A Zero-Shot LLM Pipeline for Multimodal Idiom Understanding and Ranking

Atakan Site, Oğuz Ali Arslan, Gülşen Eryiğit
Department of Artificial Intelligence and Data Engineering
Istanbul Technical University
{site21, arslanog20, gulsenc}@itu.edu.tr

Relevant UniDive working groups: WG3: Multilingual and Cross-Lingual Language Technology

1 Introduction

Idioms and potentially idiomatic expressions (PIEs) remain challenging for multilingual and cross-lingual language technology because their intended meaning is often non-compositional, context-dependent, and not directly recoverable from the surface form alone (Haagsma et al., 2020; Tayyar Madabushi et al., 2022; De Luca Fornaciari et al., 2024; Mi et al., 2025). These difficulties become sharper in multimodal settings, where a system must decide whether an image reflects the literal scene evoked by the words or the figurative meaning intended in context (Yosef et al., 2023; Saakyan et al., 2025). The problem is especially relevant for underrepresented languages, where grounded multilingual benchmarks remain scarce.

This abstract summarizes a recently published paper on the AdMIRe 2 shared task (Site et al., 2026). AdMIRe 2 evaluates multilingual multimodal idiom understanding across 15 languages and two tracks: a text-only track, where systems reason over sentences and image captions, and a multimodal track, where systems additionally inspect the images themselves (Arslan et al., 2026). The task is highly relevant to UniDive WG3 because idiomatic expressions are a natural locus of semantic divergence: translation equivalents are often partial, figurative meaning may conventionalize differently across languages, and literal versus idiomatic interpretation can hinge on subtle contextual cues.

2 Background

Earlier work on idiomaticity focused mainly on text-only datasets and classification settings. Resources such as MAGPIE established large-scale benchmarks for potentially idiomatic expressions in context (Haagsma et al., 2020), while SemEval-2022 Task 2 extended the problem to multilingual idiomaticity detection and idiom-aware sentence representations (Tayyar Madabushi et al., 2022). More recent analyses have shown that even strong LLMs often over-rely on lexical cues and still struggle when literal and idiomatic readings must be disambiguated from context (De Luca Fornaciari et al., 2024; Mi et al., 2025).

In parallel, multimodal benchmarks have shown that grounded figurative understanding is harder still. IRFL evaluates whether models can match figurative expressions to suitable images, and V-FLUTE frames visual figurative-language understanding as an explainable entailment problem (Yosef et al., 2023; Saakyan et al., 2025). The first AdMIRe task brought these lines of work together by treating multimodal idiomaticity as an image-based representation problem (Pickard et al., 2025). AdMIRe 2 broadens this setting to 15 languages and two complementary tracks, offering a useful evaluation bed for multilingual robustness, semantic divergence, and cross-lingual comparison (Arslan et al., 2026).

3 Methodology and Main Findings

The published system proposes a training-free, zero-shot inference pipeline built around large vision–language models rather than task-specific fine-tuning (Site et al., 2026). The architecture supports both AdMIRe 2 tracks: in the text-only setting, the model reasons over the context sentence and candidate image captions, while in the multimodal setting it additionally inspects the candidate images themselves. A structured chain-of-thought prompting strategy decomposes the task into five steps: deciding whether the PIE is used literally or idiomatically in context; reviewing candidate captions and, when available, images; evaluating each candidate with respect to the inferred meaning; generating a global ranking; and returning the result in

Track	All	Lit	Id	NDCG@5
Text-only	55.9	61.1	51.3	.831
Multimodal	60.1	66.9	54.8	.849

Table 1: Average AdMIRe 2 results across 15 languages from Site et al. (2026). All, Lit, and Id denote Top-1 Accuracy (%). See the main publication for detailed per-language results and model-level analyses.

a structured JSON format. A central design principle is that candidates are ranked according to how well they represent the *meaning* of the expression, not according to superficial overlap with incidental details of the sentence.

The system further combines a primary model with a fallback model to improve robustness in multilingual inference. More specifically, Gemini 2.5 Pro (Comanici et al., 2025) is used as the primary model and GPT-5.1 (OpenAI, 2025) as the fallback. The primary model is queried first; when it refuses or returns an unusable output, a fallback model is triggered. Lightweight post-processing then recovers rankings from imperfect JSON responses through direct parsing with a regex-based fallback, reducing failures caused by safety filters, formatting errors, and expression-specific instability. This design keeps the system fully zero-shot while improving robustness across multilingual inputs. To support reproducibility and future comparative work, the code for the published system is publicly available on GitHub.[1]

Table 1 summarizes the core results. The approach ranked first overall on the official AdMIRe 2 leaderboard when performance was averaged across both tracks (Site et al., 2026). Visual input improved results in 14 of the 15 languages, showing that grounding usually helps disambiguate literal and idiomatic readings, although gains vary with caption quality, visual grounding, and idiom transparency. Literal cases also remain easier than idiomatic ones in both tracks, suggesting that current large vision–language models benefit more from concrete, photographable content than from abstract figurative meaning.

Despite these strong results, several limitations remain. The approach depends on large proprietary vision–language models and API-based inference, which may limit accessibility and reproducibility. Performance also varies substantially across languages, and the benefit of multimodal grounding is not uniform across all settings. In addition, idiomatic cases remain consistently harder than literal ones, highlighting the continuing difficulty of grounded figurative understanding.

[1] github.com/oguzaliarslan/idiom-nlp.

4 Relevance to UniDive WG3

This work aligns closely with WG3 for three reasons. First, it evaluates multilingual and cross-lingual language technology on a phenomenon that is inherently divergence-sensitive: idiomatic expressions frequently involve non-compositional semantics, partial translation mismatch, and language-specific conventionalization. Second, it shows the value of evaluation settings that move beyond sentence-level labeling and require grounded semantic decisions across modalities. Third, it demonstrates that carefully designed zero-shot multilingual pipelines can already be competitive while also revealing where transfer remains fragile, especially for figurative meaning and underrepresented languages.

More broadly, the study contributes to ongoing discussions in UniDive about how semantic divergence should be operationalized and measured in multilingual NLP. AdMIRe 2 is useful not only as a strong empirical benchmark, but also as a methodological example of how multilingual multimodal evaluation can expose limitations that remain hidden in standard text-only tasks. In this sense, the work provides both a concrete published outcome and a basis for future collaboration on cross-lingual evaluation, benchmark design, and multimodal language technology within WG3.

References

Dogukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoglu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gulsen Eryigit. 2026. MWE-2026 shared task 2: AdMIRe 2 – advancing multimodal idiomaticity representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut

to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. GPT-5.1 Instant and GPT-5.1 Thinking system card. OpenAI system card. Technical report.

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.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. Preprint.

Atakan Site, Oguz Ali Arslan, and Gulsen Eryigit. 2026. ITUNLP at MWE-2026 AdMIRe 2: A Zero-Shot LLM Pipeline for Multimodal Idiom Understanding and Ranking. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 226–236.

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)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.