ATGL: An Adaptive-Threshold Global Loss for Document-level Relation Extraction

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

Abstract

Document-level relation extraction (DocRE) aims to determine which relations hold between a given entity pair within a document. As a multi-label classification task, the most commonly adopted paradigm introduces a learnable threshold to distinguish positive and negative classes for an entity pair. Under this paradigm, existing losses decouple the optimization into independent positive and negative losses, which interact solely with a shared threshold. This leads to two inherent limitations: (*i*) threshold instability caused by conflicting gradient updates from the decoupled losses; and (*ii*) optimization bias exacerbated by the severe imbalance between limited positive samples and abundant negative samples inherent in DocRE, which makes the model more likely to predict that no relation exists. To address these issues, we propose the **A**daptive-**T**hreshold **G**lobal **L**oss (ATGL). Unlike prior work, ATGL integrates positive, negative, and threshold optimization into a unified logit space and explicitly enforces ranking constraints on their contributions to the objective. Furthermore, ATGL incorporates an imbalance-aware optimization mechanism, thereby effectively addressing the severe class imbalance in DocRE. Our ATGL serves as a general optimization objective that can be readily applied to different DocRE models. Experiments on four datasets show that ATGL outperforms other DocRE losses and achieves state-of-the-art results, while consistently improving the performance of existing DocRE models. Code is available at https://github.com/xhm-code/ATGL.

1 Introduction

Relation Extraction (RE) aims to identify relations between an entity pair in a given text. Compared to sentence-level RE, document-level relation extraction (DocRE) (Yao et al., 2019) is more challenging because it requires reasoning over longer contexts where entities may appear across multiple sentences, thereby increasing the complexity of relation modeling. DocRE plays a crucial role in downstream applications, such as question answering (Baek et al., 2023; Pan et al., 2024) and knowledge graph construction (Zhang and Soh, 2024).

Given that DocRE is a multi-label classification task where an entity pair may have multiple relations, most existing works distinguish between *positive* and *negative classes*[1] for a given entity pair by applying a threshold. To achieve this, the Binary Cross-Entropy (BCE) loss (Goodfellow et al., 2016) is adopted for the DocRE task (Yao et al., 2019; Zeng et al., 2020). BCE decomposes the multi-label classification task into multiple independent binary subtasks and determines a fixed threshold for all entity pairs. However, such a fixed threshold cannot effectively adapt to different entity pairs. To address this inflexibility, Zhou et al. (2021) propose the Adaptive Threshold Loss (ATL). ATL introduces an adaptive *threshold class* $\mathcal{TH}$, which allows each entity pair to apply an adaptive threshold and decomposes the optimization into two independent parts[2]: positive loss $\mathcal{L}_{\mathcal{P}_T}$ and negative loss $\mathcal{L}_{\mathcal{N}_T}$.

Building upon the success of ATL, subsequent loss optimization strategies primarily focus on learning a clearer separation between the threshold class and positive/negative classes by enhancing discriminative margins (e.g., NCRL (Zhou and Lee, 2022), AML (Wei and Li, 2022), HingeABL (Wang et al., 2023), AFL (Tan et al., 2022a), AMTL (Xu et al., 2025a), ARPDL (Xu et al., 2025b), and CMM (Duan et al., 2025)). Despite these advances,

* Corresponding author.

[1] Given a predefined set of relations $\mathcal{R}$, the *positive classes* $\mathcal{P}_T \subseteq \mathcal{R}$ for an entity pair represent the relations that exist, while the *negative classes* $\mathcal{N}_T \subseteq \mathcal{R}$ represent the relations that do not exist.

[2] For ATL-based losses, a *threshold class* (denoted as $\mathcal{TH}$) is used to divide $\mathcal{R}$ into $\mathcal{P}_T$ and $\mathcal{N}_T$. The positive loss $\mathcal{L}_{\mathcal{P}_T}$ measures the distance between $\mathcal{TH}$ and $\mathcal{P}_T$, while the negative loss $\mathcal{L}_{\mathcal{N}_T}$ measures the distance between $\mathcal{TH}$ and $\mathcal{N}_T$.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34702–34716

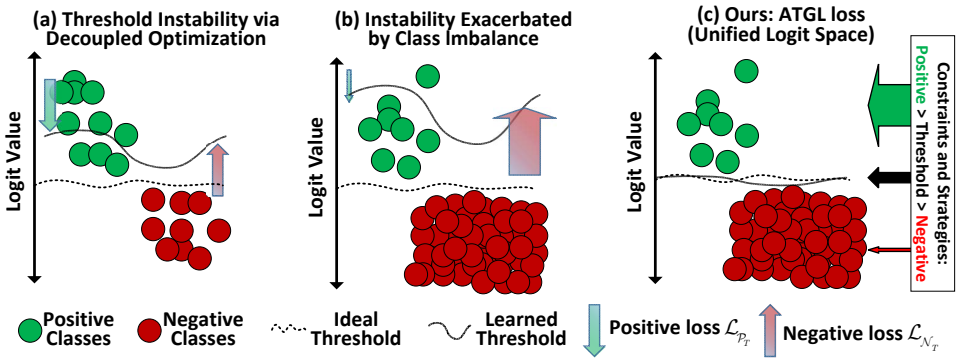


Figure 1: The degree of deviation between the learned threshold and the ideal threshold reflects the level of stability, and the thickness of the arrows represents the magnitude of the loss. **(a)** Existing ATL-based losses optimize positive and negative samples independently, inducing conflicting updates on the shared threshold and causing threshold instability. **(b)** The severe class imbalance in DocRE further amplifies this instability: abundant no-relation samples dominate the negative loss and bias the model toward predicting no relation for an entity pair. **(c)** Our proposed ATGL optimizes all classes in a unified logit space, explicitly enforcing a global ranking among them while strategically suppressing negative classes and amplifying positive ones.

several key challenges remain to be fully addressed:

(i) Threshold Instability via Decoupled Optimization. Existing ATL-based losses *independently* optimize the positive loss $\mathcal{L}_{\mathcal{P}_T}$ and the negative loss $\mathcal{L}_{\mathcal{N}_T}$, interacting solely with the shared threshold class $\mathcal{TH}$. As shown in **Fig. 1(a)**, this decoupled optimization makes the *threshold unstable*: on one hand, to minimize the positive loss $\mathcal{L}_{\mathcal{P}_T}$, the threshold may be pushed too low; on the other hand, to minimize the negative loss $\mathcal{L}_{\mathcal{N}_T}$, the threshold may be excessively raised. Consequently, this leads to an ultimately learned threshold that is either high or low compared to the ideal threshold, which causes a shift in the decision boundary.

(ii) Instability Exacerbated by Class Imbalance. As shown in **Fig. 1(b)**, this instability is further exacerbated by the severe *class imbalance*[3] inherent to DocRE (e.g., 94% of entity pairs in ReDocRED (Tan et al., 2022b) have no relation). The large number of samples with no relation dominates the optimization of the negative loss $\mathcal{L}_{\mathcal{N}_T}$, making the model more likely to predict that no relation exists. Under the decoupling optimization paradigm, the positive loss $\mathcal{L}_{\mathcal{P}_T}$ struggles to counterbalance this dominance, resulting in a sub-optimal separation threshold that fails to effectively distinguish between positive and negative classes.

To address these issues, we propose a novel loss, **A**daptive-**T**hreshold **G**lobal **L**oss (ATGL), as shown in **Fig. 1(c)**. Unlike previous approaches that decouple the optimization of positive and negative classes, ATGL integrates positive, negative, and threshold classes into a unified logit space. Through a unified optimization paradigm that explicitly enforces ranking constraints on the contributions of positive, negative, and threshold classes to the objective, ATGL ensures the stability of the threshold. Furthermore, by strategically suppressing the contribution of negative classes while amplifying positive ones, ATGL effectively counteracts the gradient dominance of the negative loss $\mathcal{L}_{\mathcal{N}_T}$ caused by class imbalance.

Our main contributions are as follows:

- We propose ATGL, a unified loss that resolves the threshold instability issue inherent in ATL-based losses by establishing a direct, global interaction between positive, negative, and threshold classes.
- We further incorporate an imbalance-aware optimization mechanism into ATGL that mitigates the gradient dominance of the negative loss, thereby addressing severe class imbalance in DocRE.
- We provide detailed theoretical analysis and formal proofs for ATGL, offering insights into its optimization behavior under threshold instability and class imbalance.
- Extensive evaluations on four DocRE datasets show that ATGL loss consistently surpasses other DocRE losses, achieving state-of-the-art

[3]Class imbalance: The number of entity pairs containing at least one valid relation is significantly smaller than those without any relation, resulting in severe class imbalance (often termed the $\mathcal{NA}$ problem).

results and exhibiting strong generalization across various DocRE backbone models.

2 Related Work

Document-level relation extraction (DocRE) aims to predict one or more relations for an entity pair and is typically formulated as a multi-label classification problem. Traditionally, DocRE tasks commonly adopt the Binary Cross-Entropy (BCE) (Goodfellow et al., 2016) loss, which models each relation independently using a sigmoid function and applies a fixed threshold during inference. However, since a fixed threshold is not suitable for all entity pairs, Zhou et al. (2021) propose the **Adaptive Threshold Loss** (ATL) to overcome this limitation. ATL introduces a threshold class $\mathcal{TH}$, requiring positive classes to yield scores higher than this threshold, while negative classes remain below it. During inference, each entity pair is assigned an adaptive threshold, and relations whose scores exceed this threshold are regarded as valid. Due to the effectiveness of ATL in DocRE tasks, subsequent DocRE models commonly adopt ATL as their optimization objective, including ALTOP (Zhou et al., 2021), DREEAM (Ma et al., 2023), SA-KD (Zhang et al., 2023), and SRF (Zhang et al., 2024).

To further enhance the discrimination between positive and negative classes, **subsequent studies extend ATL** by incorporating margins or strengthening the separation between positive and negative logits, leading to variants such as Balanced-Softmax (Zhang et al., 2021), AML (Wei and Li, 2022), NCRL (Zhou and Lee, 2022), and Hinge loss (Wang et al., 2023). Specifically, Balanced-Softmax is adopted in DocuNet (Zhang et al., 2021), the AML loss is applied in the SagDRE (Wei and Li, 2022) model, while NCRL is adopted in the REwNCRL (Xu et al., 2024) model.

In parallel, ATL remains limited in handling the severe class imbalance in DocRE, where the large number of negative samples tend to dominate its optimization. To mitigate this issue, several studies propose losses that rebalance the contributions of positive and negative classes or incorporate prior knowledge, such as AFL (Tan et al., 2022a), ARPDL (Xu et al., 2025b), and CMM (Duan et al., 2025). Meanwhile, other works aim to enhance the discriminability of positive classes or introduce multi-threshold mechanisms, as seen in PEMSCL (Guo et al., 2023) and AMTL (Xu et al., 2025a). Specifically, AFL is applied in the KD-DocRE (Tan et al., 2022a) model, and PEMSCL is used in the VaeDiff-DocRE (Tran et al., 2025) model.

Although these methods improve different aspects of ATL, their losses still optimize the positive and negative parts independently, interacting solely through the shared threshold class, as detailed in **Section 1**. This can lead to a learned threshold that deviates from the ideal threshold (either too high or too low), causing instability and a shift in the decision boundary. Additionally, this instability is further exacerbated by severe class imbalance. To address these issues, we propose the **A**daptive **T**hreshold **G**lobal **L**oss (ATGL).

3 Preliminary

3.1 Problem Formulation

Given a document $\mathcal{D}$ and a set of entities $\{e_i\}_{i=1}^{n}$ it contains, where n denotes the total number of entities, DocRE is formulated as a multi-label task that aims to identify relations in $\mathcal{R} \cup \{\mathcal{NA}\}$ that hold between each entity pair (e_s, e_o). Here, $\mathcal{R}$ denotes a predefined set of relations, and e_s and e_o correspond to the subject and object entities, respectively. For each pair $T = (e_s, e_o)$, we denote by $\mathcal{P}_T \subseteq \mathcal{R}$ the set of relations that hold, and by $\mathcal{N}_T \subseteq \mathcal{R}$ the set of relations that do not. If $\mathcal{P}_T = \emptyset$, the pair is assigned the $\mathcal{NA}$ label.

3.2 Adaptive Threshold Loss (ATL) for Multi-label Classification

ATL (Zhou et al., 2021) is designed for multi-label classification in DocRE by introducing a learnable threshold class $\mathcal{TH}$. To realize this mechanism, the ATL loss optimizes positive loss $\mathcal{L}_{\mathcal{P}_T}$ and negative loss $\mathcal{L}_{\mathcal{N}_T}$ independently, interacting solely with the shared threshold class $\mathcal{TH}$, as shown in **Eq. (1)**.

$$\begin{aligned} \mathcal{L}_{\mathcal{P}_T} &= -\sum_{r \in \mathcal{P}_T} \log \left(\frac{\exp(\text{logit}_r)}{\sum_{r' \in \mathcal{P}_T \cup \{\mathcal{TH}\}} \exp(\text{logit}_{r'})} \right), \\ \mathcal{L}_{\mathcal{N}_T} &= -\log \left(\frac{\exp(\text{logit}_{\mathcal{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\mathcal{TH}\}} \exp(\text{logit}_{r'})} \right), \\ \mathcal{L}_{ATL} &= \mathcal{L}_{\mathcal{P}_T} + \mathcal{L}_{\mathcal{N}_T}. \end{aligned} \tag{1}$$

During training, the positive loss $\mathcal{L}_{\mathcal{P}_T}$ encourages logits of positive classes $\mathcal{P}_T$ to exceed threshold class $\mathcal{TH}$, while the negative loss $\mathcal{L}_{\mathcal{N}_T}$ ensures logits of negative classes $\mathcal{N}_T$ remain below threshold class $\mathcal{TH}$. At inference, relations whose logits exceed the threshold are predicted to hold.

4 Methodology

4.1 Our Analysis of ATL-based Losses

Threshold Instability via Decoupled Optimization. As shown in **Eq. (1)**, the ATL loss consists of two decoupled components: the positive loss $\mathcal{L}_{\mathcal{P}_T}$ and the negative loss $\mathcal{L}_{\mathcal{N}_T}$. These two losses are optimized independently, each within its own logit space, interacting solely with the shared threshold class $\mathcal{TH}$. This design introduces an inherent limitation: ***independent optimization of the positive and negative losses can lead to inconsistent updates of the shared threshold***.

As shown in **Eq. (2)** and **Eq. (3)**, the gradients of the threshold $\text{logit}_{\mathcal{TH}}$ with respect to the positive and negative losses are given by:

$$\nabla_{\mathcal{TH}}\mathcal{L}_{\mathcal{P}_T} = -\sum_{r\in\mathcal{P}_T}\frac{\partial}{\partial \text{logit}_{\mathcal{TH}}} \log\frac{\exp(\text{logit}_r)}{\exp(\text{logit}_{\mathcal{TH}})+\sum_{r'\in\mathcal{P}_T}\exp(\text{logit}_{r'})}, \tag{2}$$

$$\nabla_{\mathcal{TH}}\mathcal{L}_{\mathcal{N}_T} = -\frac{\partial}{\partial \text{logit}_{\mathcal{TH}}} \log\frac{\exp(\text{logit}_{\mathcal{TH}})}{\exp(\text{logit}_{\mathcal{TH}})+\sum_{r'\in\mathcal{N}_T}\exp(\text{logit}_{r'})}. \tag{3}$$

Thus, the total gradient with respect to the threshold can be expressed as:

$$\nabla_{\mathcal{TH}}\mathcal{L}_{ATL} = \nabla_{\mathcal{TH}}\mathcal{L}_{\mathcal{P}_T} + \nabla_{\mathcal{TH}}\mathcal{L}_{\mathcal{N}_T}. \tag{4}$$

Intuitively, $\nabla_{\mathcal{TH}}\mathcal{L}_{\mathcal{P}_T} > 0$ encourages the threshold to decrease relative to logits of positive classes $\mathcal{P}_T$, whereas $\nabla_{\mathcal{TH}}\mathcal{L}_{\mathcal{N}_T} < 0$ encourages it to increase relative to logits of negative classes $\mathcal{N}_T$. These opposing gradients can interfere with each other, leading to instability in the threshold. As a result, the ultimately learned threshold is either higher or lower than the ideal threshold, which causes a shift in the decision boundary.

Instability Exacerbated by Class Imbalance. This instability is further exacerbated by the severe class imbalance (Tan et al., 2022a,b). As observed by Wang et al. (2023), $\mathcal{L}_{\mathcal{N}_T}$ in Eq. (5) $\to 0$ when $\text{logit}_{r'} - \text{logit}_{\mathcal{TH}} \to -\infty$, indicating that $\text{logit}_{\mathcal{TH}} \gg \text{logit}_{r'}$. This implies that ATL tends to learn a threshold well above the scores of most candidate relations, which results in a sub-optimal separation threshold that fails to effectively distinguish between positive and negative classes. For clarity, the negative loss $\mathcal{L}_{\mathcal{N}_T}$ can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\mathcal{N}_T} &= -\log\left(\frac{\exp(\text{logit}_{\mathcal{TH}})}{\sum_{r'\in\mathcal{N}_T\cup\{\mathcal{TH}\}}\exp(\text{logit}_{r'})}\right)\\ &= -\log\left(\frac{1}{1+\sum_{r'\in\mathcal{N}_T}\exp(\text{logit}_{r'}-\text{logit}_{\mathcal{TH}})}\right).\end{aligned} \tag{5}$$

4.2 Adaptive-Threshold Global Loss

To overcome the above limitations of ATL-based losses, we propose the **A**daptive-**T**hreshold **G**lobal **L**oss (ATGL), which integrates positive, threshold, and negative classes into a unified logit space and explicitly enforces a global ranking among them. ATGL is formally defined as:

$$w_r = \begin{cases}\alpha > 1, & r \in \mathcal{P}_T\\ 1, & r = \mathcal{TH}\\ 0, & r \in \mathcal{N}_T\end{cases},$$

$$\mathcal{L}_{ATGL} = -\sum_{r\in\mathcal{R}\cup\{\mathcal{TH}\}} w_r \cdot \log\left(\frac{\exp(\text{logit}_r)}{\sum_{r'\in\mathcal{R}\cup\{\mathcal{TH}\}}\exp(\text{logit}_{r'})}\right). \tag{6}$$

For each relation $r \in \mathcal{R}\cup\{\mathcal{TH}\}$, logit_r denotes the unnormalized logit for the relation r, and w_r denotes the weight associated with the relation r.

To further explain the ATGL loss, we answer the following three questions. ***Additional theoretical analysis of ATGL is provided in Appendix A***.

(i) How can positive, threshold, and negative classes be represented in a unified logit space while establishing a global ranking among them? To place all classes in a unified logit space, ATGL integrates positive, threshold, and negative classes into a single softmax function, where each class's logit contributes to a shared normalization. This ensures that their probabilities are mutually dependent and directly comparable. To further establish a global ranking, ATGL assigns larger weights to positive classes $\mathcal{P}_T$ and smaller weights to negative classes $\mathcal{N}_T$, with the threshold class lying in between, thereby encoding the desired ordering within the unified logit space.

(ii) What is the rationale for assigning a weight of 1 to the threshold class and 0 to the negative classes? Due to severe class imbalance inherent to DocRE, the large number of examples with no relation dominates loss optimization, making the model more likely to predict that no relation exists. By assigning a weight of 1 to the threshold class $\mathcal{TH}$ and 0 to the negative classes $\mathcal{N}_T$, ATGL removes the explicit loss contributions from negative classes and prevents them from dominating the

training objective. In this setting, if the prediction for an entity pair is that there is no relation, the loss term associated with the threshold $\mathcal{TH}$ relative to negative classes $\mathcal{N}_T$ can be written as a standard cross-entropy objective:

$$\mathcal{L}_{\mathcal{TH}} = -\log \frac{\exp(\text{logit}_{\mathcal{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\mathcal{TH}\}} \exp(\text{logit}_{r'})}. \quad (7)$$

Minimizing this term encourages the logit of threshold to rise above the logits of negative classes $\mathcal{N}_T$, thereby stabilizing the learned threshold without requiring explicit supervision on each negative class.

Moreover, fixing the weights of the threshold $\mathcal{TH}$ and negative classes $\mathcal{N}_T$ to $w_{\mathcal{TH}} = 1$ and $w_r = 0$ for $r \in \mathcal{N}_T$ yields a simple and interpretable weighting scheme: positive classes $\mathcal{P}_T$ are emphasized more strongly with $w_r = \alpha > 1$, the threshold $\mathcal{TH}$ serves as an anchor with intermediate importance, and negative classes $\mathcal{N}_T$ contribute only through the softmax normalization term in the unified logit space.

(iii) How to alleviate class imbalance? As discussed above in **(ii)**, assigning $w_r = 0$ for $r \in \mathcal{N}_T$ removes the explicit loss contributions from negative classes and prevents them from dominating the training objective. Furthermore, assigning a larger weight $\alpha > 1$ to positive classes than to the threshold ($w_{\mathcal{TH}} = 1$) and negative classes ($w_r = 0$) amplifies the gradients from scarce positive examples and encourages the model to focus on correctly identifying positive classes.

5 Experiments

Datasets. To comprehensively evaluate our proposed ATGL loss, we conduct experiments on four DocRE datasets: CDR (Li et al., 2016), DWIE (Zaporojets et al., 2021), Re-DocRED (Tan et al., 2022b) and DocGNRE (Li et al., 2023). These datasets cover both general-domain and biomedical DocRE. We report basic statistics in **Table 1**, and provide more detailed descriptions in **Appendix B**.

Implementation Details and Evaluation Metrics. We adopt $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019) as encoders, and conduct experiments on a GeForce RTX 3090 GPU. Experimental results are averaged over multiple random seeds to ensure robustness.

Following Yao et al. (2019), we adopt the **F1** score and the **Ign F1** score for performance evaluation. The Ign F1 score is calculated by removing from the dev/test evaluation set the relational facts that overlap with the training set.

Dataset	Split	#Docs.	#Rels.	#Triples.
CDR	train	500	2	5240
	dev	500	2	5087
	test	500	2	5204
DWIE	train	602	65	14,403
	dev	98	65	2,624
	test	99	65	2,495
Re-DocRED	train	3,053	96	85,932
	dev	500	96	17,284
	test	500	96	17,448
DocGNRE	train	3,053	96	96,505
	dev	500	96	17,284
	test	500	96	19,526

Table 1: Statistics of datasets.

6 Main Results and Analysis

6.1 Different DocRE Models with ATGL

To evaluate the generality and effectiveness of ATGL, we apply it to various DocRE models by *replacing their original losses*. As shown in **Table 2**, ATGL consistently improves both F1 and Ign-F1 scores across different datasets and encoder backbones. On the Re-DocRED dataset, comparatively earlier models such as ATLOP, DocuNet, and DREEAM obtain substantial gains of about 2 to 5 F1 points, while more recent and stronger models like TTM-RE and VaeDiff-DocRE are further boosted to F1 scores of 82.13 and 79.37, respectively. Similar trends are observed on the DocGNRE dataset, where ATGL yields consistent improvements in Test F1 ranging from 1.9 to 4.7 points. On the DWIE dataset, the performance is also improved, with average gains of 2.62 in Test F1 and 3.86 in Test Ign-F1. Quantitatively, as indicated by the "**Avg.**", ATGL achieves a remarkable average improvement of **2.80** points in both Test F1 and Ign-F1 across all experiments. These results suggest that ATGL serves as a general and effective loss, yielding consistent performance gains across different models, encoders, and datasets.

6.2 Comparison of Different Losses

To further evaluate the effectiveness of ATGL, we compare it with the existing state-of-the-art losses for DocRE. ATGL loss consistently achieves the best performance across four datasets, as shown in **Table 3**. On Re-DocRED dataset, it reaches 77.26 F1 and 76.05 Ign-F1, outperforming the previous best loss (CMM) by +1.14 and +1.29 points, respectively. On DocGNRE dataset, ATGL obtains

Model	Dev				Test			
	F1	F1 with ATGL	Ign-F1	Ign-F1 with ATGL	F1	F1 with ATGL	Ign-F1	Ign-F1 with ATGL
Re-DocRED Dataset with BERT$_{base}$								
ATLOP (Zhou et al., 2021)	74.22 *	**77.01** (+2.79)	73.35 *	**75.75** (+2.40)	74.02 *	**77.26** (+3.24)	73.22 *	**76.05** (+2.83)
DocuNet (Zhang et al., 2021)	74.62 *	**76.93** (+2.31)	73.60 *	**75.71** (+2.11)	74.48 *	**77.09** (+2.61)	73.53 *	**75.92** (+2.39)
KD-DocRE (Tan et al., 2022a)	74.66 *	**76.80** (+2.14)	73.68 *	**75.58** (+1.90)	74.55 *	**76.68** (+2.13)	73.64 *	**75.51** (+1.87)
DREEAM (Ma et al., 2023)	74.13 *	**76.85** (+2.72)	73.68 *	**75.67** (+1.99)	73.75 *	**76.87** (+3.12)	73.33 *	**75.76** (+2.43)
TTM-RE (Gao et al., 2024)	75.51 *	**80.40** (+4.89)	74.31 *	**79.26** (+4.95)	75.71 *	**80.58** (+4.87)	74.55 *	**79.50** (+4.95)
VaeDiff-DocRE (Tran et al., 2025)	75.89 †	**77.51** (+1.62)	74.96 †	**76.29** (+1.33)	75.07 †	**77.44** (+2.37)	74.13 †	**76.26** (+2.13)
Re-DocRED Dataset with RoBERTa$_{large}$								
ATLOP (Zhou et al., 2021)	77.63 *	**80.35** (+2.72)	76.88 *	**79.28** (+2.40)	77.73 *	**80.68** (+2.95)	76.94 *	**79.64** (+2.70)
DocuNet (Zhang et al., 2021)	78.16 *	**79.97** (+1.81)	77.53 *	**78.96** (+1.43)	77.92 *	**79.85** (+1.93)	77.27 *	**78.88** (+1.61)
KD-DocRE (Tan et al., 2022a)	78.65 *	**80.09** (+1.44)	77.92 *	**79.07** (+1.15)	78.35 *	**80.13** (+1.78)	77.63 *	**79.18** (+1.55)
DREEAM (Ma et al., 2023)	77.60 *	**80.55** (+2.95)	77.20 *	**79.50** (+2.30)	77.94 *	**80.67** (+2.73)	77.34 *	**79.65** (+2.31)
TTM-RE (Gao et al., 2024)	78.13 *	**81.68** (+3.55)	78.05 *	**80.75** (+2.70)	79.95 *	**82.13** (+2.18)	78.20 *	**81.25** (+3.05)
VaeDiff-DocRE (Tran et al., 2025)	79.19 †	**79.48** (+0.29)	78.35 †	**78.49** (+0.14)	79.03 †	**79.37** (+0.34)	78.22 †	**78.41** (+0.19)
DocGNRE Dataset with BERT$_{base}$								
ATLOP (Zhou et al., 2021)	73.89 *	**76.90** (+3.01)	73.07 *	**75.64** (+2.57)	68.74 *	**72.80** (+4.06)	68.06 *	**71.74** (+3.68)
DocuNet (Zhang et al., 2021)	74.95 †	**77.00** (+2.05)	73.99 †	**75.84** (+1.85)	69.98 †	**72.68** (+2.70)	69.16 †	**71.70** (+2.54)
KD-DocRE (Tan et al., 2022a)	75.19 †	**76.86** (+1.67)	74.20 †	**75.65** (+1.45)	70.21 †	**72.13** (+1.92)	69.38 †	**71.07** (+1.69)
DREEAM (Ma et al., 2023)	74.23 *	**76.99** (+2.76)	73.76 *	**75.73** (+1.97)	68.24 *	**72.74** (+4.50)	68.89 *	**71.68** (+2.79)
TTM-RE (Gao et al., 2024)	75.44 *	**80.12** (+4.68)	74.33 *	**78.88** (+4.55)	71.14 *	**75.85** (+4.71)	70.19 *	**74.80** (+4.61)
VaeDiff-DocRE (Tran et al., 2025)	75.60 †	**77.28** (+1.68)	74.61 †	**75.98** (+1.37)	71.05 †	**73.52** (+2.47)	70.21 †	**72.41** (+2.20)
DocGNRE Dataset with RoBERTa$_{large}$								
ATLOP (Zhou et al., 2021)	77.61 *	**80.53** (+2.92)	76.96 *	**79.42** (+2.46)	72.90 *	**76.31** (+3.41)	72.36 *	**75.38** (+3.02)
DocuNet (Zhang et al., 2021)	77.70 †	**79.85** (+2.15)	76.97 †	**78.79** (+1.82)	73.29 †	**75.60** (+2.31)	72.71 †	**74.73** (+2.02)
KD-DocRE (Tan et al., 2022a)	77.62 †	**79.83** (+2.21)	76.87 †	**78.82** (+1.95)	72.95 †	**75.65** (+2.70)	72.32 †	**74.83** (+2.51)
DREEAM (Ma et al., 2023)	77.75 *	**80.56** (+2.81)	77.28 *	**79.49** (+2.21)	72.90 *	**76.29** (+3.39)	72.97 *	**75.40** (+2.43)
TTM-RE (Gao et al., 2024)	78.16 *	**81.83** (+3.67)	77.30 *	**80.91** (+3.61)	73.72 *	**77.56** (+3.84)	73.01 *	**76.78** (+3.77)
VaeDiff-DocRE (Tran et al., 2025)	78.06 †	**79.62** (+1.56)	77.24 †	**78.37** (+1.13)	73.57 †	**75.64** (+2.07)	72.90 †	**74.59** (+1.69)
DWIE Dataset with BERT$_{base}$								
ATLOP (Zhou et al., 2021)	69.96 ‡	**72.45** (+2.49)	63.57 ‡	**67.37** (+3.80)	74.36 ‡	**77.34** (+2.98)	67.56 ‡	**70.55** (+2.99)
KD-DocRE (Tan et al., 2022a)	71.78 ‡	**75.53** (+3.75)	65.84 ‡	**69.63** (+3.79)	77.01 ‡	**78.45** (+1.44)	70.27 ‡	**71.32** (+1.05)
TTM-RE (Gao et al., 2024)	73.01 †	**76.35** (+3.34)	65.33 †	**70.90** (+5.57)	75.59 †	**78.89** (+3.30)	65.98 †	**72.01** (+6.03)
DWIE Dataset with RoBERTa$_{large}$								
ATLOP (Zhou et al., 2021)	76.65 *	**76.97** (+0.32)	72.47 *	**72.77** (+0.30)	81.39 *	**81.60** (+0.21)	76.83 *	**77.04** (+0.21)
KD-DocRE (Tan et al., 2022a)	76.55 *	**76.94** (+0.39)	72.01 *	**72.24** (+0.23)	80.92 *	**81.40** (+0.48)	75.67 *	**76.41** (+0.74)
TTM-RE (Gao et al., 2024)	72.63 †	**80.18** (+7.55)	64.35 †	**76.00** (+11.65)	75.21 †	**82.49** (+7.28)	65.40 †	**77.54** (+12.14)
Avg.	75.84	**78.44** (+2.60)	73.99	**76.56** (+2.57)	74.72	**77.52** (+2.80)	72.73	**75.53** (+2.80)

Table 2: Performance of various DocRE models when *their original losses are replaced with the ATGL loss*. ATL loss (Zhou et al., 2021) for ATLOP and DREEAM, Balanced-Softmax loss (Zhang et al., 2021) for DocuNet, AFL loss (Tan et al., 2022a) for KD-DocRE, SSR-PU loss (Wang et al., 2022) for TTM-RE, and PEMSCL loss (Guo et al., 2023) for VaeDiff-DocRE. † indicates our reproduction; * indicates results from Xu et al. (2025a); ‡ from Zhang et al. (2023). **Avg.** reports the average performance scores and improvements across all models and datasets.

Loss Function	Re-DocRED		DocGNRE		DWIE		CDR
	F1	Ign-F1	F1	Ign-F1	F1	Ign-F1	F1
ATL (Zhou et al., 2021)	73.29 *	72.46 *	68.74 *	68.06 *	74.36 §	67.56 §	68.84 †
Balanced-Softmax (Zhang et al., 2021)	73.68 *	72.85 *	68.84 *	68.13 *	67.50 †	57.68 †	68.26 †
AML (Wei and Li, 2022)	72.60 *	71.78 *	67.86 *	67.11 *	72.36 †	63.69 †	68.71 †
AFL (Tan et al., 2022a)	74.15 *	73.20 *	69.45 *	68.69 *	75.83 †	67.75 †	69.29 †
NCRL (Zhou and Lee, 2022)	73.87 *	72.79 *	69.20 *	68.27 *	74.55 †	65.96 †	68.82 †
PEMSCL (Guo et al., 2023)	73.98 *	73.06 *	69.46 *	68.70 *	75.39 †	67.12 †	68.96 †
HingeABL$_{SAT}$ (Wang et al., 2023)	73.46 *	72.61 *	69.15 *	68.41 *	75.18 †	67.14 †	69.08 †
HingeABL$_{MeanSAT}$ (Wang et al., 2023)	74.68 *	72.90 *	70.83 *	69.25 *	59.40 †	45.82 †	67.18 †
HingeABL (Wang et al., 2023)	75.15 *	73.84 *	70.98 *	69.90 *	73.38 †	63.81 †	68.84 †
AMTL (Xu et al., 2025a)	75.63 *	74.44 *	71.34 *	70.34 *	75.71 †	67.85 †	-
ARPDL (Xu et al., 2025b)	75.90 ‡	74.81 ‡	-	-	-	-	-
CMM (Duan et al., 2025)	76.12 ‡	74.76 ‡	-	-	-	-	-
ATGL (Ours)	**77.26**(+1.14)	**76.05**(+1.24)	**72.80**(+1.46)	**71.74**(+1.40)	**77.34**(+1.51)	**70.05**(+2.20)	**70.38**(+1.09)

Table 3: Performance of *different losses* on four datasets: Re-DocRED, DocGNRE, DWIE, and CDR. * marks results from Xu et al. (2025a), † indicates our reproduction, § from Zhang et al. (2023), and ‡ refers to the original paper. Consistent with prior work, all experiments employ the model ATLOP (Zhou et al., 2021) for representation and BERT$_{base}$ for encoding. Since the implementation of ARPDL and CMM are not publicly available and their results are only reported on Re-DocRED, we compare with ARPDL and CMM only on Re-DocRED.

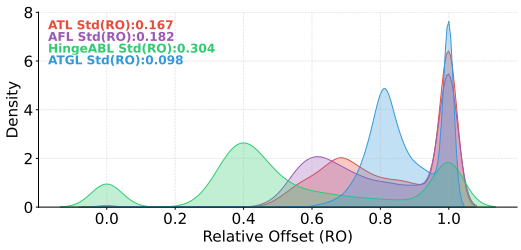


Figure 2: Distribution of the Relative Offset (RO) across different losses for *evaluating threshold instability*. RO quantifies the position of a threshold relative to the positive and negative classes. Density reflects the occurrence frequency of the RO values associated with entity pairs.

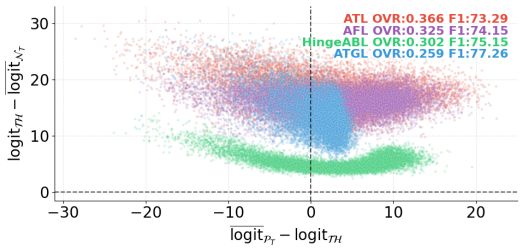


Figure 3: Comparison of *global ranking consistency* between ATL-based and ATGL losses. A lower Ordering Violation Rate (OVR) indicates better consistency.

72.80 F1 and 71.74 Ign-F1, surpassing the strongest baseline (AMTL) by +1.46 and +1.40 points. For DWIE and CDR datasets, ATGL also improves Test F1 by +1.51 and +1.09 points compared to the best reported results. ATGL consistently surpasses other losses across four datasets, improving performance by about 1.30 F1 and 1.61 Ign-F1 on average, further demonstrating its effectiveness.

7 Further Analysis

7.1 Empirical Verification of Threshold Instability

Since the thresholds in the evaluated losses are adaptive and each entity pair has its own threshold value, we first define the relative position of a threshold using the **Relative Offset (RO)**, which is defined as:

$$\text{RO} = \frac{\text{logit}_{\mathcal{TH}} - \text{logit}_{\mathcal{N}_T}}{\text{logit}_{\mathcal{P}_T} - \text{logit}_{\mathcal{N}_T}}. \tag{8}$$

RO thus provides a measure of where the threshold lies between the negative and positive classes. To quantify the stability of these adaptive thresholds across samples, we further compute the standard deviation of RO, i.e., Std(RO). *A smaller* Std(RO) *indicates that the threshold is more stably* positioned between the positive and negative classes.

Fig. 2 shows the distribution of RO across entity pairs with existing relations for different losses. ATL, AFL, and HingeABL exhibit relatively wide and dispersed RO distributions, reflecting higher variability in threshold positioning across entity pairs. In contrast, the proposed ATGL loss produces a distribution that is sharply concentrated near 1.0, suggesting that the thresholds are much more consistently positioned across samples.

Furthermore, as shown in the legend of **Fig. 2**, ATGL achieves the lowest Std(RO) value (0.098), followed by ATL (0.167), AFL (0.182), and HingeABL (0.304). Lower Std(RO) values indicate that the threshold is more stably positioned between the positive and negative classes across samples, which empirically confirms that the global ordering enforced by ATGL leads to more stable thresholds.

7.2 Global Ranking Consistency Analysis

To further examine whether different losses preserve the desired global ranking among positive, threshold, and negative logits, we use the **Ordering Violation Rate (OVR)**, which quantifies the consistency of this ranking. For each entity pair, let $\overline{\text{logit}}_{\mathcal{P}_T}$ and $\overline{\text{logit}}_{\mathcal{N}_T}$ denote the average logits of positive and negative classes, respectively, and $\text{logit}_{\mathcal{TH}}$ denote logit of the threshold class. A violation occurs if $\overline{\text{logit}}_{\mathcal{P}_T} < \text{logit}_{\mathcal{TH}}$ or $\text{logit}_{\mathcal{TH}} < \overline{\text{logit}}_{\mathcal{N}_T}$, and the overall OVR is defined as:

$$\text{OVR} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{I}\Big(\overline{\text{logit}}^{(i)}_{\mathcal{P}_T} < \text{logit}^{(i)}_{\mathcal{TH}} \vee \text{logit}^{(i)}_{\mathcal{TH}} < \overline{\text{logit}}^{(i)}_{\mathcal{N}_T}\Big), \tag{9}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition inside is satisfied and 0 otherwise, and N is the total number of entity pairs.

Fig. 3 reports the OVR performance of various losses. Among the baselines, HingeABL demonstrates relatively better consistency with an OVR of 0.302, outperforming AFL (0.325) and ATL (0.366). Notably, our proposed ATGL loss achieves the lowest OVR of 0.259, significantly reducing the violation rate by 14.2% compared to the strongest baseline (HingeABL). This substantial reduction demonstrates that ATGL effectively enforces a more rigorous global ranking. Furthermore, we observe that lower OVR values generally align with higher F1 scores, suggesting that maintaining global ranking consistency is a critical factor for improving DocRE performance.

Loss	FP ↓	FN ↓	FN _NA ↓	FN _Rel ↓	FN_NA /(FP+FN) ↓	FN /(FP+FN) ↓
ATL *	1887	6253	5498	755	67.54	76.82
AML *	2032	6363	5515	848	65.69	75.80
AFL *	2300	5744	4898	846	60.89	71.41
NCRL *	2770	5603	4872	731	58.19	66.92
PEMSCL *	2264	5746	4926	820	61.50	71.74
SAT *	**1749**	6241	5363	878	67.12	78.11
HingeABL *	2935	5083	4306	777	53.70	63.39
ARPDL *	2781	5076	4426	650	56.33	64.60
AMTL †	2890	5066	4295	771	53.98	63.68
ATGL (Ours)	3159	**4373**	**3730**	**643**	**49.52**	**58.06**

Table 4: Distribution of prediction errors for ATGL and baseline losses, illustrating effectiveness in mitigating *class imbalance*. * marks results from ARPDL (Xu et al., 2025b). † indicates our reproduction. FP (False Positive): a negative sample incorrectly predicted as positive. FN (False Negative): a positive sample incorrectly predicted as negative. FN_NA: a positive sample misclassified as negative, with the predicted label being $\mathcal{NA}$. FN_Rel: a positive sample misclassified as negative, with the label belonging to the negative classes.

Relation	FP (ATL→ATGL)	ATGL FP@NA / ATGL FP	ATGL type-consistent FP / ATGL FP
P131	350→638	551/638	637/638
P17	229→383	323/383	383/383
P27	113→210	200/210	208/210
P150	37→101	96/101	101/101
P361	37→70	46/70	70/70
P1001	20→56	24/56	56/56

Table 5: Relation-level attribution of FP increases on the Re-DocRED test set. FP@NA denotes cases where the gold label is $\mathcal{NA}$ but the model predicts a relation.

Model	F1 (Dev)	F1 (Test)	Ign-F1 (Dev)	Ign-F1 (Test)
ATLOP				
+ CMM *	76.30	76.10	75.00	74.80
+ ATGL	**77.01** (+0.71)	**77.26** (+1.16)	**75.75** (+0.75)	**76.05** (+1.25)
DocuNet				
+ CMM *	76.40	76.30	75.00	75.00
+ ATGL	**76.93** (+0.53)	**77.09** (+0.79)	**75.71** (+0.71)	**75.92** (+0.92)
KD-DocRE				
+ CMM *	75.90	76.10	74.30	74.60
+ ATGL	**76.80** (+0.90)	**76.68** (+0.58)	**75.58** (+1.28)	**75.51** (+0.91)

Table 6: Results on Re-DocRED comparing our ATGL loss with the prior SOTA CMM loss, evaluated with $\text{BERT}_{\text{base}}$. * denotes results reported in the original CMM paper.

7.3 Analysis of Class Imbalance

To investigate the effectiveness of ATGL in alleviating class imbalance in DocRE, we analyze the distribution of prediction errors. **Table 4** reports four types of false prediction patterns. As shown, prior losses such as ARPDL and PEMSCL produce a large number of false negatives. ARPDL predicts 5076 FN and 4426 FN_NA, while PEMSCL predicts 5746 FN and 4926 FN_NA. In contrast, our ATGL loss predicts 4373 FN and 3730 FN_NA, showing a clear reduction in false negatives. The number of FN_Rel is also lower for ATGL (643) compared to HingeABL (777), while FP increases slightly to 3159. Moreover, ATGL achieves the lowest FN_NA/(FP+FN) and FN/(FP+FN) ratios among all methods (49.52 and 58.06, respectively), indicating that a larger proportion of positive samples are correctly recovered. These results suggest that ATGL effectively mitigates class imbalance.

7.4 Analysis of False Positive

While ATGL substantially reduces false negatives, it also increases false positives, as shown in **Table 4**. To better understand this trade-off, we further analyze the relation-level distribution of the additional false positives. The increase is not evenly distributed across relations, but concentrated in a few frequent ones: P131, P17, and P27 account for 38% of the total FP increase from ATL to ATGL (**Table 5**). Moreover, nearly all of these additional false positives are type-consistent, with about 99%–100% satisfying type constraints across major relations. This suggests that many of these errors are plausible near-misses rather than arbitrary noise, indicating that ATGL mainly gains by recovering missed positive relations.

7.5 Analysis of Hyperparameter α

To evaluate the impact of the hyperparameter α in **Eq. (6)** on ATGL's performance, we report detailed results across four datasets in **Appendix C** and provide a practical guideline for selecting α.

7.6 ATGL vs. the Previous SOTA Loss CMM Across DocRE Models

To directly compare ATGL against the previous SOTA loss CMM (Duan et al., 2025), we evaluate both losses on several representative DocRE backbones. As shown in **Table 6**, replacing CMM with ATGL consistently improves both F1 and Ign-F1.

On ATLOP, ATGL improves test F1 from 76.10 to 77.26 and test Ign-F1 from 74.80 to 76.05 (+1.16 and +1.25, respectively). DocuNet and KD-DocRE also benefit from ATGL, with gains of up to +0.92 F1 and +1.28 Ign-F1. These consistent improvements indicate that ATGL is a strong and broadly applicable loss, serving as a reliable drop-in replacement for current SOTA DocRE losses.

7.7 Computational Cost of ATGL

To assess the computational cost of the proposed ATGL, we compare its training time against sev-

Category	Entity pair ($h \rightarrow t$)	Rel. r	Change (ATL→ATGL)	Margin (ATL→ATGL)
FN→TP	Jerry Garcia → Grateful Dead	P463	¬P463 → P463	$-0.31 \rightarrow 3.95$
FN→TP	Salesis house → Appenzell	P131	NA → P131	$-1.98 \rightarrow 3.60$
FN→TP	The Late Edition → BBC Radio	P449	NA → P449	$-1.93 \rightarrow 3.91$
TP→FN	Scandinavian → Danish	P527	P527 → NA	$7.16 \rightarrow -0.30$
TP→FN	Russia → Magnitogorsk	P150	P150 → NA	$5.79 \rightarrow -0.83$
FP@NA	Australia → Australian	P172	NA → P172	$-3.23 \rightarrow 3.41$
FP@NA	Marlborough → Worcester County	P131	NA → P131	$-3.41 \rightarrow 3.05$

Table 7: Representative qualitative examples of ATL and ATGL on the Re-DocRED test set. Margin is defined as logit_r - $\text{logit}_{\mathcal{TH}}$. FN→TP denotes cases where a relation exists but ATL misses it and ATGL predicts it correctly. TP→FN denotes cases where ATL predicts the relation correctly but ATGL misses it. FP@NA denotes cases where no labeled relation exists ($\mathcal{NA}$) but ATGL predicts relation r.

Loss Function	Training Time
ATL	40.37 minutes
Balanced-Softmax	40.03 minutes
AML	40.46 minutes
AFL	40.43 minutes
NCRL	39.44 minutes
PEMSCL	40.66 minutes
HingeABL	41.50 minutes
AMTL	40.46 minutes
ATGL (Ours)	40.71 minutes

Table 8: Comparison of losses in terms of training time on the ATLOP model with a $\text{BERT}_{\text{base}}$ encoder, trained for 30 epochs on Re-DocRED with a batch size of 4.

Model	PLM	Test	
		Ign-F1	**F1**
DocGNRE *	Llama3-8B	11.04	11.12
LMRC *	Llama3-8B	52.15	52.45
D-F *	Llama3-8B	52.50	53.33
D-R-F *	Llama3-8B	54.35	54.84
AutoRE *	Llama3-8B	58.33	59.29
EP-RSR *	Llama3-8B	63.03	64.24
DocKG-RAG †	Llama3-8B	<u>74.32</u>	<u>75.49</u>
ATLOP + ATGL loss (ours)	$\text{BERT}_{\text{base}}$	76.05	77.26
ATLOP + ATGL loss (ours)	$\text{RoBERTa}_{\text{large}}$	79.64	80.68
TTM-RE + ATGL loss (ours)	$\text{BERT}_{\text{base}}$	**79.50**	**80.58**
TTM-RE + ATGL loss (ours)	$\text{RoBERTa}_{\text{large}}$	**81.25**	**82.13**

Table 9: Performance of LLMs on Re-DocRED. * from Zhang et al. (2025); † from Xu et al. (2025c).

eral baseline losses. As shown in **Table 8**, training ATLOP with ATGL takes 40.71 minutes, which is comparable to competitive losses such as ATL (40.37 minutes), AFL (40.43 minutes), and HingeABL (41.50 minutes). These results demonstrate that ATGL achieves performance gains without introducing an additional computational burden.

7.8 Comparison with LLM-based Models

To contextualize ATGL beyond the ATL-based losses, we further compare it with representative LLM-based DocRE models on the Re-DocRED dataset, as shown in **Table 9**. ATLOP trained with ATGL achieves 77.26/80.68 F1 on the test set with $\text{BERT}_{\text{base}}$/$\text{RoBERTa}_{\text{large}}$, outperforming DocKG-RAG (75.49) and substantially exceeding earlier prompting-based methods such as EP-RSR (64.24). With a stronger backbone, TTM-RE combined with ATGL further improves performance to 80.58/82.13 F1, indicating that the effectiveness of ATGL persists across different encoder strengths.

7.9 Case Study

To further understand the behavior of ATGL, we present representative qualitative examples from the Re-DocRED test set in **Table 7**. Compared with ATL, ATGL mainly improves performance by correcting high-logit misses, i.e., cases where the target relation has a negative margin under ATL but a large positive margin under ATGL. We also include cases where ATL is correct but ATGL fails, as well as representative false positives made by ATGL on entity pairs labeled as $\mathcal{NA}$. These examples provide a balanced view of both the gains and trade-offs of ATGL.

8 Conclusion

We propose a novel **A**daptive-**T**hreshold **G**lobal **L**oss, ATGL, which effectively mitigates the threshold instability caused by decoupled optimization and the optimization bias exacerbated by severe class imbalance in DocRE tasks. ATGL proposes to integrate positive, negative, and threshold classes into a unified logit space and explicitly enforces global ranking constraints on their contributions. Experiments on four datasets show that ATGL consistently surpasses other DocRE losses, and improves different DocRE models. As a loss that does not depend on any specific model and only changes the optimization objective, we expect ATGL to be readily applicable to broader multi-label classification tasks beyond DocRE.

Limitations

Despite the substantial improvements that ATGL brings to DocRE, it still exhibits noteworthy limitations. In particular, while ATGL significantly enhances overall performance and reduces the false negative (FN) rate, these gains come at the cost of an increased false positive (FP) rate. As shown in **Section 7.3 Analysis of Class Imbalance (Table 4)**, although the overall comparative results show that ATGL effectively mitigates class imbalance compared to prior work, ATGL yields more false positives than other losses such as ATL, PEMSCL, and HingeABL. This indicates that, by strategically suppressing the contribution of negative classes while amplifying positive ones, ATGL behaves more aggressively when predicting the existence of relations for an entity pair, thereby inducing a trade-off between precision and recall. Such a trade-off may require careful calibration depending on the needs of specific application scenarios.

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

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C Park, and Sung Hwang. 2023. Knowledge-augmented language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1720–1736.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT)*, pages 4171–4186.

Zhichao Duan, Tengyu Pan, Zhenyu Li, Xiuxing Li, and Jianyong Wang. 2025. Comm: Concentrated margin maximization for robust document-level relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 23841–23849.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. TTM-RE: memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*, volume 1.

Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2598–2609.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5495–5505.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv, abs/1907.11692*.

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)*, pages 1971–1983.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 36(7):3580–3599.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8472–8487.

Khai Phan Tran, Wen Hua, and Xue Li. 2025. Vaediff-docre: End-to-end data augmentation framework for document-level relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 7307–7320.

Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. 2023. Adaptive hinge balance loss for document-level relation extraction. In *Findings of Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 3872–3878.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4123–4135.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2000–2008.

Huangming Xu, Fu Zhang, and Jingwei Cheng. 2025a. An adaptive multi-threshold loss and a general framework for collaborating losses in document-level relation extraction. In *Findings of the Association for Computational Linguistics (ACL)*, pages 20996–21007.

Huangming Xu, Fu Zhang, Jingwei Cheng, and Xin Li. 2025b. ARPDL: adaptive relational prior distribution loss as an adapter for document-level relation extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8313–8321.

Xiaolong Xu, Chenbin Li, Haolong Xiang, Lianyong Qi, Xuyun Zhang, and Wanchun Dou. 2024. Attention based document-level relation extraction with none class ranking loss. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6569–6577.

Xiaolong Xu, Yibo Zhou, Haolong Xiang, Xiaoyong Li, Xuyun Zhang, Lianyong Qi, and Wanchun Dou. 2025c. Docks-rag: Optimizing document-level relation extraction through llm-enhanced hybrid prompt tuning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 764–777.

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)*, pages 1630–1640.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9820–9836.

Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, and Yongxue Wu. 2024. Srf: enhancing document-level relation extraction with a novel secondary reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15426–15439.

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*, pages 4022–4037.

Liang Zhang, Zijun Min, Jinsong Su, Pei Yu, Ante Wang, and Yidong Chen. 2023. Exploring effective inter-encoder semantic interaction for document-level relation extraction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5278–5286.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3999–4006.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4538–4544.

A Theoretical Analysis of ATGL

Overview. Our goal is to formalize three key properties: **(i)** ATGL corresponds to matching a structured target distribution that encodes the desired group-wise ordering; **(ii)** the optimization landscape in logits is convex (modulo translation) and admits a unique minimizer in probability space; **(iii)** the optimum induces a strict group-wise logit ranking (e.g., $s_{r_p} > s_{\mathcal{TH}} > s_{r_n}$) and yields stable, interpretable gradient dynamics for the threshold.

A.1 Group-Structured Objective of ATGL

Purpose (A.1). We first rewrite ATGL in a unified softmax form and show that its class weights induce a structured target distribution over $\mathcal{C}$, where positive classes, the threshold, and negative classes play distinct roles. This view will be the basis for the convexity and ranking analyses below.

Recall that for an entity pair $T = (e_s, e_o)$, ATGL integrates positive, threshold, and negative classes into a single softmax (**Eq.** (6)). For notational convenience, we define

$$\mathcal{R} = \mathcal{P}_T \cup \mathcal{N}_T, \tag{10}$$

$$\mathcal{C} = \mathcal{R} \cup \{\mathcal{TH}\}, \tag{11}$$

and for each class $r \in \mathcal{C}$,

$$s_r = \text{logit}_r, \qquad p_r = \frac{\exp(s_r)}{\sum_{r' \in \mathcal{C}} \exp(s_{r'})}. \tag{12}$$

The per-instance ATGL loss can be written as

$$\mathcal{L}_{\text{ATGL}} = -\sum_{r \in \mathcal{C}} w_r \log p_r, \tag{13}$$

where the weights are

$$w_r = \begin{cases} \alpha > 1, & r \in \mathcal{P}_T, \\ 1, & r = \mathcal{TH}, \\ 0, & r \in \mathcal{N}_T, \end{cases} \qquad W = \sum_{r \in \mathcal{C}} w_r = \alpha|\mathcal{P}_T| + 1. \tag{14}$$

We normalize these weights as:

$$\tilde{w}_r = \frac{w_r}{W}, \quad r \in \mathcal{C}. \tag{15}$$

Then $\sum_{r \in \mathcal{C}} \tilde{w}_r = 1$, and **Eq.** (13) can be rewritten:

$$\mathcal{L}_{\text{ATGL}} = -W \sum_{r \in \mathcal{C}} \tilde{w}_r \log p_r. \tag{16}$$

The normalized weights $\tilde{w}_r$ can be interpreted as the relative contribution of positive, threshold, and negative classes for a given entity pair:

$$\tilde{w}_r = \begin{cases} \alpha/W, & r \in \mathcal{P}_T, \\ 1/W, & r = \mathcal{TH}, \\ 0, & r \in \mathcal{N}_T. \end{cases} \tag{17}$$

In particular, the positive classes $r \in \mathcal{P}_T$ share the same normalized weight $\tilde{w}_r = \alpha/W$, the threshold class satisfies $\tilde{w}_{\mathcal{TH}} = 1/W$, and the negative classes $r \in \mathcal{N}_T$ have $\tilde{w}_r = 0$ while still entering the softmax normalization through p_r.

A.2 Convexity, Hessian, and Uniqueness

Purpose (A.2). We analyze the optimization properties of ATGL: we derive the gradient and Hessian with respect to logits, establish convexity (modulo the translation invariance of softmax), and characterize the uniqueness of the optimal distribution in the probability simplex together with its attainability under finite logits.

Gradient. Using $\log p_r = s_r - \log Z$ with $Z = \sum_{j \in \mathcal{C}} \exp(s_j)$, we can rewrite

$$\mathcal{L}_{\text{ATGL}} = -\sum_{r \in \mathcal{C}} w_r (s_r - \log Z) \tag{18}$$

$$= -\sum_{r \in \mathcal{C}} w_r s_r + W \log Z. \tag{19}$$

Here $j, k \in \mathcal{C}$ index classes (relations and the threshold), and s_k denotes the logit of class k. Taking the derivative with respect to s_k gives

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_k} = -w_k + W \frac{\exp(s_k)}{Z} = W p_k - w_k. \tag{20}$$

Equivalently,

$$\frac{1}{W} \frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_k} = p_k - \tilde{w}_k, \tag{21}$$

i.e., the gradient is proportional to the difference between model probability and target probability.

Hessian and convexity. The Jacobian of p with respect to s is the standard softmax Jacobian:

$$\frac{\partial p_k}{\partial s_j} = p_k(\mathbb{I}[k = j] - p_j), \tag{22}$$

so the Hessian of $\mathcal{L}_{\text{ATGL}}$ reads

$$\frac{\partial^2 \mathcal{L}_{\text{ATGL}}}{\partial s_j \partial s_k} = W \frac{\partial p_k}{\partial s_j} = W\big(p_k \mathbb{I}[k = j] - p_k p_j\big). \tag{23}$$

In matrix form,

$$\nabla_s^2 \mathcal{L}_{\text{ATGL}} = W\big(\text{Diag}(p) - pp^\top\big). \tag{24}$$

For any vector $v \in \mathbb{R}^{|\mathcal{C}|}$,

$$v^\top \nabla_s^2 \mathcal{L}_{\text{ATGL}} v = W\big(\sum_r p_r v_r^2 - (\sum_r p_r v_r)^2\big) \tag{25}$$

$$= W \cdot \text{Var}_{r \sim p}(v_r) \geq 0. \tag{26}$$

Therefore $\nabla_s^2 \mathcal{L}_{\text{ATGL}}$ is positive semi-definite, and $\mathcal{L}_{\text{ATGL}}$ is convex in s. Moreover, the Hessian

is not positive definite: its nullspace is spanned by the all-ones vector, reflecting the standard softmax translation invariance $s \mapsto s + c \cdot \mathbf{1}$, i.e.,

$$\mathcal{L}_{\text{ATGL}}(s + c\mathbf{1}) = \mathcal{L}_{\text{ATGL}}(s), \quad \forall c \in \mathbb{R}. \tag{27}$$

Hence $\mathcal{L}_{\text{ATGL}}$ is convex but not strictly convex in s (modulo the translation subspace); in particular, any local minimizer (if it exists) is globally optimal up to a global shift.

Unique minimizer in probability space. Consider the constrained optimization problem

$$\begin{aligned} \min_{\{p_r\}} \quad & -\sum_{r\in\mathcal{C}} w_r \log p_r, \\ \text{s.t.} \quad & \sum_{r\in\mathcal{C}} p_r = 1, p_r \geq 0 \quad \forall r \in \mathcal{C}. \end{aligned} \tag{28}$$

Since $w_r = 0$ for all $r \in \mathcal{N}_T$, the objective is independent of $\{p_r\}_{r\in\mathcal{N}_T}$; if some $p_r > 0$ with $r \in \mathcal{N}_T$, we can shift an arbitrarily small amount of probability from p_r to a class in $\mathcal{P}_T \cup \{\mathcal{TH}\}$ to strictly decrease the objective while preserving the constraints, so at any minimizer we must have

$$p_r^* = 0, \quad \forall r \in \mathcal{N}_T. \tag{29}$$

Restricting to the remaining classes $\mathcal{P}_T \cup \{\mathcal{TH}\}$, the problem in **Eq.** (28) reduces to

$$\begin{aligned} \min_{\{p_r\}} \quad & -\sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} w_r \log p_r \\ \text{s.t.} \quad & \sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} p_r = 1, \\ & p_r \geq 0 \quad \forall r \in \mathcal{P}_T \cup \{\mathcal{TH}\}. \end{aligned} \tag{30}$$

The associated Lagrangian is

$$\begin{aligned} \mathcal{J}(p, \lambda) = & -\sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} w_r \log p_r \\ & + \lambda \left(\sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} p_r - 1 \right). \end{aligned} \tag{31}$$

Taking derivatives with respect to p_r and setting them to zero yields, for all $r \in \mathcal{P}_T \cup \{\mathcal{TH}\}$,

$$\frac{\partial \mathcal{J}}{\partial p_r} = -\frac{w_r}{p_r} + \lambda = 0 \quad \Rightarrow \quad p_r = \frac{w_r}{\lambda}. \tag{32}$$

Enforcing the constraint $\sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} p_r = 1$ gives

$$\begin{aligned} & \sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} \frac{w_r}{\lambda} = 1 \\ & \Rightarrow \quad \lambda = \sum_{r\in\mathcal{P}_T\cup\{\mathcal{TH}\}} w_r = W. \end{aligned} \tag{33}$$

Therefore the unique minimizer in the probability simplex is

$$p_r^* = \begin{cases} \alpha/W, & r \in \mathcal{P}_T, \\ 1/W, & r = \mathcal{TH}, \\ 0, & r \in \mathcal{N}_T, \end{cases} \tag{34}$$

which coincides with the normalized weights in **Eq.** (17).

In the probability simplex, p^* in **Eq.** (34) is the unique minimizer. Under the softmax parameterization, the solution p^* is not attained by any finite logits s; rather, $\inf_s \mathcal{L}_{\text{ATGL}}(s)$ is approached in the limit $s_{r_n} \to -\infty$ for all $r_n \in \mathcal{N}_T$. The relative logits of the positive and threshold classes are nevertheless unique up to a global shift, since $\mathcal{L}_{\text{ATGL}}(s) = \mathcal{L}_{\text{ATGL}}(s + c\mathbf{1})$ for any $c \in \mathbb{R}$.

A.3 Global Ranking

Purpose (A.3). Based on the optimal probability distribution obtained in Appendix A.2, we show that ATGL enforces a strict group-wise ordering among logits. In particular, positive classes are separated from the threshold by a logit gap controlled by α, while negative classes are pushed below the threshold in the softmax limit.

At the optimal probability distribution p^* in **Eq.** (34), we obtain a strict group-wise ranking in the logit space. For any positive class $r_p \in \mathcal{P}_T$, we have

$$\frac{p_{r_p}^*}{p_{\mathcal{TH}}^*} = \frac{\alpha/W}{1/W} = \alpha > 1. \tag{35}$$

Under the softmax parameterization,

$$\frac{p_{r_p}^*}{p_{\mathcal{TH}}^*} = \frac{\exp(s_{r_p})}{\exp(s_{\mathcal{TH}})} = \exp(s_{r_p} - s_{\mathcal{TH}}), \tag{36}$$

which yields

$$s_{r_p} - s_{\mathcal{TH}} = \log \alpha > 0. \tag{37}$$

Hence $s_{r_p} > s_{\mathcal{TH}}$ for all $r_p \in \mathcal{P}_T$.

For any negative class $r_n \in \mathcal{N}_T$, **Eq.** (34) implies $p_{r_n}^* = 0$. In the softmax parameterization, this corresponds to the limit $s_{r_n} \to -\infty$, and therefore for any finite $s_{\mathcal{TH}}$ we have $s_{\mathcal{TH}} > s_{r_n}$. Combining the two yields

$$s_{r_p} > s_{\mathcal{TH}} > s_{r_n}, \quad \forall r_p \in \mathcal{P}_T, \ \forall r_n \in \mathcal{N}_T, \tag{38}$$

which formalizes the global ranking between positive, threshold, and negative classes at optimum.

A.4 Gradient Dynamics and Threshold Stability

Purpose (A.4). Finally, we study the gradient dynamics implied by ATGL to explain why the threshold is stable during training. We show that the logit of threshold is driven toward a fixed target probability, positive classes are amplified toward a larger target mass, and negative classes are monotonically suppressed, interacting with the threshold only through the shared normalization.

The gradients in **Eq.** (20) describe how positive classes, the threshold, and negative classes interact during optimization.

Group-specific gradients. From **Eq.** (20) and **Eq.** (14), we have

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_r} = W p_r - \alpha, \qquad r \in \mathcal{P}_T, \tag{39}$$

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_{\mathcal{TH}}} = W p_{\mathcal{TH}} - 1, \tag{40}$$

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_r} = W p_r, \qquad r \in \mathcal{N}_T. \tag{41}$$

Assuming gradient descent with learning rate $\eta > 0$,

$$s_r^{(t+1)} = s_r^{(t)} - \eta \frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_r^{(t)}}. \tag{42}$$

Self-correcting threshold. **Eq.** (40) shows that the sign of the threshold gradient is determined solely by $p_{\mathcal{TH}}$:

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_{\mathcal{TH}}} \begin{cases} > 0, & p_{\mathcal{TH}} > 1/W, \\ = 0, & p_{\mathcal{TH}} = 1/W, \\ < 0, & p_{\mathcal{TH}} < 1/W. \end{cases} \tag{43}$$

Consequently, under gradient descent, the threshold logit is self-correcting around its target probability $1/W$: if $p_{\mathcal{TH}} > 1/W$, $s_{\mathcal{TH}}$ is decreased, and if $p_{\mathcal{TH}} < 1/W$, $s_{\mathcal{TH}}$ is increased.

Amplified positive classes. Similarly, for any positive class $r \in \mathcal{P}_T$,

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_r} \begin{cases} < 0, & p_r < \alpha/W, \\ = 0, & p_r = \alpha/W, \\ > 0, & p_r > \alpha/W, \end{cases} \tag{44}$$

thus the logits of positive classes are increased when their probabilities are below the target α/W and decreased otherwise, giving positive classes a larger target probability than the threshold (α/W vs. $1/W$), which is beneficial when positive examples are scarce.

Monotone suppression of negatives. For any negative $r \in \mathcal{N}_T$, **Eq.** (41) implies

$$\frac{\partial \mathcal{L}_{\text{ATGL}}}{\partial s_r} = W p_r \geq 0. \tag{45}$$

Whenever $p_r > 0$, gradient descent satisfies

$$s_r^{(t+1)} = s_r^{(t)} - \eta W p_r^{(t)} < s_r^{(t)}, \tag{46}$$

so the logits of negative classes are always pushed downward. Importantly, negative classes never contribute negative gradients to $s_{\mathcal{TH}}$ or the positive classes; they influence them only through the shared softmax normalization.

B Details of Datasets

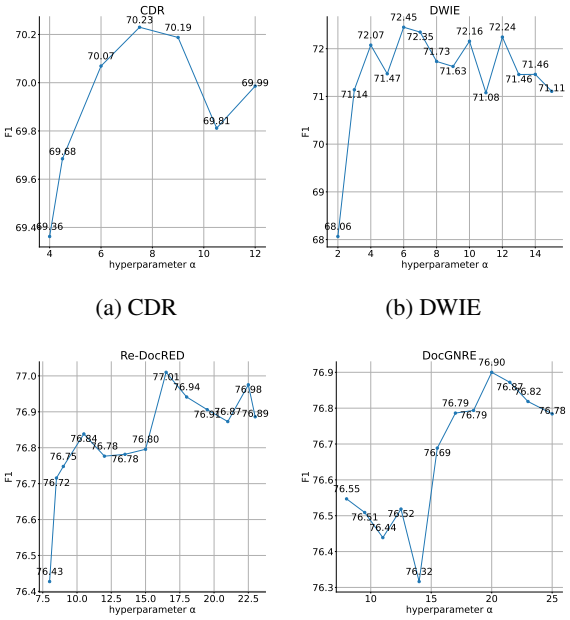
- CDR (Li et al., 2016) is a high-quality, community-annotated corpus that is essential for biomedical text-mining and relation extraction.
- DWIE (Zaporojets et al., 2021) is a document-level, entity-centered information extraction dataset covering four subtasks. It is sourced from Deutsche Welle's English online content, providing realistic annotations and rule labels for evaluating DocRE methods, with 602 training, 98 development, and 99 test documents.
- Re-DocRED (Tan et al., 2022b) is constructed based on the DocRED (Yao et al., 2019) dataset to address the annotation incompleteness in the original dataset and undergoes further manual verification. Its training set is derived from the training set of DocRED, comprising a total of 3,053 documents. The development and test sets are split from the original development set, each containing 500 documents, and these subsets are also manually re-verified.
- DocGNRE (Li et al., 2023) is an enhanced version of Re-DocRED (Tan et al., 2022b) dataset. It is built upon the original Re-DocRED by semi-automatically supplementing missing relation triples. DocGNRE follows the same document split as Re-DocRED, with 3,053 documents in the training set and 500 documents in the test set.

C Analyzing Hyperparameters

To analyze the effect of the hyperparameter α on ATGL, we conduct experiments on four DocRE datasets. As shown in **Fig. 4**, ATGL exhibits stable performance over a wide range of α values. Specifically, the F1 first increases as α grows, reaches a peak, and then shows slight degradation when α becomes excessively large. For instance, the best

Dataset	Task Type	#Rels	#Triples	#Triples/#Rels	Selected α
CDR	Binary (no NA)	2	5,240	2620.0	7.5
DWIE	Multi-label	65	14,403	221.6	6.0
Re-DocRED	Multi-label	96	85,932	895.1	16.5
DocGNRE	Multi-label	96	96,505	1005.3	20.0

Table 10: Dataset statistics for selecting α.



(a) CDR (b) DWIE

(c) ReDocRED (d) DocGNRE

Figure 4: Performance of ATGL loss across different α values on four datasets, using $\text{BERT}_{\text{base}}$ with ATLOP.

performance is achieved at α=6 on DWIE, α=16.5 on Re-DocRED, α=20 on DocGNRE and α=7.5 on CDR. These results indicate that while ATGL is generally robust to the choice of α, moderate tuning within the high-performing region yields additional gains without requiring exhaustive search.

From these results and the dataset statistics in **Table 10**, we derive a practical guideline for selecting α. For typical multi-label DocRE benchmarks with a dominant NA class, α is mainly determined by the label space size (#Rels), while #Facts/#Rels can be used as a secondary cue for minor adjustment. In practice, we first choose a default α_0 according to #Rels (e.g., around 7 for smaller label spaces and 16–20 for larger ones), and then optionally refine it using #Facts/#Rels. If a development set is available, a simple 3-point search around α_0, i.e., $\{0.5\alpha_0, \alpha_0, 2\alpha_0\}$, is usually sufficient. Since CDR is a binary task without an $\mathcal{NA}$ label, its optimal α is not directly comparable; for binary or non-NA settings, a larger α (e.g., $\alpha \approx 7.5$) is recommended.