CogMAME: A Cognitive-inspired Meta-learning Adaptive Model for Multi-modal Entity-relation Extraction

Xinyu Li (lixinyu@mail.imu.edu.cn), Yuejia Wu* (cswyj@imu.edu.cn), Jian-tao Zhou* (cszjtao@imu.edu.cn)
College of Computer Science, Inner Mongolia University, Hohhot, China
Engineering Research Center of Ecological Big Data, Ministry of Education
Inner Mongolia Engineering Laboratory for Cloud Computing and Service Software
Inner Mongolia Engineering Laboratory for Big Data Analysis Technology

Abstract

Multi-modal Named Entity Recognition (MNER) and Multi-modal Relation Extraction (MRE) are critical for knowledge extraction. Existing methods often suffer from cross-modal confusion due to semantic misalignment and data imbalance. Inspired by predictive processing theory, i.e., the brain continuously refines cognition through active prediction and error minimization, we propose a **Cog**nitive-inspired **M**eta-learning **A**daptive model for **M**ulti-modal **E**ntity-relation extraction named CogMAME. It consists of five modules: (i) multi-grained representation learning, establishing a unified semantic space; (ii) class-adaptive feature augmentation, reinforcing signals for rare categories; (iii) meta weight hypernet, dynamically calibrating modality contributions; (iv) self-adversarial perturbation, improving stability via controlled noise; (v) multi-modal information extraction, refining outputs in a joint feature space. Extensive experimental results show that our model can achieve human-like extraction capabilities and obtain state-of-the-art performance on both MNER and MRE tasks, verifying its superiority and effectiveness.

Keywords: predictive processing theory; multi-modal knowledge graphs; multi-modal named entity recognition; multi-modal relation extraction; meta-learning.

Introduction

MNER and MRE are essential components of multi-modal knowledge extraction (J. Liu et al., 2026), aiming to identify key entities and infer their semantic relationships from heterogeneous multi-modal data, such as text and images. In social media, user-generated content typically combines multiple modalities. However, the informal language style, limited contextual information, and difficulties in aligning cross-modal semantics often lead to poor performance in MNER and MRE tasks (B. Liu et al., 2026).

Based on this, some studies have incorporated accompanying images as auxiliary signals to aid entity and relation recognition (Zou et al., 2026). However, most of them treated MNER and MRE as separate tasks (Cao et al., 2026), neglecting their inherent interdependence. In recent years, joint modeling of MNER and MRE has emerged as a promising research direction, enabling bidirectional interaction and collaborative reasoning through multi-granular cross-modal feature learning (Yuan et al., 2024). Although those approaches have achieved certain progress, they still exhibit notable limitations when dealing with semantic ambiguities in text-image pairs and imbalanced sample distributions.

To improve those problems, inspired by predictive processing theory, we propose a cognitive-inspired meta-learning adaptive model, namely CogMAME. As shown in Fig. 1, this theory posits that the human brain performs adaptive multi-modal integration through a hierarchical "prediction-error minimization" loop: top-down predictions generated from high-level priors are continuously compared with bottom-up sensory inputs, and the resulting prediction errors drive iterative updates of the internal model, enabling robust perception under uncertainty. Our proposed CogMAME model computationally implements this cognitive mechanism, and its main contributions summarized as follows:

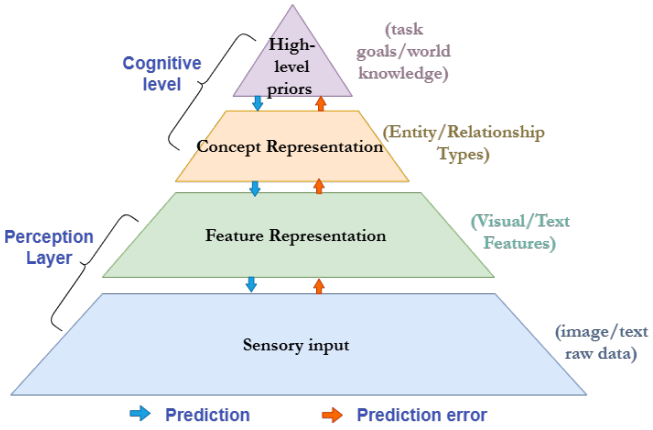

Figure 1: Overview of the predictive processing theory.

- We formulate multi-modal extraction as a hierarchical predictive coding process and propose CogMAME, which implements a meta-learning hypernetwork to dynamically weight cross-modal signals—emulating metacognitive regulation of perceptual priors in human inference.
- We design a class-adaptive feature augmentation module that amplifies prediction error signals for underrepresented classes, thereby countering recognition bias from imbalanced data through error-driven perceptual adjustment.
- We introduce a self-adversarial perturbation strategy that injects controlled noise to simulate internal prediction errors, enhancing model robustness against distribution shifts

* Co-Corresponding authors. These authors contributed equally to the work.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

and strengthening its generalization under uncertainty.

- We conduct extensive experiments on widely used datasets, and the results show that CogMAME achieves state-of-the-art (SOTA) performance on MNER and MRE benchmarks, demonstrating that explicitly modeling predictive processing mechanisms leads to more human-like and adaptive multi-modal integration.

Related Work

As key components of knowledge graph construction, MNER and MRE have attracted increasing attention in recent years.

MNER

Several studies integrated visual and textual features using transformer-based multi-modal modules, but often neglected fine-grained correspondences between images and text, leading to misclassification (J. Yu et al., 2020). Recent approaches adopted fine-grained cross-modal feature learning. For example, MKGformer (X. Chen, Zhang, Li, Deng, et al., 2022) incorporated prefix-guided interaction and fine-grained relevance-aware fusion, and UMGF (D. Zhang et al., 2021) enhanced object-entity alignment by aligning nodes. Moreover, considering the prevalence of weakly or irrelevant text-image pairs in social media, RpBERT (Sun et al., 2021) filtered out irrelevant visual representations by measuring text-image similarity. However, this strategy may discard misaligned yet complementary pairs.

MRE

Zheng et al. (Zheng, Wu, et al., 2021) introduced the first MRE dataset and established the task. Based on this, MEGA(Zheng, Feng, et al., 2021) constructed scene graphs and syntactic parse trees from images and text to capture correspondences, but loses information by treating visual relations as uniform edges. TMR (Zheng et al., 2023) utilized reverse translation in generative diffusion models for knowledge transfer. However, existing methods struggled with imbalanced class distributions and weak text-image pair similarities, leading to inaccurate entity and relation recognition.

In contrast to previous work, inspired by predictive processing theory, we propose a cognitive-inspired meta-learning adaptive approach named CogMAME that consists of five core modules: first, constructing a unified representation space through multi-grained semantic alignment to overcome semantic bias, second, employing a class-adaptive feature augmentation strategy to specifically reinforce signals for rare categories, third, introducing a meta-weight hypernetwork to dynamically calibrate multi-modal contribution weights, fourth, utilizing self-adversarial perturbation to enhance model robustness, and finally, performing multi-modal information joint reasoning and extraction in a unified feature space. This architecture systematically resolves cross-modal confusion.

Methodology

Preliminaries

Definition 1. Cross-modal Representation. Given a text-image pair, it can be represented as $G = G(t, v)$, where T and V denote the tokenized embeddings of t and v, respectively.
Definition 2. Back-Translation. For a given text t and image v, their joint probability is $p(t, v)$. Back-translation treats one modality as a "translation" of the other (H. Yu & Chien, 2025), e.g., translating text into an image modality to generate $v' \sim p(v' \mid t)$. A diffusion model then fuses v and v' for enhanced embedding.

The Proposed CogMAME Model

Overview As mentioned above, inspired by predictive processing theory, we formulate the joint task of multi-modal entity and relation extraction as a hierarchical prediction and error minimization process. As shown in Fig. 2, the proposed CogMAME model consists of five closely modules: (a) multi-grained representation learning (MRL), which extracts local and global cross-modal features to form low-level sensory evidence; (b) class-adaptive feature augmentation (CFA), which actively amplifies prediction error signals for low-frequency classes to enhance sensitivity to weakly observed patterns; (c) meta weight hypernet (MWH), which dynamically calibrates cross-modal contributions via meta-learning, simulating higher-level prior modulation of multi-channel sensory information; (d) self-adversarial perturbation (SP), which generates internal prediction errors through controlled noise injection to improve inference stability under distribution shifts; (e) multi-modal information extraction (MIE), which aligns and refines cross-modal predictions in a unified representation space to ultimately output entity and relation classifications.

Multi-Grained Representation Learning As mentioned above, inspired by the predictive processing theory which posits that cognition is refined through hierarchical error minimization, we improve the cross-modal semantic misalignment by modeling text-image pairs at varying granularities. This mimics the brain's mechanism of integrating fine-grained details with global context. According to the above definition, a given text-image pair can be represented as follows:

$$\mathrm{G} = \sum softmax\left(\frac{\mathrm{W}_d \mathrm{V}\mathrm{T}^{\mathsf{T}}}{\sqrt{d}}\right)\mathrm{T}, \tag{1}$$

where d is the dimension of textual embedding T, and W_d is a learnable cross-modal attention matrix. We consider phrase and region representations as coarse-grained, denoted as T^p and V^r, respectively, and consider word and patch representations as fine-grained, denoted as T^w and V^s, respectively. Feeding pairs (T^w, V^s) and (T^p, V^r) into "Eq.(1)" yields fine-grained G^f and coarse-grained G^c. The multi-granularity representation G is then calculated as follows:

$$\mathrm{G} = \mathrm{G}^f + \mathrm{G}^c. \tag{2}$$

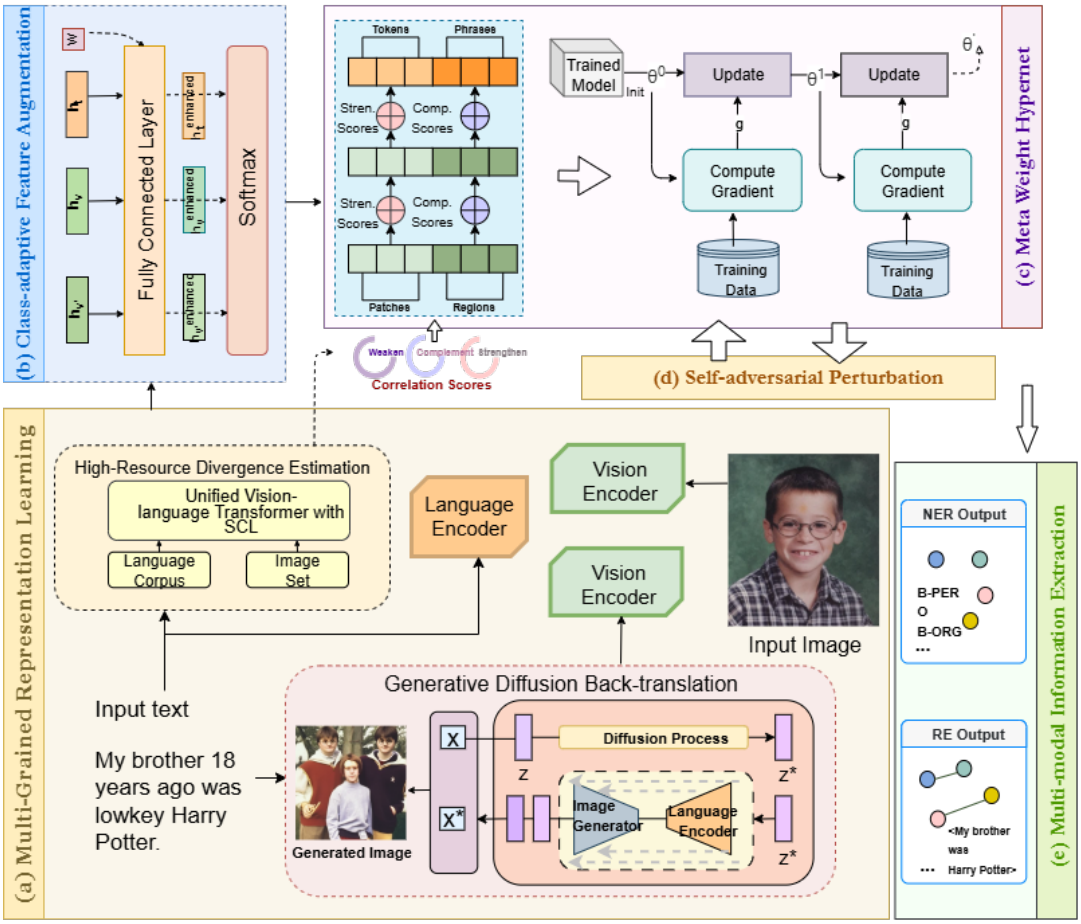

Figure 2: Overview of the proposed CogMAME model

Meanwhile, we adopt a diffusion-based generative model(Saharia et al., 2022) to train on text-image pairs, treating text t as a "translation" of image v, and generate the back-translation v' as:

$$v' = \arg\max p(\hat{v} \mid t), \tag{3}$$

where $\hat{v}$ is the image hypothesis. Consistent with (Rombach et al., 2022), we employ a stable diffusion strategy during back-translation (Zheng et al., 2023), input text as a prompt for stable diffusion to generate v' as an approximation of the back-translation of t, and then input (t, v') into the cross-modal representation yields G'.

To alleviate the issue of similarity scoring methods that ignore poorly aligned but complementary pairs, following previous work (Zheng et al., 2023), we train an independent estimator on a high-resource corpus that generates three confidence scores $(\alpha_s, \alpha_c, \alpha_w)$ for each pair in the set strengthen, complement, and weaken. This enhances the G and G' by ensuring:

$$\begin{bmatrix} \mathrm{G}^* \\ \mathrm{G}'^* \\ 0 \end{bmatrix}^\top = \begin{bmatrix} \alpha_s \\ \alpha_c \\ \alpha_w \end{bmatrix}^\top \begin{bmatrix} \mathrm{G}^f & \mathrm{G}'^f \\ \mathrm{G}^c & \mathrm{G}'^c \\ 0 & 0 \end{bmatrix}, \tag{4}$$

where $\alpha_s + \alpha_c + \alpha_w = 1$. The high-resource corpus consists of 400k training samples from the MSCOCO and LAION-400M datasets, with "strengthen", "complement", and "weaken" sample counts of 100k, 100k, and 200k, respectively. The estimator, based on VILT (Kim et al., 2021), employs supervised contrastive learning (Khosla et al., 2020) to minimize distances between anchors and positive samples from "Strengthen" and "Complement" classes. A self-supervised learning loss can be written as follows:

$$L_{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}, \tag{5}$$

where z is the estimator output, τ is a scalar temperature scalar, and i, j, a denote anchor point, positive and negative samples, respectively. $P(i)$ is the set of indices for positive pairs and $|P(i)|$ is its cardinality.

Class-adaptive Feature Augmentation To mitigate performance degradation caused by data imbalance, inspired by the error-driven learning principle in predictive processing theory, we propose a class-adaptive feature augmentation mechanism that enhances feature representations through two approaches: (i) class-weighted focal loss to prioritize the optimization of

misclassified samples, and (ii) class-aware feature gating to selectively amplify discriminative semantic signals.

(i) Class-weighted Focal Loss. We first assign inverse-frequency-based weights to each class. The weight w_c for class c is calculated as follows:

$$w_c = 1/\log(f_c + \epsilon), \tag{6}$$

where f_c represents the relative frequency of c in the training data, and ϵ is a constant to avoid numerical instability. Based on those weights, the class-weighted focal loss is defined as:

$$\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{N} w_{y_i} \cdot (1 - p_{y_i})^{\gamma} \cdot \log p_{y_i}, \tag{7}$$

where w_{y_i} is the weight for the class of the i-th sample, p_{y_i} is the predicted probability for the true class of the i-th sample, and γ denotes the focusing parameter.

(ii) Class-aware Feature Gating. We design a class-aware gating mechanism that adaptively enhances token-level features based on their class importance. For a given token representation h_i, we define the enhanced feature as follows:

$$h_i^{\text{enhanced}} = h_i + w_{y_i} \cdot \text{ReLU}(Wh_i + b), \tag{8}$$

where W and b are learnable parameters. Min-max normalization constrains $w_{y_i} \in [w_{\min}, w_{\max}]$. This gating mechanism selectively amplifies semantic representations of minority-class tokens while preserving majority-class features.

Meta Weight Hypernet Following the predictive processing principle that cognition dynamically adjusts perceptual weighting based on contextual reliability, we propose a meta-learning-driven hypernetwork to model the reliability of multi-modal correspondences. It dynamically calibrates attention allocation across modalities by generating instance-specific weight parameters, effectively reducing cross-modal confusion caused by inconsistent semantic contributions.

(i) Multi-granularity Multi-head Attention. We adopt a multi-head attention mechanism to capture multi-granularity associations between text-image pairs as follows:

$$\text{Att}_i^{(h)} = \sum_j \text{softmax}_j \left(\frac{\left(\mathbf{W}_Q^{(h)} t_i\right)^{\top} \left(\mathbf{W}_K^{(h)} v_j\right)}{\sqrt{d_h}} \right) \cdot \mathbf{W}_V^{(h)} v_j, \tag{9}$$

where $\mathbf{W}_Q^{(h)}$, $\mathbf{W}_K^{(h)}$, and $\mathbf{W}_V^{(h)}$ are learnable matrices for query, key, and value respectively, and d_h is dimension of the attention head. The final integrated representation is calculated as:

$$\tilde{t}_i = \text{Concat}\left[\text{Att}_i^{(1)}, \ldots, \text{Att}_i^{(H)}\right]. \tag{10}$$

Due to text-image inconsistencies and annotation noise, direct use of attention may overemphasize low-quality regions. We propose a meta-learning dynamic weighting mechanism to adjust the attention loss for each (t_i, v_j) pair.

(ii) Meta-learning Dynamic Weighting. We adopt a meta-learning weight hypernetwork $h_\phi(\cdot)$ and output weights based on the local matching loss ℓ_{ij}^{att} for each text-image pair as:

$$w_{ij} = h_\phi(\ell_{ij}^{\text{att}}) = \sigma\left(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \ell_{ij}^{\text{att}} + b_1) + b_2\right). \tag{11}$$

The final weighted attention loss is calculated as follows:

$$\mathcal{L}_{\text{att}} = \sum_{i,j} w_{ij} \cdot \ell_{ij}^{\text{att}}, \tag{12}$$

where ℓ_{ij}^{att} represents the prediction-supervision discrepancy of the attention for the text-image pair.

(iii) Meta-learning Training Mechanism. To train the weight network h_ϕ, we perform a virtual update on the parameters θ using a training batch, then evaluate its generalization error and optimize ϕ accordingly. The loss is defined as:

$$\mathcal{L}_{\text{train}} = \sum_{i,j} h_\phi(\ell_{ij}^{\text{att}}) \cdot \ell_{ij}^{\text{att}}. \tag{13}$$

The virtual gradient update for the learning rate α defined as follows:

$$\theta' = \theta - \alpha \cdot \nabla_\theta \mathcal{L}_{\text{train}}. \tag{14}$$

The generalization loss is computed on the validation set and backpropagation is used to optimize ϕ as follows:

$$\mathcal{L}_{\text{val}} = \mathcal{L}_{\text{val}}(D_{\text{val}}; \theta'), \quad \phi \leftarrow \phi - \eta \cdot \nabla_\phi \mathcal{L}_{\text{val}}, \tag{15}$$

where η is the inner learning rate. ∇_ϕ denotes the gradient with respect to the hypernetwork parameters ϕ for updating.

Self-adversarial Perturbation Guided by the predictive processing view that noise can drive robust perceptual refinement, we design a self-adversarial perturbation module that introduces controlled, learnable noise into both text and image representations. The perturbation intensity is dynamically regulated by a meta-controller, allowing the model to adaptively sharpen decision boundaries by focusing on hard-to-predict cross-modal samples. Perturbed features are generated by adding perturbation ϵ to text $\mathbf{T}$ and image $\mathbf{V}$ as:

$$\tilde{\mathbf{T}} = \mathbf{T} + \epsilon_t, \tilde{\mathbf{V}} = \mathbf{V} + \epsilon_v, \quad \epsilon_{t,v} = \tanh(g_{\phi_{t,v}}(\mathbf{T}, \mathbf{V})), \tag{16}$$

where $g_{\phi_\cdot}(\cdot)$ is a lightweight network that controls region-level perturbation generation, $\tanh(\cdot)$ constrains their magnitude, $\phi_\cdot$ enables dynamic token-level perturbation generation.

To prevent underfitting or overfitting from inappropriate perturbation scales, we equip a meta perturbation controller that guides the updates of ϕ_t and ϕ_v via an external objective, enabling meta-optimization of perturbation magnitude as:

$$\min_{\phi_t, \phi_v} \mathcal{L}_{\text{val}}\left(f(\mathbf{T} + \tanh(g_{\phi_t}(\mathbf{T})), \mathbf{V} + \tanh(g_{\phi_v}(\mathbf{V}))); y\right), \tag{17}$$

where y is supervision signal, evaluating the current perturbation strategy's impact on model generalization.

Multi-modal Information Extraction Based on the predictive processing framework which integrates refined predictions across modalities, we perform joint information extraction in the adaptively aligned feature space. The augmented representations G^* and G'^*, processed through our

cognitive-inspired modules, are used to execute Named Entity Recognition (NER) and Multi-modal Named Entity Recognition (MNRE) tasks within a unified reasoning paradigm.

(i) Named Entity Recognition. Following (X. Chen, Zhang, Li, Yao, et al., 2022), we implement NER with a CRF decoder. The final representation for (t, v) is obtained by combining G^* and its back-translation G'^* via multi-head extension as follows:

$$\bar{\mathcal{G}} = \text{Multihead}(G^*, G'^*) \in \mathbb{R}^{n \times d}, \tag{18}$$

which consists of n word representations from text t. Let $\mathcal{L} = \{l\}$ denote the NER label set and $Y = [y] \in \mathbb{R}^{n \times |\mathcal{L}|}$ the prediction probabilities, computed as follows:

$$p(y \mid \bar{\mathcal{G}}) = \frac{\prod_{i=1}^{n} F_i(y_{i-1}, y_i, \bar{\mathcal{G}})}{\sum_{l_j \in \mathcal{L}} \prod_{i=1}^{n} F_i(y_{i-1,j}, y_{i,j}, \bar{\mathcal{G}})}, \tag{19}$$

where $y_{i,j}$ is the probability of word i having label j, and F denotes the potential function in the CRF. Training uses maximum conditional likelihood loss.

$$L_{ner} = -\sum_{i=1}^{n} \log(p(y|\bar{\mathcal{G}})). \tag{20}$$

(ii) Relation Extraction. We merge representations of text entities, fine-grained and coarse-grained image-text pairs, and noun phrases to predict the final relationship. For entity pair (e_i, e_j) at positions i, j in t, its representation is generated as:

$$\ddot{\mathcal{G}}_{i,j} = T_i \oplus T_j \oplus \mathbf{p} \oplus \mathbf{h}, \tag{21}$$

where T_i and T_j are entity embeddings, $\oplus$ denotes concatenation, $\mathbf{p}$ the summed noun phrase features, and $\mathbf{h} = \mathcal{G}^* + \mathcal{G}'^*$ the aggregated text-image and its back-translations representation. The relation probability is $p(r|\ddot{\mathcal{G}}_{i,j}) = \text{softmax}(\ddot{\mathcal{G}}_{i,j})$ over labels $\mathcal{R} = \{r\}$. The RE loss is computed via cross-entropy. The RE loss is computed via cross-entropy as:

$$L_{re} = -\sum_{i=1}^{n} \log\left(p(r \mid \ddot{\mathcal{G}}_{i,j})\right). \tag{22}$$

Experiments and Results

Datasets and Metrics

As shown in Table 1, Table 2, and Table 3, we adopt three standard datasets widely used in MNER and MRE: i) Twitter2015 (Lu et al., 2018) and Twitter2017 (Q. Zhang et al., 2018), two MNER datasets consisting of Twitter posts from 2014–2015 and 2016–2017, respectively; ii) MNRE (Zheng, Feng, et al., 2021), a manually annotated dataset for MRE, containing text and images collected from Twitter and subsets of Twitter2015 and Twitter2017. We use precision, recall, and F1 score as the default evaluation metrics.

Baselines

We compare with two kinds of methods: (i) text-based methods, including CNN-BLSTM-CRF (Ma & Hovy, 2016), HBiLSTM-CRF (Lample et al., 2016), BERT-CRF (Devlin et al., 2019), PCNN (D. Zeng et al., 2015), MTB (Soares et al., 2019); and (ii) multi-modal methods, including OCSGA (Wu et al., 2020), RpBERT (Sun et al., 2021), UMGF (D. Zhang et al., 2021), MEGA (Zheng, Feng, et al., 2021), HVPNet (X. Chen, Zhang, Li, Yao, et al., 2022), MKGFormer (X. Chen, Zhang, Li, Deng, et al., 2022), TRFL (Q. Zeng et al., 2025), CASF (H. Li et al., 2025), LIBD (Xu et al., 2025), RAP (Z. Chen et al., 2025) etc.

Table 1: Statistics of the Twitter2015 Dataset.

Category	Train	Dev	Test	Total
Person	2217	552	1816	4583
Location	2091	522	1697	4308
Organization	928	247	839	2012
Misc	940	225	726	1881
Total Entity	6176	1546	5078	12784

Table 2: Statistics of the Twitter2017 Dataset.

Category	Train	Dev	Test	Total
Person	2943	626	621	4190
Location	731	173	178	1082
Organization	1674	375	395	2444
Misc	701	150	157	1008
Total Entity	6049	1324	1351	8724

Table 3: The Statistics of MNRE Dataset.

Statistics	Word	Sent.	Ins.	Ent.	Rel.	Img.
MNRE	258k	9201	15,485	30,970	23	9201

Experimental setup and implementation

For the MNER task, we set the batch size to 12, the learning rate to 3e-5, and train for 30 epochs, with validation starting at epoch 12. The warmup ratio is 0.01, and both the text and image hidden sizes are set to 1024, with an intermediate size of 400. The number of hidden layers is 1, and the output dimension of the meta-weight network is 256. For the MRE task, we set the batch size to 16, the learning rate to 1e-5, and train for 20 epochs. Both the text and image hidden sizes are set to 768, and the number of hidden layers is 1. The output dimension of the meta-weight network is 256.

Overall results

As shown in Table 4, our proposed model outperforms all other SOTA methods across all datasets. Compared to single-text models, the multi-modal model improves performance by 6% on MNER and 30% on MRE. This is due to the short and ambiguous nature of social media texts, making it difficult to identify entities and their relationships within limited context. Compared to multi-granularity multi-modal models, our model improves the F1 scores on the Twitter15/Twitter17/MNRE datasets from

Table 4: Compare CogMAME's overall performance with baselines on three benchmark datasets for MNER and MRE.

Modality	Methods	*Twitter2015*			*Twitter2017*			*MNER*		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Text	CNN-BLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37	-	-	-
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37	-	-	-
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44	-	-	-
	PCNN	-	-	-	-	-	-	62.85	49.69	55.49
	MTB	-	-	-	-	-	-	64.46	57.81	60.86
Text+Image	AdapCoAtt	69.87	74.59	72.15	85.13	83.20	84.10	-	-	-
	OCSGA	74.71	71.21	72.92	-	-	-	-	-	-
	RpBERT	71.15	74.30	72.69	-	-	-	-	-	-
	UMT	71.67	75.23	73.41	85.28	85.34	85.31	62.93	63.88	63.46
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51	64.38	66.23	65.29
	VisualBERT	68.84	71.39	70.09	84.06	85.39	84.72	57.15	59.48	58.30
	MEGA	70.35	74.58	72.35	84.03	84.75	84.39	64.51	68.44	66.41
	HVPNeT	73.87	76.82	75.32	85.84	87.93	86.87	83.64	80.78	81.85
	MKGFormer	-	-	-	86.98	88.01	87.49	82.67	81.25	81.95
	TMR	75.26	76.49	75.87	88.12	88.38	88.25	90.48	87.66	89.05
	TRFL	73.61	77.12	75.32	85.12	88.16	86.65	-	-	-
	CASF	73.12	75.23	74.16	87.16	86.45	86.81	-	-	-
	LIBD	75.99	77.07	76.53	87.68	88.40	88.04	-	-	-
	RAP	74.38	78.92	76.58	87.63	87.89	87.76	83.59	82.67	83.13
	CogMAME w/o CFA	74.97	77.03	75.98	88.16	88.41	88.28	90.66	89.26	89.96
	CogMAME w/o MWH+SP	74.75	77.20	75.95	88.30	88.23	88.26	90.42	88.46	89.42
	CogMAME (Ours)	**75.32**	**77.76**	**76.53**	**88.36**	**88.75**	**88.55**	**91.14**	**89.58**	**90.35**

(75.87/88.25/89.05) to (76.33/88.55/89.75), demonstrating the effectiveness of our meta-learning dynamic weighting and category-adaptive modules. We also observe more significant performance improvements in the MRE task, as the MRE dataset contains a higher proportion of complementary pairs, which can be better handled by the multi-granularity fusion module in CogMAME to reduce prediction errors.

Ablation study

To evaluate the effectiveness of the CFA, WMH, and SP in CogMAME, we conduct ablation studies in Table 4 and Fig. 3. Removing CFA and SP and replacing WMH with standard self-attention causes performance drops on both MNER and MRE tasks, confirming the crucial role of each component. Notably, the ablated models still match or surpass baselines, demonstrating the validity and superiority of our design.

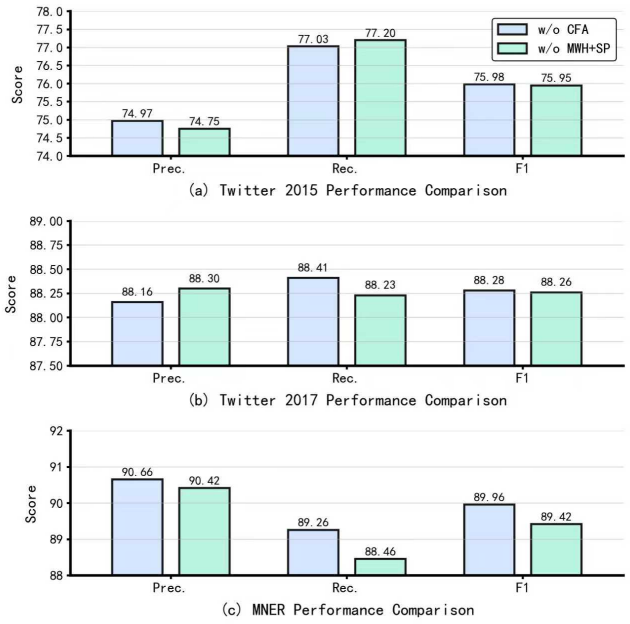

Figure 3: Ablation analysis of the CogMAME model

Conclusion

In this paper, inspired by predictive processing theory, we propose CogMAME, a cognitive-inspired meta-learning adaptive model for multi-modal entity-relation extraction with five components: (i) a multi-grained representation learning module for unified cross-modal semantic space; (ii) a class-adaptive feature augmentation module for low-frequency categories; (iii) a meta weight hypernet module for dynamic cross-modal weight adjustment; (iv) a self-adversarial perturbation module for robustness under noise and distribution shifts; and (v) a multi-modal information extraction module for joint entity and relation inference. Extensive experiments show CogMAME achieves outstanding performance in complex multi-modal scenarios. Future work will explore lightweight cognitive computing frameworks to reduce costs while maintaining performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62472236 and 62162046, the Inner Mongolia Natural Science Foundation under Grant 2019GG372 and 2025QN06009, the Inner Mongolia Science and Technology Project under Grant 2021GG0155, the Natural Science Foundation of Major Research Plan of Inner Mongolia under Grant 2019ZD15, the Inner Mongolia University Youth Academic Talent Research Launch Project under Grant 10000-A24206010, the fund of Supporting the Reform and Development of Local Universities (Disciplinary Construction) and the special research project of First-class Discipline of Inner Mongolia A. R. of China under Grant YLXKZX-ND-036, the Science and Technology Project of the Inner Mongolia Public Hospital Research Joint Fund under Grant 2025GLLH0003.

References

Cao, L., Jia, Z., Zhang, B., Bian, H., Zhang, F., Du, K., & Guo, Y. (2026). DEG-NER: dynamic graph-enhanced diffusion model for named entity recognition. *Expert Syst. Appl.*, *299*, 130094. Retrieved from https://doi.org/10.1016/j.eswa.2025.130094 doi: 10.1016/J.ESWA.2025.130094

Chen, X., Zhang, N., Li, L., Deng, S., Tan, C., Xu, C., ... Chen, H. (2022). Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR '22: The 45th international ACM SIGIR conference on research and development in information retrieval, madrid, spain, july 11 - 15, 2022* (pp. 904–915). ACM. Retrieved from https://doi.org/10.1145/3477495.3531992 doi: 10.1145/3477495.3531992

Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., ... Chen, H. (2022). Good visual guidance makes A better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *CoRR*, *abs/2205.03521*. Retrieved from https://doi.org/10.48550/arXiv.2205.03521 doi: 10.48550/ARXIV.2205.03521

Chen, Z., Li, Z., Liu, M., Zhang, C., & Ma, H. (2025). Relevance-aware prompt-tuning method for multimodal social entity and relation extraction. *Neurocomputing*, *640*, 130316. Retrieved from https://doi.org/10.1016/j.neucom.2025.130316

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/n19-1423 doi: 10.18653/V1/N19-1423

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, las vegas, nv, usa, june 27-30, 2016* (pp. 770–778). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2016.90 doi: 10.1109/CVPR.2016.90

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... Krishnan, D. (2020). Supervised contrastive learning. In *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual.*

Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th international conference on machine learning, ICML 2021, 18-24 july 2021, virtual event* (Vol. 139, pp. 5583–5594). PMLR.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *NAACL HLT 2016, the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies, san diego california, usa, june 12-17, 2016* (pp. 260–270). The Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/n16-1030 doi: 10.18653/V1/N16-1030

Li, H., Tao, Y., Wang, H., Wang, Z., & Liu, Q. (2025). CASF: correlation-alignment and significance-aware fusion for multimodal named entity recognition. *Algorithms*, *18*(8), 511. Retrieved from https://doi.org/10.3390/a18080511 doi: 10.3390/A18080511

Li, Z., Fan, Z., Tou, H., Chen, J., Wei, Z., & Huang, X. (2022). MVPTR: multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *MM '22: The 30th ACM international conference on multimedia, lisboa, portugal, october 10 - 14, 2022* (pp. 4395–4405). ACM. Retrieved from https://doi.org/10.1145/3503161.3548341 doi: 10.1145/3503161.3548341

Liu, B., Zhang, Y., & Li, J. (2026). Integrated attention using entity type constraints and relation augmentation for distantly supervised relation extraction. *Knowl. Based Syst.*, *333*, 115006. Retrieved from https://doi.org/10.1016/j.knosys.2025.115006 doi: 10.1016/J.KNOSYS.2025.115006

Liu, J., Zhong, H., Xu, M., Wu, B., Song, L., Li, Y., ... Kou, F. (2026). Dual-perspective hypergraph learning network for multimodal entity and relation extraction. *Expert Syst. Appl.*, *300*, 130290. Retrieved from https://doi.org/10.1016/j.eswa.2025.130290 doi: 10.1016/J.ESWA.2025.130290

Lu, D., Neves, L., Carvalho, V., Zhang, N., & Ji, H. (2018). Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, melbourne, australia, july 15-20, 2018, volume 1: Long papers* (pp. 1990–1999). Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1185/ doi: 10.18653/V1/P18-1185

Ma, X., & Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th annual meeting of the association for computational linguistics, ACL 2016, august 7-12, 2016, berlin, germany, volume 1: Long papers*. The Association for Computer Linguistics. Retrieved from https://doi.org/10.18653/v1/p16-1101

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, la, usa, june 18-24, 2022* (pp. 10674–10685). IEEE. Retrieved from https://doi.org/10.1109/CVPR52688.2022.01042 doi: 10.1109/CVPR52688.2022.01042

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in neural information processing systems 35: Annual conference on neural information processing systems 2022, neurips 2022, new orleans, la, usa, november 28 - december 9, 2022*.

Soares, L. B., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, florence, italy, july 28- august 2, 2019, volume 1: Long papers* (pp. 2895–2905). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/p19-1279 doi: 10.18653/V1/P19-1279

Sun, L., Wang, J., Zhang, K., Su, Y., & Weng, F. (2021). Rpbert: A text-image relation propagation-based BERT model for multimodal NER. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, february 2-9, 2021* (pp. 13860–13868). AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v35i15.17633 doi: 10.1609/AAAI.V35I15.17633

Wang, X., Ye, J., Li, Z., Tian, J., Jiang, Y., Yan, M., ... Xiao, Y. (2022). CAT-MNER: multimodal named entity recognition with knowledge-refined cross-modal attention. In *IEEE international conference on multimedia and expo, ICME 2022, taipei, taiwan, july 18-22, 2022* (pp. 1–6). IEEE. Retrieved from https://doi.org/10.1109/ICME52920.2022.9859972 doi: 10.1109/ICME52920.2022.9859972

Wu, Z., Zheng, C., Cai, Y., Chen, J., Leung, H., & Li, Q. (2020). Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *MM '20: The 28th ACM international conference on multimedia, virtual event / seattle, wa, usa, october 12-16, 2020* (pp. 1038–1046). ACM. Retrieved from https://doi.org/10.1145/3394171.3413650 doi: 10.1145/3394171.3413650

Xu, B., Huang, S., Sha, C., & Wang, H. (2022). MAF: A general matching and alignment framework for multimodal named entity recognition. In *WSDM '22: The fifteenth ACM international conference on web search and data mining, virtual event / tempe, az, usa, february 21 - 25, 2022* (pp. 1215–1223). ACM. Retrieved from https://doi.org/10.1145/3488560.3498475 doi: 10.1145/3488560.3498475

Xu, B., Wei, J., Wang, H., Du, M., Song, H., & Xiao, Y. (2025). Bridging the unseen gap: Label-enhanced information bottleneck distillation for multimodal named entity recognition. In *Proceedings of the 33rd ACM international conference on multimedia, MM 2025, dublin, ireland, october 27-31, 2025* (pp. 1500–1509). ACM. Retrieved from https://doi.org/10.1145/3746027.3755210 doi: 10.1145/3746027.3755210

Yu, H., & Chien, J. (2025). Attention disentanglement for semantic diffusion modeling in text-to-image generation. In *2025 IEEE international conference on acoustics, speech and signal processing, ICASSP 2025, hyderabad, india, april 6-11, 2025* (pp. 1–5). IEEE. Retrieved from https://doi.org/10.1109/ICASSP49660.2025.10888688 doi: 10.1109/ICASSP49660.2025.10888688

Yu, J., Jiang, J., Yang, L., & Xia, R. (2020). Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, online, july 5-10, 2020* (pp. 3342–3352). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.acl-main.306 doi: 10.18653/V1/2020.ACL-MAIN.306

Yuan, L., Cai, Y., & Huang, J. (2024). Few-shot joint multimodal entity-relation extraction via knowledge-enhanced cross-modal prompt model. In *Proceedings of the 32nd ACM international conference on multimedia, MM 2024, melbourne, vic, australia, 28 october 2024 - 1 november 2024* (pp. 8701–8710). ACM. Retrieved from https://doi.org/10.1145/3664647.3680717 doi: 10.1145/3664647.3680717

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing, EMNLP 2015, lisbon, portugal, september 17-21, 2015* (pp. 1753–1762). The Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/d15-1203 doi: 10.18653/V1/D15-1203

Zeng, Q., Yuan, M., Su, Y., Mi, J., Che, Q., & Wan, J. (2025). Improving multimodal named entity recognition via text-image relevance prediction with large language models. *Neurocomputing*, *651*, 130982. Retrieved from https://

doi.org/10.1016/j.neucom.2025.130982 doi: 10.1016/J.NEUCOM.2025.130982

Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., & Zhou, G. (2021). Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, february 2-9, 2021* (pp. 14347–14355). AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v35i16.17687 doi: 10.1609/AAAI.V35I16.17687

Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the thirty-second AAAI conference on artificial intelligence, (aaai-18), the 30th innovative applications of artificial intelligence (iaai-18), and the 8th AAAI symposium on educational advances in artificial intelligence (eaai-18), new orleans, louisiana, usa, february 2-7, 2018* (pp. 5674–5681). AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v32i1.11962 doi: 10.1609/AAAI.V32I1.11962

Zheng, C., Feng, J., Cai, Y., Wei, X., & Li, Q. (2023). Rethinking multimodal entity and relation extraction from a translation point of view. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2023, toronto, canada, july 9-14, 2023* (pp. 6810–6824). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2023.acl-long.376 doi: 10.18653/V1/2023.ACL-LONG.376

Zheng, C., Feng, J., Fu, Z., Cai, Y., Li, Q., & Wang, T. (2021). Multimodal relation extraction with efficient graph alignment. In *MM '21: ACM multimedia conference, virtual event, china, october 20 - 24, 2021* (pp. 5298–5306). ACM. Retrieved from https://doi.org/10.1145/3474085.3476968 doi: 10.1145/3474085.3476968

Zheng, C., Wu, Z., Feng, J., Fu, Z., & Cai, Y. (2021). MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE international conference on multimedia and expo, ICME 2021, shenzhen, china, july 5-9, 2021* (pp. 1–6). IEEE. Retrieved from https://doi.org/10.1109/ICME51207.2021.9428274 doi: 10.1109/ICME51207.2021.9428274

Zhong, Z., & Chen, D. (2021). A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2021, online, june 6-11, 2021* (pp. 50–61). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2021.naacl-main.5 doi: 10.18653/V1/2021.NAACL-MAIN.5

Zou, H., Wang, Y., & Huang, A. (2026). Large language model augmented joint learning framework for entity-relation extraction. *Appl. Soft Comput.*, *186*, 114094. Retrieved from https://doi.org/10.1016/j.asoc.2025.114094 doi: 10.1016/J.ASOC.2025.114094