TueCL at SemEval-2025 Task 1: Image-Augmented Prompting and Multimodal Reasoning for Enhanced Idiom Understanding

Yue Yu, Jiarong Tang, Ruitong Liu
Department of Linguistics, University of Tübingen
Tübingen, Germany
{y.yu, jiarong.tang, ruitong.liu}@student.uni-tuebingen.de

Abstract

This paper presents our approach for SemEval-2025 Task 1, **Ad**vancing **M**ultimodal **I**diomaticity **Re**presentation (AdMIRe), which focuses on idiom image ranking via semantic similarity. We explore multiple strategies, including neural networks on extracted embeddings. A key component of our methodology is the application of advanced prompt engineering techniques within multimodal in-context learning (ManyICL), leveraging GPT-4o, CLIP. Our experiments demonstrate that structured and optimized prompts significantly enhance the model's ability to interpret idiomatic expressions in a multimodal setting. The source code used in this paper is available at github.[1]

1 Introduction

Identifying and understanding idioms remain significant challenges large language models (LLMs) (Donthi et al., 2025). An idiom typically consists of multiple words, and its meaning is deeply rooted in cultural and historical contexts, making it impossible to derive solely from the meanings of its individual components (Dankers et al., 2022). Idioms often exhibit entirely different literal and figurative meanings.

Large Language Models (LLMs) has exhibited remarkable emergent abilities, typically including instruction following (Peng et al., 2023), In-Context Learning (ICL) (Brown et al., 2020), and Chain of Thought (CoT) (Wei et al., 2023). Whilist Large Vision Models (LVMs) possess strong visual perception capabilities but often lag in reasoning abilities (Shen et al., 2023). Instruction-tuning requires a large amount of task-specific data (Gu et al., 2023). GPT-4 (OpenAI, 2023).

Recent studies have shown that in-context learning (ICL) capabilities of language models can be effectively applied to vision-language-generating models. The advancement has significantly improved AI's ability to integrate visual and textual information. Models such as CLIP (Radford et al., 2021) have laid the foundation for modern MLLMs, leveraging large-scale parameterization and multimodal instruction tuning to enhance versatility.

ICL enables models to learn from few-shot examples within the input context without requiring parameter updates (Yang et al., 2023). Compared to fine-tuning, which demands significant computational resources and extensive task-specific data (Yin et al., 2024), few-shot ICL is more efficient, requiring minimal data while maintaining adaptability across different contexts.

CLIP (Radford et al., 2021) have shown excellent generalization ability to downstream tasks. This capability highlights the its potential in understanding compositional semantics. Studies have shown that designing high-quality contextual prompts can significantly enhance the performance of CLIP and other vision-language models (Jin et al., 2022)

Our methodology integrates advanced prompt engineering within multimodal in-context learning, leveraging Chain-of-Thought reasoning and self-consistency prompting. We classify idioms into literal and idiomatic cases using GPT-4o, then apply tailored textual and visual prompts for each category. For literal idioms, GPT-4o generates descriptive explanations, which are compared to images via CLIP for ranking. For idiomatic expressions, we employ Colorful Prompt Tuning (CPT) to enhance image interpretability before prompting GPT to rank them. Our approach also explores structured prompt design and annotation techniques, such as red-boxed visual cues, to improve model alignment and reasoning in multimodal tasks.

[1] https://github.com/cicl-iscl/SemEval_2025_Task1_Jiaong_Ruitong_Yue

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1753–1758

2 Data

The dataset used in this study is derived from a provided TSV (tab-separated values) and a collection of images corresponding to each idiom. The TSV file contains columns including compound (the idiom), subset (Train or Sample, Test, Dev), sentence_type (idiomatic or literal), sentence (a contextual sentence using the idiom), expected_order (the anticipated ranking of images, only provided in training dataset), and five pairs of image filenames and captions (e.g., image1_name, image1_caption).[2]

Each idiom in the dataset is associated with five images that need to be ranked based on their relevance to the idiom's interpretation. The training data also includes a Sample subset with 10 examples for initial exploration.

The images represent different levels of idiomaticity:

- A synonym for the idiomatic meaning.
- A synonym for the literal meaning.
- An image related to the idiomatic meaning but not synonymous.
- An image related to the literal meaning but not synonymous.
- A distractor image that is thematically related to the compound but unrelated to both meanings.

3 Methodology

One of our main objectives was to integrate advanced prompt engineering within multimodal in-context learning. Drawing inspiration from Chain-of-Thought (CoT) and Vision instruction prompting, we developed structured prompts to help guide the model's reasoning process.

A study by Yang et al. (2022) on prompt tuning in generative multimodal models examined how different configurations impact performance. The findings suggested that while longer prompts with more parameters generally enhance results, the improvements plateau over time, and excessively long prompts can even degrade performance.

For our experiments, we used the SemEval-2025 Task 11 dataset to evaluate different prompt engineering strategies, focusing on ranking accuracy as provided by the competition organizers[3]. Our approach incorporated both Chain-of-Thought reasoning and self-consistency prompting. Before diving into the ranking task, we first used GPT-4o as a classifier to distinguish between literal and idiomatic uses of idioms. Based on this classification, we then designed both textual and visual prompts to suit each category. For literal idioms, we had GPT-4o generate precise descriptions of their meanings, which were then compared to the images using CLIP. The five given images were ranked according to their similarity scores with these descriptions. For idiomatic expressions, instead of directly processing the text, we applied Colorful Prompt Tuning (CPT) to modify the images, making them more interpretable for large language models. With these enhanced visuals, GPT was then prompted to rank the images accordingly. A detailed breakdown of our methodology for handling literal and idiomatic idioms can be found in Sections 5.1.1 and 5.1.2.

3.0.1 Literal compounds processing

Text prompt designing We leveraged GPT-4o as an expert model to rephrase idioms into descriptive explanations based on their context within given sentences. This transformation aimed to make idiomatic expressions more interpretable for CLIP, enhancing its ability to grasp their meaning in multimodal tasks. To achieve this, we carefully designed text prompts that guided GPT to generate precise, context-aware explanations, ensuring that CLIP could associate images with their intended meanings more effectively.

As shown in Figure 3, the first prompt exhibited inconsistencies in the generated descriptions, which were sometimes excessively long or too brief. Additionally, despite the idiom being used literally in the given sentence, the description occasionally retained an idiomatic interpretation.The second prompt addressed these issues by imposing a word limit and explicitly requiring the model to generate a strictly literal interpretation when encountering literal idioms. Although one or two cases still resulted in idiomatic descriptions, the overall quality and accuracy of the generated explanations were significantly improved.The third prompt, despite providing two examples, consistently produced idiomatic interpretations even for idioms that were

[2] https://semeval2025-task1.github.io/

[3] https://www.codabench.org/competitions/4345/#/results-tab

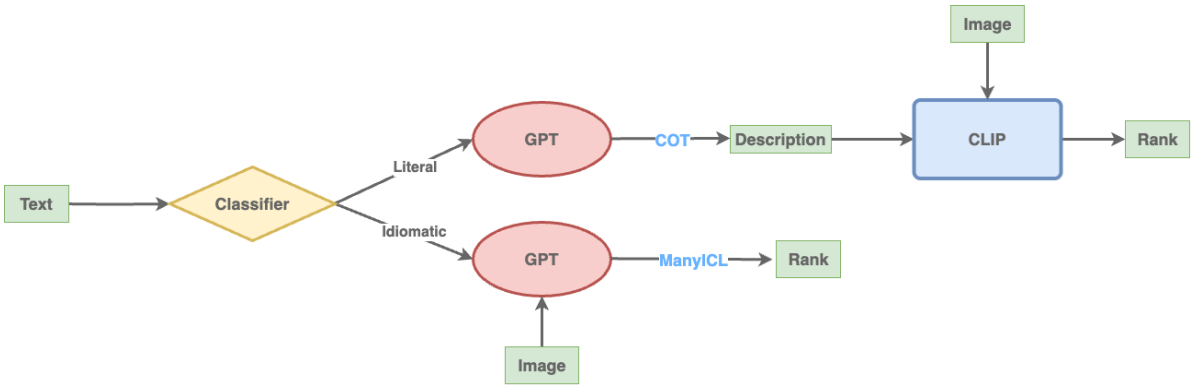

Figure 1: overview of our system framework

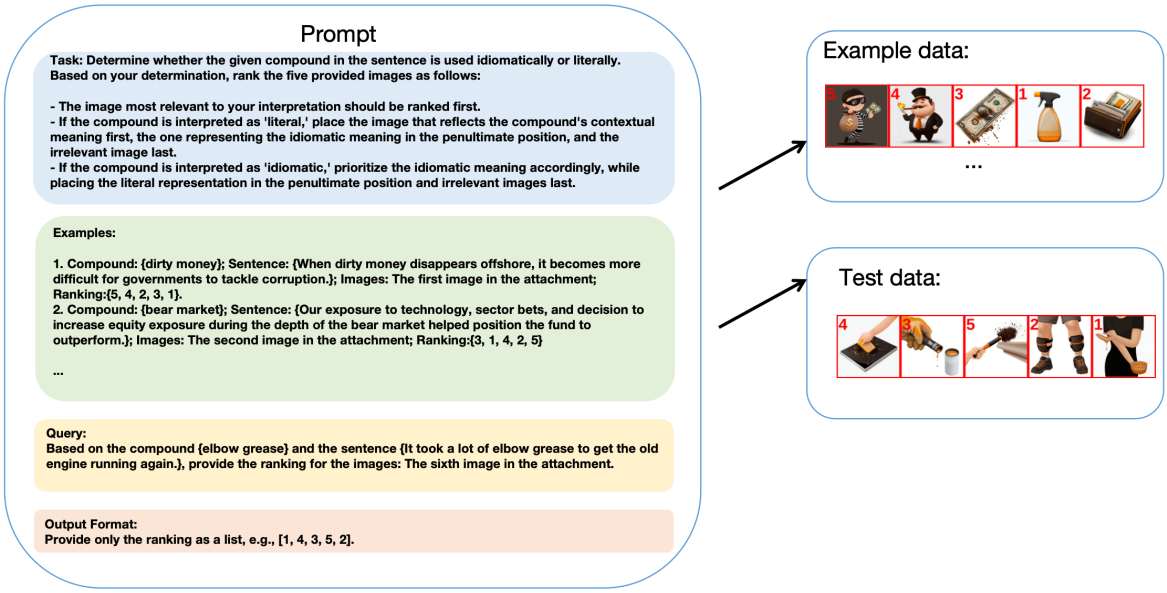

Figure 2: Few-Shot Learning output result: [5, 4, 3, 1, 2] Golden Truth: [5, 4, 3, 1, 2]

supposed to be literal. As a result, we ultimately selected the second prompt as the most effective approach.

4 Limitation

4.0.1 Idiomatic compounds processing

Visual prompt designing Colorful Prompt Tuning introduced in Yao et al. (2022), focuses on colorizing specific regions of images as visual prompts. By incorporating color cues, the model is guided to ground objects and better understand the visual context. Shtedritski et al. (2023) explores the use of annotations, such as red circles, as an innovative visual prompting design. These annotations serve as cues to guide the model's attention toward specific areas of interest, thereby enhancing its understanding of images. As illustrated in Figure 2, red boxes are used to delineate the boundaries of each image, and each image is labeled with a red number to facilitate differentiation. Additionally, we employ a combination of few-shot learning, Chain-of-Thought and self-consistency prompting to guide GPT's reasoning process.

4.1 Results and Evaluation

Our experiments showed that integrating advanced prompt engineering significantly improved performance across different evaluation metrics. We evaluated zero-shot, few-shot, and CoT-based prompting strategies to measure their effectiveness.

Experimental results indicate that simple zero-shot prompts performed poorly, as idiomatic expressions require implicit knowledge. Few-shot

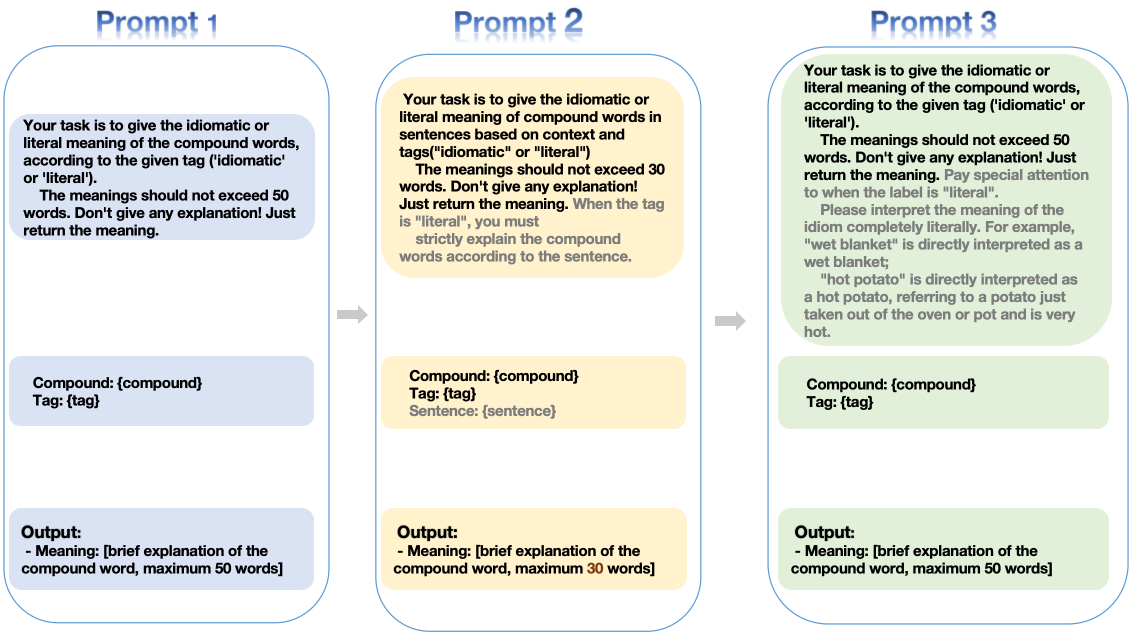

Figure 3: Three examples of generating literal compounds

Table 1: Evaluation results on the development dataset (English).

Metric	Accuracy	Score
Overall Accuracy	**0.7333**	–
Literal Accuracy	**0.875**	–
Idiomatic Accuracy	**0.5714**	–
Overall Rank Correlation	–	0.2867
Literal Rank Correlation	–	0.3875
Idiomatic Rank Correlation	–	0.1714
Overall DCG Score	–	3.1427
Literal DCG Score	–	3.3715
Idiomatic DCG Score	–	2.8813

Table 2: Evaluation results on the test dataset (English).

Metric	Accuracy	Score
Overall Accuracy	**0.6667**	–
Literal Accuracy	**0.7143**	–
Idiomatic Accuracy	**0.6250**	–
Overall Rank Correlation	–	0.2400
Literal Rank Correlation	–	0.2857
Idiomatic Rank Correlation	–	0.2000
Overall DCG Score	–	3.1168
Literal DCG Score	–	3.1950
Idiomatic DCG Score	–	3.0484

learning combined with CoT significantly improved results by providing contextual examples, enabling better model understanding. Especially when using CLIP to rank images, an accurate description of the idiom performed better than the idiom itself in conveying meaning.

To evaluate our approach on the SemEval-2025 Task 1 dataset, we follow the evaluation criteria established by the organizers, using multiple ranking metrics for model performance:

Top-1 Accuracy: The proportion of test cases where the model correctly identifies the most representative image.

Rank Correlation (Spearman's ρ): Measures the agreement between the model's ranking and the ground truth ranking.

Discounted Cumulative Gain (DCG): Evaluates ranking quality by assigning higher importance to correctly ranked top images (Pickard et al., 2025).

Our model achieved an accuracy of 67% in test dataset (Table 2) and 73% in development dataset(Table 1), indicating that it correctly identified the most representative image in the majority of test cases.

We observe that our model performed better on literal expressions compared to idiomatic ones. The model had more difficulty with idiomatic ones due to its complex semantics features.

Prompt engineering lacks interpretability, making it difficult to determine which aspects influence the model's multimodal alignment. Future work could explore these connections further, enabling more efficient experimentation. Also exploring

integrating automatic prompt engineering (APE) techniques and fine-tuning VLMs for a better interpretability.

Acknowledgements

We would like to express our sincere gratitude to Çağrı Çöltekin for his valuable guidance and support.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. 2023. Visual in-context prompting.

OpenAI. 2023. Chatgpt. Accessed: 2023-07-22.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, Carolina Scarton, and Marco Idiart. 2025. Semeval-2025 task 1: Admire – advancing multimodal idiomaticity representation.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. 2023. Aligning and prompting everything all at once for universal visual perception.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does CLIP know about a red circle? visual prompt engineering for VLMs.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Hao Yang, Junyang Lin, An Yang, Peng Wang, Chang Zhou, and Hongxia Yang. 2022. Prompt tuning for generative multimodal pretrained models.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. MM-REACT: Prompting ChatGPT for multimodal reasoning and action.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. Cpt: Colorful prompt tuning for pre-trained vision-language models.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Yifei Zhang, Bo Pan, Siyi Gu, Guangji Bai, Meikang Qiu, Xiaofeng Yang, and Liang Zhao. 2024. Visual attention prompted prediction and learning. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 5517–5525. International Joint Conferences on Artificial Intelligence Organization.

A More Analysis

In our analysis, we evaluate the impact of contextual embeddings on the understanding of idiomatic expressions. To further investigate the effectiveness

of different models, we employ a simple neural network to compute the similarity between text and image embeddings.

A.1 Text-Image Similarity via Neural Network

To quantify the alignment between text and image embeddings, we use a lightweight neural network model. Given a text embedding $\mathbf{T}$ and an image embedding $\mathbf{I}$, the model computes a similarity score S as follows:

$$S = \sigma(\mathbf{W}[\mathbf{T} \oplus \mathbf{I}] + \mathbf{b}) \tag{1}$$

where:

- σ is the sigmoid activation function,
- $\mathbf{W}$ and $\mathbf{b}$ are trainable weight and bias parameters,
- $\oplus$ represents the concatenation operation.

The network outputs a probability score S, indicating the degree of alignment between the textual and visual representations.

A.2 Ranking Images Based on Similarity

Using the computed similarity scores, we rank images based on their alignment with the given textual description:

1. Compute similarity scores S_i for all candidate images.
2. Apply softmax normalization:

$$P_i = \frac{e^{S_i}}{\sum_j e^{S_j}} \tag{2}$$

3. Rank images by descending P_i.

This ranking approach offers a structured way to evaluate embeddings from different models. As shown in Figure 4, multimodal models achieve better text-image alignment, while contextual embeddings improve idiom interpretation over isolated embeddings.

To further explore these findings, we compared text embeddings from *bert-base-uncased*, *clip-vit-large-patch14*, and *DISC* (Zeng and Bhat, 2021). The *baseline* uses embeddings from *bert-base-uncased* without context, whereas other models generate contextual embeddings from entire sentences.

This analysis assesses the impact of contextual information on compound interpretation, particularly for idioms. To ensure consistency, we fixed image embeddings across all models using *clip-vit-large-patch14* and examined their alignment with textual embeddings (Figure 4).

Results show that multimodal models yield the highest alignment, reinforcing the value of visual context. Contextual embeddings outperform isolated embeddings, indicating the importance of surrounding text. Notably, disc surpasses bert-base-uncased by 1.17% in idiom understanding, highlighting the benefits of contextualization. However, overall performance remains suboptimal, motivating further exploration of alternative approaches.

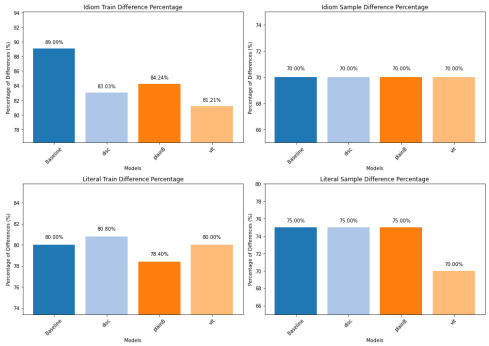

Figure 4: Alignment between text embeddings and image embeddings based on the training dataset

B Processed Image Examples

In this appendix, we present figures 5 6 of images processed using the visual in-context prompting approach. This technique improves vision reasoning (Li et al., 2023) (Zhang et al., 2024).

Figure 5: Dirty Money

Figure 6: Elbow Grease