HyperLEGNN: Hypergraph Label-Element Neural Networks for Low-Resource Chinese Legal Event Detection

Jun Zhu[1], Hanyu Yang[2], and Xin-Yu Xiao[2]✉

[1] Tianjin Normal University, Tianjin, China
[2] University of Chinese Academy of Sciences, Beijing, China
xiaoxinyu24@mails.ucas.ac.cn

Abstract. Low-resource Chinese legal event detection remains challenging because rare event types are annotated with only a few triggers, while semantically related events often share similar lexical and structural patterns. To address this, we propose HyperLEGNN, a heterogeneous label-side hypergraph neural network that learns more expressive event-type representations under limited supervision. HyperLEGNN jointly encodes event labels, trigger patterns, argument roles, behavior elements, and support instances through schema, support, and confusion hyperedges. Relation-aware propagation over the hypergraph integrates these complementary signals to generate event-specific classifier weights, while confusion-aware learning sharpens the distinction between closely related event types. On the LEVEN benchmark, HyperLEGNN attains 53.2 Macro-F1 in the 8-shot setting, outperforming the strongest hypergraph baseline by 1.8 F1 points. It further achieves 65.7 Tail Macro-F1 and 61.5 Confusable Event F1, corresponding to improvements of 1.9 and 1.8 points, respectively. Experiments on five additional event detection benchmarks show consistent gains of 0.8–1.5 Macro-F1 points over the strongest corresponding baselines. These results demonstrate the effectiveness of HyperLEGNN for low-resource legal event detection, particularly for rare and semantically confusable event types.

Keywords: Chinese legal NLP · Event detection · Low-resource learning · Hypergraph neural networks · Label representation learning

1 Introduction

Chinese legal event detection identifies event triggers in case fact descriptions and assigns them to fine-grained legal event types. It provides structured representations of case facts that can support several downstream legal NLP tasks. CAIL2018 studies criminal judgment prediction from factual descriptions [41], while LeCaRD focuses on Chinese legal case retrieval [26]. Statute retrieval from real legal queries is investigated in STARD [31], and JEC-QA evaluates question answering in the legal domain [48]. Compared with these document-level prediction and retrieval tasks, legal event detection additionally requires precise

trigger localization and fine-grained discrimination. LEVEN contains 8,116 legal documents, 150,977 event mentions, and 108 event types [45].

A central challenge is the scarcity of event-type supervision. Trigger-level annotation requires annotators to determine both trigger boundaries and event categories, which limits the number of labeled instances available for rare event types. The problem is further complicated by the semantic proximity of legal events. Fraud, theft, robbery, and extortion may all involve property transfer, yet they are distinguished by behavioral elements such as deception, secret taking, violence, and threat. Under limited supervision, conventional classifiers estimate each event type mainly from its own annotated instances. Rare labels therefore receive weak supervision, while semantically adjacent labels remain difficult to distinguish when their trigger expressions and argument structures overlap.

Existing studies improve event detection mainly through stronger contextual representations, structured event extraction, and low-resource learning. BERT provides a general contextual representation foundation for event detection [11], while MacBERT improves Chinese language representation through modified pre-training objectives [9]. For legal text, Lawformer further adapts pre-trained representations to long Chinese legal documents [40]. Beyond contextual encoding, DMCNN models trigger-centered local contexts for event extraction [5], and OneIE jointly models multiple information extraction structures [22]. Generative approaches such as Text2Event formulate event extraction as structured sequence generation [24], while UIE provides a unified structure-generation framework for information extraction [25]. Under limited supervision, ProtoNet represents classes through support-derived prototypes [30], P4E introduces prompt-guided few-shot event detection [19], and MetaEvent applies prompt-based meta-learning to zero- and few-shot settings [47]. Despite these advances, existing methods mainly improve textual representations or derive event categories from individual support examples. The structured information associated with an event type remains insufficiently modeled in a unified label-side structure.

To address this limitation, we propose HyperLEGNN, a heterogeneous label-side hypergraph neural network for low-resource Chinese legal event detection. HyperLEGNN represents each event type through event labels, trigger patterns, argument roles, behavior elements, and K-shot support instances. Schema hyperedges capture the structural composition of event types, support hyperedges incorporate supervision from labeled examples, and confusion hyperedges characterize relations among semantically adjacent event types using training-only information. Relation-aware hypergraph propagation integrates these heterogeneous relations and produces event-type representations that are used to generate classifier weights for trigger-context representations. A confusion-aware gate and margin objective further improve discrimination among event types with overlapping trigger or role information.

The main contributions of this work are as follows.

- We introduce a heterogeneous label-side hypergraph for low-resource legal event detection that jointly represents event semantics, trigger patterns, argument roles, behavior elements, and limited support instances.

– We develop relation-aware hypergraph propagation over schema, support, and confusion relations and use the resulting event-type representations to generate classifier weights for trigger-context classification.
– We introduce confusion-aware learning for semantically adjacent legal event types. Experiments on LEVEN show that HyperLEGNN achieves 53.2 Macro-F1 in the 8-shot setting, compared with 51.4 for the strongest comparable HGNN-label baseline, while improving Tail Macro-F1 from 63.8 to 65.7 and Confusable Event F1 from 59.7 to 61.5.

2 Related Work

2.1 Graph Neural Networks and Hypergraph Applications

Graph neural networks learn representations from relational structures. GCN uses normalized neighborhood aggregation [18]. GAT assigns attention weights to neighbors [33]. GraphSAGE supports inductive sampling [16]. R-GCN models typed relations in heterogeneous graphs [28]. Recent surveys summarize these architectures and their trade-offs [39]. In NLP, TextGCN builds document-word graphs [46]. Heterogeneous graph learning uses multiple textual node types [17]. Label-aware GCN propagates information between labels and textual units [4]. These methods motivate label-side modeling, but pairwise edges cannot directly express that one legal event type is jointly determined by a trigger pattern, several argument roles, and behavior elements.

Hypergraph learning represents high-order relations by connecting more than two nodes with one hyperedge. Classical hypergraph regularization was formulated by Zhou et al. [49]. HGNN performs neural propagation over the hypergraph incidence matrix [14]. HyperGCN approximates hyperedges with graph expansion [44]. HyperGAT introduces attention into hypergraph-based text classification [12]. HGNN+ improves adaptive hypergraph modeling [15]. AllSet unifies set-function based hypergraph learning [7]. HyperLEGNN differs from these models by using typed label-side hyperedges for legal event schema, few-shot support, and label confusion, and by generating classifier weights from event-label embeddings rather than using node representations only as features.

2.2 Legal Event Detection and Low-Resource Event Learning

Chinese legal NLP has moved from document-level prediction toward structured fact understanding, while domain-specific model and benchmark construction has also been explored in other specialized settings [42, 43]. Legal judgment prediction benefits from event-level constraints [13]. LEVEN defines legal event detection with fine-grained event types and trigger annotations [45]. DuEE provides a general Chinese event extraction benchmark [20]. FewFC studies Chinese financial events under limited supervision [50]. DocFEE targets document-level Chinese financial event extraction [6]. MAVEN provides a large general-domain English event benchmark [38]. ACE2005 remains a standard English event extraction

benchmark [36]. LEGAL-BERT is used for English legal text representation and is included only in auxiliary English settings [3].

Existing low-resource event methods usually use metric comparison, prompt transformation, or prototype learning. They help when each label has few examples, but they still depend on either support-instance similarity or label text alone. Long-tail losses such as focal loss [21], class-balanced loss [10], and LDAM [2] rebalance gradients but do not change the label parameterization. HyperLEGNN is complementary to these strategies. It changes the event classifier from a fixed label-index matrix to a label-side hypergraph-generated classifier.

3 Method

3.1 Problem Definition

Given a legal fact description $x = (w_1, \ldots, w_n)$, a candidate trigger is represented as a span $t = (i, j)$. For each candidate, the model predicts an event type $y \in \mathcal{Y} \cup \{\text{None}\}$, where $\mathcal{Y}$ denotes the predefined event-type set and None indicates a non-event span.

Candidate triggers are generated using a BIO-CRF tagger built on MacBERT for Chinese datasets and RoBERTa for English datasets. In the strict K-shot setting, the tagger is trained using only the sampled trigger annotations and sampled non-trigger spans. In the type-supervision setting, all training-set trigger boundaries are available to the tagger. The candidate threshold is selected from $\{0.25, 0.30, 0.35, 0.40, 0.45\}$, and the maximum candidate span length is selected from $\{4, 6, 8\}$ words for Chinese datasets and $\{3, 4, 5\}$ tokens for English datasets. Non-event candidates are sampled with a 2.5:1 None-to-gold ratio. All candidate-based baselines are evaluated using the same candidate set.

3.2 Trigger Context Encoder

For a candidate trigger $t = (i, j)$, the input is the document sentence with boundary markers around the span. A pre-trained encoder produces contextual states $\mathbf{H}$. The trigger representation is

$$\mathbf{h}_t = \left[\mathbf{H}_{[\text{CLS}]}; \text{mean}_{k=i}^{j} \mathbf{H}_k; \mathbf{H}_i; \mathbf{H}_j\right] \mathbf{W}_t. \tag{1}$$

The same encoder and candidate list are used across baselines. This makes the comparison focus on event-type representation and classifier generation.

3.3 Heterogeneous Legal Event Hypergraph

The label-side hypergraph is denoted by $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$. It contains three relation families: schema, support, and confusion. The node set contains event-label nodes, trigger-pattern nodes, role nodes, behavior-element nodes, support-instance nodes, and the None node. Table 1 reports construction statistics.

Table 1. Hypergraph statistics. Support nodes are reported as 8-shot/full-data counts.

Dataset	Event nodes	Trigger nodes	Role nodes	Element nodes	Support nodes	Schema edges	Conf. edges
LEVEN	108	684	23	41	864/94,280	108	42
DuEE	65	392	18	31	520/16,742	65	24
FewFC	26	183	19	26	208/7,562	26	13
DocFEE	15	112	23	18	120/4,860	15	8
MAVEN	168	910	21	62	1,344/67,104	168	56
ACE2005	33	244	17	22	264/4,202	33	14

Trigger-pattern nodes. Chinese text is segmented with Jieba and English text with spaCy. Stop words are removed using fixed Chinese and English lists. For each event type, candidate patterns are collected from training triggers, trigger heads, and two-token windows around triggers. We score a pattern by 0.6TFIDF + 0.4PMI, keep the top eight patterns per event type, and merge patterns whose Word2Vec cosine similarity exceeds 0.90. Chinese Word2Vec is trained on training documents and CAIL2018 fact descriptions; English Word2Vec uses the corresponding training documents with Wikipedia initialization [27].

Role and behavior-element nodes. Role nodes are read from the dataset schema. For datasets without explicit roles, we use the role inventory provided by the event extraction format after mapping synonymous names. Behavior elements are extracted from label descriptions and trigger clauses by dependency parsing. The action head and its direct object or modifier form a candidate element. A deterministic mapping table normalizes surface variants. For example, fabricate and deceive are mapped to deception, while seize and threaten are mapped to coercive taking only when a property role is present. The table is built from training data and schema descriptions; no development or test labels are used.

Schema, support, and confusion hyperedges. A schema hyperedge connects one event label with its trigger-pattern, role, and behavior-element nodes. A support hyperedge connects one event label with sampled support-trigger nodes. If a label has fewer than K triggers, all available triggers are used. Support-node features are recomputed at the beginning of each epoch with the current text encoder and detached during hypergraph propagation to reduce memory. Confusion hyperedges use only training-split signals. For event labels y_i and y_j, we compute

$$q_{ij} = 0.40q_{ij}^{\text{desc}} + 0.25q_{ij}^{\text{trig}} + 0.20q_{ij}^{\text{role}} + 0.15q_{ij}^{\text{oof}}, \tag{2}$$

where q^{desc} is label-description similarity, q^{trig} is trigger-pattern overlap, q^{role} is role overlap, and q^{oof} is confusion frequency from a five-fold out-of-fold classifier trained only on the training split. For each label, at most three neighbors with $q_{ij} > 0.55$ are retained. Connected labels form one confusion hyperedge. Hyperedges with more than five labels are split by complete-link clustering. In K-shot experiments, the procedure is rerun on each sampled split.

3.4 Relation-Aware Hypergraph Propagation

Each node $v \in \mathcal{V}$ is initialized by a textual representation of its name, description, or trigger context. For relation r, let $\mathbf{H}_r \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{E}_r|}$ denote the incidence matrix. A propagation layer is

$$\mathbf{Z}_r^{(l+1)} = \mathbf{D}_{v,r}^{-1/2}\mathbf{H}_r\mathbf{W}_{e,r}\mathbf{D}_{e,r}^{-1}\mathbf{H}_r^{\top}\mathbf{D}_{v,r}^{-1/2}\mathbf{Z}^{(l)}\mathbf{W}_r^{(l)}. \tag{3}$$

Here $\mathbf{D}_{v,r}$ and $\mathbf{D}_{e,r}$ are node-degree and hyperedge-degree matrices, and $\mathbf{W}_{e,r}$ stores relation-specific hyperedge weights. Relation outputs are fused by

$$\mathbf{Z}^{(l+1)} = \mathrm{LN}\left(\mathbf{Z}^{(l)} + \mathrm{GELU}\left(\sum_{r\in\mathcal{R}} \alpha_r^{(l)}\mathbf{Z}_r^{(l+1)}\right)\right). \tag{4}$$

All relation outputs share hidden size 256. Dropout is applied to node representations and incidence entries.

For a confusion hyperedge e, the gate is computed from its event-label members:

$$\mathbf{p}_e = [\mathrm{mean}_{y\in e}\,\mathbf{z}_y; \max_{y\in e}\mathbf{z}_y; \mathrm{std}_{y\in e}\,\mathbf{z}_y], \quad g_e = \sigma(\mathrm{MLP}(\mathbf{p}_e)). \tag{5}$$

The gate scales the corresponding confusion hyperedge weight. Its calibration term is

$$\mathcal{L}_{\mathrm{gate}} = \frac{1}{|\mathcal{E}_c|}\sum_{e\in\mathcal{E}_c}(g_e - c_e)^2, \tag{6}$$

where c_e is the mean pairwise training-only confusion score inside e. Pooling makes the gate well-defined for variable-size hyperedges and avoids pairwise computation inside large groups.

3.5 Dynamic Event Classifier

After propagation, each event label and the None node obtain an embedding $\mathbf{z}_y$. The classifier weight is generated by

$$\mathbf{w}_y = \mathbf{W}_o\mathbf{z}_y + \mathbf{b}_o, \quad s(y \mid x,t) = \mathbf{h}_t^{\top}\mathbf{w}_y. \tag{7}$$

None is a node and uses the same mapping, so it has no separate free classifier row. The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \lambda_1\mathcal{L}_{\mathrm{conf}} + \lambda_2\mathcal{L}_{\mathrm{gate}} + \lambda_3\|\Theta\|_2^2, \tag{8}$$

where $\mathcal{L}_{\mathrm{ce}}$ is class-balanced cross-entropy, and

$$\mathcal{L}_{\mathrm{conf}} = \sum_{y^-\in\mathcal{C}(y^+)} \max(0, \gamma - s(y^+ \mid x,t) + s(y^- \mid x,t)). \tag{9}$$

We select $\gamma = 0.35$, $\lambda_1 = 0.4$, $\lambda_2 = 0.1$, and $\lambda_3 = 10^{-5}$ on the development set. At inference, support features and event-type embeddings are cached. One hypergraph layer costs $O(|\mathbf{H}|d + |\mathcal{V}|d^2)$, and the gate costs $O(\sum_{e\in\mathcal{E}_c}|e|d)$.

4 Experiments

4.1 Experimental Questions and Settings

We design the experiments to examine four aspects of HyperLEGNN. RQ1 evaluates its overall effectiveness on LEVEN under type-supervision K-shot, strict K-shot, and oracle-trigger protocols. RQ2 examines whether the improvements remain consistent for tail event types and independently defined groups of confusable labels. RQ3 analyzes the contribution of the major model components, including label descriptions, support instances, schema and confusion hyperedges, gate calibration, and margin learning. RQ4 evaluates the generalizability of HyperLEGNN across additional event detection benchmarks.

LEVEN serves as the primary benchmark, while DuEE, FewFC, DocFEE, MAVEN, and ACE2005 are used for complementary evaluation across different domains and event inventories. For low-resource experiments, we consider $K \in \{8, 16, 32, 64\}$ and sample at most K gold trigger instances for each event type. If an event type contains fewer than K training instances, all available instances are retained. Unsampled trigger annotations in the selected documents are excluded from the event-type objective and are not treated as None instances. All K-shot results are averaged over five random seeds. For frequency-based analysis, event types are divided according to the number of training triggers. Head labels contain more than 100 instances, medium labels contain 20–100 instances, tail labels contain 5–19 instances, and few-shot labels contain fewer than 5 instances.

4.2 Compared Methods and Metrics

Compared methods include BERT [11], RoBERTa [23], MacBERT [9], ChineseBERT [32], Longformer [1], Lawformer [40], LEGAL-BERT [3], DMCNN [5], DyGIE++ [35], OneIE [22], CasEE [29], Text2Event [24], UIE [25], Matching Network [34], ProtoNet [30], PA-CRF [8], P4E [19], prompt tuning [37], MetaEvent [47], TextGCN [46], heterogeneous GAT [17], Label-GCN [4], R-GCN-label [28], HGNN-label [14], HyperGCN-label [44], HyperGAT-label [12], HGNN+ [15], AllSet-label [7], focal loss [21], class-balanced loss [10], and LDAM [2]. Candidate-mode baselines receive a marked trigger span and output one label in $\mathcal{Y} \cup \{\text{None}\}$. We report Trigger F1, Micro-F1, Macro-F1, Tail Macro-F1, Few-shot Macro-F1, Confusable Event F1, and Confusion Error Rate.

5 Results

5.1 RQ1: Main Results

Table 2 presents the main results on LEVEN. HyperLEGNN consistently outperforms the compared methods across different levels of event-type supervision. Compared with HGNN-label, the strongest baseline using the same label-side resources, HyperLEGNN improves Macro-F1 by 1.8, 1.8, 1.6, and 1.3 points under the 8-, 16-, 32-, and 64-shot settings, respectively. Under full supervision,

HyperLEGNN achieves 73.3 Macro-F1 and 83.0 Micro-F1, exceeding HGNN-label by 1.1 and 0.6 points. The advantage is also evident for difficult event types, with Tail Macro-F1 increasing from 63.8 to 65.7 and Confusable Event F1 from 59.7 to 61.5.

Table 2. Main results on LEVEN under the type-supervision K-shot protocol. † indicates statistical significance against the strongest non-HyperLEGNN baseline after Holm–Bonferroni correction.

Family	Method	8-shot	16-shot	32-shot	64-shot	Full-Ma	Full-Mi	Tail	Conf.
Enc.	BERT	40.7±1.3	46.2±1.0	51.8±0.8	57.7±0.7	65.9±0.5	78.2±0.4	54.5±0.9	49.1±1.0
	MacBERT	42.1±1.1	47.8±0.9	53.6±0.8	59.4±0.6	67.4±0.4	79.0±0.3	56.0±0.8	51.2±0.9
	ChineseBERT	42.8±1.2	48.5±0.9	54.0±0.7	59.9±0.6	68.0±0.4	79.4±0.3	56.6±0.8	51.8±0.8
	Lawformer	44.4±1.0	50.1±0.8	55.7±0.7	61.7±0.5	69.2±0.4	80.2±0.3	58.5±0.7	53.4±0.8
Ext.	CasEE	45.6±1.0	51.2±0.8	56.9±0.7	62.3±0.5	69.1±0.4	80.5±0.3	59.2±0.8	54.5±0.8
	Text2Event	46.1±0.9	51.8±0.8	57.5±0.6	62.8±0.5	69.6±0.4	81.0±0.3	60.0±0.7	55.1±0.7
	UIE	47.2±0.9	52.7±0.7	58.4±0.6	63.3±0.5	70.3±0.3	81.5±0.3	60.9±0.7	56.0±0.8
Few.	ProtoNet	47.7±1.3	53.0±1.0	58.5±0.8	63.3±0.7	68.6±0.5	79.9±0.4	60.3±0.9	55.3±0.9
	P4E	49.0±1.1	54.5±0.9	60.0±0.7	64.5±0.6	69.7±0.4	80.8±0.3	61.5±0.8	56.7±0.8
	MetaEvent	49.8±1.0	55.4±0.8	60.8±0.7	65.0±0.5	70.5±0.4	81.2±0.3	62.2±0.7	57.5±0.7
Graph	Label-GCN	48.1±1.0	53.7±0.8	59.5±0.7	64.4±0.6	70.5±0.4	81.1±0.3	61.3±0.8	56.5±0.8
	R-GCN-label	49.0±0.9	54.7±0.8	60.5±0.7	65.2±0.5	71.1±0.4	81.6±0.3	62.3±0.7	57.3±0.8
	LDAM	49.1±1.1	54.6±0.8	60.3±0.7	64.8±0.6	70.3±0.4	80.9±0.3	61.9±0.8	56.4±0.8
Hyper.	HyperGAT-label	50.7±0.9	56.0±0.7	62.0±0.6	65.9±0.5	71.6±0.3	82.0±0.3	63.0±0.7	58.6±0.7
	AllSet-label	51.0±0.8	56.4±0.7	62.4±0.6	66.1±0.5	71.9±0.3	82.2±0.2	63.5±0.7	59.1±0.7
	HGNN-label	51.4±0.8	56.9±0.7	62.7±0.5	66.4±0.5	72.2±0.3	82.4±0.2	63.8±0.6	59.7±0.7
	HyperLEGNN	**53.2±0.7**†	**58.7±0.6**†	**64.3±0.5**†	**67.7±0.4**†	**73.3±0.3**†	**83.0±0.2**	**65.7±0.6**†	**61.5±0.6**†

Table 3 further examines whether the improvement originates from trigger localization or event-type classification. Under strict 8-shot supervision, candidate recall decreases to 67.8, reflecting the limited trigger annotations available to the candidate generator. In this setting, HyperLEGNN improves Trigger F1 over HGNN-label by only 0.3 points, and the corresponding confidence interval includes zero. In contrast, Macro-F1 increases from 42.8 to 44.5, yielding a 1.7-point gain. When the candidate generator has access to all training trigger boundaries, HyperLEGNN improves type-supervision 8-shot Macro-F1 by 1.8 points. The gain remains 1.6 points under oracle-trigger evaluation, where trigger localization errors are completely removed. Figure 1 summarizes the performance of representative methods across the principal evaluation settings.

Table 3. Protocol-controlled results on LEVEN. Strict K-shot restricts both candidate generation and event-type supervision to K-shot annotations. Type-supervision K-shot uses full trigger-boundary supervision while restricting event-type supervision to K instances per type. Oracle trigger provides gold trigger spans and evaluates event-type classification only. Gain is computed relative to HGNN-label.

Protocol	Candidate Recall	Metric	Lawformer	MetaEvent	HGNN-label	HyperLEGNN	Gain	95% CI	p
Strict 8-shot	67.8±2.1	Trigger F1	50.6	51.4	52.1	52.4	+0.3	[-0.1, 0.7]	.058
Strict 8-shot	67.8±2.1	Macro-F1	37.4±1.4	41.9±1.2	42.8±1.0	**44.5±0.9**	+1.7	[0.9, 2.5]	.011
Type-supervision 8-shot	89.8±0.9	Macro-F1	44.4±1.0	49.8±1.0	51.4±0.8	**53.2±0.7**	+1.8	[1.0, 2.6]	.008
Oracle trigger 8-shot	100.0	Type Macro-F1	50.0±0.9	56.0±0.7	57.2±0.7	**58.8±0.6**	+1.6	[0.8, 2.3]	.014
Full-data candidate	95.7±0.4	Macro-F1	69.2±0.4	70.5±0.4	72.2±0.3	**73.3±0.3**	+1.1	[0.6, 1.5]	.018

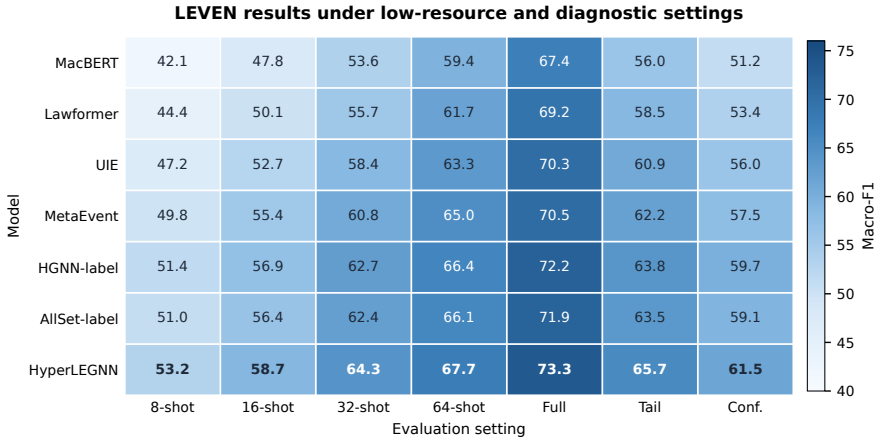

Fig. 1. Comparison of representative methods on LEVEN across K-shot, full-data, tail-label, and confusable-label evaluation settings.

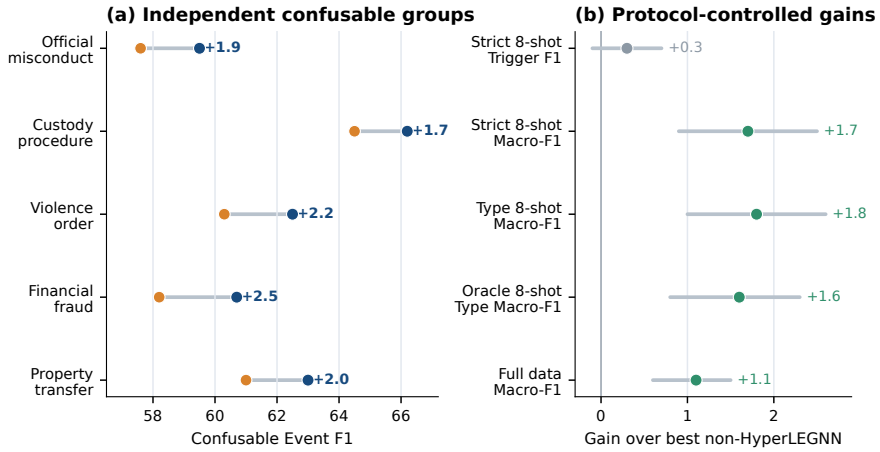

Fig. 2. Diagnostic results on LEVEN. Panel (a) compares HGNN-label and HyperLEGNN on confusable event groups, and panel (b) reports gains over the strongest non-HyperLEGNN baseline under different evaluation protocols.

5.2 RQ2: Tail Labels and Independent Confusable Groups

Figure 2 reports two diagnostic analyses. On independently defined confusable groups, HyperLEGNN improves over HGNN-label by 1.7 to 2.5 points, with the clearest gains in the violence-order groups. The smallest improvement appears in strict K-shot Trigger F1, while type-level metrics show larger and more stable gains with confidence intervals above zero.

5.3 RQ3: Ablation Study

Table 4 and Figure 3 evaluates the contribution of each variant in HyperLEGNN. Replacing the hypergraph-generated classifier with a conventional softmax classifier produces the largest performance drop, reducing 8-shot Macro-F1 from 53.2 to 49.4, Tail F1 from 65.7 to 60.8, and Confusable F1 from 61.5 to 56.5, while increasing CER from 18.4 to 23.6. Removing schema or support hyperedges mainly affects low-resource performance, whereas removing confusion hyperedges reduces Confusable Event F1 by 1.8 points and increases CER by 2.0 points.

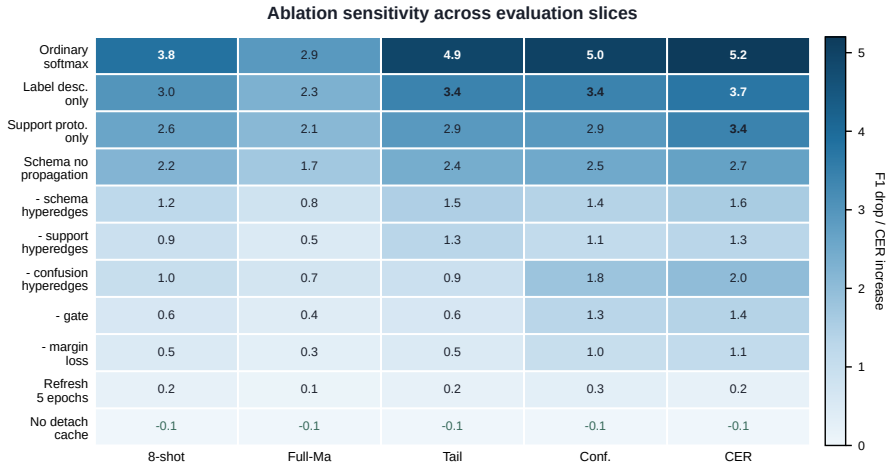

Fig. 3. Performance degradation of different ablation variants.

Table 4. Ablation results on LEVEN. Full-Ma denotes full Macro-F1, Tail denotes Tail Macro-F1, Conf. denotes Confusable F1, and CER denotes Confusion Error Rate.

Variant	8-shot	Full-Ma	Tail	Conf.	CER
Full HyperLEGNN	**53.2±0.8**	**73.3±0.3**	**65.7±0.7**	**61.5±0.8**	**18.4**
Ordinary softmax classifier	49.4±1.2	70.4±0.5	60.8±1.0	56.5±1.1	23.6
Label-description classifier only	50.2±1.1	71.0±0.5	62.3±0.9	58.1±1.0	22.1
Support-prototype classifier only	50.6±1.0	71.2±0.4	62.8±0.8	58.6±0.9	21.8
Schema nodes without propagation	51.0±0.9	71.6±0.5	63.3±0.8	59.0±0.9	21.1
Without schema hyperedges	52.0±0.9	72.5±0.4	64.2±0.7	60.1±0.8	20.0
Without support hyperedges	52.3±0.8	72.8±0.4	64.4±0.8	60.4±0.7	19.7
Without confusion hyperedges	52.2±0.9	72.6±0.4	64.8±0.7	59.7±0.9	20.4
Without confusion gate	52.6±0.8	72.9±0.3	65.1±0.7	60.2±0.8	19.8
Without margin loss	52.7±0.8	73.0±0.3	65.2±0.6	60.5±0.7	19.5
Support refresh every 5 epochs	53.0±0.7	73.2±0.3	65.5±0.6	61.2±0.7	18.6
No-detach support cache	53.1±0.9	73.2±0.4	65.6±0.7	61.3±0.8	18.5

5.4 RQ4: Error Analysis

Table 5 reports smaller gains than LEVEN. The pattern is expected because legal behavior elements are most informative for LEVEN and financial-domain FewFC. On MAVEN and ACE2005, the gain mainly comes from label descriptions, trigger patterns, and support nodes; removing behavior-element nodes causes a smaller reduction. We also inspect 300 LEVEN errors from the full-data setting, sampled uniformly from head, medium, and tail labels. The remaining errors are assigned to trigger-boundary ambiguity, missing implicit intent, confusable property-transfer labels, long-distance argument evidence, and annotation inconsistency. Trigger-boundary errors dominate strict K-shot, while type substitution dominates oracle-trigger evaluation. This split is consistent with the protocol-control table and indicates that label-side modeling mainly addresses type assignment after candidate generation.

Table 5. Auxiliary dataset results. The best baseline is identified separately for each dataset. Scores are full-data Macro-F1.

Dataset	Language / type	Best baseline	Baseline	HyperLEGNN w/o elements	HyperLEGNN	Gain
DuEE	Chinese general events	UIE	76.4±0.4	77.2±0.3	77.8±0.3	+1.4
FewFC	Chinese financial events	MetaEvent	71.8±0.5	72.7±0.4	73.3±0.4	+1.5
DocFEE	Chinese document events	Text2Event	67.2±0.5	67.6±0.4	68.0±0.4	+0.8
MAVEN	English general events	OneIE	73.0±0.4	73.8±0.3	74.4±0.3	+1.4
ACE2005	English ACE events	DyGIE++	72.7±0.5	73.2±0.4	73.8±0.4	+1.1

Conclusion

This paper presents HyperLEGNN, a label-side hypergraph neural network for low-resource Chinese legal event detection. HyperLEGNN integrates event schemas, support instances, and confusable-label relations to construct structured event-type representations and generate classifier weights for trigger-context classification. On LEVEN, it achieves 53.2 Macro-F1 in the 8-shot setting, 73.3 Macro-F1 under full supervision, 65.7 Tail Macro-F1, and 61.5 Confusable Event F1. The results show consistent improvements over strong graph and hypergraph baselines, with larger gains for sparsely supervised, tail, and semantically confusable event types. Experiments on additional benchmarks further yield Macro-F1 improvements of 0.8–1.5 points, indicating that the proposed label-side modeling remains effective across different event inventories and domains. Future work will therefore focus on extending the framework to document-level event extraction.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)

2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems. pp. 1567–1578 (2019)
3. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Findings of EMNLP 2020. pp. 2898–2904 (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.261
4. Chen, H., Xu, Y., Huang, F., Deng, Z., Huang, W., Wang, S., He, P., Li, Z.: Label-aware graph convolutional networks. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. pp. 1977–1980 (2020). https://doi.org/10.1145/3340531.3412139
5. Chen, Y., Xu, L., Liu, K., Zeng, D., Zhao, J.: Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of ACL-IJCNLP. pp. 167–176 (2015). https://doi.org/10.3115/v1/P15-1017
6. Chen, Y., Zhou, T., Li, S., Zhao, J.: A dataset for document level chinese financial event extraction. Scientific Data **12**, 772 (2025). https://doi.org/10.1038/s41597-025-05083-9
7. Chien, E., Pan, C., Peng, J., Milenkovic, O.: You are AllSet: A multiset function framework for hypergraph neural networks. In: International Conference on Learning Representations (2022)
8. Cong, X., Cui, S., Yu, B., Liu, T., Wang, Y., Wang, B.: Few-shot event detection with prototypical amortized conditional random field. In: Findings of ACL-IJCNLP. pp. 28–40 (2021). https://doi.org/10.18653/v1/2021.findings-acl.3
9. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for chinese natural language processing. In: Findings of EMNLP 2020. pp. 657–668 (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.58
10. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of CVPR. pp. 9268–9277 (2019). https://doi.org/10.1109/CVPR.2019.00949
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
12. Ding, K., Wang, J., Li, J., Li, D., Liu, H.: Be more with less: Hypergraph attention networks for inductive text classification. In: Proceedings of EMNLP. pp. 4927–4936 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.399
13. Feng, Y., Li, C., Ng, V.: Legal judgment prediction via event extraction with constraints. In: Proceedings of ACL. pp. 648–664 (2022). https://doi.org/10.18653/v1/2022.acl-long.48
14. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: Proceedings of AAAI. pp. 3558–3565 (2019). https://doi.org/10.1609/aaai.v33i01.33013558
15. Gao, Y., Feng, Y., Ji, S., Ji, R.: HGNN+: General hypergraph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3181–3199 (2023). https://doi.org/10.1109/TPAMI.2022.3182052
16. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems. pp. 1024–1034 (2017)
17. Hu, L., Yang, T., Shi, C., Ji, H., Li, X.: Heterogeneous graph attention networks for semi-supervised short text classification. In: Proceedings of EMNLP-IJCNLP. pp. 4821–4830 (2019). https://doi.org/10.18653/v1/D19-1488
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017)

19. Li, S., Liu, L., Xie, Y., Ji, H., Han, J.: P4E: Few-shot event detection as prompt-guided identification and localization. arXiv preprint arXiv:2202.07615 (2022)
20. Li, X., Li, F., Pan, L., Chen, Y., Peng, W., Wang, Q., Lyu, Y., Zhu, Y.: DuEE: A large-scale dataset for chinese event extraction in real-world scenarios. In: Natural Language Processing and Chinese Computing. pp. 534–545. Springer (2020). https://doi.org/10.1007/978-3-030-60457-8_44
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proceedings of ICCV. pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324
22. Lin, Y., Ji, H., Huang, F., Wu, L.: A joint neural model for information extraction with global features. In: Proceedings of ACL. pp. 7999–8009 (2020). https://doi.org/10.18653/v1/2020.acl-main.713
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
24. Lu, Y., Lin, H., Xu, J., Han, X., Tang, J., Li, A., Sun, L., Liao, M., Chen, S.: Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In: Proceedings of ACL-IJCNLP. pp. 2795–2806 (2021). https://doi.org/10.18653/v1/2021.acl-long.217
25. Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., Wu, H.: Unified structure generation for universal information extraction. In: Proceedings of ACL. pp. 5755–5772 (2022). https://doi.org/10.18653/v1/2022.acl-long.395
26. Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M., Ma, S.: LeCaRD: A legal case retrieval dataset for chinese law system. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2342–2348 (2021). https://doi.org/10.1145/3404835.3463250
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop (2013)
28. Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference. pp. 593–607 (2018). https://doi.org/10.1007/978-3-319-93417-4_38
29. Sheng, J., Guo, S., Yu, B., Li, Q., Hei, Y., Wang, L., Liu, T., Xu, H.: CasEE: A joint learning framework with cascade decoding for overlapping event extraction. In: Findings of ACL-IJCNLP. pp. 164–174 (2021). https://doi.org/10.18653/v1/2021.findings-acl.14
30. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
31. Su, W., Hu, Y., Xie, A., Ai, Q., Bing, Q., Zheng, N., Liu, Y., Shen, W., Liu, Y.: STARD: A chinese statute retrieval dataset derived from real-life queries by non-professionals. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 10658–10671 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.625
32. Sun, Z., Li, X., Sun, X., Meng, Y., Ao, X., He, Q., Wu, F., Li, J.: ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information. In: Proceedings of ACL-IJCNLP. pp. 2065–2075 (2021). https://doi.org/10.18653/v1/2021.acl-long.161
33. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018)
34. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. pp. 3630–3638 (2016)

35. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, relation, and event extraction with contextualized span representations. In: Proceedings of EMNLP-IJCNLP. pp. 5784–5789 (2019). https://doi.org/10.18653/v1/D19-1585
36. Walker, C., Strassel, S., Medero, J., Maeda, K.: ACE 2005 multilingual training corpus. LDC2006T06 (2006). https://doi.org/10.35111/mwxc-vh88
37. Wang, S., Yu, M., Huang, L.: The art of prompting: Event detection based on type specific prompts. In: Proceedings of ACL. pp. 1286–1299 (2023). https://doi.org/10.18653/v1/2023.acl-short.111
38. Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., Li, J., Li, P., Lin, Y., Zhou, J.: MAVEN: A massive general domain event detection dataset. In: Proceedings of EMNLP. pp. 1652–1671 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.129
39. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems **32**(1), 4–24 (2021). https://doi.org/10.1109/TNNLS.2020.2978386
40. Xiao, C., Hu, X., Liu, Z., Tu, C., Sun, M.: Lawformer: A pre-trained language model for chinese legal long documents. AI Open **2**, 79–84 (2021). https://doi.org/10.1016/j.aiopen.2021.06.003
41. Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., Xu, J.: CAIL2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)
42. Xiao, X.Y., Liu, Y., Liu, X., Li, Z., Yin, E., Xia, Q.: Lunar Twins: We choose to go to the moon with large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 1325–1339. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.findings-acl.69
43. Xiao, X.Y., Tian, Y., Yin, E., He, Z., Wang, S., Liu, Y., Xia, Q.: Lunar-Bench: Towards evaluating task-oriented reasoning of LLMs in lunar exploration scenarios. In: Findings of the Association for Computational Linguistics: ACL 2026. pp. 1668–1705. Association for Computational Linguistics, San Diego, California, United States (2026). https://doi.org/10.18653/v1/2026.findings-acl.83
44. Yadati, N., Nimishakavi, M., Yadav, P., Nitin, V., Louis, A., Talukdar, P.: HyperGCN: A new method for training graph convolutional networks on hypergraphs. In: Advances in Neural Information Processing Systems. pp. 1509–1520 (2019)
45. Yao, F., Xiao, C., Wang, X., Liu, Z., Hou, L., Tu, C., Li, J., Liu, Y., Shen, W., Sun, M.: LEVEN: A large-scale chinese legal event detection dataset. In: Findings of ACL 2022. pp. 183–201 (2022). https://doi.org/10.18653/v1/2022.findings-acl.17
46. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of AAAI. pp. 7370–7377 (2019). https://doi.org/10.1609/aaai.v33i01.33017370
47. Yue, Z., Zeng, H., Lan, M., Ji, H., Wang, D.: Zero- and few-shot event detection via prompt-based meta learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 7928–7943 (2023). https://doi.org/10.18653/v1/2023.acl-long.440
48. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: JEC-QA: A legal-domain question answering dataset. In: Proceedings of AAAI. pp. 9701–9708 (2020). https://doi.org/10.1609/aaai.v34i05.6519
49. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: Advances in Neural Information Processing Systems. vol. 19, pp. 1601–1608 (2006)

50. Zhou, Y., Chen, Y., Zhao, J., Wu, Y., Xu, J., Li, J.: What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14638–14646 (2021). https://doi.org/10.1609/aaai.v35i16.17720