Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SA-RGLP: Syntax-Aware Residual Graph Label Propagation with LLM Soft Priors for Metaphor Detection

Xiangyi Jia
Xinjiang University
Wenzhong Yang
yangwenzhong@xju.edu.cn
Xinjiang University

Research Article

Keywords: Metaphor Detection, Graph Label Propagation, Syntax-Aware Modeling, Large Language Models, Global Semantic Graph

Posted Date: August 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10641869/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SA-RGLP: Syntax-Aware Residual Graph Label Propagation with LLM Soft Priors for Metaphor Detection

Abstract

Metaphor detection aims to identify semantic mappings across conceptual domains in text and is fundamental to deep semantic understanding in natural language processing. Previous studies typically treat each target-word usage as an isolated instance and model lexical, syntactic, and conceptual-domain incongruity only within individual sentences. Consequently, they struggle to organize dispersed high-dimensional instance representations into a global semantic structure across samples, limiting their ability to capture systematic conceptual mappings. Moreover, existing methods rely mainly on contextual semantic evidence without sufficiently integrating class-level regularities revealed by labeled examples with open-domain conceptual priors provided by large language models, hindering the fine-grained discrimination of complex metaphorical usages with ambiguous semantic boundaries. To address these limitations, we propose a Syntax-Aware Residual Graph Label Propagation (SA-RGLP) framework with LLM-derived soft priors for metaphor detection. The framework first learns instance representations and initial predictions through syntax-aware conceptual-domain incongruity modeling. It then constructs a global semantic graph over the instances, incorporates continuous soft priors generated by an large language models (LLM), and jointly models local semantics, class-level regularities, and open-domain conceptual knowledge through residual graph label propagation (RGLP). In this way, SA-RGLP extends metaphor detection from isolated instance classification to cross-sample relational reasoning. Experiments on MOH-X, TroFi, and VUA_Verb show that SA-RGLP achieves competitive overall performance and state-of-the-art results on TroFi and VUA_Verb.

Keywords: Metaphor Detection, Graph Label Propagation, Syntax-Aware Modeling, Large Language Models, Global Semantic Graph

1 Introduction

Metaphor is a pervasive linguistic phenomenon and a fundamental cognitive mechanism through which abstract concepts are understood in terms of more concrete conceptual domains. Automatic metaphor detection determines whether a target word

expresses a semantic mapping across conceptual domains in a given context. It supports deep natural language understanding and downstream tasks such as machine translation [1], sentiment analysis [2–4], information extraction [5, 6], and dialogue understanding [7].

Early approaches relied on handcrafted features derived from lexical resources, abstractness, selectional preferences, and syntactic patterns. Although interpretable, they depended heavily on expert knowledge and external resources. With the development of pretrained language models, methods such as MelBERT [8], MrBERT [9], and MisNet [10] improved local semantic modeling [11, 12] by comparing the basic and contextual meanings of target words and capturing their syntactic and semantic interactions. More recent methods, including BasicBERT [13] and WPDM [14], further incorporate linguistic theories such as MIP [15–17], SPV [18, 19], and CMT [20] through basic-meaning comparison, word-pair interaction, and conceptual-domain mining.

Despite this progress, existing methods still largely follow an isolated instance-level inference paradigm [14], in which each target-word usage is encoded and classified independently. Metaphor, however, is a systematic conceptual mapping phenomenon rather than a random semantic anomaly. Usages with similar target words, argument structures, or source–target domain combinations often exhibit recurring metaphorical patterns. Relying only on local context makes it difficult to organize dispersed instance-level representations into a global semantic topology or exploit related usages as complementary evidence for ambiguous cases. Consequently, existing methods struggle to capture systematic conceptual mappings across samples and to leverage cross-sample relational evidence for disambiguating metaphorical usages.

Meanwhile, LLMs [21–24] encode extensive knowledge of word senses, commonsense, and conceptual relations from open-domain corpora, providing useful conceptual priors for metaphor understanding. Existing studies have prompted LLMs with linguistic procedures such as MIPVU [16] for metaphor identification and explanation. However, these approaches typically convert LLM outputs into discrete predictions or hard pseudo-labels, discarding the relative confidence between literal and metaphorical interpretations and limiting their integration with local semantics and class-level regularities learned from task data. As a result, LLM-derived knowledge remains largely isolated from task-specific representations and cross-sample semantic structures, failing to provide soft guidance that preserves the inherent gradience of metaphorical language.

These limitations share a common root: existing paradigms treat metaphor detection as independent per-instance classification and lack a unified framework that jointly models local semantics, external conceptual knowledge, and cross-sample semantic relations. Such integration is essential for leveraging complementary evidence across instances and for preserving the semantic gradience that characterizes metaphorical language. However, no prior work has successfully combined these complementary information sources in a coherent manner for target-word-level metaphor detection.

GLPN-LLM [25] demonstrates that LLM judgments can be organized together with instance representations and human annotations in a graph and used as relational

evidence for global label propagation. Nevertheless, directly transferring this paradigm to target-word-level metaphor detection is challenging. Graph nodes must represent specific target-word usages rather than entire sentences, metaphor gradience is better represented by continuous soft priors than hard pseudo-labels, and global propagation should refine strong local predictions rather than replace them.

Inspired by the above observations, we propose SA-RGLP, a metaphor detection framework that integrates LLM soft priors. In the first stage, SA-RGLP jointly models sentence context [26], target-word semantics [14], conceptual-domain-incongruent word pairs [27], and syntactic category information [28, 29] to produce task-specific instance-level representations and initial predictions. In the second stage, these representations are organized into a global semantic instance graph over target-word usages, while structured adversarial prompting generates continuous LLM soft priors. RGLP then integrates local semantic representations, partially visible labels, and open-domain conceptual priors to incrementally refine local predictions through cross-sample relations.

In summary, the main contributions of this work are as follows:

(1) We propose SA-RGLP, a two-stage metaphor detection framework that integrates syntax-aware conceptual-domain incongruity modeling with RGLP. It extends metaphor detection from local semantic judgments over isolated instances to local–global joint reasoning that incorporates cross-sample semantic relations.

(2) We introduce, for the first time in metaphor detection, an LLM-assisted global label propagation paradigm and adapt it to target-word-level metaphor detection using continuous soft priors instead of hard pseudo-labels.

(3) Extensive experiments on three widely used benchmark datasets, MOH-X, TroFi, and VUA-Verb, demonstrate that SA-RGLP achieves competitive performance across all three datasets and state-of-the-art results on TroFi and VUA-Verb.

2 Related Work

2.1 Large Language Models for Metaphor Detection

The development of large language models (LLMs) has introduced a new source of semantic and conceptual knowledge for metaphor detection. Unlike conventional supervised detectors that primarily learn decision boundaries from task-specific annotations, LLMs acquire broad knowledge of lexical meanings, commonsense relations, and conceptual associations from large-scale open-domain corpora, enabling them to explicitly reason about the basic meaning, contextual meaning, and potential conceptual mappings of a target word.

Early studies mainly investigated whether established linguistic procedures for metaphor identification could be directly implemented through LLM prompting. Reimann and Scheffler [30] reformulated MIPVU as a structured prompting procedure that guides an LLM to identify the contextual meaning of a target word, determine its basic meaning, and compare the two before producing a metaphor judgment. This approach demonstrates that linguistically motivated identification procedures can be partially transformed into natural-language reasoning steps performed by an LLM.

However, direct zero-shot or few-shot prediction remains sensitive to the knowledge encoded by the LLM and to the way the reasoning process is elicited.

Subsequent studies therefore began to introduce more explicit linguistic and conceptual knowledge into LLM-based metaphor reasoning. Yang et al. [31] use ChatGPT to generate literal collocations for a target verb and combine them with subject–object information from the current context, mapping these semantic units into a shared topic space to determine whether the verb exhibits cross-domain usage. TSI [32] further organizes metaphor theories, including SPV, MIP, and CMT, into a metaphor knowledge graph and derives a sequence of scaffolding questions that guide the LLM through an incremental reasoning process. When a reasoning step exceeds the estimated capability of the LLM, relevant metaphor knowledge is supplied as additional scaffolding support, making the metaphor judgment more explicitly grounded in linguistic theory. DMD [33] extends theory-guided LLM reasoning from both implicit and explicit perspectives. Its implicit component retrieves annotated examples according to theory-related representations, while its explicit component combines dictionary knowledge with metaphor-theory-guided multi-step reasoning; the resulting responses are subsequently evaluated through a self-judgment mechanism to obtain a more reliable final prediction.

Recent work has also considered the semantic gradience of metaphorical interpretation. The boundaries among literal usage, conventional polysemous extension, and cross-domain metaphor are not always categorical, and reducing an LLM judgment to a single binary label may discard useful information about competing interpretations. Prompting Metaphoricity [34] therefore represents metaphoricity through continuous LLM-generated probabilities, preserving the model's uncertainty between literal and metaphorical readings.

Overall, LLM-based metaphor detection has progressed from direct prompting toward theory-guided reasoning, external knowledge augmentation, structured multi-step inference, and graded metaphor judgments. These developments demonstrate the potential of LLMs to provide open-domain conceptual evidence that is difficult to obtain solely from task-specific annotations. Nevertheless, most existing approaches still process each target-word instance independently and use the LLM output directly as a prediction or reasoning result. Consequently, LLM-derived knowledge remains largely separated from the task-specific representations learned by supervised metaphor detectors and from semantic regularities distributed across different samples. Moreover, because LLM judgments can be affected by prompting strategies, knowledge coverage, and semantic ambiguity, directly treating them as final labels may introduce unreliable supervision.

Inspired by these studies, SA-RGLP repositions the LLM from a stand-alone metaphor classifier to an auxiliary source of conceptual knowledge. Structured adversarial prompting encourages the LLM to consider competing cross-domain metaphorical and literal polysemous interpretations, and the resulting continuous probability is retained as a soft prior rather than converted into a hard decision. This formulation preserves both open-domain conceptual knowledge and uncertainty information while providing a more suitable signal for subsequent integration with task-specific local representations and cross-sample relational evidence.

2.2 Graph Label Propagation and Relational Reasoning

Graph Label Propagation (GLP) provides a basic mechanism for exploiting relations between labeled and unlabeled samples beyond independent instance-level classification. Zhu and Ghahramani [35] introduced the classical Label Propagation (LP) method, which iteratively propagates observed labels over a data graph under the smoothness assumption that neighboring nodes tend to share similar labels. Zhou et al. [36] subsequently proposed Local and Global Consistency (LGC), which further balances local neighborhood smoothness with global consistency to the initial supervision. In natural language processing, Niu et al. [37] applied label propagation to word sense disambiguation, allowing sense information to diffuse among semantically related usages and demonstrating the value of cross-sample relations for context-dependent semantic classification.

With the development of deep representation learning and graph neural networks, label propagation has increasingly been combined with task-specific features and model predictions. Iscen et al. [38] proposed a deep semi-supervised framework that constructs a nearest-neighbor graph from learned embeddings and uses propagated labels as pseudo-supervision for unlabeled samples. Wang and Leskovec [39] proposed GCN-LPA, which unifies Graph Convolutional Networks and Label Propagation by relating feature smoothing to label smoothing and uses label information to learn more task-oriented graph connections. Huang et al. [40] further proposed Correct and Smooth (C&S), which first obtains predictions from a base classifier and then propagates correlated prediction errors and smooths the corrected outputs. These methods show that graph propagation can complement rather than replace informative instance-level predictions.

Recent studies have further developed task-specific label propagation frameworks. Zhao et al. [41] proposed FCN-LP for multimodal fake news detection, constructing a cross-modal tweet graph from text–image representations and exploiting relations among news instances to refine predictions. Hu et al. [25] proposed GLPN-LLM, which integrates task representations, human annotations, and LLM-generated pseudo-labels into a global graph and employs mask-based label propagation to utilize cross-sample relations while reducing label leakage. More recently, Chen et al. [42] proposed TR-GLP, where Residual Graph Label Propagation incorporates global cross-video context while preserving the discriminative information learned by the instance-level detector. These developments extend conventional label diffusion toward relational reasoning that coordinates local predictions, external supervision, and global sample structure.

This paradigm is highly relevant to metaphor detection but cannot be directly transferred to the target-word-level setting. Metaphoricity is associated with a particular target-word usage in context rather than an entire sentence or document, requiring graph nodes and similarities to be defined at the sentence–target-word usage level. Moreover, reliable propagation requires task-specific representations capable of distinguishing contextual, syntactic, and conceptual differences among metaphor instances. Existing LLM-assisted propagation methods also mainly rely on discrete pseudo-labels, whereas metaphorical interpretation is semantically graded, and strong supervised metaphor detectors already provide informative local predictions that should be refined rather than discarded.

Motivated by these observations, SA-RGLP adapts relational label propagation to metaphor detection by constructing a graph over sentence–target-word usages with task-specific representations, incorporating continuous LLM probabilities as soft priors, and using residual graph label propagation to learn incremental corrections to the Stage-1 local logits. This design jointly exploits local metaphor evidence, LLM-derived conceptual knowledge, and global cross-sample relations.

3 Methodology

3.1 Problem Formulation

Formally, let $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{te}}$ denote the complete metaphor detection dataset, where $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{te}}$ represent the training, validation, and test sets, respectively.

Each metaphor detection instance is represented as a triplet (S, w_t, y), where

$$S = (w_1, w_2, \ldots, w_n) \tag{1}$$

denotes a contextual sentence consisting of n tokens, $w_t \in S$ denotes the target word whose metaphoricity is to be determined, and $y \in \{0, 1\}$ denotes the corresponding ground-truth label. Specifically, $y = 1$ indicates that the target word is used metaphorically in the given context, whereas $y = 0$ indicates a literal usage.

Accordingly, the objective of target-word-level metaphor detection is to learn a mapping function

$$F : (S, w_t) \rightarrow y, \tag{2}$$

which predicts the metaphorical label of the target word according to its usage in the contextual sentence.

In the proposed SA-RGLP framework, the mapping function F is instantiated as a two-stage graph-based model. The first stage learns syntax-aware contextual representations and initial instance-level predictions by modeling the semantics of the target word, its sentence-level context, and conceptual-domain incongruity. The second stage constructs a global semantic graph over metaphor instances and jointly exploits cross-sample semantic relations and LLM-derived soft priors to refine the initial predictions. Therefore, the overall mapping can be further expressed as

$$F_{\text{SA-RGLP}}\left(S, w_t, \mathcal{G}, p^{\text{llm}}\right) \rightarrow \hat{y} \tag{3}$$

where $\mathcal{G}$ denotes the global semantic graph constructed over metaphor instances, $p^{\text{llm}} \in [0, 1]$ denotes the LLM-derived soft prior for the current instance, and $\hat{y} \in \{0, 1\}$ denotes the final predicted label.

3.2 Framework Overview

We propose SA-RGLP, a metaphor detection framework that integrates LLM soft priors, as illustrated in Figure 1. SA-RGLP adopts a two-stage learning paradigm following the reasoning path of local semantic modeling [11, 12], open-domain prior augmentation [43], and global relational refinement [44]. In addition to capturing

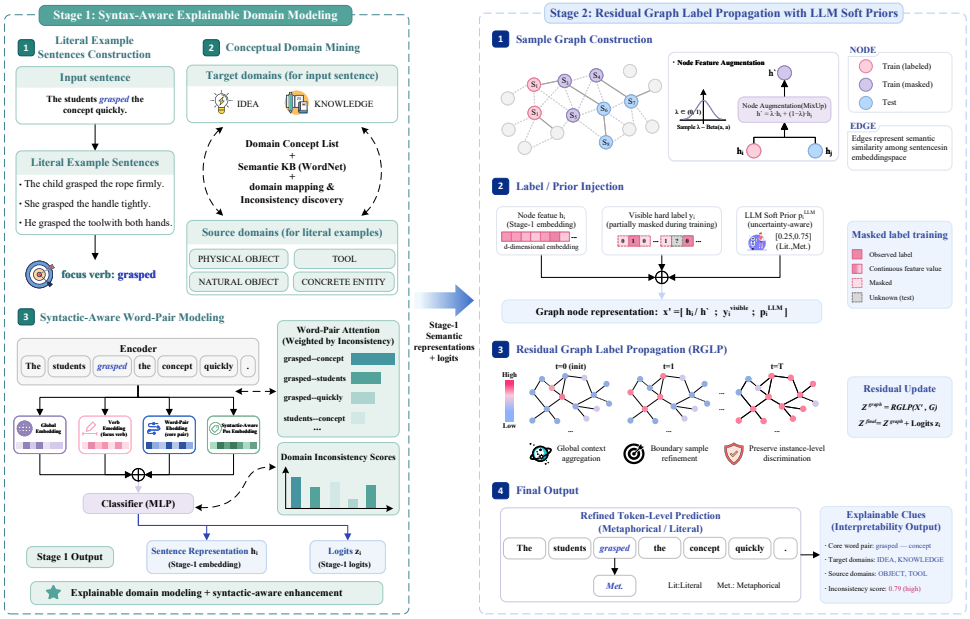


Fig. 1 Overall architecture of SA-RGLP. Stage 1 learns task-specific instance representations and initial metaphor predictions through literal-example construction, conceptual-domain mining, syntax-aware word-pair modeling, and local semantic fusion. Stage 2 generates continuous LLM soft priors, augments the graph nodes in the feature space, constructs a global semantic graph over target-word usages, and performs residual GLP. The graph-derived correction is added to the local logits to refine target-word-level metaphor predictions using cross-sample semantic relations.

semantic and syntactic cues within individual sentences [45], it explicitly models cross-sample relations among different target-word usages by constructing a global semantic instance graph. Specifically, the first stage operates on individual sentence–target-word instances and jointly models the global sentence context, contextualized target-word semantics, conceptual-domain-incongruent word pairs, and the syntactic category of the target word, producing high-dimensional instance-level representations and initial metaphor predictions. In the second stage, structured adversarial prompting guides the LLM to compare cross-domain metaphorical and literal polysemous interpretations and generate continuous soft priors. A global semantic instance graph is then constructed from the Stage-1 representations. The local high-dimensional representations, partially visible hard labels, and LLM soft priors are jointly encoded as node features, and residual GLP aggregates cross-sample relational information to incrementally refine the Stage-1 predictions.

3.2.1 Syntax-Aware Local Domain Modeling

The first stage aims to characterize the semantic state of a target word in its current context and to produce task-specific representations suitable for subsequent graph construction and relational propagation. For the i-th instance, let $S_i = (w_{i1}, w_{i2}, \ldots, w_{in_i})$

denote the contextual sentence and w_i^t the corresponding target word. A pretrained language model [46, 47] encodes the input sequence as

$$\mathbf{H}_i = \text{Encoder}(S_i) \tag{4}$$

where $\mathbf{H}_i \in \mathbb{R}^{L_i \times d_e}$, L_i is the tokenized sequence length, and d_e is the encoder hidden dimension.

The hidden state associated with the special classification token is used as the global sentence representation:

$$\mathbf{h}_i^g = \mathbf{H}_i[0] \tag{5}$$

When the target word is split into multiple subword units, their hidden states are averaged to obtain its contextualized representation:

$$\mathbf{h}_i^t = \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \mathbf{H}_i[j] \tag{6}$$

where $\mathcal{I}_i$ denotes the set of subword positions corresponding to w_i^t.

To explicitly model the incongruity between the literal usage domain of the target word and the conceptual domain activated in the current context, we construct literal examples for the target word and identify candidate context words according to semantic-role relations. The target word and each candidate context word form a candidate word pair. Let $\mathbf{h}_{ij}^c$ denote the representation of the j-th candidate context word. Its interaction with the target word is modeled through cross-attention:

$$\mathbf{h}_{ij}^p = \text{CrossAtt}\left(\mathbf{h}_i^t, \mathbf{h}_{ij}^c\right) \tag{7}$$

Let q_{ij} denote the conceptual-domain incongruity score of the j-th candidate pair. For an instance containing M_i candidate pairs, the normalized pair weight is defined as

$$\alpha_{ij} = \frac{q_{ij}}{\sum_{k=1}^{M_i} q_{ik}} \qquad j = 1, 2, \ldots, M_i \tag{8}$$

The conceptual-domain-incongruent word-pair representation is then obtained through weighted aggregation:

$$\mathbf{h}_i^p = \sum_{j=1}^{M_i} \alpha_{ij} \mathbf{h}_{ij}^p \tag{9}$$

This aggregation emphasizes contextual components that provide stronger evidence of cross-domain usage, instead of treating all context words as equally informative.

In addition to semantic evidence, SA-RGLP incorporates the syntactic category of the target word. Let r_i denote its part-of-speech index. The corresponding syntax-aware representation is obtained from a learnable embedding matrix:

$$\mathbf{e}_i^{\text{pos}} = \mathbf{E}_{\text{pos}}[r_i] \tag{10}$$

The global sentence representation, contextualized target-word representation, conceptual-domain-incongruent word-pair representation, and POS embedding are concatenated:

$$\mathbf{u}_i = \left[\mathbf{h}_i^g; \mathbf{h}_i^t; \mathbf{h}_i^p; \mathbf{e}_i^{\text{pos}}\right] \tag{11}$$

An instance-level representation and the corresponding local classification logits are subsequently computed as

$$\mathbf{h}_i = \text{MLP}(\mathbf{u}_i) \tag{12}$$

$$\mathbf{z}_i^{\text{loc}} = \mathbf{W}_c \mathbf{h}_i + \mathbf{b}_c \tag{13}$$

Let $\mathbf{y}_i = [y_{i0}, y_{i1}]^\top$ denote the one-hot label vector of the i-th training instance. The first-stage model is optimized using the average cross-entropy loss:

$$\mathcal{L}_{\text{loc}} = -\frac{1}{N_{\text{tr}}} \sum_{i \in \mathcal{D}_{\text{tr}}} \sum_{c=0}^{1} y_{ic} \log \left[\text{Softmax}\left(\mathbf{z}_i^{\text{loc}}\right)\right]_c \tag{14}$$

After Stage 1 training, $\mathbf{h}_i$ provides the task-specific semantic basis for graph construction, while $\mathbf{z}_i^{\text{loc}}$ serves as the initial prediction to be refined in Stage 2.

3.2.2 LLM Soft Prior Generation via Structured Adversarial Prompting

The local model learns metaphor-discriminative evidence from task-specific labeled data, whereas an LLM provides complementary open-domain knowledge about word senses, polysemy, commonsense, and conceptual relations. To preserve this complementary information, we design a structured adversarial prompting strategy that requires the LLM to compare two competing semantic hypotheses for each instance.

To improve reproducibility, the prompt templates used in the structured adversarial prompting procedure are summarized in Table 1.

The metaphorical hypothesis $\mathcal{H}_i^{\text{met}}$ assumes that the target-word usage involves a mapping across distinct conceptual domains. The literal hypothesis $\mathcal{H}_i^{\text{lit}}$ assumes that the usage remains within its conventional conceptual scope and corresponds to either a literal meaning or a conventional polysemous extension. Given the sentence and the target word, the LLM estimates the probability of the metaphorical interpretation:

$$p_i^{\text{met}} = f_{\text{LLM}}\left(S_i, w_i^t, \mathcal{H}_i^{\text{met}}, \mathcal{H}_i^{\text{lit}}\right) \qquad p_i^{\text{met}} \in [0,1] \tag{15}$$

Instead of converting this value into a hard pseudo-label, SA-RGLP retains the complete binary probability distribution:

$$\mathbf{p}_i^{\text{LLM}} = \left[1 - p_i^{\text{met}}, p_i^{\text{met}}\right] \tag{16}$$

The two components represent the LLM's relative preference for the literal and metaphorical interpretations, respectively. This continuous representation distinguishes confident cases from usages near the literal–metaphorical decision boundary

Table 1 Prompt templates used in the structured adversarial prompting procedure.

Prompt Type	Prompt Content
Moderator	**[Input]:** Sentence: "{sentence}"; Target Word: "{core_word}"; Target Word Index: "{core_position}". **[User Prompt]:** You are an expert linguistic judge evaluating a deep semantic debate. Your task is to determine whether the target word in a sentence is used metaphorically or literally. **[Output Format]:** Final Probability: $\langle 0.0 \sim 1.0 \rangle$.
Metaphor	The target word "{core_word}" induces a cognitive structural projection from a concrete, physical source domain to an abstract target domain. Its presence dynamizes cross-domain mapping, meaning that the semantic baseline has shifted fundamentally away from its literal physical definition.
Literal	The target word "{core_word}" remains strictly within its native domain. This specific usage is merely a standard polysemous extension, abstract generalization, or a result of semantic fuzzy matching for structural alignment. No conceptual leap across distinct domains has occurred.

and therefore preserves the semantic gradience of metaphor. The LLM soft priors are generated offline and subsequently used as independent node features in Stage 2.

3.2.3 Feature-Space Node Augmentation and Global Semantic Graph Construction

The Stage-1 representations describe individual target-word usages but do not explicitly organize semantic relations across samples. To capture recurring conceptual mapping patterns, we construct a global semantic instance graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}) \tag{17}$$

where each node represents a specific sentence–target-word usage rather than a context-independent lexical type. Consequently, two occurrences of the same target word are treated as distinct nodes when they appear in different contexts.

Because metaphor detection datasets are typically small, we augment the training-node distribution through within-class interpolation in the feature space. For two training instances i and j satisfying $y_i = y_j$, an interpolation coefficient $\lambda \in [0, 1]$ is sampled and an augmented representation is generated as

$$\mathbf{h}_{ij}^{\text{aug}} = \lambda \mathbf{h}_i + (1 - \lambda)\mathbf{h}_j \tag{18}$$

To preserve consistency across the multi-source node information, the local logits and LLM soft priors are interpolated using the same coefficient:

$$\mathbf{z}_{ij}^{\text{aug}} = \lambda \mathbf{z}_i^{\text{loc}} + (1-\lambda)\mathbf{z}_j^{\text{loc}} \tag{19}$$

$$\mathbf{p}_{ij}^{\text{aug}} = \lambda \mathbf{p}_i^{\text{LLM}} + (1-\lambda)\mathbf{p}_j^{\text{LLM}} \tag{20}$$

After combining the original and augmented nodes, semantic affinity is computed using cosine similarity:

$$s_{ij} = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\|_2 \|\mathbf{h}_j\|_2} \tag{21}$$

To suppress weak or noisy relations, an edge is retained only when the node pair satisfies both the similarity-threshold constraint and the k-nearest-neighbor constraint:

$$A_{ij} = \begin{cases} s_{ij} & s_{ij} > \tau \text{ and } j \in \mathcal{N}_k(i) \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where $\mathcal{N}_k(i)$ denotes the set of the k nearest semantic neighbors of node i, and τ is the similarity threshold.

The resulting adjacency matrix is symmetrized to obtain $\mathbf{A}^{\text{sym}}$. Self-loops are then introduced:

$$\widetilde{\mathbf{A}} = \mathbf{A}^{\text{sym}} + \mathbf{I} \tag{23}$$

Let $\widetilde{\mathbf{D}}$ be the degree matrix of $\widetilde{\mathbf{A}}$. Symmetric normalization gives

$$\widehat{\mathbf{A}} = \widetilde{\mathbf{D}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{D}}^{-\frac{1}{2}} \tag{24}$$

The normalized matrix $\widehat{\mathbf{A}}$ determines the strength with which information is aggregated from neighboring nodes during graph propagation.

3.2.4 Residual Graph Label Propagation

The final module jointly models three complementary information sources: local task-specific semantics, class-level regularities provided by human annotations, and open-domain conceptual priors generated by the LLM. For each training node i, the binary label is converted into a one-hot vector:

$$\mathbf{y}_i = \text{OneHot}(y_i) \in \{0,1\}^2 \tag{25}$$

During each graph-training iteration, the hard-label features of a randomly selected subset of training nodes are masked. Let $m_i \in \{0,1\}$ indicate whether the label feature of node i is visible. The label input is defined as

$$\widetilde{\mathbf{y}}_i = m_i \mathbf{y}_i \tag{26}$$

where $m_i = 1$ retains the hard-label feature and $m_i = 0$ replaces it with the zero vector. Labels of validation and test nodes are never exposed to the graph model:

$$\widetilde{\mathbf{y}}_i = \mathbf{0} \qquad i \in \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{te}} \tag{27}$$

This masking mechanism prevents a node from trivially copying its own label and encourages inference from its semantic representation, LLM soft prior, and labeled semantic neighborhood.

The multi-source input feature of node i is constructed as

$$\mathbf{x}_i = \left[\mathbf{h}_i; \widetilde{\mathbf{y}}_i; \mathbf{p}_i^{\text{LLM}}\right] \tag{28}$$

Stacking all node features yields the graph input matrix $\mathbf{X}$.

A two-layer graph convolutional network propagates information over the normalized semantic graph. The first layer computes

$$\mathbf{H}^{(1)} = \text{ReLU}\left(\widehat{\mathbf{A}}\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1\right) \tag{29}$$

The second layer produces a graph-based correction to the local classification logits:

$$\Delta\mathbf{Z}^{\text{graph}} = \widehat{\mathbf{A}}\,\text{Dropout}\left(\mathbf{H}^{(1)}\right)\mathbf{W}_2 + \mathbf{b}_2 \tag{30}$$

Instead of treating graph propagation as an independent classifier, SA-RGLP interprets the graph output as an incremental correction contributed by cross-sample relations. The final logits are obtained through residual fusion:

$$\mathbf{Z}^{\text{final}} = \mathbf{Z}^{\text{loc}} + \Delta\mathbf{Z}^{\text{graph}} \tag{31}$$

For node i, the corresponding expression is

$$\mathbf{z}_i^{\text{final}} = \mathbf{z}_i^{\text{loc}} + \Delta\mathbf{z}_i^{\text{graph}} \tag{32}$$

The predicted label is

$$\widehat{y}_i = \arg\max_{c \in \mathcal{C}} \left[\text{Softmax}\left(\mathbf{z}_i^{\text{final}}\right)\right]_c \qquad \mathcal{C} = \{0, 1\} \tag{33}$$

Let $\mathcal{M}_{\text{mask}}$ denote the set of training nodes whose hard-label features are masked in the current iteration. For each masked node, the classification loss is

$$\ell_i = -\sum_{c \in \mathcal{C}} y_{ic} \log \left[\text{Softmax}\left(\mathbf{z}_i^{\text{final}}\right)\right]_c \tag{34}$$

The graph-stage objective is defined as

$$\mathcal{L}_{\text{graph}} = \frac{1}{|\mathcal{M}_{\text{mask}}|} \sum_{i \in \mathcal{M}_{\text{mask}}} \ell_i. \tag{35}$$

By computing supervision only over masked training nodes, the graph model learns to recover target labels from local semantics, LLM-derived soft priors, and class information propagated through the semantic neighborhood. The residual formulation further preserves the discriminative capability of the Stage-1 classifier while allowing global relational evidence to correct locally ambiguous predictions.

3.3 Prevention of Label Leakage

In graph-based representation learning, label leakage is a critical issue where a node inappropriately utilizes its own ground-truth label during the neighborhood aggregation phase to predict itself. To strictly prevent this in our SA-RGLP framework, we implement a decoupled topological masking mechanism:

- **Feature-Only Topology Construction:** The graph adjacency matrix $\mathbf{A}$ is constructed exclusively based on high-order semantic feature similarities (e.g., cosine similarity of embeddings). It is completely isolated from any label-space information or external soft labels, preventing structural shortcut learning.
- **Masked Node Aggregation:** To avoid self-label leakage during message passing, we apply a stochastic masking strategy during training. Specifically, for training nodes, a portion of their ground-truth label components in the input feature representation is randomly zeroed out or masked according to a predefined ratio. This forces the model to predict target labels by aggregating neighborhood structural context and semantic features, while unmasked nodes retain sufficient supervision signals for parameter optimization via cross-entropy loss.
- **Independent LLM Prior Generation:** The continuous LLM priors are generated in a separate inference stage using only the input sentence and the designated target word, without access to ground-truth labels. Rather than serving as pseudo-labels or direct supervision targets, these probability distributions are introduced as auxiliary node features during graph reasoning. For instances without valid LLM outputs, a label-independent neutral prior is used, thereby preventing gold-label information from entering the prior-generation process.

3.4 Training Algorithm

The preceding subsections describe the individual components of SA-RGLP, including local representation learning, LLM soft-prior generation, semantic graph construction, and residual graph label propagation. To clarify how these components are coordinated during model optimization, Algorithm 1 summarizes the complete training procedure. Specifically, SA-RGLP follows a sequential two-stage optimization strategy: the first stage learns metaphor-discriminative local representations and initial predictions, which subsequently provide the semantic basis for graph construction and

Algorithm 1 Training Procedure of SA-RGLP

Require: Training set $\mathcal{D}_{\text{tr}}$, evaluation set $\mathcal{D}_{\text{ev}}$, neighbor number k, similarity threshold τ, masking ratio ρ

Ensure: Predicted labels $\hat{\mathbf{Y}}_{\text{ev}}$

Stage 1: Syntax-Aware Explainable Domain Modeling

Initialize the pretrained encoder, POS embeddings, feature-fusion module, and local classifier
while Stage 1 has not converged **do**
    Encode each sentence–target-word instance
    Obtain global, target-word, and word-pair representations $\mathbf{h}_i^g$, $\mathbf{h}_i^t$, and $\mathbf{h}_i^p$
    Obtain the POS embedding $\mathbf{e}_i^{\text{pos}}$
    Fuse $\mathbf{h}_i^g$, $\mathbf{h}_i^t$, $\mathbf{h}_i^p$, and $\mathbf{e}_i^{\text{pos}}$ to obtain $\mathbf{h}_i$
    Compute local logits $\mathbf{z}_i^{\text{loc}}$
    Update Stage-1 parameters using $\mathcal{L}_{\text{loc}}$
end while

Stage 2: Residual Graph Label Propagation with LLM Soft Priors

Freeze the trained Stage-1 representation module
Extract $\mathbf{h}_i$ and $\mathbf{z}_i^{\text{loc}}$ for all graph nodes
Generate LLM soft priors $\mathbf{p}_i^{\text{LLM}}$ via structured adversarial prompting
Perform within-class feature-space node augmentation
Construct the global semantic graph using cosine similarity, k-NN connectivity, and threshold τ
Normalize the graph adjacency matrix to obtain $\widehat{\mathbf{A}}$
while Stage 2 has not converged **do**
    Randomly mask a proportion ρ of training-node hard-label features
    Construct node features $\mathbf{x}_i = [\mathbf{h}_i; \widetilde{\mathbf{y}}_i; \mathbf{p}_i^{\text{LLM}}]$
    Compute graph-based residual corrections $\Delta\mathbf{Z}^{\text{graph}}$
    Compute final logits $\mathbf{Z}^{\text{final}} = \mathbf{Z}^{\text{loc}} + \Delta\mathbf{Z}^{\text{graph}}$
    Update graph parameters using $\mathcal{L}_{\text{graph}}$ on the masked training nodes
end while
Set all training-node labels as visible and keep evaluation-node label features zero
Predict $\hat{\mathbf{Y}}_{\text{ev}} = \text{argmax}(\mathbf{Z}_{\text{ev}}^{\text{final}})$
return $\hat{\mathbf{Y}}_{\text{ev}}$

relational reasoning in the second stage. The second-stage graph module is optimized while preserving the learned local representations, allowing global relational evidence and LLM-derived priors to refine rather than overwrite the local predictions. This procedure provides a compact algorithmic view of the information flow and parameter optimization throughout SA-RGLP.

Table 2 Statistics of the metaphor detection datasets. #Sent. denotes the number of sentences; #Verb denotes the number of target-verb instances; %Met. denotes the percentage of metaphorical instances; and Avg. Len. denotes the average sentence length.

Dataset	Split	#Sent.	#Verb	%Met.	Avg. Len.
MOH-X	–	647	647	48.69	8.0
TroFi	–	3,737	3,737	43.54	28.3
VUA_Verb	Train	7,479	15,516	27.90	20.2
VUA_Verb	Validation	1,541	1,724	26.91	25.0
VUA_Verb	Test	2,694	5,873	29.98	18.6

4 Experiments

4.1 Datasets and Experimental Settings

4.1.1 Datasets

To comprehensively evaluate the metaphor detection capability of SA-RGLP across different data scales and corpus settings, we conduct experiments on three widely used benchmark datasets for English verb metaphor detection: MOH-X, TroFi, and VUA_Verb. These datasets differ substantially in corpus size, sentence length, and metaphor-class distribution, and their statistics are summarized in Table 2.

MOH-X is derived from the MOH corpus introduced by Mohammad et al. [48], which annotates metaphoricity for 1,639 WordNet[49] example sentences covering 440 verbs. We adopt its 647-instance subset widely used in subsequent metaphor detection studies. Given its relatively small size and nearly balanced class distribution, MOH-X provides a useful benchmark for evaluating model robustness under limited supervision.

TroFi was introduced by Birke and Sarkar [50] and collects literal and nonliteral usages of 50 English verbs from the 1987–1989 Wall Street Journal Corpus. Compared with MOH-X, its instances have considerably longer sentences and richer sentential contexts, making TroFi suitable for evaluating whether the model can exploit both local contextual evidence and cross-sample semantic relations.

VUA_Verb is the verb-only subset of the VU Amsterdam Metaphor Corpus, covering multiple registers such as academic writing, news, fiction, and conversation. The corpus was manually annotated following the MIPVU procedure [16] and used in the VUA metaphor detection shared task [51]. Compared with MOH-X and TroFi, VUA_Verb is substantially larger and exhibits a more imbalanced class distribution, thereby providing a more challenging large-scale benchmark for evaluating metaphor detection under realistic class imbalance.

Data Splitting Protocol. For MOH-X and TroFi, we follow the commonly used 10-fold cross-validation protocol. Each dataset is divided into ten stratified folds. In each run, nine folds are used for model training and the remaining fold is held out for evaluation. This process is repeated until every fold has been used once as the

test fold, and the final performance is reported as the average over all ten runs. For VUA_Verb, we follow its predefined corpus partition. The original test set is retained for final evaluation, while 10% of the original training data is held out as the validation set. The resulting training, validation, and test partitions are reported in Table 2. To ensure reproducibility, the random seeds used for data preprocessing, sampling, and graph-related stochastic operations are fixed across all experimental runs.

4.1.2 Evaluation Metrics

Evaluation Metrics.

Following widely adopted evaluation protocols in metaphor detection [9, 13, 14], we evaluate SA-RGLP using Accuracy (Acc.) and the F_1-score (F_1) as the main performance measures, while Precision (P) and Recall (R) are additionally reported in the ablation experiments. Accuracy reflects the overall correctness of metaphor and literal classification, whereas Precision measures the reliability of metaphor predictions by indicating how many instances predicted as metaphorical are truly metaphorical. Recall evaluates the model's ability to recover metaphorical usages and is particularly useful for revealing missed metaphors, while F_1 jointly considers Precision and Recall and therefore provides a more balanced assessment of metaphor recognition performance.

We regard F_1 as the primary evaluation metric, consistent with common practice in metaphor detection and the VUA metaphor detection shared tasks [51]. This is particularly important for datasets such as VUA_Verb, where metaphorical instances constitute only a minority of the data, because a model may achieve relatively high Accuracy by favoring the dominant literal class while still failing to identify many metaphors. Accordingly, F_1 better reflects the balance between false metaphor detections and missed metaphorical usages, while Precision and Recall are further examined in the ablation study to clarify how individual components affect these two types of prediction errors.

4.1.3 Implementation Details

SA-RGLP is implemented using PyTorch v2.9.1 and the Transformers v4.57.3 library. All experiments are conducted on a computing cluster equipped with eight NVIDIA A40 GPUs. The overall training schedule contains 15 epochs.

In Stage 1, RoBERTa-base is employed as the pretrained text encoder. The syntax-aware local metaphor detection model is optimized using AdamW. To balance stable transfer of pretrained linguistic knowledge with sufficient adaptation of the newly introduced task-specific components, different learning rates are assigned to the pretrained encoder and the task-specific modules. Specifically, the encoder learning rate is set to 3×10^{-5}, whereas the learning rate of the task-specific modules is set to 3×10^{-3}.

In Stage 2, the instance representations and local classification logits produced by Stage 1 are used to construct the global semantic graph and initialize RGLP. The LLM soft priors are generated offline using GLM-4-Flash and remain fixed during subsequent model training. The number of semantic neighbors is set to $k = 5$, and the cosine-similarity threshold for graph construction is set to $\tau = 0.80$. Following the analysis in Section 4.5, the hard-label masking ratio is fixed at $\rho = 0.5$.

Table 3 Main hyperparameter settings of SA-RGLP.

Configuration or parameter	Value
General experimental settings	
Deep learning framework	PyTorch v2.9.1
Transformer library	Transformers v4.57.3
Computing hardware	8 NVIDIA A40 GPUs
Training epochs	15
MOH-X and TroFi evaluation	10-fold cross-validation
VUA_Verb evaluation	Official data split
LLM for soft-prior generation	GLM-4-Flash
Soft-prior generation mode	Offline and fixed
Stage 1: Syntax-Aware Local Domain Modeling	
Pretrained text encoder	RoBERTa-base
Maximum sequence length	160
Batch size	32
Encoder learning rate	3×10^{-5}
Task-module learning rate	3×10^{-3}
Optimizer	AdamW
Dropout rate	0.5
POS embedding dimension	64
Stage 2: Residual Graph Label Propagation with LLM Soft Priors with LLM Soft Priors	
Number of graph neighbors k	5
Graph similarity threshold τ	0.80
Graph hidden dimension	128
Hard-label masking ratio ρ	0.5
Graph learning rate	0.01

For MOH-X and TroFi, we employ 10-fold cross-validation and report the average performance across the ten folds. For VUA_Verb, we follow the official training, validation, and test splits. The principal hyperparameter settings are summarized in Table 3.

4.2 Baseline Models

To comprehensively evaluate SA-RGLP, we compare it with representative metaphor detection methods covering multi-task learning, linguistically informed pretrained models, conceptual-domain modeling, and LLM-based reasoning.

MUL_GCN [52] employs multi-task learning to transfer knowledge from word sense disambiguation to metaphor detection with graph convolutional networks. MelBERT

[8] incorporates the Metaphor Identification Procedure (MIP) [15–17] and Selectional Preference Violation (SPV) [18, 19] into contextualized representation learning to capture semantic discrepancies associated with metaphorical usage. MrBERT [9] formulates verb metaphor detection as a relation classification task and models the relations between the target verb and its grammatical, sentential, and semantic contexts. MisNet [10] introduces a linguistically enhanced Siamese network inspired by MIP and SPV to model the semantic relations between the target word, its basic meaning, and its contextual usage. AdMul [53] adopts adversarial multi-task learning to transfer knowledge from basic sense discrimination to metaphor detection. BasicBERT [13] explicitly models the basic meaning of a target word from literal annotations in the training data and compares it with its contextual meaning for metaphor identification. WPDM [14], guided by Conceptual Metaphor Theory (CMT) [20], bridges word-pair-level and token-level metaphor detection by treating word pairs as explainable intermediate clues and identifying core word pairs through conceptual-domain mining based on source–target domain inconsistency.

We further include representative LLM-based methods to evaluate SA-RGLP against recent knowledge-driven metaphor reasoning approaches. Yang et al. [31] use ChatGPT to generate literal collocations and subject–object pairs for target verbs and detect metaphors by analyzing their relations in a shared topic space. TSI [32] constructs a metaphor-theory knowledge graph and uses theory-guided scaffolding questions to progressively guide LLM-based metaphor reasoning. DMD [33] guides LLMs from both implicit and explicit theory-driven perspectives and integrates the resulting judgments through a self-judgment mechanism.

4.3 Results and Analysis

Table 4 compares SA-RGLP with representative baseline methods on MOH-X, TroFi, and VUA_Verb. All values are reported as percentages. The average results are macro-averaged across the three datasets and are calculated only for methods with available results on all three benchmarks. A dash ("–") denotes an unavailable or unreliably reproduced result.

On MOH-X, SA-RGLP achieves an Accuracy of 85.00% and an F_1 score of 84.77%. Although these results are lower than those of the GPT-4o-based DMD model, SA-RGLP outperforms all reported non-LLM supervised baselines. MOH-X contains only 647 instances, resulting in a relatively sparse cross-sample topology. This limited graph density constrains the amount of complementary relational evidence available during propagation and helps explain why the improvement is smaller than on the larger datasets.

On TroFi, SA-RGLP establishes new state-of-the-art results, achieving 76.85% Accuracy and a 76.50% F_1 score. These results exceed the previous best values by 0.55 and 2.77 percentage points, respectively. TroFi contains more instances and longer contexts than MOH-X, allowing the global graph to form a more informative semantic topology. Semantically related target-word usages can consequently provide complementary evidence for ambiguous instances. This result supports the effectiveness of extending metaphor detection beyond isolated instance-level inference.

Table 4 Comparison with baseline methods on MOH-X, TroFi, and VUA_Verb. All scores are reported as percentages. Average scores are macro-averaged across the three datasets and are reported only for methods with available results on all three datasets. Bold and underlined values indicate the best and second-best results, respectively. A dash ("–") indicates unavailable or unreliably reproduced results.

Model	Year	MOH-X		TroFi		VUA_Verb		Average	
		Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
MUL_GCN	2020	78.00	78.70	74.70	71.60	81.80	69.60	78.17	73.30
MrBERT	2021	81.90	82.10	76.00	72.90	84.90	74.40	80.93	76.47
MelBERT	2021	77.89	77.88	62.89	62.36	–	–	–	–
MisNet	2022	77.11	77.08	62.67	62.01	**86.00**	<u>75.90</u>	75.26	71.66
AdMul	2023	79.89	79.74	62.67	60.54	<u>85.20</u>	74.60	75.92	71.63
BasicBERT	2023	79.20	79.90	69.50	69.40	83.00	72.80	77.23	74.03
TSI	2024	82.93	82.59	66.89	66.07	–	–	–	–
Yang et al.	2024	61.20	70.10	45.80	60.10	45.40	50.00	50.80	60.07
WPDM	2024	84.20	83.80	<u>76.30</u>	73.40	84.40	74.10	<u>81.63</u>	<u>77.10</u>
DMD (GPT-4o)	2025	**91.53**	**91.71**	70.78	<u>73.73</u>	–	–	–	–
SA-RGLP (Ours)	2026	<u>85.00</u>	<u>84.77</u>	**76.85**	**76.50**	84.86	**81.83**	**82.24**	**81.03**

Table 5 Ablation results of SA-RGLP on MOH-X. All scores are reported as percentages. Bold and underlined values indicate the best and second-best results, respectively.

Configuration	Acc.	F_1	P	R
SA-RGLP (Ours)	**85.00**	**84.77**	**85.19**	<u>84.69</u>
w/o LLM Prior	<u>84.37</u>	<u>84.22</u>	<u>84.59</u>	84.28
w/o RGLP	82.20	83.35	79.82	**88.28**
w/o Syntax-Aware	84.22	83.88	84.50	83.85

On VUA_Verb, SA-RGLP achieves a new state-of-the-art F_1 score of 81.83%, exceeding the previous best result by 5.93 percentage points, while maintaining a competitive Accuracy of 84.86%. The substantial F_1 improvement indicates more balanced recognition of metaphorical instances under class imbalance. It also demonstrates that partially visible human labels and LLM soft priors provide complementary evidence when propagated through task-specific semantic neighborhoods.

Overall, the results reveal a consistent relationship between graph structure and model effectiveness. Small datasets restrict relational reasoning because only sparse cross-sample connections can be formed, whereas larger datasets support a richer global topology. By integrating local syntax-aware semantics, cross-sample graph propagation, and LLM-derived soft priors, SA-RGLP extends metaphor detection from isolated instance classification to local–global joint reasoning.

4.4 Ablation Studies

To evaluate the contribution of the principal components of SA-RGLP, we construct three variants by removing the LLM soft priors, RGLP, and the syntax-aware module,

Table 6 Ablation results of SA-RGLP on TroFi. All scores are reported as percentages. Bold and underlined values indicate the best and second-best results, respectively.

Configuration	Acc.	$\mathbf{F}_1$	P	R
SA-RGLP (Ours)	**76.85**	**76.50**	**76.50**	**76.67**
w/o LLM Prior	76.58	76.32	76.24	76.40
w/o RGLP	76.00	72.98	71.82	74.60
w/o Syntax-Aware	<u>76.75</u>	<u>76.40</u>	<u>76.48</u>	<u>76.63</u>

Table 7 Ablation results of SA-RGLP on VUA_Verb. All scores are reported as percentages. Bold and underlined values indicate the best and second-best results, respectively.

Configuration	Acc.	$\mathbf{F}_1$	P	R
SA-RGLP (Ours)	**84.86**	**81.83**	82.09	**81.59**
w/o LLM Prior	84.19	<u>81.79</u>	**82.88**	<u>80.73</u>
w/o RGLP	84.54	74.77	78.43	71.44
w/o Syntax-Aware	<u>84.65</u>	81.47	<u>82.33</u>	80.63

respectively. The corresponding results on MOH-X, TroFi, and VUA_Verb are reported in Tables 5–7.

Overall, the complete SA-RGLP consistently achieves the highest Accuracy and F_1 scores across all three datasets, demonstrating that the three components provide complementary information for metaphor detection. Among them, removing RGLP causes the most substantial degradation, whereas removing the syntax-aware module or LLM soft priors leads to smaller but generally consistent performance decreases. The detailed Precision and Recall results further reveal that these components influence not only overall classification performance but also the balance between correctly identifying metaphorical instances and avoiding false-positive predictions.

MOH-X: As shown in Table 5, the complete SA-RGLP achieves an Accuracy of 85.00% and an F_1 score of 84.77%. Removing RGLP reduces Accuracy by 2.80 percentage points and F_1 by 1.42 percentage points, representing the largest degradation among the three variants. Interestingly, the Recall increases from 84.69% to 88.28% after removing RGLP, whereas Precision decreases sharply from 85.19% to 79.82%. This result suggests that, without global relational refinement, the local classifier becomes more inclined to predict instances as metaphorical, recovering more positive samples at the cost of substantially more false-positive predictions. RGLP therefore contributes not only to overall performance but also to maintaining a more balanced trade-off between Precision and Recall through cross-sample relational evidence. Removing the syntax-aware module decreases Accuracy and F_1 by 0.78 and 0.89 percentage points, respectively, indicating that syntactic information provides useful local evidence for constructing metaphor-discriminative representations. Removing

the LLM soft priors results in smaller reductions of 0.63 points in Accuracy and 0.55 points in F_1, showing that the external conceptual knowledge provided by the LLM offers additional complementary information even on this relatively small and nearly balanced dataset.

TroFi: Table 6 shows a similar pattern on TroFi, where the full model obtains 76.85% Accuracy and 76.50% F_1. Removing RGLP decreases F_1 by 3.52 percentage points, from 76.50% to 72.98%, while Precision and Recall drop from 76.50% and 76.67% to 71.82% and 74.60%, respectively. The simultaneous degradation of Precision and Recall indicates that graph propagation contributes robust relational evidence rather than merely shifting the classification threshold toward one class. Compared with MOH-X, the larger performance loss after removing RGLP suggests that TroFi provides a richer cross-sample semantic structure from which global graph reasoning can benefit. In contrast, removing the syntax-aware module causes only a 0.10-point decrease in F_1, while removing the LLM soft priors reduces F_1 by 0.18 points. Although these improvements are relatively modest on TroFi, both components consistently improve the complete model, indicating that local syntactic cues and LLM-derived conceptual priors remain complementary to the dominant graph-based relational signal.

VUA_Verb: The effect of RGLP becomes particularly pronounced on the larger and more imbalanced VUA_Verb dataset, as shown in Table 7. The complete model achieves an F_1 score of 81.83%, whereas removing RGLP reduces it to 74.77%, corresponding to a substantial decrease of 7.06 percentage points. Notably, the Accuracy decreases by only 0.32 percentage points, from 84.86% to 84.54%, while Recall drops sharply from 81.59% to 71.44%. This discrepancy between Accuracy and F_1 is particularly informative under class imbalance, because a model can preserve relatively high overall Accuracy by favoring the majority literal class while failing to identify a considerable proportion of metaphorical instances. The large Recall degradation therefore demonstrates that RGLP plays an important role in recovering metaphorical samples by propagating information from semantically related target-word usages. Removing the syntax-aware module decreases F_1 by 0.36 percentage points and primarily affects Recall, which falls from 81.59% to 80.63%, suggesting that syntactic cues help preserve metaphor-sensitive local representations. Removing the LLM soft priors results in only a 0.04-point decrease in F_1; however, its Precision increases from 82.09% to 82.88% while Recall decreases from 81.59% to 80.73%. This behavior suggests that the LLM soft prior slightly shifts the model toward recovering additional metaphorical instances, improving the Precision–Recall balance rather than simply increasing one individual metric.

Taken together, the ablation results reveal a clear functional division among the three components of SA-RGLP. The syntax-aware module strengthens task-specific local representations by incorporating syntactic evidence, while the LLM soft priors provide complementary open-domain conceptual knowledge that helps resolve semantically ambiguous instances. RGLP contributes the largest performance gains by organizing independently encoded sentence–target-word usages into a global semantic topology and propagating relational evidence across samples. Its contribution

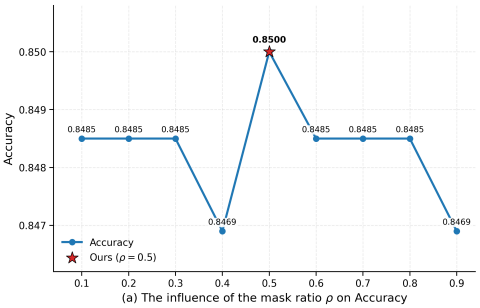


Fig. 2 Effect of the hard-label masking ratio ρ on the Accuracy of SA-RGLP on MOH-X. The highest Accuracy is achieved at $\rho = 0.5$.

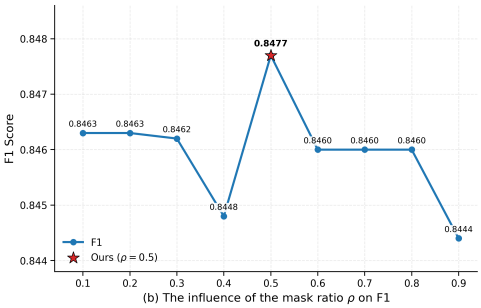


Fig. 3 Effect of the hard-label masking ratio ρ on the F_1 score of SA-RGLP on MOH-X. The highest F_1 score is achieved at $\rho = 0.5$.

becomes especially evident on VUA_Verb, where the substantial F_1 and Recall degradation without RGLP indicates that global relational reasoning is particularly valuable under larger-scale and class-imbalanced settings. The combination of these components therefore supports the central design of SA-RGLP, in which reliable local semantic representations are progressively complemented by external conceptual priors and global cross-sample relational information.

4.5 Analysis of the Label Masking Ratio

We further investigate the effect of the hard-label masking ratio ρ on the performance of SA-RGLP. This parameter controls the proportion of training nodes whose hard-label features are masked during each graph-training iteration. Figures 2 and 3 present the variations in Accuracy and F_1, respectively, on MOH-X as ρ ranges from 0.1 to 0.9.

A similar trend is observed for the F_1 score in Fig. 3. SA-RGLP achieves its highest F_1 score of 84.77% at $\rho = 0.5$. A moderate masking ratio provides a more effective balance between label-recovery supervision and visible class information. Specifically, it masks a sufficient number of nodes to encourage the graph model to infer labels from semantic neighborhoods, while retaining enough labeled nodes to support class-level information propagation.

Based on these results, we use $\rho = 0.5$ as the default hard-label masking ratio in all reported experiments.

4.6 Effect of Different LLM Prior Forms

To evaluate the utility of LLM information and compare continuous distributions with discrete pseudo-labels, we examine different LLM prior forms on MOH-X. Keeping all Stage-1 representations, graph structures, and RGLP parameters identical, we vary only the two-dimensional prior features: **No LLM Prior** assigns the zero vector $[0, 0]$; **Uniform Prior** uses the class-neutral distribution $[0.5, 0.5]$; **Hard Pseudo-Label** thresholds LLM metaphor probabilities into discrete one-hot labels; and **Continuous Soft Prior** retains the complete distribution $[1-p_i^{\text{met}}, p_i^{\text{met}}]$. The results are presented in Table 8.

Table 8 Comparison of different LLM prior forms on MOH-X. All scores are reported as percentages. Bold and underlined values indicate the best and second-best results, respectively.

LLM Prior Form	Acc.	F1	P	R
No LLM Prior	<u>84.85</u>	<u>84.63</u>	<u>85.02</u>	<u>84.56</u>
Uniform Prior	<u>84.85</u>	<u>84.63</u>	<u>85.02</u>	<u>84.56</u>
Hard Pseudo-Label	83.75	83.56	83.98	83.69
Continuous Soft Prior(ours)	**85.00**	**84.77**	**85.19**	**84.69**

As shown in Table 8, **No LLM Prior** and **Uniform Prior** yield identical results, both achieving an F1 score of 84.63%. This observation indicates that assigning the same uniform probability distribution to all samples does not provide effective sample-specific discriminative information. In other words, model performance does not improve merely by adding a two-dimensional prior feature; the effectiveness of LLM priors depends on whether they provide differentiated semantic judgments for specific target-word usages.

When the continuous probabilities are discretized into **Hard Pseudo-Labels**, the F1 score decreases to 83.56%, which is 1.21 percentage points lower than that of **Continuous Soft Prior**. Accuracy, Precision, and Recall also decrease by 1.25, 1.21, and 1.00 percentage points, respectively. These results suggest that compressing an LLM judgment into a single class discards its relative confidence in literal and metaphorical interpretations and may introduce uncertain or incorrect judgments into graph propagation as deterministic labels. By contrast, **Continuous Soft Prior** achieves the best overall performance, exceeding both **No LLM Prior** and **Uniform Prior** by 0.14 percentage points in F1. Although the improvement is modest, retaining continuous probabilities avoids the substantial degradation caused by discrete pseudo-labels. This finding indicates that treating LLM outputs as soft conceptual preferences, rather than deterministic class supervision, is more suitable for integration with local semantic representations and cross-sample graph relations.

Therefore, these results do not imply that LLM priors independently dominate model predictions. Instead, they demonstrate that continuous soft priors provide external conceptual knowledge in a more robust manner, whereas hard pseudo-labels may amplify noise in LLM judgments and interfere with the original local semantic and graph relational information.

4.7 Effect of Different LLM Backbones

To examine whether the effectiveness of SA-RGLP depends strongly on the LLM used for soft-prior generation, we further compare two different LLM backbones under the same experimental setting. Specifically, we use the locally deployed Llama-3.1-8B-Instruct and the API-based GLM-4-Flash to generate continuous metaphor probabilities, while keeping all other components and hyperparameters unchanged. The results are reported in Table 9.

Table 9 Performance with different LLMs for soft-prior generation.

LLM	Acc.	F_1	P	R
Llama-3.1-8B-Instruct	84.85	84.63	85.02	84.56
GLM-4-Flash	**85.00**	**84.77**	**85.19**	**84.69**

As shown in Table 9, both LLMs provide effective soft priors for SA-RGLP, with only minor differences in downstream performance. GLM-4-Flash achieves the best results across all four metrics, improving Accuracy and F_1 by approximately 0.15 percentage points over Llama-3.1-8B-Instruct, together with gains of 0.17 and 0.14 points in Precision and Recall, respectively. Nevertheless, the relatively small performance gap indicates that SA-RGLP is not highly sensitive to a particular LLM backbone. Even the locally deployed Llama-3.1-8B-Instruct produces competitive results, suggesting that the proposed framework can effectively exploit continuous conceptual priors from LLMs with different deployment settings and capacities. We therefore use GLM-4-Flash in the main experiments due to its slightly better and more consistent performance.

4.8 Case Study

To qualitatively demonstrate the internal execution mechanism of our proposed SA-RGLP framework, we present a step-by-step case study in Table 10. Instead of merely comparing isolated features, this analysis illustrates how the framework dynamically processes a complex instance through successive reasoning stages, progressively mitigating semantic ambiguity via its specialized modules.

By tracing a representative target word from its initial ambiguous state to its final classification, we can observe the distinct contributions of syntax-aware incongruity modeling, LLM-derived soft priors, and global graph propagation, as well as their complementary roles in refining the final prediction.

As shown in Table 10, the traditional baseline struggles due to the lack of cross-sample relational context. However, as the instance progresses through the SA-RGLP pipeline, local conceptual conflicts are first identified, followed by the injection of LLM soft priors. Finally, RGLP resolves the remaining ambiguity by situating the usage within a broader semantic topology, producing a more confident and accurate prediction.

5 Limitations and Future Work

Although the framework we proposed demonstrated strong performance in the benchmark tests, there are still some limitations, as follows:

Table 10 Qualitative case study illustrating the step-by-step reasoning process of SA-RGLP.

Component	Content / Action
Input Text	"The students grasped the concept quickly."
Target Element	**Target Word:** *grasped* **Contextual Interaction:** *grasped* ↔ *concept*
Ground Truth	*Metaphorical Usage*
1. Stage-1 Prediction	**Top-1 Prediction:** Literal Usage ($P = 0.58$) **Top-2 Prediction:** Metaphorical Usage ($P = 0.42$) **Status:** The local classifier relies solely on contextual evidence and predicts the instance independently, resulting in an uncertain literal prediction.
2. Syntax-Aware Modeling	**Action:** POS-aware contextual modeling identifies the conceptual incongruity between the physical action expressed by *grasped* and the abstract object *concept*. **Explainable Clue:** The core word pair *grasped* ↔ *concept* reveals a source–target conceptual-domain mismatch that provides explicit evidence for metaphor reasoning.
3. LLM Soft Priors	**Action:** Structured adversarial prompting guides the LLM to distinguish between a cross-domain metaphorical mapping and a literal or conventional polysemous interpretation. **Output:** The LLM soft prior is $$\mathbf{p}^{\mathrm{LLM}} = [0.25,\ 0.75],$$ where the two dimensions correspond to literal and metaphorical interpretations, respectively. The distribution indicates a stronger preference for the metaphorical interpretation.
4. Global Graph Propagation	**Node Representation:** Stage-1 representation + visible label information + LLM soft prior. **Retrieved Neighbor:** "The child grasped the rope firmly." (Literal; similarity = 0.12). **Action:** RGLP aggregates cross-sample semantic evidence and produces an incremental correction to the local classification result.
5. Final Fusion	**Residual Fusion:** The graph-derived relational correction is added to the local prediction, reducing the ambiguity of the original instance-level decision. **Final Output:** *Metaphorical Usage* (correct prediction).

Limitations in Data-Scarce Scenarios.

The SA-RGLP framework is primarily designed to leverage the rich semantic relations found in adequately sized datasets, where its global topology thrives. Consequently, a natural limitation arises in highly constrained data-scarce scenarios. As observed on the smaller MOH-X dataset, when the available instances are highly limited, the cross-sample semantic relations become inherently sparse. While the model maintains robust performance and outperforms traditional supervised baselines, this sparsity restricts the full potential of the GLP module, making it less advantageous compared to few-shot generative baselines powered by massive LLMs. To extend the framework's applicability to extreme low-resource scenarios, future work could explore dynamic graph densification strategies. For instance, incorporating external structured knowledge graphs or cross-dataset anchor nodes could help bridge isolated instances when intra-dataset relations are sparse. Additionally, introducing an adaptive weighting mechanism would allow the framework to dynamically rely more heavily on local syntax-aware representations and LLM soft priors when global semantic edges are insufficient.

Limitations in Cross-Lingual and Cross-Cultural Transfer.

The current experiments are mainly conducted on English benchmark datasets, and metaphorical expressions in cross-lingual and cross-cultural contexts have not been sufficiently evaluated. Because metaphors are deeply associated with cultural experience, social context, and language-specific conventions, the transferability of the current findings to other linguistic environments requires more fine-grained investigation. Future work may conduct systematic experiments on multilingual datasets and incorporate language-specific knowledge resources and cultural-semantic alignment strategies to progressively improve the generalization ability of SA-RGLP in cross-lingual metaphor detection.

6 Conclusion

This work presents SA-RGLP, a metaphor detection framework designed to address the limitations of existing methods in local semantic modeling, cross-sample relation utilization, and open-domain knowledge integration. SA-RGLP first employs a syntax-aware local semantic modeling module to jointly capture target-word contextual information, conceptual-domain incongruity, and syntactic features, producing linguistically interpretable instance-level representations. It then incorporates continuous LLM soft priors and uses RGLP to integrate local semantic representations, partially visible label information, and open-domain conceptual knowledge within a unified global semantic space. This design extends metaphor detection from single-instance classification to cross-sample relational reasoning. Experiments on three widely used metaphor detection benchmarks, MOH-X, TroFi, and VUA_Verb, demonstrate that SA-RGLP achieves competitive overall performance and state-of-the-art results on TroFi and VUA_Verb. Further ablation studies verify the effectiveness of the syntax-aware representations, LLM soft priors, and RGLP module. Among these components, GLP produces the most substantial performance improvements across multiple datasets, indicating that latent semantic relations among target-word usages effectively compensate for the limitations of conventional instance-level classifiers that rely only on local contextual information. Meanwhile, LLM soft priors provide additional open-domain conceptual knowledge, enabling the model to better handle the semantic ambiguity prevalent in metaphorical expressions. Overall, this work explores a new paradigm for metaphor detection by combining task-specific semantic representations, external knowledge from LLMs, and global relational reasoning. It extends metaphor identification from local semantic conflict detection to joint reasoning over multiple sources of cognitive evidence. Future work will investigate more fine-grained modeling of conceptual-domain relations and explore the application of multimodal LLMs and dynamic knowledge graph structures to complex metaphor understanding, with the aim of further improving generalization to metaphorical expressions in open-domain contexts.

Declarations

Funding. This work is a research achievement supported by the National Natural Science Foundation of China under Grant 62262065 and Grant 62476233, the "Tianchi Talent" Research Project of Xinjiang under Grant No. 51052501821, and the Key Research and Development Program of Xinjiang under Grant 2024B03041.

Data availability. The three benchmark datasets used in this study are available for research use from their original sources or previously released research repositories. The processed versions of MOH-X, TroFi, and VUA_Verb used in closely related metaphor detection studies are available in the public WPDM repository at https://github.com/TIAN-viola/WPDM/tree/main/datasets, which provides separate directories for MOH-X, TroFi, and VUA_Verb. MOH-X is derived from the metaphor-annotated WordNet examples released by Mohammad et al.; the original resource and its terms of use are available at https://saifmohammad.com/WebPages/SentimentEmotionLabeledData.html. TroFi is publicly hosted by Simon Fraser University at https://github.com/sfu-natlang/trofi-metaphor-data, and its use is subject to the accompanying TroFi and Wall Street Journal corpus licensing conditions. VUA_Verb is derived from the VU Amsterdam Metaphor Corpus; information on the original corpus and its access conditions is available through the VU Amsterdam Metaphor Corpus Online at https://research.vu.nl/en/publications/vu-amsterdam-metaphor-corpus/. For reproducibility, additional processed versions of these metaphor detection benchmarks can also be found in the MisNet repository at https://github.com/SilasTHU/MisNet and the Neural Metaphor Detection in Context repository at https://github.com/gao-g/metaphor-in-context. Users should comply with the licensing and redistribution requirements of the original datasets when obtaining or using these resources.

Conflict of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics declaration. This study uses only publicly available metaphor detection datasets and does not collect private or personally identifiable information. Since metaphor interpretation can be context-dependent and subjective, and LLM-derived priors may contain cultural or linguistic biases, the proposed method should be used with appropriate human review in sensitive applications.

References

[1] Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1179 . https://aclanthology.org/D14-1179/

[2] Wankhade, M., Rao, A.C.S., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review **55**(7), 5731–5780

(2022) https://doi.org/10.1007/s10462-022-10144-1

[3] Agarwal, B., Mittal, N., Bansal, P., Garg, S.: Sentiment analysis using common-sense and context information. Computational Intelligence and Neuroscience **2015**, 715730 (2015) https://doi.org/10.1155/2015/715730

[4] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics **37**(2), 267–307 (2011) https://doi.org/10.1162/COLI_a_00049

[5] Cowie, J., Lehnert, W.: Information extraction. Communications of the ACM **39**(1), 80–91 (1996) https://doi.org/10.1145/234173.234209

[6] Jiang, J.: Information extraction from text. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 11–41. Springer, New York, NY (2012). https://doi.org/10.1007/978-1-4614-3223-4_2 . https://doi.org/10.1007/978-1-4614-3223-4$_2$

[7] Zhong, M., Liu, Y., Xu, Y., Zhu, C., Zeng, M.: DialogLM: Pre-trained model for long dialogue understanding and summarization. Proceedings of the AAAI Conference on Artificial Intelligence **36**(10), 11765–11773 (2022) https://doi.org/10.1609/aaai.v36i10.21432

[8] Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., Lee, J.: MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1763–1773. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.141 . https://aclanthology.org/2021.naacl-main.141/

[9] Song, W., Zhou, S., Fu, R., Liu, T., Liu, L.: Verb metaphor detection via contextual relation learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4240–4251. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.327 . https://aclanthology.org/2021.acl-long.327/

[10] Zhang, S., Liu, Y.: Metaphor detection via linguistics enhanced Siamese network. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 4149–4159. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (2022). https://aclanthology.org/2022.coling-1.364/

[11] Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step. In: Image and Video Retrieval. Lecture Notes in Computer Science, vol. 3115, pp. 207–215. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27814-6_27 . https://doi.org/10.1007/978-3-540-27814-6$_2$7

[12] Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd International Conference on World Wide Web Companion, pp. 373–374. Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2567948.2577348 . https://doi.org/10.1145/2567948.2577348

[13] Li, Y., Wang, S., Lin, C., Guerin, F.: Metaphor detection via explicit basic meanings modelling. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 91–100. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-short.9 . https://aclanthology.org/2023.acl-short.9/

[14] Tian, Y., Zhang, R., Xu, N., Mao, W.: Bridging word-pair and token-level metaphor detection with explainable domain mining. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13311–13325. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.719 . https://aclanthology.org/2024.acl-long.719/

[15] Pragglejaz Group: MIP: A method for identifying metaphorically used words in discourse. Metaphor and Symbol **22**(1), 1–39 (2007) https://doi.org/10.1080/10926480709336752

[16] Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T., Pasma, T.: A Method for Linguistic Metaphor Identification: From MIP to MIPVU. Converging Evidence in Language and Communication Research, vol. 14. John Benjamins Publishing Company, Amsterdam/Philadelphia (2010). https://doi.org/10.1075/celcr.14 . https://doi.org/10.1075/celcr.14

[17] Haagsma, H., Bjerva, J.: Detecting novel metaphor using selectional preference information. In: Proceedings of the Fourth Workshop on Metaphor in NLP, pp. 10–17. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/W16-1102 . https://aclanthology.org/W16-1102/

[18] Wilks, Y.: A preferential, pattern-seeking, semantics for natural language inference. Artificial Intelligence **6**(1), 53–74 (1975) https://doi.org/10.1016/0004-3702(75)90016-8

[19] Wilks, Y.: Making preferences more active. Artificial Intelligence **11**(3), 197–223 (1978) https://doi.org/10.1016/0004-3702(78)90001-2

[20] Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago, IL (1980)

[21] Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., Hubinger, E., Lukošiūtė, K., Nguyen, K., Joseph, N., McCandlish, S., Kaplan, J., Bowman, S.R.: Studying large language model

generalization with influence functions. arXiv preprint arXiv:2308.03296 (2023) https://doi.org/10.48550/arXiv.2308.03296 arXiv:2308.03296 [cs.LG]

[22] Wang, Y., Le, H., Gotmare, A., Bui, N., Li, J., Hoi, S.: CodeT5+: Open code large language models for code understanding and generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 1069–1088. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.68 . https://aclanthology.org/2023.emnlp-main.68/

[23] Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: MemoryBank: Enhancing large language models with long-term memory. Proceedings of the AAAI Conference on Artificial Intelligence **38**(17), 19724–19731 (2024) https://doi.org/10.1609/aaai.v38i17.29946

[24] Wang, H., Du, X., Yu, W., Chen, Q., Zhu, K., Chu, Z., Yan, L., Guan, Y.: Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. Neurocomputing **618**, 129063 (2025) https://doi.org/10.1016/j.neucom.2024.129063

[25] Hu, S., Hu, J., Zhang, H.: Synergizing LLMs with global label propagation for multimodal fake news detection. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1426–1440. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-long.72 . https://aclanthology.org/2025.acl-long.72/

[26] Jia, K., Li, R.: Metaphor detection with context enhancement and curriculum learning. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 2726–2737. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.naacl-long.149 . https://aclanthology.org/2024.naacl-long.149/

[27] Gong, S.-P., Ahrens, K.: Congruent and incongruent metaphorical sentences processing in discourse: A reaction-time study. In: Huang, S.-F. (ed.) Proceedings of the 9th International Symposium on Chinese Languages and Linguistics, pp. 367–382. Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan (2004)

[28] Rauh, G.: Syntactic Categories: Their Identification and Description in Linguistic Theories. Oxford Surveys in Syntax and Morphology, vol. 7. Oxford University Press, Oxford (2010)

[29] Ohori, T.: Syntactic categories and grammatical relations: The cognitive organization of information, by william croft. English Linguistics **11**, 318–341 (1994) https://doi.org/10.9793/elsj1984.11.318

[30] Reimann, S., Scheffler, T.: Using large language models to perform MIPVU-inspired automatic metaphor detection. In: Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pp. 10–21. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.analogyangle-1.2 . https://aclanthology.org/2025.analogyangle-1.2/

[31] Yang, C., Chen, P., Huang, Q.: Can ChatGPT's performance be improved on verb metaphor detection tasks? bootstrapping and combining tacit knowledge. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1016–1027. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.57 . https://aclanthology.org/2024.acl-long.57/

[32] Tian, Y., Xu, N., Mao, W.: A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 7738–7755. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.naacl-long.428 . https://aclanthology.org/2024.naacl-long.428/

[33] Lin, Y., Liu, J., Gao, Y., Wang, A., Su, J.: A dual-perspective metaphor detection framework using large language models. In: 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Hyderabad, India (2025). https://doi.org/10.1109/ICASSP49660.2025.10889383 . https://doi.org/10.1109/ICASSP49660.2025.10889383

[34] Sánchez-Montero, A., Bel-Enguix, G., Ojeda-Trueba, S.-L., Sierra, G.: Prompting metaphoricity: Soft labeling with large language models in popular communication of science tweets in Spanish. In: Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pp. 45–56. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.analogyangle-1.5 . https://aclanthology.org/2025.analogyangle-1.5/

[35] Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, Pittsburgh, PA, USA (2002)

[36] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, vol. 16, pp. 321–328. MIT Press, Cambridge, MA, USA (2003)

[37] Niu, Z.-Y., Ji, D.-H., Tan, C.L.: Word sense disambiguation using label propagation based semi-supervised learning. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 395–402. Association for Computational Linguistics, Ann Arbor, Michigan (2005). https://doi.

org/10.3115/1219840.1219889 . https://aclanthology.org/P05-1049/

[38] Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5065–5074. IEEE, Long Beach, CA, USA (2019). https://doi.org/10.1109/CVPR.2019.00521 . https://doi.org/10.1109/CVPR.2019.00521

[39] Wang, H., Leskovec, J.: Unifying graph convolutional neural networks and label propagation. arXiv preprint arXiv:2002.06755 (2020) https://doi.org/10.48550/arXiv.2002.06755 arXiv:2002.06755 [cs.LG]

[40] Huang, Q., He, H., Singh, A., Lim, S.-N., Benson, A.R.: Combining label propagation and simple models out-performs graph neural networks. In: International Conference on Learning Representations (2021). https://arxiv.org/abs/2010.13993

[41] Zhao, W., Nakashima, Y., Chen, H., Babaguchi, N.: Enhancing fake news detection in social media via label propagation on cross-modal tweet graph. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 2400–2408. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3612086 . https://doi.org/10.1145/3581783.3612086

[42] Chen, J., Yang, W., Yin, Y., Lv, H., Wei, F., He, J., Luo, Z., Wu, J.: TR-GLP: A two-stage framework with temporal reprogramming and residual graph label propagation for short video fake news detection. In: Proceedings of the 2026 International Conference on Multimedia Retrieval, pp. 2533–2542. Association for Computing Machinery, New York, NY, USA (2026). https://doi.org/10.1145/3805622.3810626 . https://doi.org/10.1145/3805622.3810626

[43] Bose, S., Jha, A., Kandala, H., Banerjee, B.: Beyond boundaries: A novel data-augmentation discourse for open domain generalization. Transactions on Machine Learning Research (2023)

[44] Chklovski, T., Pantel, P.: Path analysis for refining verb relations. In: Proceedings of the KDD Workshop on Link Analysis and Group Detection (LinkKDD-04), Seattle, WA, USA, pp. 92–97 (2004)

[45] Tian, Y., Xu, N., Mao, W., Zeng, D.: Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7736–7752. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.480 . https://aclanthology.org/2023.emnlp-main.480/

[46] Koroteev, M.V.: BERT: A review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943 (2021) https://doi.org/10.48550/arXiv.2103.11943 arXiv:2103.11943

[47] 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) https://doi.org/10.48550/arXiv.1907.11692 arXiv:1907.11692 [cs.CL]

[48] Mohammad, S., Shutova, E., Turney, P.: Metaphor as a medium for emotion: An empirical study. In: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, pp. 23–33. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/S16-2003 . https://aclanthology.org/S16-2003/

[49] Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38**(11), 39–41 (1995) https://doi.org/10.1145/219717.219748

[50] Birke, J., Sarkar, A.: A clustering approach for nearly unsupervised recognition of nonliteral language. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 329–336. Association for Computational Linguistics, Trento, Italy (2006). https://aclanthology.org/E06-1042/

[51] Leong, C.W.B., Beigman Klebanov, B., Hamill, C., Stemle, E., Ubale, R., Chen, X.: A report on the 2020 VUA and TOEFL metaphor detection shared task. In: Proceedings of the Second Workshop on Figurative Language Processing, pp. 18–29. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.figlang-1.3 . https://aclanthology.org/2020.figlang-1.3/

[52] Le, D., Thai, M., Nguyen, T.: Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. Proceedings of the AAAI Conference on Artificial Intelligence **34**(5), 8139–8146 (2020) https://doi.org/10.1609/aaai.v34i05.6326

[53] Zhang, S., Liu, Y.: Adversarial multi-task learning for end-to-end metaphor detection. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 1483–1497. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.findings-acl.96 . https://aclanthology.org/2023.findings-acl.96/