing "polyphonic synthesizers", which is semantically isolated from the bridge entity (*Lady Gaga*). Critically, this distraction causes ECRIM to miss the pivotal bridging sentence (S3), which explicitly links "Lady Gaga" to the "new era of synth-pop". Although

Context Selection Process	ECRIM	DEBAR
Step 1: Head-Side Evidence *(Bridge ↔ Head)*		
[S1] "**Sexxx Dreams**" is a song by American singer **Lady Gaga** from her third studio album, Artpop.	✓	✓
Step 2: Tail-Side Evidence *(Bridge ↔ Tail)*		
[S2 - Noise] ... The development of **polyphonic synthesizers** led to a more commercial sound ... *(Error: Captures other Bridge)*	✓	✗
[S3 - Critical Link] Following the breakthrough of **Lady Gaga**... *(Success: Captures Bridge-Tail co-occurrence)*	✗	✓
[S4] ... media proclaimed a new era of female **synth-pop** stars ...	✓	✓

Table 9: Case study for (Sexxx Dreams, Synth-pop, *genre*). The table illustrates the **Relation Transfer Bias**: ECRIM captures the head-side evidence (S1) but gets distracted by tail-side noise (S2) due to the lack of bridge constraints. DEBAR properly enforces the bridge entity (Lady Gaga) on both sides, correctly filtering noise and retrieving the critical link (S3).

ECRIM captures the final context (S4), the absence of S3 results in a broken evidence chain ($Head \rightarrow Bridge \ldots \text{Gap} \ldots Tail$), leading to inference failure.

4.10 Hyperparameter Analysis

We further analyze the impact of key hyperparameters and present detailed results in **Appendix A**.

5 Conclusion

In this paper, we present DEBAR, a framework designed to tackle the relation transfer bias and optimization ambiguity prevalent in Cross-document RE. By strategically decoupling bidirectional reasoning and enforcing dynamic ranking constraints, DEBAR effectively prevents one-sided evidence propagation and resolves prediction order inconsistencies. Our results on CodRED not only establish a new state-of-the-art but also demonstrate the generalizability of our ranking objective, offering a robust paradigm for balanced and precise relational inference.

Limitations

First, the separate encoding of head and tail contexts in our dual-stream architecture introduces a

marginal training latency, representing a trade-off for unbiased reasoning. Second, our reliance on explicit bridge entities to construct evidence paths limits adaptability in instances where such intermediate connections are sparse or absent. Future work intends to address this by incorporating implicit reasoning mechanisms and leveraging auxiliary context from non-bridge entities to supplement the evidence chain.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly improved this paper. This work is supported by the National Natural Science Foundation of China (62276057).

References

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 745–758. Association for Computational Linguistics.

Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. 2024. Knowledge-driven cross-document relation extraction. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3787–3797. Association for Computational Linguistics.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1551–1560. International Committee on Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023a. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15453–15464. Association for Computational Linguistics.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2023b. Multi-hop evidence retrieval for cross-document relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10336–10351. Association for Computational Linguistics.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1963–1975. Association for Computational Linguistics.

Byeonghu Na, Suhyeon Jo, Yeongmin Kim, and Il-Chul Moon. 2024. Reward-based input construction for cross-document relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9254–9270. Association for Computational Linguistics.

Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. 2022. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9871–9881. Association for Computational Linguistics.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Haoran Wu, Xiuyi Chen, Zefa Hu, Jing Shi, Shuang Xu, and Bo Xu. 2023. Local-to-global causal reasoning for cross-document relation extraction. *IEEE CAA J. Autom. Sinica*, 10(7):1608–1621.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autore: Document-level relation extraction with large language models. *CoRR*, abs/2403.14888.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4452–4472. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Hao Yue, Shaopeng Lai, Chengyi Yang, Liang Zhang, Junfeng Yao, and Jinsong Su. 2024. Towards better graph-based cross-document relation extraction via non-bridge entity enhancement and prediction debiasing. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 680–691. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1630–1640. Association for Computational Linguistics.

Fu Zhang, Hongsen Yu, Jingwei Cheng, and Huangming Xu. 2025. Entity pair-guided relation summarization and retrieval in llms for document-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4022–4037. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 161–168. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620. AAAI Press.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1331–1339. Association for Computational Linguistics.

A Hyperparameter Analysis

Hyperparameter settings. We implement DEBAR using BERT-base and RoBERTa-large as backbones. The model is optimized via AdamW (Loshchilov and Hutter, 2019) with a linear warm-up (first 5% of steps) followed by a linear decay. We set the learning rate to 3e-5 and train for 10 epochs. The hidden dimension is set to 768 for BERT-base and 1024 for RoBERTa-large. For input construction, we set the granularity weights as $\alpha = 5, \beta = 3$, and $\gamma = 1$, and retain the top-ranked sentences to fill the maximum sequence length of 512. The GRN consists of $L = 2$ layers. For optimization, the ranking loss weight λ is set to 0.1 with a ranking margin $\delta = 0.3$. Regarding the curriculum schedule, the positive lower bound $m_p^{(t)}$ linearly increases from 8 to 10, while the negative upper bound $m_n^{(t)}$ decreases from 10 to 8 throughout training.

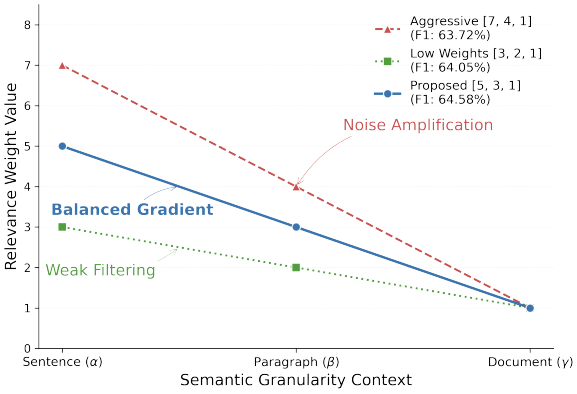

Figure 5: Performance comparison under different hierarchical weight settings (α, β, γ).

Analysis of the hyper-parameters α, β, and γ in BIC module. To validate the hierarchical filtering mechanism in our Bridge-aware Input Construction (BIC) module, we conduct a sensitivity analysis on the hyper-parameters α, β, and γ. These parameters determine the relevance weights for co-occurrences at the sentence, paragraph, and document levels, respectively. Fixing the document baseline $\gamma = 1$, we vary α and β to examine the impact of different weighting gradients. The performance comparison is visualized in Fig. 5. We observe that the pro-

posed configuration $[5, 3, 1]$ achieves the best performance (64.58), followed by the conservative setting $[3, 2, 1]$ (64.05) and the aggressive setting $[7, 4, 1]$ (63.72). This ranking reveals a critical trade-off between discrimination capability and noise amplification: **Optimal Balance** ($[5, 3, 1]$). As illustrated by the blue line in Fig. 5, this setting establishes a balanced gradient. The transition from $\alpha = 5$ to $\beta = 3$ provides sufficient discrimination to prioritize strong syntactic dependencies (sentence-level) over looser associations. Crucially, the moderate paragraph weight retains necessary context without allowing heuristic connections to dominate the reasoning path, effectively achieving an optimal signal-to-noise ratio.

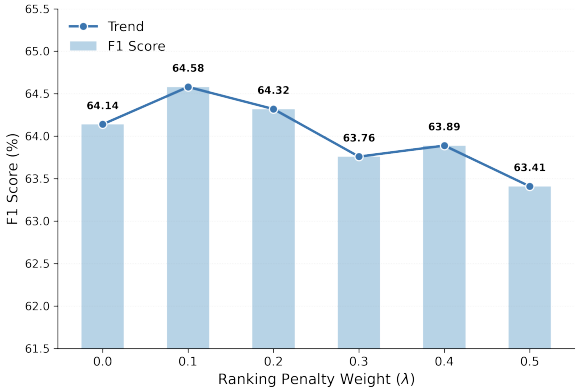

Figure 6: Impact of the ranking penalty weight λ on F1 scores.

Analysis of the ranking penalty weight λ. We investigate the influence of the ranking penalty weight λ, which balances the pairwise ranking objective against the global classification loss. We vary λ within the range $[0, 0.5]$ and report the F1 scores in Fig. 6. The results show that performance peaks at $\lambda = 0.1$ (64.58) and then exhibits a downward trend as λ increases. This trajectory reveals a critical trade-off between order refinement and optimization conflict. Specifically, we observe that an auxiliary constraint with $\lambda = 0.1$ serves as a beneficial regularizer in balancing these two objectives. The improvement from $\lambda = 0.0$ to $\lambda = 0.1$ confirms that introducing a mild ranking penalty is beneficial. A small weight is sufficient to act as a "soft corrector", refining the relative order of positive and negative pairs in the decision space without disrupting the primary feature learning driven by the global curriculum loss.

Analysis of the Ranking Margin δ. We investigate the sensitivity of the ranking margin δ in equation 12 by varying it within the range $\{0.1, 0.2, \ldots, 0.5\}$. As shown in Fig. 7, the model performance initially improves as δ increases, peaking at $\delta = 0.3$. A smaller margin (e.g., $\delta < 0.2$) fails to enforce a sufficient "safety zone" between positive and negative scores, resulting in weak discrimination. Conversely, an excessively large margin (e.g., $\delta > 0.4$) imposes overly aggressive constraints, which may lead to optimization instability or overfitting to hard negatives. Consequently, we set $\delta = 0.3$ as the optimal threshold to balance class separability and training stability.

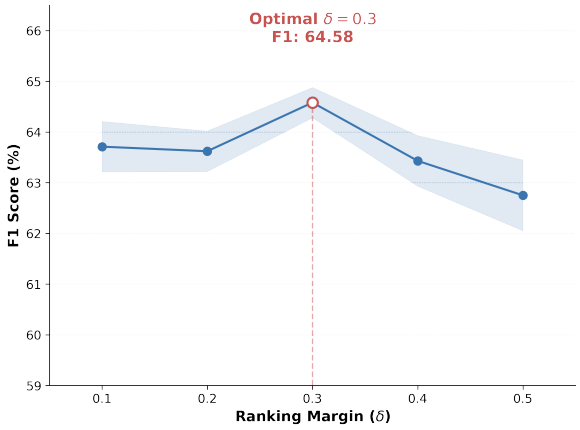

Figure 7: Impact of the ranking margin δ on F1 scores.