VisAffect at MWE-2026 AdMIRe 2: IMMCAN Idiom Multimodal Cross-Attention Network

Barış Bilen[1,2]*, Ali Azmoudeh[1]*, Hazım Kemal Ekenel[1,3], Hatice Köse[2]

[1]Dept. of Computer Engineering, Istanbul Technical University,
[2]Dept. of Artificial Intelligence and Data Science, Istanbul Technical University,
[3]Division of Engineering, NYU Abu Dhabi
{bilenb20,azmoudeh22,ekenel,hatice.kose}@itu.edu.tr
he2244@nyu.edu

Abstract

We address AdMIRe 2.0, a static image ranking task where a sentence containing a potentially idiomatic expression is paired with five image–caption candidates, and the goal is to rank the candidates by semantic compatibility with the intended idiomatic or literal meaning. We propose IMMCAN, which keeps XLM-R and Jina-CLIP-v2 frozen and learns a lightweight two-stage cross-attention fusion, caption–image grounding followed by idiom-to-multimodal conditioning, to predict a compatibility score per candidate. We also evaluate caption-only augmentation via back-translation and synonym substitution, and compare regression and rank-class formulations. On AdMIRe 1.0, text-only achieves higher test top-image accuracy than VLM-grounded modeling. In contrast, on AdMIRe 2.0 zero-shot, adding visual patch grounding improves both accuracy and NDCG indicating better cross-lingual ranking transfer. Github.com/AliAZ98/IMMCAN

1 Introduction

Idiomatic expressions are central to natural language, yet their meanings are often non-compositional and cannot be inferred from the literal meanings of their component words, which makes them difficult for both human learners and natural language processing (NLP) systems, especially in multilingual and cross-lingual settings (Villavicencio et al., 2005; Zeng and Bhat, 2022; Madabushi et al., 2022). The AdMIRe 2.0 shared task (Arslan et al., 2026; Torunoğlu-Selamet et al., 2026) addresses this in a multimodal setup: given a context sentence containing a potentially idiomatic expression (PIE) and five candidate images with captions, systems must rank the images by how well they reflect the idiomatic or literal meaning intended in that sentence, with supervision provided as a relative ordering rather than a single gold image, capturing the graded nature of idiom–image compatibility (Pickard et al., 2025). This setting is challenging because idiomaticity is highly context-dependent, i.e., the same expression can be idiomatic in one sentence and literal in another (Madabushi et al., 2022; Haagsma et al., 2020). The task is multilingual and largely zero-shot beyond English, requiring transferable cross-lingual representations without language-specific supervision (Conneau et al., 2019; Madabushi et al., 2022). In addition, models must integrate textual and visual information. They align the sentence–idiom meaning with both the image content and the accompanying captions. The system then ranks fine-grained candidates (idiomatic, literal, related, and distractor). Because supervised data is limited, the risk of overfitting increases. Therefore, effective use of pretrained language and vision–language models is essential (Radford et al., 2021).

In this work, we present a multimodal system for AdMIRe 2.0 with three core contributions: (i) a multilingual XLM-RoBERTa idiomaticity detector trained on MAGPIE dataset where we tested on AdMIRe 1.0 English and Portuguese, (ii) we propose the Idiom Multimodal Cross-Attention Network (IMMCAN), which fuses frozen XLM-R idiom embeddings and frozen Jina-CLIP-v2 image–caption features via two-stage cross-attention to score and rank candidates, and (iii) caption-only augmentation (back-translation and synonym substitution), with zero-shot results showing consistent improvements of the multimodal model over text-only baselines in top-image accuracy and NDCG.

2 Related Work

Pickard et al. (Pickard et al., 2025) introduced the AdMIRe (SemEval-2025) shared task, which provides a multilingual dataset of English and Portuguese sentences labeled as idiomatic or literal and paired with multiple candidate images. Par-

*These authors contributed equally.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 149–153
March 28, 2026 ©2026 Association for Computational Linguistics

ticipants are asked to rank the images according to how well they match the intended idiomatic or literal meaning of the sentence, a setting that is very close to our task. Within this framework, Pan et al. (Pan et al., 2025) propose a multimodal idiom ranking system that combines pretrained text encoders such as BERT or XLM-RoBERTa with a Vision Transformer for images, and then trains a regression model to rank images by semantic alignment with the idiomatic context. Khatoon et al. (Khatoon et al., 2025) also participate in AdMIRe and focus on vision–language modeling; they introduce an "Idiom Visual Understanding Dataset" and show that a CLIP-based model, which jointly embeds images and text, clearly outperforms text-only baselines for ranking images by idiomatic meaning. Their findings support the idea that integrating visual information with text can improve idiom interpretation, which aligns with our multimodal design.

Beyond AdMIRe, several works address idiomaticity from a mainly textual perspective. Chu et al. (Chu et al., 2022) treat idiom identification as a sentence-level classification problem and fine-tune a pretrained language model on sentences containing target multiword expressions, deciding whether each expression is used idiomatically or literally. Oh (Oh, 2022) proposes NEAMER, a model that exploits the similarity between idioms and named entities; which uses XLM-RoBERTa enriched with additional surface features and transfer learning from a named entity recognition (NER) task, and achieves strong multilingual idiom classification performance. At the benchmark level, Madabushi et al. (Madabushi et al., 2022) introduce SemEval-2022 Task 2, a multilingual idiomaticity task that includes English, Portuguese, and Galician data, with subtasks on idiom detection and semantic similarity in context. This line of work shows that multilingual encoders such as XLM-R can generalize idiom recognition across languages and provides an important foundation for our use of XLM-R-based components in a multimodal idiom-understanding setting.

3 System Overview

This work targets the AdMIRe 2.0 multimodal idiom task by scoring how well each of five image–caption options matches the idiomatic or literal meaning of a sentence that contains a compound idiom. Each dataset example includes the idiom, a sentence labeled as idiomatic or literal, and an expected ranking of the five images; the label is based on their relative semantic fit to the sentence meaning, not on visual correctness alone, so our dataloader turns each example into five text–image pairs and assigns numeric targets from this order. Our system uses a multimodal model with two frozen encoders: XLM-R produces contextual embeddings for the idiom in its sentence, and Jina-CLIP-v2 encodes each image with its caption into text tokens and visual patch features. We combine these signals with a two-step attention design: TextImageCoAttention links caption text to image features, and IMMCAN updates the idiom representation using this fused multimodal information. Finally, we pool the idiom-aware representation and use a regression or classification head to output the prediction per image, allowing the model to learn the fine-grained relationship between idiom meaning and the visual–textual candidates.

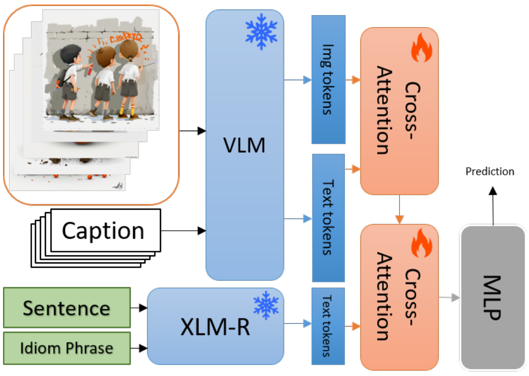

Figure 1: Overall representation of proposed IMMCAN

3.1 VLM Model

Our vision–language module uses Jina-CLIP-v2 (Koukounas et al., 2024), a multilingual dual-encoder that learns shared text–image representations by pairing a Jina XLM-RoBERTa text encoder (561M) with an EVA02-L/14 ViT vision encoder (304M), totaling 865M parameters. It supports 89 languages and long text inputs (up to 8,192 tokens) and processes images up to 512 × 512 using rotary positional embeddings and efficient attention. Both encoders produce 1024-d embeddings trained with a multi-task contrastive objective over text–text and text–image pairs, and we use the resulting token-level and global features in a frozen manner.

3.2 Idiomaticity Detection

A binary XLM-RoBERTa (Conneau et al., 2019) classifier has been trained to predict whether an expression is used idiomatically or literally, choosing XLM-R because its multilingual representations support zero-shot transfer to other languages even when training is done only on English. The model is trained on the MAGPIE dataset (Haagsma et al., 2020) using sentence–idiom pairs, where the full sentence and the target phrase are provided as separate segments to explicitly model idiom–context interaction. After training, we evaluate on the AdMIRe 1.0 English, Portuguese, and combined test sets (Table 1), and the strong Portuguese accuracy without additional supervision indicates meaningful zero-shot generalization for idiomaticity detection.

Lang.	Accuracy	Macro F1	Weighted F1
EN	0.7647	0.7607	0.7632
PT	0.6000	0.5746	0.5908
EN+PT	0.7077	0.7077	0.7077

Table 1: Model accuracy for English, Portuguese and combined on AdMIRe 1.0 dataset

3.3 IMMCAN

Figure 1 represents the proposed IMMCAN, which combines a pretrained language model and a pretrained vision–language model to build a multimodal representation for idiom understanding. The IdiomEmbedder (XLM-R) and JinaCLIPEmbedder are fully frozen and used only as feature extractors: XLM-R produces contextual idiom–sentence token embeddings, while the VLM provides caption tokens and image patch tokens from its transformer. These frozen features are fed to a small set of trainable modules, allowing us to leverage strong language and vision priors while updating few parameters. Fusion is performed with a two-step attention pipeline: TextImageCoAttention cross-attends between caption tokens and image tokens to form grounded multimodal features, which are pooled into global text and image vectors and concatenated. IMMCAN then conditions idiom tokens on this multimodal vector. Finally, a masked-mean pool produces a fixed-size embedding that an MLP head predicts the rank.

3.4 Data Augmentation

Data augmentation is used to improve robustness and reduce overfitting, but to keep idioms consistent we do not modify the idioms or their context sentences. Instead, we augment only the image captions using two methods, back-translation and synonym substitution.

Model Setup	Val	Test
TT-Reg-Base	0.4400	0.2143
TT-Reg-Aug	0.3600	0.3929
TT-Cla-Base	0.4800	0.4643
TT-Cla-Aug	**0.6400**	**0.6786**
VTT-Reg-Base	**0.4000**	**0.2857**
VTT-Reg-Aug	0.1200	0.1429
VTT-Cla-Base	0.3600	0.2500
VTT-Cla-Aug	0.0800	0.0714

Table 2: Validation and test performance of different models

For back-translation, each example's five captions are translated to an intermediate language and then translated back, using English→French→English for English captions and Portuguese→English→Portuguese for Portuguese captions, which varies syntax while keeping meaning. For synonym substitution, we rewrite captions by replacing randomly selected words at about a 15% rate, selecting candidate words by simple token filtering, e.g., minimum length, and retrieving English and Portuguese synonyms from WordNet.

4 Experimental Setup

AdMIRe 2.0 is a static image ranking task where each example provides a context sentence with a potentially idiomatic expression (PIE) and five image–caption candidates, and the system must rank the images by how well they match the intended meaning. For supervised learning, we use AdMIRe 1 (English and Portuguese) and expand each instance into five text–image candidates so the model outputs one score per candidate and recovers a full ranking at inference. Since labels are given as an ordering, we map them to either regression targets $\{1.0, 0.75, 0.50, 0.25, 0.0\}$ or classification labels $\{0, 1, 2, 3, 4\}$, and we evaluate with top image accuracy and NDCG@5 using relevance weights $[3, 1, 0, 0, 0]$. In the AdMiRe 1.0 dataset, there are 70 training, 15 validation, and 15 test samples in total. We do not use any AdMIRe 2.0 training or validation labels. All hyperparameters are tuned on AdMIRe 1.0 only, and AdMIRe 2.0 is used strictly for zero-shot evaluation.

We evaluate two variants under the same fusion design. The Vision-Text-Text (VTT) setting uses both VLM caption tokens and image patch tokens together with the idiom text stream, while Text-Text (TT) removes image patches and applies cross-attention only between caption tokens and idiom tokens to isolate the effect of visual grounding. For both settings, we compare a regression head

Lang.	**TT-based**				**VTT-based**			
	Reg-Base	Reg-Aug	Cla-Base	Cla-Aug	Reg-Base	Reg-Aug	Cla-Base	Cla-Aug
KA	0.283 (0.696)	0.327 (0.703)	0.310 (0.705)	0.248 (0.669)	0.336 (0.717)	0.106 (0.580)	**0.407 (0.747)**	0.381 (0.743)
PT-BR	0.298 (0.699)	0.303 (0.703)	0.307 (0.712)	0.237 (0.672)	0.298 (0.706)	0.044 (0.571)	0.342 (0.724)	**0.395 (0.738)**
UZ	0.325 (0.712)	0.250 (0.692)	0.258 (0.691)	0.233 (0.672)	**0.417 (0.749)**	0.117 (0.573)	0.342 (0.723)	0.383 (0.729)
ES-EC	0.167 (0.628)	0.167 (0.626)	0.229 (0.658)	0.188 (0.642)	0.333 (0.695)	0.188 (0.636)	**0.333 (0.727)**	0.333 (0.716)
NO	0.262 (0.689)	**0.302 (0.707)**	0.257 (0.705)	0.198 (0.665)	0.287 (0.698)	0.129 (0.606)	0.287 (0.705)	0.277 (0.706)
IG	0.322 (0.742)	0.374 (0.750)	**0.383 (0.739)**	0.296 (0.695)	0.296 (0.725)	0.035 (0.537)	0.365 (0.752)	0.339 (0.748)
SK	0.305 (0.712)	0.278 (0.704)	0.278 (0.706)	0.192 (0.656)	0.338 (0.723)	0.093 (0.579)	**0.411 (0.742)**	0.325 (0.716)
TR	**0.313 (0.714)**	0.308 (0.719)	0.269 (0.702)	0.264 (0.686)	0.308 (0.720)	0.099 (0.576)	0.308 (0.725)	0.280 (0.711)
RU	0.336 (0.721)	0.350 (0.725)	0.386 (0.746)	0.279 (0.673)	0.364 (0.732)	0.057 (0.572)	**0.400 (0.745)**	0.379 (0.731)
EL	0.250 (0.687)	0.245 (0.689)	0.260 (0.691)	0.236 (0.671)	0.274 (0.685)	0.120 (0.604)	**0.327 (0.712)**	0.269 (0.693)
SR	0.256 (0.695)	0.240 (0.690)	0.284 (0.703)	0.196 (0.666)	0.355 (0.718)	0.094 (0.583)	0.333 (0.722)	**0.358 (0.717)**
ZH	0.246 (0.678)	0.201 (0.661)	0.246 (0.685)	0.190 (0.640)	0.302 (0.714)	0.067 (0.580)	**0.363 (0.735)**	0.324 (0.719)
PT-PT	0.255 (0.691)	0.277 (0.694)	0.268 (0.691)	0.236 (0.674)	0.300 (0.706)	0.068 (0.576)	0.277 (0.696)	**0.318 (0.712)**
SL	0.246 (0.686)	0.254 (0.687)	0.300 (0.725)	0.238 (0.668)	0.283 (0.700)	0.083 (0.579)	**0.350 (0.726)**	0.329 (0.715)
KK	0.282 (0.706)	0.289 (0.709)	0.282 (0.705)	0.231 (0.671)	0.397 (0.739)	0.077 (0.565)	0.417 (0.752)	**0.442 (0.770)**
ALL	0.276 (0.697)	0.278 (0.697)	0.288 (0.704)	0.231 (0.668)	0.326 (0.715)	0.091 (0.578)	**0.350 (0.727)**	0.336 (0.721)

Table 3: AdMIRe Subtask A zero-shot results across TT-based and VTT-based models. Each cell reports Acc (NDCG).

(Reg) that predicts a scalar compatibility score and a classification head (Cla) that predicts rank classes, and we test two data conditions, Base captions and Aug captions generated with the method in Section 3.4.

For training, we use SmoothL1Loss with $\beta = 1.5$ for regression to reduce sensitivity to large errors, label-smoothed cross-entropy with $\epsilon = 0.3$ for VTT classification to avoid overconfident rank predictions, and a ListNet-style KL loss for TT to learn rankings in a list-wise manner. All models are trained with SGD (learning rate 0.003, weight decay 0.001), batch size 4, and an exponential learning-rate scheduler. Experiments are run on an NVIDIA RTX 3080 Ti GPU.

5 Results

Performance on AdMIRe 1.0 and multilingual zero-shot transfer results on AdMIRe 2.0 have been summarized here.

5.1 Test Results

The AdMIRe 1.0 validation/test split (see Table 2) indicated that the most reliable improvements have been obtained with the text-only classification. The highest test score has been achieved by TT-Cla-Aug implying that caption augmentation has been beneficial when a discrete rank-class target has been learned. For the multimodal variant, the best test result has been observed with VTT-Reg-Base, and performance has dropped sharply for augmented VTT configurations, which has shown that the current caption augmentation has not remained compatible with the visual-token fusion.

5.2 Zero-Shot Results

Zero-shot transfer results across AdMIRe 2.0 languages are summarized in Table 3: TT variants remain unchanged across settings, suggesting that text-only pairing provides limited cross-lingual transfer in our current setup, whereas VLM-based variants yield clearer gains across multiple languages, e.g., KA, UZ, RU, KK, indicating that explicit visual grounding improves both top-image selection and overall ranking quality. Caption augmentation shows mixed effects in zero-shot transfer, with Cla-Aug staying competitive and improving several languages, while Reg-Aug consistently degrades, which suggests that paraphrastic caption changes may add noise that harms regression-style scoring. Finally, we find that top image accuracy can be low while NDCG@5 remains high. This indicates that the model often places a highly relevant image near the top, even when it does not rank the single best image first. Because NDCG rewards near-correct ordering under graded relevance, it is less sensitive than top accuracy to swaps between the best and second-best candidates.

6 Conclusion

We presented IMMCAN for multimodal idiom understanding in AdMIRe 2.0, combining a zero-shot XLM-R idiomaticity detector with a lightweight fusion module over frozen XLM-R and Jina-CLIP-v2 features. Our results indicate that explicit visual grounding yields clearer gains in zero-shot transfer than text-only pairing, improving both top-image selection and NDCG@5 across several languages. We also found that caption augmentation has mixed effects: it tends to help when learning discrete rank classes, but it can degrade regression-based scoring, likely due to added paraphrastic noise. Future work includes designing augmentation methods that preserve fine-grained ranking signals, exploring stronger list-wise objectives for multimodal settings, and training or adapting the fusion layers with more multilingual supervision to further improve cross-lingual generalization.

Acknowledgment

This work was partially funded by the ITU-Turkcell research scholarship.

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.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. Hit at semeval-2022 task 2: Pretrained language model for idioms detection. *arXiv preprint arXiv:2204.06145*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arxiv. *arXiv preprint arXiv:1911.02116*, 10.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

Maira Khatoon, Arooj Kiyani, Tehmina Farid, and Sadaf Abdul-Rauf. 2025. Fjwu_squad at semeval-2025 task 1: An idiom visual understanding dataset for idiom learning. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1759–1765.

Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

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. *arXiv preprint arXiv:2204.10050*.

Min Sik Oh. 2022. kpfriends at semeval-2022 task 2: Neamer–named entity augmented multi-word expression recognizer. *arXiv preprint arXiv:2204.08102*.

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Umuteam at semeval-2025 task 1: Leveraging multimodal and large language model for identifying and ranking idiomatic expressions. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 743–749.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire–advancing multimodal idiomaticity representation. *arXiv preprint arXiv:2503.15358*.

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. In *International conference on machine learning*, pages 8748–8763. PmLR.

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.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut.

Ziheng Zeng and Suma Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.