ITUNLP2 at MWE-2026 AdMIRe 2: Modular Zero-Shot Pipelines for Multimodal Idiom Grounding and Ranking

Özge Umut and Bora Şenceylan
Istanbul Technical University
Türkiye
{umut24,senceylan19}@itu.edu.tr

Abstract

We describe a zero-shot system for AdMIRe 2, a shared task on multimodal understanding of potentially idiomatic expressions (PIEs) (Arslan et al., 2026). Given a context sentence with a PIE and five candidate images, the system predicts whether the usage is literal or idiomatic and ranks the images based on how well they match the intended meaning. We use closed-source large multimodal models and compare prompting pipelines from direct one-step ranking to modular multi-step pipelines that separate sense prediction, PIE-focused image semantics, and final ranking. All steps produce constrained JSON outputs to enable deterministic parsing and composition. In the official AdMIRe 2 evaluation on CodaBench, our best pipeline achieves an average Top-1 accuracy of 0.52 and an average nDCG score of 0.70 across the 12 languages we submitted. We obtain the best score among submitted systems in 10 of these languages (cod, 2026).

1 Introduction

Idioms are hard because the intended meaning is often not compositional. In AdMIRe 2, each example provides a context sentence containing a potentially idiomatic expression (PIE) and five candidate images; systems must decide whether the PIE is literal or idiomatic and rank the images by semantic match (Arslan et al., 2026). This setting links sense disambiguation to grounded retrieval, and errors can come from either the meaning decision or the fine-grained ordering among several plausible images.

Recent prompting work shows that forcing intermediate steps can improve reliability on tasks that require multi-step reasoning (Wei et al., 2022; Kojima et al., 2022), and similar ideas have been applied to multimodal reasoning (Zhang et al., 2023). Following this direction, we build modular pipelines where the model first commits to a PIE sense, then produces PIE-centered image semantics, and only then produces a final ranking. We evaluate multiple pipeline variants with OpenAI-o3 and Gemini-3-pro-preview, and we analyze when ordering is the main remaining failure mode. Complete code for our pipeline variants is publicly available [1] .

Our contributions are:

- a set of zero-shot modular pipelines for multimodal idiom grounding, implemented with structured JSON interfaces;
- an ablation-style comparison across pipeline depths and aggregation strategies;
- official multi-language evaluation results from the AdMIRe 2 CodaBench phase (cod, 2026).

2 Related Work

Idiomaticity has been studied widely in text-only settings, including idiom corpus construction tasks (Eryigit et al., 2021), (Arslan et al., 2025) and multilingual idiom detection shared tasks (Tayyar Madabushi et al., 2022). AdMIRe extends this line by requiring multimodal grounding: a PIE in context must be matched to images that represent either idiomatic or literal meaning (Pickard et al., 2025), and many strong systems use multi-stage designs rather than a single direct ranking.

Several AdMIRe system papers explicitly separate sense disambiguation from visual ranking. AlexUNLP-NB predicts literal vs. idiomatic usage, derives a literalized meaning signal, and then performs retrieval-style ranking (Badran et al., 2025). PALI-NLP refines image descriptions with PIE-relevant details and applies a revision step before ranking (You et al., 2025). Other approaches strengthen the retrieval backbone with fine-tuning

[1] GitHub repository for pipeline codes

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 248–253
March 28, 2026 ©2026 Association for Computational Linguistics

or ensembling (Wang et al., 2025). For the vision–language similarity component, CLIP (Radford et al., 2021) and newer contrastive objectives such as SigLIP (Zhai et al., 2023) are common choices, but performance still depends strongly on the text representation used for matching. Our work stays in a training-free setting and focuses on how stepwise prompting and intermediate semantic signals affect ranking quality.

3 Methodology

3.1 Dataset

To design and evaluate our pipelines, we used the SemEval-2025 Task 1 (AdMIRe) dataset (Pickard et al., 2025) which includes examples in both English and Portuguese. Each example consists of a context sentence containing a potentially idiomatic expression (PIE) and 5 candidate images encompassing : Idiomatic Synonym, Idiomatic Related, Literal Synonym, Literal Related and a Distractor. The English subset comprises 200 sentences covering 100 unique PIEs, while the Portuguese subset contains 110 sentences with 55 unique PIEs. After finalizing the pipeline, final evaluation was conducted using the AdMIRe 2 blind test set via CodaBench (cod, 2026), which extends the task to 15 diverse languages (Torunoğlu-Selamet et al., 2026).

3.2 System Overview

For addressing the challenge of ranking images based on their relevance to a context sentence containing a PIE, we employed a zero-shot framework. Analysis of earlier SemEval-2025 Task 1: AdMIRe - Advancing Multimodal Idiomaticity Representation results indicated that closed-source multimodal models demonstrated strong performance. Therefore, we used Gemini-3-pro-preview and OpenAI-o3 for our experiments. As it was demonstrated in recent research into Large Multimodal Models (LMMs) (Khot et al., 2023; Khan et al., 2023), decomposing complex tasks into modular sub-tasks significantly improves performance in both text-only and vision-language settings. Consistent with these studies, most participating systems in SemEval Task 1 adopted multi-step reasoning pipelines. Our methodology builds upon this principle of modular reasoning by decomposing the reasoning process of the models into discrete, structured steps iteratively, improving image-text matching accuracy. We developed and evaluated seven distinct pipelines, progressively increasing the granularity of the reasoning process. For each step of the pipelines, all model outputs were constrained to JSON format with pre-defined fields depending on its sub-task to ensure reproducible data extraction and pipeline integration. Additionally, for each sub-step, models were prompted to provide explanations of its reasoning for analytical purposes. The explanation of each pipeline is given below:

3.2.1 Pipeline A: One Step Direct Ranking (Baseline)

In this single-step approach, the model is given the context sentence, the PIE, and the five candidate images simultaneously and asked to directly output a ranked list and an explanation of its reasoning. The results of this pipeline are utilized to assess the impact of breaking the task into sub-tasks on performance.

3.2.2 Pipeline B: Two-step (Sense Prediction + Ranking)

We next introduce a two-step pipeline to explicitly see how model interprets the PIE in the context sentence. In the first step model is asked to predict whether the PIE is used literally or idiomatically in the given context sentence. In the second step, model is asked to rank the images utilizing the predicted sense from step 1. In each step model is prompted to provide a brief explanation on its reasoning.

When these explanations are inspected in detail, it was noticed that during image interpretation, models tend to over-analyze other aspects of the sentence, rather than PIE itself. For example, in the sentence “Lyn says that her relationship with Paul is ancient history, but Steph thinks that she should go and see him to give him a chance to apologize", the model ranked an image depicting a couple holding hands last (as distractor), justifying the decision by noting that it showed an ongoing relationship. This reasoning reflects that model paid attention to the general relationship theme more than PIE "ancient history". To mitigate this, we introduced next three-stage modular pipeline:

3.2.3 Pipeline C: Three-step (Sense Prediction + Image Semantics + Ranking)

As in pipeline B, the model was asked to identify literal or idiomatic usage. In step 2, with the aim of the model to focus only on the PIE, given each image and their captions, we ask the model to categorize them into five predefined categories: Idiomatic Synonym, Idiomatic Related, Literal Synonym, Literal Related, Distractor; provided only the PIE, not the context sentence. In the third step the model receives the outputs from steps 1 and 2 and asked to produce the final ranking. This stage acts as an aggregator, weighing the specific PIE-to-image classification against the broader sentence context.

3.2.4 Pipeline D: Three-step 2 (Sense Prediction + Image Semantics + Manual Ranking)

This pipeline extends the three-step approach by replacing model-based ranking with a rule-based manual ranking strategy in the final stage. The first and second steps are the same as the previous pipeline, in the third step images are ranked deterministically based on following rule: If the sentence is predicted as literal, images are ordered as: literal synonym, literal related, idiomatic related, idiomatic synonym, distractor. If the sentence is predicted as idiomatic, images are ordered as: idiomatic synonym, idiomatic related, literal related, literal synonym, distractor. The objective of this pipeline is to determine whether aggregating the outputs of the first two steps using explicit rule yields better performance than relying on an LLM to perform the final ranking.

3.2.5 Pipeline E: Four-Step 1 (Caption Refinement and Extension)

This pipeline introduces an explicit image caption refinement stage prior to image classification step to test whether richer visual descriptions improve the image semantic classification. Models were asked to expand the original caption into a more detailed description and add concrete visual details. The rest of the steps are the same as pipeline C.

3.2.6 Pipeline F: Four-Step 2 (Explicit PIE–Image Grounding)

This pipeline focuses on strengthening explicit semantic grounding between the PIE and each image before ranking. In step 1, model is asked to predict whether the PIE is used literally or idiomatically, in step 2 model explicitly explains the relation between the PIE and each image, in step 3, using the explanations generated in the previous step, each image is assigned to its semantic category. In the last step, the model is asked to rank the images based on the information from previous steps.

3.2.7 Pipeline G: Four-Step 3 (Explicit Grounding + Manual Ranking)

This pipeline mirrors the previous four-step approach but replaces model-based ranking with manual rule-based aggregation as described in pipeline D.

3.3 Evaluation Metrics

Each pipeline is evaluated using Top-1 Accuracy, Exact Match Accuracy, Top-2 Accuracy and Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002). Top-1 Accuracy measures the proportion of instances in which the most relevant image is predicted correctly, Top-2 Accuracy measures how often the images predicted in the first and second positions are correct. Exact-Match Accuracy requires the entire predicted ranking of images to exactly match the ground-truth order. It measures how well the model orders all images correctly. Normalized Discounted Cumulative Gain used with weights: [3, 1, 0, 0, 0] to assess overall ranking quality and break ties.

4 Results

The experimental results for our different pipeline configurations in the English portion of the SemEval-2025 Task 1 (AdMIRe) dataset (Pickard et al., 2025) with OpenAI-o3 and Gemini-3-pro-preview are summarized in Tables 1 and 2. Overall, the results show that multi-step pipelines consistently outperform simpler 1-step and 2-step approaches, supporting the previous research demonstrating decomposing complex tasks into modular sub-tasks significantly improves performance in vision-language settings.

OpenAI-o3 Results: For the OpenAI-o3 model (Table 1), the 1-step baseline (Pipeline A) achieved a Top-1 Accuracy of 0.82. This was exceeded by all modular pipelines, with the Pipeline F: 4-step 2 approach reaching the highest Top-1 score of 0.90, followed closely by Pipeline C: 3-step 1 at 0.89. A similar trend appeared in the Discounted Cumulative Gain (DCG) scores, which rose from 2.83 (Pipeline A: 1-step) to a peak of 3.16 (Pipeline C: 3-step 1).

Due to the small performance difference between the 3-step and 4-step pipelines, and considering the significantly higher computational cost and latency of 4-step pipelines, we focused our experiments with Gemini-3-pro-preview on 1-step and 3-step configurations.

Table 1: Zero-Shot Evaluation Results for OpenAI-o3 Model

Metric	Pipeline A	Pipeline B	Pipeline C	Pipeline D	Pipeline E	Pipeline F	Pipeline G
Top-1 Acc.	0.82	0.83	0.89	0.84	0.88	**0.90**	0.87
Exact-Match Acc	0.09	0.09	0.235	**0.63**	0.21	0.14	0.61
Top-2 Acc.	0.60	0.60	**0.76**	0.75	0.74	0.72	0.75
Average DCG	2.83	2.87	**3.16**	2.99	3.11	3.15	3.06

Gemini-3-pro-preview Results: As shown in Table 2, the performance improvements from multi-step decomposition are even more pronounced for Gemini than for OpenAI-o3; the Pipeline C: 3-step 1 configuration improved Top-1 Accuracy from 0.81 to 0.92. Overall, Gemini outperformed o3 in most metrics, except for exact-match accuracy in specific 3-step settings.

Table 2: Zero-Shot Evaluation Results for Gemini-3-pro-preview

Metric	Pipeline A	Pipeline C	Pipeline D
Top-1 Acc.	0.81	**0.92**	0.91
Exact-Match Acc.	0.08	0.04	**0.75**
Top-2 Acc.	0.60	0.78	**0.83**
Avg. DCG	2.79	3.26	**3.27**

Hybrid Evaluation: To investigate whether the two models provide complementary strengths, we also evaluated a hybrid zero-shot configuration (Table 3), using Gemini-3-pro-preview for sense prediction and image classification and OpenAI-o3 for final ranking. While this achieved a competitive Top-1 Accuracy of 0.904, it did not outperform the standalone Gemini 3-step pipeline.

Table 3: Hybrid Zero-Shot Results (Gemini + o3)

Metric	Score
Top-1 Acc.	0.90
Exact-Match Acc.	0.20
Top-2 Acc.	0.72
Avg. DCG	3.17

Cross-Lingual Transfer: To assess cross-lingual generalization, we applied the same 3-step pipeline to a Portuguese language split without modifying the overall approach. As shown in Table 4, Gemini again achieves higher Top-1 accuracy and DCG, while OpenAI-o3 performs slightly better on exact-match accuracy. The overall gap between models is relatively small, suggesting that ranking quality, rather than language-specific understanding, remains the primary issue.

Table 4: Zero-Shot Pipeline C: 3 step 1 Results on Portuguese Data

Metric	Gemini	o3
Top-1 Acc.	**0.88**	0.84
Exact-Match Acc.	0.05	**0.11**
Top-2 Acc.	**0.62**	0.49
Avg. DCG	**3.06**	2.86

The "Ordering" Challenge: Model vs. Heuristic Ranking: When overall results were analyzed for both models, it is observed that the exact-match accuracy is the most challenging metric across all experiments, as it requires the entire image ranking to be correct. In fully model-driven pipelines, exact-match remains low for both models. However, when manual rule-based ranking is introduced (Pipeline D: 3-step 2 and Pipeline G: 4-step 3), exact-match accuracy increases substantially. For example, OpenAI-o3 improves from 0.09 in the 1-step baseline to 0.63 in the 3-step 2 pipeline, while Gemini reaches 0.75 in the same setting. This sharp improvement indicates that many remaining errors arise not from incorrect PIE interpretation or image classification, but from ordering decisions among multiple partially relevant images.

Final System Selection: Based on these findings, we selected Gemini-3-pro-preview with the Pipeline C: 3-step 1 pipeline as our final system for generating predictions on the AdMIRe 2 competition test set via CodaBench. This configuration offered the best trade-off between performance and computational cost. In our official submission, we evaluated 12 languages; Greek, Serbian, and Slovak were not included because our final three-step setup requires three separate LLM calls per instance, and completing the full blind test across all 15 languages was not feasible within the submission deadline given API latency and rate limits. Our final submission achieved 3rd place overall in the competition and ranked 1st in 10 out of the 12 languages we participated in, indicating that the proposed modular reasoning setup performs reliably across languages. The complete results are presented in Table 5. Performance varies notably across languages. The system performs strongest

Table 5: Competition Results Across Languages

Metric	AVG	ZH	KA	IG	KK	NO	PT-BR	PT-PT	RU	SL	ES-EC	TR	UZ
Top-1 Accuracy	0.52	0.53	0.56	0.56	0.70	0.82	0.88	0.72	0.71	0.82	0.40	0.65	0.52
nDCG Score	0.70	0.81	0.82	0.84	0.89	0.94	0.96	0.91	0.91	0.94	0.79	0.88	0.83

on Brazilian Portuguese (PT-BR), Norwegian (NO), Slovenian (SL), where both Top-1 accuracy and nDCG exceed 0.80, suggesting robust cross-lingual generalization for these languages. In particular, PT-BR achieves the highest scores (Top-1: 0.88, nDCG: 0.96), indicating effective semantic alignment in this language. Worst performance is observed in Ecuadorian Spanish (ES-EC). The limited data for this language with fewer than 50 samples may explain its position as the lowest performer in the dataset.

5 Conclusion

We presented a zero-shot, modular prompting system for multimodal idiom grounding in AdMIRe 2 (Arslan et al., 2026). Across two closed-source multimodal models, multi-step pipelines improve Top-1 accuracy and ranking quality compared to direct one-step ranking. Manual aggregation experiments show that a large part of the remaining gap comes from ordering decisions among partially relevant images, not only from sense disambiguation. In the official evaluation phase on CodaBench, the selected pipeline obtained strong cross-lingual results across the 12 languages we entered (cod, 2026).

Limitations

Our approach relies entirely on closed-source large multimodal models in a zero-shot setting. As a result, the reproducibility of our pipelines depend on the availability, behavior, and pricing policies of external APIs. Differences in model versions, hidden system prompts, or parameter updates may affect performance over time.

The modular design increases the number of calls per instance. This improves performance in our experiments, but it also increases cost and latency, and it becomes harder to run large batches under API rate limits. For the official evaluation, these constraints affected our submission, and we evaluated 12 languages instead of all 15.

References

2026. Admire 2.0 shared task: Codabench competition results. https://www.codabench.org/competitions/10547/#/results-tab. Accessed: 2026-01-06.

Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit, and Joakim Nivre. 2025. Using LLMs to advance idiom corpus construction. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Mohamed Badran, Youssef Nawar, and Nagwa El-Makky. 2025. AlexUNLP-NB at SemEval-2025 task 1: A pipeline for idiom disambiguation and visual representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 546–550, Vienna, Austria. Association for Computational Linguistics.

Gülsen Eryigit, Ali Sentas, and Johanna Monti. 2021. Gamified crowdsourcing for idiom corpora construction. *CoRR*, abs/2102.00881.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker, and Yun Fu. 2023. Exploring question decomposition for zero-shot VQA. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

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 Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Yanan Wang, Dailin Li, Yicen Tian, Bo Zhang, Jian Wang, and Liang Yang. 2025. dutir914 at SemEval-2025 task 1: An integrated approach for multimodal idiomaticity representations. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1198–1203, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Runyang You, Xinyue Mei, and Mengyuan Zhou. 2025. PALI-NLP at SemEval-2025 task 1: Multimodal idiom recognition and alignment. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1211–1216, Vienna, Austria. Association for Computational Linguistics.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Zhuosheng Zhang, Aston Zhang, and Mu Li. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.