DAGS: A Dependency-Based Dual-Attention and Global Semantic Improvement Framework for Metaphor Recognition

Puli Chen[1], Cheng Yang[1], Xingmao Zhang[3]*, Qingbao Huang[1,2,4,*]
[1]School of Electrical Engineering, Guangxi University, Nanning 530004, China
[2] School of Artificial Intelligence at Guangxi University, Nanning 530004, China
[3]College of General Education, Guangxi Arts University, Nanning 530022, China
[4]Guangxi Key Laboratory of Multimedia Communications and Network Technology, China
{2312391007, 2212391065}@st.gxu.edu.cn, 20170024@gxau.edu.cn,
qbhuang@gxu.edu.cn

Abstract

Current metaphor recognition mainly rely on Metaphor Detection Theory (MDT), such as the Metaphor Identification Procedure, which recognizes metaphors by comparing the basic meaning of target word with context meaning. Existing studies have gradually adopted literal annotations to model basic meanings, rejecting the aggregated meanings of target words. However, these methods ignore the problem of interference caused by literal annotations, and do not make full use of semantic expression relations of MDT, making the models difficult to detect and generalize. To address these challenges, we propose a dependency-based **D**ual-**A**ttention and **G**lobal **S**emantic Improvement (DAGS) framework. DAGS first extracts literal annotations of target words as basic meaning from several mainstream corpora. Then, we apply dependency tree and dual-attention while filtering on input sentences and basic meanings. Finally, we improve the MDT to further consider the global semantic relationship on contexts. The DAGS can not only extract features from multiple information sources but also effectively removes redundancy, while focusing on mission-critical information. We achieve state-of-the-art on several mainstream metaphor datasets (e.g., VUA ALL, VUAverb, TroFi and PSUCMC), which suggests that filtering and global semantic improvement of contexts is crucial for enhancing metaphor recognition performance. Our code is available at https://github.com/VILAN-Lab/Metaphor-DAGS.

1 Introduction

As a universal linguistic phenomenon, metaphor is widely found in daily communication, literary works and media reports. According to the Conceptual Metaphor Theory (CMT), metaphor is defined as a mapping between the source domain and target domain, i.e., the expression of a deeper meaning beyond its literal meaning through one or more words (Lakoff and Johnson, 2008; Lagerwerf and Meijers, 2008). For example, in "*He **hit** the nail on the head in the meeting.*" The "hit" means pinpointing the core point of issue rather than actual body part that was struck. Metaphor recognition plays an important role in cognition and communication, and is widely used in NLP tasks, such as machine translation (Babieno et al., 2022; Mao et al., 2018), paraphrase generation (Chakrabarty et al., 2020; Li et al., 2022b) and sentiment analysis (Li et al., 2022a; Cambria et al., 2017).

Several strategies have been proposed to detect metaphors. For example, Mao et al. (2018); Gao et al. (2018) models the complete sentence context, whereas Choi et al. (2021) introduces the Metaphor Identification Procedure (MIP) (Group, 2007; Steen et al., 2010) and Selectional Preference Violation (SPV) (Wilks, 1975). For MIP, a word can be identified as metaphor when its literal meaning contrasts with its context meaning of a given sentence. To SPV, the target word is metaphorical when it occurs less frequently in its context or is semantically mismatched (Choi et al., 2021). For example, in "*he **ignites** inspiration*", the contextual meaning of "ignites" is "stimulates creativity", which is different from the literal meaning of "ignites a flame". Moreover, the "ignite" is metaphorical, which is rare in the context of "inspiration". However, the challenge of accurately handling contextual noise and extracting the basic meanings of target words remains an open research question. Recently, Wang et al. (2023) considers only part of the contextual noise problem and still has redundant information and aggregated meanings (i.e., proximate meaning). Zhang and Liu (2022); Li et al. (2023a) propose an improved method for MIP by abandoning the traditional aggregated meaning of the target word and taking the literal annotation as the basic meaning (e.g., replacing "attack" with "February the Germans attacked Verdun."). They

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 10459–10476
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

argue that the use of aggregated meanings weakens the validity of MIP and SPV. Although transforming basic meanings in MIP by adding context may help in metaphor recognition (Cheng et al., 2021), it may also introduce redundant interference that prevents the model from focusing on key semantic information. Furthermore, these methods fail to fully utilize the semantic relationships of contexts when using MDT. Current SPVs only consider the difference between the target word and a single context, while ignoring its relationship with other contexts (e.g., literal annotations). For MIP, these methods also lack the understanding of the global context. Therefore, filtering redundancy in context, focusing on task-relevant information and improving semantic relationships remain serious challenges for metaphor recognition.

To solve the above problems, we propose a new metaphor recognition framework, which combines dependency-based dual-attention and semantic improvement, namely DAGS. First, we utilizes GPT-4o (*Version GPT-4o-2024-08-06*) to obtain literal annotations of target words from mainstream metaphor corpora, serving as basic meanings. Subsequently, we construct a dependency tree structure and perform dependency analysis, pruning input sentences and basic meanings simultaneously. Meanwhile, DAGS introduce a dual-attention module to further extract contextual task-relevant information. Finally, DAGS improves SPV and MIP, named global-SPV (G-SPV) and global-MIP (G-MIP), respectively, to further detect metaphors by considering the semantic relations of contexts. Additionally, DAGS also integrates multiple linguistic features, such as part-of-speech (POS), position and local context.

In summary, our contributions are as follows:

1. We propose DAGS, a structure based on RoBERTa-base that is capable of filtering both the input sentence and basic meaning.

2. We construct a unique dual-attention module on metaphor, which is able to focus on important features and information in different streams and make connections.

3. Compared with traditional SPV and MIP, our G-SPV and G-MIP can capture the semantic information of context more effectively.

4. Experiments demonstrate that DAGS achieves the best performance on several mainstream metaphor datasets, including English and cross-linguistic, and shows significant advantages in other experiments such as zero-shot.

2 Related Work

Earlier studies used aggregated meanings as a substitute for basic meanings. However, Zhang and Liu (2022); Li et al. (2023a) argued that aggregated meanings violated the MIP principle and instead began using literal annotations as the basic meaning. Building on this, Tian et al. (2024b) continued to use literal annotations to construct sets of example sentences to recognize metaphors. But these approaches fail to adequately consider redundancy and key information in multiple contexts. Previously, Wang et al. (2023) mitigated the noise in the context by optimizing dependency parsing despite the use of aggregated meanings, which provides important inspiration for dual-dependencies in DAGS. Recently, Jia and Li (2024) introduced internal and external semantics and multiple metaphor recognition mechanisms to learn differences between sentences. Similarly, Qiao et al. (2024) modeled literal meaning uncertainty in MIP through a density matrix. And Uduehi and Bunescu (2024) used SPV and MIP to construct expectation and realization components to evaluate the meaning of target words. While Wang et al. (2025) refined the conceptual knowledge of inter-word relationships and explored the similarity of cross-domain concepts. These studies further demonstrate the importance of semantics in metaphor detection, but most of the approaches ignore the role of global semantic information.

In addition, Zhang and Liu (2022); Li et al. (2023a) used aggregated meanings on some of the data despite considering literal annotations. In recent years, LLM has shown significant potential in metaphor research. For example, Wachowiak and Gromann (2023) employed GPT-3 to detect metaphor expressions in a given sentence and predict their source domains. Yang et al. (2024) detected verb metaphors using GPT-3.5, aided by literal collocations and entailment relationship analysis. Similarly, Chen et al. (2024); Tian et al. (2024a) explored the metaphor identification and reasoning capabilities of multiple LLMs. We believe that this problem can be solved by combining multiple sources and GPT-4o to extract literal annotations.

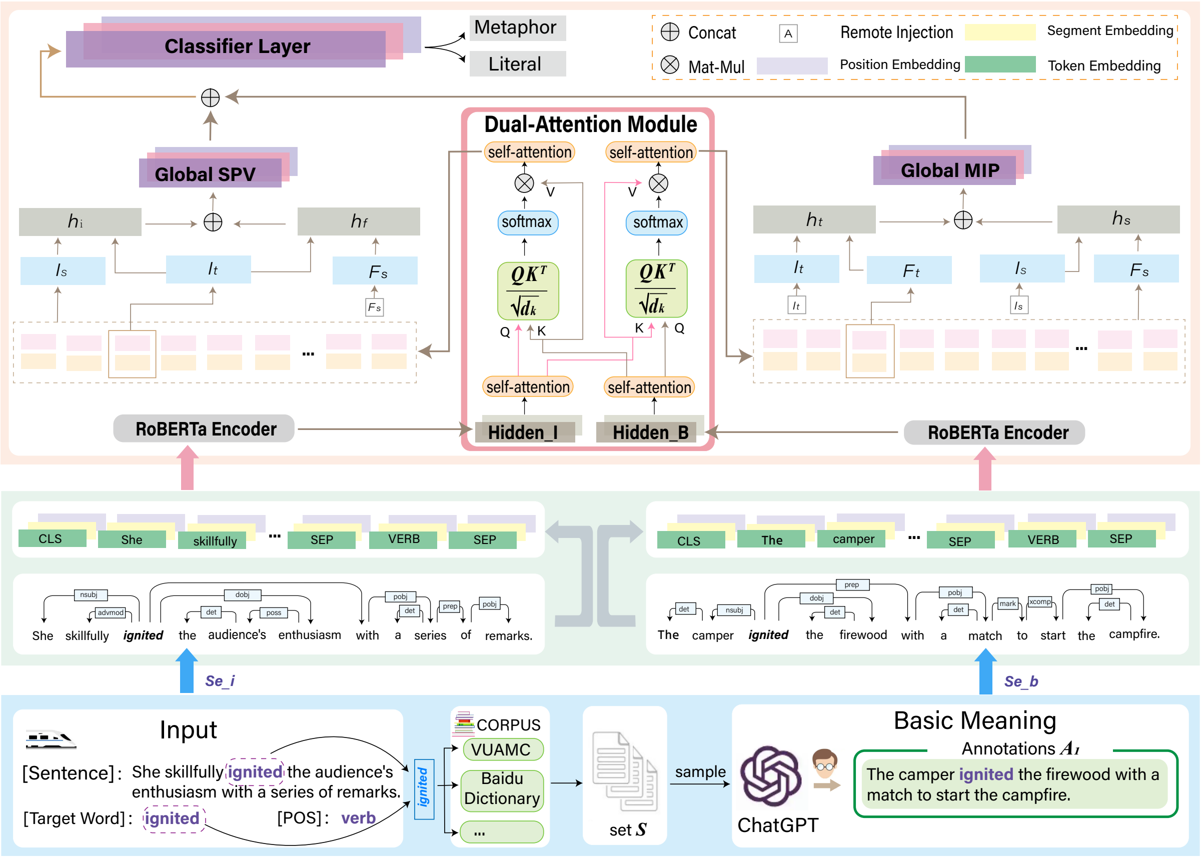


Figure 1: The overall architecture of DAGS, which we use two RoBERTa encoders with shared weights.

3 Method

We propose a new metaphor recognition method, DAGS, that combines dependency-based dual-attention and semantic improvement. The Figure 1 shows an overview of our framework.

3.1 Basic Meaning Acquisition

The VUAMC[1] contains a number of English texts covering a wide range of domains and genres, such as news reports and novels. We use VUAMC as the main source of basic meanings. For Chinese data, the literal annotations are filtered from the Baidu Dictionary[2]. Similarly to (Li et al., 2023a; Tian et al., 2024b), we construct a literal annotations set $S = \{A_1, \ldots, A_n\}$. Then, we use GPT-4o to sample literal annotations of target words from S and manually evaluate them (see Appendix G for details), representing basic meanings.

3.2 Dependency Parsing and Encoding

Inspired by (Wang et al., 2023), our goal is to focus on the context words relevant to target word. First, we set the target word as the root and utilize the spaCy parser (Honnibal and Montani, 2017) to obtain the dependencies of the target word in the context. Next, based on the depth of the tree (the distance between the root and leaves), we prune the parsed sentences. Specifically, we mask the input sentences (retaining words with depth "1") and feed them into the encoder.

The input sentence to be detected is denoted as $Se\_i = \{i_1, \ldots, i_t, \ldots, i_n\}$, and the basic meaning is denoted as $Se\_b = \{f_1, \ldots, f_t, \ldots, f_m\}$. Here, i_t and f_t both represent the same target word. Furthermore, we use the RoBERTa encoder to encode (Enc) the input features to obtain hidden layer outputs *Hidden_I* and *Hidden_B*:

$$Hidden\_I = \text{Enc}([\text{CLS}], Se\_i, [\text{SEP}], \text{POS}) \quad (1)$$

$$Hidden\_B = \text{Enc}([\text{CLS}], Se\_b, [\text{SEP}], \text{POS}) \quad (2)$$

where CLS represents a special classification token, and SEP is a segment separator token. *Hidden_I* is a matrix $\in \mathbb{R}^{n \times k}$, and *Hidden_B* is a matrix $\in \mathbb{R}^{m \times k}$, where k is the hidden size of the encoder.

[1] http://www.vismet.org/metcor/documentation/home.html
[2] https://dict.baidu.com

3.3 Dual-Attention Module

Dual-attention mechanism is able to process two different types of information streams in parallel and automatically learn the correlations between different parts (Vaswani, 2017; Zhao and Gu, 2024; Khan et al., 2023). Inspired by this, we design a novel dual-attention module that processes input information streams in parallel and later integrates them. Specifically, we first compute the intra-stream attention for each flow. Next, we use the cross-attention mechanism to enable interaction between the streams. The prior filtering of dependency trees allows the attention mechanism to focus better on cross-sentence and cross-semantic-level associations. Finally, the module returns to the single-stream processing step and outputs high-quality semantic representations. The calculation formulas are as follows:

$$h^{(l+1)} = \text{Atten}\left(h^{(l)}\right) \odot \text{Cross}\left(h^{(l)}, h_{\text{o}}^{(l)}\right), \quad (3)$$

where $h^{(l)}$ and $h^{(l+1)}$ represent the hidden states of the lth and l+1th layers, respectively, while $h_{\text{o}}^{(l)}$ refers to the hidden state of **other** layer. Furthermore, $\odot$ signifies the macroscopic integration of self-attention and cross-attention mechanisms. Moreover, we analyze the effect of input order in Cross Attention in detail on Appendix I.

3.4 Metaphor Recognition

According to the ouput of dual-attention module, we can obtain the output vectors I_t and F_t for the target word in input sentence and basic meaning, respectively. We then compute the output representations I_s and F_s of the sentence, respectively:

$$I_s = \frac{1}{n} \sum_{j \in P_n} h_j, F_s = \frac{1}{n} \sum_{m \in Q_n} h_m, \quad (4)$$

where P_n represents the "n" words within the range of neighbors, and h_j is the hidden state of the corresponding word, similarly Q_n and h_m.

3.4.1 Global-SPV

Traditional SPV focus only on semantic difference between current target word and a single context, while ignoring the influence of other contexts, which can lead to under-learning of features and metaphorical misclassification. Therefore, we consider semantic features of global contexts. A target word can be recognized as a metaphor when its difference in its context (I_s and I_t) is large and its difference in a specific context (F_s and I_t) is small. We compute the difference representation of target word with different contexts separately:

$$h_{\text{i}} = W_{\text{i}}^{\top} \left[I_s; I_t; |I_s - I_t|; I_s \odot I_t\right] + b_{\text{i}}, \quad (5)$$

$$h_{\text{f}} = W_{\text{f}}^{\top} \left[F_s; I_t; |F_s - I_t|; F_s \odot I_t\right] + b_{\text{f}}, \quad (6)$$

here, w and b represent the weight and bias, respectively. The notation $[\cdot]$ is used to denote the reading method, $|\cdot|$ indicates the absolute value, signifies concatenation, and $\odot$ represents the Hadamard product. Subsequently, we compute the hidden vector $H_{\text{G-SPV}}$ by connecting h_i and h_f:

$$H_{\text{G-SPV}} = g_1([h_i; h_f]), \quad (7)$$

where $H_{\text{G-SPV}} \in \mathbb{R}^{h \times 1}$ and $g_1(.)$ is a function of the learning vector gap in the MLP layer.

3.4.2 Global-MIP

According to (Steen et al., 2010), MIP also requires an understanding of the overall content and context, whereas the current approach focuses only on the relevant meaning of the target word. Therefore, we consider both the semantic differences h_t of the target word and the semantic differences h_s of the overall context to recognize metaphors:

$$h_{\text{t}} = W_{\text{t}}^{\top} \left[F_t; I_t; |F_t - I_t|; F_t \odot I_t\right] + b_{\text{t}}, \quad (8)$$

$$h_{\text{s}} = W_{\text{s}}^{\top} \left[F_s; I_s; |F_s - I_s|; F_s \odot I_s\right] + b_{\text{s}}, \quad (9)$$

where w and b represent the weight and bias, respectively. Similarly, we use h_t and h_s to compute the hidden vector $H_{\text{G-MIP}}$:

$$H_{\text{G-MIP}} = g_2([h_s; h_t]), \quad (10)$$

where $g_2(.)$ also is a learning vector gap function.

3.4.3 Post Computation

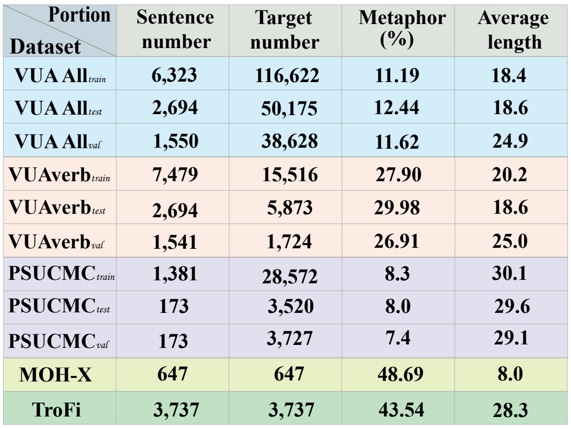
We use $H_{\text{G-SPV}}$ and $H_{\text{G-MIP}}$ to determine whether the target word is used metaphorically:

$$\hat{y} = \sigma(W^{\top}[H_{\text{G-SPV}}; H_{\text{G-MIP}}] + b), \quad (11)$$

where W and b are weights and biases, respectively, and σ is softmax function. $\hat{y} \in \mathbb{R}^2$ represents the predicted label. Finally, we compute the loss L:

$$L = -\sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)], \quad (12)$$

where N is the number of samples in the training set, while y_i and $\hat{y}_i$ are the true and predicted labels of the i-th sample in the training set.



Portion / Dataset	Sentence number	Target number	Metaphor (%)	Average length
VUA All$_{train}$	6,323	116,622	11.19	18.4
VUA All$_{test}$	2,694	50,175	12.44	18.6
VUA All$_{val}$	1,550	38,628	11.62	24.9
VUAverb$_{train}$	7,479	15,516	27.90	20.2
VUAverb$_{test}$	2,694	5,873	29.98	18.6
VUAverb$_{val}$	1,541	1,724	26.91	25.0
PSUCMC$_{train}$	1,381	28,572	8.3	30.1
PSUCMC$_{test}$	173	3,520	8.0	29.6
PSUCMC$_{val}$	173	3,727	7.4	29.1
MOH-X	647	647	48.69	8.0
TroFi	3,737	3,737	43.54	28.3

Figure 2: Sample statistics for dataset.

4 Experiment Design

4.1 Datasets

We use several current mainstream metaphorical datasets with statistics, seeing Figure 2.

- **VUA ALL and VUAverb** (Leong et al., 2020): VUA ALL has been applied to the shared task of metaphor recognition. And the VUAverb is a verb part extracted from VUA ALL.

- **TroFi** (Birke and Sarkar, 2006): TroFi contains literal and metaphorical usage of 50 English verbs from Wall Street Journal corpus.

- **MOH-X** (Mohammad et al., 2016): MOH also focuses on verb metaphors, comprising 1,639 sentences. MOH-X is a subset of MOH.

- **PSUCMC** (Nacey et al., 2019): PSUCMC is composed of text samples from the Lancaster Corpus of Mandarin Chinese.

4.2 Baseline

RoBERTa_SEQ (Leong et al., 2020) is used in the VUA2020 shared task. And **DeepMet** (Su et al., 2020) incorporates multiple linguistic features into RoBERTa. **MelBERT** (Choi et al., 2021) detect metaphor by interactively computing the outputs of MIP and SPV. For **MrBERT** (Song et al., 2021), metaphors are recognized by extracting dependency relationships in sentences and embedding. **CATE** (Lin et al., 2021) is to increase the distance between literal and metaphorical meanings of target word. While **MDGI** (Wan et al., 2021) explains metaphors by annotating them. **MisNet** (Zhang and Liu, 2022) model transforms MIP and SPV into a semantic matching task, and calculates the similarity. **MRW** (Babieno et al., 2022) collects dictionary definitions to extract non-metaphorical word meanings. Then, **AAAS** (Feng and Ma, 2022) model transforms a categorization task into a keyword extraction to capture metaphor features. Instead, **RoPPT** (Wang et al., 2023) focuses on semantically relevant information. **AdMul** (Zhang and Liu, 2023) migrates basic sense discrimination (BSD) knowledge to metaphor recognition. And **BasicBERT** (Li et al., 2023a) models the basic meanings and compares them with contextual meanings to identify metaphors. Additionally, **FrameBERT** (Li et al., 2023b) incorporates FrameNet. In **ContrastWSD** (Elzohbi and Zhao, 2024), the contextual and basic meanings are extracted by using WSD and analyzed in comparison. Also to the **ER** (Uduehi and Bunescu, 2024), the target word representation is obtained by constructing different components. While **MiceCL** (Jia and Li, 2024) utilizes sentence external differences to better handle semantic relations. **QMM** (Qiao et al., 2024) does a fine-grained match recognition by modeling uncertainty in literal meanings through density matrices. What's more, **CKEMI** (Wang et al., 2025) further designs graph networks with concept mapping functions to detect metaphors.

4.3 Implementation Details

We set the learning rate uniformly to 3e-5 and the dropout to 0.2. The learning rate gradually increases from 0 to 3e-5 during two training epochs and decreases linearly until the last epoch. All experiments use the cross-entropy loss function with weights set to 3. The training process uses the AdamW optimizer, and the number of epoch for each set of experiments is 20. Furthermore, we set different batch sizes for each dataset: 100 for VUA ALL dataset, 50 for VUAverb and PSUCMC, 20 for TroFi, and 8 for MOH-X. Three sets of random seeds are set for each set of experiments, and the final result is the average of the three experiments. For TroFi's supervised assessment, we provide details in Appendix H. All experiments use eight NVIDIA RTX A6000 GPUs for computation.

5 Experiment Overall Results

Table 1 presents the performance of DAGS and other models on mainstream datasets. We primarily focus on the core metric F1. Clearly, DAGS is superior to other strong baselines. Compared to earlier RoBERT_SEQ that did not incorporate MDT,

Model (Source Year)	**VUA ALL**			**VUAverb**			**TroFi**			**PSUCMC**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
RoBERTa_SEQ (Fig-Lang 2020)	80.4	74.9	77.5	79.2	69.8	74.2	-	-	-	71.6	73.6	72.6
DeepMet (Fig-Lang 2020)	82.0	71.3	76.3	79.5	70.8	74.9	-	-	-	73.8	73.2	73.5
MrBERT (ACL 2021)	**82.7**	72.5	77.2	*80.8*	71.5	75.9	-	-	-	-	-	-
MelBERT (NAACL 2021)	80.1	76.9	78.5	78.7	72.9	75.7	-	-	-	79.6	76.4	77.9
MRW (Applied Sciences 2022)	79.3	78.5	78.9	60.9	77.7	68.3	53.2	72.8	61.4	-	-	-
CATE (EMNLP 2021)	79.3	78.8	79.0	78.1	73.2	75.6	*74.4*	74.8	74.5	-	-	-
MDGI (ACL 2021)	*82.5*	72.5	77.2	78.9	70.9	74.7	-	-	-	**89.0**	70.6	78.7
MisNet (COLING 2022)	79.8	77.0	78.7	78.9	71.1	73.4	-	-	-	79.2	70.6	73.8
AAAS (EMNLP 2022)	81.6	77.4	79.4	**81.6**	71.1	76.0	72.5	77.5	*74.8*	-	-	-
FrameBERT (EACL 2023)	**82.7**	75.3	78.8	-	-	-	70.7	*78.2*	74.2	-	-	-
AdMul (ACL 2023)	78.4	*79.5*	79.0	78.5	78.1	*78.3*	70.5	**79.8**	74.7	-	-	-
RoPPT (EACL 2023)	80.0	78.2	79.1	-	-	-	-	-	-	79.3	*79.0*	*79.1*
BasicBERT (ACL 2023)	79.1	77.7	78.3	76.7	77.5	76.8	-	-	-	75.0	74.9	75.0
ER (Fig-Lang 2024)	80.2	77.5	78.8	-	-	-	72.2	73.5	72.8	-	-	-
MiceCL (NAACL 2024)	80.4	75.2	78.5	75.1	78.0	75.9	-	-	-	-	-	-
ContrastWSD (COLING 2024)	75.5	72.9	74.2	79.1	66.9	72.5	-	-	-	-	-	-
QMM (COLING 2024)	80.9	77.8	79.3	73.9	*79.0*	76.4	-	-	-	-	-	-
CKEMI (IPM 2025)	80.9	78.7	*79.8*	-	-	-	-	-	-	-	-	-
DAGS (our)	81.7	**80.6**	**80.9**	80.6	**80.0**	**80.4**	**76.9**	76.1	**76.1**	*80.4*	**80.0**	**80.1**

Table 1: The overall results of DAGS and other baseline models on VUA ALL, VUAverb, and TroFi datasets (Best is in **Bold**, followed by *italic underlined*). The "-" indicates that the original paper did not conduct this experiment.

DAGS significantly improves performance (e.g., on F1, 3.4%, 6.2%, and 7.5% on VUA ALL, VUAverb, PSUCMC respectively). MelBERT, which integrates both MIP and SPV, achieves the 78.5% (-2.4% vs. DAGS). This indicates that metaphor detection theories have some efficacy. When compared with CKEMI, our model demonstrates better filtering capabilities (increasing 1.1% on VUA ALL). Moreover, on VUA ALL, DAGS shows improvements of 2.2% and 1.5% over the current strong baselines MisNet and AAAS, respectively. In VUAverb, which contains more complex verbs including auxiliary verbs and linking verbs, the prediction task is more challenging. While DAGS captures the most relevant information about the verb through a dependency-based dual-attention module , and improves the semantics. In comparison, DAGS outperforms the AdMul baseline by 2.1%. This result underscores the importance of acquiring key information from basic meanings to enhance metaphor recognition performance. Furthermore, DAGS is not only applicable to English dataset, but also shows strong competition on Chinese dataset (e.g., 80.1% on F1). We calculate the p-value differences between DAGS and other baselines (e.g., for F1, DAGS vs. MisNet with a p-value of 0.0001 on VUA ALL, DAGS vs. AdMul with a p-value of < 0.0001 on VUAverb). This suggests that DAGS is more effective than other methods.

6 POS Experiments

For VUA ALL, we conduct fine-grained experiments based on POS (Adjective, Verb, Noun, and Adverb). Table 2 shows the results, where DAGS consistently achieves the best performance on F1. Compared to the current strong baseline ContrastWSD, DAGS realizes improvements of 3.7% (Adjective), 1.4% (Noun), 3.2% (Verb), and 2.5% (Adverb). These results indicate that, regardless of the complexity of POS, simultaneously filtering and semantic improvement of both input sentences and basic meanings can further improve the model's recognition performance. Except for the POS, we achieve promising results in another breakdown experiment (see Appendix A and Table 7).

7 Zero-Shot Transfer

We design zero-shot transfer experiments across English datasets. Specifically, we train on VUA ALL and test on entire TroFi and MOH-X.

The results are shown in Table 3. It can be found that DAGS not only performs well in supervised experiments, but also achieves the best in zero-shot. In TroFi, the precision (Prec) of DAGS is higher than other models, reflecting its strong prediction ability for positive classes. The same trend is also

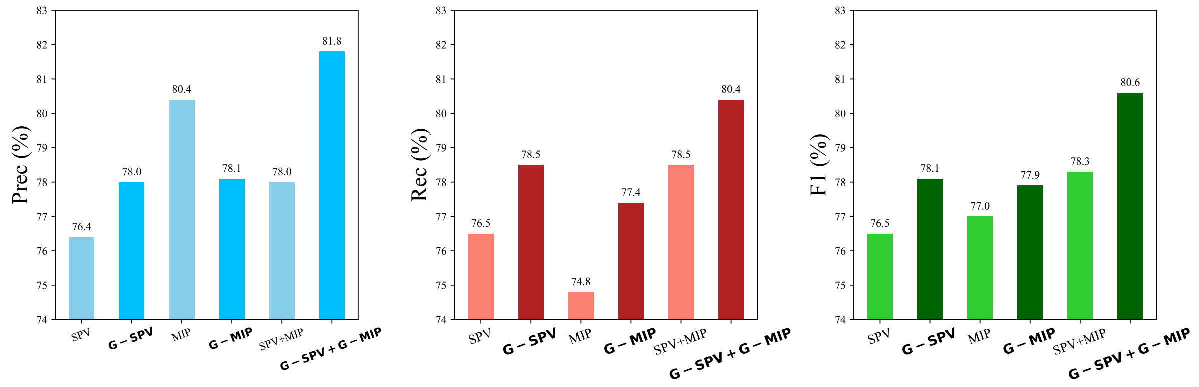


Figure 3: Semantic improvement visualization.

Model	Adjective			Noun			Verb			Adverb		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
DeepMet	**79.0**	52.9	63.3	**76.5**	57.1	65.4	*78.8*	68.5	73.3	*79.4*	66.4	72.3
MelBERT	69.4	60.1	64.4	*75.4*	66.5	70.7	78.7	72.9	75.7	**80.2**	69.7	*74.6*
MisNet	67.9	65.5	66.4	73.8	68.4	69.5	76.7	*78.0*	77.1	75.9	71.2	64.1
MiceCL	68.5	68.9	*68.8*	70.6	70.7	70.7	75.1	*78.0*	75.9	73.3	71.3	72.2
ContrastWSD	65.7	*70.8*	68.1	66.2	**76.3**	*70.9*	*78.8*	75.7	*77.2*	68.5	**77.4**	72.6
DAGS (our)	*72.2*	**71.6**	**71.8**	72.2	*72.5*	**72.3**	**80.6**	**80.0**	**80.4**	74.6	*75.8*	**75.1**

Table 2: The POS results for Adjective, Noun, Verb and Adverb (Best is in **Bold**, followed by *italic underlined*).

Model	VUA to TroFi			VUA to MOH-X		
	Prec	Rec	F1	Prec	Rec	F1
DeepMet	53.7	73.9	60.7	*79.9*	76.5	77.9
MrBERT	53.8	75.0	62.7	75.9	**84.1**	79.8
MelBERT	53.4	74.1	62.0	79.3	79.7	79.2
MisNet	52.9	*75.1*	62.1	-	-	-
RoPPT	*54.2*	**76.2**	*63.3*	77.0	*83.5*	*80.1*
MiceCL	*54.2*	75.0	62.9	-	-	-
DAGS (our)	**55.3**	73.5	**64.1**	**81.5**	80.4	**80.6**

Table 3: Transfer performance on TroFi and MOH-X (Best is in **Bold**, followed by *italic underlined*).

Model	VUAverb			TroFi		
	Prec	Rec	F1	Prec	Rec	F1
w/o DAM	76.0	76.3	76.1	73.2	72.0	72.3
w/o CP	*79.3*	77.6	*78.9*	*73.9*	*73.5*	*73.6*
w/o G-SPV	*77.4*	78.1	*77.9*	72.7	72.8	72.7
w/o G-MIP	78.0	*78.5*	78.1	72.3	72.2	72.2
DAGS (our)	**80.6**	**80.0**	**80.4**	**76.9**	**76.1**	**76.1**

Table 4: Ablation experiment performance of Context Pruning (CP), Dual-Attention Module (DAM), Global-SPV (G-SPV) and Global-MIP (G-MIP). The "w/o" indicates that the part is removed (Best is in **Bold**, followed by *italic underlined*).

observed in MOH-X (e.g., DAGS vs. RoPPT with a p-value of < 0.001 on Prec). In addition, DAGS also reaches the optimum on F1, demonstrating its strong robustness, especially in its migration ability on small-scale datasets. Noting that although models perform relatively close to each other on two datasets, none of them outperforms DAGS (e.g., RoPPT and MiceCL). This further proves DAGS's excellent generalization ability on different datasets. Addition to same-language transfer, DAGS has shown strong competitiveness in cross-language (see Appendix F and Table 11).

8 Semantic Improvement

To evaluate the effectiveness of semantic improvement, we conduct comparative experiments on VUAverb. For each set of comparative results, we provide a data visualization, which is shown in Figure 3. G-SPV and G-MIP show significant enhancement in both models. For example, for DAGS, G-SPV enhances F1 from 76.5% to 78.1%, while G-MIP improves it from 77.0% to 77.9%. In addition, G-SPV and G-MIP not only have independent effects in terms of their respective semantic improvement, but also show synergistic effects. These

Length Range	DAGS			RoPPT			CKEMI			BasicBERT		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
< 20 Tokens	75.9	78.1	**77.8**	76.4	74.8	75.6	77.2	76.8	*76.9*	77.3	76.3	76.7
20 - 40 Tokens	81.5	81.0	**81.2**	81.8	79.9	*80.8*	80.4	79.9	80.1	75.7	75.1	75.6
> 40 Tokens	82.8	84.0	**82.9**	82.3	80.0	81.1	81.5	81.3	*81.4*	75.4	74.7	75.2

Table 5: The results of models on different length range sentences (Best is in **Bold**, followed by *italic underlined*).

results suggest that incorporating global semantics can further utilize the semantic information, especially in complex language environments.

9 Ablation

The ablation results for each module are shown in Table 4. DAGS consistently performs best. When the DAM and CP modules are removed, the performance metrics of the model on the VUAverb dataset drop to 76.1% and 78.9%, respectively. This phenomenon may stem from the fact that model learns a number of redundant features without filtering context. A similar trend is observed on TroFi, indicating that DAGS is able to effectively improve the overall performance of the model by filtering the interference and ambiguity in the input data. In addition, further removal of G-SPV and G-MIP modules also significantly degraded the model performance, which verifies the necessity of incorporating more global contextual semantics, consistent with the results in Figure 3 of semantic improvement experiments.

10 Sentence Length Experiments

Sentence length variations affect the focus of model's attention and interference degree. To explore this phenomenon in depth, we design experiments to analyze the effect of sentence length on model performance. As shown in Table 5, the performance of DAGS progressively improves with increasing sentence length and is consistently better than RoPPT and CKEMI. Longer sentences tend to contain more redundant information and noise, which constitutes a significant interference in the metaphor recognition task, verifying the importance of focusing on task-relevant semantics. The results show that DAGS can effectively filter out irrelevant contexts in sentences and basic meanings, and its performance improvement shows a significant positive correlation with sentence length (e.g., F1 with 82.9% > 81.2% > 77.8%).

Target Word	Sentence
back	He leaned back on the more beautiful and more comfortable cushion, which was placed conveniently near the window.
bogged	A major initiative aimed at enhancing the city's public transportation network to include more eco-friendly options became bogged down.

Table 6: Case Study. The red indicates the target word, while the blue shows the most relevant word.

11 Case Study

See Table 6, previous methods may be influenced by redundant context that not filtered. For instance, the aggregated meaning of the word "back" is "the posterior surface of the human body", which is not its basic meaning; terms like "back up" and "back on" occur more frequently in the corpora (Li et al., 2023a). However, relying solely on literal annotations may cause the model to fail in recognizing the key information of target word. For example, in the sentence "*He leans back on the more beautiful and more comfortable cushion, which is placed conveniently near the window.*", the target word "back" is followed by a comparative phrase, which might lead the model to learn an incorrect representation. Similarly, in "*A major initiative aimed at enhancing the city's public transportation network to include more eco-friendly options becomes bogged down.*", the target word "bogged" is separated from the subject by a long phrase, which might prevent the model from identifying the metaphor. DAGS filters both the input sentence and basic meanings, and extracts key information about the target word, resulting in high-precision processing.

12 Conclusion

In this paper, we propose a novel metaphor recognition model, DAGS, which introduces dependency-based dual-attention module to filter context and focus on key content. And then, DAGS further enhances metaphor recognition through global semantic. Experiments show that DAGS achieves state-of-the-art performance on metaphor recognition compared to existing models.

Limitations

Experimental results show that our method exhibits superior results in the metaphor recognition task, a result that is in line with our expectations. To contrast with previous work, our generalization experiments are based on the VUA ALL and TroFi datasets. However, the TroFi dataset is older and some of the metaphors may have been transformed into literal meanings. Although DAGS achieves the best results in comparison to other baseline models, there is still significant room for improvement. We believe that utilizing external knowledge (e.g., dictionaries) may be helpful for generalization experiments. This is something we plan to explore further in future research.

Ethics Statement

In this study, we strictly adhere to academic and research ethics, emphasizing transparency and openness. We explicitly cite public data sources to respect the original authors and data providers in metaphor recognition research. Throughout our research, we do not intentionally criticize or plagiarize others' work, aligning with the principles of academic integrity. We prioritize authenticity, transparency, and fairness in every step of the research process. We believe this commitment will positively contribute to the growth of the academic community.

Acknowledgments

We would like to thank all reviewers for their insightful comments and suggestions to help improve the paper. This work was supported by the National Natural Science Foundation of China (62276072), Guangxi Natural Science Foundation Key Project (2025GXNSFDA069017), the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology, and Innovation Project of Guangxi Graduate Education (JGY2023016).

References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. Applied Sciences, 12(4):2081.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In 11th Conference of the European chapter of the association for computational linguistics, pages 329–336.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. IEEE Intelligent Systems, 32(6):74–80.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6455–6469, Online. Association for Computational Linguistics.

Puli Chen, Cheng Yang, and Qingbao Huang. 2024. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 5850–5860.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. 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), pages 5968–5978, Online. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. 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, pages 1763–1773, Online. Association for Computational Linguistics.

Mohamad Elzohbi and Richard Zhao. 2024. ContrastWSD: Enhancing metaphor detection with word sense disambiguation following the metaphor identification procedure. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3907–3915, Torino, Italia. ELRA and ICCL.

Huawen Feng and Qianli Ma. 2022. It's better to teach fishing than giving a fish: An auto-augmented structure-aware generative model for metaphor detection. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 656–667.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. Metaphor and symbol, 22(1):1–39.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear, 7(1):411–420.

Kaidi Jia and Rongsheng Li. 2024. 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), pages 2726–2737.

Zulfiqar Ahmad Khan, Tanveer Hussain, and Sung Wook Baik. 2023. Dual stream network with attention mechanism for photovoltaic power forecasting. Applied Energy, 338:120916.

Luuk Lagerwerf and Anoe Meijers. 2008. Openness in metaphorical and straightforward advertisements: Appreciation effects. Journal of Advertising, 37(2):19–30.

George Lakoff and Mark Johnson. 2008. Metaphors we live by. University of Chicago press.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In Proceedings of the second workshop on figurative language processing, pages 18–29.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2022a. The secret of metaphor on expressing stronger emotion. In Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 39–43, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022b. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling. In Proceedings of the 29th international conference on computational linguistics, pages 6468–6479.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023a. Metaphor detection via explicit basic meanings modelling. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023b. FrameBERT: Conceptual metaphor detection with frame embedding learning. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1558–1563, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. Cate: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In Proceedings of the 2021 conference on empirical methods in natural language processing, pages 3888–3898.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In Proceedings of the 56th annual meeting of the association for computational linguistics. Association for Computational Linguistics (ACL).

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In Proceedings of the fifth joint conference on lexical and computational semantics, pages 23–33.

Susan Nacey, W Gudrun Reijnierse, Tina Krennmayr, and Aletta G Dorst. 2019. Metaphor identification in multiple languages.

Wenbo Qiao, Peng Zhang, and ZengLai Ma. 2024. A quantum-inspired matching network with linguistic theories for metaphor detection. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1435–1445.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. 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), pages 4240–4251.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. A method for linguistic metaphor identification: From MIP to MIPVU, volume 14. John Benjamins Publishing.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In Proceedings of the second workshop on figurative language processing, pages 30–39.

Yuan Tian, Nan Xu, and Wenji Mao. 2024a. 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), pages 7731–7748.

Yuan Tian, Ruike Zhang, Nan Xu, and Wenji Mao. 2024b. 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), pages 13311–13325.

Oseremen Uduehi and Razvan Bunescu. 2024. An expectation-realization model for metaphor detection. In Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), pages 79–84, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. Advances in Neural Information Processing Systems.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1018–1032.

Hai Wan, Jinxia Lin, Jianfeng Du, Dawei Shen, and Manrong Zhang. 2021. Enhancing metaphor detection by gloss-based interpretations. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1971–1981.

Dian Wang, Yang Li, Suge Wang, Xin Chen, Jian Liao, Deyu Li, and Xiaoli Li. 2025. Ckemi: Concept knowledge enhanced metaphor identification framework. Information Processing & Management, 62(1):103946.

Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. 2023. Metaphor detection with effective context denoising. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1404–1409, Dubrovnik, Croatia. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. Artificial intelligence, 6(1):53–74.

Cheng Yang, Puli Chen, and Qingbao Huang. 2024. 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), pages 1016–1027.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced siamese network. In Proceedings of the 29th International Conference on Computational Linguistics, pages 4149–4159.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. In Findings of the Association for Computational Linguistics: ACL 2023, pages 1483–1497, Toronto, Canada. Association for Computational Linguistics.

Yimin Zhao and Jin Gu. 2024. Feature fusion based on mutual-cross-attention mechanism for eeg emotion recognition. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 276–285. Springer.

A VUA All Breakdown

The VUA ALL dataset not only supports fine-grained analysis based on parts-of-speech (POS) but also encompasses various domain types, including Academic, Conversation, Fiction, and News.

In Table 7, although DeepMet achieves an impressive precision of 88.4% in the Academic domain, its recall rate is somewhat lacking, leading to a lower overall F1 score. Instead, DAGS, with its higher recall rate, secures an F1 score of 85.2%, demonstrating a more balanced performance. For Conversation, DAGS excels, particularly in handling colloquial and fragmented sentences, achieving an F1 score of 75.1%, significantly surpassing other models. This highlights its remarkable capability in dealing with more informal or irregular language structures. To the Fiction, where unique linguistic structures and proper names often pose challenges, DAGS outperforms other models with an F1 score of 76.9%, showcasing its superiority in processing complex texts. Furthermore, in News domain, although DAGS's performance is close to that of BasicBERT, it still achieves the highest F1 score of 82.0%, demonstrating its ability to handle formal written language while maintaining competitive performance in news-related texts. Overall, the experimental results indicate that DAGS strikes a good balance between precision and recall, particularly excelling in more complex text types such as conversations and fictional works.

B Domain Transfer

According to the four domains classified by VUA ALL, there is a large gap in supervised testing performance of the models across different domains. We further explore the model's adaptability to each domain. Specifically, we let the model be trained on VUA ALL, and then separately on Academic, Conversation, Fiction, and News domains for testing.

The experimental results are shown in Table 8. Compared with Table 7, most models perform best in the Academic domain, achieving higher F1 values for both single-domain tests and cross-domain transfer, which may be due to the relatively explicit use of metaphors in academic language. Despite the relatively more complex data in the Conversation and Fiction categories, most of the models are trained with VUA ALL instead, which may be the VUA ALL training set is able to cover different types of metaphorical structures in conversations, both in terms of data type and quantity, than the Conversation subset alone. In addition, as a whole, the DAGS model excels in all domains, especially in Academic and News, with stable performance and strong domain transfer capability. In contrast, MelBERT and RoPPT perform better on single domains, but are not as robust as DAGS in the more challenging domains of Conversation and Fiction.

C Target Word Depth Range Experiment

We assume that other words with different ranges of the target word in the sentence will interfere with the model to varying degrees. The "*depth*" is defined as the distance between the target word and the root node in the dependency tree. Based on the dependency tree parsing results, we categorize it into four levels, which are "$depth = 1$" to "$depth = 4$", with larger values representing the more range word we introduce. We conduct experiments on the VUA ALL and TroFi datasets.

The experimental results are shown in Figure 4. We observe an increase in the F1 score as "*depth*" moves from 0 to 1. At "$depth = 0$", where only aggregated meanings are used, the model fails to capture relevant conceptual information. However, starting from "$depth = 2$", the model's performance begins to decline, which may be due to the redundancy introduced by longer texts. In other words, deeper syntactic structures add more noise and complexity, making it difficult for the model to focus on relevant features. Key contextual information and features are more beneficial for the model in detecting metaphors, as they help the model effectively identify and distinguish metaphors.

D Stand-alone Modular Ablation

In addition to ablation experiments on DAGS for module removal, we also evaluate the performance of each module when used independently and compare it to the full model. The results of the experiments are shown in Table 9. DAGS continues to exhibit the best performance. It is worth noting that the performance of the DAM module on both datasets is quite competitive with existing strong baselines. Furthermore, the optimal performance of DAGS shows that relying solely on metaphor detection theory is not sufficient to achieve optimal results, and the introduction of the dual-attention module significantly improves the model's performance. The dual-attention module reduces redundancy by dynamically assigning weights to make

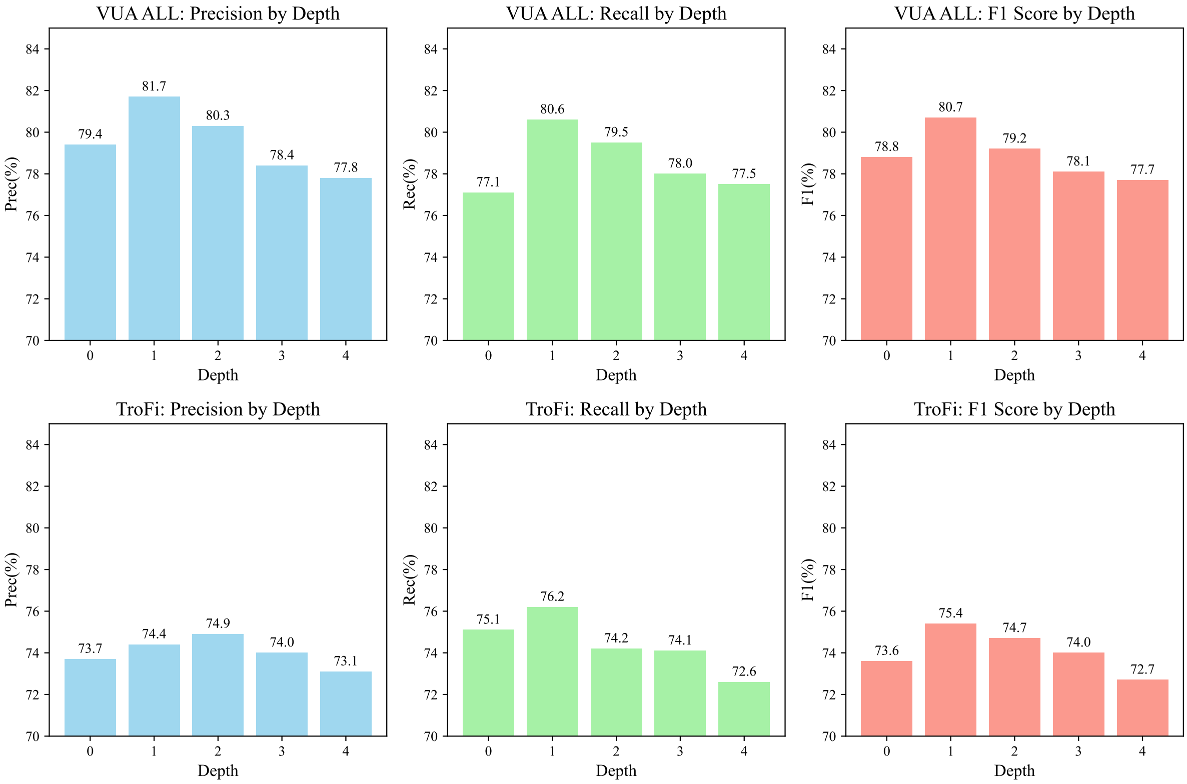


Figure 4: Effect of different sentence depth on DAGS.

Model	Academic			Conversation			Fiction			News		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
DeepMet	**88.4**	74.7	81.0	71.6	71.1	71.4	*76.1*	70.1	73.0	**84.1**	67.6	75.0
MelBERT	85.3	82.5	83.9	70.1	71.7	70.9	74.0	76.8	75.4	81.0	73.7	77.2
MisNet	83.4	81.2	82.1	70.6	72.8	71.1	74.1	**78.0**	*75.7*	*83.1*	75.4	78.7
BasicBERT	*85.4*	**85.5**	**85.4**	70.9	68.9	69.8	73.5	73.3	73.4	81.8	*82.1*	*81.9*
RoPPT	85.1	85.0	85.0	*74.7*	*73.1*	*73.7*	73.4	73.8	73.5	80.5	80.2	80.4
MiceCL	84.5	83.9	84.2	73.8	72.1	72.8	72.9	73.6	73.1	78.4	78.6	78.5
DAGS (our)	85.2	*85.3*	*85.2*	**75.0**	**75.3**	**75.1**	**76.8**	*77.3*	**76.9**	81.3	**82.3**	**82.0**

Table 7: VUA All Breakdown Experiment results. Breakdown genres include Academic, Conversation, Fiction, and News. The metrics in each interval include Precision (Prec), Recall (Rec), and F1-score (F1) (Best is in **Bold**, followed by *italic underlined*).

Model	VUA to Acad			VUA to Conv			VUA to Fict			VUA to News		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
MelBERT	83.7	83.3	83.5	75.7	75.2	75.4	74.6	74.2	74.4	73.0	73.2	73.1
MisNet	81.8	82.0	81.9	74.0	72.8	73.3	74.0	73.7	73.8	79.9	80.1	80.0
BasicBERT	83.1	83.3	83.1	74.5	74.9	74.7	74.1	74.3	74.2	81.1	*81.5*	81.4
RoPPT	**84.4**	*84.6*	*84.5*	**76.1**	*75.9*	*76.0*	*75.1*	*75.1*	*75.1*	*83.1*	81.3	81.1
MiceCL	82.6	82.8	82.7	74.3	72.5	74.2	74.3	73.8	73.8	81.8	80.9	*81.5*
DAGS (our)	*84.2*	**85.5**	**84.9**	*76.0*	**76.7**	**76.4**	**75.7**	**75.4**	**75.4**	**83.5**	**83.7**	**83.5**

Table 8: The model's performance on domain transfer. The genres include Academic (**Acad**), Conversation (**Conv**), Fiction (**Fict**), and **News**, where "**VUA**" stands for VUA ALL. And the metrics in each interval include Precision (Prec), Recall (Rec), and F1-score (F1) (Best is in **Bold**, followed by *italic underlined*).

the model more focused on task-relevant information, while reducing the computational burden. Further analysis reveals that the G-MIP and G-SPV modules also perform relatively well (e.g., 1.5% and 1.7% improvement over CP on the VUAverb dataset, respectively), which suggests that it is beneficial to consider global semantic information in context. The experimental results further validate the findings of the module removal ablation experiments. We design DAGS to achieve more efficient module synergy by integrating modules such as DAM, G-SPV and G-MIP.

Model	VUAverb			TroFi		
	Prec	Rec	F1	Prec	Rec	F1
DAM	*79.5*	*77.7*	*78.2*	*73.7*	*73.2*	*73.3*
CP	75.4	76.2	75.7	70.6	71.2	71.0
G-SPV	77.7	77.2	77.2	71.6	71.7	71.6
G-MIP	77.4	77.4	77.4	72.2	71.0	71.4
DAGS (our)	**80.6**	**80.0**	**80.4**	**76.9**	76.1	**76.1**

Table 9: Model performance on Stand-alone Modular Ablation Experiment (Best is in **Bold**, followed by *italic underlined*), including Context Pruning (CP), Dual-Attention Module (DAM), Global-SPV (G-SPV) and Global-MIP (G-MIP).

E Word Frequency Distribution

To explore the impact of different target word frequencies on model performance, we partition the VUAverb dataset based on target word frequency (< 100, 100 - 200, 200 - 300, > 300) and conduct experiments with several models. The experimental results are shown in Table 10. Compared to the overall results of VUAverb (see Table 1), the performance of each model varies across different frequency ranges. Among them, the DAGS model consistently performs well across all frequency ranges, especially in the low-frequency range (< 100) and the high-frequency range (200 - 300), achieving F1-scores of 82.3% and 88.3%, respectively, demonstrating strong generalization ability. In the high-frequency range (> 300), BasicBERT has a slight advantage with an F1-score of 84.5%, but DAGS maintains stable performance on other metrics. The RoPPT model excels in the high-frequency range but shows some shortcomings in the low-frequency range, while MelBERT and BasicBERT perform relatively average in the medium to low-frequency ranges. The overall performance of RoBERTa_SEQ is weaker, particularly struggling with low-frequency word handling.

In addition, we further analyze DAGS for its different performance on low-frequency words and high-frequency words. Low-frequency target words are often diverse, and DAGS shows strong generalization ability in cross-word frequency scenarios and can effectively adapt to the detection task in different word frequency intervals. In contrast, high-frequency target words occur frequently in the training data, and the feature boundaries between their metaphorical and literal usages may be more ambiguous, and some models may tend to rely on memory and thus have a slight advantage in high-frequency intervals. Although DAGS is slightly inferior to the best performing model in the detection of high-frequency target words, it is still firmly in the second place, demonstrating strong robustness.

F Cross-Language Transfer

In previous studies, all our experiments are based on the same language. To explore the performance of the DAGS model in cross-language metaphor recognition tasks, we design and implement zero-shot migration experiments across language datasets. Specifically, we train on the VUA ALL and VUAverb datasets and test on the PSUCMC dataset, respectively; in addition, training on the PSUCMC dataset and testing on the VUAverb and TroFi datasets. We select a variety of cross-language pre-trained models (e.g., mBERT, mDeBERTa, mRoBERTa, and XLM-RoBERTa), as well as a strong baseline model of metaphor recognition. The experimental parameter settings are kept consistent with the previous experiments.

The results of the experiments are presented in Table 11. It can be seen that DAGS outperforms all other models on several cross-linguistic metaphor recognition tasks, e.g., in terms of recall (70.2%) and F1 value (61.4%). BasicBERT, although slightly higher in terms of precision (53.3%), has a lower F1 value than DAGS. In comparison to the supervised experiments (see Table 1) the model's performance decrease more (e.g., on VUA ALL, from 80.9% to 61.4% on DAGS and from 78.3% to 57.9% on BasicBERT). This performance degradation may be attributed to the differences in cultural context and metaphorical structure between English and Chinese, resulting in the features learned by the model in the English corpus not being able to be directly and efficiently migrated to the Chinese corpus. In addition, most purely pre-trained

Model	< 100			100 - 200			200 - 300			> 300		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
RoBERTa_SEQ	70.7	73.7	71.9	71.2	72.1	71.4	75.3	75.2	75.2	80.7	75.3	77.7
MelBERT	76.3	74.4	75.3	63.6	**76.7**	69.5	80.0	73.3	76.8	79.3	72.7	75.8
BasicBERT	80.7	80.9	80.8	**74.8**	73.7	*74.0*	87.6	89.5	88.2	**84.1**	*85.2*	**84.5**
RoPPT	*81.4*	79.7	80.2	73.7	73.8	73.7	**89.0**	*90.1*	**88.6**	79.5	82.8	79.8
ER	80.5	*81.8*	*81.1*	72.5	72.1	72.3	86.7	89.5	87.1	78.5	79.4	78.9
MiceCL	79.7	80.3	79.9	70.7	70.0	70.2	83.4	82.6	83.0	80.1	79.9	80.0
DAGS (our)	**82.2**	**82.7**	**82.3**	*74.6*	*74.7*	**74.6**	*87.9*	**90.7**	*88.3*	*83.4*	**85.9**	*83.2*

Table 10: The performance of the model on different intervals of target word frequency, including < 100, 100 - 200, 200 - 300, and > 300. The metrics in each interval include Precision (Prec), Recall (Rec), and F1-score (F1) (Best is in **Bold**, followed by *italic underlined*).

Model	ALL to PSU			Verb to PSU			PSU to Verb			PSU to TroFi		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
mBERT	34.2	38.1	36.5	41.2	53.1	49.9	49.0	60.0	51.6	31.8	56.4	40.7
mDeBERTa	38.7	53.7	48.1	*58.2*	38.0	46.0	53.7	56.8	54.8	53.9	53.8	53.8
mRoBERTa	34.6	52.6	44.6	47.5	53.0	50.7	55.2	56.4	55.2	51.2	*57.2*	54.1
XLM-RoBERTa	38.1	51.5	47.1	55.2	56.0	55.6	56.4	57.4	56.4	53.3	53.4	53.3
MelBERT	50.4	68.3	57.0	55.7	61.6	55.9	54.7	*69.5*	57.7	50.3	51.3	50.5
MisNet	50.1	67.9	*58.8*	52.5	68.9	57.6	55.1	66.4	56.5	52.2	54.3	52.7
BasicBERT	*53.3*	*69.4*	57.9	51.2	*70.1*	57.1	57.3	68.7	58.3	*54.2*	55.3	*54.2*
RoPPT	47.7	69.0	56.4	*61.8*	56.9	*58.4*	59.1	58.2	58.8	50.5	49.2	49.4
ER	47.0	68.7	55.7	**65.4**	52.2	53.2	*59.4*	67.7	*59.1*	51.4	55.4	53.1
MiceCL	**57.3**	63.6	57.5	57.3	52.7	54.0	58.2	**69.7**	58.6	51.7	55.2	52.9
DAGS (our)	52.7	**70.2**	**61.4**	50.0	**70.8**	**60.6**	**59.8**	66.7	**60.3**	**57.2**	**59.4**	**57.5**

Table 11: Model performance on cross-language transfer. Here, "**ALL**" indicates the VUA ALL dataset, "**Verb**" represents the VUAverb dataset, and "**PSU**" stands for the PSUCMC dataset (Best is in **Bold**, followed by *italic underlined*).

language models (PLMs) perform poorly in cross-lingual metaphor recognition tasks, which may be due to their failure to adequately capture the more complex semantic features behind the metaphors. In contrast, DAGS is able to effectively consider and capture key metaphor features by employing a dual-attention module and semantic improvement strategies (e.g., G-SPV and G-MIP). The experimental results show that the cross-language metaphor recognition task is highly challenging. Although the model performs well in the recognition task in a single language, it still suffers from insufficient generalization ability in the cross-language migration task.

In addition, we further consider cross-language experimental designs for low-resource languages, such as Slovene (KOMET dataset). For this, we conduct transfer experiments between English-Slovene and Chinese-Slovene respectively, and the results are shown in Table 14. It can be seen that DAGS (our) significantly outperforms the other models on both tasks, suggesting that the approach is able to integrate cross-corpus features more efficiently, thus improving the migration performance. Meanwhile, MiceCL and ER also show strong generalization ability, while mBERT, mRoBERTa and XLM-RoBERTa, as the base models, perform relatively weakly in the cross-dataset task.

G Manual Evaluation

We invite five volunteers to participate in evaluating the basic meanings in Section 3.1. To ensure consistency and scientific rigor, we provide systematic training for the volunteers. The training covers the theoretical foundations of literal meaning, including basic knowledge of linguistics and semantics,

Sample Range	Sampling Interval	Total Samples Verified	Approved Samples	Resampling Count
1–5000	50	100	92	8
5001–10000	50	100	96	4
10001–15000	50	100	95	5

Table 12: GPT-4o generation and manual review statistics for literal meaning extraction on PSUCMC. We take samples for manual evaluation at intervals of 50 steps and tally the results in 100 manual spot checks.

Sample Range	Sampling Interval	Total Samples Verified	Approved Samples	Resampling Count
1–3737	30	124	118	6

Table 13: GPT-4o generation and manual review statistics for literal meaning extraction on TroFi. We adopt an equally spaced sampling method consistent with PSUCMC, but adjust the sampling interval from 50 to 30 for PSUCMC (i.e., $3,737/30 \approx 124.6$, with 124 samples actually drawn).

Model	ALL to K			PSU to K		
	Prec	Rec	F1	Prec	Rec	F1
mBERT	37.5	39.3	38.4	30.0	30.8	30.4
mRoBERTa	39.2	41.0	40.1	31.5	34.0	32.7
XLM-RoBERTa	43.0	46.0	44.5	35.0	38.7	36.8
MelBERT	48.8	50.4	49.6	39.0	42.0	40.5
MisNet	50.3	52.0	51.5	41.5	43.5	42.3
BasicBERT	51.5	54.0	52.8	40.5	42.5	41.0
RoPPT	49.8	50.6	50.2	40.0	42.5	41.3
ER	51.0	51.5	50.9	42.0	45.0	43.5
MiceCL	52.5	53.7	53.1	43.0	46.5	44.7
DAGS (our)	**53.5**	**56.0**	**54.7**	**45.5**	**48.3**	**46.9**

Table 14: Results of Slovene's cross-language experiments. Here, "**ALL**" indicates the VUA ALL dataset, "**K**" represents the KOMET dataset, and "**PSU**" stands for the PSUCMC (Best is in **Bold**).

as well as methods for distinguishing literal annotations and understanding their manifestations in different contexts. Volunteers also receive detailed guidance on evaluation criteria and voting principles, with a focus on mastering the majority rule in the voting mechanism. Additionally, we use case studies to demonstrate the process of evaluating various annotations, helping volunteers familiarize themselves with practical workflows and methods.

Regarding the use of GPT-4o, we follow the methodologies outlined in (Tian et al., 2024a; Chen et al., 2024). Specifically, we design a **Prompt** (i.e., *Please provide the literal meaning of the target word in a non-metaphorical context, and avoid including any metaphorical explanations.*) to guide GPT-4o in obtaining sampling results for literal annotations. Subsequently, we manually evaluate the results and determine the final results based on a minority-majority voting rule. For each sampled example, we use an equally spaced sampling kernel. We present some statistics for PSUCMC and TroFi in Table 12 and Table 13.

The mainstream corpora we use contain relevant examples of target words, but we remove entries or sentences of metaphorical, abstract, or ambiguous interpretations through sampling with GPT-4o and manual review to ensure that the selected meanings align with the "basic" definition. For instance, in the sentence "The idea caught fire in the community," the verb "caught" does not refer to the literal meaning of "grabbing something," but metaphorically expresses that an idea quickly gained attention or became popular. We exclude sentences with such metaphorical meanings, and focus on dynamic examples with contextual information (e.g., "*Your body needs time to digest the meal.*") (Zhang and Liu, 2022; Li et al., 2023a). In cases where target words lack corresponding annotations in the VUAMC dataset, previous studies typically use aggregated meanings. In contrast, our approach incorporates lexical knowledge from *Etymology Online* [3] to enhance annotation accuracy. To the literal annotations that fail to meet requirements, volunteers resample and reevaluate them, ensuring the accuracy and reliability of the evaluation results.

What's more, to demonstrate the extent of human annotators' contribution, we design experiments under fully automated conditions and test them on the VUA ALL and PSUCMC datasets. The results of the experiments are shown in Table 15. Comparing with the results of the paper, we observe that the F1 scores of the models decrease to varying degrees under fully automatic conditions. For instance, BasicBERT achieves an F1 score of 78.3% on VUA ALL and 76.8% on VUAverb, whereas under fully automatic conditions, its F1 scores slightly

[3] https://www.etymonline.com

Model	VUA ALL (Orig)			VUA ALL (Auto)			VUAverb (Orig)			VUAverb (Auto)		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BasicBERT	79.1	77.7	**78.3**	78.0	76.5	77.2	76.7	77.5	**76.8**	75.5	76.2	75.8
MiceCL	80.4	75.2	**78.5**	79.3	74.1	76.6	75.1	78.0	**75.9**	74.2	77.0	75.1
ContrastWSD	75.5	72.9	**74.2**	74.5	71.5	73.0	79.1	66.9	**72.5**	77.8	65.8	71.2
DAGS (our)	81.7	80.6	**80.9**	80.6	78.9	79.1	80.6	80.0	**80.4**	79.0	78.5	78.7

Table 15: Comparison of results under manual and fully automatic conditions. In this case, "**Orig**" represents the results under the original manual conditions, while "**Auto**" represents the results under fully automated conditions (Best is in **Bold**).

Datasets	DAGS (our)			DAGS (conversely)		
	Prec	Rec	F1	Prec	Rec	F1
VUA ALL	81.7	80.6	**80.9**	80.9	74.6	77.5
VUAverb	80.6	80.0	**80.4**	80.1	75.2	76.2
PSUCMC	80.4	80.0	**80.1**	75.0	76.3	75.0

Table 16: Comparative experimental results of the opposite attention in DAGS (Best is in **Bold**).

drop to 77.2% and 75.8%, respectively. This indicates that high-quality human annotations provide accurate basic meanings. Although automatically generated basic meanings can support the models in completing most tasks, cumulative errors and semantic ambiguities are difficult to fully avoid in fully automated settings. Furthermore, although the performance of all models decreases, DAGS shows the smallest decline and remains superior to other models.

H Implementation Supplement

Table 1 in our paper presents the results of supervised experiments. Following previous studies (Choi et al., 2021; Tian et al., 2024b; Jia and Li, 2024), we also use a development dataset to determine the optimal hyperparameter settings for evaluating TroFi and. Since TroFi does not have an official data split, we adopt 10-fold cross-validation for evaluation. Specifically, we evenly divide the dataset into 10 subsets, using 9 subsets for training and 1 subset for testing in each iteration. This process is repeated 10 times, and we report the average performance across all runs. The evaluation metrics remain consistent with those used for the other datasets.

In addition, for the setting of the learning rate, we also refer to the previous method of increasing and then decreasing. This strategy, commonly referred to as learning rate warm-up and linear decay, effectively balances the stability in the early stages of training with convergence in the later stages. This method shows good performance in pre-trained language models (e.g., BERT and RoBERTa), and this setting also allows for better comparison with previous baselines.

I Cross Attention Input Order Comparison Experiment

We obtain updated representations through a combination of self-attention and cross-attention. In Figure 1, the input of the current sentence is used as Q (Query), while the input of the other sentence is used as K (Key) and V (Value). Q represents the information that currently requires attention, while K and V provide contextual support. Since the task requires the current sentence to focus on capturing the key information of the other sentence, we chose this input direction in our design. In order to investigate the effect of different input orders on the model results, we design opposite experiments and perform a comparative analysis.

Observing the Table 16, DAGS (our) significantly outperforms DAGS (conversely) on all datasets. For instance, on the VUA ALL dataset, the F1 score of DAGS (our) is 80.9%, significantly higher than the 77.5% of DAGS (conversely). Similarly, on the VUAverb and PSUCMC datasets, the F1 scores improved by 4.2% and 5.1%, respectively. This difference indicates that our method is more efficient in capturing key information between the current sentence and another sentence, while the converse design may lead to poor semantic extraction performance. This may be due to the fact that the opposite design leads to a dispersion of model attention, which reduces the characterization.

J Details of Evaluation Metrics

In this section, we provide a detailed introduction to the evaluation metrics used in the paper, including

Precision (**Prec**), Recall (**Rec**), and the **F1** score. Precision measures the proportion of true positives among all predicted positive instances, reflecting the model's exactness in identifying positive samples. A higher precision indicates better performance in predicting positive cases. Recall, on the other hand, assesses how many true positives the model can identify, reflecting the model's sensitivity in recognizing positive samples. A higher recall means that the model is capable of finding more positive instances. The F1 score is the harmonic mean of Precision and Recall, aiming to balance these two metrics. Particularly in cases of class imbalance, the F1 score provides a more comprehensive assessment of performance than using Precision or Recall alone.