Fine-Grained Prototype Adaptation for Few-Shot Document-Level Relation Extraction

1st Hongxia Jin
School of Computer Software
Tianjin University
Tinjin, China
jinhongxia@tju.edu.cn

2nd Dazhuang Wang
School of Computer Software
Tianjin University
Tinjin, China
wdz@tju.edu.cn

3rd Xiaowang Zhang ∗
School of Computer Software
Tianjin University
Tinjin, China
xiaowangzhang@tju.edu.cn

***Abstract*—Few-shot document-level relation extraction (FSDLRE) is often hindered by prototype dilution, where critical relational evidence is obscured by irrelevant document-level context and semantic overlap. To address this, we propose a framework that enhances prototype discriminability through three key innovations: (1) A Query-driven Evidence Adaptation (QE) module that dynamically aligns support instances with query-specific evidence to filter contextual noise; (2) A Relation-guided Feature Attention (RF) module that leverages relational semantic anchors to dynamically recalibrate feature importance, thereby enhancing the discriminative power for overlapping relations; and (3) A Task-specific NOTA Representation (TN) that adaptively refines decision boundaries for relation-absent instances. Experiments on FREDo and ReFREDo benchmarks show our model achieves state-of-the-art results, outperforming competitive baselines. Visualization and case studies further confirm our model's robustness in capturing intra-class consistency and inter-class distinction in complex document contexts.**

***Index Terms*—Few-shot Learning, Relation Extraction, Prototypical Networks, Relation Overlap**

I. Introduction

Few-shot document-level relation extraction (FSDLRE) has emerged as a critical task, requiring models to infer complex, inter-sentential relationships from only a few labeled documents [1]–[3]. Unlike sentence-level RE, FSDLRE necessitates reasoning over long-range contexts where entity mentions are scattered across sentence boundaries [4], [5]. As illustrated in Fig. 1, determining a "Place of birth" relation may require synthesizing information from non-adjacent sentences. Formally, the task is organized into episodes: given a small support set and a query document, the model must categorize query entity pairs into candidate relations or a "None of the Above" (NOTA) class [6], [7]. While existing meta-learning approaches reduce the dependence on extensive annotations , the combination of long-range dependencies and limited supervision makes robust representation learning particularly difficult.

Despite their potential, conventional FSDLRE models primarily rely on uniform averaging to construct relation prototypes [8]–[10]. Unlike sentence-level tasks, document-level contexts are inherently noisier; such coarse-grained aggregation leads to "prototype dilution," where discriminative relational features are masked by redundant document-level noise. Furthermore, existing methods often treat relation labels as isolated indices or simple embeddings, failing to resolve "semantic overlap" where distinct relations share identical entity mentions or discursive patterns. Our diagnostic analysis on the FREDo dataset reveals a critical bottleneck: over 35% of misclassification errors occur between relations sharing similar semantic roots. This confirms that representation ambiguity—the inability to decouple overlapping semantics in high-dimensional space—remains a primary challenge that current static prototypical networks fail to address.

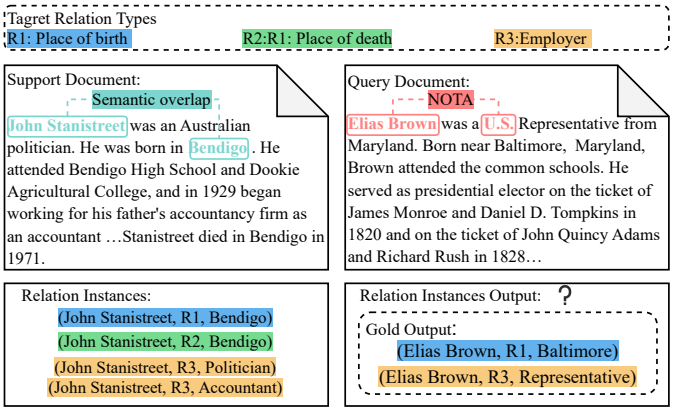

Fig. 1: Description of an episode in the FSDLRE. Gold output refers to the ground-truth relation label assigned to each query instance in a given episode.

To resolve these bottlenecks, we propose a Fine-Grained Prototype Adaptation Framework for FSDLRE. Diverging from previous models that assume all support instances contribute equally, our approach employs Query-driven Evidence Adaptation (QE) to perform instance-level alignment, dynamically weighting support evidence to mitigate prototype dilution. To further decouple overlapping semantics, unlike methods that rely on raw feature distributions, our Relation-guided Feature Attention (RF) module calibrates the latent representation space by leveraging LLM-generated semantic priors to amplify discriminative feature subspaces. Finally, to stabilize the decision boundary, we introduce a Task-specific

∗: Corresponding author.

NOTA Representation (TN) to handle the pervasive "None-of-the-Above" noise more robustly than standard zero-vector or mean-based baselines.

The main contributions of this work are summarized as follows: (1) We develop a hierarchical representation learning framework that replaces traditional static averaging with a query-aware adaptation mechanism (QE). This systematically refines relation prototypes at the instance level, specifically targeting the prototype dilution problem in document-level contexts. (2) We introduce a Relation-guided Feature Attention (RF) mechanism that, unlike prior label-embedding techniques, utilizes fine-grained external relation descriptions. This provides a principled way to recalibrate feature importance and decouple semantically overlapping relations in few-shot scenarios. (3) We propose a Task-specific NOTA (TN) modeling strategy that builds a hybrid noise representation. This ensures a more stable decision boundary for relation-absent instances compared to existing FSDLRE baselines. Extensive experiments on the FREDo and ReFREDo benchmarks demonstrate that our method achieves new state-of-the-art performance, with in-depth analyses confirming its effectiveness in resolving inter-class ambiguity.

II. Related work

FSDLRE builds upon two research pillars. (1) Document-Level RE focuses on long-range dependencies using sequence-based encoding [5] or graph-based modeling [11]–[13]. However, these paradigms typically rely on dense supervision, making them ill-suited for low-resource scenarios. (2) Few-Shot RE addresses data scarcity via prototypical networks [9] and metric learning [14], [15]. Yet, these are primarily designed for sentence-level inputs and struggle with the high-dimensional noise inherent in document-level contexts.

Recent FSDLRE methods like RAPL [16] and TPN [17] utilize regularized prototypes to capture document patterns. Despite their progress, two gaps remain: (i) Representation Collapse, where uniform averaging fails to decouple overlapping relations, and (ii) NOTA Instability, as extremum-based approximations for NOTA lacks robustness for task-specific negatives. Our work bridges these gaps by introducing hierarchical collaboration between instance-level adaptation and feature-level calibration.

III. Methodology

A. Problem Formulation

Each FSDLRE episode comprises a support set $\mathcal{S}$ consisting of a small number of labeled documents and a query document $\mathcal{Q}$. Let $\mathcal{R}$ be the set of target relation types in the current episode. For each relation $r \in \mathcal{R}$, the set of support instances is denoted as $\mathcal{H}_r = \{h^s_{r,i}\}_{i=1}^{k_r}$, where k_r is the total number of instances of relation r extracted from the support documents, and $h^s_{r,i} \in \mathbb{R}^d$ is the contextual embedding of the i-th instance. Similarly, each query instance from $\mathcal{Q}$ is represented by its embedding $h^q \in \mathbb{R}^d$. The objective is to learn a discriminative mapping $f(\mathcal{H}_\mathcal{R}, h^q) \rightarrow \mathcal{R} \cup \{NOTA\}$, which assigns the query instance to either a target relation or NOTA class. To evaluate generalization, the relation sets used during training and testing are strictly disjoint.

B. Framework Overview

As illustrated in Fig. 2, our proposed framework addresses the representation collapse in FSDLRE via a fine-grained prototype adaptation architecture.

C. Query-driven Evidence Adaptation

Standard prototypical networks construct relation representations via uniform averaging, making them vulnerable to prototype dilution from noisy or irrelevant document contexts. To mitigate this, we propose a query-driven mechanism to dynamically recalibrate each support instance's contribution. For a query h^q and target relation r, we first compute the fine-grained interaction between h^q and each support instance $h^s_{r,i} \in \mathcal{H}_r$ via an element-wise Hadamard product:

$$u_{r,i} = \tanh(h^q \odot h^s_{r,i}) \tag{1}$$

This interaction is then aggregated into a scalar matching score $s_{r,i} = \sum_{d=1}^{D}(u_{r,i})_d$, where D is the embedding dimension. Theoretically, this operation is equivalent to measuring the empirical mean of the activated joint features. Then, we normalize these scores across the k_r instances within the support set of relation r using a softmax function to obtain the query-driven weight distribution:

$$\alpha_{r,i} = \frac{\exp(s_{r,i})}{\sum_{l=1}^{k_r}\exp(s_{r,l})} \tag{2}$$

The final enhanced prototype $\tilde{p}^r$ is then formulated as a weighted sum of the support instances:

$$\tilde{p}^r = \sum_{i=1}^{k_r} \alpha_{r,i} h^s_{r,i} \tag{3}$$

D. Relation-guided Feature Attention

To resolve semantic ambiguity, we leverage natural language descriptions $desc_r$ generated by a Large Language Model (LLM) to provide global semantic priors. For each relation r in the dataset, we provide its name and a set of representative example sentences to the LLM. The prompt explicitly instructs the model to generate a concise, feature-level definition that emphasizes how this relation differs from semantically similar ones. These descriptions are encoded as $e_r = \text{Encoder}(desc_r)$, where $e_r \in \mathbb{R}^D$ serves as the semantic anchor for relation.

We compute the feature-wise attention $\theta_r \in \mathbb{R}^D$ by fusing the global description e_r with the instance-level prototype $\tilde{p}^r$:

$$\theta_r = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1[e_r \oplus \tilde{p}^r])) \tag{4}$$

where $\oplus$ denotes concatenation and $\mathbf{W}_1, \mathbf{W}_2$ are learnable parameters. The final metric between query h^q and prototype $\tilde{p}^r$ is then defined as a weighted squared Euclidean distance:

$$d(\tilde{p}^r, h^q) = \sum_{d=1}^{D} \theta_{r,d} \cdot (\tilde{p}^r_d - h^q_d)^2 \tag{5}$$

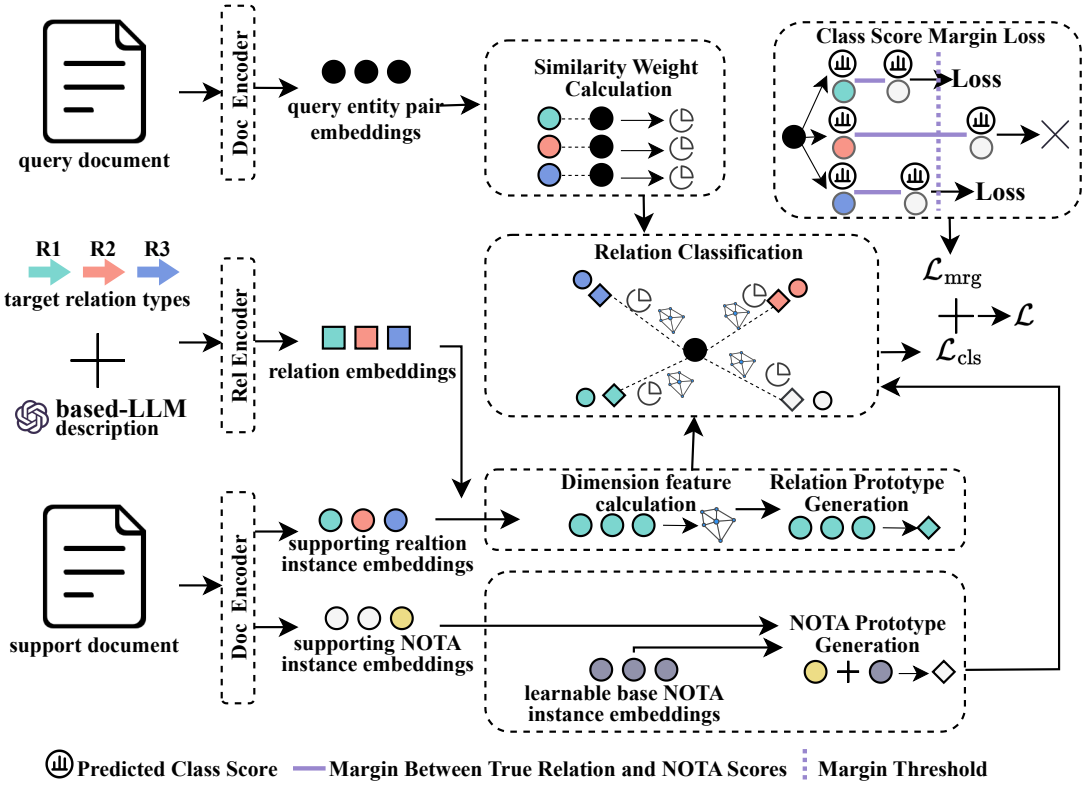

Fig. 2: The overall architecture of our method.

where $\theta_{r,d}$ denotes the learned importance weight for the d-th feature dimension of relation r. By selectively amplifying informative feature subspaces indicated by the relation description, the model can effectively decouple semantically overlapping relations.

E. Task-specific NOTA Representation

Task-specific NOTA representation fuses a Base NOTA Prototype with an Episode-Specific Correction. During training, we maintain a stable foundation p^{base} by averaging a fixed number of sampled NOTA embeddings: $p^{\text{base}} = \frac{1}{N_{sam}} \sum h_i^{\text{NOTA}}$. In each episode, we identify "hard" NOTA instances that are most representative of the current task's negative space. For each task-specific NOTA instance $h_j^s \in \mathcal{H}_{\text{NOTA}}$, we compute a discriminative score s_j:

$$s_j = \text{sim}(h_j^s, p^{\text{base}}) - \max_{r \in \mathcal{R}} \text{sim}(h_j^s, \tilde{p}^r) \tag{6}$$

where $\text{sim}(\cdot)$ denotes cosine similarity. This scoring mechanism assigns higher priority to instances that align with the global NOTA prior while remaining distant from the current task's true relations. The corrected component p^{spec} is then:

$$\beta_j = \frac{\exp(s_j)}{\sum_{l=1}^{n_{\text{NOTA}}} \exp(s_l)} \tag{7}$$

where n_{NOTA} denotes the total number of task-specific NOTA instances in the current episode.

$$p^{\text{spec}} = \sum_j \beta_j h_j^s \tag{8}$$

The final robust NOTA prototype $\tilde{p}^{\text{NOTA}}$ is computed via a rectification factor $\lambda \in [0, 1]$:

$$\tilde{p}^{\text{NOTA}} = \lambda p^{\text{base}} + (1 - \lambda) p^{\text{spec}} \tag{9}$$

F. Loss Function

To enhance the discriminability of the NOTA boundary, we optimize the model using a combination of binary cross-entropy and a max-margin constraint.

For each query instance $q \in \mathcal{Q}$, we compute the logit difference $x_{r,q} = \text{sc}_r(q) - \text{sc}_{\text{NOTA}}(q)$, where $\text{sc}(\cdot)$ is the similarity metric. The probability of relation r being present is $P_{r,q} = \sigma(x_{r,q} - \eta)$, where η is a learnable bias representing the global decision threshold. The classification loss is defined as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|\mathcal{Q}|_R} \sum_{q \in \mathcal{Q}} \sum_{r=1}^{R} [y_{r,q} \log P_{r,q} + (1 - y_{r,q}) \log(1 - P_{r,q})] \tag{10}$$

where $y_{r,q} \in \{0, 1\}$ is a binary indicator (ground truth) that equals 1 if the query instance q carries the relation r, and 0 otherwise. Note that in the FSDLRE task, if q belongs to the NOTA category, $y_{r,q} = 0$ for all $r \in \mathcal{R}$

To enforce a clear separation between positive relations and the NOTA category, we introduce a margin penalty δ:

$$\mathcal{L}\text{mrg} = \frac{1}{\sum y_{r,q}} \sum_{q,r} y_{r,q} \cdot \max(0, \delta - x_{r,q}) \tag{11}$$

This term explicitly pushes true relation scores to be at least δ higher than the NOTA score, effectively clustering positive instances away from the NOTA base.

IV. Experiments

A. Experimental Setup

Benchmarks and evaluation metric. We evaluate our framework on FREDo [6] and its refined version ReFREDo [16]. To the best of our knowledge, these are the only publicly available benchmarks specifically designed for the FSDLRE task. These datasets cover both In-domain (Wikipedia) and Cross-domain (Academic papers) scenarios, providing a comprehensive assessment of generalization capabilities. Following [6], we adopt the Macro F1-score as our primary metric to ensure an unbiased assessment of imbalanced distributions. All results are averaged over five random seeds.

Baselines. We compare our model against several representative FSDLRE baselines, categorized as follows: (1) Basic RE Models: DL-Base directly utilizes the pre-trained BERT; DL-MNAV [5] and its variants (DL-MNAV$_{SIE}$, DL-MNAV$_{SIE+SBN}$) adapt document-level RE for few-shot settings by incorporating support instance embeddings and a basic NOTA prototype [6]. (2) Attention-based Models: CHAN [18] employs a hierarchical attention mechanism to aggregate information from sentence to document levels. (3) Advanced Prototypical Networks: RAPL [16] enhances prototypes by fusing label semantics to improve discriminative power. TPN [17] integrates transfer learning with prototypical networks to enhance semantic propagation, particularly in cross-domain scenarios. (4) LLM-based Baselines: ChatGPT (ICL) evaluates the in-context learning capability of Large Language Models [19].

Implementation details. We implement our framework using PyTorch and utilize BERT-base-cased as the primary encoder. Relation descriptions are generated by GPT-4o and encoded into 768-dimensional vectors. During training, we use the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 4 episodes. For the proposed modules, the rectification factor λ is set to 0.85, and the loss balance coefficient γ is 1.0. The margin δ is empirically fixed at 0.1. To prevent overfitting, we apply a dropout rate of 0.1 and use a linear learning rate scheduler with a warmup ratio of 0.1. All models are trained on a single NVIDIA RTX 4090 GPU, and the best checkpoints are selected based on the performance on the development set.

B. Main Result

Table I summarizes the performance across eight settings, where our model consistently achieves competitive results. We observe more pronounced gains in 1-Doc settings (e.g., 4.85% in FREDo cross-domain) than in 3-Doc settings, suggesting our architecture is particularly effective under extreme data sparsity. By leveraging semantic anchors via the attention mechanism, the model better identifies discriminative features within noisy document-level contexts. Furthermore, steady improvements on the ReFREDo dataset—reaching 18.71% in the in-domain 3-Doc task—demonstrate the model's robustness against semantic overlap. Finanlly, while LLM like ChatGPT have shown promise in general NLP, their performance in strict FSDLRE settings remains limited without fine-tuning [19]–[21]. For instance, on the ReFREDo benchmark, ChatGPT (in-domain, 3-Doc) achieves a Macro F1 of only 5.39%, significantly trailing our specialized framework. These results indicate that the task-specific NOTA representation and feature recalibration effectively prevent prototype dilution, ensuring a stable decision boundary and strong generalization during cross-domain transfer.

C. Ablation Study

We conduct ablation experiments on the ReFREDo dataset to evaluate the contribution of each module (Table II). (1) Removing the QE module leads to a decline in in-domain performance. This suggests that grounding prototypes in query-specific evidence effectively filters out irrelevant context in long documents. (2) The removal of RF results in significant degradation, particularly in the 1-Doc cross-domain setting. This underscores that LLM-derived anchors are crucial for identifying discriminative feature subspaces when support instances are extremely sparse. (3) Notably, excluding the TN module incurs the most consistent and substantial drop across all scenarios. This confirms that a fixed NOTA prototype is insufficient for the FSDLRE task; instead, dynamically adapting the NOTA boundary to the current task's semantic distribution is essential for reducing misclassifications in relation-absent pairs.

D. Analyses and Discussions

Robustness Against NOTA Instances We evaluate our task-specific NOTA representation (TN), which fuses a stable Base NOTA Prototype with an Episode-Specific Correction. As illustrated in Figure 3, incorporating the TN module significantly reduces FP-1 errors (misclassifying NOTA as a target relation) by an average of 12.88% across all settings. This improvement validates the effect our hybrid design.

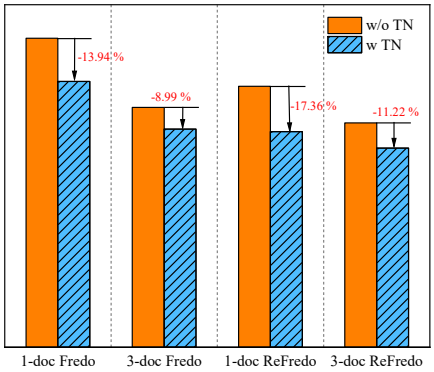

Fig. 3: Effect of task-specific NOTA representation (TN).

Effect of hyperparameters. We evaluate the impact of hyperparameters on the ReFREDo 3-Doc task (Figure 4): (1) Performance peaks at $\lambda = 0.85$, demonstrating that

TABLE I: Performance comparison of different models.

Model	FREDo				ReFREDo			
	In-domain		Cross-domain		In-domain		Cross-domain	
	1-Doc	3-Doc	1-Doc	3-Doc	1-Doc	3-Doc	1-Doc	3-Doc
DL-Base	0.60	0.89	1.76	1.98	1.38	1.84	1.76	1.98
DL-MNAV	7.05±0.18	8.42±0.64	0.84±0.16	0.48±0.21	12.97±0.88	12.43±0.36	1.12±0.38	2.28±0.19
DL-MNAV$_{\text{SIE}}$	7.06±0.15	6.77±0.21	1.77±0.60	2.51±0.66	13.37±0.98	12.00±0.80	1.39±0.74	2.92±0.41
DL-NAV$_{\text{SIE-SBN}}$	1.71±0.04	2.79±0.24	2.85±0.12	3.72±0.14	4.59±0.30	5.43±0.24	2.84±0.24	3.86±0.27
CHAN	7.91	8.65	1.95	2.78	–	–	–	–
RAPL	8.75±0.80	10.6±0.77	3.33±0.50	5.35±0.72	15.20±0.82	16.35±0.60	3.51±0.79	5.48±0.63
TPN	9.12±0.31	8.64±0.44	3.98±0.28	4.48±0.29	15.54±0.14	15.73±0.53	4.72±0.43	5.02±0.50
ChatGPT$_{\text{ICL}}$	2.25	2.95	5.22	5.83	2.86	5.39	5.22	5.83
Ours	**10.08±0.85**	**11.75±0.32**	**4.85±0.46**	**5.46±0.78**	**16.81±0.60**	**18.71±0.15**	**5.43±0.65**	**5.52±0.34**

TABLE II: Ablation study results on ReFREDo.

Model	In-domain		Cross-domain	
	1-Doc	3-Doc	1-Doc	3-Doc
Ours	16.81	18.71	5.43	5.52
w/o QE	16.23	17.36	5.71	5.12
w/o RF	15.78	18.34	4.21	4.46
w/o TN	15.63	17.98	4.92	4.34

balancing the global NOTA prior with task-specific instances is superior to relying on either alone. Deviations toward $\lambda = 1.0$ or $\lambda = 0.8$ cause performance drops, as the former lacks task-adaptation while the latter is prone to local noise. (2) The model achieves optimal results at $\gamma = 1.0$, with higher sensitivity observed in in-domain scenarios. Increasing γ beyond 1.0 gradually degrades performance, suggesting that an excessive margin penalty may lead to overfitting on support distributions and hinder generalization to diverse query instances. Beyond the primary hyper-parameters λ and γ, the stability of the learnable threshold η also plays a crucial role in maintaining a clear NOTA boundary. Our empirical observations suggest that η converges to a stable range across different K-shot tasks, effectively acting as an automated calibrator for 'rejection' confidence.

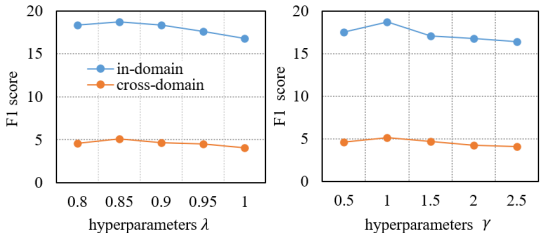

Fig. 4: Effect of hyperparameters λ and γ under the 3-Doc task setting in ReFREDo.

Support Embeddings Visualization. To intuitively evaluate our model's discriminative power, we visualize support embeddings for three semantically overlapping relations (P17, P495, and P1001) using t-SNE (Figure 5). Comparing our full model (a) with variants (b) and (c), we observe that our method produces more compact clusters and clearer decision boundaries, despite the inherent jurisdictional semantic similarities. This enhancement is driven by two mechanisms. First, the QE module dynamically aligns support instances with query-specific evidence, effectively filtering out non-relational context and promoting intra-class consistency. Second, the RF module acts as a dynamic feature mask that highlights informative semantic dimensions while suppressing noise. This implicitly refines the embedding space, ensuring that even under domain shift, the model can isolate the most distinctive relational patterns. These results confirm that our task-specific representation learning successfully mitigates relation confusion in complex document-level contexts.

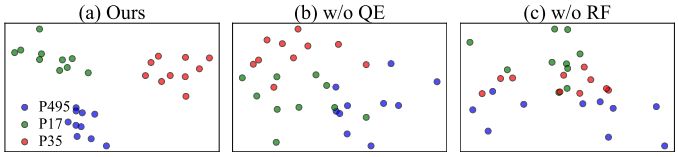

Fig. 5: Three t-SNE visualizations comparing different model settings.

Computational Complexity We evaluate the computational overhead of the QE module compared to the standard prototype baseline. In our QE module, for each query, the model computes pairwise interactions with all relevant instances associated with the target relation within the support documents. While this introduces a linear computational increase proportional to the number of relational instances per support document, the absolute cost remains minimal in few-shot settings. Since the number of instances in 1-Doc or 3-Doc support sets is typically very small, this change represents a nearly constant-level increase in complexity. Our empirical tests indicate that the QE module only increases the average training time per epoch by approximately 4.2%. This marginal cost is well-justified by the improvement in Macro F1, as the fine-grained alignment effectively filters out irrelevant document-level context that simple averaging fails to capture.

Case Study. As shown in Figure 6, our model effectively resolves semantic overlap between P17 (geographic containment) and P495 (country of origin). For the query (Phnom Penh, Cambodia) , the QE module prioritizes the locational evidence in (Athens, Greek) , rather than the institutional

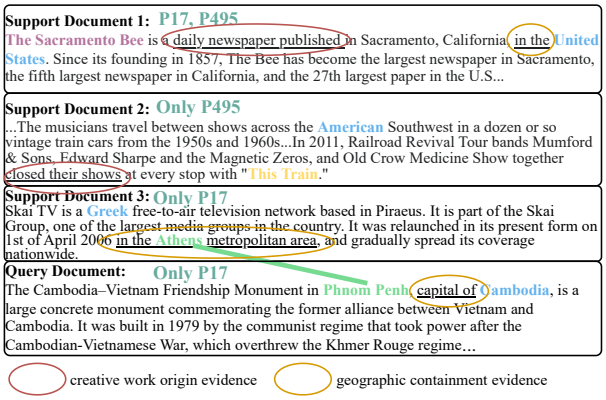

Fig. 6: Case study illustrating the impact in our model.

context in (The Sacramento Bee, US). Simultaneously, the RF module recalibrates features to highlight spatial inclusion while suppressing "organization" dimensions. To further demystify this internal mechanism, we analyzed the attention weight distribution and similarity scores. Quantitatively, the QE module increases the similarity score between the query and Support Document 3 (P17) by 24.3% compared to the vanilla prototype, effectively "denoising" the irrelevant institutional information in Support Document 1. Furthermore, the RF module's recalibration leads to a significant shift in the feature space: the activation of spatial-related dimensions is enhanced by 1.8×, while the overlap with the P495 (origin) category is reduced by 15.6%. These data-driven insights confirm that our model's superiority stems from its ability to perform fine-grained evidence alignment and task-specific feature filtering, rather than simple pattern matching.

V. Conclusion

We proposed a fine-grained prototype adaptation framework for FSDLRE to address prototype dilution and semantic overlap. By integrating QE and RF modules, the model dynamically aligns prototypes with query-specific evidence and recalibrates feature importance using relational priors. Additionally, the TN module establishes a robust decision boundary for relation-absent instances. Experiments on FREDo and ReFREDo benchmarks demonstrate that our model achieves state-of-the-art performance. Despite these gains, the current study primarily focuses on gold-standard entity metadata. Future research will investigate the model's robustness against real-world perturbations, such as label noise and the absence of explicit coreference information, by developing more resilient prototype recalibration mechanisms.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 23 716–23 736.

[2] E. Perez, D. Kiela, and K. Cho, "True few-shot learning with language models," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 11 054–11 070.

[3] H. Cheng, S. Yang, J. T. Zhou, L. Guo, and B. Wen, "Frequency guidance matters in few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 814–11 824.

[4] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "Docred: A large-scale document-level relation extraction dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 764–777.

[5] W. Zhou, K. Huang, T. Ma, and J. Huang, "Document-level relation extraction with adaptive thresholding and localized context pooling," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 14 612–14 620.

[6] N. Popovic and M. Färber, "Few-shot document-level relation extraction," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5733–5746.

[7] H. Wang, K. Qin, R. Y. Zakari, G. Lu, and J. Yin, "Deep neural network-based relation extraction: an overview," *Neural Computing and Applications*, vol. 34, pp. 1–21, 2022.

[8] K. Ding, J. Wang, J. Li, K. Shu, C. Liu, and H. Liu, "Graph prototypical networks for few-shot learning on attributed networks," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, p. 295–304.

[9] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 1877–1901.

[10] B. Mao, X. Zhang, L. Wang, Q. Zhang, S. Xiang, and C. Pan, "Learning from the target: Dual prototype network for few shot semantic segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 1953–1961, 2022.

[11] W. Xu, K. Chen, and T. Zhao, "Discriminative reasoning for document-level relation extraction," in *Findings of the Association for Computational Linguistics*, 2021, pp. 1653–1663.

[12] T.-J. Fu, P.-H. Li, and W.-Y. Ma, "GraphRel: Modeling text as relational graphs for joint entity and relation extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 1409–1418.

[13] H. Zhu, Y. Lin, Z. Liu, J. Fu, T.-S. Chua, and M. Sun, "Graph neural networks with generated parameters for relation extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 1331–1339.

[14] H. Liu, F. Zhang, X. Zhang, S. Zhao, J. Sun, H. Yu, and X. Zhang, "Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, p. 1079–1087.

[15] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[16] S. Meng, X. Hu, A. Liu, S. Li, F. Ma, Y. Yang, and L. Wen, "RAPL: A relation-aware prototype learning approach for few-shot document-level relation extraction," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5208–5226.

[17] Y. Zhang and Z. Kang, "Tpn: Transferable proto-learning network towards few-shot document-level relation extraction," in *2024 International Joint Conference on Neural Networks*, 2024, pp. 1–9.

[18] D. Wang, S. Wu, X. Zhang, and Z. Feng, "Multi-relation identification for few-shot document-level relation extraction," in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 52–64.

[19] D. Pan, Y. Sun, B. Xu, J. Li, Z. Yang, L. Luo, H. Lin, and J. Wang, "Prototype tuning: A meta-learning approach for few-shot document-level relation extraction with large language models," in *Findings of the Association for Computational Linguistics*, 2025, pp. 1112–1128.

[20] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!" in *Findings of the Association for Computational Linguistics*, 2023, pp. 10 572–10 601.

[21] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," 2023.