

Multi-Level Implicit Discourse Relation Classification via Sequence-Level Reinforcement Learning

Wei Liu¹, Xiaoyu Shen², Xiyan Fu^{3*}

¹Heidelberg Institute for Theoretical Studies

²Institute of Digital Twin, Eastern Institute of Technology, Ningbo

³Nanyang Technological University

wei.liu.llm@gmail.com | xyshen@eitech.edu.cn | xiyan.fu@ntu.edu.sg

Abstract

Implicit Discourse Relation Classification (IDRC) identifies discourse relations between text spans when connectives are absent. The task is typically evaluated on the Penn Discourse Treebank (PDTB), where discourse senses are organized in a hierarchical structure. Recent unified models predict discourse labels level by level under teacher forcing, which leads to exposure bias and cascading errors across the hierarchy. Moreover, token-level training objectives fail to assess the correctness of the entire label sequence. We reformulate multi-level IDRC as hierarchical label sequence generation and optimize the full label sequence using sequence-level reinforcement learning. By sampling and ranking complete label sequences, the proposed framework directly optimizes global sequence correctness rather than isolated token-level decisions. Experiments on PDTB 2.0 and 3.0 demonstrate consistent improvements over strong baselines. Further analysis shows that fine-to-coarse decoding outperforms the conventional coarse-to-fine strategy, highlighting the importance of decoding order in hierarchical discourse modeling.

1 Introduction

Discourse relations, such as *Contrast* and *Instantiation*, characterize the logical relationship between two text spans (Pitler et al., 2009). They can either be signaled explicitly with connectives, as shown in (1), or expressed implicitly, as in (2):

- (1) [It's raining outside.]_{Arg1} **but** [I forgot to bring an umbrella.]_{Arg2}
—— Comparison. Contrast
- (2) [Tom was selected to handle the issue.]_{Arg1} (**Implicit=Yet**)
[Many people doubted whether he had the ability to do so.]_{Arg2}
—— Comparison. Contrast

In the Penn Discourse Treebank, discourse relations are organized in a hierarchical structure, where

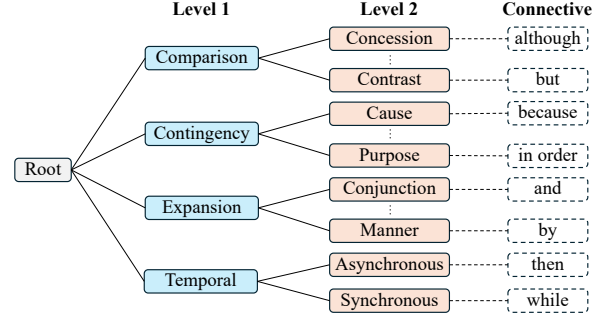


Figure 1: The PDTB 3.0 Sense Hierarchy. For clarity, we show only the top two levels of discourse senses, together with a representative discourse connective for each Level 2 sense.

coarse senses (e.g., *Comparison*) are at higher levels and fine-grained senses (e.g., *Contrast*) are at lower levels. Figure 1 illustrates the sense hierarchy in PDTB 3.0, where commonly associated discourse connectives are shown as an auxiliary layer to facilitate semantic interpretation of relations (Chan et al., 2023).

Implicit Discourse Relation Classification (IDRC) identifies the discourse relation between two arguments when discourse connectives are absent. Most previous work on this task focuses on enhancing the representations of arguments and training separate classifiers for each level (Braud and Denis, 2015; Qin et al., 2017; Kishimoto et al., 2020; Liu and Strube, 2023, 2025a), while ignoring the hierarchical structure of discourse relations. More recent studies attempt to incorporate hierarchical dependencies by jointly modeling multiple levels of discourse relations, a setting often referred to as **multi-level IDRC** (Wu et al., 2022; Chan et al., 2023). Despite these advances, existing approaches typically rely on token-level training objectives with teacher forcing (Lamb et al., 2016), where predictions at each level are conditioned on gold labels during training. This design introduces two key limitations. First, models are

*Corresponding author.

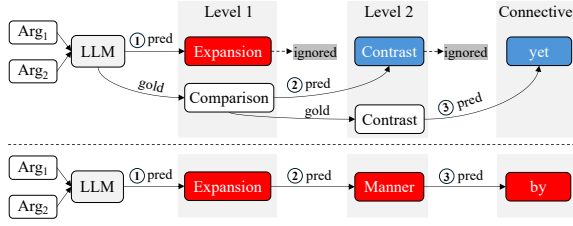


Figure 2: Illustration of exposure bias. **Training (top):** under teacher forcing, the model is conditioned on gold labels, so errors in its intermediate predictions (e.g., *Expansion* instead of *Comparison*) do not affect subsequent decisions. **Eval (bottom):** without gold guidance, an error at Level-1 propagates through the hierarchy, causing all subsequent predictions to be incorrect.

never exposed to their own intermediate predictions during training, making them fragile when errors occur during inference (i.e., exposure bias; see Figure 2). In multi-level prediction, an incorrect decision at a higher level can cascade through the hierarchy and lead to incorrect predictions at subsequent levels. Second, token-level supervision optimizes each level independently and does not evaluate the correctness of the entire relation label sequence. For example, given a gold sequence *Comparison* $\rightarrow$ *Contrast* $\rightarrow$ *yet*, token-level supervision cannot distinguish that *Comparison* $\rightarrow$ *Contrast* $\rightarrow$ *but* is preferable to *Expansion* $\rightarrow$ *Contrast* $\rightarrow$ *yet*, even though the former preserves the correct coarse-level relation while the latter fails at Level 1.

To address these challenges, we reformulate multi-level IDRC as hierarchical sequence generation over discourse labels. Given a pair of arguments, instead of predicting each level independently under teacher forcing, our model generates the complete relation label sequence corresponding to the hierarchy path. This formulation enables sequence-level optimization via Reinforcement Learning (RL), allowing the model to directly optimize the correctness of the entire label sequence rather than isolated token-level decisions. Specifically, we leverage Group Relative Policy Optimization (GRPO, Shao et al., 2024) to optimize the label generation process with hierarchy-informed dense rewards that provide graded supervision signals. By sampling and ranking complete label sequences, our framework mitigates exposure bias and encourages globally correct predictions. Furthermore, we investigate the impact of decoding order. While conventional approaches follow a coarse-to-fine order (Wu et al., 2022), our experi-

ments reveal that a fine-to-coarse strategy (*Conn* $\rightarrow$ *L2 Rel* $\rightarrow$ *L1 Rel*) consistently leads to better performance, highlighting the importance of decision ordering in hierarchical discourse modeling.

In summary, our contributions are as follows:

- We reformulate multi-level IDRC as hierarchical label sequence generation and optimize it with sequence-level reinforcement learning to mitigate exposure bias and cascading errors.
- We design hierarchy-informed dense rewards that provide fine-grained supervision signals for effective sequence-level optimization.
- We empirically¹ demonstrate consistent improvements on PDTB 2.0 and 3.0 and show that predicting fine-grained relations before coarse ones leads to better performance.

2 Related Work

Implicit discourse relation classification (IDRC), a key component of shallow discourse parsing, has been extensively studied since the release of the Penn Discourse Treebank (PDTB) 2.0 (Prasad et al., 2008). Early approaches relied on linguistically motivated feature engineering (Lin et al., 2009; Rutherford and Xue, 2014), while later work adopted neural architectures (Qin et al., 2017; Kishimoto et al., 2018), pretrained language models (Kishimoto et al., 2020; Long and Webber, 2022; Cai et al., 2024), and large language models (Chan et al., 2024; Wang et al., 2025). Despite strong performance, most methods treat discourse sense prediction at each level as an independent classification task and do not explicitly model the hierarchical sense structure defined in PDTB.

Recent work has incorporated the PDTB hierarchy more explicitly. Costa and Kosseim (2025) leverage it for multi-task and multi-label discourse relation classification, while other studies use implicit connectives as auxiliary signals that provide lexical cues to the underlying discourse relation (Shi and Demberg, 2019; Liu and Strube, 2023; Liu et al., 2024; Fan et al., 2026). Another line of work jointly predicts discourse relations across hierarchical levels with a connective, capturing dependencies within the PDTB hierarchy (Wu et al., 2022; Chan et al., 2023; Zeng et al., 2024). However, these joint frameworks typically rely on teacher forcing, where predictions condition on

¹https://github.com/liuwei1206/IDRC_RL

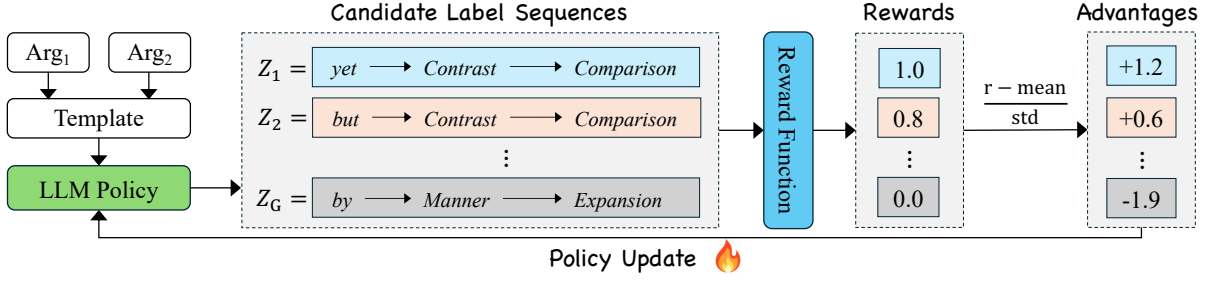


Figure 3: Overview of the proposed sequence-level reinforcement learning framework for multi-level IDRC. The LLM policy takes the input arguments formatted using a prompt template and generates G candidate label sequences $\{Z_1, Z_2, \dots, Z_G\}$ in a fine-to-coarse order. Each candidate is evaluated at the sequence level by comparing the predicted label sequence with the ground truth (e.g., **yet** $\rightarrow$ **Contrast** $\rightarrow$ **Comparison**). Relative advantages are computed within the candidate group and used to update the policy, encouraging globally better label sequences.

gold labels from other levels, leading to exposure bias (Lamb et al., 2016). Moreover, token-level supervision optimizes predictions independently, providing limited guidance for distinguishing globally correct label sequences from inferior ones. In contrast, our work formulates multi-level discourse relation prediction as a hierarchical sequential decision process and directly optimizes the entire label sequence.

Our work is also related to sequence-level optimization, which has been widely explored in NLP tasks requiring evaluation of global output quality. For example, prior studies leverage sequence-level feedback to improve factual consistency in summarization (Roit et al., 2023) and align machine translation outputs with human preferences (Ramos et al., 2024). While these approaches apply sequence-level objectives in different settings, we introduce sequence-level optimization to multi-level IDRC, enabling the model to mitigate exposure bias and better capture hierarchical dependencies in the PDTB sense structure.

3 Method

Figure 3 illustrates our sequence-level reinforcement learning framework for multi-level IDRC. We first initialize the policy via supervised fine-tuning (SFT), and then perform sequence-level optimization using Group Relative Policy Optimization (GRPO). We also investigate alternative decoding orders for label sequence generation.

Formally, given arguments Arg_1 and Arg_2 , we construct the model input X using a prompt template (see Appendix A). Let y_1^* and y_2^* denote the Level-1 and Level-2 relations, and let c^* denote the annotated implicit connective used as an auxiliary training signal. Each training instance consists of

(X, Y^*) , where $Y^* = (y_1^*, y_2^*, c^*)$ denotes the gold discourse labels. At inference time, only the arguments are provided, and the model predicts the label sequence without access to gold connectives or discourse relations.

3.1 Problem Reformulation

We reformulate multi-level IDRC as hierarchical label sequence generation and optimize it with sequence-level reinforcement learning.

Given input X , the model autoregressively generates a token sequence $Z = (z_1, \dots, z_{|Z|})$. We treat the language model as a stochastic policy π_θ . At each step t , the model observes the state $s_t = (X, z_{<t})$ and selects an action z_t from the vocabulary according to $\pi_\theta(z_t | X, z_{<t})$. The sequence probability factorizes as:

$$\pi_\theta(Z | X) = \prod_{t=1}^{|Z|} \pi_\theta(z_t | X, z_{<t}) \quad (1)$$

The generated sequence is deterministically parsed (using a parser function f) into a label tuple:

$$Y = f(Z) = (y_1, y_2, c) \quad (2)$$

where y_1 , y_2 , and c correspond to the predicted Level-1 relation, Level-2 relation, and implicit connective extracted according to the predefined output format. After generating a complete sequence, a sequence-level reward $R(f(Z), Y^*)$ is computed and used to update the policy.

3.2 Stage 1: Supervised Policy Initialization

We first perform supervised fine-tuning to produce valid label sequences. The model is trained under teacher forcing to maximize the likelihood of the gold output sequence Z^* , which is constructed

from the gold discourse labels Y^* according to the predefined output template (see Appendix A):

$$\mathcal{L}_{\text{SFT}}(\theta) = - \sum_{t=1}^{|Z^*|} \log \pi_{\theta}(z_t^* | X, z_{<t}^*) \quad (3)$$

This stage provides a stable initialization (Ouyang et al., 2022) but optimizes token-level likelihood rather than sequence-level correctness.

3.3 Stage 2: Sequence-Level Optimization

After supervised initialization, we further optimize the policy using sequence-level reinforcement learning, maximizing the expected reward over generated label sequences for each input X :

$$\mathcal{J}(\theta) = \mathbb{E}_{Z \sim \pi_{\theta}(\cdot | X)} [R(f(Z), Y^*)] \quad (4)$$

where Y^* denotes the gold discourse labels and $R(\cdot)$ evaluates the parsed label sequence. In this work, we instantiate $R(\cdot)$ using either a sparse reward R_s or a hierarchy-informed dense reward R_d , described in Section 3.3.2.

3.3.1 Sequence Sampling

For each instance, we sample a group of G label sequences from the policy without teacher forcing:

$$\mathcal{G} = \{Z_1, Z_2, \dots, Z_G\} \quad (5)$$

where each sequence is generated autoregressively according to the LLM policy, i.e., $Z_i \sim \pi_{\theta}(\cdot | X)$. Rewards are assigned only after full sequence generation, ensuring that early decisions influence the final reward and reflecting inference-time behavior.

3.3.2 Reward Functions

Sequence-level reinforcement learning requires a reward over complete label sequences. A natural choice is an exact-match reward (sparse), which gives positive feedback only to fully correct sequences. As a result, most sampled sequences receive zero reward, providing little guidance for improving intermediate decisions. We therefore design a hierarchy-informed dense reward that decomposes supervision across label-sequence components and rewards partially correct predictions.

Sparse Reward. A sparse reward provides supervision only when both discourse relation levels are correctly predicted:

$$R_s(f(Z), Y^*) = \mathbb{I}[y_1 = y_1^*] \cdot \mathbb{I}[y_2 = y_2^*] \quad (6)$$

where $\mathbb{I}[\cdot]$ is the indicator function. This directly optimizes sequence-level relation exact match but

does not distinguish partially correct predictions from completely incompatible ones.

Hierarchy-Informed Dense Reward. To provide richer supervision, we introduce a dense reward that provides supervision for different components of the predicted label sequence:

$$\begin{aligned} R_d(f(Z), Y^*) = & \lambda_1 \mathbb{I}[y_1 = y_1^*] \\ & + \lambda_2 \mathbb{I}[\mathcal{H}(y_2) = \mathcal{H}(y_2^*)] \\ & + \lambda_3 \mathbb{I}[y_2 = y_2^*] \\ & + \lambda_4 \mathbb{I}[c \sim y_2^*] \end{aligned} \quad (7)$$

Here, $\mathcal{H}$ maps a Level-2 relation to its Level-1 parent in the PDTB hierarchy. The first and third terms reward exact correctness at Level 1 and Level 2, while the second term assigns partial reward when the predicted Level-2 relation belongs to the same coarse-level category as the gold label. The indicator $c \sim y_2^*$ denotes semantic compatibility between the predicted connective and the gold Level-2 relation, i.e., whether the connective is annotated in PDTB as a valid signal for that relation (e.g., *yet* for *Contrast*). This term rewards connectives that are semantically consistent with the intended relation, even when the full label sequence is not exactly matched. As a result, semantically plausible predictions (e.g., predicting *but* instead of *yet* for a *Contrast* relation) are distinguished from entirely incompatible ones.

Overall, the dense reward provides denser and more informative supervision signals than exact-match rewards, enabling the model to learn from partially correct label sequences and improving optimization stability. In our experiments, we set $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.3, 0.15, 0.5, 0.05)$, assigning larger weights to exact Level-2 and Level-1 correctness while using smaller weights for coarse-level consistency and connective compatibility.²

3.3.3 Group Relative Policy Optimization.

To update the policy, we adopt Group Relative Policy Optimization (GRPO), which estimates advantages by comparing multiple sequences sampled from the same input. This group-relative comparison reduces dependence on absolute reward scales and provides stable learning signals across candidate label sequences, making it appropriate for sequence-level optimization (Guo et al., 2025).

For each sampled group of sequences $\{Z_i\}_{i=1}^G$, we compute their rewards $\{R_i\}$ using one of the

²We conduct a sensitivity analysis over alternative reward-weight configurations in Appendix E.

sequence-level reward functions described above. We then compute the group-relative advantages:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_i\})}{\text{std}(\{R_i\}) + \delta}. \quad (8)$$

where δ is a small constant for numerical stability. The policy is updated using a clipped surrogate objective with KL regularization:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G} \sum_{i=1}^G \frac{1}{|Z_i|} \sum_{t=1}^{|Z_i|} \text{CLIP}(r_{i,t}, \hat{A}_i, \epsilon) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \quad (9)$$

where the clipped surrogate function is defined as

$$\text{CLIP}(r, \hat{A}, \epsilon) = \min(r\hat{A}, \text{clip}(r, 1 - \epsilon, 1 + \epsilon)\hat{A}) \quad (10)$$

Here, $r_{i,t}$ denotes the importance sampling ratio between the current policy and the sampling policy. The clipping operator, with threshold ϵ , limits overly large policy updates following PPO-style stabilization (Schulman et al., 2017). The KL term regularizes the policy toward the SFT initialization π_{ref} , with β controlling the regularization strength.

3.4 Decoding Order

In sequential multi-level IDRC, discourse labels are generated step by step, so the decoding order acts as an inductive bias that affects both prediction and sequence-level RL optimization. Our framework supports flexible decoding orders, enabling us to explore different prediction strategies. We consider two alternatives:

- **Coarse-to-fine (C2F):** Following the hierarchical structure of PDTB and common practice in prior IDRC work (Wu et al., 2022), the model first predicts the Level-1 relation, then the Level-2 relation, and finally the implicit connective, i.e., $L1 \text{ Rel} \rightarrow L2 \text{ Rel} \rightarrow \text{Conn}$.
- **Fine-to-coarse (F2C):** The model begins with the most fine-grained signal, the implicit connective, followed by the Level-2 and Level-1 relations, i.e., $\text{Conn} \rightarrow L2 \text{ Rel} \rightarrow L1 \text{ Rel}$.

Since each prediction conditions on earlier decisions, the decoding order may affect information flow and error propagation. Intuitively, F2C grounds relation prediction on connective-based lexical cues, which can provide strong semantic signals for higher-level discourse relations. We compare both strategies empirically in Section 5.2.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our model on PDTB 2.0 (Prasad et al., 2008) and 3.0 (Webber et al., 2019). Both datasets annotate discourse relations using a three-level sense hierarchy. We follow previous works (Ji and Eisenstein, 2015; Kim et al., 2020) to use top-level 4-way and second-level 11-way classification for PDTB 2.0, and top-level 4-way and second-level 14-way classification for PDTB 3.0. We adopt the standard data split from Ji and Eisenstein (2015), using sections 2–20 for training, 0–1 for development, and 21–22 for testing.³

Implementation Details. We implement our method using verl⁴, an open-source reinforcement learning framework. All experiments are conducted on Llama-3.1-8B-Instruct.⁵ During policy initialization, we perform supervised fine-tuning (SFT) with LoRA for three epochs, using a learning rate of 3×10^{-5} , a LoRA rank of 12, and a LoRA scaling factor (alpha) of 24. For sequence-level optimization with GRPO, we set the KL regularization coefficient β to 0.01. We use a rollout batch size of 128 and an update batch size of 64. During rollout generation, we sample 16 trajectories for each input with a sampling temperature of 1.0. To account for training variability on PDTB, we report the mean performance over 5 random restarts. Performance is evaluated using standard accuracy (Acc, %) and macro F1 (F1, %).

Baselines. To demonstrate the effectiveness of our model, we compare our method with two LLM-based baselines that reflect different modeling assumptions for multi-level IDRC:

- **Level-wise SFT:** Fine-tune LLMs to predict each discourse relation level independently. This baseline does not model hierarchical dependencies across levels.
- **Joint SFT:** Fine-tune LLMs to jointly generate all levels as a single sequence under teacher forcing. This setting captures dependencies, but relies on token-level supervision with gold previous labels during training.
- **Joint SFT + SS:** We replace teacher forcing in Joint SFT with Scheduled Sampling (Mihaylova and Martins, 2019), partially expos-

³We also show results on the Lin split in Appendix C.

⁴<https://github.com/verl-project/verl>

⁵We also conduct experiments on Qwen-2.5-7B-Instruct and observe similar performance; see Appendix G for details.

		PDTB 2.0				PDTB 3.0			
		L1		L2		L1		L2	
Models		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Wu et al. (2022)		71.18	63.73	60.33	40.49	-	-	-	-
Liu and Strube (2023)		74.59	68.64	62.75	42.36	76.23	71.15	65.51	54.92
Chan et al. (2023)		73.80	67.79	61.41	44.04	-	-	-	-
Zeng et al. (2024)		73.89	67.83	61.33	46.14	75.53	71.59	64.87	56.50
Wang et al. (2025)		78.20	71.14	62.46	46.38	76.93	73.33	64.35	55.04
Fan et al. (2026)		77.15	73.89	66.61	53.17	78.61	73.88	68.86	62.29
F2C	Level-wise SFT	68.73 _{2.1}	61.32 _{1.6}	59.75 _{1.5}	45.90 _{2.3}	73.21 _{1.9}	68.11 _{2.0}	63.48 _{1.8}	53.58 _{1.7}
	Joint SFT	72.45 _{1.4}	68.54 _{1.2}	61.08 _{1.1}	48.66 _{0.6}	76.38 _{0.9}	72.63 _{1.1}	65.96 _{1.4}	56.75 _{1.5}
	Joint SFT + SS	72.86 _{1.6}	68.87 _{1.7}	60.92 _{1.4}	48.53 _{1.1}	75.86 _{1.2}	72.35 _{1.5}	65.54 _{1.0}	56.33 _{0.9}
	Our Model - R_s	77.64 _{1.3}	74.51 _{0.7}	67.48 _{1.5}	51.52 _{0.6}	79.03 _{1.3}	75.29 _{0.7}	70.14 _{1.4}	62.64 _{1.1}
	Our Model - R_d	78.37 _{0.9}	75.04 _{0.7}	68.39 _{1.1}	52.73 _{0.7}	79.99 _{0.6}	76.54 _{0.9}	70.64 _{0.8}	63.16 _{1.0}
C2F	Joint SFT	70.72 _{1.7}	64.87 _{1.6}	59.60 _{1.8}	46.08 _{2.0}	73.46 _{1.7}	69.15 _{1.7}	62.79 _{1.5}	52.68 _{1.3}
	Joint SFT + SS	70.48 _{1.6}	65.22 _{1.7}	58.84 _{2.0}	45.78 _{2.4}	73.88 _{1.9}	69.52 _{1.8}	63.08 _{1.8}	53.42 _{1.9}
	Our Model - R_s	67.42 _{1.9}	66.55 _{1.6}	56.81 _{1.7}	43.72 _{2.1}	71.24 _{2.1}	68.72 _{1.2}	60.12 _{2.0}	51.94 _{1.7}
	Our Model - R_d	75.45 _{1.3}	72.44 _{1.1}	66.82 _{1.6}	49.60 _{1.0}	77.49 _{1.3}	73.59 _{1.6}	68.18 _{1.3}	59.13 _{1.2}

Table 1: Mean accuracy and F1 scores (with standard deviations) on PDTB 2.0 and PDTB 3.0.

ing the model to its own predictions during training to mitigate exposure bias.

Furthermore, we report results from representative prior approaches for reference. See Appendix B for more details about the training setup and baselines.

4.2 Overall Results

Table 1 shows results on PDTB 2.0 and 3.0. The **F2C** block reports results under the Fine-to-Coarse decoding order (i.e., $Conn \rightarrow L2\ Rel \rightarrow L1\ Rel$), while the **C2F** block corresponds to the Coarse-to-Fine strategy (i.e., $L1\ Rel \rightarrow L2\ Rel \rightarrow Conn$).

Under the **F2C** strategy, Joint SFT consistently outperforms Level-wise SFT on both datasets, with gains of approximately 3–4 percentage points in accuracy and F1-score. This confirms that jointly modeling hierarchical discourse labels is beneficial for multi-level IDRC, consistent with prior studies (Wu et al., 2022; Chan et al., 2023). Replacing teacher forcing with scheduled sampling (Joint SFT + SS), which partially exposes the model to its own predictions during training, yields only marginal improvements. By contrast, incorporating sequence-level optimization yields substantial improvements over the Joint SFT baseline, regardless of whether the sparse reward R_s or the dense reward R_d is used. For example, with the dense reward, our model achieves an accuracy of 68.39% on PDTB 2.0 L2 and 70.64% on PDTB 3.0 L2, outperforming Joint SFT by approximately 5 percentage

points. It also surpasses previously reported representative models, including hierarchy-aware token-level methods (Wu et al., 2022; Chan et al., 2023; Zeng et al., 2024), auxiliary connective generation methods (Liu and Strube, 2023), subtext-enhanced generative IDRC methods (Wang et al., 2025), linguistic cues-augmented models (Fan et al., 2026). These results indicate that optimizing relation prediction at the sequence level is more effective than token-level supervision for modeling hierarchical discourse relations. Moreover, the dense reward consistently outperforms⁶ the sparse variant (e.g., 68.39% vs. 67.48% accuracy on PDTB 2.0 L2), showing that richer reward signals further improve sequence-level optimization.

In contrast, under the **C2F** strategy, Joint SFT only outperforms Level-wise SFT at the L1 level, while achieving comparable or slightly worse performance for L2 relation classification. The scheduled sampling variant (Joint SFT + SS) shows similar results to Joint SFT, indicating that this strategy brings limited improvement in the coarse-to-fine setting. When applying sequence-level optimization with sparse rewards, performance does not improve and can even fall below the SFT baseline. This likely occurs because the sparse reward provides only binary feedback over the entire sequence, offering limited guidance for improving

⁶The improvement is statistically significant over five seed-matched runs ($p < 0.01$, two-sided paired t -test).

Models	L1		L2	
	Acc	F1	Acc	F1
Joint SFT	72.45	68.54	61.08	48.66
RLOO- R_s	74.86	71.87	65.20	50.15
RLOO- R_d	77.06	74.12	66.92	51.28
GRPO- R_s	77.64	74.51	67.48	51.52
GRPO- R_d	78.37	75.04	68.39	52.73

Table 2: RLOO ablation on PDTB 2.0 using Llama-3.1-8B-Instruct under the F2C decoding order.

intermediate predictions. In contrast, using the hierarchy-informed dense reward leads to substantial improvements over both Level-wise SFT and Joint SFT. Specifically, our model achieves an accuracy of 66.82% on PDTB 2.0 L2 and 68.18% on PDTB 3.0 L2, exceeding Joint SFT by approximately 7 and 5 percentage points, respectively. These results highlight the importance of informative reward signals when applying sequence-level optimization under the C2F decoding order.

5 Analysis

5.1 Sequence-level Optimization

To better understand sequence-level optimization, we use an RLOO ablation to separate sequence-level reward optimization from GRPO-specific components and conduct two complementary analyses focusing on robustness to intermediate errors and global sequence ranking.

RLOO Ablation. We add an RLOO baseline on PDTB 2.0 under F2C, using the same policy initialization and sequence-level rewards as our GRPO models. RLOO is a REINFORCE-style policy-gradient method that samples complete label sequences, computes sequence-level rewards, and uses a leave-one-out reward baseline to reduce variance (Ahmadian et al., 2024). Unlike GRPO, it does not use normalized group-relative advantages or a clipped surrogate objective. As shown in Table 2, both RLOO variants improve over Joint SFT: RLOO- R_s increases PDTB 2.0 L2 accuracy from 61.08 to 65.20, while RLOO- R_d improves it to 66.92. These results indicate that sequence-level reward optimization is beneficial without relying on GRPO-specific updates. However, both RLOO variants remain below their GRPO counterparts, suggesting that GRPO provides a stronger optimizer in this setting.

Robustness to Intermediate Errors. We analyze error propagation using the conditional error rate

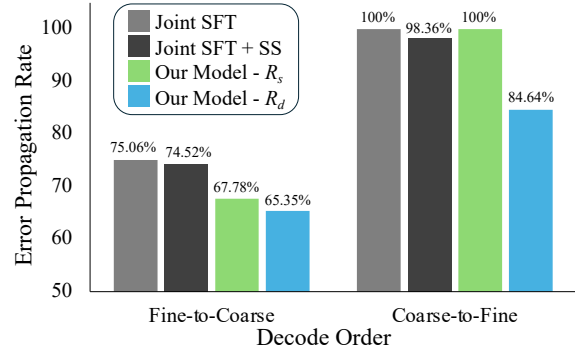


Figure 4: Error propagation rates for Joint SFT, Joint SFT + SS, and our models under different decoding orders. Under F2C, propagation is defined as $P(L1 \text{ wrong} \mid L2 \text{ wrong})$, while under C2F it is defined as $P(L2 \text{ wrong} \mid L1 \text{ wrong})$. Scheduled sampling yields only marginal reductions in propagation, whereas sequence-level optimization substantially lowers propagation rates, demonstrating improved robustness to intermediate prediction errors.

of a later prediction given an earlier error in the label sequence (Error Propagation Rate). Under the C2F setting, we measure the probability that the Level-2 prediction is incorrect given an incorrect Level-1 prediction; under F2C, we compute the probability that the Level-1 prediction is incorrect given an incorrect Level-2 prediction. Figure 4 reports results on PDTB 2.0.

Joint SFT exhibits strong error propagation. Under F2C, 75.06% of instances with incorrect Level-2 predictions also have incorrect Level-1 predictions, and scheduled sampling (Joint SFT + SS) only slightly reduces this rate to 74.52%. The effect is even stronger under C2F, where the propagation rate reaches 100% for Joint SFT and remains high with scheduled sampling at 98.36%. These results suggest that partially exposing the model to its own predictions during training is insufficient to mitigate cascading errors. In contrast, sequence-level optimization substantially reduces error propagation under F2C, lowering the rate to 67.78% with sparse rewards and 65.35% with dense rewards. Under C2F, sparse rewards fail to reduce error propagation, with the rate remaining at 100%, whereas dense rewards effectively reduce the rate to 84.64%. Overall, sequence-level optimization helps the model recover from intermediate errors, and denser signals further improve robustness.

Global Sequence Ranking. We evaluate Joint SFT and our model using $\text{pass}@k$, which measures whether the gold label sequence appears among k sequences sampled from the model, re-

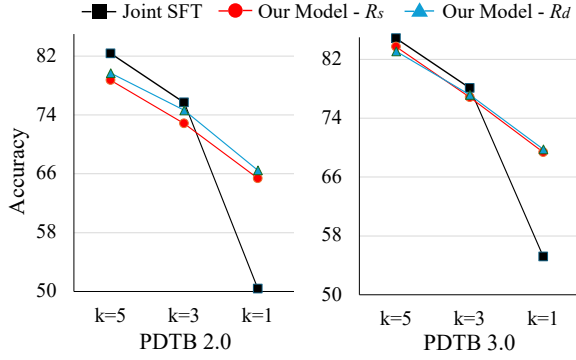


Figure 5: Pass@ k results on PDTB 2.0 and 3.0 under the F2C setting. Our models achieve substantially higher pass@1 while maintaining comparable performance at larger k , indicating improved ranking of globally correct label sequences through sequence-level optimization.

flecting their ability to rank globally correct label sequences. For each input X , we independently sample k sequences $\{\hat{Z}^{(1)}, \hat{Z}^{(2)}, \dots, \hat{Z}^{(k)}\}$ and check whether any parsed relation sequence matches the gold labels Y^* : $\text{pass}@k = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\exists j \leq k \text{ s.t. } f(\hat{Z}_i^{(j)}) = Y_i^*)$, where N is the number of evaluation instances and $\mathbf{1}(\cdot)$ is the indicator function. Figure 5 reports results on PDTB 2.0 and 3.0 under F2C.

When k is large, such as $k = 5$, Joint SFT and our model achieve comparable pass@ k , indicating that both can generate the correct label sequence among sampled outputs. However, when $k = 1$, our model with either sparse or dense rewards substantially outperforms Joint SFT, suggesting that sequence-level optimization ranks globally correct sequences higher and makes them more likely to be selected as the top prediction. By contrast, token-level supervision provides less consistent preference for globally optimal sequences. Overall, these results show that sequence-level optimization better ranks globally correct label sequences.

5.2 Decoding Order

We observe in Table 1 that the Fine-to-Coarse (F2C) decoding order consistently outperforms Coarse-to-Fine (C2F). One reason is that fine-grained relations and discourse connectives provide more specific semantic cues, whereas coarse-level relations are more abstract and ambiguous. Predicting fine-grained labels first therefore offers stronger grounding for subsequent coarse-level predictions, while coarse-level predictions provide weaker guidance for fine-grained decisions. This is consistent with the PDTB 2.0 annotation man-

ual (Prasad et al., 2008), which notes that inserting a connective facilitates relation identification, as well as connective-enhanced IDRC studies showing that connective prediction improves discourse relation classification (Shi and Demberg, 2019; Liu and Strube, 2023). F2C is also more robust to error propagation. As shown in Figure 4, incorrect Level-1 predictions in C2F almost always lead to incorrect Level-2 predictions, indicating severe cascading errors. In contrast, F2C yields substantially lower propagation rates, suggesting that Level-2 errors are less destructive to subsequent predictions. This is because fine-grained predictions, even when incorrect, often remain semantically close to the correct relation: for example, predicting “*Concession*” instead of “*Contrast*” can still lead to the correct Level-1 relation “*Comparison*”.

More generally, these results suggest that fine-to-coarse decoding may provide an effective ordering principle for tree-structured hierarchical prediction. Fine-grained, information-rich labels can ground more abstract parent categories, while errors among nearby fine-grained labels may still preserve the correct higher-level path. This perspective may extend beyond IDRC to other tasks with tree-like label hierarchies, such as hierarchical text classification and fine-grained entity or event typing, where a GRPO-style sequence-level formulation can be used to optimize the entire predicted path rather than independent local decisions.

5.3 Reward Design

To evaluate the impact of reward design, we compare sparse and hierarchy-informed dense reward functions. As shown in Table 1 and Figure 4, the dense reward consistently achieves better performance than the sparse reward, with larger gains under the C2F decoding order. To examine optimization behavior, Figure 6 presents training reward curves on PDTB 3.0 under C2F. When using sparse rewards, the average reward fails to improve beyond its initialization level and exhibits unstable optimization dynamics. In contrast, the dense reward leads to steady reward increases and higher convergence levels. This difference is likely due to the more informative learning signal provided by the dense reward during training. Sparse rewards provide only binary sequence-level feedback, making partially correct predictions indistinguishable from completely incorrect ones. The issue is especially pronounced under C2F, where early coarse-level decisions strongly constrain downstream pre-

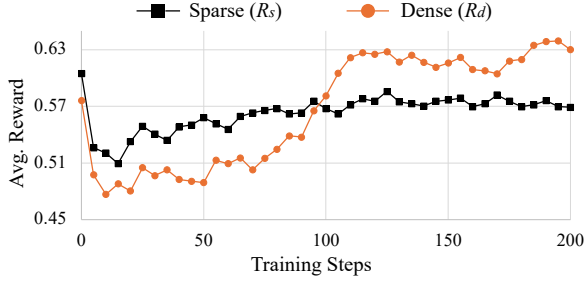


Figure 6: Training reward curves on PDTB 3.0 under the C2F decoding order. Sparse rewards fail to surpass their initialization level and exhibit unstable optimization, whereas dense rewards lead to steady reward improvements and higher convergence.

dictions. By decomposing the label sequence into hierarchical components, the dense reward provides informative intermediate feedback and enables more stable sequence-level optimization.

6 Conclusion

In this paper, we reformulate multi-level implicit discourse relation classification as hierarchical label sequence generation and optimize it using sequence-level reinforcement learning. We design both sparse and hierarchy-informed dense rewards to guide training. Experiments on PDTB 2.0 and 3.0 demonstrate that our approach consistently outperforms strong baselines. Through detailed analyses, we show that sequence-level optimization mitigates error propagation caused by token-level supervision and improves the ranking of globally correct label sequences. Furthermore, we find that hierarchy-informed dense rewards provide more informative learning signals than sparse rewards, particularly under challenging decoding orders. Finally, our results reveal that a fine-to-coarse decoding strategy is better suited for multi-level IDRC, highlighting the importance of decision ordering in hierarchical discourse modeling.

7 Limitations

Despite the strong empirical results, our study has several limitations. First, handling ambiguous instances remains an open issue. In PDTB 2.0, 3.45% of implicit instances are annotated with multiple labels, and the proportion rises to 5.0% in PDTB 3.0. Following prior work (Shi and Demberg, 2019; Liu and Strube, 2023, 2025b), we use only the first label during training, which simplifies supervision but may not fully capture annotation ambiguity. Exploring more principled treatments of multi-label

cases, such as ambiguity-aware objectives or distributional supervision, remains an important direction for future work. Second, our approach requires more training computation than Joint-SFT. In particular, sequence-level reinforcement learning samples multiple candidate label sequences for each instance. However, this additional cost only affects training and does not increase inference cost. Moreover, sequence-level optimization leads to consistent performance gains on both PDTB 2.0 and PDTB 3.0. Exploring more efficient sequence-level optimization strategies remains an important direction for future work. Third, following prior multi-level IDRC work (Wu et al., 2022; Chan et al., 2023; Zeng et al., 2024), we restrict our study to L1 and L2 discourse relations and focus on English. Nevertheless, our sequence-generation formulation can naturally be extended to L3 relations by generating a longer label sequence, and it would also be valuable to explore its generalizability (Fu and Frank, 2023) in multilingual settings (Liu et al., 2023). We leave these extensions to future work.

References

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, Bangkok, Thailand. Association for Computational Linguistics.
- Chloé Braud and Pascal Denis. 2015. [Comparing word representations for implicit discourse relation classification](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2201–2211, Lisbon, Portugal. Association for Computational Linguistics.
- 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, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. [Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian’s, Malta. Association for Computational Linguistics.

- 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.
- Nelson Filipe Costa and Leila Kosseim. 2025. [A multi-task and multi-label classification model for implicit discourse relation recognition](#). In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 231–245, Avignon, France. Association for Computational Linguistics.
- Yi Fan, Michael Strube, and Wei Liu. 2026. [Linguistic cues for LLM-based implicit discourse relation classification](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4585–4602, Rabat, Morocco. Association for Computational Linguistics.
- Xiyan Fu and Anette Frank. 2023. [SETI: Systematicity evaluation of textual inference](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4101–4114, Toronto, Canada. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- 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.
- Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. [Implicit discourse relation classification: We need to talk about evaluation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.
- Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. [A knowledge-augmented neural network model for implicit discourse relation classification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595, Santa Fe, New Mexico, USA. 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.
- Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. [Professor forcing: A new algorithm for training recurrent networks](#). *Advances in neural information processing systems*, 29.
- 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.
- Wei Liu, Yi Fan, and Michael Strube. 2023. [HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.
- Wei Liu and Michael Strube. 2023. [Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.
- Wei Liu and Michael Strube. 2025a. [Discourse relation-enhanced neural coherence modeling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4748–4762, Vienna, Austria. Association for Computational Linguistics.
- Wei Liu and Michael Strube. 2025b. [Joint modeling of entities and discourse relations for coherence assessment](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21910–21926, Suzhou, China. Association for Computational Linguistics.
- Wei Liu, Stephen Wan, and Michael Strube. 2024. [What causes the failure of explicit to implicit discourse relation recognition?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.
- Wanqiu Long and Bonnie Webber. 2022. [Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tsvetomila Mihaylova and André F. T. Martins. 2019. [Scheduled sampling for transformers](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 351–356, Florence, Italy. Association for Computational Linguistics.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Emily Pitler, Annie Louis, and Ani Nenkova. 2009. [Automatic sense prediction for implicit discourse relations in text](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, 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).
- Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. [Adversarial connective-exploiting networks for implicit discourse relation classification](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.
- Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. [Aligning neural machine translation models: Human feedback in training and inference](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).
- Paul Roit, Johan Ferret, Lior Shani, Roei Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. [Factually consistent summarization via reinforcement learning with textual entailment feedback](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.
- Attapol Rutherford and Nianwen Xue. 2014. [Discovering implicit discourse relations through brown cluster pair representation and coreference patterns](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden. Association for Computational Linguistics.
- 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, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#).
- Wei Shi and Vera Demberg. 2019. [Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification](#). In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.
- 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](#). *Philadelphia, University of Pennsylvania*, 35:108.
- Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. [A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11486–11494.
- Lei Zeng, Ruifang He, Haowen Sun, Jing Xu, Chang Liu, and Bo Wang. 2024. [Global and local hierarchical prompt tuning framework for multi-level implicit discourse relation recognition](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7760–7773, Torino, Italia. ELRA and ICCL.

Input Template in the C2F Setting
You are an expert in {{ task }} implicit discourse relation classification. Your task is to analyze the relationship between two arguments based on the Penn Discourse Treebank {{ version }} annotation manual. Given Arg1 and Arg2, predict: 1. Level-1 Sense: One of [{{ 11_labels }}]. 2. Level-2 Sense: One of [{{ 12_labels }}]. 3. Connective: A natural implicit connective that best links the two arguments. Input: "Arg1": "{{ arg1 }}" "Arg2": "{{ arg2 }}" Output Format: Level1: <Level-1 label> Level2: <Level-2 label> Connective: <connective>

Figure 7: Input template in the C2F setting, where $\text{task} \in \{\text{pdtb 2.0, pdtb 3.0}\}$ and $\text{version} \in \{2.0, 3.0\}$.

A Prompt Template

Figures 7 and 8 show the input templates for the C2F and F2C settings, respectively, while Figures 9 and 10 show the corresponding output templates.

B Additional Training Details

For the Level-wise SFT and Joint SFT baselines, we experimented with both full-parameter fine-tuning and LoRA fine-tuning, and observed comparable performance. Therefore, we adopt LoRA fine-tuning for these baselines for improved training efficiency. Specifically, we use a learning rate of 3×10^{-5} , a LoRA rank of 32, a LoRA scaling factor (alpha) of 64, and a maximum epoch number of 8 for training.

For the Joint SFT + SS baseline, we apply the same LoRA configuration. Scheduled sampling follows the linear decay strategy proposed by [Mihaylova and Martins \(2019\)](#), where the probability of sampling ground-truth tokens at training step i is defined as $t(i) = \max(\epsilon, k - ci)$. Here $0 \leq \epsilon < 1$ denotes the minimum teacher forcing probability, while k and c control the offset and slope of the decay. This function determines the teacher forcing ratio at each training step.

For our model, most hyper-parameters are de-

Input Template in the F2C Setting
You are an expert in {{ task }} implicit discourse relation classification. Your task is to analyze the relationship between two arguments based on the Penn Discourse Treebank {{ version }} annotation manual. Given Arg1 and Arg2, predict: 1. Connective: A natural implicit connective that best links the two arguments. 2. Level-2 Sense: One of [{{ 12_labels }}]. 3. Level-1 Sense: One of [{{ 11_labels }}]. Input: "Arg1": "{{ arg1 }}" "Arg2": "{{ arg2 }}" Output Format: Connective: <connective> Level2: <Level-2 label> Level1: <Level-1 label>

Figure 8: Input template in the F2C setting, where $\text{task} \in \{\text{pdtb 2.0, pdtb 3.0}\}$ and $\text{version} \in \{2.0, 3.0\}$.

Output Template in the C2F Setting
Level1: {{ 11_label }} Level2: {{ 12_label }} Connective: {{ connective }}

Figure 9: Output template in the C2F setting.

Output Template in the F2C Setting
Connective: {{ connective }} Level2: {{ 12_label }} Level1: {{ 11_label }}

Figure 10: Output template in the F2C setting.

scribed in Section 4.1. In addition, we use the Adam optimizer for all models. All experiments are conducted on 8 H20 GPUs. Training the baselines takes approximately 3–5 hours, while our models require about 15 hours on PDTB 2.0 and about 20 hours on PDTB 3.0.

C Results on the Lin Split

To show that the improvements of our methods are robust to different data splits, we additionally evaluate on PDTB 2.0 using the Lin split ([Lin et al., 2009](#)). Table 3 shows the results of the Joint

Models	L1		L2	
	Acc	F1	Acc	F1
Joint SFT	71.80	65.42	60.31	45.23
Our Model- R_s	76.11	70.18	67.10	48.86
Our Model- R_d	77.42	72.85	67.49	49.32

Table 3: Results on PDTB 2.0 using the Lin split under the F2C decoding order.

Setting ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$)	L1		L2	
	Acc	F1	Acc	F1
Original (0.30, 0.15, 0.50, 0.05)	78.37	75.04	68.39	52.73
L1-heavy (0.40, 0.15, 0.40, 0.05)	78.78	75.48	67.69	52.19
L2-heavy (0.20, 0.15, 0.60, 0.05)	77.15	74.16	68.74	53.14
Hierarchy-heavy (0.25, 0.25, 0.45, 0.05)	77.53	74.42	67.88	52.32
No connective (0.30, 0.15, 0.55, 0.00)	78.20	74.86	68.45	52.82

Table 4: Reward-weight sensitivity on PDTB 2.0 using Llama-3.1-8B-Instruct under the F2C decoding order. The original setting is used in the main experiments.

SFT baseline and our methods with Llama-3.1-8B-Instruct under the F2C decoding order. Both sequence-level optimization variants improve over Joint SFT, with R_d achieving the best performance. In particular, R_d improves over Joint SFT by 7.18 points in L2 accuracy and 4.09 points in L2 F1. These results suggest that the improvements are not specific to the Ji split (Ji and Eisenstein, 2015) used in the main experiments.

D RLOO Ablation

To disentangle the effect of sequence-level reward optimization from the GRPO-specific update rule, we add an RLOO baseline on PDTB 2.0 using Llama-3.1-8B-Instruct under the F2C decoding order. RLOO is a REINFORCE-style policy-gradient method that samples complete label sequences and optimizes the same sequence-level rewards used by our GRPO models. For variance reduction, each sampled sequence uses the mean reward of the other samples from the same input as a leave-one-out baseline (Ahmadian et al., 2024). Unlike GRPO, RLOO does not use normalized group-relative advantages or a clipped surrogate objective.

Table 2 reports the results. Both RLOO variants substantially improve over Joint SFT, confirming that sequence-level reward optimization itself is beneficial. GRPO further improves over RLOO, especially under the sparse reward, indicating that GRPO provides more effective optimization for this sequence-level objective.

Relations	Joint SFT	Our Model- R_d
Comparison	65.40	68.18
Contingency	66.89	70.79
Expansion	76.78	84.03
Temporal	66.67	71.90

Table 5: Per-relation F1 on PDTB 2.0 at Level 1.

E Reward Weight Sensitivity

To demonstrate the robustness of our dense reward, we conduct sensitivity experiments on PDTB 2.0 using Llama-3.1-8B-Instruct under the F2C decoding order with alternative configurations of $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ in Equation 7. Table 4 shows that all variants outperform the Joint SFT baseline in Table 1, indicating that the gains are robust across reasonable weight choices. The results show clear trade-offs across hierarchy levels: increasing the L1 weight improves L1 performance while slightly reducing L2 performance, whereas increasing the L2 weight improves L2 performance while slightly reducing L1 performance. The original setting provides balanced performance across levels. Removing the connective compatibility term yields comparable final performance, suggesting that the final L1/L2 gains are not driven solely by connective-level supervision.

F Relation Analysis

We investigate which relations benefit from sequence-level optimization. Tables 5 and 7 report the per-relation F1 scores of Joint SFT and our dense-reward model with Llama-3.1-8B-Instruct on PDTB 2.0 under F2C decoding order using the Ji split. Our model with R_d improves F1 across all L1 relations and across nearly all L2 relations with non-zero performance. The only exceptions are extremely sparse relations such as Pragmatic Cause and Concession, where both models obtain zero F1. These results indicate that the improvements are not solely driven by frequent labels.

G Results Using Qwen

To further demonstrate the generality of our approach, we conduct additional experiments using Qwen-2.5-7B-Instruct. Table 6 reports the results on PDTB 2.0 and PDTB 3.0. Overall, the results show trends consistent with those observed using Llama-3.1-8B-Instruct. First, Joint SFT consistently outperforms Level-wise SFT, confirming that modeling dependencies across hierarchical levels

		PDTB 2.0				PDTB 3.0			
		L1		L2		L1		L2	
Models		Acc	F1	Acc	F1	Acc	F1	Acc	F1
	Level-wise SFT	70.25 _{1.9}	63.64 _{1.6}	58.51 _{1.6}	44.24 _{1.6}	73.11 _{1.4}	68.32 _{1.6}	62.80 _{2.0}	54.88 _{1.2}
F2C	Joint SFT	70.93 _{1.5}	65.04 _{1.7}	59.38 _{1.8}	48.37 _{1.9}	75.92 _{2.1}	72.12 _{2.2}	64.84 _{1.8}	57.20 _{1.7}
	Joint SFT + SS	71.35 _{1.5}	65.62 _{2.1}	59.74 _{2.1}	48.92 _{1.5}	75.51 _{1.4}	72.36 _{1.5}	64.34 _{2.1}	56.84 _{2.1}
	Our Model - R_s	74.27 _{1.5}	70.34 _{2.1}	64.66 _{2.1}	53.02 _{1.5}	78.56 _{1.4}	75.00 _{1.5}	68.32 _{2.1}	59.30 _{2.1}
	Our Model - R_d	75.80 _{1.3}	72.91 _{1.3}	65.13 _{1.6}	53.68 _{1.6}	79.34 _{1.9}	75.34 _{1.4}	69.28 _{1.5}	60.63 _{1.7}
C2F	Joint SFT	70.08 _{0.8}	65.49 _{1.2}	57.82 _{0.6}	43.71 _{0.8}	71.16 _{1.1}	67.63 _{1.2}	60.60 _{1.2}	52.22 _{1.2}
	Joint SFT + SS	70.64 _{1.7}	66.12 _{1.9}	57.23 _{1.6}	42.46 _{1.7}	70.74 _{1.5}	67.15 _{1.8}	61.18 _{1.8}	60.63 _{1.9}
	Our Model - R_s	69.87 _{0.9}	65.23 _{1.2}	55.54 _{0.9}	40.93 _{1.1}	70.66 _{0.6}	66.38 _{1.0}	58.25 _{0.8}	50.43 _{1.2}
	Our Model - R_d	75.19 _{0.6}	71.56 _{0.5}	64.17 _{0.9}	48.79 _{0.8}	76.25 _{0.7}	71.51 _{0.4}	66.38 _{0.4}	58.09 _{0.6}

Table 6: Mean accuracy and F1 scores (with standard deviations) on PDTB 2.0 and 3.0 using Qwen-2.5-7B-Instruct.

Relations	Joint SFT	Our Model- R_d
Asynchronous	58.62	67.29
Synchrony	78.82	82.61
Cause	64.67	70.45
Pragmatic Cause	0.00	0.00
Contrast	57.11	66.12
Concession	0.00	0.00
Conjunction	54.54	64.66
Instantiation	68.50	75.52
Restatement	62.38	67.53
Alternative	66.67	78.26
List	18.07	25.47

Table 7: Per-relation F1 on PDTB 2.0 at Level 2.

is beneficial for multi-level IDRC. Second, introducing scheduled sampling (Joint SFT + SS) brings only marginal improvements, suggesting that this strategy alone is insufficient to address exposure bias in this task. Third, our method significantly outperforms the Joint SFT baseline under the F2C setting when using either sparse or dense rewards, while under the C2F setting, only the dense reward leads to substantial improvements. Finally, the F2C decoding order consistently yields better performance than the C2F order, further supporting our analysis in the main text.