3K2T at MWE-2026 AdMIRe 2: CARIM– Category-Aware Reasoning for Idiomatic Multimodality

Kubilay Kağan KÖMÜRCÜ
Istanbul Technical University
komurcu17@itu.edu.tr

Tuğçe TEMEL
Istanbul Technical University
temel21@itu.edu.tr

Abstract

Idiomatic expressions pose a fundamental challenge for multimodal understanding due to their non-compositional semantics, while pretrained vision–language models tend to overrely on literal visual alignments. We address this issue in the context of the AdMIRe 2.0 multimodal idiomatic image ranking task (Arslan et al., 2026) by introducing CARIM (Category-Aware Reasoning for Idiomatic Multimodality), an inference-time framework that injects structured semantic reasoning without end-to-end retraining.Experiments on the official Codabench leaderboard demonstrate that CARIM achieves competitive Top-1 Accuracy and nDCG across multiple languages. Additional post-competition evaluation on the released test annotations further shows that CARIM maintains robust multilingual performance, highlighting the effectiveness of inference-time category-aware reasoning for multimodal idiomatic grounding.

1 Introduction

Idiomatic expressions are pervasive in natural language and pose a fundamental challenge for natural language processing due to their non-compositional semantics, where the figurative meaning cannot be inferred from the meanings of individual words. Correctly interpreting idioms is essential for downstream tasks such as machine translation, visual-language understanding, and language grounding.

While recent contextualized language models have improved idiom detection and interpretation by leveraging surrounding context, they remain largely text-centric and often struggle to distinguish figurative meaning from literal interpretations in ambiguous settings(Shwartz and Dagan, 2019; Garcia and García-Serrano, 2018). This limitation becomes more evident in multimodal scenarios, where models must associate an idiomatic expression with a visual representation that reflects its figurative meaning rather than its literal constituents (Liu et al., 2022; Ma et al., 2023).

Motivated by the strong imagery evoked by idiomatic language, recent work has explored grounding idioms in visual space using vision-language models (Liu et al., 2022). However, models pretrained on large-scale literal image–text pairs tend to align idioms with literal visual concepts, leading to incorrect or ambiguous mappings. This highlights the need for learning objectives that explicitly separate idiomatic and literal meanings across modalities (Ma et al., 2023).

In this work, we study idiom understanding from a multimodal perspective and formulate idiom-image association as an image-text ranking problem and our contributions are:

- We introduce CARIM, an inference-time category-aware framework for multimodal idiom–image ranking that separates literal and idiomatic visual evidence without retraining.
- We propose a structured category model and category-conditioned ranking rules that penalize misleading literal visual evidence under idiomatic usage.
- We evaluate CARIM on the AdMIRe 2.0 shared task and through a post-competition multilingual analysis, demonstrating robust performance across diverse languages.

2 Related Work

Idiomatic expressions pose a challenge for NLP due to their non-compositional semantics (Sag et al., 2002). Early work relied on lexical resources and rule-based identification of multiword expressions (Baldwin and Kim, 2010), but these approaches struggled with contextual variability.

Distributional methods showed that standard word embeddings are biased toward literal meanings and insufficient for modeling idioms (Salehi

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 160–164
March 28, 2026 ©2026 Association for Computational Linguistics

et al., 2015). Contextualized language models such as BERT significantly improved idiom understanding by leveraging context, commonly framing the task as idiom detection or literal–figurative classification (Shwartz and Dagan, 2019; Garcia and García-Serrano, 2018). Several studies further explored idiom representation learning and paraphrase-based supervision (King and Cook, 2016; Peng and Feldman, 2018).

More recently, idiom learning has been extended to multimodal settings, motivated by the strong imagery associated with figurative language. Vision–language models have been used to align idiomatic expressions with figurative images while contrasting them against literal visual interpretations (Liu et al., 2022; Ma et al., 2023). These works typically formulate the problem as image–text ranking or contrastive learning, demonstrating that visual grounding provides complementary semantic signals beyond text-only models.

3 Methodology

In this section, we introduce CARIM, our inference-time category-aware reasoning model for multimodal idiomatic ranking.

3.1 CARIM Overview

At inference time, our system performs category-aware image ranking conditioned on the contextual usage of a compound expression. While the overall framework may incorporate learned components in earlier stages, the method described in this section operates exclusively at **prediction time** and does not update model parameters. Instead, it applies a structured reasoning procedure that encodes domain knowledge about idiomatic image categories directly into the inference process.

Given a compound expression c, a context sentence s, and a set of candidate images $\mathcal{I} = \{I_1, \ldots, I_5\}$ (with optional captions), the inference module outputs (i) a ranked list over the images and (ii) a compound-type label indicating whether c is used literally or idiomatically.

Inference Decomposition: The inference procedure is decomposed into two sequential steps:

1. **Compound-type inference**, which determines whether the compound is used literally or idiomatically in context.
2. **Category-aware image ranking**, which applies expert-defined ranking rules conditioned on the inferred sentence type.

This decomposition explicitly models the dependency between contextual meaning and visual relevance, reducing ambiguity compared to single-step ranking.

Sentence-Type Inference: In the first step, the model infers a sentence-type label

$$y \in \{\textsc{Literal}, \textsc{Idiomatic}\}$$

for the compound expression c as used in sentence s. A Literal usage denotes reference to the physical objects or properties described by the component words of c, whereas an Idiomatic usage denotes a figurative meaning that cannot be derived compositionally from the literal referents.

This decision is produced at inference time via a constrained prediction that conditions on c and s, and serves as a control signal for the subsequent ranking step.

Category Model: The ranking strategy is grounded in a category model that reflects a consistent structure in the candidate image sets. Each image is assumed to fall into one of five semantic categories relative to the compound expression:

1. **Literal (L):** depicts the literal referents of the component words of c.
2. **Literal-related (LR):** partially matches the literal meaning.
3. **Idiomatic (I):** directly depicts the figurative meaning of c.
4. **Idiomatic-related (IR):** loosely supports the figurative meaning.
5. **Distractor (D):** superficially related but incorrect for the intended meaning.

Literal object detection: To support category assignment, the compound is decomposed into its component words

$$c \rightarrow \{w_1, \ldots, w_m\}.$$

At inference time, the Chatgpt 5.1 evaluates whether an image contains salient visual evidence corresponding to these literal referents. This signal is used differently depending on the inferred sentence type.

Category-Aware Ranking Rules: The final ranking is produced by applying deterministic rules conditioned on y.

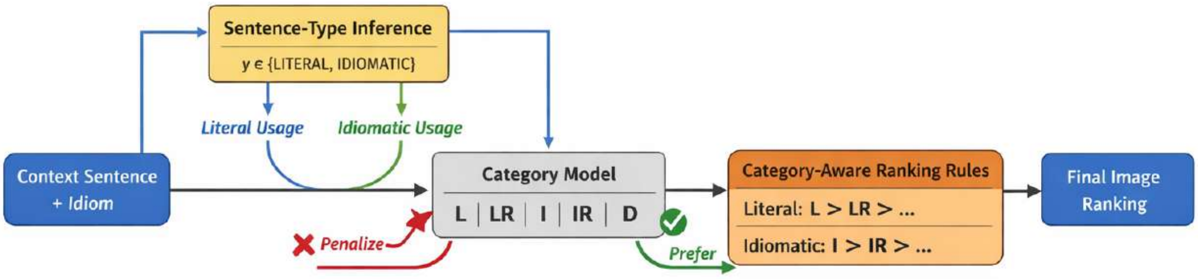

Figure 1: Overview of the proposed inference-time category-aware ranking framework, CARIM. Given a context sentence and a compound expression, the system first infers whether the expression is used literally or idiomatically. Based on this decision, candidate images are categorized into semantic types (L, LR, I, IR, D) and ranked using category-conditioned rules that penalize misleading literal visual evidence under idiomatic usage.

Literal usage. If $y = \textsc{Literal}$, images are ranked according to:

$$\mathbf{L} \succ \mathbf{LR} \succ \mathbf{IR} \succ \mathbf{I} \succ \mathbf{D}.$$

Images depicting the literal referents are preferred, while purely idiomatic depictions are down-ranked.

Idiomatic usage: If $y = \textsc{Idiomatic}$, the ranking order is:

$$\mathbf{I} \succ \mathbf{IR} \succ \mathbf{LR} \succ \mathbf{L} \succ \mathbf{D}.$$

Crucially, images with strong literal evidence (**L**) are explicitly penalized under idiomatic usage, as they typically correspond to incorrect surface-level interpretations.

Inference-Time Composition: The inference module combines the two steps as:

$$y = f_{\text{infer}}(c, s), \qquad \pi = f_{\text{rank}}(c, s, \mathcal{I} \mid y),$$

where f_{infer} denotes sentence-type inference and f_{rank} denotes category-aware ranking. Although parameter learning may occur elsewhere in the system, this module performs structured reasoning exclusively at inference time, interpretable and context-sensitive image ranking.

Illustrative Examples: For the compound *bad apple*, literal usage prioritizes images depicting a spoiled apple, whereas idiomatic usage prioritizes images depicting a corrupting individual and penalizes literal apples. Similarly, for *green fingers*, idiomatic usage favors gardening-related imagery while down-ranking images that merely depict green-colored fingers.

4 Results

We evaluate CARIM using the official Codabench leaderboard of the AdMIRe 2.0 shared task. Results are reported in terms of Top-1 Accuracy and nDCG, following the task evaluation protocol. Since our submission was evaluated on a subset of languages, we report results for Igbo (IG), Kazakh (KK), Turkish (TR), and Uzbek (UZ), together with the average across these languages. Full leaderboard results are shown in Table 1.

Overall, our system,CARIM achieves competitive performance across all evaluated languages without relying on additional end-to-end training or task-specific fine-tuning. This suggests that structured inference-time reasoning can be effective for multimodal idiomatic ranking even when operating on frozen backbone models.

Following the AdMIRe 2.0 shared task, we conducted a post-competition evaluation using the released test annotations to assess multilingual generalization. As shown in Table 2, CARIM achieves a macro-average Top-1 Accuracy of 57.2% and an average nDCG of %83.5 across all evaluated languages, without additional fine-tuning. These results demonstrate consistent multilingual performance of inference-time category-aware reasoning beyond the competitive setting.

4.1 Language-Specific Performance

As shown in Table 1, the proposed method performs strongest on Kazakh (KK), achieving its highest Top-1 Accuracy and nDCG among the evaluated languages. This indicates that the category-aware ranking strategy is particularly effective in settings where literal visual cues act as strong distractors, and where explicit penalization of literal imagery under idiomatic usage provides clearer

Participant	AVG		IG		KK		TR		UZ	
	Acc.	nDCG	Acc.	nDCG	Acc.	nDCG	Acc.	nDCG	Acc.	nDCG
ozgeumut	0.6075	0.86	0.56	0.84	0.70	0.89	0.65	0.88	0.52	0.83
ITUNLP	0.56	0.8375	0.43	0.78	0.61	0.84	0.68	0.90	0.52	0.83
kkkomurcu	**0.48**	**0.8025**	**0.41**	**0.76**	**0.56**	**0.84**	**0.54**	**0.83**	**0.41**	**0.78**
davidcotiga	0.4675	0.78	0.39	0.74	0.53	0.80	0.62	0.84	0.33	0.74
tiberiucarp	0.445	0.78	0.37	0.74	0.51	0.81	0.48	0.80	0.42	0.77
bilenbaris	0.3575	0.735	0.30	0.73	0.40	0.74	0.31	0.72	0.42	0.75
nikoniko	0.3225	0.7225	0.33	0.73	0.33	0.74	0.34	0.71	0.29	0.71
utkucolakitu	0.275	0.7025	0.22	0.69	0.28	0.71	0.29	0.70	0.31	0.71
akkurt_buzlu	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 1: Official Codabench leaderboard results for AdMIRe 2.0. Scores are reported as Top-1 Accuracy and nDCG, averaged across four evaluated languages (AVG) and for individual languages IG, KK, TR, and UZ .

Language	Acc.	nDCG
ZH	0.497	0.792
KA	0.531	0.808
EL	0.639	0.868
IG	0.409	0.761
KK	0.564	0.838
NO	0.644	0.873
PT-BR	0.864	0.942
PT-PT	0.636	0.866
RU	0.714	0.893
SR	0.579	0.826
SK	0.556	0.833
SL	0.708	0.887
ES-EC	0.292	0.727
TR	0.538	0.833
UZ	0.408	0.782
Average	**0.572**	**0.835**

Table 2: Top-1 Accuracy and nDCG scores of **CARIM** across languages on the AdMIRe 2.0 evaluation set (Torunoğlu-Selamet et al., 2026). The last row reports the macro-average over all languages.

separation.

Performance on TR, IG, and UZ language datsets remain competitive relative to other submissions on the leaderboard. While absolute scores vary across languages, the consistency between Top-1 Accuracy and nDCG suggests stable ranking behavior rather than isolated correct predictions. This aligns with the design goal of producing coherent, context-sensitive rankings instead of optimizing solely for the top-ranked image.

Notably, these results are obtained without modifying model parameters at inference time. In contrast to approaches that rely on contrastive retraining or language-specific adaptation, our method incorporates task knowledge through an explicit category model and deterministic ranking rules. This allows the system to systematically down-rank literal visual interpretations when idiomatic usage is inferred, addressing a common failure mode of pretrained vision-language models.

5 Conclusion

We presented an inference-time, category-aware ranking approach,CARIM for multimodal idiomatic understanding, motivated by the tendency of pretrained vision–language models to favor literal visual alignments. By explicitly decomposing inference into sentence-type prediction and category-conditioned image ranking, our method injects structured task knowledge without requiring end-to-end retraining or parameter updates. This design enables interpretable control over ranking behavior and directly addresses literal bias in idiom–image association.

Evaluation on the AdMIRe 2.0 Codabench leaderboard demonstrates that this lightweight reasoning framework achieves competitive performance across multiple languages, despite operating on frozen backbone models and limited supervision. The results suggest that explicit category reasoning at inference time can serve as an effective complement to learned multimodal representations, particularly in figurative language settings where surface-level visual similarity is misleading.

Acknowledgments

This work was conducted as part of the course BLG505-NLP at ITU. The authors would like to thank Prof. Gülşen Eryiğit for helpful comments and suggestions.

References

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.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing*.

Marcos Garcia and Ana García-Serrano. 2018. Towards deep learning of idiomaticity. In *Proceedings of EMNLP*.

Milton King and Paul Cook. 2016. Verifying semantic compositionality of multiword expressions. In *Proceedings of ACL*.

Fangyu Liu, Xiaowei Zhai, and Joyce Chai. 2022. Multimodal idiom understanding. In *Proceedings of ACL*.

Yukun Ma, Xiang Chen, and Zhiyuan Liu. 2023. Figurative language grounding with vision–language models. In *Proceedings of EMNLP*.

Jing Peng and Anna Feldman. 2018. Neural network models for idiom detection. In *Proceedings of LREC*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of NAACL*.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. In *Proceedings of ACL*.

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.