Modular Zero-Shot Pipelines for Multimodal Idiom Grounding and Ranking

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

Relevant UniDive working groups: WG3

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. In the official AdMIRe 2 evaluation, 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 challenging for multimodal systems because their meanings are often non-compositional. The AdMIRe 2 shared task requires systems to determine if a context sentence containing a potentially idiomatic expression (PIE) is literal or idiomatic and then rank five candidate images based on semantic match (Arslan et al., 2026). We propose a zero-shot framework using closed-source Large Multimodal Models —OpenAI-o3 and Gemini-3-pro-preview—to solve this via modular, multi-step reasoning.

Our contributions include a set of modular pipelines for multimodal idiom grounding. We present an ablation-style comparison across varying pipeline depths and aggregation strategies to identify the most effective reasoning chains (Umut and Şenceylan, 2026).

2 Methodology

2.1 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).

2.2 System Overview

For addressing ADMIRE 2 challenge, we employed a zero-shot framework. We developed and evaluated seven distinct pipelines, progressively increasing the granularity of the reasoning process.

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.

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

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, given each image, their captions and PIE, we ask the model to categorize them into five predefined categories: Idiomatic Synonym, Idiomatic Related, Literal Synonym, Literal Related, Distractor. In the third step the model receives the outputs from steps 1 and 2 and asked to produce the final ranking.

Pipeline D: Three-step 2 (Sense Prediction + Image Semantics + Manual Ranking): This pipeline extends the three-step approach by replac-

ing model-based ranking with a rule-based manual ranking strategy in the final stage. 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.

Pipeline E: Four-Step 1 (Caption Refinement and Extension): This pipeline introduces an explicit image caption refinement stage prior to image classification step. Models were asked to expand the original caption into a more detailed description.

Pipeline F: Four-Step 2 (Explicit PIE–Image Grounding): This pipeline focuses on strengthening explicit semantic grounding between the PIE and each image. Model is explicitly asked to explain the relation between the PIE and each image before image classification step.

Pipeline G: Four-Step 3 (Explicit Grounding + Manual Ranking): This pipeline mirrors the previous four-step approach but uses manual rule-based aggregation as described in pipeline D.

2.3 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.

3 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.

For the OpenAI-o3 model, 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 reaching the highest Top-1 score, followed closely by Pipeline C. A similar trend appeared in the Discounted Cumulative Gain (DCG) scores.

As shown in Table 2, the performance improvements from multi-step decomposition are even more pronounced for Gemini than for OpenAI-o3. Overall, Gemini outperformed o3 in most metrics.

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

Pipeline	Top-1 Acc.	Exact-Match Acc.	Top-2 Acc.	Average DCG
Pipeline A	0.82	0.09	0.60	2.83
Pipeline B	0.83	0.09	0.60	2.87
Pipeline C	0.89	0.24	**0.76**	**3.16**
Pipeline D	0.84	**0.63**	0.75	2.99
Pipeline E	0.88	0.21	0.74	3.11
Pipeline F	**0.90**	0.14	0.72	3.15
Pipeline G	0.87	0.61	0.75	3.06

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

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

Cross-Lingual Transfer: To assess cross-lingual generalization, we applied the same 3-step pipeline to a Portuguese language split. As shown in Table 3, Gemini again achieves higher Top-1 accuracy and DCG.

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

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

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. In our official submission, we evaluated 12 languages. Our final submission achieved 3rd place overall in the competition and ranked 1st in 10 out of the 12 languages we participated in. The complete results are presented in Table 4. Performance varies notably across languages. The system performs strongest on Brazilian Portuguese, Norwegian, Slovenian, 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:

Table 4: Competition Results Across Languages

Language	Top-1 Accuracy	nDCG Score
Chinese	0.53	0.81
Georgian	0.56	0.82
Igbo	0.56	0.84
Kazakh	0.70	0.89
Norwegian	0.82	0.94
Brazilian Portuguese	0.88	0.96
Portugal Portuguese	0.72	0.91
Russian	0.71	0.91
Slovenian	0.82	0.94
Ecuadorian Spanish	0.40	0.79
Turkish	0.65	0.88
Uzbek	0.52	0.83
Average	**0.52**	**0.70**

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.

4 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.

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, and behavior of external APIs. Differences in model versions, hidden system prompts, or parameter updates may affect performance over time.

References

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

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.

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

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.

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, Carlos Manuel Hidalgo-Ternero, Chaya Liebeskind, Danka Jokić, Diego Alves, Eleni Triantafyllidi, Erik Velldal, Fred Philippy, Giedre Valunaite Oleskeviciene, Ieva Rizgeliene, Inguna Skadina, Irina Lobzhanidze, Isabell Stinessen Haugen, Jauza Akbar Krito, Jelena M. Marković, Johanna Monti, Josue Alejandro Sauca, Kaja Dobrovoljc, Kingsley O. Ugwuanyi, Laura Rituma, Lilja Øvrelid, Maha Tufail Agro, Manzura Abjalova, Maria Chatzigrigoriou, María del Mar Sánchez Ramos, Marija Pendevska, Masoumeh Seyyedrezaei, Mehrnoush Shamsfard, Momina Ahsan, Muhammad Ahsan Riaz Khan, Nathalie Carmen Hau Norman, Nilay Erdem Ayyıldız, Nina Hosseini-Kivanani, Noémi Ligeti-Nagy, Numaan Naeem, Olha Kanishcheva, Olha Yatsyshyna, Daniil Orel, Petra Giommarelli, Petya Osenova, Radovan Garabik, Regina E. Semou, Rozane Rebechi, Salsabila Zahirah Pranida, Samia Touileb, Sanni Nimb, Sarfraz Ahmad, Sarvinoz Nematkhonova, Shahar Golan, Shaoxiong Ji, Sopuruchi Christian Aboh, Srdjan Sucur, Stella Markantonatou, Sussi Olsen, Vahide Tajalli, Veronika Lipp, Voula Giouli, Yelda Yeşildal Eraydın, Zahra Saaberi, and Zhuohan Xie. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding.

Özge Umut and Bora Şenceylan. 2026. Itunlp2 at mwe-2026 admire 2: Modular zero-shot pipelines for multimodal idiom grounding and ranking. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 248–253.