

PALI-NLP at SemEval-2025 Task 1: Multimodal Idiom Recognition and Alignment

Runyang You* Xinyue Mei* Mengyuan Zhou

Ping An Life Insurance Company of China, Ltd.

{yourunyang013, meixinyue001, zhoumengyuan425}@ping.com.cn

Abstract

Understanding idioms in multimodal contexts poses significant challenges due to data scarcity, idiomatic ambiguity, and the need for effective alignment of visual and textual inputs. In this work, we introduce MIRA (Multimodal Idiom Recognition and Alignment), a training-free framework designed to address these challenges on the SemEval-2025 Task 1 (AdMIRE) benchmark. MIRA leverages powerful closed-source large language models (LLMs) and integrates three key innovations: bias correction via in-context learning, multi-step semantic-visual fusion, and a self-revision mechanism that iteratively refines its outputs through backward verification. By systematically processing and fusing multimodal inputs, MIRA generates high-quality, fine-grained image-text representations that enhance idiom comprehension across different languages and cultural contexts. Experimental evaluations in both English and Portuguese demonstrate that our approach achieves robust performance without the need for additional training, setting a new standard for multimodal idiom recognition.

1 Introduction

The SemEval-2025 Task 1 (AdMIRE) (Pickard et al., 2025) presents a new benchmark for understanding idioms in both visual and textual forms. It poses three main challenges: 1. Data Scarcity: With limited data (102 samples for Subtask A and 20 for Subtask B), traditional training methods won't work, so we need more efficient solutions. 2. Idiomatic Ambiguity: Phrases like "panda car" can mean different things (e.g., a police car or a toy) depending on context, making classification difficult. This is further complicated by cultural and domain-specific differences. 3. Multimodal Alignment: Combining visual and textual information requires new ways to integrate these different types of data, beyond just merging features.

To address these challenges, we propose **MIRA** (Multimodal Idiom Recognition and Alignment), a training-free framework that leverages powerful closed-source large language models (LLMs). MIRA is built on three key innovations: Bias Correction via In-Context Learning – employing diverse in-context learning techniques to mitigate biases inherent in the closed-source LLM; Multi-Step Semantic-Visual Fusion – to ensure the retention of fine-grained visual details without excessive computational overhead; Self-Revision Mechanism – leveraging a backward verification process that diagnoses discrepancies, reconstructs justification chains for reliable outputs.

Our novel pipeline is designed to support cross-lingual and cross-modal idiom comprehension by systematic processing of multimodal inputs, with code available¹. In addition to the core system, our contributions include:

- By first extracting image information and step-by-step fusing it with textual data, we can effectively obtain high-quality, fine-grained image-text representations not only enhance the performance of downstream tasks.
- Synergies in-context learning with in-domain knowledge, LLM can interpret semantics in accordance with the data distribution, yielding robust and accurate results without the need for additional training.
- Through combination of advanced test-time-scaling approaches and casual inference pipeline, MIRA can accurately interpret cross-lingual, cross-modal semantics in cross-cultural contexts, ultimately securing top rankings in both English and Portuguese evaluations.

* Equal contribution.

¹<https://github.com/xinyuem1/mira>.

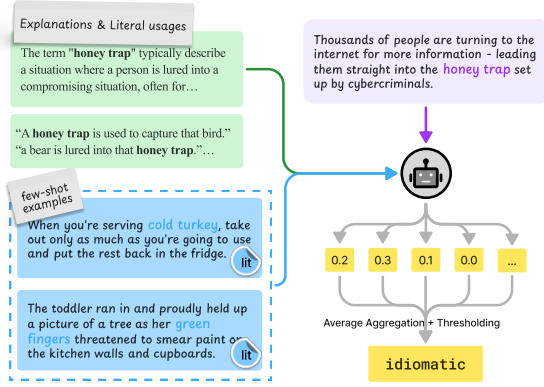


Figure 1: Pipeline for Sentence type interpolation

2 Subtask A System overview

In Subtask A, given a sentence containing a potentially idiomatic nominal compound (NC), the task is to rank five candidate images based on their relevance to the NC’s meaning in context. This requires both accurate interpretation of the NC’s sense (literal or idiomatic) and an effective ranking mechanism that aligns images with the intended meaning.

We introduce a bi-step approach to address this challenge: (1) Sentence Interpolation, and (2) Image Understanding and Ranking. This structured decomposition ensures that the system first establishes a clear understanding of the NC’s meaning before attempting image ranking, allowing for more reliable alignment between textual interpretation and visual selection.

2.1 Sentence Interpolation

2.1.1 In-context Learning

Training-free methods face inherent limitations due to their unfamiliarity with dataset-specific statistical patterns and label distributions when applied outside domain-adapted fine-tuning (Li et al., 2025; Snell et al., 2024).

Idiomatic Bias Our analysis (see Figure 2) reveals that LLMs are systematically biased toward figurative interpretations of NCs, often failing to accurately identify their literal usage. This shortcoming can be attributed to a training data bias that emphasizes prevalent figurative instantiations over the less frequent literal ones, thereby impairing the models’ ability to capture nuanced semantic distinctions in contextually ambiguous cases. We address this limitation through 2 in-context learning strategies – by providing additional exemplars and by

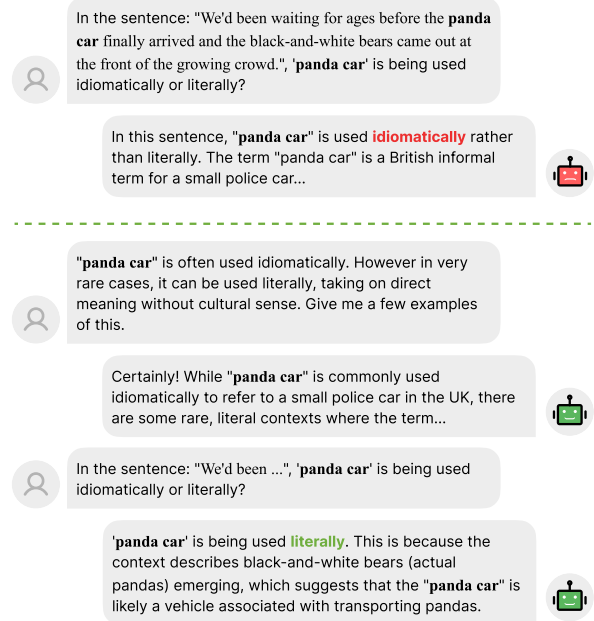


Figure 2: Case study on idiomatic bias of Zero-shot Inference

offering literal explanations of the NC, which (1) align model reasoning with domain-specific label distributions and (2) enforce output format compliance via syntactic templates.

Literal Exemplars We prompt the LLM for literal use cases before classification to counteract LLM’s figurative bias. This simple addition boosts accuracy from 91.4% to 98.6% on the English dataset, demonstrating the power of contextual exemplars.

Challenging Cases Using diagnostic sampling on the training data without exemplars, we identified 15 misclassified cases as challenging examples. At inference, we include literal cases, their explanations, and a random selection (0–2 examples) from this set to guide ambiguous predictions. Additionally, if a compound appears only once in training, its instance is added as an extra exemplar.

2.1.2 Self-Consistency Reasoning

To secure robustness and reduce prompt sensitivity, we employ self-consistency reasoning (SCR) (Wang et al., 2023), which generates multiple divergent reasoning paths with varied prompt formulations, then selects the most coherent classification via majority voting, as illustrated in Figure 1. This approach leverages the principle that diverse problem-solving trajectories often converge on the

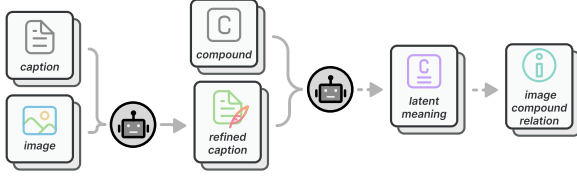


Figure 3: Multi-step image understanding pipeline

correct answer, enhancing robustness and reducing the impact of local minima.

2.2 Image Understanding

We introduce a unified pipeline that enhances compound-image relevance assessment through a 3-step pipeline – enabling fine-grained interpretation of multimedia content through interconnected stages (see Figure 3).

Cross-Modal Caption Refinement Given an image I and its associated textual caption T_I , we refine the caption by integrating both visual and textual features. Specifically, an LLM is prompted to generate an enriched caption $\hat{T}_I = \text{LLM}(T_I, f(I))$ where $f(I)$ denotes the extracted visual features of I . This multimodal alignment ensures that the refined caption \hat{T}_I captures intricate semantic details beyond the original textual description.

Latent Meaning Inference To infer deeper semantic intent, the refined caption \hat{T}_I is further analyzed to determine implicit or non-literal meanings regarding the compound C . The LLM processes \hat{T}_I and outputs a latent interpretation L : $L = \text{LLM}(C, \hat{T}_I)$, which incorporates idiomatic, metaphorical, or culturally specific insights essential for downstream classification and ranking.

Relation Scoring Leveraging the previous output L , the LLM is prompted to estimate ternary relation probabilities over three categories: *Literal* (L), *Idiomatic* (I), and *N/A* (N). The probability distribution is denoted as: $P = \{p_L, p_I, p_N\} = \text{LLM}(L)$. These probabilities serve as fine-grained relational signals, enriching the final ranking process with nuanced semantic information.

2.3 Reliable Ranking

Let $\mathbf{E}_i = \{P_{ij}, L_{ij}\}_{j=1}^5$ denote the evidence tuples for data instance i , where P_{ij} and L_{ij} represent relation probabilities and latent interpretations for the j -th candidate image. These tuples are fed into an LLM via a unified prompt to generate an initial ranking. To enhance stability, we adopt a

simple verification step inspired by (Weng et al., 2023), in which K independent forward passes are performed, followed by diagnosing ranking discrepancies, reconstructing justification chains. Finally output the refined ranking.

3 Subtask B System overview

The goal is twofold: to extend a visual narrative by selecting the most appropriate candidate image from a set of four, and to classify the NC usage as either idiomatic or literal. We address this through a two-stage pipeline: multimodal story analysis followed by usage classification. Similar to Subtask A, we also include literal use cases to tackle idiomatic bias, as described in Section 2.1.1.

Target Image Selection The LLM is prompted to describe and continue the story. Based on this continuation, the system scores the candidate images based on their likelihood to fit the narrative. SCR (explained in Section 2.1.2) is used along with average probability aggregation to enhance the robustness of the image selection process.

Sense Classification After selecting the most appropriate image, the system classifies the NC usage as idiomatic or literal, providing probability scores for both interpretations. SCR and average probability aggregation are applied to ensure accurate and robust classification.

4 Experiment

In this section, we perform a series of experiments to address the following research questions:

- How does in-context learning optimize results?
- How does visual information enhance textual features?
- How does Self-Consistency Reasoning work?

4.1 Experimental Settings

In this work, we utilize training data to conduct experiments and fine-tune our models, while also exploring a training-free framework that employs GPT-o1 as the underlying large language model (LLM) for inference.

To evaluate the performance of our approaches, we adopt a comprehensive set of metrics, including top-1 accuracy, Discounted Cumulative Gain (DCG), Normalized Discounted Cumulative Gain (NDCG), ensuring a robust assessment across both subtasks and experimental settings.

4.2 Primary Results

Subtask	Modality	Metric	Test Set	Extended Eval Set
A	English	Top 1 Acc	0.9333	0.83
		DCG Score	3.522991	3.425982
	Portuguese	Top 1 Acc	0.6923	0.7636
		DCG Score	3.207992	3.225982
B	English	Image Acc	0.6	0.9333
		Sentence Type Acc	0.8	1.0

Table 1: Performance Results for Subtask A and B

The main results of task A and B are presented in Table 1. Our approach demonstrates strong performance across both tasks, particularly in the extended evaluation set, which has a larger data volume, making the improvements even more significant. For Subtask A, our method achieved first-place rankings in all three extended evaluations. Notably, the performance for Subtask B also secured the top position, with an accuracy of 1.0 on sentence type classification task.

4.3 How does in-context learning optimize results?

Dataset	FewShot	Idiomatic	Literal	Overall
Train	None	89.7% (35/39)	64.5% (20/31)	78.6%
	+ Hard	87.2% (34/39)	87.1% (27/31)	87.1%
	+ + Same	N/A		N/A
Eval (ex)	None	87.0% (40/46)	64.8% (35/54)	75.0%
	+ Hard	89.1% (41/46)	81.5% (44/54)	85.00%
	+ + Same	93.5% (43/46)	87.0% (47/54)	90.0%

Table 2: Result on In-Context Learning Variants for Sentence Classification. Binary classification accuracy results, with the specific number of correct predictions indicated in parentheses. "None": no example, "Hard": the include hard example, and "Same": using data instance with identical NC as detailed in Section 2.1.1.

To explore optimal in-context learning strategies for this task and assess the impact of few-shot exemplars as discussed in Section 2.1.1, we examine three configurations of the in-context approach: (1) exclusion of all examples, (2) inclusion of exclusively hard examples, and (3) additional integration of training-set examples with identical NC values. Throughout all experimental conditions, the quantity of "Hard" instances remains fixed at one (serving as single-shot exemplars).

Table 2 reveals that incorporating hard examples consistently enhances overall performance across all metrics. Notably, the "Hard" configuration demonstrates greater improvements compared to baseline (+26%), indicating that hard examples

sharing distributional characteristics with the training data yield superior contextual learning benefits.

A pronounced distinction emerges in performance gains between literal and idiomatic classification accuracy. This phenomenon correlates with fundamental principles of in-context learning mechanisms - LLMs unfamiliar with the dataset’s distribution exhibit inherent discrepancies in interpreting idiomatic NCs, which typically function semantically differently than their literal counterparts. The strategic provision of representative examples addresses this representational gap by aligning the model’s contextual reasoning with the target data distribution’s statistical properties.

4.4 How does visual information enhance textual features?

To evaluate the contribution of visual information in enriching textual features, we assess the impact of Cross-Modal Caption Refinement and Latent Meaning Inference in Section sec: self-consistency through controlled experiments, where each ablation progressively removes a key processing step. The compared variants are:

- **w/o refine & latent:** directly ranks candidate images based only on their captions and the given compound, bypassing both caption refinement and latent meaning inference.
- **w/o refine:** Instead of Cross-Modal Caption Refinement, the model performs latent meaning inference using raw captions without multimodal enhancement.
- **w/o latent:** Excludes Latent Meaning Inference, directly using the refined caption and compound for Relation Scoring.

As shown in Figure 4, incorporating visual information significantly enhances the quality of textual features. Specifically, Cross-Modal Caption Refinement leads to substantial improvements in Acc@1, with gains of 15.8% and 13.0% for English and Portuguese, respectively. This highlights the limitations of raw captions and the necessity of integrating image-based enhancements.

While Latent Meaning Inference does not yield substantial improvements in Acc@1, it plays a crucial role in optimizing the overall ranking quality. By leveraging compound semantics for deeper interpretation, this step improves the DCG score by 3.0% and 4.0% in English and Portuguese, respectively. These findings suggest that multimodal

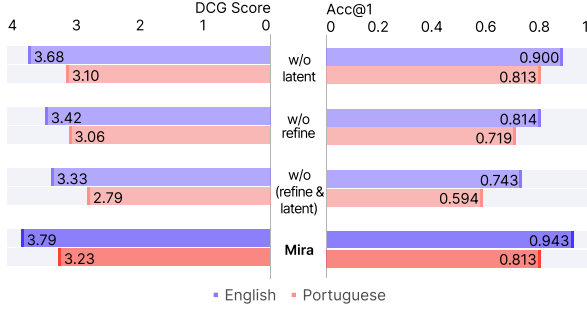


Figure 4: Impact of visual information on ranking performance. The left plot shows DCG Score, and the right plot shows Acc@1. Purple represents English, while red represents Portuguese. Incorporating visual features significantly improves Acc@1, while latent meaning inference enhances overall ranking quality (DCG Score).

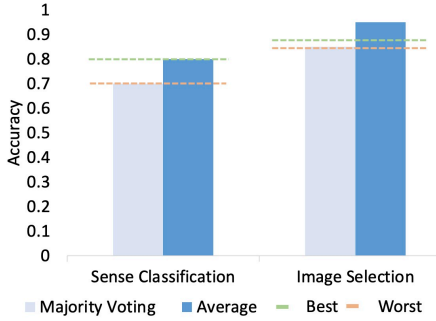


Figure 5: Comparison of Majority Voting and Mean Aggregation for task 1: sense classification and task 2: image selection accuracy. The red line represents the worst performance among individual results before aggregation, while the green line highlights the best performance.

alignment enhances both direct selection accuracy and the ranking consistency of candidate images.

4.5 How does Self-Consistency Reasoning work?

Table 5 below presents a comparison of the performance of Majority Voting and Mean Aggregation for 2 tasks in Subtask B.

From the results, we observe that Average Aggregation consistently outperforms Majority Voting across both tasks. For Sense Classification, Average Aggregation achieves a mean score of 0.8, compared to Majority Voting’s 0.7. Similarly, in the Image Selection task, Average Aggregation shows a stronger performance, with a mean of 0.95 versus Majority Voting’s 0.85. Additionally, the Worst and Best values for both tasks reflect the relative stability of Average Aggregation.

5 Conclusion

This work presents MIRA, a training-free framework for multimodal idiom comprehension. Leveraging powerful closed-source language models, MIRA overcomes data scarcity, idiomatic ambiguity, and multimodal alignment challenges through three core components: visual-text fusion, in-context learning, and self-consistency reasoning. Visual-text fusion extracts fine-grained visual details and integrates them with text to create high-quality representations. In-context learning leverages in-domain knowledge to ensure semantic interpretations align with the data distribution, while self-consistency reasoning aggregates multiple reasoning paths to mitigate errors and enhance reliability. Together, these components form a robust causal inference pipeline that has achieved top rankings in both English and Portuguese evaluations and adapts efficiently to low-resource multilingual scenarios. Future work will explore enhanced chain-of-thought reasoning and zero-shot debiasing to further expand its applicability.

6 Related Work

Multi-modal Understanding Multimodal understanding, the integration of visual and textual data, faces significant challenges in alignment issues, noise resilience, and disparities in feature representation (Masry et al., 2025; Li and Tang, 2024). While studies have highlighted the importance of leveraging complementary information across modalities for accuracy and applicability, training-free methods remain limited (Chen et al., 2025). This framework have explored a pipeline path-way to tackle these challenges through multi-steps thinking, thus dynamically align semantic features without fine-tuning.

Test-time scaling and LLM Reasoners The concept of test-time scaling, where increased compute at test time leads to better results, has gained traction in the context of LLMs (Xu et al., 2025). Recent work using models like OpenAI’s o1 and Deepseek-r1 (Sui et al., 2025) demonstrated superior performance through scaled test-time computation. Our work leverages these insights by employing self-consistency reasoning (Wang et al., 2023; Yao et al., 2023) to enhance performance without fine-tuning, aligning with the goal of efficient and scalable reasoning in LLMs.

References

- Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. 2025. [Multimodal Representation Alignment for Image Generation: Text-Image Interleaved Control Is Easier Than You Think](#). *arXiv preprint*. ArXiv:2502.20172 [cs].
- Songtao Li and Hao Tang. 2024. [Multimodal Alignment and Fusion: A Survey](#). *arXiv preprint*. ArXiv:2411.17040 [cs].
- Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. 2025. [Test-Time Preference Optimization: On-the-Fly Alignment via Iterative Textual Feedback](#). *arXiv preprint*. ArXiv:2501.12895 [cs].
- Ahmed Masry, Juan A. Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, Sathwik Tejaswi Madhusudan, Marco Pedersoli, Bang Liu, Nicolas Chapados, Yoshua Bengio, Enamul Hoque, Christopher Pal, Issam H. Laradji, David Vazquez, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. 2025. [Align-VLM: Bridging Vision and Language Latent Spaces for Multimodal Understanding](#). *arXiv preprint*. ArXiv:2502.01341 [cs].
- Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters](#). *arXiv preprint*. ArXiv:2408.03314 [cs].
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025. [Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models](#). *arXiv preprint*. ArXiv:2503.16419 [cs].
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. [Large Language Models are Better Reasoners with Self-Verification](#). *arXiv preprint*. ArXiv:2212.09561 [cs].
- Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. [Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models](#). *arXiv preprint*. ArXiv:2501.09686 [cs].
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of Thoughts: Deliberate Problem Solving with Large Language Models](#). *arXiv preprint*. ArXiv:2305.10601 [cs].

Reproducibility Details

Hyper-parameter Configuration

Subtask	Stage	T	p	n
Task A	Sentence Interpolation	1	1	16
	Cross-Modal Caption Refinement	0.1	0.1	–
	Latent Meaning Inference	0.1	0.2	–
	Relation Scoring	0.1	0.2	–
Task B	Reliable Ranking	1	0.1	16
	Target Image Selection	1	1	16
	Sense Classification	1	1	16

Table 3: Key hyper-parameters (T : temperature, p : top- p). n denotes the number of reasoning paths sampled for each input. All other parameters are left at OpenAI defaults.

Given in Table 3 are the hyperparameters used. Seed is set to 42.

Implementation Details

- **Prompt Templates:** Full prompt specifications for Latent Meaning Inference, Relation Scoring, and Self-Revision are available in the repository.
- **Preprocessing:** Input texts undergo base64 encoding without additional transformations.
- **Training Configuration:** LLM was used in inference-only mode; cross-validation was omitted consistent with zero-shot evaluation.