Data Augmentation via Relation-Variant Discourse Recasting for Implicit Discourse Relation Recognition

Yikang Yang[†], **Weihao Sun**[†], **Yu Hong**[*], **Zhipang Wang, Guodong Zhou**
School of Computer Science and Technology, Soochow University, Suzhou, China
{ykyang203922, whsun16, tianxianer, zhipangwang}@gmail.com; gdzhou@suda.edu.cn

Abstract

Implicit discourse relation recognition (abbr., IDRR) is a task of automatically classifying semantic relations between pairwise arguments in the absence of explicit connectives. IDRR suffers from imbalanced data distribution of different relations. The current data augmentation approaches have relieved this problem to some extent. However, the incorporated discourse samples during augmentation lack customized contrastive effects. To address this issue, we propose to construct topic-relevant relation-inconsistent arguments for augmentation. Specifically, we implement a **R**elation-**V**ariant **D**iscourse **R**ecasting (RVDR) approach. RVDR utilizes Qwen3-4B to paraphrase discourses under the forcible constraints of preserving inherent topics and transforming relation senses toward infrequent target relations. To avoid excessive divergent thinking of RVDR, we optimize the Qwen3-4B via reinforcement learning, where Group Relative Policy Optimization (GRPO) is used along with multiple task-specific rewards. Furthermore, rule-based filtering and confidence estimation are used to ensure data quality, which additionally enable two application modes, i.e., Confidence-based and Progressive Data Augmentation (i.e., CDA and PDA). Experiments on PDTB 2.0 and 3.0 show that RVDR-based CDA consistently improves IDRR performance across multiple backbones, including both encoder-only and decoder-only cases. It also allows the strong IDRR frameworks CP-KD and SCDN to be enhanced further. The code is publicly available.[1]

1 Introduction

IDRR aims to identify the semantic relation between two text spans (*Arg1* and *Arg2*), when explicit discourse connectives such as *because* or *however* are absent (Lin et al., 2009; Cai et al., 2024). Compared with explicit discourse relation recognition, IDRR is substantially more difficult because the relation must be inferred from latent semantic cues instead of using explicit connectives (Lin et al., 2009). Current IDRR studies still suffer from two issues as follows:

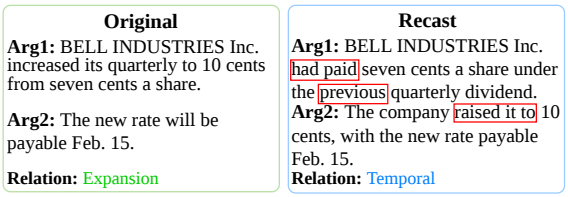


Figure 1: Example of RVDR for IDRR.

- The PDTB (Prasad et al., 2008; Webber et al., 2019) training datasets are highly imbalanced. In PDTB 2.0, *Expansion* accounts for 53.8%, whereas *Temporal* accounts for only 5.3%.
- Existing augmentation approaches can alleviate the imbalance (Jiang et al., 2021; Omura et al., 2024), but lack customized contrastive effects for relation transformation.

Although existing data augmentation approaches can introduce more training data, the introduced samples do not provide sufficient discriminative information for improving decision boundaries. Therefore, we need customized contrastive samples that are more challenging for models by keeping similar content while changing the discourse relation sense. Such samples encourage models to perceive subtle cues for relation discrimination. As shown in Figure 1, the red boxes highlight the subtle cues that support the relation change. In conclusion, customized contrastive samples refer to topic-relevant relation-inconsistent argument pairs, which preserve the inherent topics of the source arguments but express different relation senses.

To generate such samples with customized contrastive effects, we propose **R**elation-**V**ariant

[†]These authors contributed equally to this work.
[*]Corresponding author.
[1]https://github.com/yyk203922/RVDR

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 20430–20444
October 24-29, 2026 ©2026 Association for Computational Linguistics

Discourse **R**ecasting (RVDR). Given an argument pair, RVDR recasts it into a new argument pair expressing an infrequent target relation, while preserving the inherent topics. In this way, the original sample and the recast sample form a topic-relevant relation-inconsistent pair, providing customized contrastive supervision for IDRR. To avoid excessive divergence during recasting, we optimize Qwen3-4B (Yang et al., 2025) with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and multiple task-specific rewards. These rewards cover format validity, topic preservation, relation transformation and shortcut prevention. After recasting, we apply rule-based filtering and confidence estimation to retain reliable recast samples.

Based on the retained recast samples, RVDR supports two augmentation modes. Confidence-based Data Augmentation (CDA) incorporates only high-confidence recast samples into the training set, aiming to maximize sample reliability. Progressive Data Augmentation (PDA) further explores whether lower-confidence samples can be gradually introduced during training. In our experiments, we use RVDR-based CDA as the default setting because high-confidence recast samples provide more reliable supervision for IDRR.

We evaluate RVDR on PDTB 2.0 and 3.0. Experimental results show that RVDR-based CDA consistently improves IDRR performance across multiple backbones, covering both encoder-only and decoder-only settings. Moreover, RVDR can further enhance strong IDRR frameworks CP-KD (Wu et al., 2023) and SCDN (Wang et al., 2025), showing that our method is not tied to a specific backbone or training framework.

Our contributions are summarized as follows:

- We propose RVDR, a novel relation-variant discourse recasting framework. It constructs topic-relevant but relation-inconsistent argument pairs to provide customized contrastive effects. RVDR provides a controllable way to alleviate data imbalance in IDRR.
- Experiments on PDTB show that RVDR consistently outperforms existing augmentation methods. Notably, it achieves superior performance without relying on external data.
- Further experiments show that RVDR generalizes across multiple backbones, and can further enhance strong IDRR frameworks such as CP-KD and SCDN.

2 Approach

In this section, we first introduce the task formulation and the overall RVDR framework in Section 2.1. We next introduce relation-variant discourse recasting in Section 2.2, followed by GRPO-based recasting optimization in Section 2.3. Finally, we introduce recast sample filtering and the CDA/PDA augmentation modes for supervised IDRR training in Sections 2.4 and 2.5.

2.1 Task Formulation and Overview

We formulate IDRR as the task of predicting a relation sense from a predefined label set $\mathcal{Y}$. Our primary evaluation considers four top-level relation senses: *Comparison*, *Contingency*, *Expansion*, and *Temporal*. We also evaluate second-level classification in an 11-way setting for PDTB 2.0 and a 14-way setting for PDTB 3.0.

We consider both encoder-only and decoder-only IDRR settings. For encoder-only models, IDRR is implemented as a standard multiclass classification task, where the two arguments are encoded as a paired input. For decoder-only models, IDRR is formulated as a multiple-choice generation task. Given *Arg1*, *Arg2*, and the candidate relation senses in $\mathcal{Y}$, the model is prompted to select one option from a predefined answer space as follows:

What is the relation between *Arg1* and *Arg2*?
A. *Contingency* B. *Expansion*
C. *Comparison* D. *Temporal*

For encoder-only and decoder-only IDRR settings, RVDR enhances model training through data augmentation. As illustrated in Figure 2, the framework consists of GRPO-based discourse recasting, rule-based and confidence-based sample filtering, and IDRR training with the retained samples.

2.2 Relation-Variant Discourse Recasting

RVDR constructs customized contrastive samples. Given a source sample $x_s = (a_1, a_2, y_s)$, RVDR selects an infrequent target relation sense y_t different from y_s and recasts it:

$$\tilde{x}_t = (R_\theta(a_1, a_2, y_s, y_t), y_t), \quad (1)$$

where R_θ outputs the recast argument pair. Rather than freely generating a new discourse sample, RVDR constructs a topic-relevant relation-inconsistent sample that preserves the source content while expressing the target relation sense.

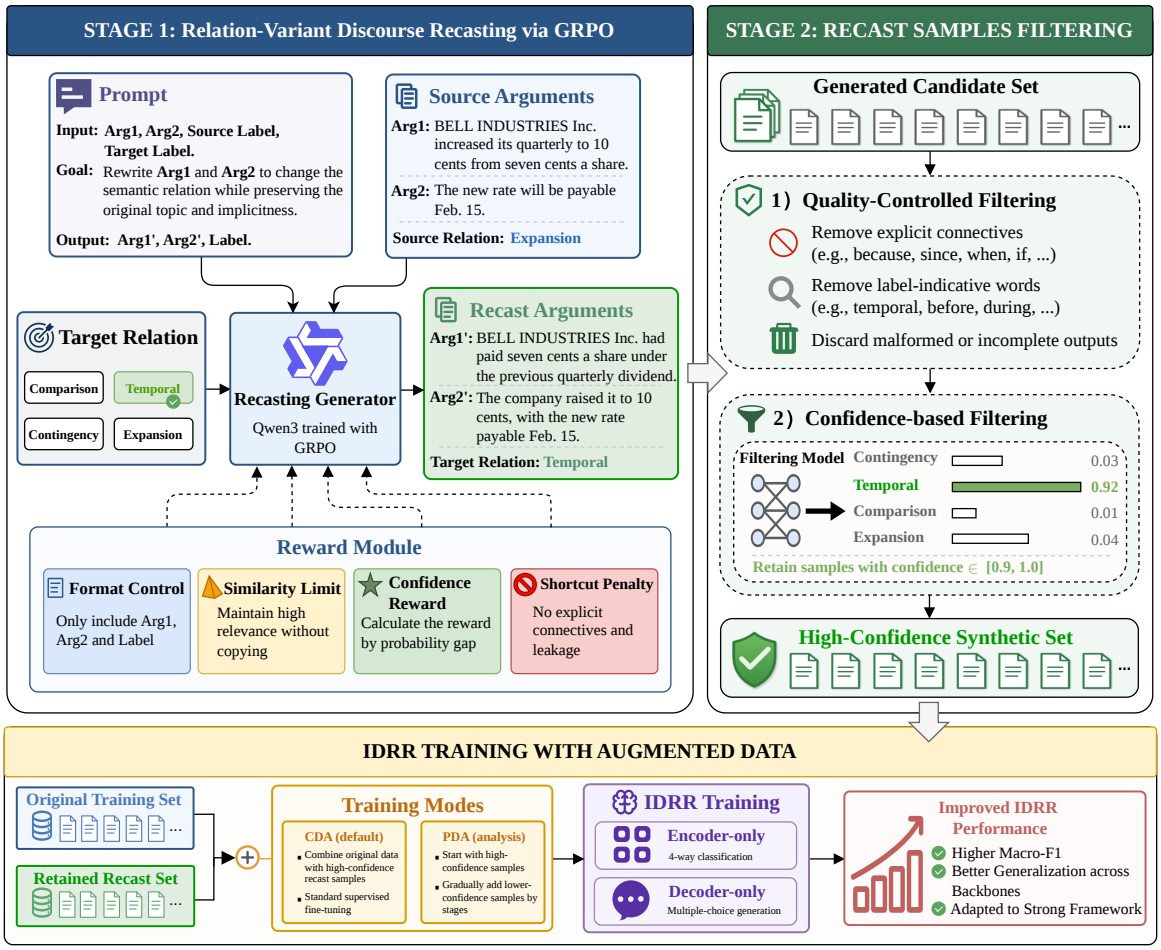


Figure 2: Overview of our proposed **R**elation-**V**ariant **D**iscourse **R**ecasting (RVDR) framework. RVDR first generates topic-relevant relation-inconsistent candidates, then filters them with rule-based and confidence-based criteria, and finally incorporates the retained recast samples into IDRR training through CDA or PDA.

Concretely, for a source sample $x = (a_1, a_2, y_s)$ and an infrequent target relation y_t with $y_t \neq y_s$, the recaster generates a recast sample $\tilde{x} = (\tilde{a}_1, \tilde{a}_2, y_t)$. A useful recast sample should satisfy three requirements. First, it should express the infrequent target relation sense. Second, it should remain grounded in the source topics and events. Third, it should remain implicit and free from explicit discourse connectives, label-indicative cues, or rigid template statements.

We use Qwen3-4B as the recasting backbone. For top-level classification, RVDR trains a dedicated recaster for each infrequent target relation sense. For second-level classification, which involves more target senses, we train a single recaster conditioned on the target sense to reduce computational cost. Standard supervised fine-tuning is not directly applicable because no gold recast samples are available, and prompting provides limited control over recasting quality. Therefore, we optimize the recasting process with GRPO and multiple task-specific rewards, as described in next section.

2.3 GRPO-based Recasting Optimization

RVDR optimizes the recasting process with GRPO. For each source sample and target relation, the recasting model samples a group of G candidate recasts and assigns the g-th candidate a scalar reward R_g. Unlike Proximal Policy Optimization (PPO) (Schulman et al., 2017), which relies on a learned value model for advantage estimation, GRPO computes the advantage from normalized rewards within the same sampled group:

$$A_g = \frac{R_g - \mu_R}{\sigma_R + \epsilon}, \tag{2}$$

where μ_R and σ_R are the mean and standard deviation of rewards in the same group, and ϵ is a small constant for numerical stability. This group-relative advantage does not require an extra critic and reduces instability from inaccurate value estimates, making it particularly suitable for our

reward-guided discourse recasting setting.

The reward is designed to make recast samples valid, topic-relevant, relation-variant, and implicit:

$$R = \lambda_{\text{fmt}} r_{\text{fmt}} + \lambda_{\text{sim}} r_{\text{sim}} + \lambda_{\text{conf}} r_{\text{conf}} + r_{\text{pen}}, \quad (3)$$

where r_{fmt}, r_{sim}, and r_{conf} denote format, similarity, and confidence rewards. λ_{fmt}, λ_{sim}, and λ_{conf} are their weights; and r_{pen} is the penalty term.

The format reward encourages valid structured outputs with *Arg1*, *Arg2*, and *Label* fields. The similarity reward keeps the recast sample grounded in the source sample while avoiding near-copying. The confidence reward measures whether the recast sample moves toward the infrequent target relation sense. The penalty term suppresses shortcut behaviors, including explicit connectives and label-indicative cues. Detailed reward definitions and settings are provided in Appendix C.

2.4 Recast Sample Filtering

After recasting, RVDR generates candidate recast samples for each infrequent target relation. For each target relation, we apply the corresponding recaster to source samples with different source labels. The resulting outputs are collected into a candidate set. To further ensure data quality for training, we apply two filtering stages: rule-based filtering and confidence-based filtering.

Rule-based filtering. The first stage removes malformed outputs, obvious shortcut cases, and rigid or repetitive expressions. We discard a candidate if either recast argument contains a forbidden explicit connective within its first few words. We also remove candidates that introduce label-indicative cues directly revealing the target relation sense, use rigid templates that do not preserve the source content, or have invalid outputs.

Confidence-based filtering. The second stage uses an IDRR model trained on the corresponding PDTB training set as the filter model. Given a remaining recast candidate $\tilde{x}$, the filter model assigns a confidence score $C_f(y \mid \tilde{x})$ to each relation sense $y \in \mathcal{Y}$. For encoder-only classifiers, this score is the softmax probability over relation senses. For decoder-only multiple-choice models, it is computed from the normalized likelihood of generating the option corresponding to each relation.

For the top-level experiments, we apply two criteria to each candidate with target relation sense y_t. First, the predicted relation sense should match the target relation sense:

$$\arg\max_{y \in \mathcal{Y}} C_f(y \mid \tilde{x}) = y_t, \quad (4)$$

Second, the confidence assigned to the target relation sense should be no lower than a threshold:

$$C_f(y_t \mid \tilde{x}) \geq \eta, \quad (5)$$

where C_f denotes the filter model and η is the confidence threshold. We retain recast samples whose target-relation confidence falls in $[0.9, 1.0]$, since high-confidence samples provide more reliable supervision than lower-confidence ones.

We include second-level classification as an additional evaluation of RVDR. For this setting, we do not conduct a separate grid search for the filtering threshold. Because the second-level label space differs from the top-level label space, we also do not reuse the top-level confidence threshold. Instead, after rule-based filtering, we rank the candidates for each target sense according to the confidence assigned to that sense by the filter model and retain the top 5% for training. We use a fixed 5% retention ratio without further tuning, as the experiments are intended to evaluate whether RVDR can be effectively extended to a fine-grained setting.

We use a matched filter model, where the filter model is fine-tuned on the original data and belongs to the same model family as the target IDRR models. Additional analysis of filter-model choice is provided in Section 4.4. After filtering, we obtain the retained recast set, which contains high-confidence recast samples for infrequent relation senses. The retained recast samples are then used by CDA and PDA, as described in Section 2.5.

2.5 Training with Recast Data: CDA and PDA

RVDR expands the imbalanced IDRR training data by incorporating recast samples with customized contrastive effects. After relation-variant discourse recasting and data quality control, the retained recast samples can be used in two application modes:

CDA. Under CDA, the original training set is expanded with high-confidence recasts retained after filtering. Since lower-confidence recasts may contain weak relation shifts or misleading cues, high-confidence recasts are expected to provide more reliable contrastive supervision for IDRR training. This design keeps the augmentation process simple while reducing the potential noise introduced by recast samples. Overall, CDA tends to rely on high-confidence recast data during augmentation.

PDA. To explore whether lower-confidence samples can be effectively utilized to increase data diversity, we introduce PDA as a progressive variant. While a lower confidence threshold retains more recast samples, directly incorporating them risks introducing misleading semantic noise. Therefore, PDA employs a mechanism inspired by curriculum learning, where training starts with strict high-confidence recasts and gradually adds samples from lower confidence ranges in later stages. This design allows models to first learn from cleaner recasts and then gradually incorporate more diverse samples during the learning process.

3 Experiments

3.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate RVDR on two standard IDRR benchmarks, PDTB 2.0 and PDTB 3.0. Top-level classification covers four relations in both datasets: *Comparison*, *Contingency*, *Expansion*, and *Temporal*. Second-level classification covers 11 relations in PDTB 2.0 and 14 relations in PDTB 3.0. Following Ji and Eisenstein (2015), we use sections 2–20 for training, sections 0–1 for development, and sections 21–22 for testing. We report macro-F_1 and accuracy as evaluation metrics. For experiments repeated over three random seeds, we report the mean performance, with smaller trailing numbers denoting sample standard deviations. Detailed dataset statistics and split information are provided in Appendix A.

Implementation Details. We use Qwen3-4B (Yang et al., 2025) as the recasting model backbone trained with GRPO. For IDRR classifier training, we use BERT-base (Devlin et al., 2019), RoBERTa-base, RoBERTa-large (Liu et al., 2019), LLaMA3.1-8B-Instruct, LLaMA3-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025). For top-level experiments, we retain synthetic samples in the confidence range $[0.9, 1.0]$. For second-level experiments, we retain the top 5% within each target sense after rule-based filtering. Additional training details and synthetic data settings are provided in Appendix B.

Baselines. We compare RVDR with three types of data augmentation baselines: (i) EDA (Wei and Zou, 2019), which applies word-level perturbations while keeping the original relation sense unchanged; (ii) Strong Prompted Recasting, which uses Qwen3-Max as the prompted generator with the same source-sample selection, target relation senses, and filtering pipeline as RVDR, but without GRPO-based optimization; and (iii) other representative data augmentation methods from recent IDRR literature, including SDE (Zhou et al., 2020), CVAE (Dou et al., 2021), Omura et al. (2024), and Song et al. (2024). For a fair comparison, EDA and Strong Prompted Recasting add the same number of training instances as RVDR for every dataset, backbone, and target relation.

Range	Mode	PDTB 2.0		PDTB 3.0	
		F_1	*Acc*	F_1	*Acc*
–	Vanilla	61.87	69.63	67.80	73.20
$[0.9, 1.0]$	CDA	**65.73**	**71.19**	**70.25**	**75.03**
$[0.7, 1.0]$	CDA	64.21	70.65	68.49	72.82
	PDA	64.30	71.03	69.09	74.15
$[0.5, 1.0]$	CDA	63.36	69.50	67.67	73.36
	PDA	62.60	70.27	67.52	72.73

Table 1: Effect of CDA and PDA with different confidence ranges using RoBERTa-base.

3.2 Comparison of CDA and PDA

We first examine how the retained recast samples should be incorporated into IDRR training. As shown in Table 1, CDA with only high-confidence samples achieves the best overall performance. When medium-confidence samples are included by expanding the range to $[0.7, 1.0]$, PDA outperforms direct CDA, suggesting that progressive incorporation helps the model exploit additional recast samples while reducing moderate noise. However, further expanding the range to $[0.5, 1.0]$ leads to performance degradation, indicating that low-confidence samples introduce weaker or misleading relation cues that PDA cannot fully offset. We therefore use CDA as the default augmentation mode in the following experiments, with the confidence threshold selected on the development set. The corresponding threshold sensitivity is further analyzed in Section 4.3. Unless otherwise specified, RVDR refers to this default setting.

3.3 Main Results

We compare RVDR with the three types of baselines described above to examine whether it can provide more effective augmentation without relying on external discourse samples.

As shown in Table 2, RVDR achieves the

Backbone Model	Method	External Data	PDTB 2.0		PDTB 3.0	
			F_1	*Acc*	F_1	*Acc*
BERT-base	EDA (Wei and Zou, 2019)	None	54.23*	63.54*	61.66*	67.73*
	SDE (Zhou et al., 2020)	Required	51.16	–	–	–
	Strong Prompted Recasting	None	54.77	62.53	63.34	68.77
	RVDR	None	**56.53**	**63.86**	**64.10**	**69.13**
RoBERTa-base	EDA (Wei and Zou, 2019)	None	62.01*	69.12*	67.37*	72.93*
	CVAE (Dou et al., 2021)	Required	65.06	70.17	–	–
	Omura et al. (2024)	None	–	–	70.00	74.70
	Song et al. (2024)	Required	64.29	69.75	69.94	74.98
	Strong Prompted Recasting	None	62.50	70.08	68.07	73.61
	RVDR	None	**65.73**	**71.19**	**70.25**	**75.03**

Table 2: Comparison with representative data augmentation baselines for top-level IDRR. Results are grouped by downstream classifier backbone; * denotes reproduced results.

Method	PDTB 2.0		PDTB 3.0	
	F_1	*Acc*	F_1	*Acc*
CP-KD-base	68.05*	74.19*	71.08*	75.51*
+ RVDR	**69.76**	**75.24**	**72.43**	**77.07**
CP-KD-large	71.19*	75.72*	74.62*	77.82*
+ RVDR	**71.73**	**76.39**	**75.08**	**78.02**
SCDN	67.15*	74.47*	72.67*	76.66*
+ RVDR	**68.03**	**75.05**	**73.52**	**77.82**

Table 3: Top-level compatibility results of RVDR with strong IDRR frameworks on PDTB 2.0 and PDTB 3.0. Within each framework pair, better results are bolded and * denotes reproduced results.

Method	PDTB 2.0		PDTB 3.0	
	F_1	*Acc*	F_1	*Acc*
BERT-base				
Vanilla	53.79 0.28	63.10 0.35	62.66 0.11	68.02 0.55
RVDR	**56.53** 0.62	**63.86** 0.29	**64.10** 0.38	**69.13** 0.36
RoBERTa-base				
Vanilla	61.87 1.58	69.63 0.87	67.80 1.14	73.20 0.51
RVDR	**65.73** 0.46	**71.19** 0.73	**70.25** 0.13	**75.03** 0.14
LLaMA3.1-8B-Instruct				
Vanilla	65.58 0.96	72.66 1.01	70.81 0.20	75.14 0.41
RVDR	**67.50** 0.62	**73.14** 0.29	**72.67** 0.86	**76.59** 0.79
Qwen3-8B				
Vanilla	62.74 0.80	69.38 1.20	69.68 0.85	74.43 0.85
RVDR	**64.63** 0.34	**71.70** 0.51	**70.98** 0.80	**75.37** 0.82

Table 4: Generalization of RVDR across multiple backbones on PDTB 2.0 and PDTB 3.0.

strongest results on both datasets under both classifier backbones. Compared with EDA, RVDR improves macro-F_1 by 2.30–3.72 points with BERT-base and RoBERTa-base on both datasets. This result shows that RVDR provides more effective augmentation than generic word-level perturbations. RVDR also outperforms Strong Prompted Recasting in every matched setting, indicating that direct prompting with Qwen3-Max does not provide the same augmentation benefit as task-specific reward-guided recasting. RVDR further outperforms representative IDRR augmentation baselines across the reported metrics. Taken together, these results show that RVDR provides more effective supervision than the representative baselines without requiring additional discourse data.

3.4 Compatibility with Strong Frameworks

To examine whether RVDR can complement strong IDRR frameworks, we add the retained recast samples to the top-level training data of CP-KD and SCDN. Table 3 compares each framework with and without RVDR on PDTB 2.0 and PDTB 3.0.

The results show that RVDR improves both macro-F_1 and accuracy when integrated with CP-KD and SCDN. CP-KD-base obtains the largest gains, with macro-F_1 increases of 1.71 points on PDTB 2.0 and 1.35 points on PDTB 3.0, accompanied by accuracy gains of 1.05 and 1.56 points. The improvements for CP-KD-large and SCDN are more modest but remain positive on both datasets, showing that RVDR can provide complementary supervision when combined with strong IDRR frameworks.

3.5 Generalization to Multiple Backbones

RVDR generalizes well across different backbone architectures. As shown in Table 4, RVDR improves all four representative backbones on both PDTB 2.0 and PDTB 3.0, covering both encoder-only PLMs and decoder-only LLMs.

Backbone Model	Parameters	Method	PDTB 2.0		PDTB 3.0	
			F_1	*Acc*	F_1	*Acc*
RoBERTa-base	125M	Full RVDR	**65.73**0.46	**71.19**0.73	**70.25**0.13	**75.03**0.14
		w/o rule filtering	63.74 0.35	70.65 0.44	69.16 0.78	73.88 0.72
		w/o conf.-based filtering	61.68 0.52	69.50 0.58	67.76 0.88	73.07 0.83
		w/o GRPO	62.37 0.27	68.87 0.31	67.73 0.63	72.64 0.66
LLaMA3.1-8B-Instruct	8.03B	Full RVDR	**67.50**0.62	**73.14**0.29	**72.67**0.86	**76.59**0.79
		w/o rule filtering	66.73 1.22	72.75 1.13	71.60 0.68	75.78 1.22
		w/o conf.-based filtering	65.74 0.47	71.91 0.77	71.85 1.38	75.98 0.43
		w/o GRPO	65.60 0.91	72.56 1.32	70.68 1.12	74.49 0.96

Table 5: Top-level ablation results of RVDR on encoder-only and decoder-only backbones.

Method	PDTB 2.0		PDTB 3.0	
	F_1	*Acc*	F_1	*Acc*
RoBERTa-base				
Vanilla	39.02 1.63	58.04 0.25	49.68 1.11	61.89 0.40
RVDR	**43.26**0.25	**60.41**0.53	**55.77**0.42	**64.57**0.66
LLaMA3.1-8B-Instruct				
Vanilla	43.40 0.09	59.80 0.64	54.07 1.62	63.55 1.24
RVDR	**47.33**0.87	**62.69**0.39	**58.15**0.46	**66.80**0.80

Table 6: Generalization of RVDR to fine-grained relation classification on PDTB 2.0 and PDTB 3.0.

Dataset	Setting	Comp.	Cont.	Exp.	Temp.
PDTB2	Vanilla	62.71 1.77	62.63 0.71	77.33 0.74	44.83 3.86
	RVDR	**65.10**2.17	**64.49**0.79	**77.92**1.09	**55.43**4.68
PDTB3	Vanilla	61.35 3.03	74.64 0.45	78.67 0.21	56.56 1.51
	RVDR	**64.46**2.25	**77.21**0.35	**78.95**0.25	**60.38**2.31

Table 7: Binary classification macro-F_1 of RoBERTa-base for top-level relations on PDTB 2.0 and PDTB 3.0.

The gains are observed in both macro-F_1 and accuracy. These improvements across all evaluated backbones and both datasets indicate that the benefit is not restricted to a particular dataset, metric, or model family. Together, these results suggest that RVDR can serve as a stable and general data augmentation strategy for different IDRR backbones.

3.6 Generalization to Fine-Grained Relations

RVDR also improves both macro-F_1 and accuracy in second-level classification across both backbones and datasets. As reported in Table 6, the largest macro-F_1 gains occur with RoBERTa-base, reaching 4.24 points on PDTB 2.0 and 6.09 points on PDTB 3.0. Consistent improvements are also observed with LLaMA3.1-8B-Instruct on both datasets. These results show that the recast samples remain useful for distinguishing fine-grained relation senses.

4 Analysis

4.1 Ablation Study

To evaluate the contribution of the main components in RVDR, we conduct ablation experiments on encoder-only and decoder-only backbones. Table 5 reports results on RoBERTa-base and LLaMA3.1-8B-Instruct. The w/o GRPO variant uses only prompting with the base Qwen3-4B, while the other variants remove rule-based filtering or confidence-based filtering.

Full RVDR achieves the best macro-F_1 across both backbones on PDTB 2.0 and PDTB 3.0. Removing GRPO causes consistent degradation, showing that using only prompting is insufficient for relation-variant recasting. Removing either filtering stage also reduces performance, confirming that rule-based filtering and confidence-based filtering are both important for retaining useful recast samples. The degradation is especially clear when confidence-based filtering is removed, indicating that target-relation confidence is critical for data augmentation. These trends are consistent across encoder-only and decoder-only settings, suggesting that both recasting optimization and filtering are necessary for generating reliable recast data.

4.2 Binary Classification Performance

To examine how RVDR affects each relation sense under imbalanced data distribution, Table 7 reports binary classification macro-F_1 scores of RoBERTa-base for each top-level relation. This evaluation shows whether the overall gain is mainly driven by improved recognition of infrequent relation senses.

RVDR improves all relations on PDTB 2.0 and PDTB 3.0, showing that the recast samples do not improve one relation by sacrificing another. On

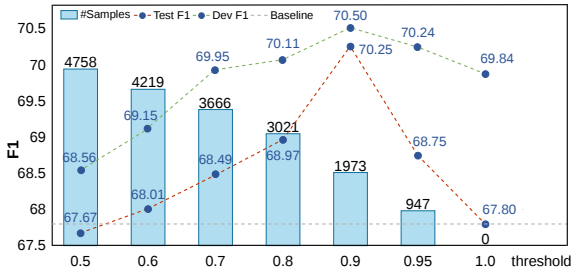


Figure 3: Effect of confidence-based filtering on PDTB 3.0 for top-level relation classification with RoBERTa-base. Results are averaged over three random seeds.

PDTB 2.0, the largest gain appears on *Temporal*, increasing from 44.83 to 55.43, while *Comparison*, *Contingency*, and *Expansion* also improve. On PDTB 3.0, RVDR brings clear gains on *Comparison* and *Temporal*, with only a small change on *Expansion*. These results suggest that RVDR strengthens recognition of infrequent relation senses while maintaining performance on frequent ones.

4.3 Effect of Confidence-based Filtering

Our main setting keeps recast samples in the confidence range $[0.9, 1.0]$, with the lower threshold tuned on the development set. Figure 3 further examines the sensitivity of macro-F_1 to this threshold by gradually raising it from 0.5 to 1.0. Development performance reaches its best result at 0.9, which is therefore adopted in our main experiments. On the test set, as the threshold increases from 0.5 to 0.9, fewer recast samples are retained while macro-F_1 generally improves, suggesting that many low-confidence samples introduce harmful noise. Raising the threshold further to 0.95 reduces performance, indicating that overly strict filtering limits the diversity of retained recast samples and weakens the augmentation effect. When the threshold reaches 1.0, no recast samples are retained and performance returns to the vanilla RoBERTa-base level. Overall, the test-set sensitivity trend is consistent with the development-set selection, suggesting that the selected threshold effectively filters low-confidence samples without excessively reducing data coverage.

4.4 Effect of Model Choice for Filtering

We further examine whether RVDR depends on a specific filter model. By default, RVDR uses a backbone-matched filter, where the filter model belongs to the same model family as the target IDRR model. This setting is intuitive because different backbones may assign confidence scores with different calibration behaviors. As shown in Table 8, matched filters generally achieve the best or near-best performance. Meanwhile, cross-model filters remain competitive, although their confidence scores are produced by different model families. This suggests that the benefit of RVDR is not solely caused by model-specific sample filtering. Instead, the retained recast samples have general utility across different models.

Backbone Model	Filter	F_1	*Acc*
RoBERTa-base	RoBERTa	**70.25**0.13	**75.03**0.14
	LLaMA	69.580.64	74.080.36
	Qwen	69.630.46	74.220.77
LLaMA3.1-8B-Instruct	RoBERTa	71.920.68	75.920.92
	LLaMA	**72.67**0.86	**76.59**0.79
	Qwen	72.310.74	76.121.14
Qwen3-8B	RoBERTa	70.330.42	74.900.48
	LLaMA	**71.01**0.94	75.240.80
	Qwen3	70.980.80	**75.37**0.82

Table 8: Top-level results of different filters applied to various backbone models on PDTB 3.0.

5 Related Work

5.1 Implicit Discourse Relation Recognition

Early studies on IDRR mainly focused on improving model architectures and training strategies. With the development of pretrained language models, subsequent work improved relation recognition through stronger contextual representations and prompt-based learning (Kishimoto et al., 2020; Chan et al., 2023). Beyond representation learning, SCIDER (Cai et al., 2024) enhances logical semantics through semantics confrontation, while CP-KD (Wu et al., 2023) introduces connective prediction as auxiliary supervision and uses knowledge distillation to improve IDRR.

With the rapid development of LLMs, recent studies have begun to apply LLMs to IDRR by exploiting their generation and reasoning abilities. For example, SCDN (Wang et al., 2025) uses generated subtext to provide auxiliary semantic signals, while EIDRR (Wang and Wu, 2026) uses LLM-generated natural language explanations to improve both performance and interpretability for IDRR.

Despite these advances, to alleviate the data imbalance problem in IDRR, some studies have explored data augmentation methods that construct additional discourse instances and exploit noise-control strategies. Zhou et al. (2020) generate

pseudo implicit samples and use policy-based strategy to filter instances. Dou et al. (2021) propose a latent-space data augmentation method with risk-aware sampling control. Song et al. (2024) exploit explicit discourse instances for data augmentation and select those most similar to natural implicit instances. Omura et al. (2024) generate synthetic data from hard negative samples. However, existing data augmentation methods provide limited guidance for models to capture the subtle cues that distinguish relation differences, particularly for infrequent relation senses. To address this limitation, RVDR generates topic-relevant relation-inconsistent samples for IDRR.

5.2 Text Data Augmentation via RL

Reinforcement learning (RL) has been explored to make text data augmentation more task-oriented. Instead of relying on heuristic perturbations or unconstrained generation, RL-based methods use task-specific rewards to guide augmented sample construction. Data Boost uses RL to guide conditional generation in low-resource classification (Liu et al., 2020). AugmenToxic applies PPO-based optimization to generate toxic-text samples for imbalanced toxicity detection (Bodaghi et al., 2025). Event AutoAugment learns word-level augmentation policies for event detection (Gao et al., 2025). Unlike prior work, where rewards primarily serve as outcome-level signals, RVDR treats rewards as a behavioral specification for discourse relation recasting. Concretely, RVDR combines complementary rewards with shortcut penalties, encouraging helpful recasts while suppressing reward hacking.

6 Conclusion

In this paper, we propose Relation-Variant Discourse Recasting (RVDR), a novel data augmentation framework for IDRR. RVDR generates topic-relevant relation-inconsistent samples by recasting source samples toward infrequent relation senses, and uses GRPO-based training and additional filtering to control sample quality. Experiments on PDTB 2.0 and PDTB 3.0 show that RVDR-based CDA consistently improves encoder-only and decoder-only backbones, and further enhances strong IDRR frameworks. These results demonstrate the effectiveness of RVDR without relying on external discourse data. In future work, we will explore whether our method can generalize to multimodal discourse relation studies.

Limitations

In this section, we discuss several limitations of our work under the current setting, which can be summarized in the following two aspects.

First, our work focuses on an augmentation method for improving IDRR models. Although the recast samples are used effectively for model training, we do not make the categorization of recasting patterns a central focus of this paper. Some relation transformations may be realized through local cue changes in one argument, while others may involve coordinated changes across both arguments. A more detailed qualitative analysis of such recasting behaviors could provide additional insight into how relation-variant samples are formed.

Second, the GRPO-based recasting stage introduces higher computational overhead than conventional prompting-based augmentation methods. Although the recasting model is used only for offline synthetic data construction, reinforcement learning requires additional rollout and reward computation, increasing both training time and hardware cost.

Acknowledgments

This research was supported by National Natural Science Foundation of China under No.62376182.

References

Arezo Bodaghi, Benjamin C. M. Fung, and Ketra A. Schmitt. 2025. Augmentoxic: Leveraging reinforcement learning to optimize llm instruction fine-tuning for data augmentation to enhance toxicity detection. *ACM Trans. Web*, 19(4).

Mingyang Cai, Zhen Yang, and Ping Jian. 2024. Improving implicit discourse relation recognition with semantics confrontation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8828–8839, Torino, Italia. ELRA and ICCL.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

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 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zujun Dou, Yu Hong, Yu Sun, and Guodong Zhou. 2021. CVAE-based re-anchoring for implicit discourse relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1275–1283, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jun Gao, Huan Zhao, Wei Wang, Changlong Yu, and Ruifeng Xu. 2025. Event autoaugment: Word-level data augmentation for textual event detection. *Expert Systems with Applications*, 288:127740.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An empirical study of synthetic data generation for implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Wei Song, Hongfei Han, Xu Han, Miaomiao Cheng, Jiefu Gong, Shijin Wang, and Ting Liu. 2024. Discriminative explicit instance selection for implicit discourse relation classification. *Frontiers of Computer Science*, 18:184340–.

Heng Wang and Changxing Wu. 2026. Improving implicit discourse relation recognition with natural language explanations from llms. *Preprint*, arXiv:2602.21763.

Zhipang Wang, Yu Hong, Weihao Sun, and Guodong Zhou. 2025. Using subtext to enhance generative IDRR. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 464–473, Vienna, Austria. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse Treebank 3.0 Annotation Manual*. University of Pennsylvania, Philadelphia.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Meilin Zhou, Qi Liang, Lu Ma, Dan Luo, Peng Zhang, and Bin Wang. 2020. Towards selective data enhanced implicit discourse relation recognition via reinforcement learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

A Dataset Statistics

We evaluate our method on PDTB 2.0 (Prasad et al., 2008) and PDTB 3.0 (Webber et al., 2019). Detailed top-level dataset statistics are shown in Table 9 and Table 10. For the fine-grained experiments, we further report the training-set distributions of the 11-way and 14-way second-level relation inventories in Table 11 and Table 12, respectively.

Relation	Train	Dev	Test
Comparison	1,894	191	146
Contingency	3,281	287	276
Expansion	6,792	651	556
Temporal	665	54	68
Total	12,632	1,183	1,046

Table 9: Data statistics of PDTB 2.0 for top-level relation classification.

Relation	Train	Dev	Test
Comparison	1,830	190	154
Contingency	5,896	579	529
Expansion	7,941	748	643
Temporal	1,418	136	148
Total	17,085	1,653	1,474

Table 10: Data statistics of PDTB 3.0 for top-level relation classification.

Second-level relation sense	**Train**
Comparison.Concession	180
Comparison.Contrast	1,566
Contingency.Cause	3,227
Contingency.Pragmatic cause	51
Expansion.Alternative	146
Expansion.Conjunction	2,805
Expansion.Instantiation	1,061
Expansion.List	330
Expansion.Restatement	2,376
Temporal.Asynchronous	517
Temporal.Synchrony	147
Total	12,406

Table 11: Training-set distribution for the 11-way second-level relation inventory of PDTB 2.0.

Second-level relation sense	**Train**
Comparison.Concession	1,164
Comparison.Contrast	639
Contingency.Cause	4,469
Contingency.Cause+Belief	157
Contingency.Condition	152
Contingency.Purpose	1,102
Expansion.Conjunction	3,584
Expansion.Equivalence	252
Expansion.Instantiation	1,117
Expansion.Level-of-detail	2,493
Expansion.Manner	191
Expansion.Substitution	278
Temporal.Asynchronous	985
Temporal.Synchronous	433
Total	17,016

Table 12: Training-set distribution for the 14-way second-level relation inventory of PDTB 3.0.

B Additional Implementation Details

Recasting model training. We use Qwen3-4B as the recasting model backbone and train it with GRPO using LoRA (Hu et al., 2021). The actor learning rate for RL training is set to 5e-5, and training is run for 3 epochs. We use a relatively compact recasting model backbone to balance generation quality and the computational cost of reward-guided training. The RL-trained recasting model is used only for offline recast data construction and is not used as the IDRR classifier.

LoRA and systems settings. We use LoRA with rank 32, and apply LoRA to all linear layers. Rollout is implemented with vLLM (Kwon et al., 2023) using tensor model parallel size 2. RL training is conducted on two GPUs. All experiments are conducted on RTX 4090 GPUs.

Multiple backbone models. For supervised IDRR training, we use BERT-base, RoBERTa-base, RoBERTa-large, Qwen3-8B, LLaMA3.1-8B-Instruct, and LLaMA3-8B-Instruct as backbone models. All models are trained for 5 epochs. For BERT-base, RoBERTa-base, and RoBERTa-large, we use a batch size of 16, with learning rates of 1e-5, 1e-5, and 5e-6, respectively. For Qwen3-8B, LLaMA3.1-8B-Instruct, and LLaMA3-8B-Instruct, we use LoRA-based fine-tuning with rank 16, batch size 1, and learning rate 5e-5.

C Additional Details on Reward Design

This section provides additional details of the reward components used in RVDR. The reward is computed after a complete recast sample is generated, because topic preservation, relation sense transformation, implicitness, and output validity are sequence-level properties.

Format reward. The format reward ensures that the recasting model consistently produces outputs

that can be parsed and evaluated. Since all filtering and reward computation depend on the generated *Arg1*, *Arg2*, and *Label* fields, malformed or incomplete outputs are unusable for the rest of the pipeline. In preliminary trials, outputs without explicit format control sometimes contained explanations, extra text, or missing fields, making reward computation unstable. Concretely, we first attempt to parse the model output as JSON, and only retain outputs that contain non-empty *Arg1*, *Arg2*, and *Label* fields. Valid structured outputs receive a positive format reward, while invalid or incomplete outputs receive an immediate negative score. This reward serves as a basic training stabilizer and makes reward computation reliable.

Similarity reward. The similarity reward balances topic preservation against trivial copying. A useful recast sample should remain grounded in the source content, but it should not copy the source arguments. If the recast sample is too dissimilar, it may drift away from the source discourse. If it is too similar, it provides little new supervision.

In implementation, we compute semantic similarity separately between source *Arg1* and recast *Arg1*, and between source *Arg2* and recast *Arg2*, using BGE-large embeddings (Xiao et al., 2023) and cosine similarity. The lower of the two scores is used as the final similarity score s, so that both arguments are required to remain semantically grounded. We reward a moderate similarity interval $[\tau_l, \tau_u]$ rather than copying:

$$r_{\text{sim}} = \begin{cases} s - \beta(\tau_l - s), & s < \tau_l, \\ s, & \tau_l \le s \le \tau_u, \\ \tau_u - \alpha(s - \tau_u), & s > \tau_u, \end{cases} \quad (6)$$

where α and β control the penalty strengths. Scores below τ_l are penalized as semantic drift, while scores above τ_u are penalized as near-copying.

Relative confidence reward. The confidence reward encourages the recast sample to move toward the target relation. A straightforward objective is to directly maximize the classifier confidence on the target relation. In preliminary experiments, however, this absolute confidence objective produced less reliable recasts: the recasting model tended to increase classifier confidence through explicit markers, label-indicative cues, or rigid template statements, rather than through natural implicit relation transformation. This is undesirable for IDRR, because the sample should remain implicit.

We therefore use relative confidence gain instead of absolute confidence. During GRPO training, the confidence score is provided by a frozen RoBERTa-large reward model that is fine-tuned on the corresponding PDTB training split. This reward model is used only to compute the confidence reward and is not updated during recasting optimization. Let p_{old} and p_{new} denote the reward-model confidence on the target relation sense before and after recasting, respectively, and let $\Delta p = p_{\text{new}} - p_{\text{old}}$. The confidence reward is defined as:

$$r_{\text{conf}} = \lambda_p \sqrt{[\Delta p]_+} - \lambda_n \sqrt{[-\Delta p]_+} - \lambda_o [p_{\text{new}} - c]_+. \quad (7)$$

where $[z]_+ = \max(z, 0)$ and c is the confidence ceiling. The first term rewards positive confidence gain, the second term penalizes confidence decrease, and the third term penalizes overconfidence.

The confidence ceiling is used to reduce shortcut-driven overconfidence during reward optimization. During GRPO training, multiple recasts for the same source sample and target relation sense are compared within a group. If target-relation confidence becomes excessively high across candidates, the relative reward signal may be dominated by confidence maximization and may favor shortcut patterns rather than meaningful differences in discourse quality. Therefore, we penalize confidence values above a ceiling to prevent the recasting model from over-optimizing this reward.

This design does not conflict with confidence-based filtering. The confidence ceiling regularizes the training objective by limiting excessive confidence during reward optimization, whereas confidence-based filtering selects reliable recast samples. In other words, the ceiling discourages degenerate overconfidence in the recasting model, while the filtering step still retains samples whose target relation is confidently recognized after passing validity and quality constraints.

Anti-shortcut constraints. The confidence reward is further combined with anti-shortcut constraints. For each target relation sense, we define a lexicon of label-indicative cues, including target-relation words and close variants. If a recast sample newly introduces such cues relative to the source sample, a leakage violation is triggered. This discourages the recasting model from making the target relation obvious through lexical shortcuts.

We also penalize recast arguments that begin with newly introduced forbidden explicit discourse

Model	PDTB2			PDTB3	
	Contingency	Temporal	Comparison	Temporal	Comparison
BERT-base	261	170	676	632	795
RoBERTa-base	238	271	823	1119	854
RoBERTa-large	394	938	395	689	1778
LLaMA3.1-8B-Instruct	319	876	1027	215	1245
LLaMA3-8B-Instruct	308	913	1001	280	879
Qwen3-8B	275	751	822	475	1000

Table 13: Final retained synthetic data used for training under diverse backbone-matched filter models.

connectives. This provides another defense against shortcuts that expose the target relation through explicit markers. The connective check focuses on the argument prefix, because explicit connectives at the beginning of an argument are especially likely to reveal the discourse relation.

Penalty terms. The penalty term suppresses three types of observable violations: explicit connectives, target-relation label leakage, and output-label mismatch. We denote their corresponding binary indicators by v_{conn}, v_{leak}, and v_{label}:

$$r_{\text{pen}} = -\gamma_1 v_{\text{conn}} - \gamma_2 v_{\text{leak}} - \gamma_3 v_{\text{label}}. \quad (8)$$

where v_{conn} corresponds to a forbidden explicit connective in the argument prefix, v_{leak} corresponds to newly introduced label-indicative cues, and v_{label} corresponds to a generated label that does not match the target relation sense.

Reward combination. In implementation, the final reward combines the weighted format, similarity, and confidence rewards with the penalty term. We set $\lambda_{\text{fmt}} = 0.5$, $\lambda_{\text{sim}} = 1.5$, and $\lambda_{\text{conf}} = 3.0$.

Why the reward components are combined. The reward components address different failure cases in relation-variant discourse recasting. A reliable recast sample should have a valid output format, preserve the source topic, express the target relation sense, and avoid shortcut behaviors such as explicit connectives, label-indicative cues, label mismatch, and rigid templates. Therefore, the format reward discourages malformed outputs, the similarity reward keeps the recast sample close to the source content without copying it, the relative confidence reward encourages the target relation, and the penalty terms suppress these shortcut behaviors. We combine these rewards because removing any part may lead to invalid, off-topic, weakly transformed, or shortcut-driven recasts.

D Additional Analysis

D.1 Synthetic Data Statistics

The number of retained recast samples varies across backbone-matched filter models. Table 13 reports the final retained synthetic data used for IDRR classifier training. These results show that the data budget is determined not only by generation volume, but also by filtering strictness and model-specific confidence computation.

We further report the quantity of synthetic data at different stages of the pipeline. Table 14 summarizes the number of generated candidates before and after the first-stage rule-based filtering.

Dataset	Relation	Generated	After Stage 1
PDTB2	Contingency	9,351	6,940
	Temporal	11,967	8,176
	Comparison	10,738	7,083
PDTB3	Temporal	15,667	9,933
	Comparison	15,255	11,322

Table 14: Synthetic data quantity before and after rule-based filtering for each target relation.

D.2 Case Study

We further examine the role of reward design through representative qualitative examples. The reward components in RVDR serve as complementary constraints for relation-variant discourse recasting. Without the similarity reward, the recasting model may drift away from the source topic and produce samples detached from the original content. Without the relative confidence reward, the recast samples may fail to clearly express the target relation sense. Without the penalty terms, the recasting model may exploit explicit connectives or label-indicative cues, violating the implicitness requirement of IDRR.

Example

Source Argument Pair: *[its amended pension program will nearly double to 34,000 the number of managers eligible to retire with immediate pension payments.]*$_{Arg1}$ *[about one-third of its eligible managers to retire under the new program.]*$_{Arg2}$
Recasting goal: *Expansion → Comparison*
Full RVDR: [*its amended pension program will nearly double the number of managers eligible to retire with immediate pension payments.*]$_{Arg1}$ [*the current program applies to only about one-third of eligible managers.*]$_{Arg2}$
w/o similarity reward: [*one side shows improvement and the other side shows limitation.*]$_{Arg1}$ [*this difference forms a contrast between the two situations.*]$_{Arg2}$
w/o relative confidence reward: [*the amended pension program changes the retirement situation for managers.*]$_{Arg1}$ [*the earlier situation was different in some respects.*]$_{Arg2}$
w/o penalty terms: [*the amended pension program nearly doubles the number of eligible managers.*]$_{Arg1}$ [*in comparison, the current program covers only about one-third of eligible managers.*]$_{Arg2}$

The example above shows an *Expansion → Comparison* recasting case. It illustrates the main failure modes when different reward components are removed in RVDR. With all reward components included, the recast pair preserves the source topic and expresses the target comparison relation without explicit connectives. In contrast, removing the similarity reward produces generic contrastive text detached from the pension-program event. Removing the relative confidence reward yields a vague recast pair that only weakly reflects the target relation sense. Finally, removing the penalty terms introduces an explicit marker such as *in comparison*. These observations show that the reward components work together to support controlled relation-variant discourse recasting.

D.3 Reward-weight sensitivity.

We further examine the sensitivity to the confidence-reward weight λ_{conf}, while keeping $\lambda_{\text{fmt}} = 0.5$ and $\lambda_{\text{sim}} = 1.5$ fixed. As shown in Table 15, reducing λ_{conf} from 3.0 to 2.0 decreases downstream macro-F_1 by only 0.36 points, indicating limited sensitivity around the default setting. When increasing λ_{conf} to 5.0, we observed severe output templating and repetitive constructions during generation, and therefore did not proceed with downstream evaluation. This suggests that an excessively large confidence-reward weight may overemphasize relation confidence at the expense of generation diversity.

Setting	λ_{fmt}	λ_{sim}	λ_{conf}	Macro-F_1
Low conf. weight	0.5	1.5	2.0	69.89
Default	0.5	1.5	3.0	**70.25**
High conf. weight	0.5	1.5	5.0	–

Table 15: Sensitivity analysis of the confidence-reward weight on PDTB 3.0 with RoBERTa-base.