Zhoumou at SemEval-2025 Task 1: Leveraging Multimodal Data Augmentation and Large Language Models for Enhanced Idiom Understanding

Yingzhou Zhao, Bowen Guan, Liang Yang*, Hongfei Lin
School of Computer Science and Technology, Dalian University of Technology, China
{zyz2020dllg, 20201071138}@mail.dlut.edu.cn
{liang, hflin}@dlut.edu.cn

Abstract

This paper describes our system developed for SemEval-2025 Task 1 : Advancing Multimodal Idiomaticity Representation. This task focuses on ranking images based on the relevance of their visual content and descriptive text to the specific meaning of a given noun compound. Leveraging parameter-efficient fine-tuning of the BLIP-2 model and external knowledge injected through the DeepSeek, our method enables effective ranking of images based on semantic relevance to noun compounds. Specifically, we fine-tune BLIP-2 with LoRA on the provided training dataset to generate descriptive captions for candidate images. The generated captions are then integrated with the original image descriptions using a large language model to create a unified textual representation, which, along with the target noun compound and its sentential context, serves as input for the DeepSeek. Our system achieves a classification accuracy of 0.73 on the English dataset and 0.85 on the Portuguese dataset.

1 Introduction

In SemEval-2025 Task 1 Subtask A (Pickard et al., 2025), participants are challenged to rank images based on their relevance to a given noun compound (NC) within a specific sentential context. This task aims to address the limitations of current large language models, which, compared to humans, often struggle with figurative expressions such as idioms (Tayyar Madabushi et al., 2021; Chakrabarty et al., 2022; Phelps et al., 2024). Building on the premise that human understanding of idioms relies on multisensory interactions with the real world (Lakoff and Johnson, 1980), Subtask A leverages visual representations to encourage the development of models that can better capture the semantic meaning of idioms, a crucial aspect of natural language understanding.

To address this challenge, we propose a system that leverages image-to-text techniques for data augmentation and utilizes advanced large language models to improve classification accuracy by combining parameter-efficient fine-tuning of a multimodal model with external knowledge injection.

We employ LoRA to fine-tune the BLIP-2 model (Li et al., 2023), a pre-trained multimodal model that excels at vision-language tasks, in order to generate descriptive captions for candidate images. BLIP-2 leverages a frozen image encoder and a learned query transformer to efficiently bridge the gap between visual and textual representations, making it well-suited for capturing the interplay between images and idiomatic expressions.

With the recent surge in popularity of DeepSeek, we utilized it for text integration and relevance analysis in the later stages of our experiments.The generated captions are then combined with the original image descriptions using DeepSeek to produce comprehensive and richly detailed textual representations, which are input to the DeepSeek API for inference in order to rank the images based on their contextual relevance to the target noun compound. By combining visual grounding with the reasoning capabilities of a large language model, our system aims to enhance idiomaticity representation and understanding, while also demonstrating the potential of multimodal models for this task.

Our experimental results on the SemEval-2025 Task 1 Subtask A dataset demonstrate that our system achieves promising ranking performance, obtaining high accuracy scores on both the English (0.73) and Portuguese (0.85) datasets. These findings highlight the potential of leveraging multimodal models for visually grounded language understanding and contribute to the development of more robust and context-aware techniques for idiomaticity representation.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 931–936

2 Background

2.1 Dataset Description

The dataset used in the experiment is divided into two parts, English and Portuguese. The English part contains 70 data and the Portuguese part contains 32 data. Each data is composed of: a compound word, a sentence using the compound word, the meaning type of the compound word in the sentence, five comic pictures associated with the compound word, and a text description of each picture.

2.2 Related Work

With the rapid development of machine learning, how to make machines learn to understand colloquialisms has become a very interesting topic. In recent years, many works have used various methods to explore this direction, such as synonym knowledge enhancing (Long et al., 2020), multi-granularity reasoning (Dai et al., 2023), and multi-semantic contrasting (Wu et al., 2024).

The superior ability of large language models has brought new possibilities for the development of colloquialism understanding. Some works (Donthi et al., 2024) have been thinking about how to enhance the understanding ability of large language models for colloquialisms while taking advantage of the basic reasoning ability of models. At the same time, many works (Khoshnevisan, 2019) are also exploring whether multimodal information and models can enhance the understanding ability of colloquialisms. The series of studies shows that the use of multimodal large models for colloquialism understanding tasks is worthy of in-depth exploration, and can provide more fresh ideas for the improvement of this task.

3 System Overview

Our system aims to enhance ranking accuracy by transforming the multimodal task into a purely text-based one, leveraging image-to-text models in the initial stage. Given that current large language models (LLMs) exhibit a comparatively limited capacity for image understanding compared to their proficiency in processing textual data, we strategically employ BLIP-2 for initial data augmentation. This allows us to transform the images into text, thereby enabling the use of more sophisticated LLMs for downstream processing and ultimately achieving superior ranking outcomes.Following this rationale, our system is structured into two primary components: Data Augmentation and Relevance Assessment. The overall architecture of the system is illustrated in Figure 1.

3.1 Data Augmentation

In this task, we employ BLIP-2 for data reprocessing to achieve the goal of data augmentation.

3.1.1 Fine-tuning

In this stage, we extracted all image-text pairs from the provided training and development datasets. To enhance the quality of image descriptions, we fine-tuned the BLIP-2 model using LoRA on these extracted pairs. Additionally, to assess the impact of dataset size on performance, we conducted a comparative experiment by randomly selecting 100 image-text pairs and fine-tuning a separate BLIP-2 model on this smaller subset. This allowed us to evaluate the effectiveness of our data augmentation strategy and explore the trade-off between dataset size and model accuracy.

Our results indicate that the BLIP-2 model fine-tuned on the smaller 100-pair dataset, while maintaining adherence to accurate image content descriptions, exhibited a greater capacity for generating creative and detailed textual descriptions compared to the model fine-tuned on the full dataset.

3.1.2 Caption Integration

A naive approach of simply concatenating the BLIP-2 generated image descriptions with the original captions often resulted in redundancy and logical inconsistencies, thereby hindering effective downstream relevance ranking. To address these limitations, we opted to leverage a large language model (LLM) for text integration.

Specifically, we employed the DeepSeek API, prompting it with carefully designed instructions to synthesize the two textual sources into a coherent and concise description. These instructions were carefully crafted, assigning DeepSeek the role of a "master of textual integration and reasoning." Beyond ensuring information completeness and logical coherence, the instructions also encouraged DeepSeek to infer symbolic meanings from the descriptive details. This strategic prompting aimed to enrich the integrated text with novel insights, ultimately enhancing the performance of downstream relevance assessment. This approach ensured that the resulting textual representation was not merely a concatenation of information, but rather a logi-

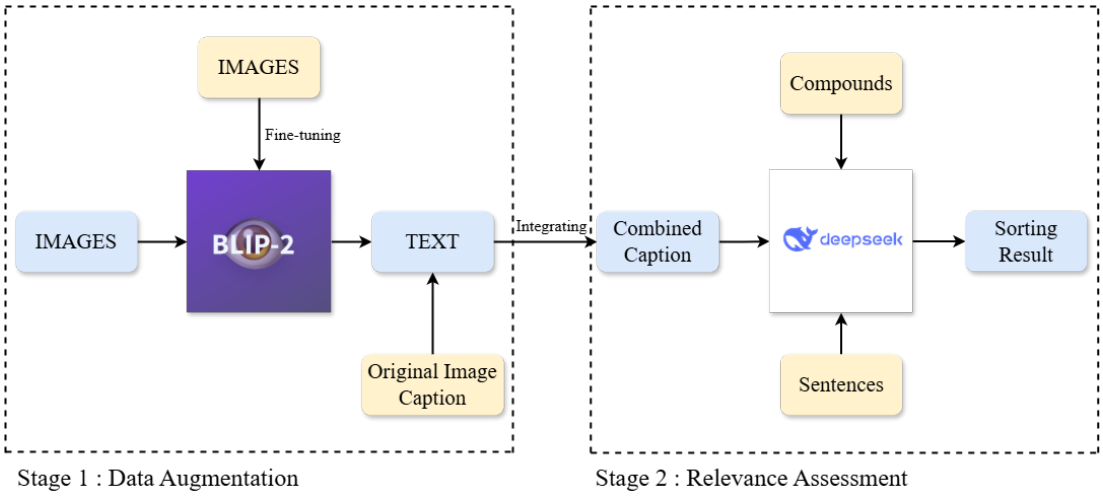

Figure 1: The overall architecture of our proposed system.

cally consistent and semantically rich summary, facilitating improved comprehension by subsequent ranking models.

3.2 Relevance Assessment

In the second component of our system, we explored two distinct approaches to achieve effective relevance ranking. The first approach involved direct utilization of a large language model (LLM) API, prompting the model with carefully crafted queries to elicit a ranking based on contextual relevance. The second approach focused on training a smaller, resource-efficient LLM, such as Llama 3.1 or Qwen 2.5, through fine-tuning on the generated textual data, which was obtained in the previous steps. This allowed us to compare the effectiveness of a zero-shot approach leveraging a powerful LLM API with a fine-tuning approach using a more specialized and resource-conscious model.

3.2.1 Direct API calls

For the relevance assessment component, we again utilized the DeepSeek API to analyze the degree of relatedness between the generated image descriptions and the target compound phrases. In this context, DeepSeek was prompted with meticulous instructions, assuming the role of an "image relevance ranking expert." These instructions directed DeepSeek to consider both the intrinsic meaning of the compound phrase and its nuanced interpretation within the provided sentential context. Furthermore, DeepSeek was explicitly instructed to account for any potential idiomatic or homophonic meanings of the compound, while providing a detailed rationale for the resulting image ranking. This deliberate approach ensured a thorough and context-aware evaluation of image relevance, moving beyond simple keyword matching to incorporate a deeper understanding of semantic relationships.

To generate relevance rankings, we formatted the compound phrase, the corresponding sentence, the five integrated descriptions, and the names of each image from the newly created dataset into a structured input for the DeepSeek API. This input, combined with the aforementioned instructions, enabled DeepSeek to produce both a ranked list of images and a detailed rationale outlining its reasoning process. This output was then used for subsequent evaluation.

3.2.2 Fine-tuning LLMs

In addition to leveraging the DeepSeek API, we explored a second strategy for relevance assessment: fine-tuning smaller, more resource-efficient large language models (LLMs) on the generated textual data. This approach involved training both Llama-3.1-8B and Qwen2.5-7B models on a dataset comprising the combined image descriptions (synthesized as previously described), the corresponding compound phrase, and the sentence providing contextual information. The objective of this fine-tuning process was to directly instill within these models the ability to discern and rank images based on their relevance to the target compound phrase.

While this fine-tuning approach offered the advantage of creating specialized models with potentially lower inference costs compared to relying on an external API, our experimental results indicated a clear performance gap compared to the

DeepSeek API-based ranking. Specifically, the models fine-tuned on Llama-3.1-8B and Qwen2.5-7B, while demonstrating some ability to capture semantic relationships, struggled to achieve the same level of accuracy and coherence in image ranking as the DeepSeek API, particularly in nuanced cases requiring a deeper understanding of idiomatic meanings. This suggests that, for this specific task and dataset, the reasoning capabilities and broader knowledge base inherent in larger, externally hosted LLMs offer a distinct advantage over the knowledge and skills that can be acquired through fine-tuning smaller models.

Despite the reduced performance, we believe that exploring fine-tuning approaches remains valuable. Further research could investigate techniques such as more extensive fine-tuning, the incorporation of more diverse training data, or the use of specialized loss functions to better optimize smaller models for this challenging relevance ranking task.

4 Experimental setup

4.1 DeepSeek API-based Ranking

To leverage the DeepSeek API for relevance ranking, we designed a meticulous prompt that combined the task instruction, the target noun compound, the contextual sentence, and the generated image descriptions. The prompt was structured to explicitly guide the LLM to assess the semantic similarity between the images and the compound, while considering both literal and figurative interpretations of the compound within the given context. Furthermore, the prompt requested a clear and detailed rationale for the model's ranking decision, allowing for a qualitative analysis of the model's reasoning process.

We accessed the DeepSeek API through the OpenAI Python library, configuring the API client with our unique API key and the appropriate base URL. The model parameter was set to "deepseek-reasoner" directing the API to utilize DeepSeek's general-purpose conversational model. For each data instance, the meticulously crafted prompt was packaged into a message with the "user" role and sent to the DeepSeek API. The resulting JSON response, containing the ranked list of images and the associated rationale, was then parsed to extract the predicted ranking.

4.2 Fine-tuning LLAMA and Qwen

For our fine-tuning experiments, we employed the Llama-Factory framework [1], a user-friendly and efficient tool for adapting large language models. Llama-Factory provides a streamlined interface for managing the fine-tuning process, encompassing data loading, model configuration, and training loop management. This framework facilitated experimentation with diverse hyperparameters and training strategies while ensuring consistent and reproducible results. We leveraged this framework to fine-tune both Llama-3.1-8B and Qwen2.5-7B.

Prior to fine-tuning, the BLIP-2 generated textual descriptions were combined with the original image descriptions as described previously to create a more comprehensive data format. We structured this data into a JSON file with a distinct instruction format that incorporated task descriptions. The fine-tuning process was conducted using optimized parameters saved within the model directory. Subsequently, we utilized the fine-tuned Llama-3.1-8B and Qwen2.5-7B models directly for inference, generating ranking predictions via their respective APIs. These predictions were then used for evaluation.

4.3 Evaluation Method

We employed the officially provided CodaBench website [2] for accuracy evaluation.

5 Results

The results of our experiments demonstrate a significant performance disparity between the two relevance ranking approaches. Specifically, the ranking accuracy achieved by directly utilizing the DeepSeek-R1 API substantially outperformed that of the Llama-3.1-8B and Qwen2.5-7B models after fine-tuning. The ranking accuracy of both methods on the given dataset is summarized in Table 1. This suggests that, for the task of visually grounded noun compound ranking, the inherent reasoning capabilities and extensive knowledge base of large, externally hosted LLMs offer a distinct advantage over models that have been fine-tuned on a limited dataset.

Furthermore, we observed that the DeepSeek API, when provided with the textually augmented

[1] https://github.com/hiyouga/LLaMA-Factory

[2] https://www.codabench.org/competitions/4345/#/pages-tab

	Deepseek API	Llama-3.1-8B	Qwen2.5-7B
EN	0.73	0.37	0.27
PT	0.85	0.34	0.29

Table 1: The ranking accuracy of both methods on the given dataset.

data, generated more accurate relevance rankings compared to both the original image-text data and the results obtained without any data augmentation. This indicates that our data augmentation strategy, which leverages image-to-text generation to enrich the textual representation of images, effectively enhances the ability of LLMs to discern nuanced semantic relationships and perform robust relevance assessments.

This observation highlights the challenge of effectively transferring knowledge acquired through fine-tuning to complex tasks that require nuanced semantic understanding. While the fine-tuned models demonstrated some capacity for capturing relationships between images and text, they appear to have struggled to generalize to the complexities of idiomatic expressions and the subtle contextual cues required for accurate relevance assessment. This underscores the importance of leveraging the vast pre-existing knowledge and sophisticated inference mechanisms embodied in advanced LLM APIs for tasks demanding a high degree of semantic understanding and reasoning.

6 Conclusion

In this work, we successfully addressed the challenge of visually grounded idiom understanding and ranking by combining data augmentation with fine-tuned BLIP-2 and leveraging the DeepSeek-R1 API. Our experimental results demonstrated that this synergistic approach yields high accuracy in ranking images according to their relevance to idiomatic expressions. Furthermore, our investigation of alternative strategies illuminated the performance gap between fine-tuning smaller parameter models and directly utilizing large language model APIs for this complex task.

Future research will focus on exploring further possibilities for multimodal idiom understanding. Key areas of investigation include enhancing the direct comprehension capabilities of multimodal LLMs for both image and text information, and exploring the use of text-to-image generation techniques to facilitate more effective text-based ranking approaches.

References

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*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Dai, Yuqiao Liu, Lei Yang, and Yufan Fu. 2023. An Idiom Reading Comprehension Model Based on Multi-Granularity Reasoning and Paraphrase Expansion. *Applied Sciences*.

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Doh, and Eid Rodan. 2024. Improving LLM Abilities in Idiomatic Translation. *ArXiv*, abs/2407.03518.

Babak Khoshnevisan. 2019. Spilling the Beans on Understanding English Idioms Using Multimodality: An Idiom Acquisition Technique for Iranian Language Learners. *International Journal of Language, Translation and Intercultural Communication*.

George Lakoff and Mark Johnson. 1980. The Metaphorical Structure of the Human Conceptual System. *Cogn. Sci.*, 4:195–208.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*.

Siyu Long, Ran Wang, Kun Tao, Jiali Zeng, and Xinyu Dai. 2020. Synonym Knowledge Enhanced Reader for Chinese Idiom Reading Comprehension. *ArXiv*, abs/2011.04499.

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

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. *ACL*.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingmin Wu, Yuxue Hu, Yongcheng Zhang, Zeng Zhi, Guixin Su, and Ying Sha. 2024. Mitigating Idiom Inconsistency: A Multi-Semantic Contrastive Learning Method for Chinese Idiom Reading Comprehension. In *AAAI Conference on Artificial Intelligence*.

A Appendix

Table 2 gives the prompts used during experimentation.

phase	Prompt
integrative phase	You are now a master of image information integration, and I need you to help me integrate two descriptions of the same image. It is required that no key or detailed information be omitted, and at the same time, you can think and speculate on the content or scene present in the image, such as expressing emotions or possible metaphors.
sort phase	You are now an excellent expert in matching graphic and textual information and sorting relevance. Please help me complete an image sorting task. Below, I will provide you with a phrase (compound), a sentence that uses this phrase (sentence), the type of meaning of the phrase in the sentence (sentence_type), as well as the names of five images and their respective descriptive texts(image_caption). Please evaluate the relevance of these five images to the phrase I have given based on their descriptive texts, and perform a sorting task, requiring them to be sorted in descending order of relevance, with the most relevant ones written at the beginning. When evaluating relevance, you can refer to the meanings of the phrases I provided in the sentence to better understand their true meanings. Please carefully analyze and perform the sorting task. Sort the images by their names, with output formats similar to ['35234427395. png ',' 53378381715. png ',' 39938261459. png ',' 7485253662. png ',' 54879908369. png ']

Table 2: The prompts used during experimentation.