FL-MSCL: A Unified Figurative Language Detection Model Driven by Multi-Type Signals and Contrastive Learning

Shijia Lu[1,3], **Fumiyo Fukumoto**[2], **Xiaoxi Huang**[1], **and Yoshimi Suzuki**[4]
[1]School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
[2]School of Interdisciplinary Mathematical Sciences, Meiji University, Tokyo, Japan
[3]Graduate School of Engineering, University of Yamanashi, Kofu, Japan
[4]Interdisciplinary Graduate School, University of Yamanashi, Kofu, Japan
{231050366, huangxx}@hdu.edu.cn
fukumoto@meiji.ac.jp, ysuzuki@yamanashi.ac.jp

Abstract

Figurative language recognition poses significant challenges in NLP, particularly when distinguishing between fine-grained rhetorical categories such as metaphor, metonymy, and simile. This paper formulates the problem as a four-way sentence-level classification task and proposes **FL-MSCL**, a unified framework integrating prompt-based knowledge injection with supervised contrastive learning. Experiments across both unified and single-class benchmarks demonstrate that FL-MSCL achieves competitive performance compared to State-of-the-Art (SOTA) methods, indicating consistent advantages in cross-category generalization and category-specific detection. Our code is available at https://github.com/LSxianxianxian/FL-MSCL.

1 Introduction

Figurative language—such as metaphor, metonymy, and simile—serves as a powerful tool for expressing abstract meanings. While crucial for language understanding, detecting these figures is complicated by their shared semantic deviation from literal meaning despite distinct cognitive drivers.

Despite significant progress in figurative language detection, two fundamental gaps persist. First, existing approaches are overwhelmingly framed as single-category binary classifiers; even multi-task extensions target broad comprehension tasks (e.g., NLI) rather than the fine-grained disambiguation of cognitively related figures. Second, no unified framework has been proposed to jointly resolve the inherent ambiguity among the confusable trio of metaphor, metonymy, and simile.

To bridge these gaps, we introduce FL-MSCL, a unified framework for four-way sentence-level classification. FL-MSCL strategically integrates prompt-based knowledge injection with supervised contrastive learning to enforce explicit class distinctions. Our main contributions are three-fold: (1) We formulate a unified task to address semantic ambiguity among metaphor, metonymy, and simile; (2) We propose a novel framework integrating dynamic prompts with supervised contrastive learning; (3) Extensive evaluations show FL-MSCL achieves new SOTA results on benchmarks like MOH-X and competitive performance on six other datasets.

2 Related Work

General Figurative Language Detection. Grounded in the Metaphor Identification Procedure (MIP), early work relied on lexical feature engineering (Mohammad et al., 2016), before the field shifted through RNN-based contextual encoders (Gao et al., 2018) and graph-based syntactic models (Le et al., 2020) to Transformer-dominant paradigms (Leong et al., 2020; Choi et al., 2021). Recent work has also explored figurative language as a strategy in multi-agent communication (Xu and Zhong, 2025). Despite this progress, the framing remains largely binary and single-category, leaving inter-category ambiguity among metaphor, metonymy, and simile unaddressed.

Multi-Task and Multi-Modal Figurative Modeling. MTL approaches have explored joint modeling of related figure pairs (Do Dinh et al., 2018; Badathala et al., 2023), while broader benchmarks frame figurative language as entailment—textual (Chakrabarty et al., 2022) or multimodal (Saakyan et al., 2025). These efforts target comprehension rather than fine-grained detection, and none specifically addresses unified disambiguation of metaphor, metonymy, and simile.

Knowledge-Enhanced and LLM-Based Approaches. Integrating cognitive theories into pre-trained encoders has advanced single-task metaphor detection: MelBERT (Choi et al., 2021)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 694–702

adopts MIP-based interaction, BasicBERT (Li et al., 2023) models literal meanings explicitly, and WPDM (Tian et al., 2024) leverages domain-level conceptual mining. Despite their effectiveness, these models rely on static, metaphor-only knowledge. Meanwhile, even advanced LLMs struggle to match supervised approaches on fine-grained detection (Sanchez-Bayona and Agerri, 2025; Reimann and Scheffler, 2025), typically due to reliance on surface features rather than cognitive grounding. FL-MSCL addresses both limitations by combining dynamic exemplar retrieval with supervised contrastive learning for unified four-way classification.

3 Task Design

We formulate figurative language recognition as a unified four-way sentence-level classification task. Given an input sentence x, the model predicts a label $y \in \{0, 1, 2, 3\}$. Unlike binary detection, the task resolves ambiguity among the following categories:

0 - Literal: Expressions used in their literal sense.
Ex: "The water is boiling on the stove."
1 - Metaphor: Abstract meanings via conceptual mappings.
Ex: "His words cut deeper than a knife."
2 - Metonymy: Reference via conceptual contiguity.
Ex: "The White House issued a statement."
3 - Simile: Explicit comparisons using "like" or "as".
Ex: "She ran like the wind."

4 FL-MSCL Framework

4.1 Model Overview

Figure 1 illustrates the FL-MSCL framework. Following the line of research on definition-based knowledge injection (Yu et al., 2021; Fukumoto and Asakawa, 2023), we adapt this mechanism for multi-class figurative language recognition. We innovate by replacing static definitions with dynamically retrieved exemplars and incorporating supervised contrastive learning. The pipeline operates as follows:

Input Construction. As shown at the bottom right of Figure 1, the model takes a composite sequence x as input. This sequence is constructed by the Prompt Mechanism (detailed in Sec. 4.2 and the left side of Fig. 1) and then mapped through Input and Position Embeddings:

$$x = \{x_1, x_2, \ldots, x_n\}, \tag{1}$$

where n is the sequence length.

Encoding & Representation. The sequence x is fed into a DeBERTa encoder(right side of Fig. 1) to generate contextual embeddings $H \in \mathbb{R}^{n \times d}$. We extract the sentence representation z from the `[CLS]` token via a linear transformation:

$$z = W \cdot h_{[\text{CLS}]} + b, \tag{2}$$

where W and b are learnable parameters. This z serves as the core feature for both classification and contrastive objectives.

Prediction & Optimization. As shown in the top-right part of Fig. 1, the class probability is computed via a softmax classifier:

$$P(y|x) = \text{Softmax}(W_c \cdot z + b_c). \tag{3}$$

As shown in the middle-right right part of Fig. 1, the training objective combines standard cross-entropy ($\mathcal{L}_{\text{CE}}$) with a supervised contrastive term ($\mathcal{L}_{\text{SCL}}$):

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{M}\sum_{j=1}^{4} y_{ij} \log \hat{y}_{ij}, \tag{4}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{SCL}}, \tag{5}$$

where M is batch size and λ balances the losses.

4.2 Prompt Mechanism

Visualized in the prompt part of Fig. 1, to bridge rhetorical categories with contextual semantics, we augment target sentences with retrieved examples. Each prompt follows this structure:

> `[CLS] [target] [SEP]`
> *Ex (Literal):* [sent] `[SEP]` *Ex (Metaphor):* [sent] `[SEP]`
> ...
> **Question:** Rhetorical type of this sentence?

We adopt a **Dynamic Prompts combined with Multi-Example** strategy. Using Sentence-BERT and FAISS, we retrieve the most semantically similar exemplars each category. This reformulates the dictionary-based injection paradigm (Yu et al., 2021; Fukumoto and Asakawa, 2023) into an exemplar-based analogical signal. Comparisons of prompt strategies are in Appendix A.

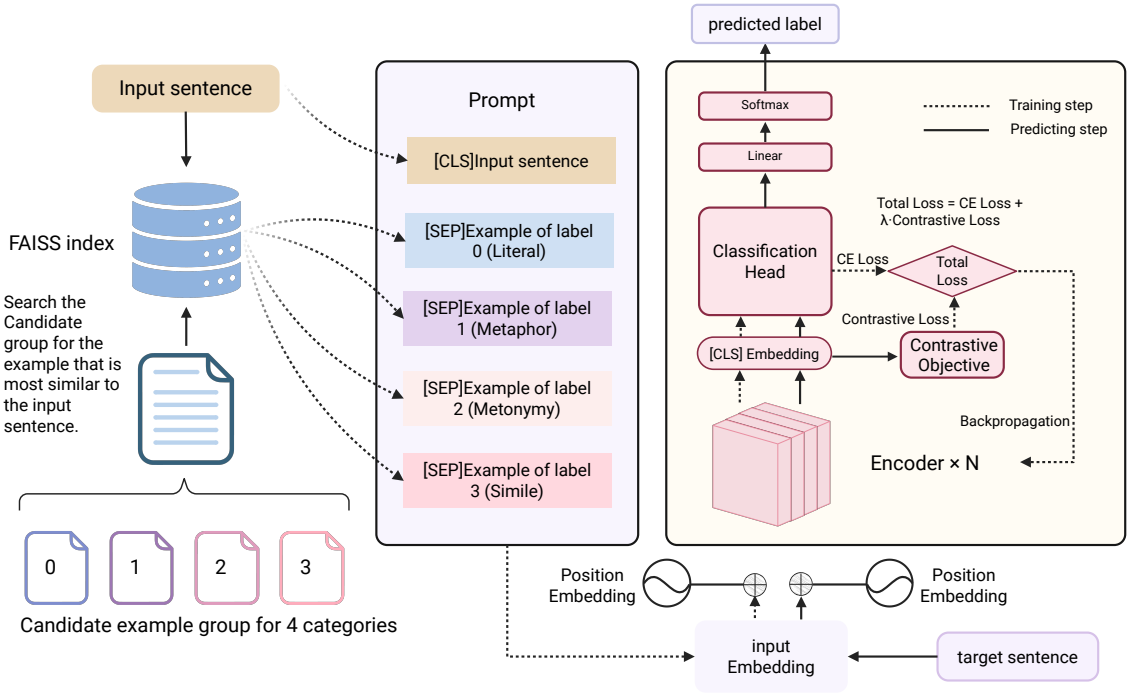

Figure 1: The FL-MSCL framework. Left: Dynamic prompt construction via FAISS retrieval. Right: DeBERTa-based architecture with dual-objective optimization (Cross-Entropy + Supervised Contrastive Learning).

4.3 Supervised Contrastive Learning (SCL)

To enforce clearer boundaries between the four confusable categories, we apply SCL on the representation z (corresponding to the Contrastive Objective block in Fig. 1). The loss brings same-class samples closer while pushing different classes apart. For a sample i, the InfoNCE loss is:

$$T = \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)},$$

$$\mathcal{L}_{\text{SCL}}^{(i)} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} T, \qquad (6)$$

where $P(i)$ is the set of positive samples, $A(i)$ includes all batch samples excluding i, and τ is the temperature. The final batch loss $\mathcal{L}_{SCL}$ is the average over all samples, which is added to the total objective as defined in Eq. (5).

5 Experiments

5.1 Experimental Setup

Datasets We evaluate FL-MSCL on representative benchmarks across three figurative categories: (1) **Metaphor**: VUA-18/20/Verb (Leong et al., 2020), MOH-X (Mohammad et al., 2016), and TroFi (Birke and Sarkar, 2006); (2) **Metonymy**: ReLocaR (Gritta et al., 2017); (3) **Simile**: PLMs-Simile (He et al., 2022). To construct the **Unified Training Corpus**, we merge these datasets into a single collection, applying down-sampling to balance the four classes. For evaluation on **Standard Binary Benchmarks** (e.g., VUA, MOH-X), we strictly follow the official train/test splits from prior literature to ensure fair comparison.

Protocols & Metrics. We adopt two evaluation settings with task-specific metrics: **Unified 4-Way Classification:** We report Macro-F1 and Accuracy to evaluate the model's overall ability to distinguish among all categories while mitigating class imbalance. **Binary Detection:** For comparison with baselines, we report the standard positive-class F1, along with Precision and Recall. Detailed training configurations and data statistics are provided in Appendix B.

Baselines. We compare FL-MSCL against three groups of strong baselines: (1) **Classic & Contextual**: RNN_ELMo, RNN_HG (Gao et al., 2018), MUL_GCN (Le et al., 2020), RoBERTa (Leong et al., 2020), and MelBERT (Choi et al., 2021); (2) **Advanced SOTA**: Specialized metaphor detection models including BasicBERT (Li et al., 2023), ContrastWSD (Elzohbi and Zhao, 2024), MiceCL (Jia

Method	R	P	F1	Acc
w/o Pro + w/o Cont	83.8	80.7	80.7	78.6
Prompt + w/o Cont	**84.1**	82.1	81.8	**83.5**
w/o Pro + Cont	82.9	83.2	83.0	81.4
FL-MSCL (Full)	**84.1**	**88.8**	**86.1**	83.3

Table 1: Results on the Unified 4-way Classification Task. **Pro**: Prompt, **Cont**: Contrastive.

and Li, 2024), KEG (Heyan et al., 2023), and WPDM (Tian et al., 2024); (3) **LLMs**: GPT-3.5 and GPT-4 in zero-shot and 5-shot settings (Tian et al., 2024). For non-metaphor tasks, we compare with task-specific SOTAs: PreWin (Gritta et al., 2017) for metonymy and PLMs-RoBERTa (He et al., 2022) for simile. For all baselines, we report the best published results on standard splits.

Implementation. We use `DeBERTa-v3-base` as the encoder. The model is trained using a combined Cross-Entropy and Supervised Contrastive Loss. Detailed hyperparameters and training configurations are provided in Appendix B.

5.2 Results

Unified Multi-Class Detection. Table 1 presents the performance on the unified 4-way classification task, which challenges the model to resolve ambiguity among Literal, Metaphor, Metonymy, and Simile simultaneously. FL-MSCL achieves the best overall performance (86.1 F1). Comparing ablation variants (rows 1-3) reveals a complementary mechanism: while adding supervised contrastive learning alone (*w/o Pro + Cont*) significantly boosts Precision (80.7→83.2) by sharpening decision boundaries, it slightly hurts Recall (83.8→82.9) due to stricter rejection. Crucially, the full model incorporating dynamic prompts restores and improves Recall to 84.1 while pushing Precision to 88.8. This confirms that prompts provide necessary semantic grounding to compensate for the strict discrimination of SCL, resulting in a balanced and robust unified model.

Performance on Metaphor Benchmarks. Table 2 summarizes the performance on standard metaphor benchmarks. FL-MSCL demonstrates robust generalization, setting a new SOTA on MOH-X (86.2 F1) and achieving 88.9 F1 on the challenging VUA-Verb dataset, surpassing the domain-mining based WPDM by a substantial margin (+14.8). The ablation results (bottom of Table 2) reveal the critical driver behind this success: removing dynamic prompts leads to a sharp performance drop on context-sensitive tasks like VUA-Verb (from 88.9 down to 81.9). This confirms that retrieved exemplars act as necessary "semantic anchors," providing superior grounding for complex verb metaphors compared to static knowledge or contrastive learning alone.

Generalization on Metonymy & Simile. Table 3 validates the model's robustness on non-metaphor tasks. Table 3 validates the model's robustness on non-metaphor tasks. On Metonymy (ReLocaR), FL-MSCL achieves 89.5 F1, outperforming PreWin (84.8) and approaching feature-engineered BiLSTMs (90.1) without relying on explicit NER/POS tags. Mechanism analysis (detailed in Appendix Table 9) attributes this success to the supervised contrastive objective, which proves essential for distinguishing metonymic mappings (e.g., specific entities used as agents) from their literal counterparts. On Simile, the model achieves near-perfect detection (99.9 F1).

5.3 Error Analysis

Despite strong performance, error analysis reveals distinct patterns: (1) **Conservative Metaphor Prediction:** On VUA-18, the model shows high Precision but lower Recall compared to some baselines (see Appendix table 5). This suggests that when retrieved exemplars are not sufficiently similar, the model tends to predict "Literal" to be safe. (2) **Pragmatic Ambiguity:** In Metonymy, errors often cluster around institutional references (e.g., *"Washington said"*). Distinguishing these from literal locations requires world knowledge beyond current semantic retrieval. (3) **Implicit Similes:** While explicit similes are handled well, the model occasionally misses creative comparisons lacking standard markers like "like" or "as".

6 Conclusion

We introduced FL-MSCL, a unified framework for fine-grained figurative language detection. By integrating prompt-based knowledge injection and supervised contrastive learning, FL-MSCL not only effectively handles the unified 4-way classification task but also achieves new SOTA performance on VUA-Verb and MOH-X. Our experiments validate that explicitly modeling the boundaries between confusable rhetorical categories enhances both category-specific accuracy and cross-category generalization.

Method	VUA Datasets (F1)			MOH-X	TroFi
	VUA-18	VUA-20	VUA-Verb	F1	F1
Advanced SOTA & LLMs					
KEG (Heyan et al., 2023)	-	-	-	81.8	**76.7**
MiceCL (Jia and Li, 2024)	-	-	-	85.2	62.9
WPDM (Tian et al., 2024)	-	-	74.1	83.8	73.4
GPT-4 (5-shot) (Tian et al., 2024)	-	-	67.6	75.3	68.3
BasicBERT (Li et al., 2023)	79.0	73.3	-	-	-
ContrastWSD (Elzohbi and Zhao, 2024)	**79.5**	73.1	77.3	-	-
*Ablation Study (**FL-MSCL**)*					
w/o Prompt + w/o Contrast	77.6	73.5	81.9	77.3	64.1
Prompt + w/o Contrast	78.4	**77.7**	88.0	81.7	66.2
w/o Prompt + Contrast	74.0	73.9	87.7	84.5	68.5
FL-MSCL (Pro + Cont)	79.2	76.4	**88.9**	**86.2**	72.7

Table 2: Consolidated F1 scores on all metaphor benchmarks. We compare FL-MSCL against classic models, recent specialized SOTAs (e.g., ContrastWSD, WPDM, KEG), and LLMs. "-" indicates results not reported in the original papers. Best results are bolded, second-best are underlined. Note: Due to space limits, full metrics (P/R) are provided in Appendix C

Task	Method	Acc	F1
Metonymy	PreWin	83.6	84.8
	FL-MSCL	**85.8**	**89.5**
Simile	RoBERTa	89.4	-
	FL-MSCL	**99.9**	**99.9**

Table 3: Results on Metonymy (ReLocaR) & Simile.

Acknowledgments

We would like to thank the anonymous reviewers and the meta-reviewer for their helpful comments and suggestions. This work is partially supported by JKA (2024M-557), JSPS KAKENHI (No.24K15085, 26K14959), the MOE(China) Humanities and Social Sciences Grant (No.18YJA740016), and NSSFC (No.18ZDA290).

Limitations

Despite encouraging results, this study has several limitations. Our experiments are limited to English datasets, and the manually designed prompts with static examples may not fully generalize to diverse metaphorical expressions. We also focused on supervised contrastive learning, leaving other strategies such as hard negative mining unexplored. Finally, our model is primarily evaluated at the sentence level, without extending to discourse-level figurative detection.

References

Saiteja Badathala, Ashish Goswami, and Radhika Mamidi. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

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, Trento, Italy. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. 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.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. 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.

Fumiyo Fukumoto and Shou Asakawa. 2023. Knowledge injection with perturbation-based constrained attention network for word sense disambiguation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 171–177, Nusa Dua, Bali. Association for Computational Linguistics.

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.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7875–7887, Dublin, Ireland. Association for Computational Linguistics.

Huang Heyan, Liu Xiao, and Liu Qian. 2023. Knowledge-enhanced graph encoding method for metaphor detection in text. Journal of Computer Research and Development, 60(1):140–152.

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, Mexico City, Mexico. Association for Computational Linguistics.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8139–8146.

Chee Wee (Ben) 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, Online. Association for Computational Linguistics.

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

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, Berlin, Germany. Association for Computational Linguistics.

Sebastian Reimann and Tatjana Scheffler. 2025. 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)*, pages 10–21, Vienna, Austria. Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477, Vienna, Austria. Association for Computational Linguistics.

Yuan Tian, Ruike Zhang, Nan Xu, and Wenji Mao. 2024. 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)*, Bangkok, Thailand. Association for Computational Linguistics.

Shuhang Xu and Fangwei Zhong. 2025. CoMet: Metaphor-driven covert communication for multi-agent language games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7892–7917, Vienna, Austria. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Ziqing Yang, and Zengchang Liu. 2021. Dict-BERT: Enhancing language representation pre-training with dictionary. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1428–1439, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A prompt design strategies

four construction strategies:

(1)**Static Prompts**: Fixed exemplars from each category are used for all inputs throughout training.

(2)**Random Prompts**: Random selection of one exemplar per category during each training step.

(3)**Dynamic Prompts**: Semantic retrieval of the most similar exemplar per category during each training step using Sentence-BERT (all-MiniLM-L6-v2) and FAISS indexing.

(4)**Dynamic Prompts + Multi-Ex**: Enhanced dynamic approach with three retrieved examples specifically for metaphor category.

Prompt Design	Accuracy	F1
Without prompt	78.6	80.67
Fixed prompt	<u>81.74</u>	82.55
Random prompt	80.22	82.26
Dynamic prompt	81.26	<u>82.99</u>
Dynamic prompt + Multi-ex	**81.8**	**83.47**

Table 4: Comparison of Prompt Design Strategies. Best results are in bold, second-best underlined.

Table 4 demonstrates that prompting strategies universally surpass no-prompt baselines, with dynamic variants outperforming static/random approaches and the Dynamic+Multi-Ex configuration achieving optimal performance. These findings confirm dynamic strategies provide superior semantic guidance, with targeted metaphor reinforcement through multi-example optimization delivering the most substantial gains. We consequently adopt Dynamic+Multi-Ex for all subsequent experiments.

B Detailed Setup

B.1 Unified Data Construction

For the Multi-class task, we merge all the aforementioned datasets (VUA series, MOH-X, TroFi, ReLocaR, PLMs-Simile) into a unified collection. Since the combined dataset is heavily imbalanced (dominated by literal and metaphor samples), we apply a downsampling strategy on the majority classes to ensure the ratio of samples for Literal, Metaphor, Metonymy, and Simile is balanced during training. We then perform a proportional split into training, validation, and test sets.

B.2 Training Configuration

We use a DeBERTa-v3-base encoder for all experiments. The loss function combines cross-entropy and supervised contrastive loss, with a weighting coefficient $\lambda = 0.25$ and contrastive temperature $\tau = 0.28$. Optimization is performed using AdamW with a learning rate of 1×10^{-5} and a batch size of 16. A linear learning rate scheduler is applied. Each model is trained for 40 epochs, and the checkpoint with the highest macro-F1 score on the validation set is selected for testing. All experiments are conducted on a single NVIDIA A6000 GPU. The total training time ranges from 4 to 20 hours depending on the specific prompt design (static vs. dynamic).

C Detailed Experimental Results

This appendix provides the comprehensive experimental results omitted from the main text due to space constraints. We report Precision (P), Recall (R), and F1 scores for all baselines across the evaluated datasets.

C.1 Metaphor Recognition Breakdown

Table 5 presents the full comparison on the sentence-level VUA-18 and VUA-20 datasets. Table 6 details the results on the VUA-Verb dataset, where FL-MSCL demonstrates significant improvements. Finally, Tables 7 and 8 provide the breakdown for the MOH-X and TroFi datasets, respectively.

Metaphor Recognition Results

C.2 Generalization Details (Metonymy & Simile)

Table 9 and Table 10 provide the detailed metrics for Metonymy and Simile recognition, respectively, validating the cross-category robustness of FL-MSCL.

Method	VUA-18			VUA-20		
	R	P	F1	R	P	F1
Classic/Contextual Baselines						
RNN_ELMo (2018)	73.6	71.6	72.6	73.6	71.6	72.6
RNN_BERT (2019)	71.9	71.5	71.7	71.9	71.5	71.7
RNN_HG (2019)	76.3	71.8	74.0	**76.3**	71.8	74.0
RNN_MHCA (2019)	75.7	73.0	74.3	75.7	73.0	74.3
MUL_GCN (2020)	/	/	/	75.5	74.4	75.1
Transformer-based & Advanced Baselines						
RoBERTa_BASE (2020)	75.0	79.4	77.1	68.0	74.9	71.2
RoBERTa_SEQ (2020)	74.9	80.4	77.5	66.7	76.9	71.4
DeepMet (2020)	71.3	82.0	76.3	65.9	76.7	70.9
MelBERT (2021)	76.9	80.1	78.5	68.6	76.4	72.3
ContrastWSD (2024)	**78.9**	80.2	**79.5**	69.8	76.6	73.1
Recent SOTA Comparison						
BasicBERT (2023)	78.5	79.5	79.0	73.2	73.3	73.3
w/o BasicMIP	75.1	81.7	78.3	74.8	74.8	69.8
*Ablation Study (**FL-MSCL**)*						
FL-MSCL w/o Pro, w/o Cont	66.8	95.6	77.6	62.0	90.1	73.5
FL-MSCL Pro, w/o Cont	66.1	96.3	78.4	66.8	92.8	**77.7**
FL-MSCL w/o Pro, Cont	59.6	**97.6**	74.0	61.0	93.7	73.9
Pro + Cont	70.7	93.5	79.2	64.9	**94.3**	76.4

Table 5: Comparison of results on VUA datasets (VUA-18 and VUA-20). The best results are in bold, second-best are underlined.

Method	VUA-verb		
	R	P	F1
Classic Baselines			
RNN_ELMo (2018)	71.3	68.2	69.7
RNN_BERT (2019)	66.7	69.0	69.0
RNN_HG (2019)	72.3	69.3	70.7
RNN_MHCA (2019)	72.5	66.3	70.5
MUL_GCN (2020)	70.9	72.5	71.7
Transformer-based & Advanced Baselines			
RoBERTa_BASE (2020)	72.8	76.9	74.7
RoBERTa_SEQ (2020)	69.8	79.2	74.2
DeepMet (2020)	70.8	79.5	74.9
MelBERT (2021)	72.9	78.7	75.7
ContrastWSD (2024)	79.2	66.9	77.3
Recent SOTA & LLM Comparison			
GPT-3.5 (Zero-Shot)	/	/	50.5
GPT-3.5 (5-Shot)	/	/	53.6
GPT-4 (Zero-Shot)	/	/	66.9
GPT-4 (5-Shot)	/	/	67.6
WPDM (2024)	/	/	74.1
Ablation Study (FL-MSCL)			
FL-MSCL w/o Pro, w/o Cont	80.2	83.6	81.9
FL-MSCL Pro, w/o Cont	**91.3**	84.9	88.0
FL-MSCL w/o Pro, Cont	85.6	**89.8**	87.7
Pro + Cont	89.8	88.1	**88.9**

Table 6: Comparison of results on VUA-verb dataset metaphor detection. Best results are in bold, second-best are underlined.

Method	R	P	F1	Acc
RNN_ELMo (2018)	73.5	79.1	75.6	77.2
RNN_BERT (2019)	81.8	75.1	78.2	78.1
RNN_HG (2019)	79.8	79.7	79.8	79.7
RNN_MHCA (2019)	83.1	77.5	80.0	79.8
MUL_GCN (2020)	80.5	79.7	79.6	79.9
MelBERT (2021)	82.7	79.7	81.1	81.6
MisNet (2022)	84.0	84.2	83.4	83.6
KEG (2023)	81.6	82.1	81.8	81.6
MiceCL(2024)	**87.7**	83.2	85.2	**85.2**
GPT-3.5(Standard Zero-Shot)	/	/	64.4	67.3
GPT-3.5(5-Shot)	/	/	70.1	72.5
GPT-4(Standard Zero-Shot)	/	/	72.6	77.0
GPT-4(5-Shot)	/	/	75.3	79.0
WPDM	/	/	83.8	84.2
w/o Pro + w/o Cont	82.4	72.8	77.3	76.9
Prot + w/o Cont	82.7	80.7	81.7	79.1
w/o Pro + Cont	83.9	85.0	84.5	82.4
Pro + Cont	86.4	**85.9**	**86.2**	84.9

Table 7: Comparison on the MOH-X dataset. Best results are in bold, second-best are underlined.

Method	R	P	F1	Acc
RoBERTa_BASE(2020)	74.3	54.6	62.9	/
RoBERTa_SEQ(2020)	70.1	53.6	60.7	/
DeepMet(2020)	72.9	53.7	61.7	/
MelBERT(2021)	74.1	53.4	62.0	/
MrBERT(2021)	75.0	53.8	62.7	61.1
MiceCL(2024)	75.0	54.2	62.9	61.5
KEG(2024)	77.8	**75.6**	**76.7**	<u>76.2</u>
GPT-3.5(Standard Zero-Shot)	/	/	50.1	56.0
GPT-3.5(5-Shot)	/	/	57.6	60.0
GPT-4(Standard Zero-Shot)	/	/	67.3	67.5
GPT-4(5-Shot)	/	/	68.2	68.3
WPDM(2024)	/	/	<u>73.4</u>	**76.3**
w/o Pro + w/o Cont	68.7	60.1	64.1	66.5
Prot + w/o Cont	76.5	58.0	66.2	68.7
w/o Pro + Cont	<u>77.9</u>	61.1	68.5	70.8
Pro + Cont	**80.1**	<u>66.6</u>	72.7	73.1

Table 8: Comparison on the TroFi dataset. Best results are in bold, second-best are underlined.

Method	**Accuracy**	**F1**
PreWin	83.6	84.8
LSTM (GLoVE)	78.4	78.4
+NER+POS	80.6	80.6
BiLSTM (GLoVE)	83.0	82.9
+NER+POS	84.2	84.2
BiLSTM (ELMo)	<u>90.0</u>	**90.1**
+NER+POS	**90.1**	**90.1**
w/o Pro + w/o Cont	82.3	85.1
Prot + w/o Cont	84.2	86.6
w/o Pro + Cont	83.3	88.2
Pro + Cont	85.8	<u>89.5</u>

Table 9: Performance on the ReLocaR metonymy dataset. Best results are in bold, second-best are underlined.

Method	**Accuracy**	**F1**
PLMs-RoBERTa-LARGE	89.4	/
w/o Prompt + w/o contrast	<u>99.88</u>	<u>99.99</u>
Prompt + w/o contrast	<u>99.88</u>	<u>99.99</u>
w/o Prompt + contrast	99.53	99.76
Contrast+Prompt	**99.88**	**99.99**

Table 10: Performance on the Simile dataset. Best results are in bold, second-best are underlined.